



University of Melbourne
Faculty of Science

MAST 90107 Data Science Project

**Relationship Between Climate Events & Corporate
Financial Performance**

Group 22

Andrew STRINGER	694968	<code>astringer@student.unimelb.edu.au</code>
Chao JIA	958973	<code>chajia@student.unimelb.edu.au</code>
Wei LI	956833	<code>liwl1@student.unimelb.edu.au</code>
Xin WEI	980496	<code>xwwei@student.unimelb.edu.au</code>


October 31, 2021

Abstract

With the increased presence of climate change in everyday life, it has become apparent that climate events may impact the financial performance of markets and companies. However, there is a lack of research into the links between climate factors (temperature variations, bushfires) and accounting-based measures of corporate financial performance. This study addresses this gap by constructing a dataset for 26 U.S. agriculture companies and connecting measures of their reported financial performance with monthly state average temperature and yearly state bushfire data for years 2002-2020. This dataset is used to assess whether it contains evidence of linear regression or nonlinear classification relationship with limited evidence uncovered. Limitations of this study and future research directions are discussed, with the constructed dataset appearing to be a solid foundation for future research.

Declaration

We certify that this report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text. The report is 8880 words in length (excluding text in images, tables, bibliographies and appendices).

Name: Andrew STRINGER Signature: 

Name: Chao JIA Signature: 

Name: Wei LI Signature: 

Name: Xin WEI Signature: 

Acknowledgement

We are grateful to our Course Coordinators, Dr. Michael Kirley, Dr. Jeremy Sliver and Dr. Karim Seghouane who lead the subject and support our project during the past year. In addition, we would like to thank Dr. Ziad Al Bkhetan who supervised our project and inspired us throughout the year.

We could also like to thank our client, Dr. Laura Rusu and the Lensell, providing us the necessary resources, valuable suggestions and support.

Finally, we would like to thank our families and friends for being with us in the tough Covid situation, and anyone who has helped and supported us in this project.

Contents

1	Introduction	6
2	Literature Review	7
2.1	Previous Literature Category	7
2.2	CFP Measurement	8
2.3	Natural Disaster Data Utilization	8
3	Dataset	9
3.1	Financial Dataset	10
3.1.1	New Financial Data Source	10
3.1.2	Financial Data Re-collection Process	10
3.1.3	New Financial Data	11
3.1.4	Remaining issues	12
3.2	Climate Dataset	13
3.2.1	Climate datasets Problem in Semester 1	13
3.2.2	New climate Data Source	14
3.2.3	New Climate Data	15
3.2.4	Remaining Issues	17
3.3	Joining Dataset	18
4	Methodology	18
4.1	Measurements	18
4.1.1	CFP Measurement Selection	18
4.1.2	Climate Measurement	20
4.2	Hypotheses and Modelling	20
4.3	Dimension Reduction	20
4.3.1	Functional Principle Component Analysis	21
4.3.2	Independent Component Analysis	21
4.4	Regression Model	23
4.4.1	Linear Model	23
4.4.2	Hypothesis Test	23
4.5	Classification Model	24
4.5.1	Logistic Regression	24
4.5.2	Support Vector Machine	25

4.5.3	Random Forest	26
5	Results	27
5.1	Regression Model	27
5.2	Classification Model	29
6	Discussion and Future Directions	31
7	Conclusion	33
8	Appendix	37
8.1	Document Links	37
8.2	Data Process	37
8.3	Results	39

1 Introduction

In different regions of the United States loans and derivatives related to commercial real estate and commercial agriculture are all vulnerable to climate events and other environmental changes. For example, the increased intensity and frequency of droughts, floods, fires and other environmental changes may reduce the value of damaged assets and put pressure on borrowers' ability to repay loans. Climate change and natural disaster events affect financial markets in many ways. Dealing with these natural events can help investors, companies, and governments build a more comprehensive picture of the financial market in which they operate to make more rational investment decisions. A detailed understanding of this relationship may improve an entity's ability to address climate risks. Therefore, it is important to study the relationship between climate and financial performance.

Our project aims to assess the relationship between climate and corporate financial performance. The primary goal of this work is to better understand how companies are currently affected by their climate with the directive that this work may inform future analysis of how climate change may impact corporate financial performance.

The project uses climate data and financial data to investigate the relationship between the two. Quantifying the impacts of climate events is challenging because they may have both direct and indirect effects on financial markets, and these effects will not disappear for a period of time, depending on the type and severity of the event. At the same time, it is difficult to enumerate the impact of all-climate events. In the financial market, although there are several corporate financial performance (CFP) indicators, it may not be simple to select the best indicators for analysis. In addition to feature selection and quantification, determining the appropriate research scope is also a challenge. Generally, researchers model financial markets on three levels, the company level, the industry (index) level, and the macroeconomic level. Recent studies tend to combine some of them ([Sun et al., 2020](#)). In terms of data, climate data is full of noise and periodic changes, such as temperature, which requires well-thought-out pre-processing methods to normalize the data. Financial data is similar to climate data, but has random volatility and is affected by multiple factors and systems inside and outside the market. Due to the above reasons, this research direction is very scarce, and it is difficult to find relevant information ([Diaz-Rainey et al., 2017](#)). After reading some literature and to achieve our research goals, we decided to evaluate the impact of state monthly average temperature and bushfires on corporate financial performance.

The rest of the article is organized as follows: Section 2 is based on a review of the literature. Section 3 describes the datasets used in the study including the financial datasets, climate

datasets, and joining datasets. Section 4 presents the methodology along with the measurements, methods of dimension reduction, regression models, and classification models used in this study. Section 5 describes the results, Section 6 is a discussion of these results and the future directions of the project, followed by conclusions in Section 7.

2 Literature Review

Recently, more attention has been directed towards climate change (O'Neill et al., 2020) with many scholars using different methods to tackle the problem of obtaining a deep understanding of climate change, including methods such as statistical methods (Moonen et al., 2002), machine learning methods (Rolnick et al., 2019), and deep learning methods (Ardabili et al., 2019). However, there is limited research available which focuses on the relationship between climate change and finance.

Since finance is affected by many factors and climate-related data is difficult to tackle (Diaz-Rainey et al., 2017). Dimension reduction methods have been used to decompose the temperature data, and feed the lower-dimensional representations into various model types including a classifier that achieves a good performance (GHAYOUR, 2005). Functional principle component analysis (FPCA) is a variant of PCA, which is designed to solve the smoothness issues and is more friendly to time series data (Happ and Greven, 2018), and this method is used to decompose agriculture data (Lu et al., 2017). Independent component analysis (ICA) is considered to structure climate data (Hannachi et al., 2009), since it does not pose Gaussian assumption as the PCA-based methods (Comon, 1994). Ndehedehe and Ferreira (2020) use ICA to decompose water, while Moradkhani and Meier (2010) employees it to structure typhoon and global climate data, and both of them perform well in the actual model. However, few studies are concerned about the temperature and bushfire data, this project uses FPCA and ICA to reshape temperature data.

2.1 Previous Literature Category

Despite the sparsity of climate-finance literature (Zhang et al., 2019), previous literature in the climate-finance area can be broadly categorised by the climate factors focused upon and the level of financial impact being assessed. Climate factors may be climate variables (such as temperature, precipitation, air pressure, cloud cover, El Nino-Southern Oscillation phase, etc) which are present everyday and may be thought of as continuous variables (for Environmental Information, 2021; Organization, 2021a). Climate factors may otherwise be climate events (such as bushfires, flash flooding events, cyclones, earthquakes, etc) which are isolated incidents whose

damages may become more impactful ([Organization, 2021b](#)). Financial impacts considered are typically on the level of company, industry, or index level impacts. Additionally, financial impacts considered are returns-based, risk-based, or both. Our work fits into this segmentation as work that considers one climate variable, one climate event and focuses on company-level impacts to financial returns.

2.2 CFP Measurement

Before we are able to assess these relationships we must first decide on how to measure corporate financial performance. CFP measures may be accounting-based (Net Income, Return on Assets (ROA), Zmijewski score), market-based (stock returns, market value of a company), or composite measures ([Galant and Cadez, 2017](#)). In this project, we have data for and plan to use multiple accounting-based measures. These data are available for all companies (not just publicly listed firms) and these measures can be used for reliable comparisons between companies of different sizes within the same industry. The drawback of accounting-based measures is that they are typically only available on a yearly basis while some market-based measures are available on a daily basis. Yearly data appear to be sufficient as climate change is a phenomenon observed on the time-scale of decades, however, it has been argued that higher frequency data produces more reliable results ([Tzouvanas, 2019](#)). The argument for this is that periods of hot and cold temperatures will not appear in the yearly mean value. This perspective is incorporated into our analysis via the use of monthly temperature data observations which describe the seasonal patterns of temperature as well as give specific insight into temperature variations and patterns not described by yearly mean value alone. If even higher frequency data is desired then each monthly average value could be supplemented by distribution-based statistics (such as standard deviation, or extreme value measures) to address this issue, however this study uses monthly mean value data. Additionally, this tradeoff — reliable financial data vs reliable temperature data — is not an issue for bushfire and natural disaster data as there are no opposite effects from which misleading data may arise. To ensure a comprehensive analysis, the use of accounting-based measures are included in this study, with an awareness of how averaging daily data into monthly data may be influencing any results.

2.3 Natural Disaster Data Utilization

Previous use of natural disaster data has largely relied on representing daily data as binary data where 0 are days without a natural disaster and 1 are days with a natural disaster ([Worthington and Valadkhani, 2004](#); [Worthington, 2008](#); [Wang and Kutan, 2013](#)). The drawback of this

approach is that there is no distinction in the scale of impact of a natural disaster event. Moreover, this appears insufficient when considering yearly data (which we are), as there are bushfires every year the difference is in how many occurred and the extent of total damage suffered. As we have data on the impact extent (burnt area), this will be used in our study.

When considering the impacts of natural disasters on financial performance, we may distinguish between direct and indirect impacts. Direct impacts are those in which the company suffers from the natural disaster event, while indirect impacts include those which other companies (which the original company is connected to e.g. via supply chains) suffer. Previous studies have assessed the prolonged influence of climate events by including historical dependence in their models (Wang and Kutan, 2013). However, to the best of our knowledge, there are no such models which distinguish between direct and indirect influences which, in our view, impact companies on different time-scales. While direct impacts have immediate effects, indirect impacts may only be evident after days, weeks, or possibly months. This study focuses on immediate direct impacts.

Additionally, the use of historical dependence in these models obfuscates the actual prolonged impacts of natural disaster events on CFP by including implicit impact chains in the model. For example, if a company’s stock price is predicted by its previous day’s stock price and the 1-5 day impact delay of a bushfire — if a bushfire occurs on a Monday then Tuesday’s stock price is predicted by Monday’s stock price and the 1-day impact of the bushfires, but Wednesday’s stock price is predicted by Tuesday’s stock price and the 2-day impact of bushfires. In this situation, the bushfires impact Wednesday’s stock price in two ways with only one being explicitly interpreted as a result. As such, this type of model formulation may be a better representation of the underlying data, however, regression parameters alone may lead to inaccurate model interpretation. On the one hand, this is an undocumented limitation of previous work, on the other hand, this serves to inform the interpretation of our results.

3 Dataset

The dataset of this project consists of two parts, which are financial data and climate data. To be more specific, this project selects financial ratios as corporate financial performance measures which are calculated from each companies’ annual financial statements. The climate data used is temperature data and bushfire data (yearly total burnt area).

The training dataset which is used for analysis is the joined dataset of these financial ratios and climate data, where data are joined by time (month) and location (U.S. state). Note that the financial data are response variables and climate data are the predictors.

3.1 Financial Dataset

Based on the results of a correlation analysis (between company financial ratios, yearly average temperature and yearly bushfire area burnt) conducted in Semester 1, there appears to be a limited or weak relationship. By revisiting the dataset, it was found that some of financial statements on Yahoo Finance were incomplete. For example, there was no Non Current Asset Value for CF Industries Holdings Inc. in 2002 or 2003. This missing data will result in an incomplete calculation of Total Assets, which is the sum of Current Assets and Non Current Assets. More importantly, Total Assets is a very significant component in the ROA calculation. These kinds of problems will lead to incomplete and unreliable calculations of other financial ratios used for CFP measurements as well. Therefore, a complete and reliable dataset needs to be re-collected.

Another issue from the Yahoo Finance data is the industry segmentation, where Yahoo Finance classifies companies that are related to agriculture but not in the agriculture industry as agricultural companies. For example, CTA is classified as an agriculture industry company by Yahoo Finance is not an agricultural company, while the real industry of CTA is Plastic Material.

3.1.1 New Financial Data Source

To search for accurate and reliable data, the data on the US Securities and Exchange Commission (SEC) website is used. The data on the SEC website is the source of all other financial databases. To acquire the data from the SEC website, the XBRL API is applied, which was a recommendation by the client. Moreover, the SEC website also provides the industry segmentation criteria, Standard Industrial Classification (SIC) code, and the respective companies' data will be obtained according to the SIC industry segmentation criteria.

3.1.2 Financial Data Re-collection Process

In order to obtain financial data, there are several steps as follows:

- I. obtain the corresponding list of companies according to the SIC code;
- II. extract the CIK code of each company, which is a unique code for each company;
- III. acquire corresponding financial data by setting the Central Index Key (CIK) code of each company, the code of the corresponding financial data and other information we need as parameters in XBRL API. For example, 'Assets' is the code for total assets in financial statements and 'period.end' will return their reporting date;
- IV. generate a table of financial data for each company based on all returned values.

SIC Code	Office	Industry Title
100	Office of Life Sciences	AGRICULTURAL PRODUCTION-CROPS
200	Office of Life Sciences	AGRICULTURAL PROD-LIVESTOCK & ANIMAL SPECIALTIES
700	Office of Life Sciences	AGRICULTURAL SERVICES
800	Office of Life Sciences	FORESTRY
2870	Office of Life Sciences	AGRICULTURAL CHEMICALS

Table 1: Selected Industry List from SEC

3.1.3 New Financial Data

There are 5 agricultural sub-industries that are selected in the project, shown in Table 1. They are Agricultural production-crops (SIC 100), Agricultural prod-livestock animal specialties (SIC 200), Agricultural services (SIC 700), Forestry (SIC 800) and Agricultural Chemicals (SIC 2870) (SEC, 2021). The original strategy is to set the Agricultural production-crops and Agricultural prod-livestock animal specialties as first batch and the others as the second batch. This is because the above mentioned two sub-industries are the most traditional agricultural sectors which are the ideal agricultural sub-industries this project would like to analyse, while the others are agricultural related. In particular, Agricultural services and Agricultural chemicals are industries which provide services for agriculture and which are not significantly influenced by region. As a result, this may reduce the performance of the modelling.

After filtering the CIK of each company from XBRL API from the 5 mentioned agricultural sub-industries, there are 26 companies in total with returned values. This means that only those 26 companies are linked with XBRL API, shown Figure in 2. Specifically, SIC 100 has 6 companies return their financial data, and SIC 200, 700, 800 and 2870 have 2, 2, 0, 16 companies return their financial data respectively. Therefore, to make sure there is sufficient data for modelling, those 26 companies are combined together to represent the agriculture industry in this project.

The number of data points for each company varies depending on the number of their publicly published financial statements. As financial ratios are reported annually, for some companies with few valid financial statements, they do not provide sufficient data points for appropriate modelling and data analysis. For example, some companies only have two years' available financial data statements, which means there are only two data points which can be used for each financial ratio in the modelling part. Therefore, the companies with no more than

SIC Code	Industry Title	The Number of Companies with Returned Value
100	AGRICULTURAL PRODUCTION-CROPS	6
200	AGRICULTURAL PROD-LIVESTOCK & ANIMAL SPECIALTIES	2
700	AGRICULTURAL SERVICES	2
800	FORESTRY	0
2870	AGRICULTURAL CHEMICALS	1

Table 2: Number of Companies with API Returned Value Respect to Each Industry

Quantile	Min	Quantile 1	Median (Quantile 2)	Quantile 3	Max
Number of Years (Before Filtering)	2	6	11	13.5	19
Number of Years (After Filtering)	3	9	11	14.5	19

Table 3: Statistical Quartile for the Number of Years Data

2 years' valid financial data will be removed in the project modelling. After this filtering, the number of valid company reduces to 23. Table 4 shows the statistical quartile information about the number of years that can be used before and after filtering. Furthermore, some financial data before 2010 cannot be returned by XBRL API, therefore the rest of the financial data has to be manually acquired via the SEC website. This also is a time costly process and limits the number of data points in the project.

3.1.4 Remaining issues

There is still one issue in the financial data. After manually copying data from SEC website, some companies still have missing values, especially for Revenue data, which is used to calculate Net Margin and Operating Margin. To be more specific, There is no Revenue data for Kiwa Bio-tech Products Group Corp reported between 2011 to 2015 and no Revenue data for Scotts Miracle-Gro Co from 2008 to 2012. This will not only lead to Net Margin and Operating Margin being unable to be calculated for those respective years, but also means that the data points cannot be continuous. Thus, it may have an impact on the modelling part.

Another issue is that different companies have different reporting periods. Although the majority use reporting period January to December, there are other companies who report their financial statement in June, September and March. Those different reporting periods are a challenge to joining financial data with climate data on the respective year and month. Another challenge is that, there are two companies, Mosaic Co. and Suma Inc., who have changed their reporting date. This means that for some financial data it does not represent the full year's performance for the company, which is an additional challenge to matching the financial data with the climate data.

3.2 Climate Dataset

3.2.1 Climate datasets Problem in Semester 1

Our climate data set continues the idea of the first semester and is divided into two data sets, temperature and wildfire. By analyzing the unsatisfactory results of the first semester, we believe that there are two reasons: 1. The time interval of the data set is too small. In the first semester, The dataset including the land average temperature in the US is provided by our client, and the dataset is published on the Kaggle website¹. Kaggle repackaged the original data that is from the latest compilation compiled by Berkeley Earth official website, a subsidiary of Lawrence Berkeley National Laboratory. However, this data set is only updated to 2013, and data for the past 7 years is not available. 2. In the first semester, we did not consider the factors of geographical influence when we made the model. We used the data of the whole United States for analysis, but considering that the United States has a large area across the latitude and longitude, the data of each state are different. For example, the average temperature in Hawaii in January 2017 was 21 degrees, but the temperature in Alaska during the same period was -17 degrees, so the average temperature in the United States can no longer be used as an indicator variable. In addition, agricultural companies are more affected by geographic factors. In the planning for the second semester, we will consider the geographical location of the state where each company is located. Thus, we re-researched the average temperature data of each state from 2002 to 2020 on Berkeley Earth. Therefore, we re-collected and processed the two types of data after improvement.

¹Kaggle website: <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data/discussion/32275>

Abbreviation of Agency	Full Name of Agency
BIA	The Bureau of Indian affairs
BLM	The Bureau of Land Management
DOD	United States Department of Defense
DDQ	Department of Defense
FS	The Forest Service
FWS	The United States Fish and Wildlife Service
OAS	Office of Air Services
NPS	The National Park Services
USFS	The United States Forest Service

Table 4: NIFC corporation agency summary

3.2.2 New climate Data Source

In order to ensure the accuracy and consistency of the data, the wildfire data comes from the official website of the National Interagency Fire Center of the United States². NIFC compiles annual wildfire statistics for federal and state agencies. This information is collected from the incident management report, which has been in use for decades. It is reported by federal, state, local, and tribal land management agencies through established reporting channels. The wildfire data sources used in this report were collected from the Inter-Agency Fire and Aviation Management Web Application (FAMWEB) system, including the Situation Report and Incident Status Summary (ICS-209) program. This document also uses previous National Inter-Agency Coordination Center (NICC) annual reports and other sources. The statistics provided here are intended to provide a national perspective on annual fire activity, but they may not reflect official data from specific agencies. The list of cooperative agencies included in the NIFC report is shown in Figure 2. In addition, the historical year-end fire statistics by state provided by NIFC is 2002-2020.

For the temperature data, we continue to use the data provided by the Berkeley Earth website³. Berkeley Earth is an independent U.S. non-profit organization that provides comprehensive open source world air pollution data and highly user-accessible global temperature data. These data are timely, fair and verified.

²National Interagency Fire Center website: <https://www.nifc.gov/fire-information/statistics>

³State temperature in Berkeley Earth website: <http://berkeleyearth.lbl.gov/state-list/>

3.2.3 New Climate Data

The 2002-2020 wildfire data report is divided into two statistics: the number of wildfires and the total number of hectares burned. Since wildfires may be caused by many different reasons, in terms of the type of fire, during the period 2005-2020, various agencies have counted man-made fires, Rx fire. Prescribed (Rx) fire, or fire ignited under known conditions of fuel, weather, and topography to achieve specified objectives. Prescribed fire is an important tool for reducing wildfire hazards, ecosystem restoration, vegetation management and wildlife habitat improvement; It is also an important cultural resource and has applications in forest management and pasture improvement. As shown in Figure 3. In addition, the 2002-2004 report included three

National Report of Wildland Fires and Acres Burned by State in 2018 Figures from the Fire and Aviation Management Web Applications Program.

ALABAMA

Agency	Fires - Human	Acres - Human	Fires - Rx	Acres - Rx	Fires - Total	Acres - Total
FS	11	244.5	0	0	11	244.5
FWS	0	0	0	0	0	0
OTHR	7	238	0	0	7	238
ST	952	14,981.3	0	0	952	14,981.3
Totals:	970	15,463.8	0	0	970	15,463.8

Figure 1: National Report Wildland Fires and Acres Burned by State in 2008

types of fires: Wildland, Rx and WFU. Wildfire Utilization (WFU) is the management of naturally ignited wildfires (fires caused by lightning or lava) to achieve specific resource goals in a predetermined area. Goals can include maintaining healthy forests, pastures and wetlands, and supporting ecosystem diversity. as shown in picture 4.

For the choice of Bushfire indicators, The bushfire dataset contains the total number of wildfires per year and annual wildfire-burned area. We prefer to use the annual wildfire-burned area as the bushfire indicator.

According to Figure 5, the trend between the number of fires and the burning area is not the same. Although the total number of bushfires is relatively high, the total burning area is not necessarily large. In addition, the burned area is more capable of reflecting the impact of fire. For instance, the number of fires is high in 1985, but the burned-area is relatively lower than most other period. The similar example is in 2020 as well. Therefore, in the current stage, we will first use burned area as the variable to indicate the climate events.

Since only 3 years of reports include WFU type fire statistics, and few organizations calculate WFU, the data for most states are empty. Therefore, we choose to add the burning area of the

National Report Wildland Fires and Acres Burned by State in 2004							
State	Agency	Wildland		Rx		WFU	
		# Fires	# Acres	# Fires	# Acres	# Fires	# Acres
ALASKA	BIA	6	18	0	0	0	0
	BLM	37	702,783	0	0	0	0
	DDQ	12	65,211	0	0	0	0
	FWS	32	341,451	0	0	0	0
	NPS	10	133,810	0	0	0	0
	OTHR	41	53,232	0	0	0	0
	ST	384	880,143	0	0	0	0
	FS	21	17	0	0	0	0
	Totals	543	2,176,665	0	0	0	0

Figure 2: National Report Wildland Fires and Acres Burned by State in 2004

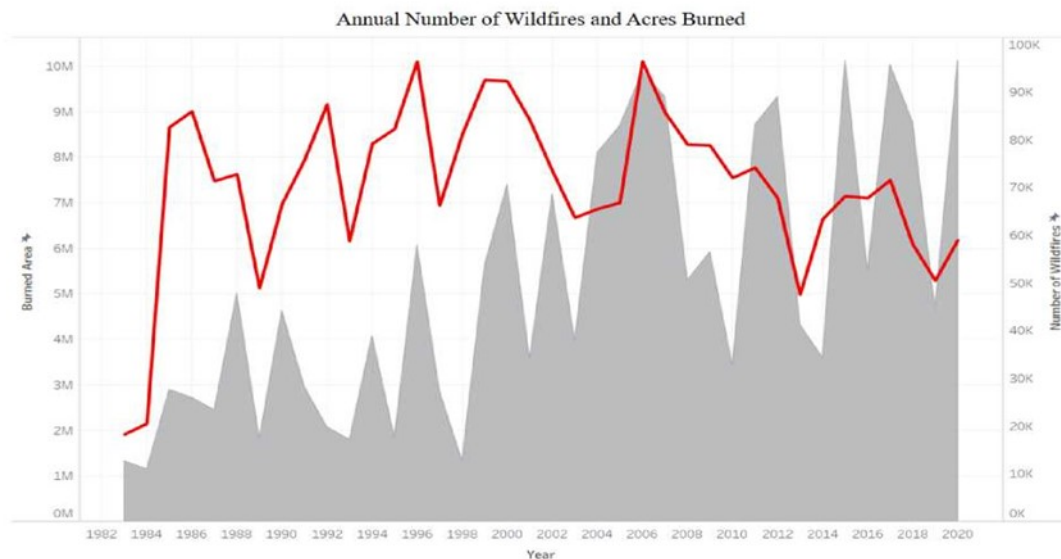


Figure 3: Annual Number of Wildfires and Wildfire-burned Area

wildland and Rx fires in the dimensions of the state according to the year as the variable.

We selected the monthly average temperature of 51 states in the United States from 2002 to 2020. Since the data for each state in the website is on a separate page, and the temperature is measured in degrees Celsius and reported as an anomaly relative to the average value from January 1951 to December 1980. Therefore, we need to manually enter the monthly data for each state from 2002 to 2020 based on the average value as the reference line. In order to ensure the accuracy of the data, we also used the corresponding crawler code to obtain and calculate the data which is compared with manually processed data. The new temperature data set used in the model this semester is the monthly average data of 51 states in the United States from 2002 to 2020. As shown in the figure, the data set contains 12 months of annual temperature

records, which also makes up for the lack of the data set in the first semester. Because in the land average temperature data, there are only 9 months of data in 2013 (2013/01-2013/09), and other years includes full 12 months.

3.2.4 Remaining Issues

The first issue of current climate datasets is the missing value and outliers. There are several missing values in the burnt area data by bushfire in each state from 2002 to 2020. As Figure 6 shown, there was no agency reporting the area of wildfires burned by Delaware in 2011 and 2012, so the values for these two years are null (red rectangle). In addition, the data for 2004, 2014, 2015, 2016, 2018, and 2019 are recorded as 0 (orange rectangle). According to the data of previous years, although wildfires in this state rarely occur, the possibility that these are outliers cannot be ruled out. The same example is the area of wildfires burned by Hawaii in 2008, 2013, and 2014. All data are 0 (blue rectangle), but according to Hawaii's tropical climate and its data from previous years (mean value of 6315.42 hectares), these three zero values should be outliers. After checking, the main reason for missing values is that there is no agency to conduct statistics for the state. There are a small number of missing values in the original temperature data set, for example, the data for the state of Idaho for May 2005 and September 2016 are null. But fortunately, the geographic location of the agricultural company we studied is not in the state where the null value is located, and the annual temperature difference between these states is almost small, so it will not have a great impact on the results.

Year	ALABAMA	ALASKA	ARIZONA	ARKANSAS	CALIFORNIA	COLORADO	CONNECTICUT	DELAWARE	FLORIDA	GEORGIA	HAWAII	IDAHO	ILLINOIS	INDIANA	IOWA	KANSAS	KENTUCKY
2002	1023679	2176665	795458	151.373	595.058	937.241	1,949	1,659	343.529	20.381	3.660	131.237	779	17.235	3.662	8.037	33.332
2003	779.992	560.887	303.760	224.014	861.184	49.893	234	42	446.845	89.807	6.710	362.989	2.298	13.087	3.874	75.126	27.964
2004	120.717	6.701.879	286.099	247.277	321.215	72.784	279	0	537.887	95.961	656	81.539	1718	20.872	8312	35.957	49.118
2005	100.410	4.440.775	1.086.908	162.285	297.849	75.603	393	456	329.963	140.706	4	498.590	4142	14.430	21.840	47.010	81.726
2006	924.998	278.305	188.113	252.752	746.597	131.144	475	220	464.534	117.280	9.863	997.007	5.767	23.545	11260	106.908	84.728
2007	857.030	545.667	199.974	257.556	1.158.513	56.579	348	318	891.785	888.869	21.030	2.028.340	13159	32.885	12915	45.454	79.737
2008	126.066	66.638	218.447	199.613	1.463.197	185.012	961	128	559.069	114.026	0	174.161	6247	14.767	2.540	80.495	55.869
2009	91199	2.951.887	410.889	214.514	499.525	76.130	322	136	559.069	113.881	7.800	56.576	15.988	17.840	41.904	57.122	62.499
2010	110.131	1.125.924	163.144	242.479	182.094	73.397	319	54	507.491	178.448	10.172	11.387	8.969	679.649	11.387	36.697	83.331
2011	14.781	302.000	1.077.598	196.535	186.127	190.498	286	N/A	357.169	245.985	2.328	427.436	9.871	8715	36.517	121.230	37.913
2012	120.525	300.113	297.049	207.040	926.429	255.220	459	N/A	305.043	59.865	1.802	1.698.718	25.719	16.223	14428	64.598	54.306
2013	138693	1322053	154772	193949	626219	211466	275	217	347028	138250	0	749098	16572	5035	33152	8451	40443
2014	142034	293152	270104	154469	555044	44756	103	0	1376867	72656	0	228480	15082	9004	30573	43588	50540
2015	130161	5116357	257126	118908	940191	48780	184	0	375178	996602	5611	830362	24365	10710	44664	62259	29369
2016	173199	530044	410270	316640	607768	160102	930	0	359358	1559911	15098	403333	23868	11963	49319	362416	87109
2017	964647	717973	563442	249483	1315746	134812	274	286	2481811	1456006	2098	719077	15915	8617	33231	483598	32673
2018	15463.8	41068	165356	24071	1823153.2	475803	40	0	138820	14236	21979	604481	120	115	8014	59234	8417
2019	22158	2498159	384942	8602	259148	40392	72	0	122500	12407	10710	284026	41	523	2020	21167	11714
2020	20557	181169	978567.5	12552	4092150.5	625357	383	1356	99413.1	5677	472	314352	239.6	313	2168	34581	7950

Figure 4: Monthly temperature data for 17 states in the U.S.

Another remaining issue is that bushfire data is only available by full year. But the CFP data is only available by full reporting period and this may not align with the bushfire data period. We tried to contact the NIFC staff, but their feedback was that the organization that provided the data could not disclose the monthly fire data to the public, so this part is currently not available to us. To address this issues, there are two assumptions were made: the first one is that the impact of bushfires on company performance is delayed and another one is that the majority

of fires occur in the summer period. Thus, the reporting periods beginning from January to October used the bushfire data from that same year. However, the reporting periods beginning in November or December used a linear interpolation between the two bushfire observations. Although this makes us more cautious of inferring bushfire impacts on companies with these reporting periods, the majority of companies report from January to December so that there is little cause for concern in the majority of cases.

3.3 Joining Dataset

The financial dataset is a collection of company data that consists of the financial ratios of that company at a specific year. The climate dataset includes the U.S.A temperature dataset and the U.S.A bushfire dataset, the two datasets both are at the state level while the reporting period of the temperature dataset is monthly and the bushfire dataset is yearly. The joining dataset aligns the financial data and climate data together over the same yearly timestamp⁴ and we apply some tricks to handle the discontinuity data issue in the financial data, inconsistent reporting period issue in financial data, and missing value issues in bushfire data in financial dataset while concatenating the data⁵. The dataset acts differently while applied in regression model and classification model (seen in section 8.2), since the regression model requires continuous variable as response variable typically and the classification model accepts binary variable only. All data in the joining dataset has been standardized, which eliminates the variance of the raw data and scales the data with a different unit in the same page (Colan, 2013). Building such dataset to analysis the relationship is one of our contributions since no one has constructed a financial-climate dataset at the company level to facilitate the research of other scholars.

4 Methodology

4.1 Measurements

4.1.1 CFP Measurement Selection

To select the CFP measurement, only 5 financial ratios are used in the modeling. They are ROA, ROE, Current Ratio, Net Margin and Operating Margin. The reason is that the element in those ratios are compulsory items in financial statements, which means every company all should include these data in their financial statements (May, 1937). The following graph will further explain why the other financial ratios will not be included.

⁴The detailed data structure of joining dataset of joining data is shown in section 8.2.

⁵The detail is shown in challenge paragraph section 8.2 .

However, other financial ratios may include optional items or bias items. Those data may be different depends on their business activities and their accounting accounts used. For example, Quick Ratio requires the use of Inventories, however inventories vary depending on the business of the companies and some companies they even do not have inventories.

Another example is debt, which is required in ROC and D/E ratios calculation. However, there are several different accounting accounts to record their real debt in financial statements. A normal way to register debt in financial statements is "Debt", while some companies may register debt as loan, operating loan or others. This results in the returned debt value through the XBRL API is not consistent and inaccurate. The last one is EPS, EPS is calculated as:

$$EPS = \frac{NetIncome}{the\ Number\ of\ Shareholders} \quad (1)$$

EPS is an important measurement to value a company's share price ([Ohlson and Juettner-Nauroth, 2005](#)). But as it varies with the change of number of share holders. So, if there is a company issues new shares with no change on net income, there EPS will be lower, however, their financial performance does not change. This can lead to distortion of the measurement. For the selected financial ratios, there are different calculations as well. Last semester, there are two ROA calculation methods have been introduced ([May, 1937](#)),

$$ROA_1 = \frac{Net\ Income_t}{Total\ Asset_t} \quad (2)$$

$$ROA_2 = \frac{Net\ Income_t}{average(Total\ Asset_t, Total\ Asset_{t-1})} \quad (3)$$

Similarly, ROE can also be calculated as ROE1 and ROE2 [May \(1937\)](#),

$$ROE_1 = \frac{Net\ Income_t}{Total\ Equity_t} \quad (4)$$

$$ROE_2 = \frac{Net\ Income_t}{average(Total\ Equity_t, Total\ Equity_{t-1})} \quad (5)$$

By investigating the formula we can find that, ROA2 and ROE2 require two year data of Asset and Equity. This is because Net Income reflect the performance of a company during the past reporting period, while Asset and Equity only reflect the financial situation on the reporting date. Therefore, an average value of two-year reporting date is imported to reflect the assets situation during the past reporting periods as objective as possible ([May, 1937](#)). Although ROA2 and ROE2 to some extent is better than ROA1 and ROE1, based on current limited data points, ROA2 and ROE2 will lead to a reduction in the amount of data points. Therefore, ROA1 and ROE1 are ultimately selected in order to provide as many data points as possible for the modeling.

4.1.2 Climate Measurement

Two climate events are used in this project, bushfire and temperature change. To represent the change of the climate, we use the 1-year difference of the data instead its actual value and some data noises can be eliminated in this way. For example, the data of 2020 is the difference between 2020's and 2019's.

4.2 Hypotheses and Modelling

The frame of our work is to assess the hypothesis that climate factors impact corporate financial performance. This suggests that our competing hypotheses are:

$$\begin{aligned} H_0 &: \text{Climate factors do not impact CFP} \\ H_A &: \text{Climate factors impact CFP} \end{aligned} \tag{6}$$

Within H_A there are three distinct hypotheses: only temperature, only bushfires, or both temperature and bushfires impact CFP which are to be explored after the null hypothesis is rejected (when that occurs). To assess the hypothesis H_A , there are two key steps: firstly to identify plausible relationships between climate factors and CFP measures, and secondly to assess the significance of these identified relationships. Given the lack of relevant work in this area multiple relationship types were explored: linear vs non-linear, and regression vs classification.

4.3 Dimension Reduction

A key challenge in this work is that each company has at most 19 data points of joined CFP, state temperature, and state bushfire data with many companies limited to ~ 10 data points. The number of climate predictors used is 13 (one temperature predictor for each month, and one bushfire predictor) which quickly leads to models being overfit. This problem is addressed in two ways: to implement dimension reduction on the 12 temperature predictors, and to select models via measures which guard against overfitting (such as adjusted R^2 and cross validation error).

Dimension reduction was implemented on the temperature data to construct global information about each year's temperature variations. Initially, it was suggested that bushfires be predicted by the temperature data to enable further dimension reduction, however this was not implemented as results are to be interpreted within the hypothesis frame mentioned in 4.2 where significant relationships may include only temperature, only bushfires, or both climate factors. This distinction is integral to our work and its significance to the literature area.

4.3.1 Functional Principle Component Analysis

The first dimension reduction approach taken was functional principle component analysis. This was chosen as the monthly temperature data is representative of a yearly temperature function. By casting the vector of monthly temperature averages to a function we improve interpretability. The two steps to achieve this are: to represent the raw data as functional data, and to conduct functional PCA on the functional data.

A function may be represented as a linear combination of basis functions, which is analogous to representing a vector as a linear combination of basis vectors⁶. Using vector notation, the temperature function $T_i(t)$ for year $i = 2002, \dots, 2020$ is expanded as:

$$T_i(t) = \mathbf{c}_i^T \mathbf{b}(t) \quad (7)$$

where $\mathbf{b}(t)$ is a vector of the basis functions and \mathbf{c}_i is the vector of coefficients for the linear combination of these basis functions for year i .

The casting of our temperature data to a function was implemented by using a Fourier basis expansion/truncated Fourier series. Specifically:

$$T_i(t) = \frac{a_{0i}}{2} + \sum_{k=1}^K a_{ki} \cos\left(\frac{2\pi}{12} kx\right) + b_{ki} \sin\left(\frac{2\pi}{12} kx\right) \quad (8)$$

Plots of some results of this interpolation procedure are provided in Figure 5 to clarify what the functional data representation used is.

The second step is to conduct fPCA on the obtained functional data which is done using the fda package in R (Ramsay and Silverman, 2005). fPCA selects the set of basis functions by finding the dominant modes of variation of the functional data (from the mean-value function).

Prior to conducting this decomposition, it was decided that we would choose the number of principle components to either retain 95% of the temperature data variation or 6 at most. As evident in Table 5, which details the cumulative proportion variance explained, the first 6 principle components were retained and they explained an average of 88.9% of the year-to-year temperature variation. Additionally, for companies with $n < 10$ data points, the number of principle components was limited to $n - 3$ to ensure model identifiability when using all retained principle components.

4.3.2 Independent Component Analysis

The second is also a linear transformation method named independent component analysis (ICA), which decomposes the data into statistical independent components (Hyvärinen, 1999).

⁶A clarifying example is the Taylor series expansion which uses the basis set $\mathcal{B} = \{x^n : n \in \mathbb{N} \cup \{0\}\}$.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
CA	0.265	0.487	0.642	0.730	0.798	0.859	0.917	0.957	0.981	0.992	0.997	1
OH	0.323	0.478	0.616	0.742	0.839	0.893	0.927	0.959	0.975	0.986	0.995	1
GA	0.270	0.473	0.632	0.773	0.837	0.895	0.938	0.969	0.984	0.993	0.998	1
FL	0.336	0.525	0.693	0.812	0.890	0.934	0.970	0.985	0.993	0.997	0.999	1
CO	0.233	0.426	0.573	0.679	0.762	0.834	0.897	0.948	0.971	0.986	0.996	1
MN	0.334	0.530	0.690	0.803	0.861	0.910	0.945	0.971	0.983	0.993	0.997	1
KS	0.297	0.467	0.632	0.739	0.828	0.884	0.923	0.951	0.973	0.990	0.997	1
PA	0.358	0.506	0.647	0.763	0.843	0.899	0.937	0.963	0.979	0.991	0.998	1
TX	0.323	0.489	0.630	0.738	0.836	0.886	0.925	0.955	0.974	0.990	0.997	1
MA	0.400	0.561	0.713	0.793	0.845	0.888	0.925	0.954	0.982	0.993	0.997	1
MS	0.301	0.507	0.651	0.758	0.840	0.895	0.934	0.965	0.979	0.991	0.997	1
Average	0.313	0.495	0.647	0.757	0.834	0.889	0.931	0.962	0.979	0.991	0.997	1

Table 5: Table: Cumulative variance explained by retaining different numbers of functional principle components for each state in which at least one company is headquartered.

Statistical independence among components s_1, \dots, s_i means the density function can be factorized as:

$$f(s_1, \dots, s_i) = f_1(s_1)f_2(s_2)\dots f_i(s_i) \quad (9)$$

ICA also assumes all the components s_1, \dots, s_i should be non-Gaussian, which is opposite to fPCA.

In this scenario, ICA decomposes the 12-month temperature data \mathbf{T} into 4 components s_1, \dots, s_4 , which corresponds to 4 seasons since different season has different temperature pattern typically, such as:

$$\mathbf{T} = \mathbf{A}\mathbf{s} \quad (10)$$

The matrix \mathbf{A} represents a constant $n \times 4$ mixing matrix, and \mathbf{W} denotes \mathbf{A}^{-1} which maps the $\mathbf{T}(n \times 12)$ matrix into $\mathbf{s}(n \times 4)$ matrix. Since we only need 4 components instead of all, one-unit contract functions are designed to find the single independent component iterative, and we uses the general contract function with $\log(\cosh)$ setting to measure the non-Gaussian and find the single component. Meanwhile, the iterative algorithms is implemented by fastICA⁷ with the

⁷sklearn package: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html#sklearn.decomposition.FastICA.transform>

form as:

$$\mathbf{w}(k) = E \{ \mathbf{t} g(\mathbf{w}(k-1)^T \mathbf{t}) \} - E \{ g'(\mathbf{w}(k-1)^T \mathbf{t}) \} \mathbf{w}(k-1) \quad (11)$$

where g is the derivation of the measurement in general contract function (Hyvärinen and Oja, 1997).

4.4 Regression Model

A linear regression model was used to assess the relationship between climate factors and CFP. Identifying a plausible relationship involved selecting the model with maximum adjusted R^2 value and conducting a permutation test to assess the likelihood of obtaining a model with this adjusted R^2 value given H_0 . This p -value is then used to assess the statistical significance of the identified relationship.

4.4.1 Linear Model

The plausible linear regression relationship was identified by selecting the predictor combinations which maximised the adjusted- R^2 value. The predictors used were: a constant term, the bushfire burnt area and the retained temperature principle components for the state where the company is headquartered.

It was decided that we explore all predictor combinations, rather than limit the temperature PC's to sequential inclusions. This decision was made on the basis that each functional principle component describes different seasonal variation patterns, and have differing connections to year-to-year temperature variations and differing connections to a temporal trend (possibly reflective of state-specific seasonal climate change patterns). For example, temperature variations within spring and summer periods may have a differing impact on an agricultural company than temperature variations within the autumn and winter periods. Additionally as an example, the fourth and sixth principle components of California's temperature data were most strongly associated with time (with respective p -values of 0.016 and 0.072 from just the 2002–2020 data and together accounting for 14.9% of the year-to-year temperature variance). By exploring all predictor combinations, we account for the situation in which specific seasonal variation patterns or climate change trends are associated with financial performance impacts, rather than limiting the model to modes of maximal variation.

4.4.2 Hypothesis Test

Although this approach enables precise identification of plausible linear relationships, the trade-off is that many models are fit to the data, and by nature of selecting the model with maximal

adjusted- R^2 value, p -values associated with each regression predictor are insufficient in determining statistical significance. To address this issue, a non-parametric permutation test is implemented to find the distribution $p\left(\max\{R_{adj}^2\}|H_0\right)$ so that the null hypothesis may be rejected if we obtain an extreme p -value. This test is interpreted as asking the question: if there is no true relationship, are we likely to identify a relationship with similar performance to what we actually found? If not, then we may reject the assumption of there being no true relationship.

4.5 Classification Model

4.5.1 Logistic Regression

Logistic regression is an in predicting (or regression) discrete results from a set of continuous and/or categorical observations. Each observation is independent, and the probability p that an observation belongs to this class is a function that describes the characteristics of the observation. Consider a set of n observations $[x_i, y_i; Z_i]$ where x_i, y_i are the eigenvalues of the i th observation. If the i th observation belongs to this category, Z_i is equal to 1, otherwise equal to 0. The probability of obtaining n such observations is simply the product of their probabilities (x_i, y_i) .

$$L = \prod_{i=1}^n P_i^{Z_i} (1 - P_i)^{(1-Z_i)} \quad (12)$$

$$\log L = \sum_{i=1}^n [Z_i \log \frac{p_i}{1 - p_i} + \log(1 - p_i)]$$

The odds ratio $p/(1-p)$ is uniform on the decision boundary because $p = 1-p = 0.5$. If we define the function $f(x, y; c)$ as:

$$f(x, y; c) = \log \frac{p}{1 - p} \quad (13)$$

$$p = \frac{1}{1 + e^{-f}}$$

Then $f(x, y; c) = 0$ will be the decision boundary. c is the m parameters in the function f . The log likelihood of f is:

$$\log L = \sum_{i=1}^n [Z_i f_i - \log(1 + e^{f_i})] \quad (14)$$

All that remains to do now is to find a set of values for the parameter c that maximize $\log(L)$ in Equation 14. We can apply optimization techniques or solve a coupled set of m nonlinear equations $\log(L)/dc = 0$ for c .

When Logistic regression solves nonlinear problems, it is necessary to perform nonlinear feature transformation on the data to make the data linearly separable in higher dimensions.

You can replace variables by adding link functions. In the logistic regression technique, variable transformation is performed to improve the fit of the model to the data.

4.5.2 Support Vector Machine

Support Vector Machine (SVM) is a linear classification function, while kernel function is a non-linear function which could convert 2 dimensional data into 3 dimensions. Therefore, a SVM with kernel trick can solve a non-linear relationship by finding a linear hyperspace through adding data dimensions ([Hofmann, 2006](#)).

SVM is a linear classifier by finding the linear boundary between two clusters. There are margins are defined as

$$W^T * X + b = 1 \quad (15)$$

and

$$W^T * X + b = -1 \quad (16)$$

For the class labeled as "1", the constrain is

$$W^T * X + b \leq 1 - e_i \quad (17)$$

and for the class labeled as "-1", the constrain is

$$W^T * X + b > -1 + e_i. \quad (18)$$

In a perfect linear separate situation where $e_i = 0$, there will be no data points between two margins and all points will be perfectly separated. The SVM function in this situation is called Hard Margin SVM. In contrast, the Soft Margin SVM is a SVM function with eased margin which means there could be data points between margins even on the wrong side of the boundary. In addition to the case of Hard Margin SVM, it also includes $e_i < 1$, classifying data correctly within the margin and $e_i \geq 1$, misclassifying the data points ([Murty and Raghava, 2016](#)).

Although SVM performs well on linear classification, the practical problems are non-linear. To solve the issue, kernels will be introduced. As a function, kernel is defined as

$$K(u, v) = \phi(u)' \phi(v) \quad (19)$$

where u and v are the coordinates of data points. In this way, data points with a adding feature can be convert from a 2 dimensions plane into a 3 dimensions space. Then a non-linear dataset can be linearly separated by a hyper-plane in the space. Therefore, a non-linear relationship can be classified by a SVM with kernel trick ([Murty and Raghava, 2016](#)).

4.5.3 Random Forest

Random forest is a simple and efficient nonlinear model, mainly used to solve regression and classification problems. It is generally generated from top to bottom. Each decision or event may lead to two or more events, leading to different results. Drawing this decision branch into a graph similar to the branches of a tree. The weak classifier of random forest uses decision tree and there are multiple decision trees in the forest. When the dependent variables of the data set are discrete values like our variables, it is a classification tree, which can solve the classification problem well. Unlike the Bagging algorithm (only the rows of the training set are sampled when sampling, the attributes are not sampled), when the random forest constructs each decision tree, the sampling needs to be completely independent. "Random" makes it resistant to overfitting, while "Forest" makes it more accurate.

At present, the more popular methods include information gain, gain rate, Gini coefficient and chi-square test. The feature selection based on GINI is used here, because the CART decision tree used by random forest selects features based on the Gini coefficient. The Gini index replaces the entropy gain in the ID3 algorithm or the information gain ratio in the C4.5 algorithm.

For a general decision tree, if there are a total of K categories, the probability that the sample belongs to the k-th category is: p_k , then the Gini index of the probability distribution is:

$$Gini(P) = \sum_{k=1}^K P_k(1 - P_k) = 1 - \sum_{k=1}^K P_k^2 \quad (20)$$

The larger the Gini coefficient, the greater the uncertainty; the smaller the Gini coefficient, the smaller the uncertainty, and the more thorough and clean data segmentation. For the CART tree, since it is a binary tree, it can be expressed as:

$$Gini(p) = 2p(1 - p) \quad (21)$$

When we traverse each segmentation point of each feature, when using feature $A=a$, divide D into two parts, namely D1 (sample set satisfying $A=a$) and D2 (sample set not satisfying $A=a$). Then under the condition of characteristic $A=a$, the Gini index of D is:

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (22)$$

$Gini(D)$: Represents the uncertainty of set D. $Gini(A, D)$: Represents the uncertainty of the set D after $A=a$ segmentation. Each CART decision tree in the random forest continuously traverses all possible split points of the feature subset of this tree, finds the split point of the

feature with the smallest Gini coefficient, and divides the data set into two subsets until the stop condition is met.

Random forest has a good anti-overfitting advantage. When selecting features for each tree, they are randomly generated from all m features, which reduces the risk and tendency of overfitting. The model will not be determined by a specific feature value or feature combination, and the increase in randomness will not infinitely increase the fitting ability of the control model.

In addition, random forest has made improvements in the establishment of decision trees. For an ordinary decision tree, an optimal feature will be selected from all m sample features on the node to divide the left and right sub-trees of the decision tree. However, each tree of RF is actually part of the selected feature. Among these features, choosing an optimal feature to divide the left and right sub-trees of the decision tree expands the influence of randomness and further enhances the generalization ability of the model.

5 Results

5.1 Regression Model

The plausible climate-CFP measure relationships identified by choice of maximal R_{adj}^2 are shown in Appendix 8.3 Table 13. The corresponding p -values obtained by the permutation test are shown in Table 6. These p -values are estimators of $\mathbb{P}\left(\max R_{adj}^2 < (\max R_{adj}^2 | H_0)\right)$. It was decided in advance to use a significance level of 0.05 meaning that $p < 0.025$ and $p > 0.975$ are statistically significant. Only two of 33 models achieved this, namely *CurrentRatio* as predicted by Ohio's *PC3* and *PC6* for company with cik 825542, and *ROE1* as predicted by Massachusetts's *PC2*. The first of these corresponds to warmer spring and early summer periods as well as cooler late summer and autumn periods. This does not have a significant correlation with time (i.e. climate change) with a t-test p -value of $p = 0.561$. The second of these corresponds to cooler winter periods and also does not have a significant correlation with time with a t-test p -value of $p = 0.255$. It is noteworthy that neither of these statistically significant relationships have bushfires (*BurntArea*) as a predictor.

The corresponding p -values obtained for each identified relationship from a t -test are shown in Table 7. These result in 11 relationships which achieve $p < 0.05$ and may be meaningful relationships for future analysis. However, based on how this study was conducted this translates to just two statistically significant relationships after accounting for the data mining of climate-CFP relationships.

Although these two relationships are statistically significant, it is possible that they are

CIK	CurrentRatio	NetMargin	OperatingMargin	ROA1	ROE1
1441693	0.816	NA	NA	0.83	NA
1159275	NA	0.33	0.262	0.332	0.574
825542	0.978	0.294	0.054	0.93	NA
1548240	NA	0.682	0.694	NA	NA
5981	0.832	NA	NA	0.816	0.842
3545	0.328	NA	NA	0.86	0.846
1705843	0.82	NA	NA	0.7	NA
1425292	0.772	NA	NA	0.914	0.948
37785	0.68	0.217	0.218	0.276	0.496
1121702	0.296	0.566	0.6	0.048	0.992

Table 6: p -values obtained for $\mathbb{P}\left(\max R_{adj}^2 < (\max R_{adj}^2 | H_0)\right)$ for each relationship. Values under 0.025 and above 0.975 are considered statistically significant.

false positives under H_0 : there is no relationship between climate and financial performance. Assuming the false positive rate is 0.05 (as determined by significance level) then we expect to see 1.65 statistically significant results from 33 tests, with $\mathbb{P}(\text{two or more significant results}) = 0.496$ under a Binomial test assumption.

CIK	CurrentRatio	NetMargin	OperatingMargin	ROA1	ROE1
1441693	0.2823	0.014	0.0342	0.1349	0.0061
1159275	0.0323	0.052	0.0472	0.1228	0.1353
825542	0.4226	0.0709	0.0689	0.3186	0.3607
1548240	0.0242	0.1067	0.122	0.0218	0.0135
5981	0.2903	0.1801	0.1425	0.2246	0.266
3545	0.022	0.0357	0.1806	0.1465	0.1251
1705843	0.401	0.367	0.3709	0.2686	0.1114
1425292	0.0661	0.1746	0.1978	0.2006	0.2744
37785	0.2002	0.1897	0.1873	0.2086	0.3646
1121702	0.0349	0.0714	0.0705	0.0097	0.4204

Table 7: Table: t test

5.2 Classification Model

In terms of classification models, we use the leave one out cross-validation (LOOCV) method to compute the prediction accuracy for all three models. SVM employs an additional measurement to quantify the model generalization, which is the proportion of support vectors in the total data length.

Model	Logistic Regression		Support Vector Machine		Random Forest		
Reduction	ICA	fPCA	ICA	fPCA	ICA	fPCA	
CFP							Avg.
CurrentRatio	<u>0.521</u>	0.465	0.480	0.453	0.491	0.462	<u>0.479</u>
NetMargin	0.373	0.340	0.389	<u>0.306</u>	0.417	0.423	0.375
OperatingMargin	0.350	0.354	0.360	0.357	0.439	0.413	0.379
ROA	0.450	0.466	0.396	0.374	0.516	0.473	0.446
ROE	0.391	0.463	0.375	0.409	0.474	0.468	0.430
Avg.	0.417	<u>0.418</u>	<u>0.400</u>	0.380	<u>0.467</u>	0.448	0.422

Table 8: The Overall Avg. Pred. Accuracy of All Classifiers

Table 8 displays the average prediction accuracy of the three models with two dimension reduction technologies. These results are computed in the following steps. First, we use LOOCV to find the prediction accuracy for each company and all CFPs at a specific model setting, such as logistic regression with 4-component ICA. Then average the prediction accuracy for one CFP. And repeat the above two steps until complete all settings. According to Table 8, the maximum prediction accuracy is 0.521 and the average prediction accuracy over all settings is 0.422. Considering that we only used two climate events, temperature and bushfire, for forecasting, and the company’s CFP is also related to many other non-climate factors, such as the company’s strategic arrangements, the country’s policies (Capon et al., 1990), and international emergencies (Horváthová, 2010), these results are acceptable. Meanwhile, ICA outperforms fPCA in SVM (kernel trick) and random forest, and has a close performance in logistic regression, which indicates the ICA captures more temperature patterns than fPCA. Random forest far exceeds SVM and logistic regression, which suggests the relationship between the climate events and CFP is non-linear instead of linear. As for CFP, the current ratio gets the best performance, and it is computed from $TotalAsset/TotalLiability$, where Total Asset and Total Liability are compulsory accounting accounts in Financial Statement with compulsory account names. Compared with other selected CFP measurement, current ratio is the only ratio that can measure the liquidity of a company and the prediction results indicate that climate

Evaluation	Avg. Pred. Acc. of fPCA					Avg. Pred. Acc. of ICA				
Quartile	Min	Q1	Median	Q3	Max	Min	Q1	Median	Q3	Max
CFP										
CurrentRatio	0.0	0.3	0.545	0.615	1.0	0.0	0.333	0.556	0.647	1.0
NetMargin	0.0	0.167	0.333	0.5	0.75	0.0	0.111	0.385	0.6	0.8
OperatingMargin	0.0	0.12	0.349	0.454	1.0	0.0	0.046	0.4	0.522	1.0
ROA	0.0	0.353	0.444	0.615	0.9	0.0	0.214	0.5	0.571	1.0
ROE	0.0	0.333	0.444	0.6	1.0	0.0	0.0	0.455	0.667	1.0
Avg.	0	0.255	0.423	0.557	0.93	0	0.141	0.459	0.601	0.96

Table 9: The Results of Logistic Regression

Evaluation	Avg. Pred. Acc. of fPCA					Avg. Pred. Acc. of ICA				
Quartile	Min	Q1	Median	Q3	Max	Min	Q1	Median	Q3	Max
CFP										
CurrentRatio	0.0	0.333	0.436	0.686	1.0	0.1	0.292	0.466	0.692	1.0
NetMargin	0.0	0.344	0.428	0.543	1.0	0.0	0.333	0.414	0.5	1.0
OperatingMargin	0.0	0.3	0.38	0.542	0.889	0.0	0.264	0.434	0.657	0.889
ROA	0.111	0.333	0.464	0.593	1.0	0.222	0.4	0.472	0.65	1.0
ROE	0.0	0.225	0.522	0.657	1.0	0.0	0.29	0.472	0.627	1.0
Avg.	0.022	0.307	0.446	0.604	0.978	0.064	0.316	0.452	0.625	0.978

Table 10: The Results of Random Forest

events have a more significant impact on the liquidity of a company rather than its profitability.

Then we analyze the results at the model level. Quartile means dividing the prediction accuracy of all companies into 4 equal parts of the same size, and $Q1$ represents the value at 25% and $Q3$ indicates the value at 75% of the population. According to Table 9, some companies show a promising prediction accuracy, 100%, and some do not support a classification model with 0 prediction accuracy. Meanwhile, Table 10 has a similar result. Both logistic regression and random forest, ICA has a more stable prediction accuracy distribution and a higher value. In terms of support vector machine (Table 11 and 12), ICA also outperforms FPCA and ICA has a fewer proportion than FPCA, which indicates the SVM with ICA has a better generalization than SVM with FPCA.

In a nutshell, classification models have an acceptable results. Random forest indicates the non-linear classifier have a stronger predictive ability, and ICA captures more temperature

Evaluation	Avg. Pred. Accuracy					Proportion				
Quartile	Min	Q1	Median	Q3	Max	Min	Q1	Median	Q3	Max
CFP										
CurrentRatio	0.0	0.222	0.429	0.667	1.0	0.929	0.972	1.0	1.0	1.0
NetMargin	0.0	0.111	0.231	0.462	0.75	0.946	1.0	1.0	1.0	1.0
OperatingMargin	0.0	0.134	0.333	0.528	1.0	0.95	1.0	1.0	1.0	1.0
ROA	0.0	0.222	0.4	0.471	0.889	0.95	1.0	1.0	1.0	1.0
ROE	0.0	0.111	0.438	0.538	1.0	0.893	1.0	1.0	1.0	1.0

Table 11: The Results of 4 fPCs and SVM

Evaluation	Avg. Pred. Accuracy					Proportion				
Quartile	Min	Q1	Median	Q3	Max	Min	Q1	Median	Q3	Max
CFP										
CurrentRatio	0.0	0.2	0.556	0.667	1.0	0.411	0.871	0.9	0.944	1.0
NetMargin	0.0	0.0	0.444	0.667	0.75	0.625	0.844	0.93	0.972	1.0
OperatingMargin	0.0	0.015	0.444	0.597	1.0	0.75	0.881	0.93	0.96	1.0
ROA	0.0	0.0	0.5	0.6	1.0	0.7	0.85	0.911	0.972	1.0
ROE	0.0	0.0	0.462	0.667	1.0	0.511	0.82	0.897	0.973	1.0

Table 12: The Results of 4 ICs and SVM

pattern than FPCA and ameliorates the model generalization.

6 Discussion and Future Directions

This study did not find evidence of a linear regression relationship or nonlinear classification relationship between corporate financial performance and temperature and/or bushfires. It is worth noting that this is not necessarily evidence for the null hypothesis, and the relationship explored was with climate factors in general not necessarily climate change.

The results described by this study are a first in exploring this relationship of interest and have attempted to address the question: is CFP impacted by temperature variations and bushfires?

A strength of this study is the use of detailed company data and detailed climate data, which are connected on a monthly time-scale and state-level spatial-scale. An additional strength of this study is the exploration of multiple relationship types for precise relationship identification.

However, this study has a few limitations which are intended to be addressed in future work.

Two limitations, which would be integral to future developments, include the fact that company operations are not limited to the state in which they are headquartered and the focus of this work on direct climate impacts without accounting for indirect impacts. Further limitations include: weak/moderate strength relationships may not achieve statistical significance with sample sizes of < 20 data points, each company CFP measure is influenced by many factors which are not explicitly accounted for in the above models, and our study explores many possible climate-CFP relationships when specific climate-CFP relationships may be of interest were we to have company-specific expertise. Each of these limitations are elaborated upon below.

Firstly, the use of company headquarters alone to locate a company appears to be a plausible assumption for companies with operations located entirely within the headquarter state and companies with operations located within the headquarter state and neighbouring states (i.e. ones from a similar climate region). However, this does not appear to be sufficient for multinational companies and companies operating in U.S. states from diverse climate regions. Additional geographic locations of interest may also be indirect such as supply chain locations and end-use or market locations. Unfortunately there is no standard for companies to report operation locations nor supply-chain, end-use, or market locations. This limitation was not addressed in this study as additional geographic locations leads to more predictors within models which are fit to a limited number of data points. However, taking an industry-wide perspective may enable this information to be better suited to analysis.

Secondly, this study focussed on direct climate impacts and did not include indirect climate impacts. Some indirect impacts which may be of interest include the impact of climate change risk on insurance premiums, and impacts to supply chain and end-users. As suggested above taking an industry-wide perspective is one way to enable analysis of these suggestions and is the direction of future work.

Finally, each company CFP measure is influenced by many factors which are not explicitly accounted for in this study's modelling. Company financial performance is determined by many factors of which climate was hypothesised to be one. Without detailing these factors their impacts are confined to the error terms within these models. By describing core CFP processes, models may better disentangle the hypothesised influence of climate factors on observed financial performance. Similar to the first two limitations, exploring this would be better done on an industry level and for which our dataset would be useful.

Future directions for this line of inquiry, exploring any climate-CFP relationships, which are suggested include assessing the connection between other natural disasters and other climate variables (such as flash-flooding events and precipitation), exploring industry-level relationships

to address limitations of this work, and expanding this study’s approach to assess possible relationships with companies of other industries such as the tourism industry and the energy industry.

7 Conclusion

In this project, we investigate the relationship between climate change and financial performance and found limited evidence supporting this hypothesis. Linear regression and nonlinear relationships were modelled with the regression models finding few statistically significant relationships of CFP with temperature which could not be distinguished from being false positives and with classifications displaying acceptable prediction accuracy especially for current ratio, which suggests that climate change correlates to liquidity of the company rather than profitability.

Moreover, we manually obtained financial data to construct a climate-finance dataset, which is the first dataset to the best of our knowledge which combines temperature and bushfire data with accounting-based financial ratios by month, by state, and at the company level for 26 agricultural companies. We believe that this dataset will be beneficial to future research and provides a framework for constructing datasets to be used in assessing hypotheses about the relationship between climate and company financial performance.

We have investigated both linear and non-linear correlation relationships using different models. Extending to causal relationships may be the subject of future work. In conducting this study many directions for future work became evident, with future work exploring both direct and indirect impacts as well as geographic locations of all company operations being of most interest. Furthermore, assessing the relationship at the industry level may provide further insights into the focus of this study as well as enabling these suggested directions for future work.

References

- Ardabili, S., Mosavi, A., Dehghani, M., and Várkonyi-Kóczy, A. R. (2019). Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. In *International Conference on Global Research and Education*, pages 52–62. Springer.
- Capon, N., Farley, J. U., and Hoenig, S. (1990). Determinants of financial performance: a meta-analysis. *Management science*, 36(10):1143–1159.
- Colan, S. D. (2013). The why and how of z scores. *Journal of the American Society of Echocardiography*, 26(1):38–40.
- Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.
- Diaz-Rainey, I., Robertson, B., and Wilson, C. (2017). Stranded research? leading finance journals are silent on climate change. *Climatic Change*, 143(1):243–260.
- for Environmental Information, N. N. C. (2021). The gcos essential climate variable (ecv) data access matrix.
- Galant, A. and Cadez, S. (2017). Corporate social responsibility and financial performance relationship: a review of measurement approaches. *Economic research-Ekonomska istraživanja*, 30(1):676–693.
- GHAYOUR, H. (2005). Classification of temperature regime of iran using pca and ca.
- Hannachi, A., Unkel, S., Trendafilov, N., and Jolliffe, I. (2009). Independent component analysis of climate data: a new look at eof rotation. *Journal of Climate*, 22(11):2797–2812.
- Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.
- Hofmann, M. (2006). Support vector machines-kernels and the kernel trick. *Notes*, 26(3):1–16.
- Horváthová, E. (2010). Does environmental performance affect financial performance? a meta-analysis. *Ecological economics*, 70(1):52–59.
- Hyvärinen, A. (1999). Survey on independent component analysis.
- Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492.

- Lu, W., Atkinson, D. E., and Newlands, N. K. (2017). Enso climate risk: predicting crop yield variability and coherence using cluster-based pca. *Modeling Earth Systems and Environment*, 3(4):1343–1359.
- May, G. O. (1937). Principles of accounting. *Journal of Accountancy (pre-1986)*, 64(000006):423.
- Moonen, A., Ercoli, L., Mariotti, M., and Masoni, A. (2002). Climate change in italy indicated by agrometeorological indices over 122 years. *Agricultural and Forest Meteorology*, 111(1):13–27.
- Moradkhani, H. and Meier, M. (2010). Long-lead water supply forecast using large-scale climate predictors and independent component analysis. *Journal of Hydrologic Engineering*, 15(10):744–762.
- Murty, M. and Raghava, R. (2016). Kernel-based svm. In *Support vector machines and perceptrons*, pages 57–67. Springer.
- Ndehedehe, C. E. and Ferreira, V. G. (2020). Identifying the footprints of global climate modes in time-variable gravity hydrological signals. *Climatic Change*, 159(4):481–502.
- Ohlson, J. A. and Juettner-Nauroth, B. E. (2005). Expected eps and eps growth as determinantsof value. *Review of accounting studies*, 10(2):349–365.
- Organization, W. M. (2021a). Essential climate variables.
- Organization, W. M. (2021b). Natural hazards and disaster risk reduction.
- O’Neill, B. C., Carter, T. R., Ebi, K., Harrison, P. A., Kemp-Benedict, E., Kok, K., Kriegler, E., Preston, B. L., Riahi, K., Sillmann, J., et al. (2020). Achievements and needs for the climate change scenario framework. *Nature climate change*, 10(12):1074–1084.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York.
- Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. (2019). Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*.
- SEC (2021). U.s. securities and exchange commission. <https://www.sec.gov/>. Accessed: 2021-09-30.
- Sun, Y., Yang, Y., Huang, N., and Zou, X. (2020). The impacts of climate change risks on financial performance of mining industry: Evidence from listed companies in china. *Resources Policy*, 69:101828.

- Tzouvanas, P. (2019). *Climate Change and Financial Performance*. PhD thesis, University of Portsmouth.
- Wang, L. and Kutan, A. M. (2013). The impact of natural disasters on stock markets: Evidence from japan and the us. *Comparative Economic Studies*, 55(4):672–686.
- Worthington, A. and Valadkhani, A. (2004). Measuring the impact of natural disasters on capital markets: an empirical application using intervention analysis. *Applied Economics*, 36(19):2177–2186.
- Worthington, A. C. (2008). The impact of natural events and disasters on the australian stock market: A garch-m analysis of storms, floods, cyclones, earthquakes and bushfires. *Global Business and Economics Review*, 10(1):1–10.
- Zhang, D., Zhang, Z., and Managi, S. (2019). A bibliometric analysis on green finance: Current status, development, and future directions. *Finance Research Letters*, 29:425–430.

8 Appendix

8.1 Document Links

- Project Github link: <https://github.com/lowspace/MAST90106>
- Meeting minutes: <https://github.com/lowspace/MAST90106/tree/main/Meeting%20Minutes>
- Meeting slides: <https://github.com/lowspace/MAST90106/tree/main/Meeting%20Slides>
- Project data: <https://github.com/lowspace/MAST90106/tree/main/data>
- Project code: <https://github.com/lowspace/MAST90106/tree/main/code>

8.2 Data Process

All data in the `joining_dataset` (shown in Code 1) has been standardized at first, and the *actual data* in the below description list means actual standardized data and *the difference* between two pieces of data is the subtraction between two standardized data instead of their raw data.

cik1 The central index key of one company, and this dataset contains all available CIKs in the list since some companies, such as 0001756180, only have two data points or only have one class data which is unacceptable for classifiers if the dataset is used to classify.

state_temp_data A 2D numpy array contains the state temperature data where the company headquarter locates, which reserves to feed to the feature engineering methods, such as FPCA and ICA mentioned in section 4.3. The row of this array is the yearly timestamp, and the columns are the same as the `temp_data`'s.

This is part is consistent either for regression or classification.

cfp1 One financial ratio of `cik`. Each `cik` shares the same CFP list mentioned in section 4.1, though different CFP may has different length due to original data issue and computation method.

cfp_data The CFP data of `cfp1`.

If the dataset works for regression task, then the data is the actual CFP data, otherwise is the binary value of CFP data which indicates the positive or negative of the difference

```

joining_dataset =
    { # a nested dictionary
      cik1:{ # the key of this dict is the CIK of each company
        state_temp_data: np.array # a 2D numpy array
        cfp1:{ # the key this dict the CFPs listing in the section 5.1
          cfp_data: list, # a list includes the cfp data at specific year
          temp_data: np.array, # a 2D numpy array includes the temp data at the
                        # same year with the cfp_data
          bush_data: list, # a list includes the bushfire data at the
                        # same year with the cfp_data
        },
      cik2{...}, ..., # other CFPs
    }
    cik2:{...}, cik3:{...}, ... # other CIKs
  }

```

Code 1: Data structure of the joining dataset

between the recent year data and its previous year data, such as 1 implies the data of 2020 is greater than 2019's, while 0 means smaller.

temp_data A 2D numpy array contains the temperature data of **cfp1**. The row of this array is the yearly timestamp corresponding to the data points of **cfp_data**, such as the first element of **cfp_data** is from 2020 and the first row of **temp_data** is from 2020, and the columns are the monthly timestamps of that reporting year, such as this **cik** publishes its report in 2020/10/31 and the columns are 2019/11, 2019/12, 2020/1, ..., and 2020/10. This part is reserved to feed to the transformer⁸ trained by FPCA and ICA, and the transformed data servers as a part of feature data of the model.

The data is the difference between recent year temperature data and its previous year data either for regression model or for classification model. Typically, the number of rows for regression is longer than the classification's.

bush_data A list includes the bushfire data of **cfp1**, the bushfire we get is yearly instead

⁸Transformer is an object to transform the original data into the processed manner. More information on <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html#sklearn.decomposition.FastICA.transform> and <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html#sklearn.decomposition.FastICA.transform>.

monthly as temperature data.

The data is the difference between recent year bushfire data and its previous year data either for regression model or for classification model. Typically, the list length for regression is longer than the classification's.

Both temperature and bushfire dataset have much more data points than the financial dataset's, which means the time range of temperature and bushfire dataset covers the financial and we can obtain sufficient corresponding climate data points for `temp_data` and `bush_data`.

We solve some challenges while grouping the two datasets. The first one is the discontinuity in financial dataset. We record the yearly timestamp for each `cfp` data points, then match the climate data with the same yearly timestamp. For example, in classification model `cfp1` has the data points in 2020, 2019, 2010, 2009, ..., and 2002, then `temp_data` and `bush_data` only stores the data with the same yearly timestamp. The second is the inconsistency of the reporting period of some companies, and the inconsistency includes the various periods of different company, such as `cik1` publishes in June while `cik2` reports in December, and the same company may change its reporting period, such as in 2020 it reports in July while in 2019 it reports in October. To handle the former, we use the month of the reporting day as the anchor and count forward 11 months, which is used to generate the full-year temperature data, `temp_data` and `state_temp_data`. Most of companies reports at the last day of one month, rounding the days if the companies with irregular reporting day, 1th to 10th servers as the previous month, otherwise this month. To handle the latter, we assume that the reporting time of two consecutive years is the same if the time difference between them is less than 32 days, and classifiers indulge in this setting due to more data points. For example, `cik1` reports at 2020/12/31 and 2019/12/13, we think it both reports in December and can do subtraction over the two data points; `cik2` reports at 2020/7/31, 2019/10/31, and 2018/10/31, we remove the 2020's data and do subtraction on 2019's and 2018's. The third one is the missing value in bushfire data. When we complete using the timestamps of the `cfp` data points to match the corresponding bushfire and temperature data, the `cfp_data`, `temp_data`, and `bush_data` all have the same length, then we exhaust the `bush_data` to check whether the element is a `nan` value, if yes, we delete index corresponding to the `nan` value for all the three. For example, the third element of `bush_data` is `nan`, then we delete the third element for `cfp_data`, `temp_data`, and `bush_data`.

8.3 Results

CIK	CurrentRatio	NetMargin	OperatingMargin	ROA1	ROE1
1441693	PC2+PC3+PC5 +PC6+BurntArea	PC1+PC3+PC4 +PC6+BurntArea	PC1+PC4+PC5 +PC6+BurntArea	PC3+PC5 +BurntArea	PC1+PC3 +BurntArea
1159275	PC1+PC5+PC6 +BurntArea	PC1+PC3+PC4 +BurntArea	PC1+PC3+PC4 +BurntArea	PC1+PC2+PC3 +PC5+PC6 +BurntArea	PC2+PC3+PC4 +PC5
825542	PC3+PC6	PC1+PC2+PC3 +PC5+BurntArea	PC1+PC2+PC3 +PC4+PC5 +BurntArea	PC3+PC5	PC4
1548240	PC1+PC2+PC3 +BurntArea	PC2+PC6 +BurntArea	PC2+PC6 +BurntArea	PC3+PC6	PC3+PC6
5981	PC1+PC3 +BurntArea	PC1	PC1+PC4	PC1+PC3	PC1+PC3
3545	PC3+PC6	PC2	PC2	PC2	PC2
1705843	PC1	PC3+BurntArea	PC3+BurntArea	PC1	PC1+PC2+PC3 +BurntArea
1425292	BurntArea	PC2	PC2	PC2	PC1
37785	PC1+PC2+PC4 +BurntArea	PC1+PC2+PC3 +PC4+BurntArea	PC1+PC2+PC3 +PC4+BurntArea	PC1+PC2+PC3 +PC4+PC5 +BurntArea	PC1+PC2+PC3 +PC4+PC5 +BurntArea
1121702	PC2+PC6 +BurntArea	PC3+PC5	PC3+PC5	PC3+PC4+PC6 +BurntArea	PC2

Table 13: Identified relationships with maximal R_{adj}^2 value for each CFP measure and for each company.

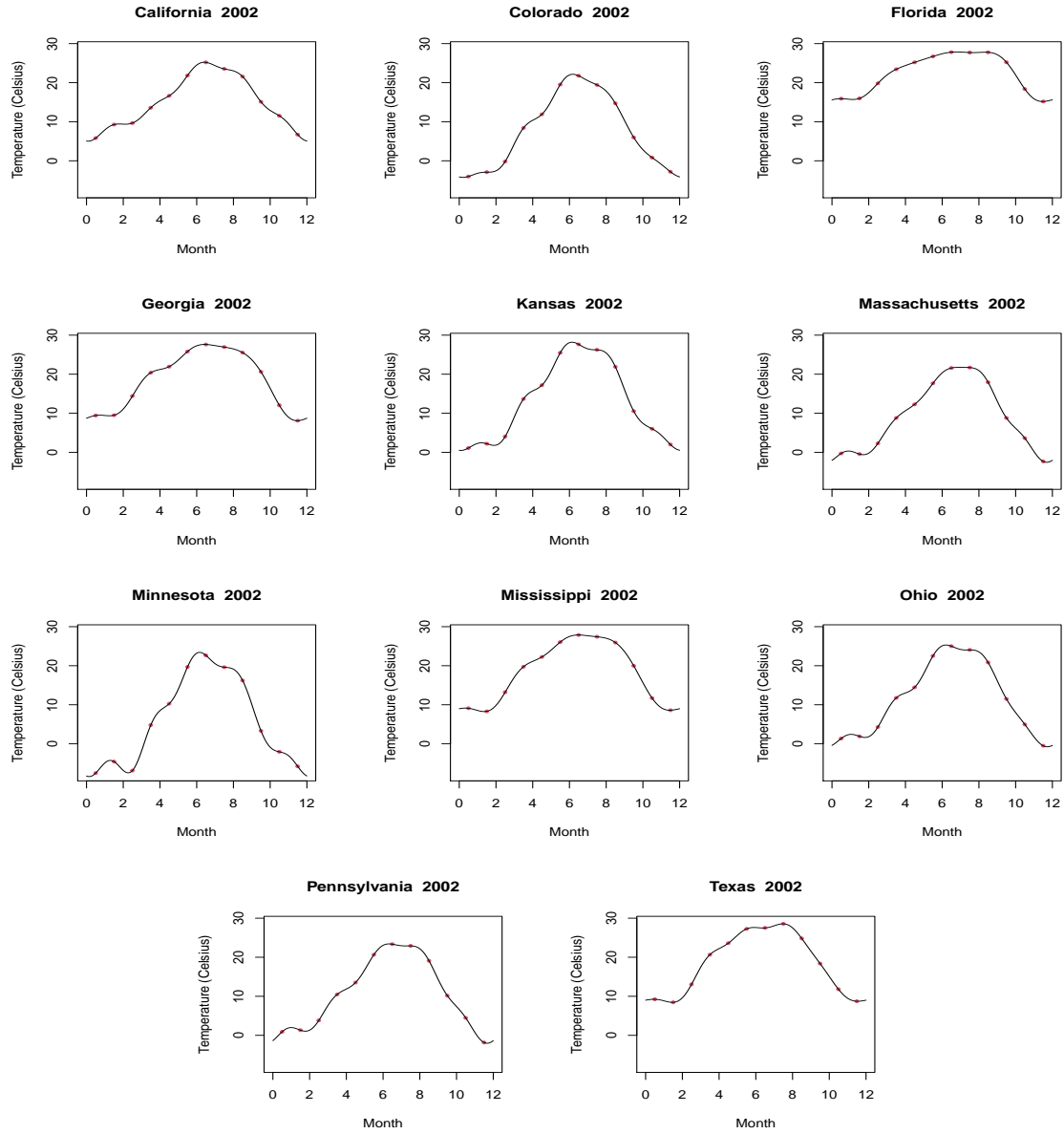


Figure 5: Functional data representations of raw temperature data for each state where at least one company from our dataset is headquartered from the year 2002.