CSC110 Project:
Impact of climate change on wildfires in California

Yuzhi Tang, Zeyang Ni, Junru Lin, Jasmine Zhuang

December 14, 2020

Problem Description and Research Question

Wildfire and climate change are two hot topics nowadays. The frequent occurrence of the catastrophic wildfire, one of the most serious ecological disasters, has aroused people's concerns around the world. Meanwhile, climate change has also caused a series of ecological and environmental problems.

Climate change has manifested itself in many ways. In recent years, there are observable changes in temperature, precipitation, soil moisture, relative humidity, wind speed, etc. at various geological levels. High temperature contributes to a better state and larger amount of the fuel. More precipitation means more plants can act as the fuel in a forest fire, and lead to less possibility for wildfire spreading. As for soil moisture, relative humidity and wind speed, these factors can also affect the occurrence and propagation of wildfires.

We then want to know whether climate change have anything to do with wildfire. We then did some research and found some paper suggesting the possible link between climate change and wildfires, which we will then investigate further.

One of the paper found shows that there can be closed relation between forest fire and climate change, that is, ”Climate change resulting from human activities has nearly doubled the area burned by forest fires in the western United States over the past three decades.” (Climate change: Wildfires burn more US forest, 2016) Also, M.D Flannigan, B.J Stocks, B.M Wotton(2000, October) thought the impact of climate change in America ”could have an almost immediate and significant impact on ecosystems, due to likely increases in area burned and fire intensity/severity. ” Furthermore, S.A. Parks, J.T. Abatzoglou(2020, October) said that ”warmer and drier fire seasons corresponded with higher severity fire, indicating that continued climate change may result in increased fire severity in future decades.”

Since climate change can possibly influence wildfires shown by the above paper found, we would like to explore this question: **how much climate change has impacted the frequency and intensity of wildfires in California?**. Among a number of aspects of climate change, we choose the change of precipitation and temperature to explore how the fire size and the occurrence vary from 1994 to 2013. California is a state in America that is shown to have the highest frequency of wildfire occurrence. Therefore, the it is chosen as the typical research object.

In our project, precipitation and temperature will be independent variables, while the occurrences and area of wildfires will be dependent variables. We choose the number of wildfires per year to represent the frequency of the occurrence of wildfires, which is the most intuitive reference factor for it.

The majority of the damage caused by climate change tends to be invisible, while wildfires are considered to caused significant socioeconomic damage every year. Thus, we would like to find one of the implicit hazards of climate change in terms of wildfires. By exploring the quantitative relationship between the two, we wish to test our hypothesis that climate change has significantly increased the severity of wildfires, which is helpful for calling for people to pay more attention to this global issue, climate change and adopt to a low-carbon lifestyle.

Dataset Description

- **Dataset 1:** Monthly Precipitation and Temperature Data for California U.S. from Jan, 1989 to Oct, 2020

Dataset description:
This dataset includes monthly precipitation and temperature(in fahrenheit) data made by lots of weather stations in California. Some relevant attributes include the mean, mean max, and mean min temperature in a given month, and total precipitation in a given month.

Subsets of data used:
In 'ca_climate.csv' which consists of data observed in California, we will use the data of temperature stored in the following columns : 'DATE', 'TAVG',"TMAX', 'TMIN', where"TAVG' stores data of mean temperature, 'TMAX' stores data of mean max temperature, and 'TMIN' store data of mean min temperature. We will also use the data of precipitation stored in the following columns:'DATE', 'PRCP', where 'PREP' stores data of total precipitation.

Source: National Centers for Environmental Information.

Website: https://www.ncdc.noaa.gov/cdo-web/search.

Data format: csv.

- **Dataset 2:** Geo-referenced Wildfire Records in the U.S. from 1992 to 2015

Dataset description:
This dataset includes 1.88 million geographic-referenced records of wildfires that occurred in the U.S. from 1992 to 2015. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. Basic error-checking was performed and redundant records were identified and removed, to the degree possible.

Subsets of data used:
In the file 'wildfire_data.csv', we will use the specific data related to the wildfire happened in California . We will use data for California that is stored in the following columns: 'STATE', 'FIRE_YEAR', 'DISCOVERY_DOY','FIRE_SIZE', where 'STATE' specifics this data belongs to California , 'FIRE_YEAR' indicates the year the fire happened, DISCOVERY_DOY' indicates the date(in a year) of fire that was discovered.
Source: Kaggle.

Website: https://www.kaggle.com/rtatman/188-million-us-wildfires.

Data format: SQLite database (extension: .sqlite).

Computational Overview

- **Data Preparation:**

Before we implement any operations on our datasets, we would need to transfer all the datasets to csv file where SQLiteStudio will be used to extract the data from sqlite file and store it in csv file.

Then, all relevant attributes will be extracted from the datasets. For Dataset 1, these include: the state where the weather station is in, the date when the monthly temperature/precipitation data was recorded, the average temperature of the month, and the total precipitation in the month. For Dataset 2, these include: the state where the wildfire occurred, the size of fire, the statistical cause of the fire, and the year when the fire occurred. The raw data extracted will be transformed into reasonable Python data types (e.g. string, float, datetime) to allow for easy computation.

For the extracted data of Dataset 1, the monthly temperature and precipitation data will be transformed into a yearly format. Average temperature in a year of a state will be calculated by taking the average of all monthly temperature data recorded in that state, in that year. Average precipitation in a year of a state will be calculated by taking the average of all monthly precipitation data recorded in that state, in that year. Then, we would have the monthly temperature and precipitation data for California, stored using a custom Python dataclass.

For the extracted data of Dataset 2, all wildfires that occurred outside of California will be filtered out using pandas.drop function. The wildfire data will also be transformed into a monthly format. Average wildfire size in a month of a state will be calculated by taking the average of the size of all wildfires in that state, in that month. Total wildfire occurrence in a month of a state will be calculated by summing up all wildfire occurrences in that state, in that month. Then, we would have the monthly wildfire size and occurrence data for California, stored using a custom Python dataclass.

- **Computational Models/Algorithms:**

To evaluate the impact of climate change on the frequency and intensity of wildfires, we would first built several possible models for the regression.

- exponential model

$$y = a \cdot (b^x) + c$$

We would use scipy.optimize.curve_fit function to find the best fitted parameters.

- quadratic model

$$y = a \cdot x^2 + b \cdot x + c$$

This model include the linear model in cases that a = 0

- inverse model

$$y = \frac{a}{x} + b$$

- logarithm model

$$y = a \cdot log(x) + b$$

- periodic model

$$a \cdot cos(b \cdot (x - c)) + d \cdot cos(e * (x - f)) + g \cdot x + h$$

We found that periodic function especially useful for temperature and precipitation which tends to have seasonal features, therefore by using this model, we could mimic the periodic fluctuation of temperature and precipitation. The first cos block is supposed to mimic the periodic fluctuation of data each year and the second cos block is supposed to mimic the fluctuation of peak values which also seems to have periodic features. The g· x term is supposed to mimic the linear change of overall temperature and precipitation data.

Criteria:

- RMSE(root-mean-square error)

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y_i})^2}$$

We would use the RMSE value to compare the accuracy of different models. RMSE can be used to compare different sizes of datasets (as long as they're all in the same unit) and to compare different models on the same dataset. We will conduct all models and compare there RMSE value, for the same dataset, the smaller the RMSE, the more accurate the model.

Implementations:

- Perform regression between time(independent variable) and average month temperature(response variable)
- Perform regression between time(independent variable) and average month precipitation(response variable)

By doing these two regressions, we would be able to predict the temperature change in the following decades and precipitation changes in the following decades.

- Perform regression between average month temperature(independent variable) and wildfire frequency(response variable)
- Perform regression between average month temperature(independent variable) and wildfire size(response variable)
- Perform regression between average month precipitation(independent variable) and wildfire frequency(response variable)
- Perform regression between average month precipitation(independent variable) and wildfire size(response variable)

By doing these four regressions, we would find the relationship between climate and wildfire, since we'd already predicated the future climate data, we could then use this model to predict the future wildfire size and frequency.

We would take the following two procedures to conduct regression.

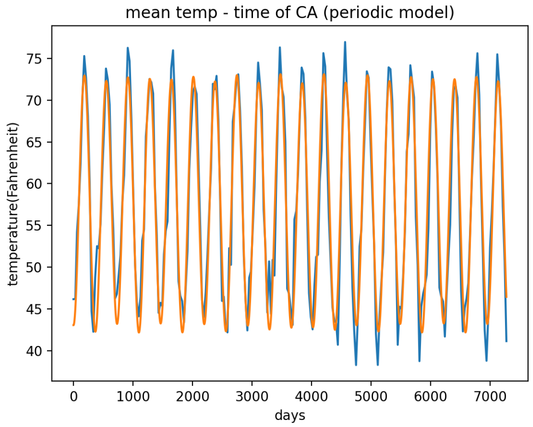
I. For each regression(the relationship between two variables), we will try every model and calculate their RMSE value then pick the model that has the smallest RMSE value.

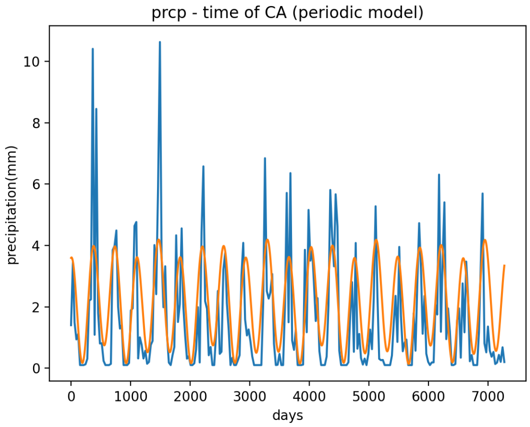
II. Using matplotlib.pyplot to visualize the model, for each figure, we draw the original data as blue line(or dots) and predicted data as yellow line(or dots).

- **Results Reporting:**

After performing the regressions and calculate the RMSE values, we found the most suitable model for each regression. The blue line(or dots) represents the original data and the yellow line represents the fitted model.

- Time - Temperature and Time - Precipitation





We found that the periodic model can best fit the data after the horizontal comparison of each model.

Note: when finding the optimized coefficients for the periodic model, we had to first provide an initial guess of the coefficient values to the scipy optimization function (so as to make the scipy optimization function return better optimized coefficients). Recall the periodic model has the following coefficients:

$$a \cdot cos(b \cdot (x - c)) + d \cdot cos(e * (x - f)) + g \cdot x + h$$

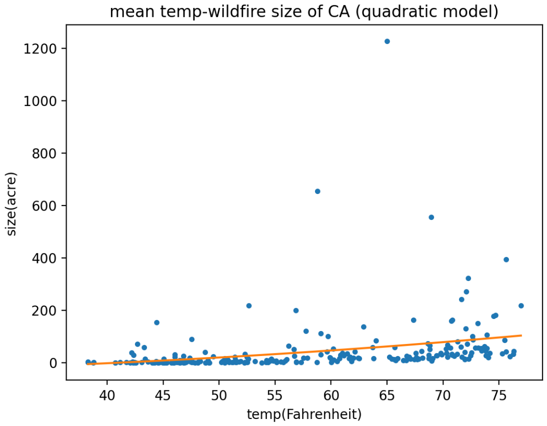
For example, for the mean temperature vs. time periodic model, we chose our initial guess of the coefficient values with the following general considerations in mind:

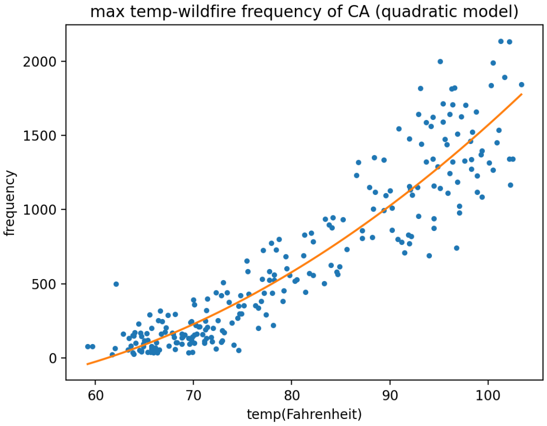
The temperature has an amplitude of about 15, so we estimate a = 15. The climate has a period of 365 days, so we estimate $b = \frac{2\pi}{365}$. The highest temperature of a year is always in July or August, so we estimate $c = 180$. Since the difference between among the highest peaks is about 3, we estimate $d = 1.5$. We find that the the peaks decreases, increase and decrease again in 18 years, so we estimate $e = \frac{2\pi}{12 \cdot 365}$. The first lowest of the high peak is in the 6 year so we estimate $f = 0.5 \cdot (6 \cdot 365)$. Since there is no obvious increase in the temperature, we estimate $g = 0.0001$. The middle point between 45 and 75 is 60, so we estimate $h = 60$.

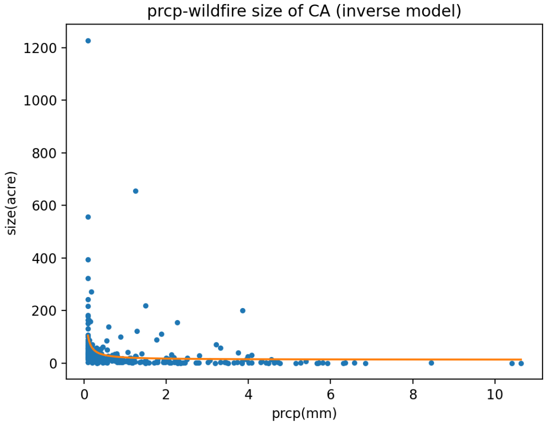
- Temperature - Fire size and Temperature - Fire Frequency

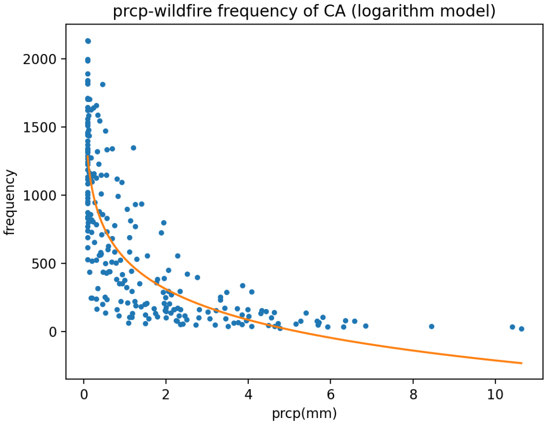
We found the quadratic model best fit the relationship between temperature, wild fire size and temperature, wild fire frequency.

- Precipitation - Fire size and Precipitation - Fire Frequency









We found that the inverse model suitable for precipitation - fire size and the logarithm model suitable for precipitation - fire frequency.

Predictions:

- Temperature and Precipitation prediction

Using our fitted models, we can predict the temperature and precipitation data for the coming years.

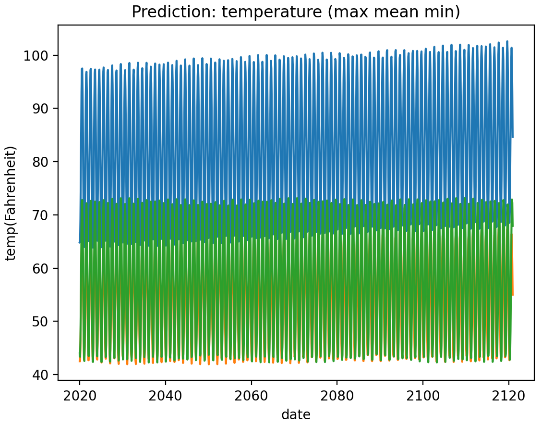
- Wild fire frequency and size prediction

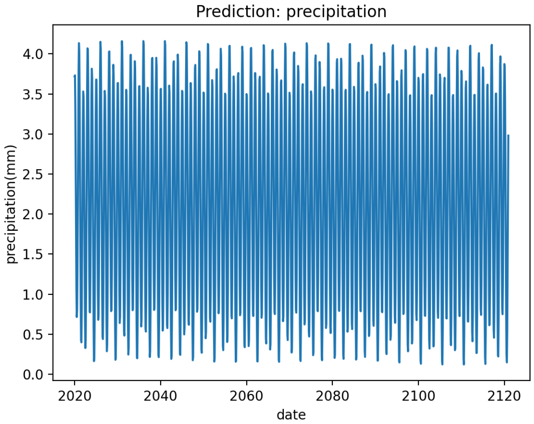
Using our fitted models, we can predict the wildfire frequency and size based on the max temperature and precipitation respectively.

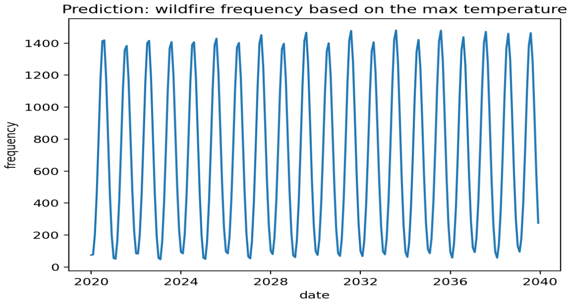
Conclusions:

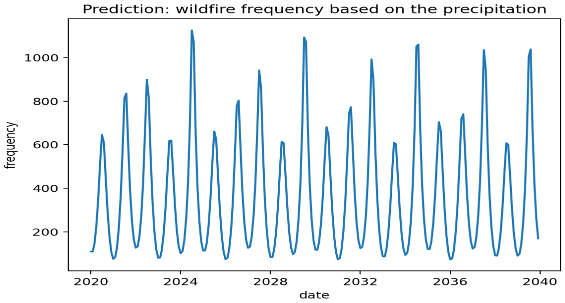
According to our predictions on the temperature and precipitation, we made predictions on the wildfire size and frequency for the coming years:
According to the prediction of temperature, for the upcoming 100 years, the temperature in California tends to increase which can cause higher frequency and slightly greater size of wildfire. And our predictions on precipitation shows that the precipitation in California tends to decrease which based on our model, will also lead to higher frequency of wildfire.

Instructions for obtaining data sets and running the program









1. obtain data sets:We obtain the first dataset from this URL, **it cannot be downloaded directly so we recommend to download it directly from MarkUS**. https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GSOM. We obtain the second dataset from this URL, **this file is quite large and we have a processed version uploaded on markus, if you want to run on the original dataset, please run the file 'data_prep' to process the data.** https://www.kaggle.com/rtatman/188-million-us-wildfires?select=FPA_FOD_20170508.sqlite.
2. filter the data sets: The data sets we use for analysis are *new_ca_climate.csv* and *new_wildfire_data.csv*, which only contain the data we need. They are generated from the data sets *ca_climate.csv* and *wildfire_data.csv*. To successfully run the program, you can download *new_ca_climate.csv* and *new_wildfire*$_d$*ata.csv* to the same folder of python files. **If you want to get *new_ca_climate.csv* and *new_wildfire*$_d$*ata.csv* from *ca_climate.csv* and *wildfire_data.csv*, you don't need to download *new_ca_climate.csv* and *new_wildfire_data.csv*. Instead, you should create two blank csv files named *new_ca_climate.csv* and *new_wildfire_data.csv* in the pyhon file folder and then go to *data_prep.py*, uncomment line 287 and then run this python file. Then two blank csv files will have the data we need.**
3. Run Python file: Go to *main.py*. Uncomment line 47 and run the file, then you will get 20 figures of modelling. After this, comment line 46 and uncomment line 48, you will get another 10 figures of prediction.

Changes to our project plan between proposal and final submission

First, we have changed the title of our project from generally saying "the impact of climate change on wildfires in America" to more specifically saying California which is consistent and effective as our TA advised.

Secondly, we discuss in more details the wildfire and climate change with some articles suggesting the possible link between them; we use inline citation for quotations included in text.

Also, we phrase our research question with question mark.

Moreover, we use more academic language like replacing the sentence "we want to demonstrate that climate change has significantly increased the severity of wildfire" that is too strong by "we wish to test our hypothesis that climate change has significantly increased the severity of wildfires", so that we aren't slating our research towards a predetermined and desire outcome.

In addition to problem description and research question, we have also made a change to our computational plan. We originally planed to use average yearly climate data to analyse the relation between wildfire and it. Then we realized that wildfire can have different kinds of outcome due to seasons(difference of climate between seasons). Thus, we turn to use monthly climate data when computing and modeling.

Moreover, we will use RMSE, an absolute measure of fit to see whether the models fit well, the smaller the value of RMSE, the better the model fits.

And instead of using linear regression, we change to use various models for fitting so that we can have the best chosen models out of multiple models as we previously had no idea how the resultant graphs will be.

Discussion

We will first take a look at the relation between temperature(max,min and mean) and wildfire(frequency and size). Below is the table showing RMSE of the models of frequency of wildfire related to max/min/mean temperature using exponential function and quadratic function, so that we can compare the accuracy of these models.

Dependent Variable	Independent Variable	model	RMSE	function
wildfire frequency	max temperature	exponential	209.347070	$y = 370.393200(1.020766^x) - 1329.127257$
wildfire frequency	min temperature	exponential	241.454917	$y = 275.412073(1.033339^x) - 1046.657859$
wildfire frequency	mean temperature	exponential	222.539099	$y = 84.776346(1.043656^x) - 455.945522$
wildfire frequency	max temperature	quadratic	208.379095	$y = 0.484187x^2 - 37.603840x + 487.663092$
wildfire frequency	min temperature	quadratic	239.079895	$y = 1.156259x^2 - 71.198213x + 1024.829793$
wildfire frequency	mean temperature	quadratic	239.079895	$y = 1.097731x^2 - 81.693246x + 1583.813078$

The overall RMSE values are quite large, which generally means our models is failing to account for important features underlying our data.Thus they are not as accurate as what we expect. The first 3 rows of the table shows that the graph of max-temperature and frequency using exponential model have the smallest RMSE value, meaning this model that use max-temp data is the most accurate.The last 3 rows of the table shows that the graph of max-temperature and frequency using quadratic model have the smallest RMSE value, meaning this model that use max-temp data is the most accurate. Since the data of max temperature is found to be larger then the data of min temperature and mean temperature, and it shows that using max temperature data gives the smallest RMSE for both models, we concludes that the models using max temperature fit the best.
By comparing the RMSE of the max-temperature-wildfire frequency of exponential model and the RMSE of the max temperature-wildfire frequency of quadratic model, we find that the quadratic model has the smaller RMSE. Thus, quadratic model fits better than exponential model for displaying the relation of temperature and fire frequency, and we will choose quadratic model to predict future relation of wildfire frequency and temperature though it is possible that temperature doesn't have much to do with fire frequency for they are not shown to be so related as what we expected.

Below is the table showing RMSE of the graph of the size of wildfire related to max/min/mean temperature using exponential and quadratic models so that we can compare the accuracy of these models.

Dependent Variable	Independent Variable	model	RMSE	function
wildfire size	max temperature	exponential	103.458272	$y = 3983.999903(1.000574^x) - 4127.625702$
wildfire size	min temperature	exponential	103.454512	$y = 1504.149378(1.002003^x) - 1633.208192$
wildfire size	mean temperature	exponential	103.081503	$y = 82.881676(1.014682^x) - 150.948525$
wildfire size	max temperature	quadratic	103.458103	$y = -0.000887x^2 + 2.539651x - 153.822627$
wildfire size	min temperature	quadratic	103.454501	$y = 0.002948x^2 + 3.037211x - 129.947644$
wildfire size	mean temperature	quadratic	103.076922	$y = 0.024695x^2 - 0.035411x - 39.442500$

It is notably smaller than the RMSE values from the previous table, these lower values of RMSE indicate these models have better degree of fitting.These models' RMSE values are quite similar, both are around 103.From the data of wildfire size, we know the majority of the data of size between 0 and 200. However, from the above table, RMSE closes to 100 still meaning that the models are under-fitting. It shows that exponential and quadratic model failed to fit the temperature-fire size data. Thus, it is possible that temperature doesn't have much to do with fire size for they are not shown to be so related as what we expected.

Next we will move to analyze the relation between precipitation and wildfire(frequency and size). Below is the table showing the RMSE of models that are designed to fit the relation between precipitation and wildfire frequency/size.

Dependent Variable	Independent Variable	model	RMSE	function
wildfire frequency	precipitation	inverse	362.483733	$y = 113.647806/x + 275.585686$
wildfire frequency	precipitation	logarithm	333.143293	$y = -325.234402\log(x) + 536.221334$
wildfire size	precipitation	inverse	101.901471	$y = 9.204338/x + 12.909558$
wildfire size	precipitation	logarithm	103.136424	$y = -22.255857\log(x) + 35.478253$

Surprisingly, the RMSE values of models of wildfire frequency-precipitation are significantly larger than the RMSE values of models of wildfire size-precipitation. Thus, the models of precipitation-fire size using inverse/logarithm function are under-fitting. It shows that fire size is possibly unrelated to precipitation. By comparing the models of wildfire frequency-precipitation using logarithm function and the models of wildfire frequency-precipitation using inverse function, we see the model of wildfire frequency-precipitation using logarithm function has the smaller RMSE value hence it fits better. Thus, we will choose the logarithm model to predict future relation of wildfire frequency and precipitation though they can be possibly not so related.

Next, we will analyze the RMSE values of future predication models, including the predication for future temperature(max/min/mean) and the predication for future precipitation.

Dependent Variable	Independent Variable	model	RMSE
max temperature	time	periodic	3.872390
min temperature	time	periodic	2.932889
mean temperature	time	periodic	2.995619
precipitation	time	periodic	1.509227

The followings are the corresponding functions with respect to the rows of the above table.

$$y = 16.855509(cos(0.017191(x - 188.895404))) - 0.655883(cos(0.009173(x - 1730.125758))) + 0.000124x + 79.560000$$

$$y = 11.913980(cos(0.017191(x - 189.993939))) - 0.581465(cos(0.009041(x - 1736.122727))) + 0.000020x + 54.136364$$

$$y = 14.922533(cos(0.017174(x - 184.738788))) - 0.545275(cos(0.009235(x - 1798.091414))) + 0.000001x + 57.640874$$

$$y = 1.724477(cos(0.017209(x - 375.542063))) - 0.319513(cos(0.006854(x - 1929.415152))) - 0.000001x + 2.148300$$

Obviously, the RMSE value of model of max temperature prediction is greatest, the RMSE value of the model of min temperature prediction is the smallest, and the model of mean temperature prediction is at the middle but closes to the min-temperature model. Notice that their RMSE is way smaller than those previous models, we can conclude that using the model of min temperature prediction can give the most accurate result. Recall that we can predict fire frequency using max temperature data while we'd better use min temperature to predict future temperature trend for accuracy. We can't predict future max,min and mean temperature at the same time using these models.
Besides, the RMSE value of model of precipitation prediction is the smallest among all the models, it can be used to predict future frequency as well.

To sum up, the major findings of our study is that there is a small but not obvious relation between climate change represented by temperature and precipitation and wildfire represented by frequency and size, and we can expect a slowly increasing temperature and a ignorably and slowly decreasing precipitation in the future.
The results of our computational exploration helps answer our research question; climate change has small impact on the frequency and intensity of wildfires in California.

Our project have chosen various models and use RMSE to see how accurate these models are at predicting future data. But there are limitations of our finding: our models are not very accurate and we just picked temperature and precipitation, factors of climate, to predict future trend of climate change, which is incomplete prediction.
For further research, we may take more factors of climate change and wildfire into account to look for more deeply hidden relations between them. Also, we may implement models using multi-dimension models other than two-dimensional models that can give more accurate result of their relatedness.

References

Climate change: Wildfires burn more US forest. (2016, October). *Nature, 538*(7625), 292+. https://link.gale.com/apps/doc/A467 CPI?u=utoronto_main&sid=CPI&xid=57b7c2d9

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 9(3), 90–95.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

M.D Flannigan, B.J Stocks, B.M Wotton(2000, October). Climate change and forest fires. *Science of The Total Environment*, 262(3), 221-229. https://www-sciencedirect-com.myaccess.library.utoronto.ca/science/article/pii/S004896970000

National Oceanic and Atmospheric Administration. (2020, November). *National centers for environmental information.* https://www.ncdc.noaa.gov/cdo-web/search

Oliphant, T. E. (2006). A guide to NumPy (Vol. 1). Trelgol Publishing USA.

S.A. Parks, J.T. Abatzoglou(2020, October).Warmer and Drier Fire Seasons Contribute to Increases in Area Burned at High Severity in Western US Forests From 1985 to 2017. *Geophysical Research Letters*, 47(22).https://doi-org.myaccess.library.utoronto.ca/10.1029/2020GL089858.

sqlite3. (2020, November). *DB-API 2.0 interface for SQLite databases.* https://docs.python.org/3/library/sqlite3.html#module-sqlite3

Tatman, R. (2020, November). *1.88 million US wildfires: 24 years of geo-referenced wildfire records.* https://www.kaggle.com/rtatman/188-million-us-wildfires

The pandas development team(2020).pandas-dev/pandas: Pandas.https://doi.org/10.5281/zenodo.3509134

Virtanen, P., Gommers, R., Oliphant,Travis E., Haberland, M., Reddy, T., Cournapeau, D., … Contributors, SciPy 1. 0. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods.

os(2020, Dec) Miscellaneous operating system interfaces. https://docs.python.org/3/library/os.html