Climate Change Analysis: A MapReduce, Hive, and Trino Approach

Jonathan Amsalem
Courant Institute of
Mathematical Sciences
New York University NY,
USA

Apoorv Singh
Courant Institute of
Mathematical Sciences
New York University NY,
USA

Divya Rallapalli
Courant Institute of
Mathematical Sciences
New York University NY,
USA

Abstract

This paper presents a comprehensive analysis of climate change patterns using big data techniques. The focus is on temperature trends at the city, state, and country levels using the Climate Change Earth Surface Temperature dataset. The study leverages advanced data processing frameworks such as MapReduce, Hive, and Trino to uncover significant temperature changes, emphasizing the urgency of addressing climate change.

1 Introduction

Climate change has emerged as one of the most critical environmental challenges of the 21st century, affecting ecosystems, economies, and societies worldwide. As global temperatures continue to rise, understanding the patterns and factors that contribute to climate change has become essential for policymakers, scientists, and environmentalists.

This study aims to investigate temperature trends across three geographical resolutions: city, state, and country. Using historical temperature data sourced from Kaggle's Earth Surface Temperature dataset, we performed an extensive analysis utilizing big data frameworks such as MapReduce, Hive, and Trino. These platforms enabled us to process large-scale datasets, extract meaningful insights, and predict future temperature changes. The primary goal is to evaluate historical temperature variations, highlight regional warming patterns, and forecast possible climate scenarios for the coming decades. The project emphasizes the use of robust data processing methods to overcome challenges related to data inconsistencies, missing records, and computational complexity. Through regression modeling and predictive analytics, this work contributes to a broader understanding of climate dynamics and supports efforts to mitigate the impacts of climate change.

This paper details the methods used, the challenges encountered and the findings uncovered during the analysis. The results reveal consistent temperature increases across all studied regions, emphasizing the growing urgency of addressing climate change through data-driven insights and informed policy decisions.

Literature Survey

The increasing global concern about climate change has necessitated the integration of big data analytics into environmental research. This literature survey highlights key studies that contribute to understanding climate change through data-driven analysis.

Masson-Delmotte et al. (2021) provide a comprehensive scientific assessment of climate change in the IPCC report, emphasizing the critical role of big data in climate modeling and forecasting. They outline how large-scale environmental data support the development of global climate policies [2]. Tamiminia et al. (2020) explore the capabilities of Google Earth Engine in geo-big data applications, highlighting its effectiveness in climate change monitoring, particularly in remote sensing and environmental management [7].

Mikalef et al. (2020) investigate how big data analytics can improve decision making in dynamic and operational environments, demonstrating that rapid data processing contributes to environmental sustainability [3].

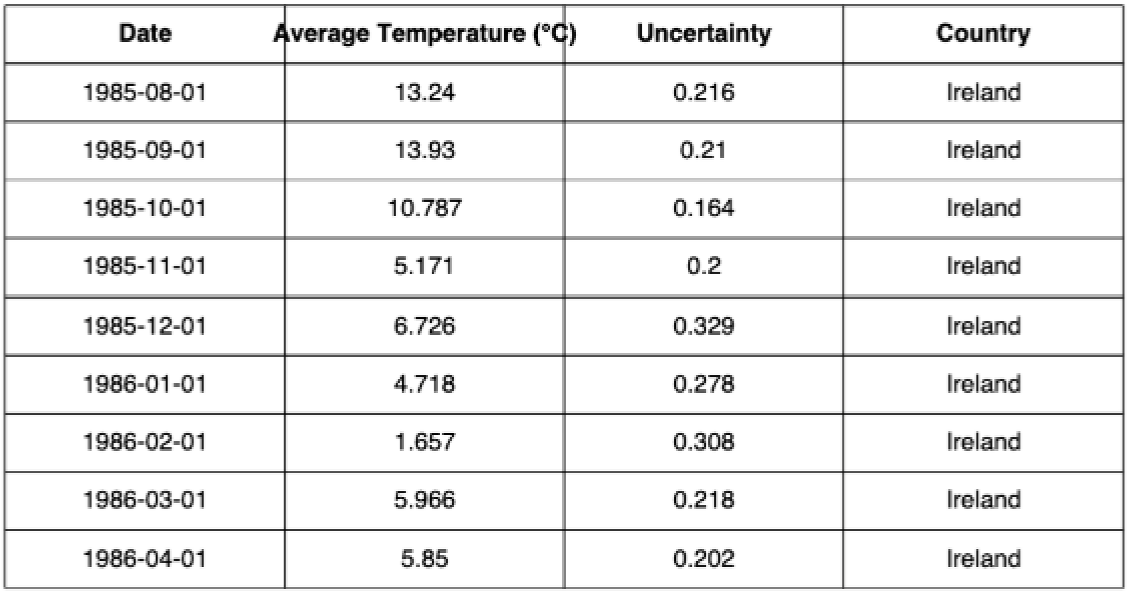
Nguyen et al. (2018) discuss big data analytics in supply chain management, emphasizing its potential to optimize resource allocation and reduce environmental impacts through logistics management [4]. Wamba et al. (2020) explore supply chain ambidexterity through big data analytics, identifying the moderating effect of environmental variability on decision-making efficiency [8].

Wang et al. (2020) demonstrate the intersection of climate change and public health by showing how big data-driven responses can mitigate crises such as pandemics, underscoring the broader applicability of data analytics to societal resilience [9]. Finally, Piao et al. (2019) examine plant phenology as an indicator of climate change. Their work underscores the growing need for comprehensive environmental monitoring driven by data analytics [5].

2 Data Sources and Preprocessing

2.1 Dataset Overview

- Source: Kaggle - Climate Change Earth Surface Temperature Data.
- Data Files: GlobalLandTemperaturesByCountry, GlobalLandTemperaturesByState, and GlobalLandTemperaturesByCity.
- Size: 13 to 31 MB per dataset.



Date	Average Temperature (°C)	Uncertainty	Country
1985-08-01	13.24	0.216	Ireland
1985-09-01	13.93	0.21	Ireland
1985-10-01	10.787	0.164	Ireland
1985-11-01	5.171	0.2	Ireland
1985-12-01	6.726	0.329	Ireland
1986-01-01	4.718	0.278	Ireland
1986-02-01	1.657	0.308	Ireland
1986-03-01	5.966	0.218	Ireland
1986-04-01	5.85	0.202	Ireland

Figure 1: Sample data from the GlobalLandTemperaturesByCountry dataset.

2.2 Data Preprocessing Techniques

- Data Cleaning: Corrupted records, including incomplete or invalid entries, were identified and removed. Date formats were standardized to ensure uniformity across all records, eliminating parsing errors during analysis.

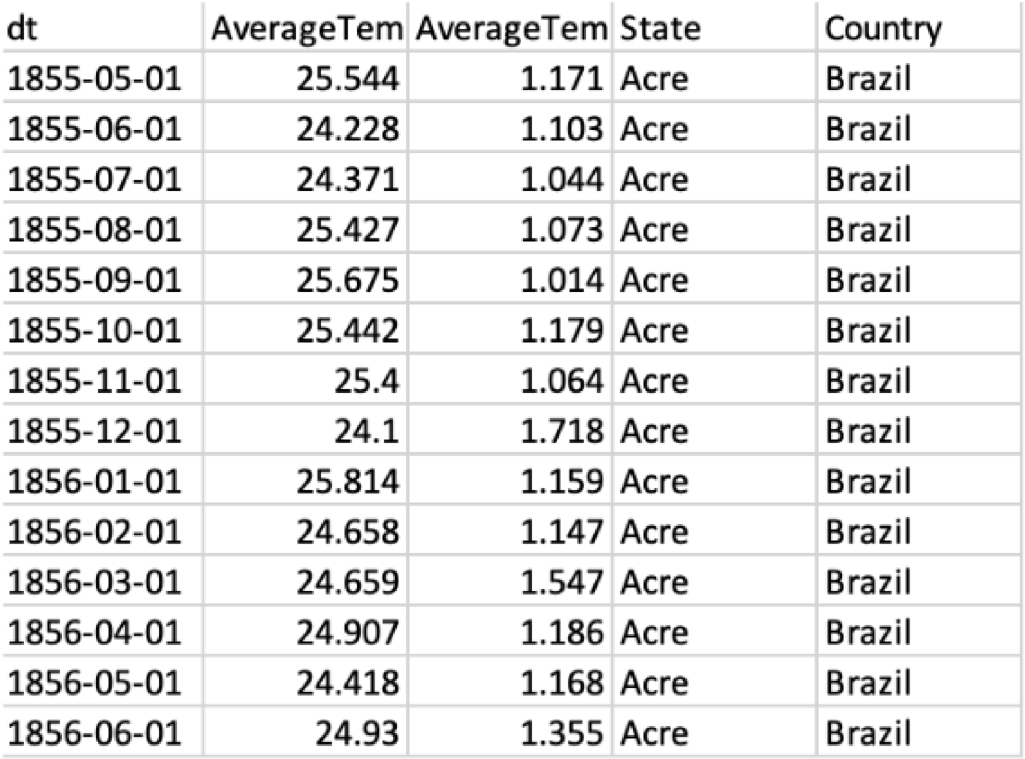
- Imputation: Used median values for missing temperatures, implemented using MapReduce.

- Feature Selection: Discarded unnecessary features to reduce noise.

3 Methodology

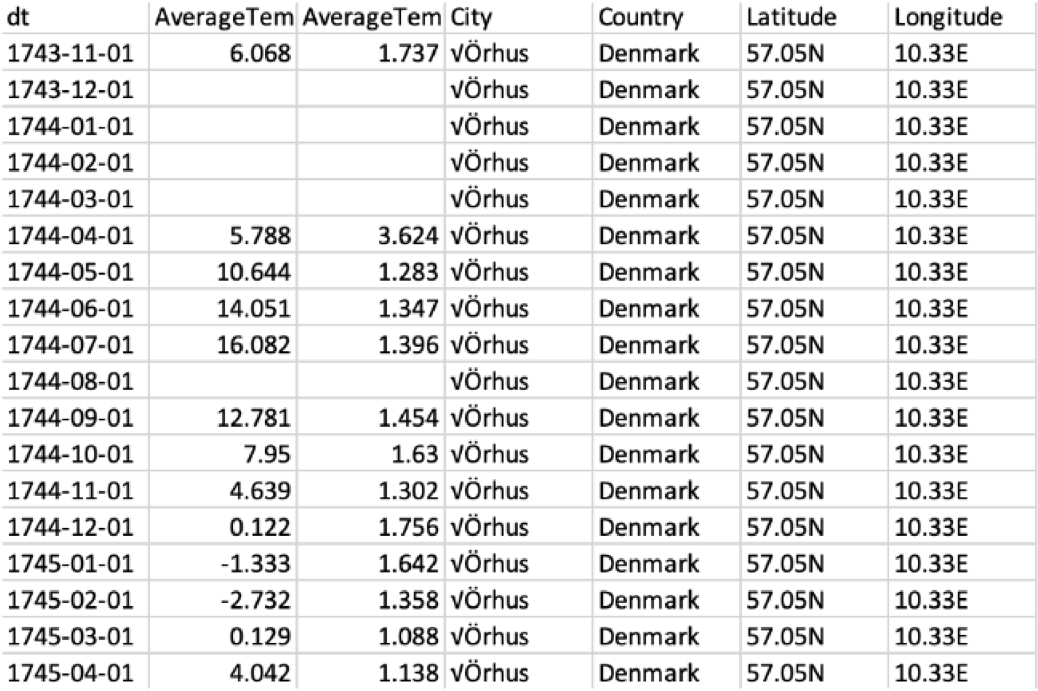
The analysis involved building regression models using big data frameworks. We employed a cloud-based platform (NYU Dataproc) for scalable data processing. The datasets were processed using MapReduce, followed by data cleaning scripts to handle missing values through median imputation. Process automation was achieved using shell scripts.

Each CSV file was uploaded to Hive tables on the NYU DataProc cluster. Trino queries were used to visualize temperature trends and perform regression analysis for predictions from 2025-2035.



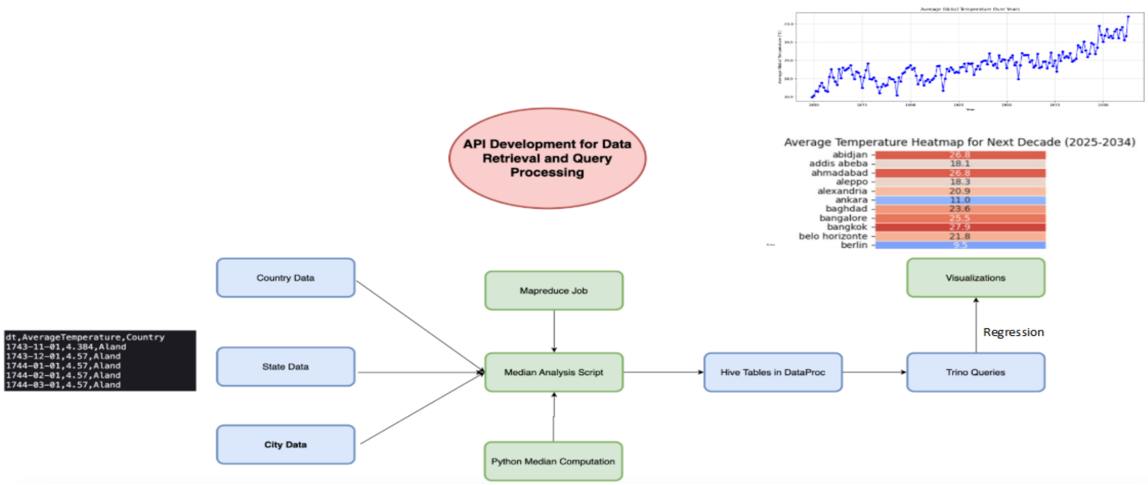
dt	AverageTem	AverageTem	State	Country
1855-05-01	25.544	1.171	Acre	Brazil
1855-06-01	24.228	1.103	Acre	Brazil
1855-07-01	24.371	1.044	Acre	Brazil
1855-08-01	25.427	1.073	Acre	Brazil
1855-09-01	25.675	1.014	Acre	Brazil
1855-10-01	25.442	1.179	Acre	Brazil
1855-11-01	25.4	1.064	Acre	Brazil
1855-12-01	24.1	1.718	Acre	Brazil
1856-01-01	25.814	1.159	Acre	Brazil
1856-02-01	24.658	1.147	Acre	Brazil
1856-03-01	24.659	1.547	Acre	Brazil
1856-04-01	24.907	1.186	Acre	Brazil
1856-05-01	24.418	1.168	Acre	Brazil
1856-06-01	24.93	1.355	Acre	Brazil

Figure 2: Sample data from the GlobalLandTemperaturesByState dataset.



dt	AverageTem	AverageTem	City	Country	Latitude	Longitude
1743-11-01	6.068	1.737	√Örhus	Denmark	57.05N	10.33E
1743-12-01			√Örhus	Denmark	57.05N	10.33E
1744-01-01			√Örhus	Denmark	57.05N	10.33E
1744-02-01			√Örhus	Denmark	57.05N	10.33E
1744-03-01			√Örhus	Denmark	57.05N	10.33E
1744-04-01	5.788	3.624	√Örhus	Denmark	57.05N	10.33E
1744-05-01	10.644	1.283	√Örhus	Denmark	57.05N	10.33E
1744-06-01	14.051	1.347	√Örhus	Denmark	57.05N	10.33E
1744-07-01	16.082	1.396	√Örhus	Denmark	57.05N	10.33E
1744-08-01			√Örhus	Denmark	57.05N	10.33E
1744-09-01	12.781	1.454	√Örhus	Denmark	57.05N	10.33E
1744-10-01	7.95	1.63	√Örhus	Denmark	57.05N	10.33E
1744-11-01	4.639	1.302	√Örhus	Denmark	57.05N	10.33E
1744-12-01	0.122	1.756	√Örhus	Denmark	57.05N	10.33E
1745-01-01	-1.333	1.642	√Örhus	Denmark	57.05N	10.33E
1745-02-01	-2.732	1.358	√Örhus	Denmark	57.05N	10.33E
1745-03-01	0.129	1.088	√Örhus	Denmark	57.05N	10.33E
1745-04-01	4.042	1.138	√Örhus	Denmark	57.05N	10.33E

Figure 3: Sample data from the GlobalLandTemperaturesByCity dataset.



```
dt,AverageTemperature,Country
1743-11-01,4.384,Aland
1743-12-01,4.57,Aland
1744-01-01,4.57,Aland
1744-02-01,4.57,Aland
1744-03-01,4.57,Aland
```

Figure 4: Design flow diagram

4 Results and Discussion

Our results consistently show an increase in average temperature across all three geographical resolutions. City temperatures seem to increase by 2.5 degrees celsius while Country temperatures show a substantial yet jumpy 12 degree incline in about 100 years (1750-1850) and then a slow steady increase. State temperatures illustrate a consistent slow increase from about 8 degrees celsius to 11 degrees celsius. We suspect these results vary due to lack of recordings/accuracy in earlier dates. Our prediction analysis, while limited, displays predicted averages of temperatures for all three resolutions.

4.1 City-Level Analysis

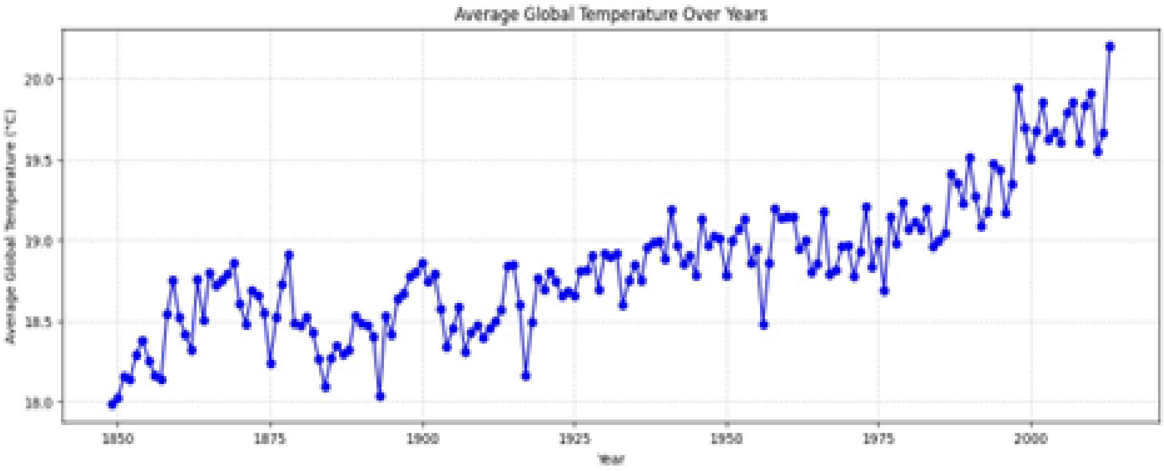


Figure 5: Average Temperature Increase: 2.5°C (from 18.0°C to 20.5°C).

4.2 State-Level Analysis

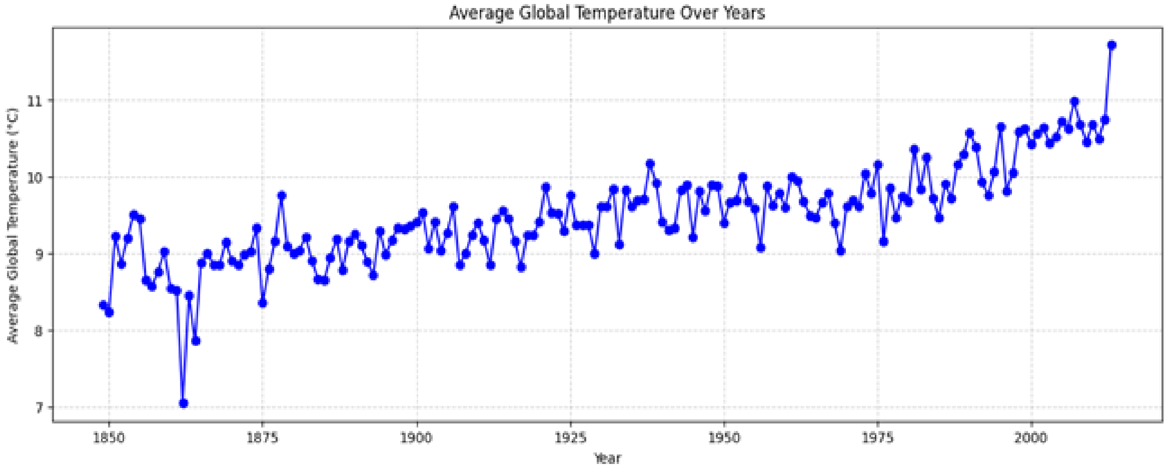


Figure 6: Average Temperature Increase: 3.6°C (from 8.2°C to 11.8°C).

4.3 Country-Level Analysis

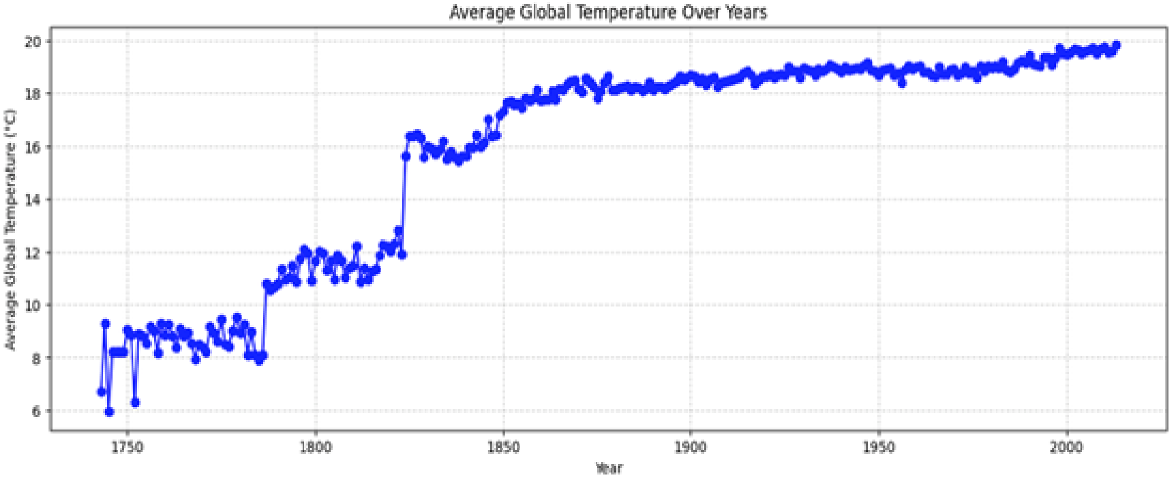


Figure 7: Most Significant Increase: 14°C overall.

4.4 Key Observations

- Missing data from 2013 onwards created gaps, affecting prediction accuracy.
- Inconsistent date formats required preprocessing for effective analysis.
- Early records exhibited inaccuracies due to limited historical data coverage.

5 Challenges and Limitations

- **Data Inconsistencies:** Due to differences in data resolution (city, state, country), it was difficult to generalize and streamline the data profiling across the datasets. For example, cities had coordinates while country-level data had some continent measurements that had to be removed. We addressed this by tailoring the profiling but also by extracting only the relevant features for the regression analysis.

- **Missing Data:** The weather data had several missing values across the time series. Data was missing from 2013 onwards which posed an issue for accurate trend prediction. On top of that several records were missing data which was solved by imputing the medians into those rows but could still have an implication on accuracy.

- **Computational Load:** The large dataset presented computational challenges when performing regression analysis, thus we chose to use the DataProc cluster to run our Trino queries.

6 Future Work

The lack of data after 2013 posed a serious limitation when trying to perform predictive analytics. Relevant and up-to-date data of all three geographical resolutions could be extremely useful in creating a more descriptive and accurate model that can predict average temperatures between 2025-2035.

7 Conclusion

The results underline a substantial rise in temperatures at all geographic levels, corroborating global warming trends. These insights can guide policy-making and awareness efforts to combat climate change. Future work should incorporate more recent datasets to enhance prediction accuracy.

References

[1] Berkeley Earth and Kristen Sissener. Climate change: Earth surface temperature data. `https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data`, 2024. Date published: unknown.

[2] Valerie Masson-Delmotte, Panmao Zhai, Hans-Otto Pörtner, et al. Ipcc report on climate change. *IPCC Sixth Assessment Report*, 2021. URL `https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_FrontMatter.pdf`.

[3] Patrick Mikalef et al. Big data analytics and decision-making in dynamic environments. *Decision Support Systems*, 2020. URL `https://www.sciencedirect.com/science/article/pii/S0378720618301022`.

[4] Tho Ph Nguyen et al. Big data analytics in supply chain management. *Supply Chain Management Review*, 2018. URL `https://kar.kent.ac.uk/62271/1/Manuscript_final.pdf`.

[5] Shilong Piao et al. Climate change and plant phenology monitoring. *Global Change Biology*, 2019. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14619`.

[6] Apoorv Singh et al. Github - jonamsalem/climate-change-analysis. `https://github.com/jonamsalem/Climate-Change-Analysis`, 2024. Accessed: 2024-12-12.

[7] H Tamiminia et al. Google earth engine for geo-big data applications. *Remote Sensing of Environment*, 2020. URL `https://www.researchgate.net/publication/341228837`.

[8] Samuel F Wamba et al. Supply chain ambidexterity through big data analytics. *Journal of Operations Management*, 2020. URL `https://www.sciencedirect.com/science/article/pii/S0925527319303184`.

[9] Yu Wang et al. Public health and big data-driven climate responses. *Journal of the American Medical Association*, 2020. URL `https://jamanetwork.com/journals/jama/article-abstract/2762689`.