Analysis of US Weather / Tempo-Spatial Global Warming Analysis

Author: Kevin Duran
Department of Information Systems, California State University Los Angeles
CIS 4560-01 Intro to Big Data
kduran26@calstatela.edu

Abstract: This research will target to analyze a U.S. weather dataset from 2011-2020 utilizing big data framework with Apache Hive and Hadoop in order to examine if Global Warming occurred within this data. By looking into this decade of data and over 1 million weather recordings from U.S. weather stations, this data was used to evaluate trends in precipitation and temperature. Analysis will maintain focus on temperature patterns, how precipitation is distributed, along with diurnal temperature variability over time.Visualizations and analysis of this focus was executed using Excel Power Map and Tableau. Results found that diurnal temperature demonstrated narrowing in between the years of 2012 and 2019, this shows temperatures in the night are rising which aligns with climate change science findings and related works. Another finding was increased precipitation variability which is another indicator of climate change.

1. Introduction

Being able to understand climate behavior over time is essential, as it remains a top concern as a challenge that's globally recognized. This is because of precipitation patterns and rising temperatures actually impacting different regions especially across the United States. This project will investigate possible signs of climate change and global warming utilizing U.S. weather station data covering 2011-2020. Data will be processed using Apache Hadoop and Hive to prepare it for analysis. The goal and focus will be on temperature and precipitation patterns in order to look into how these attributes fluctuate over time. The importance of this analysis and research stems from changing weather patterns along with rising temperatures helping identify climate change occurrence. This study will aim to answer if the observed weather in the U.S. from this decade exhibits signs of global warming and align with scientific consensus on how climate change is analyzed.

2. Related Work

With the global recognition this topic has received, there are lots of studies that help provide foundations of methods and evidence to help suggest and support the effect of climate change. Beginning with the National Centers for Environmental Information(NOAA) within their 2022 State of the Climate Report, they discuss how this time period of 2010-2022, including all but 2011, 2012, 2013, were the ten warmest years.[1] They go into this further by stating how “since global records began in 1880,...the 10 warmest years in the 143-year record have all occurred since 2010”. This aligns with my research focus, as the range I’ll be focused on is within this range measured to see if additional signals of this occurrence can be covered.

In another study this time done in 2021, Wang et al. investigate how there is a difference in seasons in the aspect of length.[2] Stating “Climate change is altering the timing and length of the seasons.” They go on essentially saying how this is evident through summers having increased length and lasting longer and even winters becoming shorter as well. They even identified this was most pronounced within the Northern Hemisphere.This supports this studies research as it looks into seasonal shifts observed through temporal analysis looking at temperature over time.

The Fourth National Climate Assessment from 2018 dives further into this helping support these findings by indicating variability along with extreme temperatures increasing aligning with the national observed warming trend over time.[3] At the result of this, there were increased shifts of precipitation intensity and even frequent heat waves observed. They discuss how, “Heavy rainfall events are becoming more common, especially in central and eastern U.S.”.This assessment aligns with my approach of looking into temperature ranges and precipitation trends for assessing global warming.

Being able to utilize Hadoop was essential for completing this project. In a related work, Silvana Greca, Ingrid Shehi and Jonuz Nuhi provide a great foundation looking into weather utilizing Hadoop and Hive in order to study climate change.[4] They followed a similar approach but instead made use of a larger range from 1700 to 2013 studying the city of Durrës of the republic Albania. Within their studies, they found “stark changes in temperature over the past century” in which they were able to provide a trendline that showed steady increase from 1800 to 2013. By utilizing these tools conclusions were stated about extreme

[1] Related work was important to access as it developed the foundation to what’s important in this type of analysis.

weather events are increasing along with global temperatures.

3. Specifications

The dataset used in this project primarily looks at weather records from weather stations within the United States. The original dataset contained weather records dating from 1992-2021. The original dataset is 8.37 GB in size and is in CSV format. To be able to analyze effectively , a clean dataset was created by filtering data to focus on the years 2010-2020. We found that there was no valid data from 2010 for use so that is why the data was focused to 2011-2020. This Table 1 shows the dataset specification about the original dataset acquired from Kaggle and the cleaned dataset after data cleaning. Following these specifications, projects can be scaled possibly further with Big data datasets that are substantially much larger than the current size that was used for this study, for example with Greca et al who performed climate analysis across a broader range.[4] Doing this process was essential for being able to efficiently manage the data and prepare it for further analysis.

Table 1 Data Specification

Original Dataset size	8.37 GB CSV format
Cleaned Dataset size	1.88 GB CSV format

The table below shows the specifications for the we are using for our project process.

Table 2 H/W

Number of Nodes	5
Cluster Version	Hadoop 3.1.2
Memory Size	31 GB
Nodes	5
PC CPU	AMD Ryzen 7 7730U (8 Cores, 2.0GHz)

4. Implementation Flow Chart

To illustrate this implementation process, I used an implementation flow chart (Figure 1). The process starts with acquiring the zipped dataset through download from Nachiket Komad on Kaggle who was able to provide 1992-2021 weather data.[5] After this, the local terminal was accessed within the location of where the zipped weather folder was stored. Doing this, now the dataset can be uploaded as zipped format to the Hadoop Distributed File System (HDFS) which grants distributed storage along with access. Now that the data was in HDFS, it was unzipped then imported within the Hadoop Cluster. This allowed for the use of Hive to help create the cleaned table schema making use of Beeline. Having this cleaned table allowed for the cleaning and filtering of the range 2010-2020 to have specified range. After exporting the contents of the cleaned Hive table into my 'hdfs' directory. This allowed the creation of the final csv that I'll download to my local machine. Having this prepared csv on the local machine now allowed for the creation of visualizations and further analysis making use of both Excel Power Maps along with Tableau to look further into this dataset in the aspect of identifying climate change.

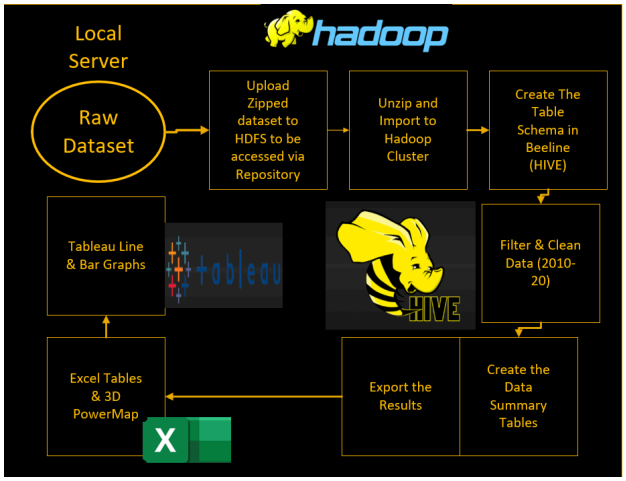

Figure 1: Implementation Flow Chart showing the process done in this project

[2] Elevation measure was kept within the data but not used.

5. Data Cleaning

During the data cleaning process there were specific decisions made to help provide support for the analysis. It begins with making use of Hive script casts 'CAST( )', this was because temperature and precipitation values were changed into floats and integers. This can be seen in 'CAST(tmax AS INT)' which helped convert the max

temperature field from string to numeric format since the measure of the attribute was tenths of a degree Celsius. Just for supporting temporal accuracy within the data, the date field was validated so that the data can maintain a 'MM/DD/YYYY' format restricted to 2010-2020 this was done with RLIKE '^[0-9]{1, 2}/[0-9]{1, 2}/20(1[0-9]|20$'. As previously stated, Nulls were addressed along with blank fields in the attributes I'll be looking to analyze further for climate change. For the derived columns, calculation methods were used to get **avg_temp_celsius** and **temp_range_celsius**. They were calculated using '(CAST(tmax AS FLOAT) + CAST(tmin AS FLOAT)) / 20.0', along with '(CAST(tmax AS FLOAT) - CAST(tmin AS FLOAT)) / 10.0' with the reasoning being as they help reflect the midpoint of these measures converted from tenths of Celsius. For **temp_range_celsius** this gave the diurnal range in degrees celsius. Categorical labels also had to be generated to help support the analysis, **temp_range_volatility** helped identify different ranges of volatility which is how much the temperature changes through the day. This gave labels of 'NORMAL RANGE'(5-15℃), 'STABLE' if temperature was less than 5℃, and 'EXTREME RANGE' if the temperature fluctuation exceeded 15℃ for the day which demonstrated high volatility. For **avg_temp_label, records were segmented into** 'COLD'(>5℃), 'MODERATE'(5℃-25℃), and 'HOT'(>25℃) which helped categorize the temperature into a scale. To keep in line to ensure data validity, 'ELSE 'UNKNOWN' was also used to help with this necessity, this just means that only records that were completed were labeled.

6. Analysis and Visualizations

In order to further analyze this data, visualizations were developed so that the data can be looked at on a closer scale. After changing the csv to an excel file now both Tableau and Excel can be used to fulfill different goals. Excel was used to help create 3D Power Map using Excel's 3D Power Map feature. This helped give an in depth look at temperature across the U.S. along with how precipitation is spread as well over the date range captured 2011-2020. As for Tableau, visuals such as line charts, and bar graphs were used to analyze global warming over time making use of comparisons between time periods of the data captured. For example, one method used was comparing early decade with late decade measures to see how these change over time. These proved to be very useful as they helped demonstrate a segmentation of the data that's being investigated across different measures and dimensions like, 'year', or even volatility classification. Using these tools and methods, allowed for global warming to be assessed and evaluated to help with understanding more about the possible effects of climate change that are happening within the U.S.

6.1 Excel 3D

The first visualization Figure 2 we made with our data was a 3D Excel map. The map helps visualize how the average temperature in celsius changes daily and over time. It gives a great look into what temperatures in different locations within the U.S. are like at different times of the day. We can use this to understand if certain factors of global warming are being visualized, such as hotter nights. It also helps with understanding how certain regions within the U.S. are being affected by this. For example, the Northern region demonstrates increased temperatures in the nights and mornings over time.

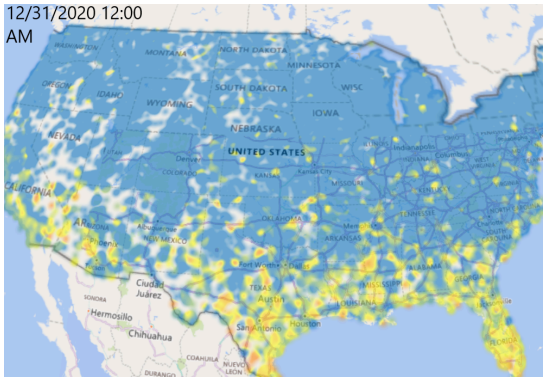

Figure 2: 3D Power Map visual demonstrates Average temp celsius throughout 2011-20

Tableau 6.2

Moving forward with analysis using Tableau, this tool was utilized to reach project goals such as, identifying volatility trends along with climate variability, diurnal temperature over time and identify other signals of climate change. These visualizations will aim to provide insight on detecting temporal trends of global warming, comparing early period measures with late period measures considering factors like precipitation accumulated at different temperature levels cold (blue), hot(orange), moderate(green).

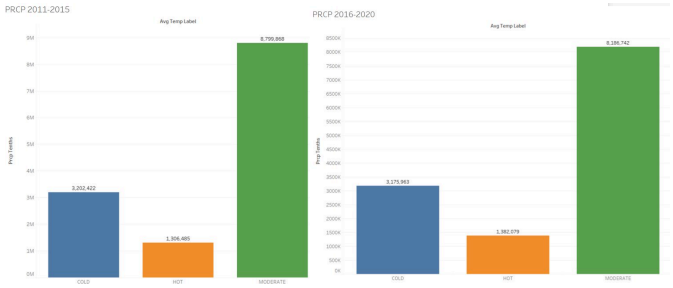

Figure 3: Precipitation early decade vs late decade

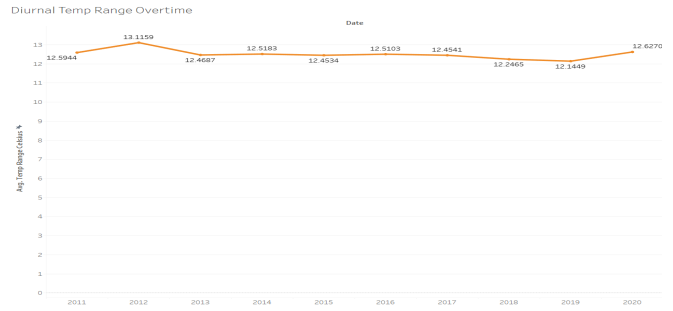

Figure 4: Diurnal temperature over time.

In Figure 3, we can see how 'MODERATE' zones accumulated the highest amount of precipitation across both comparisons. It can also be seen how during the earlier period, 'HOT' was at 1.306 million tenths, and 'COLD' 3.202 million tenths. Later in the decade, 'HOT' reached 1.382 million tenths and 'COLD' with 4.175 million tenths. Although these changes were modest and not drastic, the demonstration of 'HOT' precipitation movement increasing over time even when considering extreme weather shows a trend of warmer precipitation profiles in the weather records. For Figure 4, we look into diurnal temperature range (DTR) over time which is a key factor in measuring global warming as it's a greenhouse induced warming. The DTR is configured through the difference between daily max and min temperatures(tmax, tmin). This is a critical climate change metric to understand as this chart demonstrates the DTR in later years specifically 2017-2019 showing narrowing and the lowest DTR as well overall which can potentially indicate that the thermal retention is being decreased most likely overnight. Warming nights has been shown to reduce DTR as we see in the figure. Looking specifically at 2019 this was actually the lowest of all years within this range which can prove to be a significant indication of climate change. This aligns with the research done on this greenhouse induced warming identifying the increase of warming nights not just warming days showing less temperature fluctuation happening from night and day.

7. Conclusion

To summarize the findings, it has been concluded that:

1. DTR analysis is important to measure as this study demonstrated, decreasing values over time signify the fluctuation in temperature throughout the day.
2. Precipitation was shown to be increased in 'HOT' climates over time.
3. Volatility trends are showing to be flattening but 'EXTREME' range continues forward post 2016 signifying climate instability in these warming patterns.

Overall, with the trends of increasing 'HOT' classified precipitation, along with narrowing DTR, and extreme volatility trends demonstrate how the climate is changing. Temperatures are rising especially when looking at day and night temperature range fluctuations. This means this project was able to successfully detect global warming factors within the U.S. Covering 2011-2020. Utilizing related works to help determine a foundation of measures and metrics necessary to analyze to be able to consider climate change is occurring. With continued monitoring and increased research adoption we can continue measuring global warming in the future and expand this. I recommend further expanding the data quality and coverage along with also being able to look further at specific state metrics across the U.S.. This would help grant additional insights on why certain states may have increasing climate change.

References

[1] National Oceanic and Atmospheric Administration (NOAA). (2022). *Annual Global Climate Report – 2022*. National Centers for Environmental Information. https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202213

[2] X. Wang, R. Zhang, and Y. Huang, "Changing lengths of the four seasons by global warming," *Geophysical Research Letters*, vol. 48, no. 5, p. e2020GL091753, 2021. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020GL091753

[3] U.S. Global Change Research Program, *Fourth National Climate Assessment, Volume II: Impacts, Risks, and Adaptation in the United States*, 2018. https://nca2018.globalchange.gov/

[4] S. Greca, I. Shehi, and J. Nuhi, "Analyzing climate change impacts using big data Hadoop," in *Proc. 1st Int. Workshop on Big Data Applications and Principles (BDAP 2023)*, CEUR Workshop Proc., vol. 3402, 2023. http://ceur-ws.org/Vol-3402/paper03.pdf

[5]Dataset: N. Kamod, "Weather Dataset (US)," Kaggle, 2024. https://www.kaggle.com/datasets/nachiketkamod/weather-dataset-us/data

[1]Github Link: https://github.com/KevMD-23/CIS-4560-Weather