Weather Pattern Extraction using Statistical Methods & Machine Learning Techniques

by

Adria Binte Habib
18241027

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Adria Binte Habib
18241027

Approval

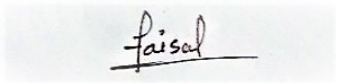
The thesis/project titled "Weather Pattern Extraction using Statistical Methods & Machine Learning Techniques" submitted by

Adria Binte Habib (18241027)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 2, 2021.

Examining Committee:

Supervisor:
(Member)



Faisal Bin Ashraf
Lecturer
Department of Computer Science Engineering
Brac University

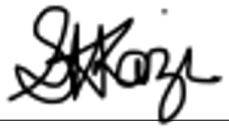
Co-Supervisor:
(Member)

Moin Mostakim
Senior Lecturer
Department of Computer Science Engineering
Brac University

Program Coordinator:
(Member)

Dr. Golam Rabiul Alam
Associate Professor
Department of Computer Science & Engineering
Brac University

Head of Department:
(Chair)



Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

It is known to all that the existence of life is highly dependent on the weather. Due to the unfavorable condition of current global weather the existence of life is in danger already. Since for the existence of lives a livable environment which lies in the weather is very much intrinsic, it should be taken care of before it is too late. This is the main context of this research. The goal of this research is to find out the future condition of temperature of some particular places of California using machine learning and statistical methods and compare which place will be more livable after two years. Currently, one of the most alarming issue in the world is the global warming. The effect of global warming is increasing rapidly every day without any sign of slowing down. As a result of this, it's very concerning and important to understand the state of the temperature of the world and the route it will take in the future. As such, the objective of this reseach is to predict the temperature conditions of the future. The research starts by collecting data of few select areas in california and hence, extracted data from 14 stations of california. The data was then fed to the ARIMA model to find the future trend with the respective ARIMA orders and other paremeters per station. The research has successfully identified the trend of the next 730 days (2 years) while considering the errors that the model creates. Furthermore, the research tried to identify the most favorable place to live, in california, by comparing the RMSE of the different stations by comparing the distance between the favorable human ambient temperature of 70∘F with the results that we got from the prediction. As such, the **'Miramar'** station gave the least RMSE value of 10.7824 while the **'lake Arrowhead'** gave the worst RMSE of 24.3605. From these RMSE values and also the learning curves it was decided, the most favorable place to live around was the **'Miramar'** station, while **'lake Arrowhead'** station was the worst in terms of favourable temperature for humans.

Keywords: Statistical Methods; Machine Learning; Weather; Prediction; ARIMA; Trend Analysis.

Dedication

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

Acknowledgement

Firstly, all the praise to the Almighty Allah for whom my under-graduation thesis have been completed without major interruption.
Secondly, to my advisor Mr. Faisal Bin Ashraf sir for his kind support and advice in our work. He helped us whenever we needed help.
Thirdly, to my co-advisor Mr. Moin Mostakim sir for his guidance.
Finally to my parents, without whose continuous support and prayers I am now on the verge of my graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

acf Auto Correlation Function

p Number of Autoregressive Terms

d Number of Non-seasonal Differences Needed for Stationarity

q Number of Lagged Forecast Error in the Prediction Equation

MSE Mean Square Error

$RMSE$ Root Mean Square Error

$uptrend$ Upward Tendency of a Trend

$downtrend$ Downward Tendency of a Trend

ϵ Epsilon

υ Upsilon

Chapter 1

Introduction

1.1 Thoughts behind the Prediction Model

Humans have always thrived for data and the meaning behind these data. Furthermore, finding patterns and being able to predict the future with this collected data is another fascination of us human beings. This is where the concept of predictive analytics comes. Predictive analytics extracts factors from the fed data and uses these factors to manage and predict our desired behavioral patterns [3] , [2] . AI, Data modeling, machine and deep learning and other applications of data mining can be considered statistical techniques of analyzing predictions [27]. This also can be termed "predictive analytics". AI, as we know, are systems that are capable of self-learning logics that are built based on the data that is fed to it. Adding to it, it can even improve certain aspects given the correct data and environment without the need of programming any logic particularly for it. Machine learning is primarily concerned with the development of software or algorithms that can use the data to learn logics and patterns. All we do is pre-process the data, format the data according to the algorithm we are using, and finally feed the data for the machine learning algorithm to do its job. The context of this research is to work with weather with support of predictive analysis.

1.2 Aims and Objectives

The aim is to prepare a model which will forecast future condition of temperature of some particular places of California using machine learning and statistical methods and compare which place will be more livable after two years. Since, currently global warming is an alarming issue and it is increasing rapidly day by day, it is really concerning where the increase of temperature will take the world. This is why, through this research it is trying to bring out the future condition of the temperature in some selected places.

1.3 Problem Statement

From the perspective of the global climate, temperature is one of the most basic and essential factor for the human beings. Also, it can have a significant effect on the entire ecosystem. In other words, if the the average air temperature crosses

the upper bound or the lower bound of the tolerance level of the living organisms (in terms of temperature), it can affect the health in a negative manner from feeling unwell to the worst case, death, or even extinction. On the other hand, if the changes in temperature occur very fast resulting in the living organisms unable to get the opportunity to get used to it, then it may entangle the overall environmental conditions [25].

Moving on, there's a increasing congregation of greenhouse gases in the atmosphere of Earth. As a result of this, the average temperature of the Earth overall is increasing. Furthermore, alarmingly, the researches expect it to be increasing non-stop in the future. Consequently, the change in this overall average temperature affects the climate itself. Like a domino-effect, it then causes the wind to change its patterns, and oceans to change its currents' pattern. These two factors also have a direct effect on the climate of Earth and hence, because of the change in patterns, the effect of temperature rise is not equal everywhere. At some places it will be more warm, while some places will be cooler than it was previously.
From the blog Wild Weather, National Geography [38],

"On the one hand, the most important influences on weather events are natural cycles in the climate. Two of the most famous weather cycles, El Niño and La Niña, originate in the Pacific Ocean and can affect weather patterns worldwide. But something else is happening too: The Earth is steadily getting warmer, with significantly more moisture in the atmosphere. The long-term accumulation of greenhouse gases in the atmosphere is trapping heat and warming up the land, oceans and atmosphere. As the oceans warm up, they produce more water vapor and this, in turn, feeds big storms, such as hurricanes and typhoons. And yet, there are ways of dealing with the effects of such extreme events. After 2003, French cities set up air-conditioned shelters for use in heat waves. In the 2006 heat wave, the death rate was two-thirds lower."
These extreme weather events are because of the dangerous, human-made changes that's damaging the Earth's climate. To get a better life in this situation, firstly, violation of weather should be stopped. In addition to that, alternative ways can be brought out.

Chapter 2

Related Work

In paper [9], the researchers collected a historical dataset to perform their research on weather forecasting. Due to the complexity of the patterns of atmosphere, non-linear and traditional methods are not effective nor is efficient to be used. As such, referring from the previous section, Deep learning seems the most fit to find and work with problems of such complexity. In this research, they use the following model.
Pre-processing of the data: After using various methods to take care of missing data, the experimenters decided to use the Spline Interpolation method to take care of the data that is missing since it proved to be the most effective out of all method.
Algorithm used: The model discussed in this paper compares between the difference in performance with respect to differing input styles. There exists a number of national and international sources having weather forecasting data. These sources has the collective data of many past years including parameters such as maximum and minimum temperature, wind direction, precipitation, wind speed and etc. However, working with all these parameters pose a major challenge due to the lack of proper and strong hardware. As a result, the experiment performed in this paper was limited to having just one single parameter for studying. This was done so by using ANN, one of the strongest algorithms to perform prediction-based researches. While traditionally ANNs are usually maintained at 10:1 ratio as the Input-Output Ratio, the method proposed here use two different ratios were used keeping the dataset same. The first ratio was 4:1 and the second test with the ratio 19:1. These were then compared against each other. There were approximately 1500 and 7000 elements respectively in a data of a single city.
The input data was divided into three parts with the ratio 8:1:1 for test, train and validation respectively. Although it seems like combining data from more cities should predict better patterns in forecasting weather, and over larger areas at that, the actual experiment and testing proved to give the opposite result. Combining the data resulted in a much higher Mean Square Error for the ANN model. This in turn proves that, there is no or less co-relation or dependency in-between the weather patterns of different areas or cities. Furthermore, the experiment proved that the cities with higher levels of pollution tends to have higher number of anomalies in their weather, although minor and shifts in their weather pattern. The model built in this research paper was a maiden approach to identify the status of Global Warming using non-conventional computing techniques.
In the paper [11], the researchers demonstrate how classifiers like Naïve Bayes and

Chi-Square algorithms can be used to predict the weather. The researchers built a user log-in based website application with a proper Graphical User Interface (GUI). Users here will input the required information (or, parameters) like, current outlook, humidity levels, temperature, wind conditions, etc. After this, the system analyzes the input parameters to predict the weather based on the growing dataset that it has in its database. The results of this application clearly show how data mining approaches can be good enough for weather forecasting.
This paper introduces a classifier approach for prediction of weather conditions and shows how Naive Bayes and Chi square algorithms can be utilized for classification purposes. This system is a web application with effective graphical User Interface. User will login to the system utilizing his user ID and password. Users will enter some information such as current outlook, temperature, humidity and wind condition. This system will take this parameter and predict weather after analyzing the input information with the information in the database. Consequently, two basic functions to be specific classification (training) and prediction (testing) will be performed. The outcomes demonstrated that these data mining procedures can be sufficient for weather forecasting.
The Chi Square test summarizes the results of their hypothesis and the they used data. This test also shows that the values obtained are substantially different from the values that were expected based on the attributes set on the training set. This model trains the probability of the study with the environment set where a chi-square statistic with more than 2 degrees of deviation is considered to be of a significant level. Figure 2(a), shows that all the values of all attributes are higher than the threshold of significant level. Next, the model classifies the data using the Naïve Bayes Algorithm. Figure 2(b) shows that the 'bad' has a better value than 'good' weather forecast overall.
Moving on to paper [6], To predict General Aviation (GA) accident rates from noisy Total Flight Hours (TFH) data, this research dives deep into figuring out the ability of the function (rate), a non-linear gamma-based function. The weighted "goodness" of fit (R2w) for non-Instrument-Rated (non-IR and Instrument-Rated (IR) pilots were calculated to be 0.654 and 0.75 respectively. This was done so by using two sets of datasets, recorded by the pilot's instrumental ratings, of National transportation Safety Board (NTSB) and Federal Aviation Administration (FAA). The model built in this paper would enable them to predict the GA accident rates directly. Furthermore, the model can also be used to find the probabilistic figures to identify flight risks in other types of models. Lastly, after applying the model on the dataset by FAA, it was seen that, the risk of accidents was much higher than anticipated previously.
In the Figure 3(a)(b), they show the curve-fits of the data. Histograms of rates of accidents, grate models are shown over the graph while estimates of parameter are shown below the graph. Figure 3(a) represents non-IR pilots while Figure 3(b) shows IR pilots.
In both the figures, it shows the accident rate on the y axis (probability of an accident occuring) and TFH at the time of accident on the x-axis (keeping the class-width at 100 units). Based on the data of the past 4.65 years, each class' raw number of accidents got divded by the total number of active pilots (accidents + non-accidents) keeping the Total Flight Hours constant at 100 hours per class. The results are then being divided by 4.65 to calculate the average per year before fitting the curves.

Relative weights are represented with green scaled to fit the y axis within the same figure constraints. Empty classes here are assigned with an weight of 0 to prevent them from affecting the fitting of the curve. Finally, the yellow curves around τ_{rate} represents the 95% of confidence interval (CI).
Next, in paper [18], the researchers studied the trend, based on Mann-Kendall and mutation trend analysis methods, of accidents and other damages in flights. They have then predicted the accidents and other casualties by using a big dataset. To perform this prediction, they built the ARIMA time-series analysis model.
Figure 4(a) gives the forward (UF) and backward (UB) trend curves of the number of accidents in flight. It can be seen from the curve of UF that, the number of accidents increased in a fluctuating manner long before 1950. Then, gradually started to reduce till 1980, while reducing rapidly after. Soon after, in 1990, it reached a level where it wasn't considered to be critical anymore. After this, the year 2003 started to see a drastic change again.
Figure 4(a) also shows the curves of UF and UB focusing on total 'number of accidents'. Starting with the UF curve, the accidents increased (fluctuating) till 1978. During this period, the curve shows that it crossed the critical value in 1961 while maintaining an increasing trend till 1978. However, in the 'fatality' phase, we see that it started to reduce since then and went below critical value in the year 2000. After this, a downward trend can be observed from the curve. Finally, the UF curve intersects with UB curve in the critical value boundary in 2013. This is when a drastic change was observed.
On the other hand, Figure 4(b) depicts the same concerns as 4(a) with the addition of Mann-Kendall test on the accidents. To start, in the 'en-route' phase, UF rose, fluctuating, till 1950 and then declining drastically afterwards. An intersection between UF and UB can be seen in the year 1980 within the boundaries of critical values. Furthermore, this point was lower than the minimum critical value threshold of 1984. After this, the curves took a downward trend and abrupt changes were seen in the year 1980. Next during the 'approach phase', gradually the UF curve kept on increasing, finally crossing the critical threshold in 1970. Finally, after lots of fluctuating episodes, the UF and UB curve again intersects during the year 2012 within the critical thresholds. Following the trend, drastic changes were observed again here.
In conclusion of the paper, the results of Mann-Kendall Analysis, accidents drastically reduced in all phases except in the landing phase. Furthermore, the fatalities reduced only in the 'en-route' phase. Although there were downward trends in fatalities in other phases, they were not as significant as en-route.
Finally, in paper [13], the researchers work with a model based on CNN (Convolutional Neural Networks). This model is trained with weather forecast data of the past. This model works by assigning a scalar confidence value to a new weather data. This value indicates the certainty of the prediction in a certain time of the year. Although this model is weaker compared to other models, its very efficient in terms of computing power and furthermore, performs better than other numerical-based model forecasts. On top of that, this trend proves that it's possible to use ML for predict uncertainty of forecasts based on past data. However, side by side, the limitation of this model is the amount of data available to train the algorithm for a solid enough result.
During the training, the researchers in this paper built the network by inputting the

atmosphere fields and output targets, also known as error or spread. While during the testing time, only inputs fields were provided where then the network predicted the output value. i.e. the uncertainty in the forecast that's being predicted. The whole experiment is done in a trial-error basis since there's no way to understand which network architecture best fits the given, new, scenario or data before performing the actual experiment. As such, a commonly used architecture was used. After choosing, the configurations of the model were varied to find the best results.
Baseline Method: To properly evaluate the network, they used Persistence/Local Dimension (This is a recent proposal to be used in weather forecasting, however, not yet used in real life) and Weather Type (Unlike Persistence/Local Dimensions, this is not operational and this is mainly for analyzing the patterns, mainly to identify the skillfulness of various weather forecast conditions) clustering that they discussed about in their literature review. Furthermore, for even more clarification, they have used Simple Nearest Neighbor approach to compare further. This is included to be used as a reference of Machine Learning methods used.
From the above discussion, we can see that paper 1 and 2 worked on the prediction of present weather forecast however according to this idea, Predict the weather pattern over the years and then predict weather for the next few years, I want to make future weather prediction. Secondly, in paper 3 and 4, they worked on air accident rate, flight risk etc. but for the idea, Predict Air accident rate, I need to find a relationship between air accidents and weather. Thirdly, in paper 5 they worked on weather uncertainty, which may help for finding a better solution for idea 1,3 and 4.

In paper [21], the authors studied the weather pattern of some areas of Bangladesh depending on 60 years of past data using data mining techniques. They analysed the trend for the given areas and found out the areas with similar trend depending on the weather attributes. They researched on monthly weather data of Maximum Temperature, Minimum Temperature and Relative Humidity for the years from 1953 to 2013 and found out the homogeneous climate zone in Bangladesh.

Chapter 3

Time-series forecasting using Statistical Methods

For any time-series based forecasting, there are several statistical methods that can be performed. They are discussed below.

- **AR:** Also known as Autoregression, is a method that builds a linear-function model based on the upcoming in the time-series sequence based on the observation of the past time steps in the data [22].

- **MA:** Moving Average method, moves to the following step in the sequence by using a linear function same as AR. However, here, this model considers the residual errors (mean) calculated using the prior time steps [22].

- **ARMA:** This method, Autoregressive Moving Average method, combines AR and MA to build its model [22].

- **ARIMA:** Autoregressive Integrated Moving Average (ARIMA) method is ARMA with the addition of pre-processing (by differencing) the steps of the sequence to make the sequence stationary (also called Integration (I)) [22].

- **Theta:** This method builds the model using simple exponential smoothing with drift [16].

- **ETS:** This method is typically used for forecasting, introduced by Gardner Jr, 1985. This is also an exponential smoothing like Theta. However, it uses a state-space model [16].

- **Tbats:** This is ETS along with Box-Cox Transformation. Furthermore, it also uses errors, seasonal components and trends found from ARMA [16].

3.1 Reasons for Choosing statistical method

The preprocessed dataset is univariate (i.e. one column). Furthermore, methods such as ARIMA, ETS perform better than machine learning and deep learning methods in one step forecasting. On top of it, ARIMA performs better than deep learning for multi-step forecasting at a time [15].
Furthermore, few of the key models to forecast weather conditions are primarily based on statistics and statistical reasoning. Furthermore, these predictions can have ranges from predicting in terms of hours to numerous seasons. To start dynamic forecasting on weather and climate conditions, these methods are one of the best suited and potential predictors. Also, the dynamic predictions can be further post-processed. i.e. remove unnecessary bias and add weights to the factors for better accuracy. The reason behind doing such kind of forecast is to eliminate the aid of other dynamic models which are less accurate. Lastly, the statistical method is even more accurate because it's impossible to add manipulated results. Rather, they are all based on pure statistical outputs [1].

3.2 Selected Method

The goal of this research is to predict the future condition of each station to identify the most livable place in the future. For predicting the future, the ARIMA algorithm has been selected for this research.
The ARIMA algorithm is one of the highest used and fairly simple techniques. The algorithm best works when the data of the past has a proper and stable inter-related trend [7]. Consequently, because of these characteristics of the algorithm, it fits quite well with weather prediction related research. Furthermore, the ARMA and ARIMA models work very well with log functions. Due to this, it is also considered to be one of the best statistical models for many other predictions. Subsequently, when speaking in terms of this research, there has been much research conducted by using the models ARMA and ARIMA to predict weather and do related work in the field of weather [12], [17], [8].

Chapter 4

Methodology

4.1 Time-series forecasting using ARIMA

Data which are observed over time is called time-series. ARIMA models can almost accurately predict the future based on a number of historical data of a single variable. This model is quite different from the others because it does not assume anything from the given historical data. From the Box-Jenkins methodology for building up ARIMA model, the following steps can be extracted: (1) Model Identification, (2) Parameter Estimation and Selection, (3) Diagnostic Checking (or Modal Validation) and (4) Model's use [5].
The model can be identified by choosing the right values for the parameters p, d and q. Collectively, these three variables are called "order" of the ARIMA model. The "order" of the algorithm helps us to fit the predicted values with our validation set to increase the accuracy of our model. However, unlike in machine and deep learning methods of achieving a higher accuracy, we here work with the RMSE value. Lower the RMSE value, the better fit our model is. Moving on, let's discuss about the three parameters in the following points:
Parameter 'p': The variable 'p' indicates the lag of the values in the y-axis. In simple terms, it's just the value that needs to be added or subtracted with the Y value of the model. Models predicted are often lagged and hence, we use the value of 'p' to reduce that lag. In more technical terms, it helps us to get a better prediction by considering the narrow periods of increase or decreasing curves of our data. Hence, going along with the "autoregressive" characteristics of the algorithm, ARIMA.
If p = 1; The data is linearly increasing or decreasing
If p = 2; The data is exponential increasing or decreasing ... and so on
Parameter 'd': This parameter controls the degree of "differencing" the data. Meaning, how many times the data selected needs to be "differenced" (to be discussed below in this section). This is a requirement for the model to give a output which has a constant mean over a given period of time. As a result, it helps us identify the "integrated" characteristics of the algorithm.
If d = 0; The data is already stationary. Furthemore, it also means the algorithm is using ARMA rather than ARIMA because there's no "integrated" nature when d = 0.
Parameter 'q': Similarly to the 'p' parameter, this also works with the Y values. Except, rather than working on raw data, it tries to make up for the error that's

made during prediction. Finally, this parameter holds true for the "moving average" characteristics of ARIMA.
All in all, the above three parameters together form the AR-I-MA of the whole algorithm [23].

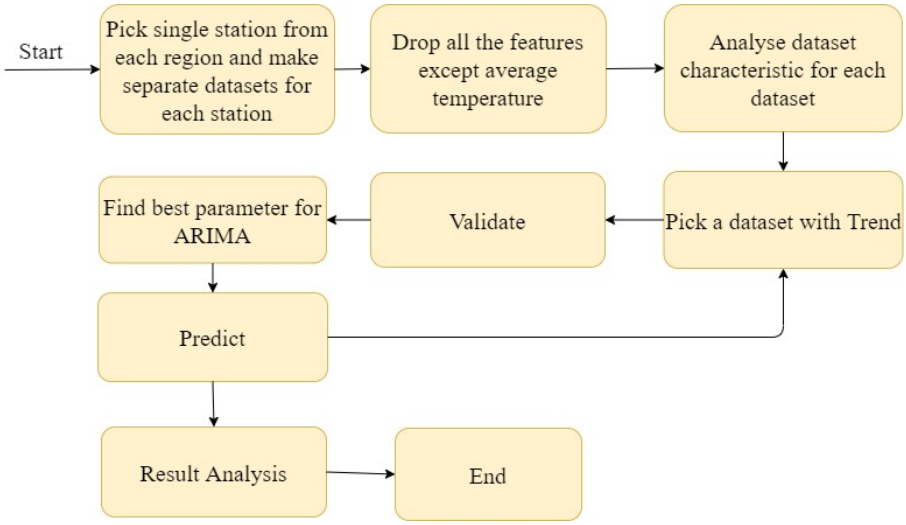


Figure 4.1: Proposed Model of Analysis

Chapter 5

Data Collection & Analysis

5.1 Data Collection

The dataset for this research has been collected from the online community named Kaggle. The name of the dataset is California Environmental Conditions Dataset [24]. It is a Weather Dataset of California sourced from CIMIS weather stations (approximately 262 stations). The stations used a selenium chrome driver to collect the data. Along with the 262 Stations, having their own unique IDs, there's 14 numerical features to perform the research on and devise various outputs and to test various ML/DL algorithms on. Lastly, this data has date and time stamps in case the research requires time series analysis. Finally, the target column states about the fires that occurred on a respective date, in a given region which was being observed.
It contains 19 columns for tracking the information. The feature names are given below.
1. Stn Id
2. Stn Name
3. CIMIS Region
4. Date
5. ETo(in)
5. Precip(in)
6. Sol Rad (Ly/day)
7. Avg Vapour Pressure(mBars)
8. Max Air Temparature(F)
9. Min Air Temparature(F)
10.Avg Air Temparature(F)
11.Max Relative Humidity (%)
12.Min Relative Humidity (%)
13.Avg Relative Humidity (%)
14.Dew Point (F)
15.Avg Wind Speed (mph)
16.Wind Run (miles)
17.Avg Soil Temperature(F)
18.Target
From a study,it was found that daily data for forecasting provides the closest prediction [29]. In this dataset, historical weather data are collected on a daily basis

which is a great indicator for getting a close enough prediction. This is why, this dataset has been selected.
For approaching any type of statistical algorithm, we need the perfectly processed numerical data. However, time series analysis follows a different type of data pre-processing format. From the statement 'Time Series Analysis' we can get the idea that this analysis is related to time which may include hours, days, months etc.

5.2 Data Pre-processing

In this research, I am going to predict the future weather conditions of California based on temperature. From the previous section, we know that data volume is huge. To increase the prediction accuracy the outliers have to be removed. Let's see it step by step.

- There are 15 CIMIS Regions and 253 stations. Since one region named 'North-east Plateau' is far away from the other regions (shown in the Figure 5.1), this region has been excluded. From these stations 14 specific stations need to be picked which will cover all the regions. Firstly, all the stations were separated into 14 sections based on their CIMIS Region. After that, from each section of stations, the station with the highest number of data-points was selected for the further research.

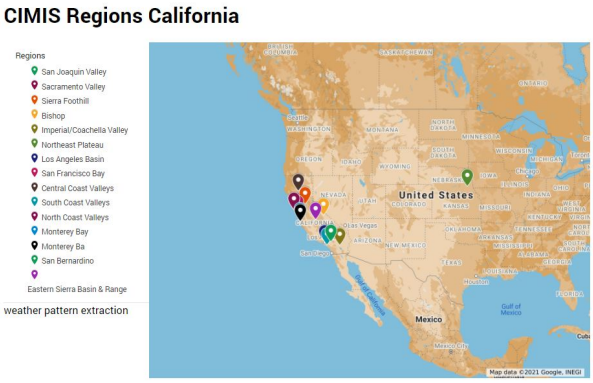


Figure 5.1: Location of the CIMIS Regions

- Next, the raw data was then splitted according to the stations (i.e. the 14 stations that we selected in the previous step based on the number of the data-points in each station). This step is necessary to predict the future per station as each station is in a different region. And each region may have a different pattern compared to others.

- Since, in this research only the temperature parameter is going to be predicted, resulting in the other feature columns in the dataset being irrelevant.

Figure 5.2: Location of the Stations along with the CIMIS Regions

As such, they were dropped and as the final input data, only the columns for date and average temperature were taken into account. On a side note, the data had minimum and maximum temperature readings as well. However, the average temperature gave a more even meaning for the prediction that is to be conducted in this research.

5.3 Dataset Characteristic Analysis

For selecting methods for future prediction, the type of the time series should be revealed properly. Here comes the concept of stationarity and the technique of differencing time series.

5.3.1 Importance of Stationarity

A time-series data, which does not depend on the given time, is identified as a stationary time-series. To add to it, time-series data having trends and/or seasonality cannot be called or identified to be stationary time-series data [33]. In the stationary time-series, the consecutive (hourly/daily/monthly) data do not depend on each other [23]. This is why the data pattern follows a discontinuous manner. However, for weather forecasting, a dataset pattern must be continuous which means there should be dependency between the consecutive data. This is the reason we need non-stationary data for weather forecasting.

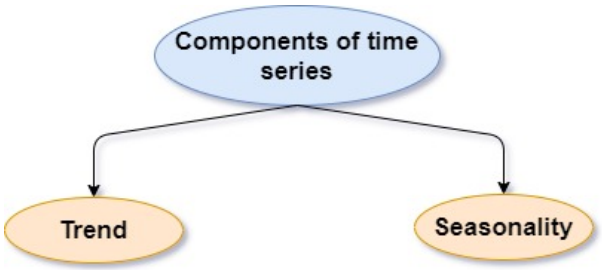


Figure 5.3: Components of Time Series

5.3.2 Trend

A trend is whenever it can be observed that there's a pattern in the data with respect to time. In other words, we can see an increase or decrease in the value of the data repeating over a period of time. Furthermore, the trend can be divided into subsections in terms of time and then we can still identify trends as long as there's any sort of pattern. i.e. It might be increasing in one period, while decreasing in another. However, it can still be called a trend depending on a trigger of some sort [26].
Trend can be linear or non-linear. In the case of the chosen dataset, it is non-linear.

5.3.3 Seasonality

Seasonality can be considered like a subset of the overall trend of a data. While trend observes data over a long period of time (having multiple repetitive periods), seasonality describes just one of that period and identifies the repeating points. By the by, when considering the weather, the seasonality can be observed over the period of 365 days or 1 year. Furthermore, to observe this seasonality in terms of weather or for any other type of data, the data must be recorded in a short period of time and regularly ranging from as minimum as starting from seconds [26].
Though in this research, I will discuss only one component which is trend. Now, diving into the analysis, firstly it was checked that the data is stationary or not.

5.3.4 Stationarity Analysis

The stationarity of the data was identified by the ACF. It was found that for a non-stationary time-series the ACF of the data should decrease slowly but never drop to zero [14]. Using this concept, the ACF plots of the 14 stations were judged and defined whether the data were stationary or not.
For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly [33]. So, for finding out the stationarity ACF plotting has been used on the 14 stations accordingly.
From the above figure, for each station, it is seen that the ACF plot decreased slowly and did not touch zero which means the data of each station is non-stationary. As a result, the corresponding data have relation with time and contain trend and seasonality.

5.3.5 Trend Analysis

A time-series data can have trends. Moreover, it's prominent that weather follows a trend on a yearly basis. This trend can be analysed by plotting graphs. For this research, we can plot the trends of the individual stations by reading the plots of the trends extracted from their respective raw data.. In the figure below the trend of 14 stations have been extracted and displayed.

By closely looking at the plots on Figure 5.5, we can see that the trend can be compared to sin or cos curve like a wave. Moreover, as stated before, this curve shows the trend of the temperature of their respective area.

To explain the graph, it can be said that we can observe local maxima at every 400 days approximately. Furthermore, seeing the repetitive trend, it can be said that the weather condition is roughly the same every 400 days. Although it looks like the curve of one of the stations, Markleeville on Figure 5.5(i), looks different, by taking a closer look, it can be seen that it has the same curve of having a period of 400 days. However, due to the lack of data, unlike the other stations, it seems different. But, it means the same thing. Since of having a lack of data ARIMA can not be applied on this particular dataset. ARIMA works better when a dataset has more than one cycle of seasonality. While for a data with having a single season, the SARIMA variate is used instead.

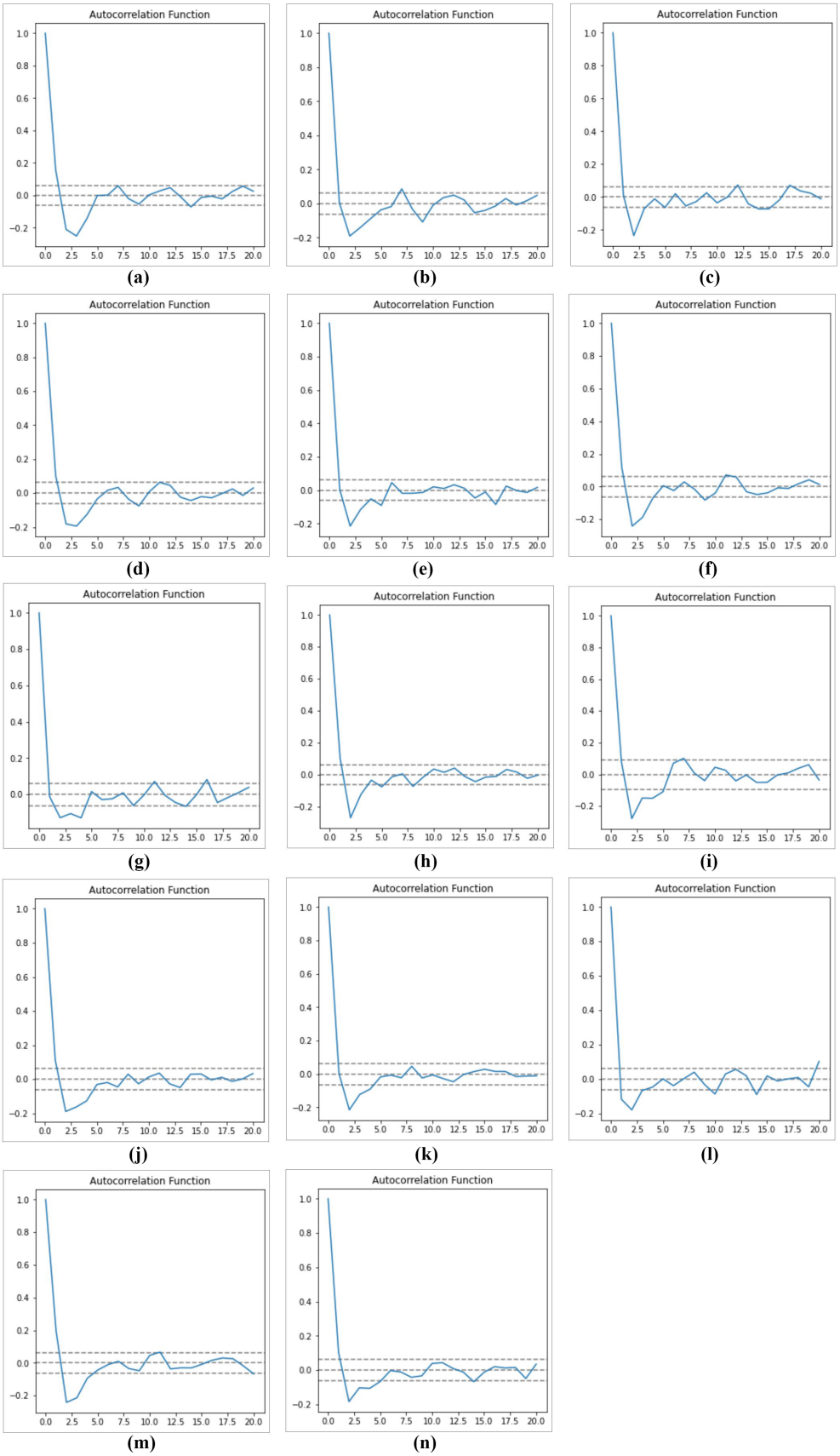


Figure 5.4: Plotting of the ACF for all stations

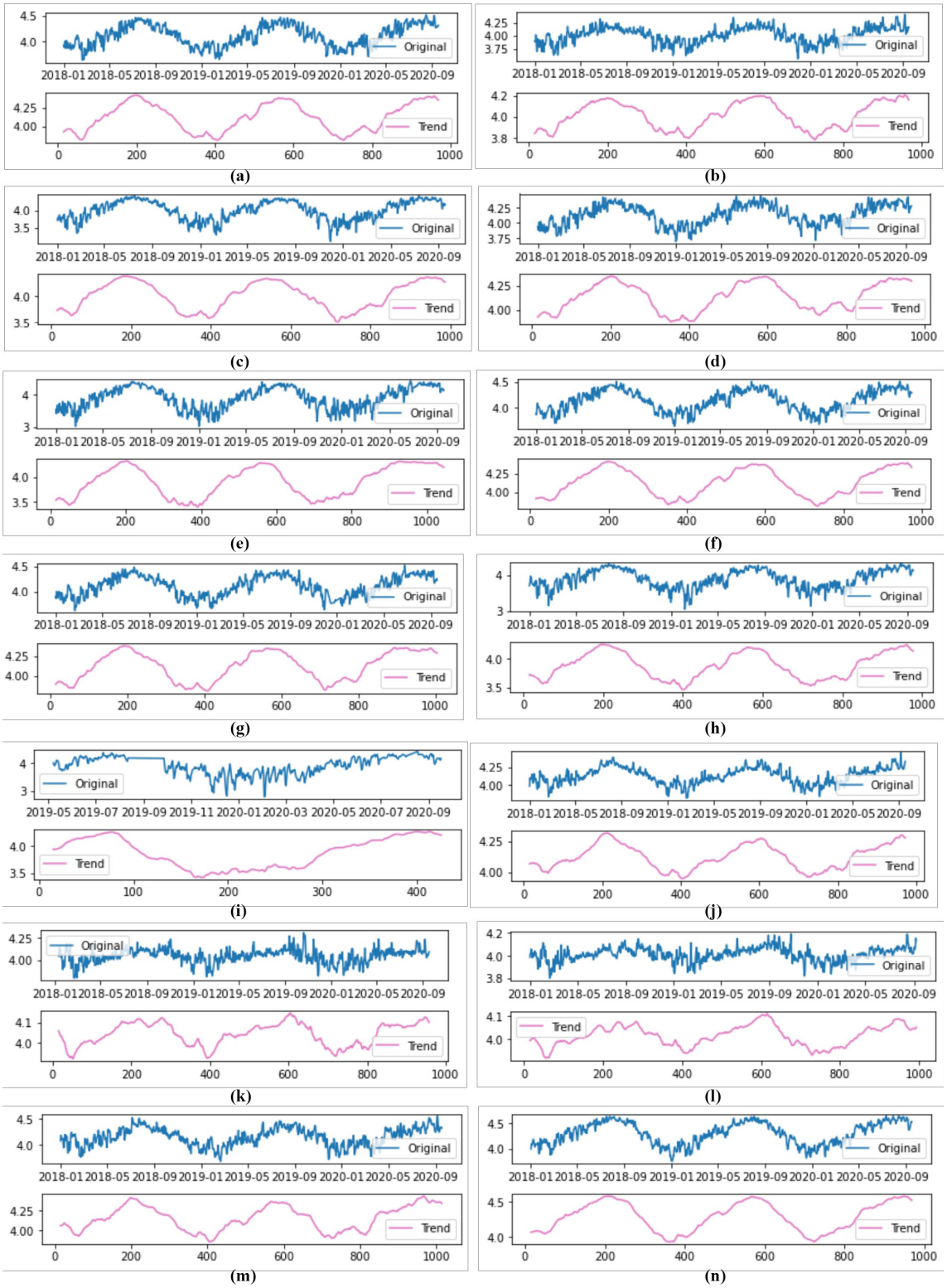


Figure 5.5: Trend Analysis for all stations

Chapter 6

Implementation

The entire implementation and experimentation were done in Python 3.8.5 environment on the operating system, Ubuntu 20.04. The hardware configuration is Processor: AMD Ryzen 5 3600 (not overclocked), RAM: 32GB 3200Mhz (2133Mhz operational), GPU: NVIDIA RTX 2060 Super 8GB, Storage: Samsung EVO Plus 500GB SSD, Cooling System: EKWB EK AIO 360.I have used python libraries: scikit-learn [35], pandas [34], statsmodels [31], numpy [28], datetime [30] and math [32] with different parameters for the calculation process. I have also used matplotlib [37] for the plotting purpose.Moreover, for forecasting the statistical algorithm ARIMA has been used [36].
For initializing the dataset, pandas library has been used. Moreover, to parse the 'Date' column the datetime library has been used. After that, for the purpose of dataset characteristic analysis, the stationarity of the dataset was analysed using the libraries numpy, matplotlib. Moreover, from the statsmodels library, the ARIMA model has been imported to predict the future temperature condition. Finally, for calculating the RMSE the scikit-learn library has been used. Firstly, from the scikit-learn library, the MSE function has been used to find out the mean squared error. After that, using sqrt function from math library, the RMSE was calculated. The RMSE has been used here for error calculation because the root mean square error is highly recommended in the temperature prediction [19], [14] , [10]. Since ARIMA is a time series forecasting algorithm, the working process of it is, iterating through each date and predicting the feature for the corresponding day. That is how the algorithm goes forward to the destination date. For achieving the process, the 'Date' column must refer to index so that the algorithm can iterate through it. This is why, the 'Date' column was changed into index while the data was being read.

6.1 Validation

Lets take a single dataset , for finding out the best fitted model, firstly the dataset was splitted into train set (75%) and test set (25%). After that the test set was predicted and validated. The best suited parameters of ARIMA were selected by using 3 nested loops, each ranging between 0 to 15, for p,d and q respectively. The best suited parameters were then selected based on the lowest RMSE score. The lower the RMSE score, the more accurate the model is [20]. The best suited parameters are given below.

Station Names	Best Fitted Model	RMSE
Auburn	ARIMA(2, 0, 9)	11.6579
Bennett Valley	ARIMA(4, 1, 0)	7.3409
Bishop	ARIMA(9, 2, 2)	8.9873
Brentwood	ARIMA(2, 1, 1)	7.3061
Buntingville	ARIMA(11, 2, 10)	11.5589
Five Points	ARIMA(1, 1, 7)	6.8057
Gerber South	ARIMA(9, 2, 2)	7.5512
Lake Arrowhead	ARIMA(2, 1, 2)	7.9786
Miramar	ARIMA(4, 1, 6)	5.9150
Nipomo	ARIMA(0, 1, 0)	11.5973
Pacific Grove	ARIMA(1, 1, 11)	3.5959
Santa Clarita	ARIMA(2, 1, 0)	9.0019
Seeley	ARIMA(2, 0, 13)	10.1027

Table 6.1: Best Parameter for the Model

6.2 Prediction

After finding out the best the best order values for the ARIMA model, the research moved on to the next step, which is, the prediction. Initially, it was decided to predict the temperature of up to 2 years. Side by side, from the table (Table 6.1) and figure (Figure 6.1) it can be seen that the lowest RMSE was seen in the 'Pacific Grove' station. This means that the prediction for this station should be the most accurate out of all the other stations. Nonetheless, the temperature of all the 13 stations were predicted and is presented in the next chapter.

Chapter 7

Result Analysis & Discussion

Finally, after the implementation of the research (discussed in the last section), the results were extracted in two parts (validation and prediction).

7.1 Validation & Prediction

To start, the code for the validation set is run first to identify the error of the model. For that, as shown in figure 1, the data was splitted in to 3:1 ratio and then validated with the last 4th of the whole data. The prediction and validation curves are plotted for each station (from Figure A7 to Figure A19) (legend given in figure). By observering, it can be seen that, the two curves barely have any overlapping values. Furthermore, at points the distance between the two curves is huge, while at some points it's minimal, while at some places it does intersect, and finally, in some places the differences are negative or positive. Although it seems that the prediction is all over the place, if we take a closer look at the curves, the overall raw data-points show a similar trend of up and down curves. Next, traditionally most of the researches use the "accuracy" factor to compare two models and to identify the credibility of the model. However, it's not possible here in the time series data because, as explained above, the data point values are all over the place. However, as, again, mentioned above, the curves do show a similar trend. As such, a better indicator of accuracy in our case is finding the RMSE value of the two curves. The best RMSE values (lower is better) [20] is presented in table below.

From the table, we see that the RMSE value of each station varies. This is because, the number of data in each dataset of the individual data varies both in terms of number and date of collection of data. For example, Station Markleeville contained 16 months of data whereas other 13 stations contained 32 months of data. As a result, For Markleeville trend was not observed but seasonality was found (shown in figure 7.1). For this particular station, seasonality can be found because most of the time seasonality repeats after almost 12 months but trend takes longer than that [26]. By following this hypothesis and also by running the code for ARIMA, it was decided that Markleeville has no trend. However, as stated above, all of the curves (from Figure A7(a) to Figure A19(a)) show that the predictions and test-set data point values follow a similar trend if not same values. This proves that the ARIMA model has correctly captured the trend and was able to produce results for the future.

Station Names	Best Fitted Model	RMSE
Auburn	ARIMA(2, 0, 9)	11.6579
Bennett Valley	ARIMA(4, 1, 0)	7.3409
Bishop	ARIMA(9, 2, 2)	8.9873
Brentwood	ARIMA(2, 1, 1)	7.3061
Buntingville	ARIMA(11, 2, 10)	11.5589
Five Points	ARIMA(1, 1, 7)	6.8057
Gerber South	ARIMA(9, 2, 2)	7.5512
Lake Arrowhead	ARIMA(2, 1, 2)	7.9786
Miramar	ARIMA(4, 1, 6)	5.9150
Nipomo	ARIMA(0, 1, 0)	11.5973
Pacific Grove	ARIMA(1, 1, 11)	3.5959
Santa Clarita	ARIMA(2, 1, 0)	9.0019
Seeley	ARIMA(2, 0, 13)	10.1027

Table 7.1: Best fitted ARIMA models used in the research

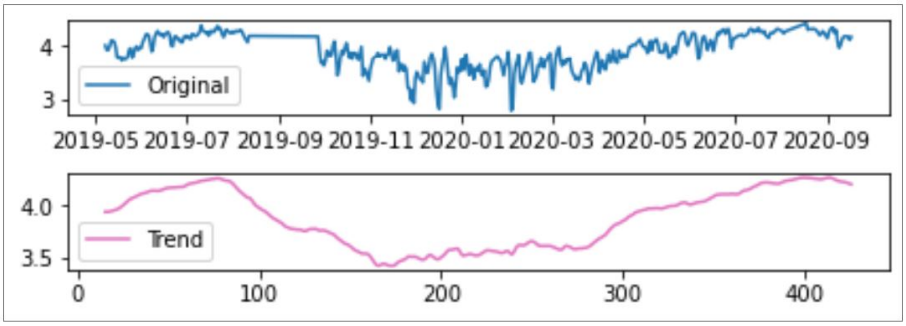


Figure 7.1: Trend and Seasonality Analysis of Markleville

Next, coming on to the prediction of the future for this research. In contrast to the validation code run, there was no validating data this time. Instead, the whole data of the station was fed as for the training and then a pure future was predicted for the next 730 days (2 years). Furthermore, since the ARIMA has full data to train itself on, we can safely assume that the the RMSE value is lower than the values obtained during the validation phase. However, there's no way to check for this since it was a pure future prediction of values. Except, we can use our observations. Looking closely at the figure (Figure A7(b)-A19(b)), it can be observed that for the period of the next 700 days (from the last date present in the dataset), it follows a similar trend to our data by following the shape of a sin curve. It follows an uptrend till around the 180 200 day (which is the month of January).After that,it follows a downtrend till around 350 380 day.
Now let's talk about the predicted temperatures of each of the station one by one. Starting with 'Auburn', in the below Table (Table 7.2), it can be seen the prediction of the temperature has been divided into 5 zones to identify the uptrend and downtrend of the predicted temperature. In September 2020, it can be seen the temperature was around 45° F to 50 ° F. In March 2021, the predicted temperature increased upto 80° F.

Moreover, in the August-September 2021, the temperature decreased around 36° F to 40° F which means the weather should be very cold at that time. Again, in April-May 2022, it is following the almost same trend alike March 2021 which is almost same timing of previous year. Moreover, in September 2022, the temperature followed the same trend like September 2021.
The Station 'Bennett Valley' is similar to 'Auburn'.
Moreover, looking at all stations (Table B(1) to B(12)), it can be be seen in the months of September the temperature stays always low. On the other hand, at the month of March, April, May the temperature stays high.
However, the temperature of the station 'Lake Arrowhead' looks extremely cold all over the year.

Stations	RMSE
Auburn	14.3726
Bennett Valley	15.5334
Bishop	21.3465
Brentwood	11.3136
Buntingville	14.9911
Five Points	14.5300
Gerber South	13.2127
Lake Arrowhead	24.3605
Miramar	10.7824
Nipomo	12.6582
Pacific Grove	14.6430
Santa Clarita	12.2020
Seeley	16.1631

Table 7.3: RMSE comparison with respect to 70° F

7.2 Favourable Living Condition

From an interview [4] of Professor Jeffery W. Walker, University of Arizona, we know that the most favorable temperature for a human being is 70'F. As such, The RMSE values were extracted by comparing the predicted values with the 70'F. The results are shown in the table below. A higher RMSE value means the temperature is overall further away from the favorable conditions. While a lower RMSE means the conditions are closer to 70'F overall.
Next, from the figures A(20) to A(32), it can be seen that the green curve is fluctuating depending on the RMSE. Moreover, taking a closer look at the figures, it can be observed, when the RMSE decreases, the Green Curve goes closer to the middle.From the above observation, it can be said, **'the curve of the stations containing the green curve close to middle along with the lower RMSE'** are mostly preferable areas to live. Adding with it, if the most preferable areas for living are pointed out then it can be seen, the **'Brentwood'**, **'Miramar'**, **'Santa Clarita'** and **'Nipomo'** are the most preferable for living for next 2 years. However, the 'Auburn', 'Bennett Valley', 'Buntingville', 'Five Points', 'Gerber South', 'Pacific Grove' and 'Seeley' holds average place.

Auburn		
Time Period	**Days**	**Temperature**
September 2020	1	45.42197714
	2	51.99700537
	3	41.95875865
	4	40.10853606
	5	41.66387221
March 2021	180	75.04288264
	181	78.86269727
	182	77.08595264
	183	76.32887382
	184	78.92018343
	185	80.06818585
August - September 2021	350	36.63836906
	351	38.1377862
	352	34.32413565
	353	34.41746722
	354	36.72719504
	355	40.63882936
April - May 2022	570	64.02827874
	571	64.82881218
	572	70.82918262
	573	74.2288607
	574	77.52833018
September 2022	726	36.9286701
	727	39.72873128
	728	37.92875952
	729	36.02871505
	730	37.02866405

Table 7.2: Predicted Temperature of Auburn

Chapter 8

Conclusion

In summary, it can be said, by using statistical methods and machine learning techniques, it is very much possible to predict the future weather conditions. However, the accuracy of the prediction models can be enhanced by changing the value of moving average and also the other parameters. Throughout the paper, it was seen that the temperature condition of each station has a different trend. That is why the future condition of each station was predicted differently. As it is seen, the predicted temperature of some stations are much livable and others are not. From this decision, the better livable place for the next two years can be assumed. Nonetheless, like any research, there are limitations to this research as well. Starting with, the dataset collected was not even. Firstly, the stations didn't have the same number of rows. Next, there were missing dates which might have directly affected the trend of the data. Moving on, one of the most lagged behind problems with ARIMA is that it works with a single column data. As a result, unlike machine learning, it only depends on one parameter rather than predicting while finding relations within the factors present in the data.

Bibliography

[1] F. W. Zwiers and H. Von Storch, "On the role of statistics in climate research," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 24, no. 6, pp. 665–680, 2004.

[2] W. W. Eckerson, "Predictive analytics," *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report*, vol. 1, pp. 1–36, 2007.

[3] C. Nyce, "Predictive analytics white paper, sl: American institute for chartered property casualty underwriters," *Insurance Institute of America*, 2007.

[4] J. W. Walker, *When air is the same temperature as our body, why do we feel hot?* Apr. 2009. [Online]. Available: https://www.scientificamerican.com/article/why-people-feel-hot/.

[5] M. Kumar and M. Anand, "An application of time series arima forecasting model for predicting sugarcane production in india," *Studies in Business and Economics*, vol. 9, no. 1, pp. 81–94, 2014.

[6] W. R. Knecht, "Predicting accident rates from general aviation pilot total flight hours," *Federal Aviation Administration*, 2015.

[7] Z. A. Farhath, B. Arputhamary, and L. Arockiam, "A survey on arima forecasting using time series model," *Int. J. Comput. Sci. Mobile Comput*, vol. 5, pp. 104–109, 2016.

[8] G. Jain and B. Mallick, "A review on weather forecasting techniques," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 12, pp. 177–180, 2016.

[9] H. Tyagi, S. Suran, and V. Pattanaik, "Weather-temperature pattern prediction and anomaly identification using artificial neural network," *International Journal of Computer Applications*, vol. 975, p. 8887, 2016.

[10] A. Mirakyan, M. Meyer-Renschhausen, and A. Koch, "Composite forecasting approach, application for next-day electricity price forecasting," *Energy Economics*, vol. 66, pp. 228–237, 2017.

[11] B. Munmun, D. Tanni, and B. Sayantanu, "Weather forecast prediction: An integrated approach for analyzing and measuring weather data," *Int J Computer Appl*, 2018.

[12] M. Murat, I. Malinowska, M. Gos, and J. Krzyszczak, "Forecasting daily meteorological time series using arima and regression models," *International agrophysics*, vol. 32, no. 2, 2018.

[13] S. Scher and G. Messori, "Predicting weather forecast uncertainty with machine learning," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 717, pp. 2830–2841, 2018.

[14] J. Zhao and X. Liu, "A hybrid method of dynamic cooling and heating load forecasting for office buildings based on artificial intelligence and regression analysis," *Energy and Buildings*, vol. 174, pp. 293–308, 2018.

[15] J. Brownlee, *Comparing classical and machine learning algorithms for time series forecasting*, Aug. 2019. [Online]. Available: https://machinelearningmastery.com/findings-comparing-classical-and-machine-learning-methods-for-time-series-forecasting/.

[16] V. Cerqueira, L. Torgo, and C. Soares, "Machine learning vs statistical methods for time series forecasting: Size matters," *arXiv preprint arXiv:1909.13316*, 2019.

[17] A. Kocharekar, B. V. Nemade, C. G. Patil, D. D. Sapkale, and S. G. Salunke, "Weather prediction for tourism application using arima," *Weather*, vol. 6, no. 11, 2019.

[18] Y. Li, "Analysis and forecast of global civil aviation accidents for the period 1942-2016," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[19] E. Lucas Segarra, H. Du, G. Ramos Ruiz, and C. Fernández Bandera, "Methodology for the quantification of the impact of weather forecasts in predictive simulation models," *Energies*, vol. 12, no. 7, p. 1309, 2019.

[20] H. Pham, "A new criterion for model selection," *Mathematics*, vol. 7, no. 12, p. 1215, 2019.

[21] F. B. Ashraf, M. R. Kabir, M. S. R. Shafi, and J. I. M. Rifat, "Finding homogeneous climate zones in bangladesh from statistical analysis of climate data using machine learning technique," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2020, pp. 1–6.

[22] J. Brownlee, *11 classical time series forecasting methods in python (cheat sheet)*, Dec. 2020. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/.

[23] J. Kohlrausch and E. A. Brin, "Arima supplemented security metrics for quality assurance and situational awareness," *Digital Threats: Research and Practice*, vol. 1, no. 1, pp. 1–21, 2020.

[24] C. Zaloumis, *California environmental conditions dataset*, Oct. 2020. [Online]. Available: https://www.kaggle.com/chelseazaloumis/cimis-dataset-with-fire-target.

[25] *Climate change indicators: U.s. and global temperature*, Apr. 2021. [Online]. Available: https://www.epa.gov/climate-indicators/climate-change-indicators-us-and-global-temperature.

[26] T. S. A. Toppr, *Components of time series*, 2021. [Online]. Available: https://www.toppr.com/guides/business-mathematics-and-statistics/time-series-analysis/components-of-time-series/.

[27] w. wikipedia wikipedia, *Predictive analytics*, May 2021. [Online]. Available: https://en.wikipedia.org/wiki/Predictive_analytics.

[28] [Online]. Available: https://numpy.org/doc/.

[29] D. Abugaber. [Online]. Available: https://ademos.people.uic.edu/Chapter23.html.

[30] *Datetime - basic date and time types¶*. [Online]. Available: https://docs.python.org/3/library/datetime.html.

[31] *Introduction¶*. [Online]. Available: https://www.statsmodels.org/stable/index.html.

[32] *Math - mathematical functions¶*. [Online]. Available: https://docs.python.org/3/library/math.html.

[33] oTexts, *Forecasting: Principlesnbsp;andnbsp;practice (2nd edition)*. [Online]. Available: https://otexts.com/fpp2/stationarity.html.

[34] *Pandas.read$_c$sv*. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html.

[35] sklearn, *Sklearn.metrics.mean$_s$quared$_e$rror*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html.

[36] *Statsmodels.tsa.arima$_m$odel.arima*. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima_model.ARIMA.html.

[37] *Visualization with python¶*. [Online]. Available: https://matplotlib.org/.

[38] *Wild weather*. [Online]. Available: https://www.ngllife.com/wild-weather-1.

Appendix A

Appendix A : Figures

Appendix A : Figures

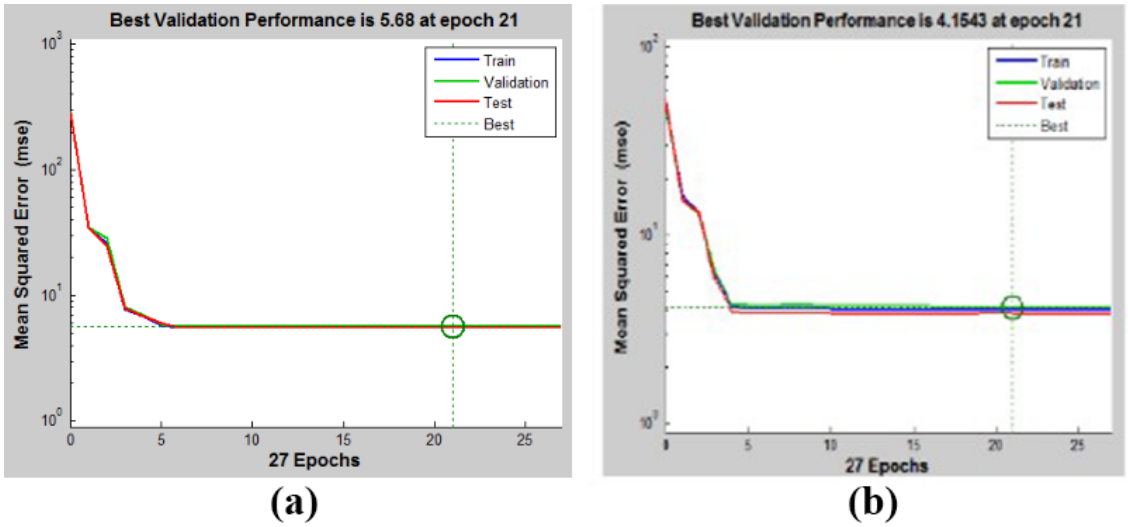


Figure 1: Evaluation of Training, Validation & Testing Performance for a Given Input-Output Data Ratio DELHI over last 20 years. (a) Left – 4:1 Ratio; (b) Right – 19:1 Ratio

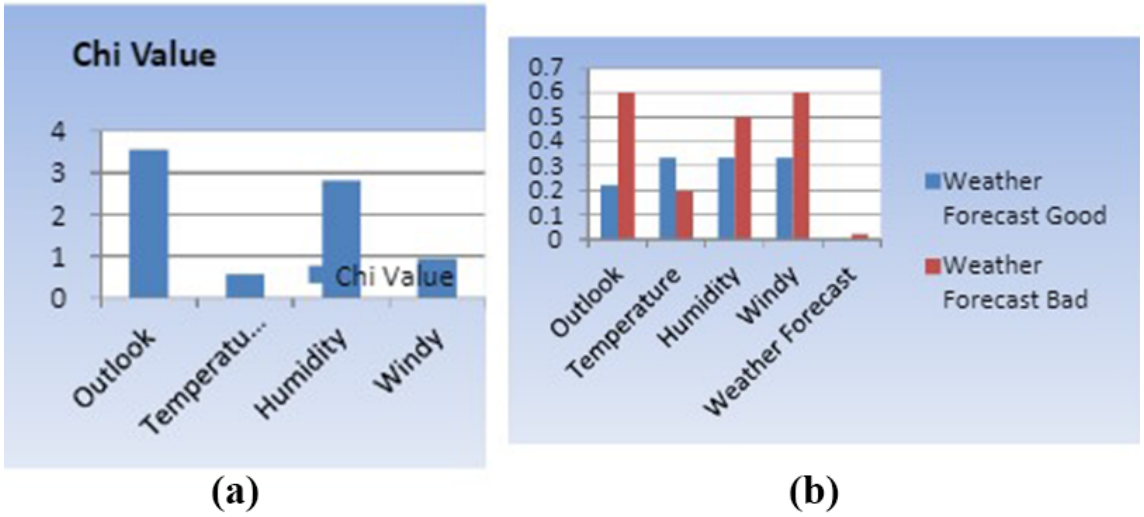


Figure 2: (a) Left - Chi square and; (b) Right - Naive Bayes Model test results

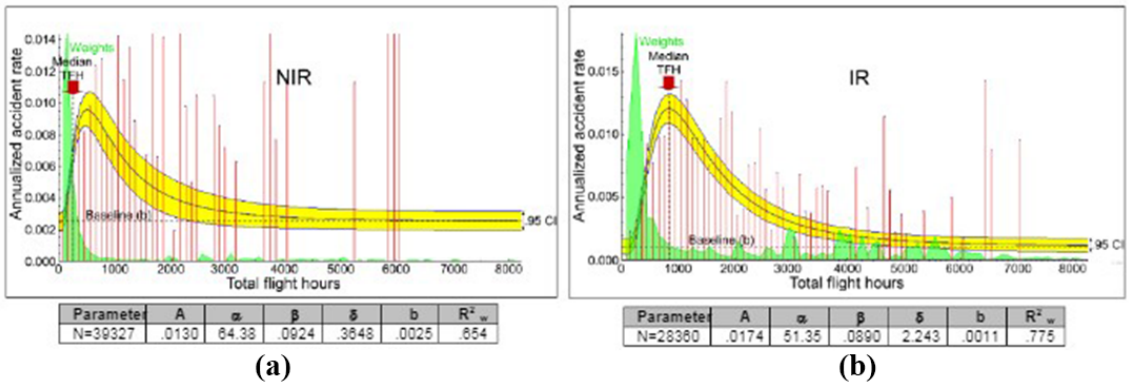


Figure 3: (a) Left - Non-IR accident rates; (b) Right - IR accident rates

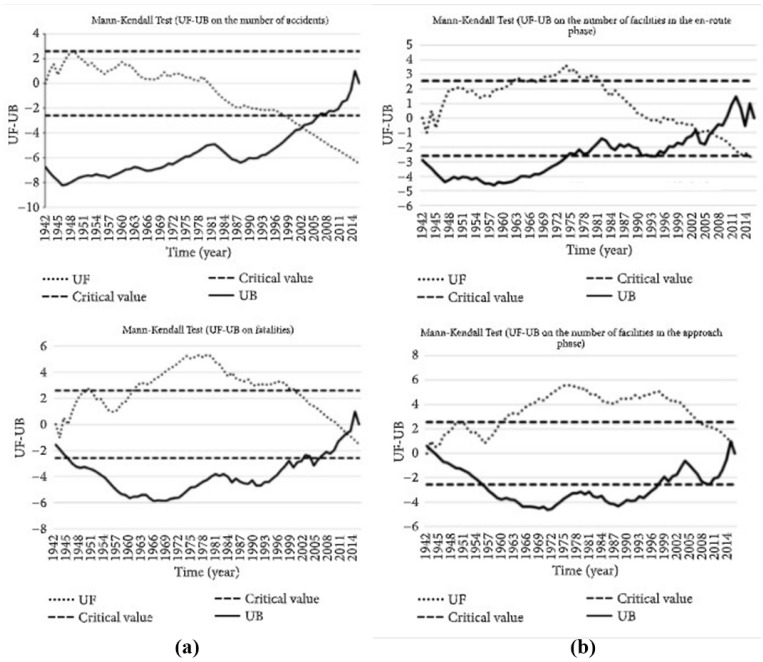


Figure 4: (a) Left – Mann-Kendall test of global civil aviation accident and casualties; (b) Right – Mann-Kendall test of number of global civil aviation accidents at different stages

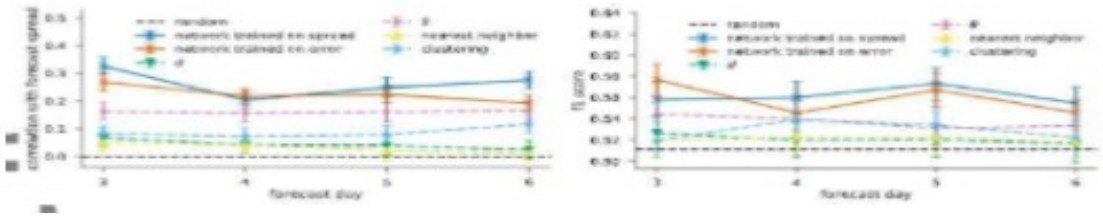


Figure 5: Comparison of 4 baseline methods

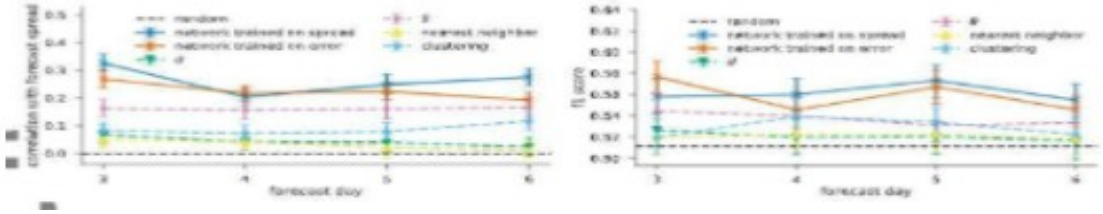


Figure 6: Comparison of 4 baseline methods

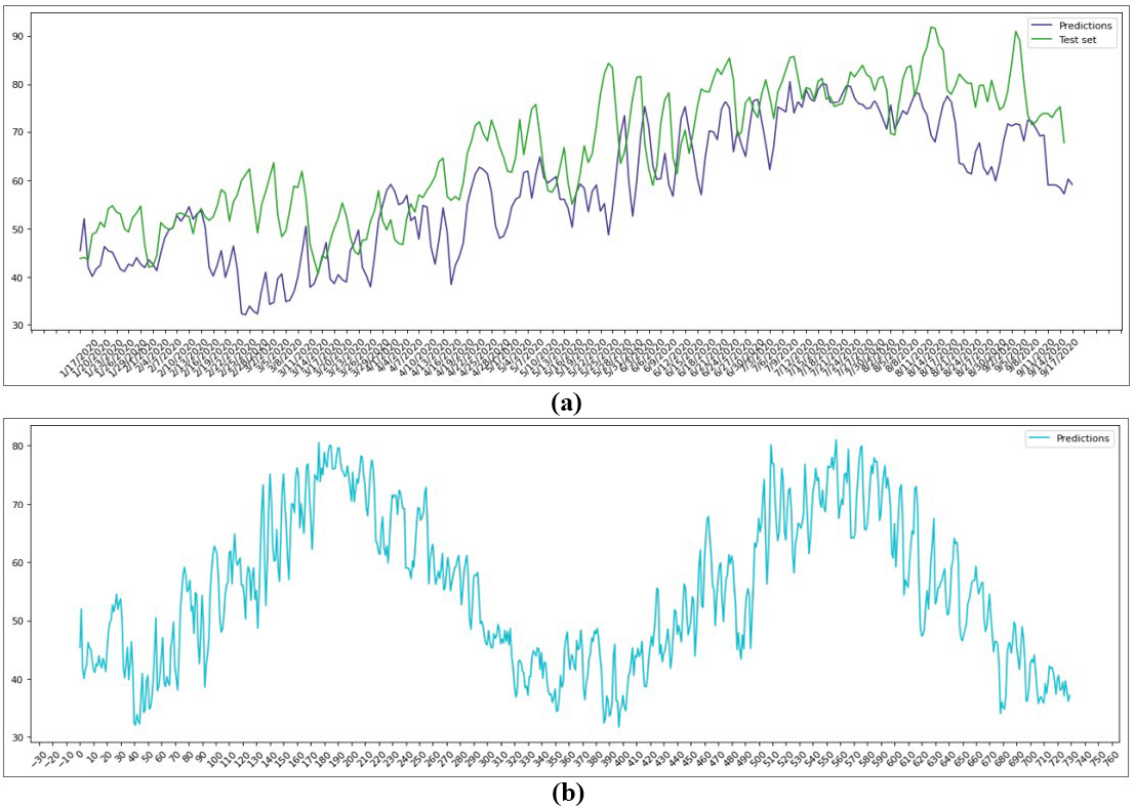


Figure 7: Validation and Prediction Learning Curve of Auburn

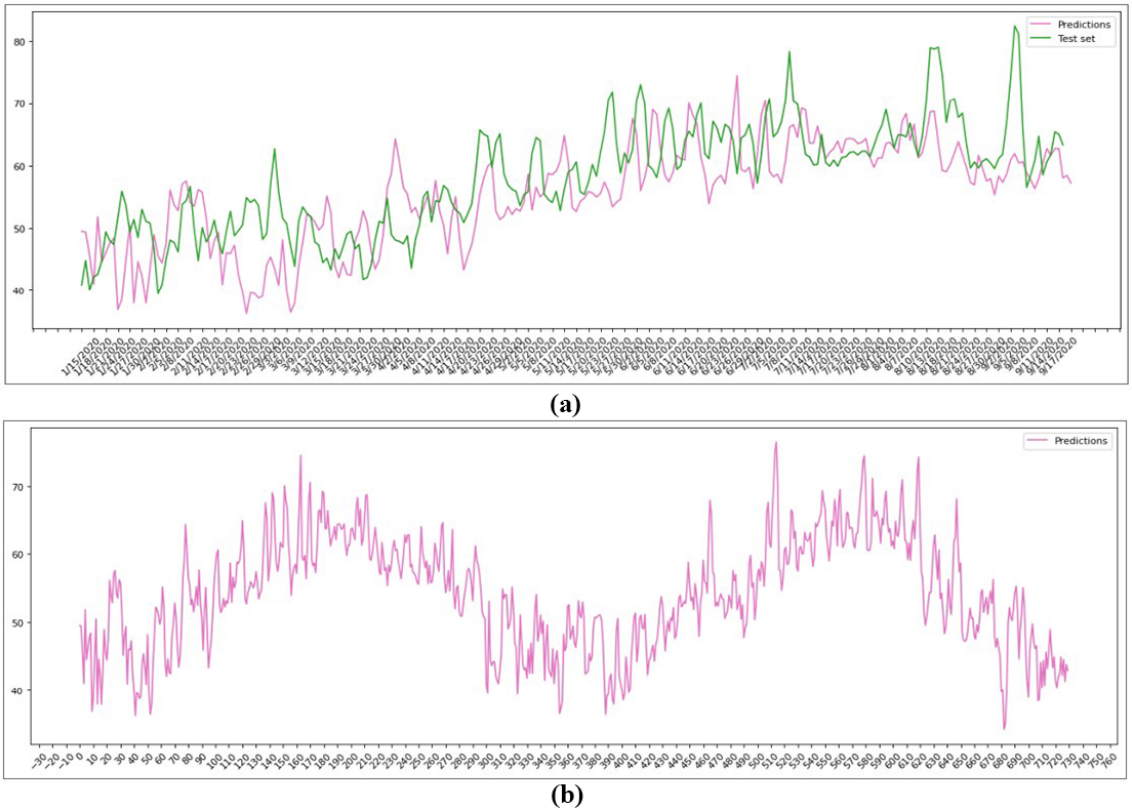
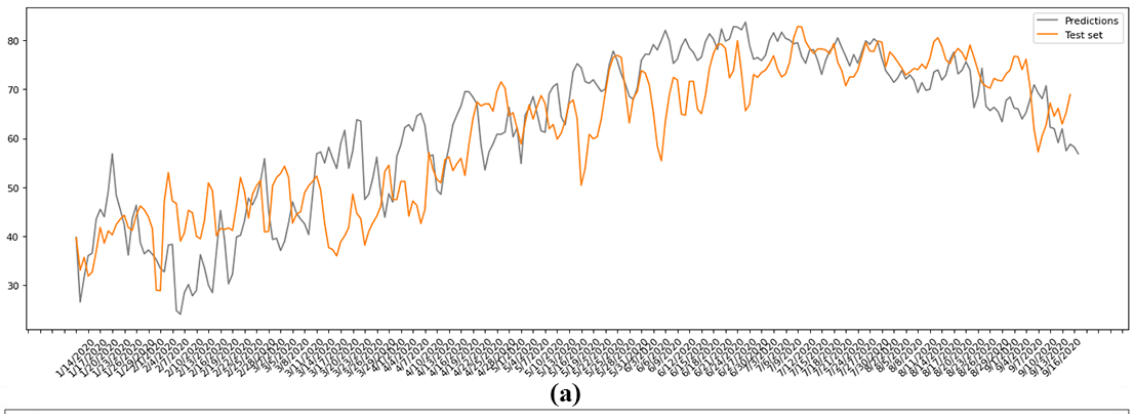
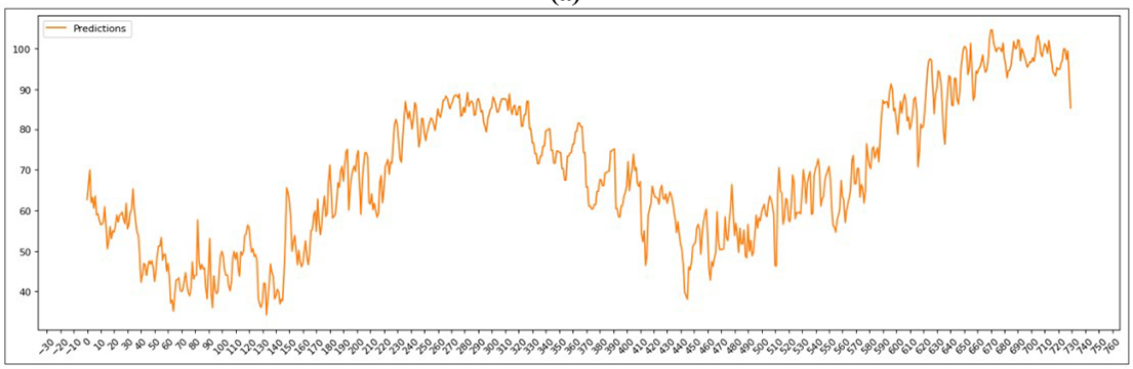


Figure 8: Validation and Prediction Learning Curve of Bennett Valley

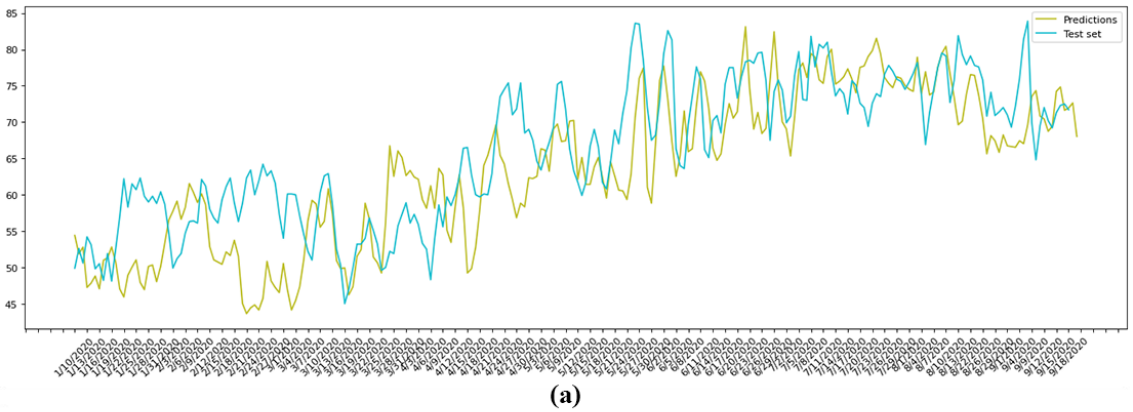


(a)

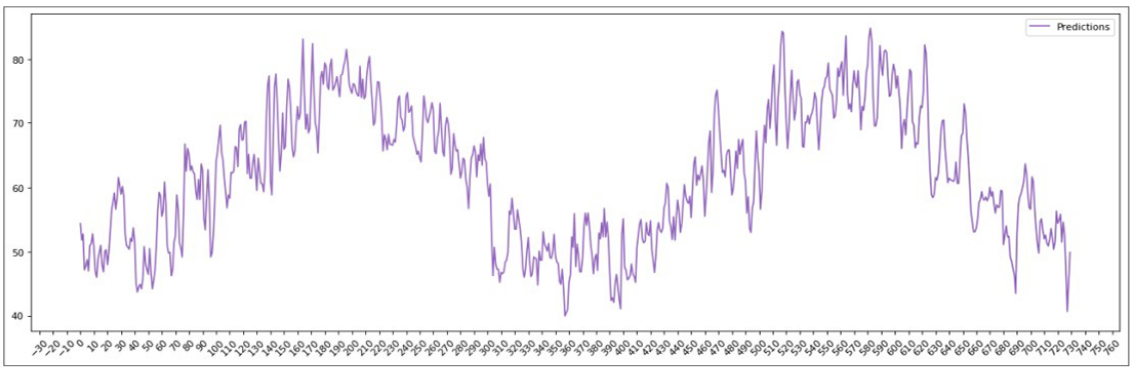


(b)

Figure 9: Validation and Prediction Learning Curve of Bishop



(a)



(b)

Figure 10: Validation and Prediction Learning Curve of Brentwood

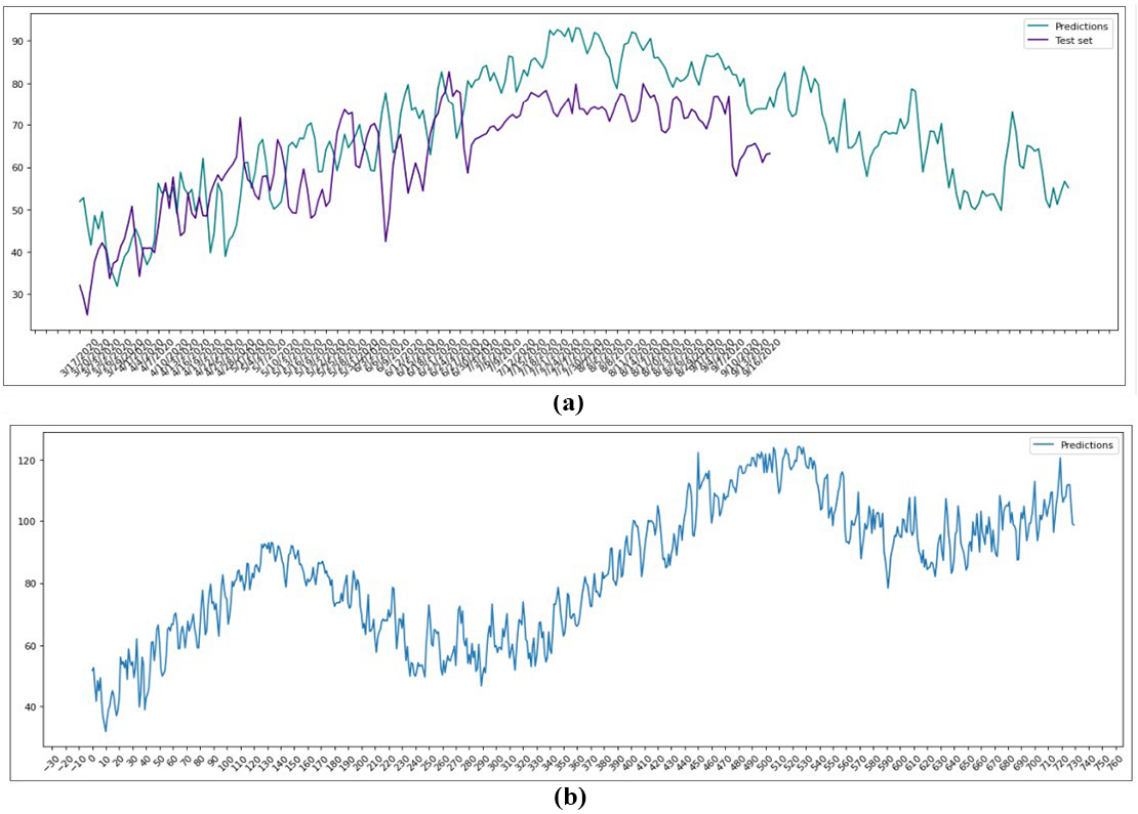


Figure 11: Validation and Prediction Learning Curve of Buntingville

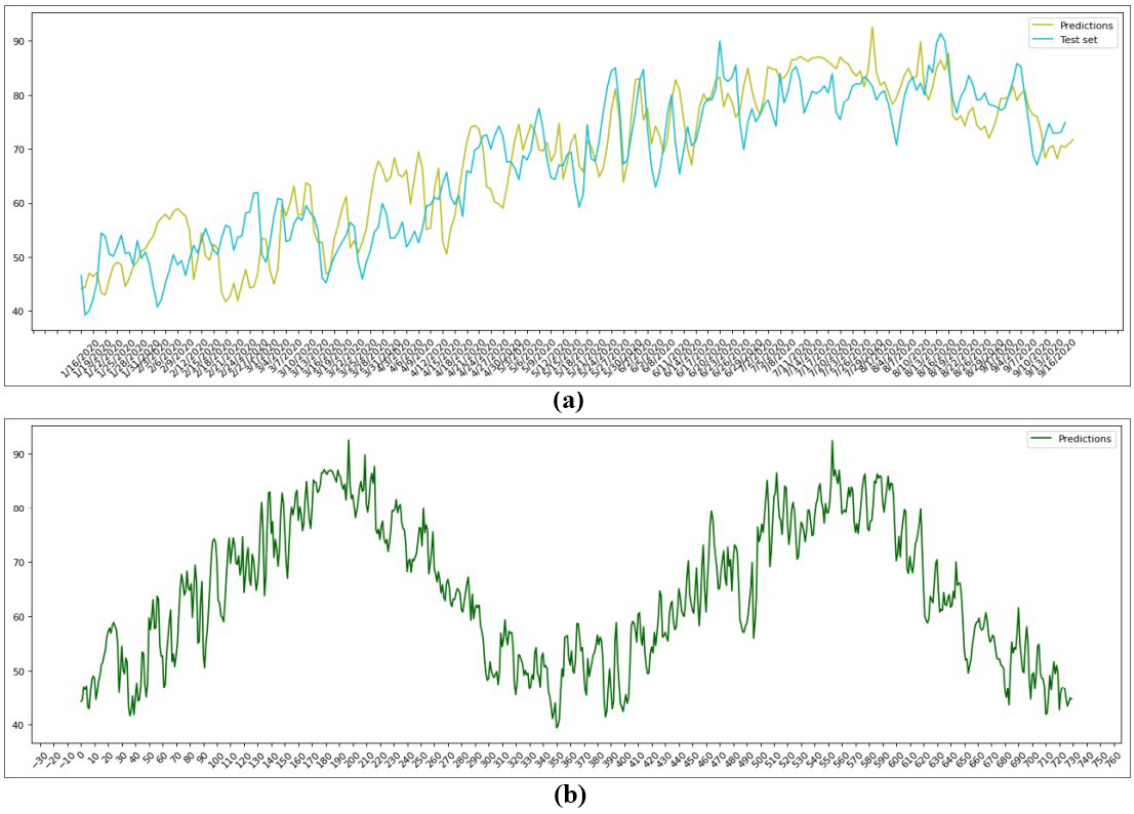
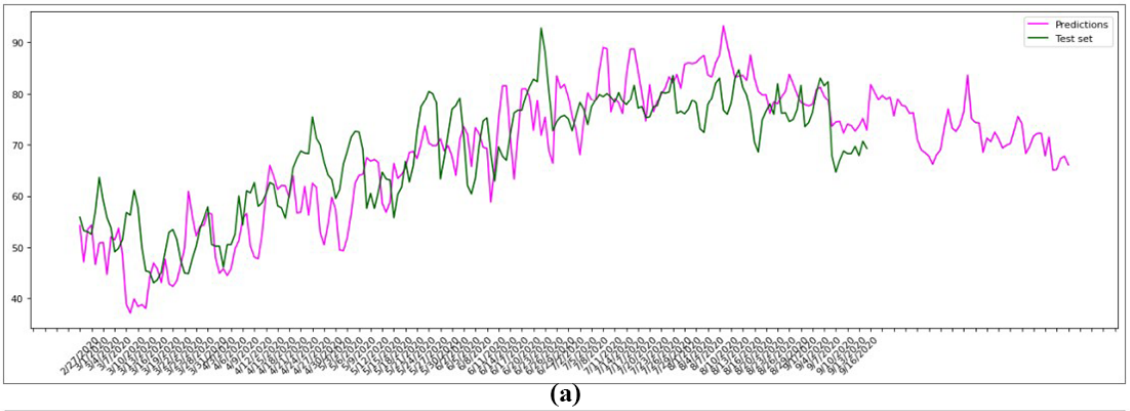
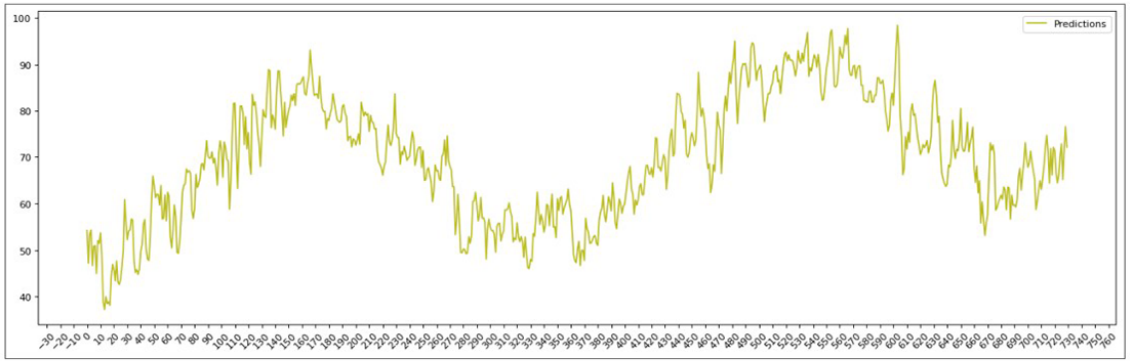


Figure 12: Validation and Prediction Learning Curve of Five Points

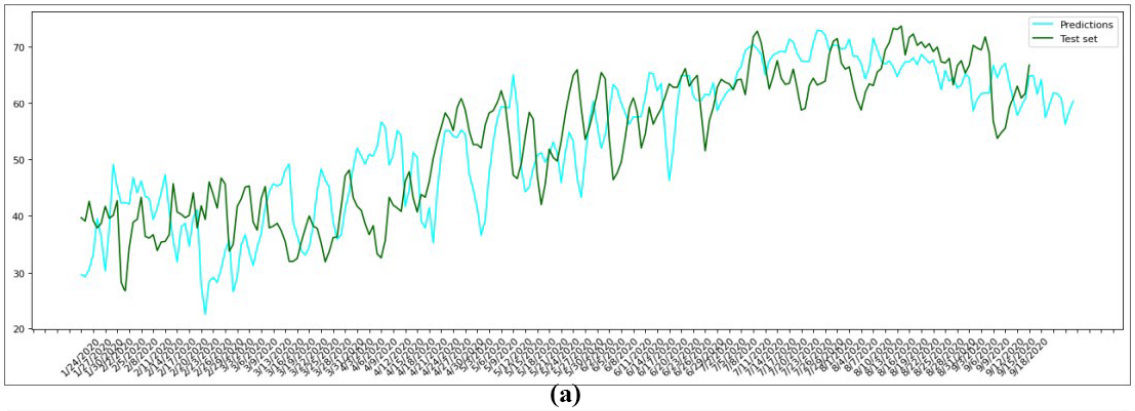


(a)

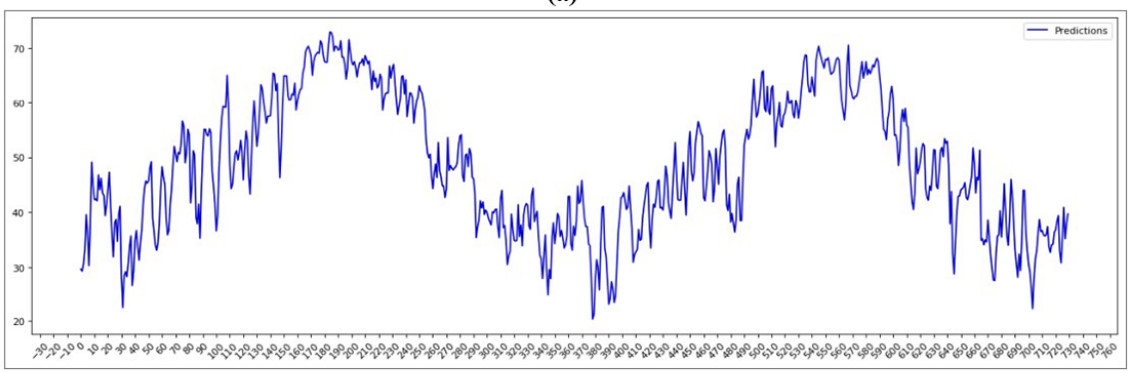


(b)

Figure 13: Validation and Prediction Learning Curve of Gerber South

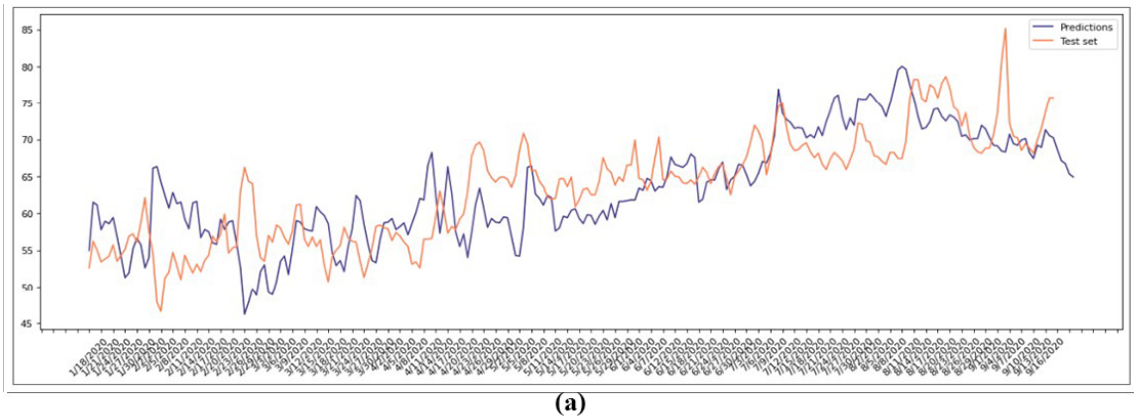


(a)

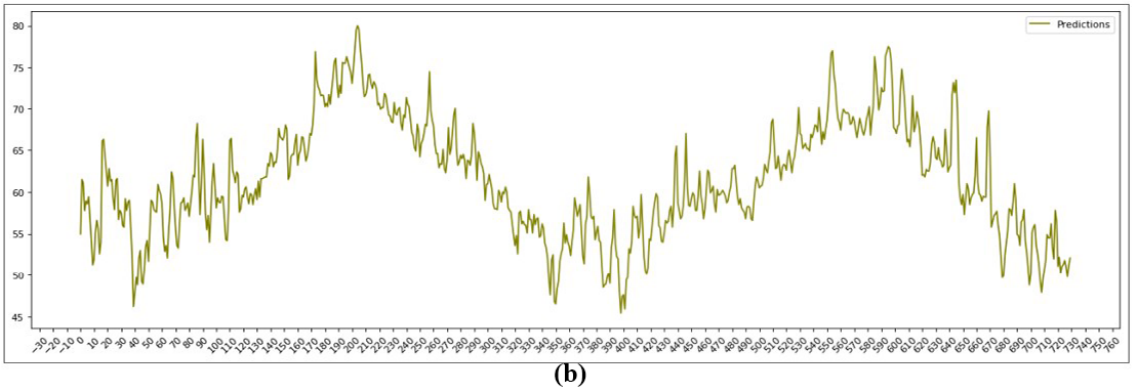


(b)

Figure 14: Validation and Prediction Learning Curve of Lake Arrowhead

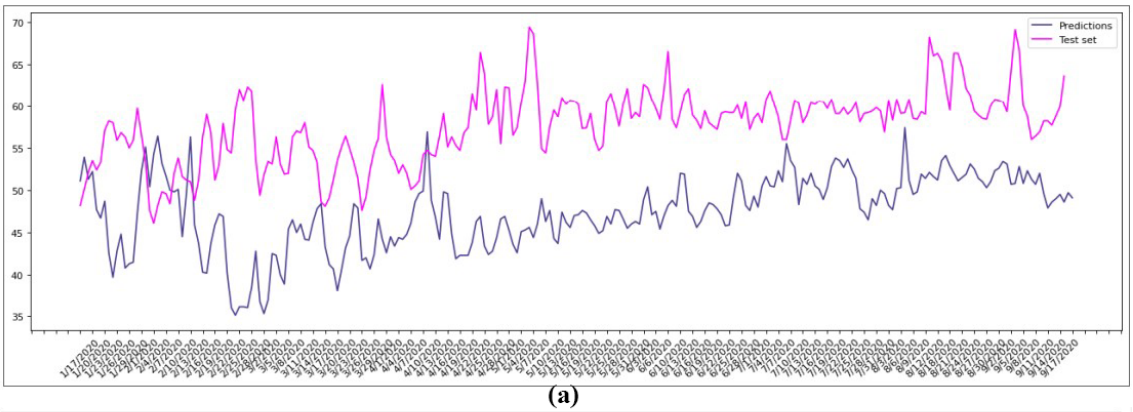


(a)

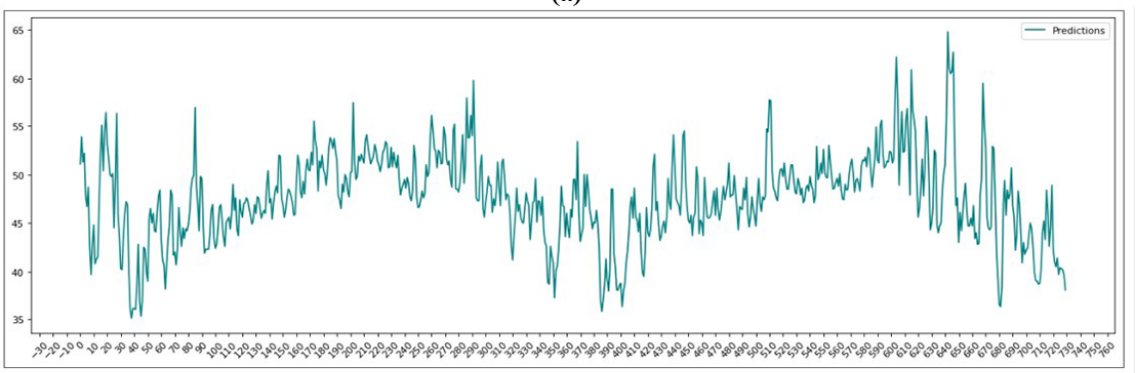


(b)

Figure 15: Validation and Prediction Learning Curve of Miramar

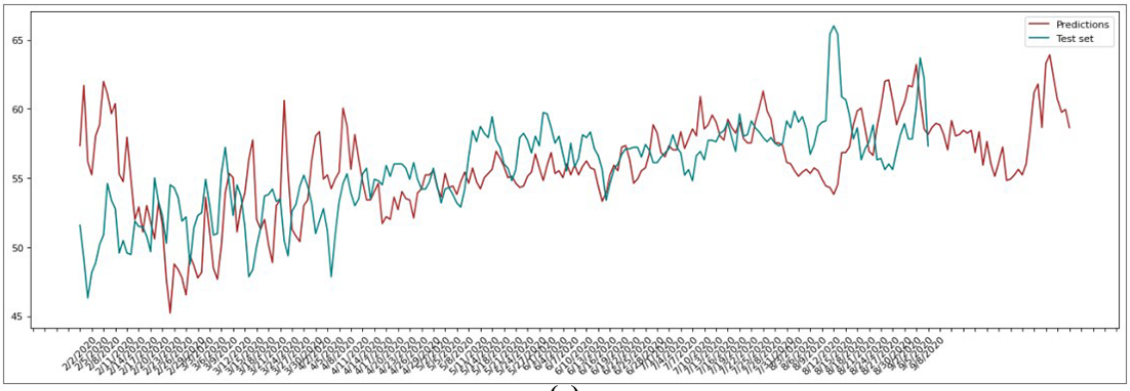


(a)

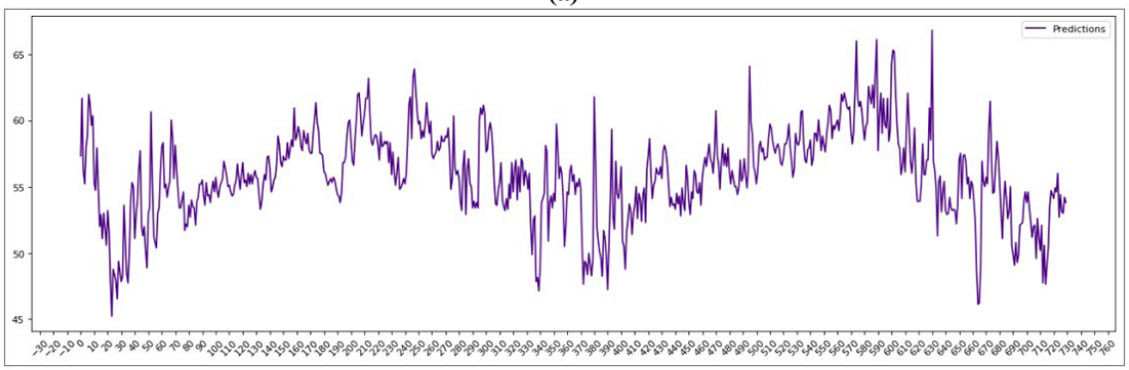


(b)

Figure 16: Validation and Prediction Learning Curve of Nipomo

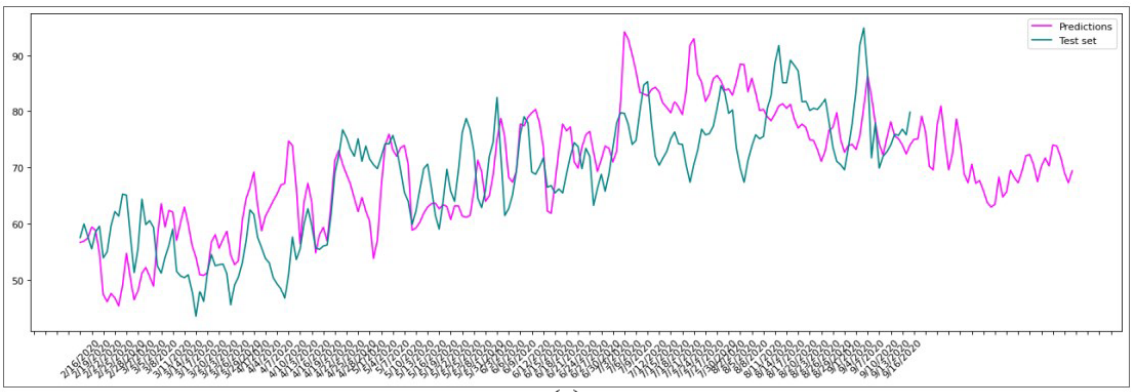


(a)

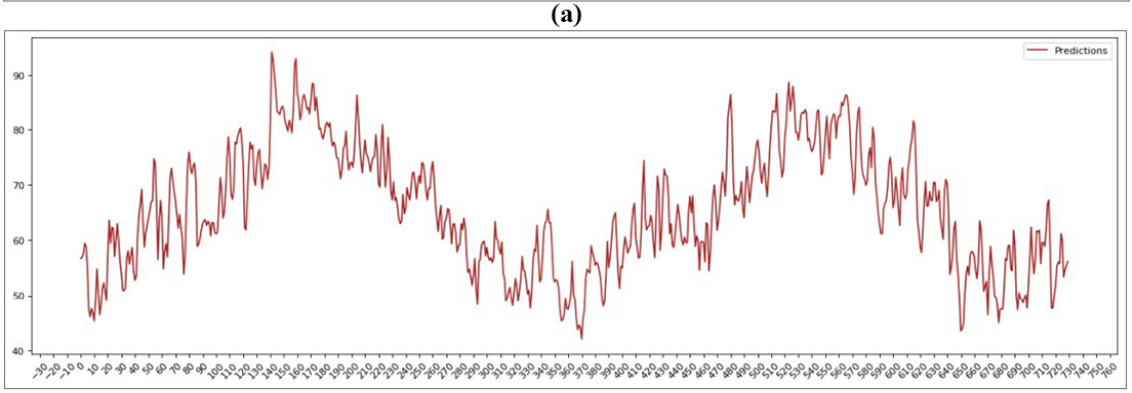


(b)

Figure 17: Validation and Prediction Learning Curve of Pacific Grove



(a)



(b)

Figure 18: Validation and Prediction Learning Curve of Santa Clarita

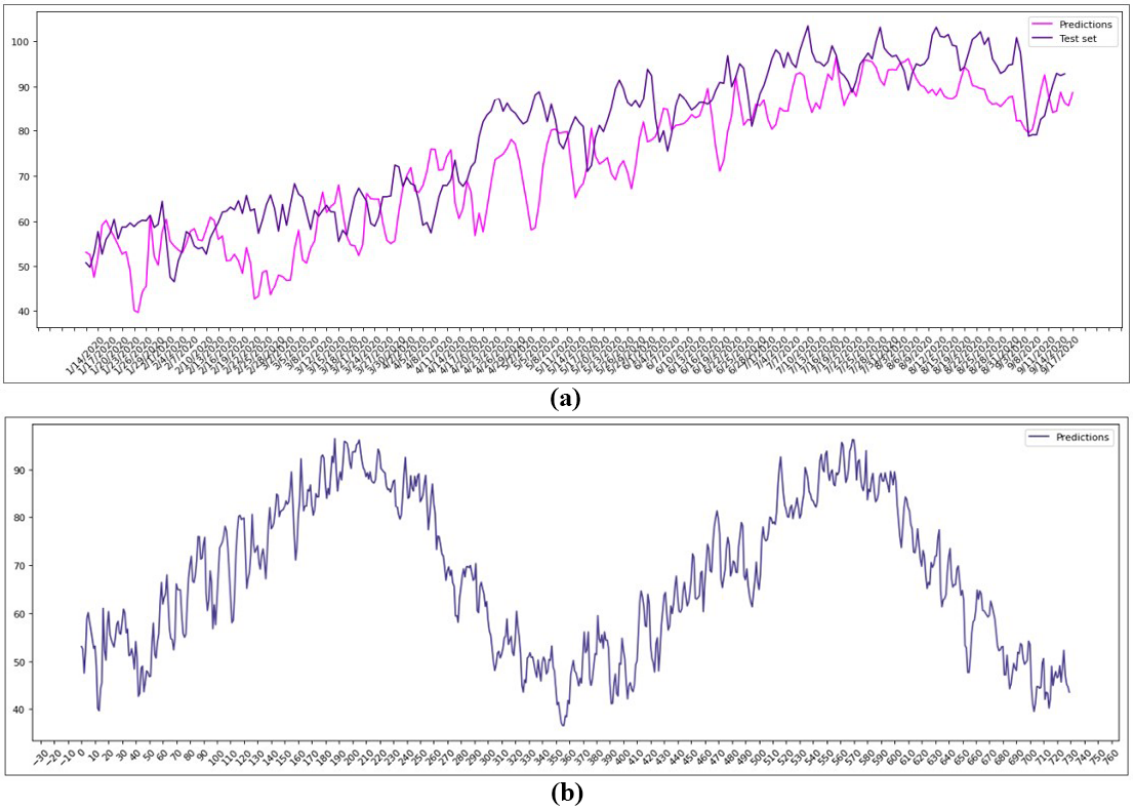


Figure 19: Validation and Prediction Learning Curve of Seeley

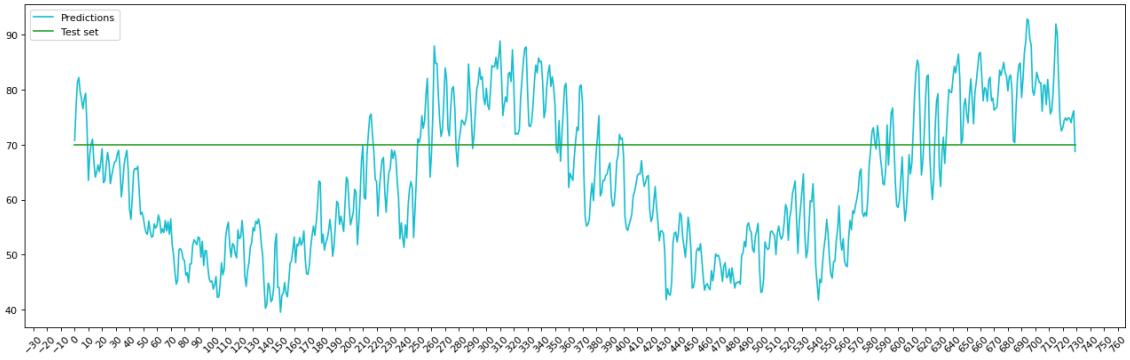


Figure 20: Favourable Temperature vs. Predicted Temperature of Auburn

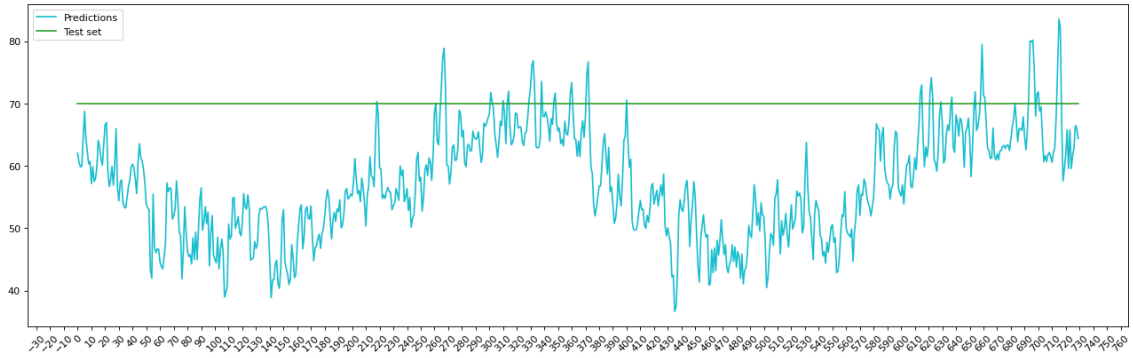


Figure 21: Favourable Temperature vs. Predicted Temperature of Bennett Valley

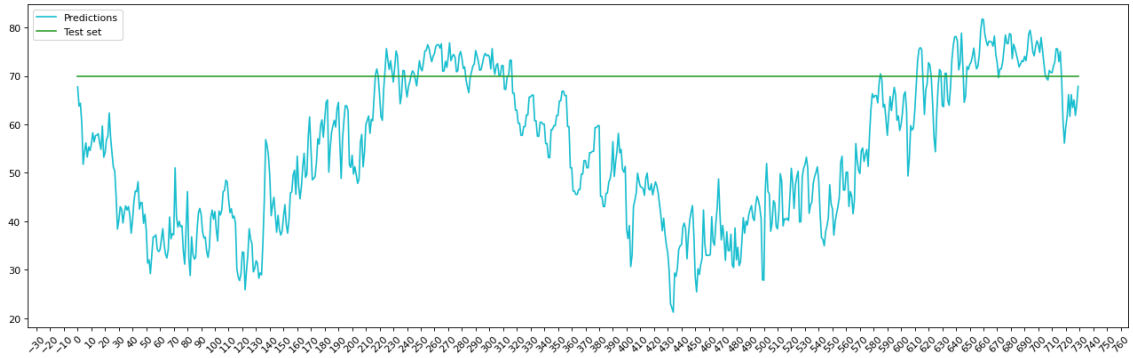


Figure 22: Favourable Temperature vs. Predicted Temperature of Bishop

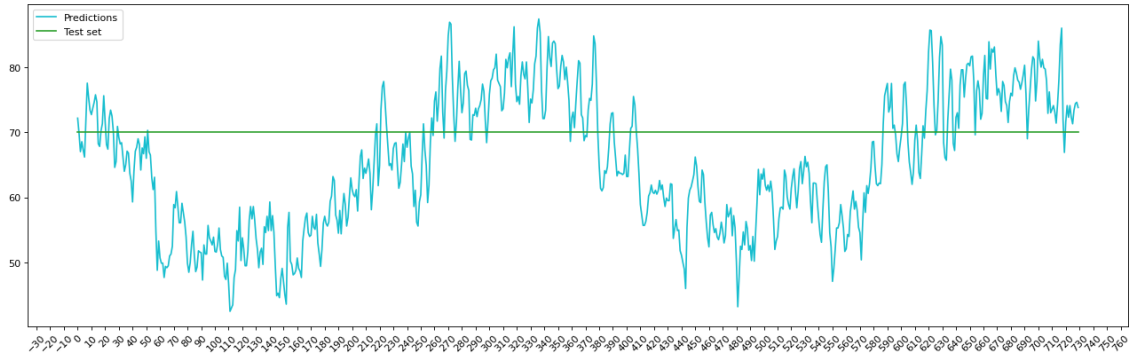


Figure 23: Favourable Temperature vs. Predicted Temperature of Brentwood

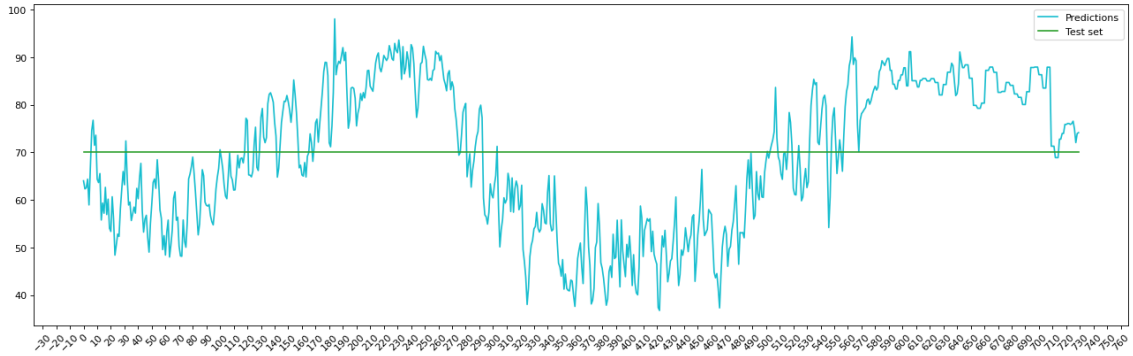


Figure 24: Favourable Temperature vs. Predicted Temperature Buntingville

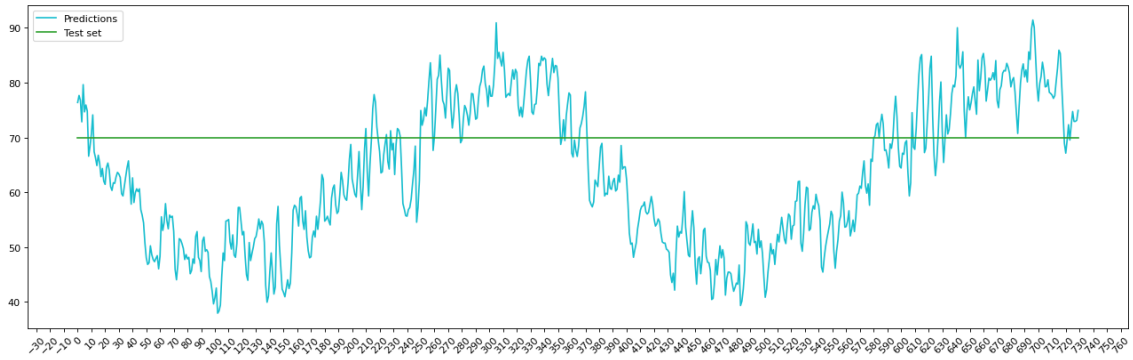


Figure 25: Favourable Temperature vs. Predicted Temperature of Five Points

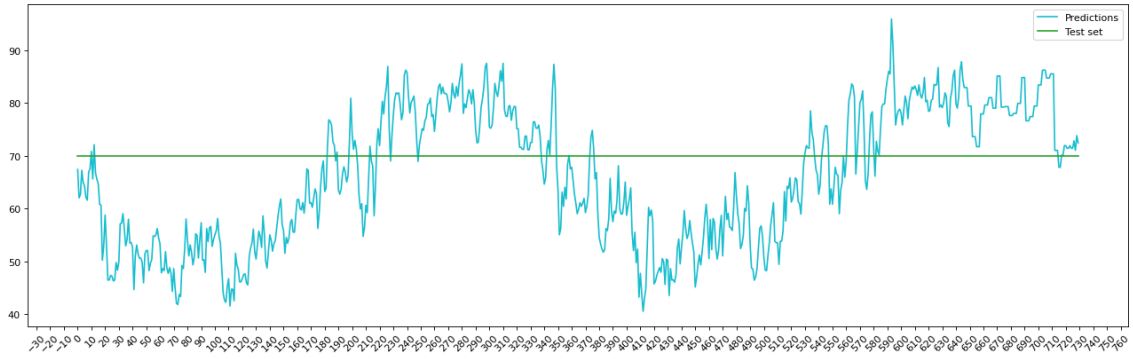


Figure 26: Favourable Temperature vs. Predicted Temperature of Gerber South

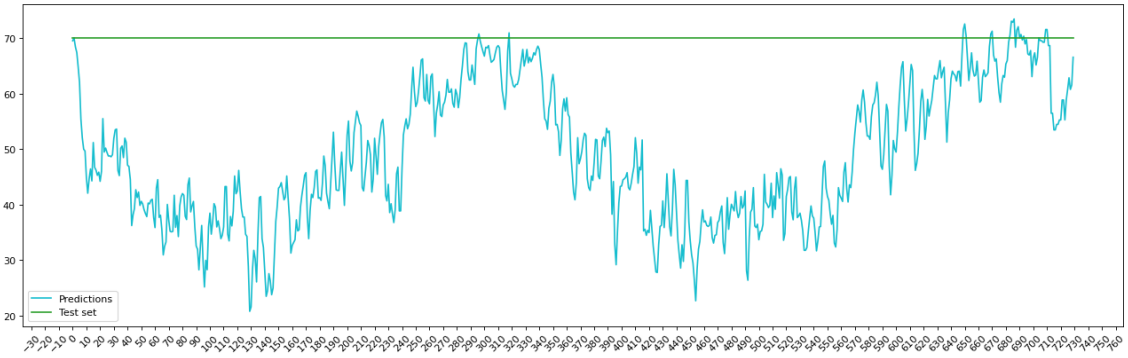


Figure 27: Favourable Temperature vs. Predicted Temperature of Lake Arrowhead

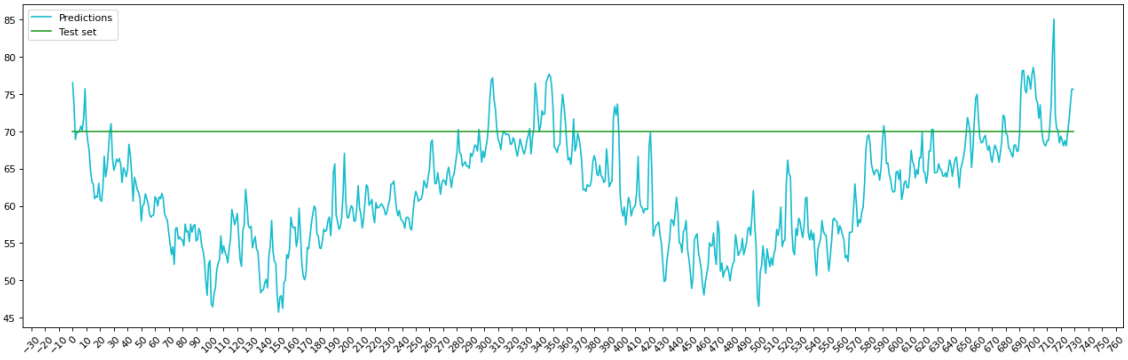


Figure 28: Favourable Temperature vs. Predicted Temperature of Miramar

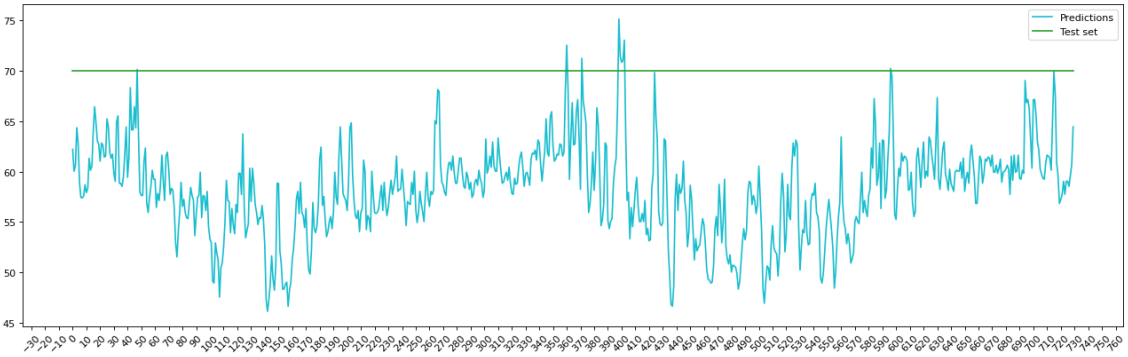


Figure 29: Favourable Temperature vs. Predicted Temperature of Nipomo

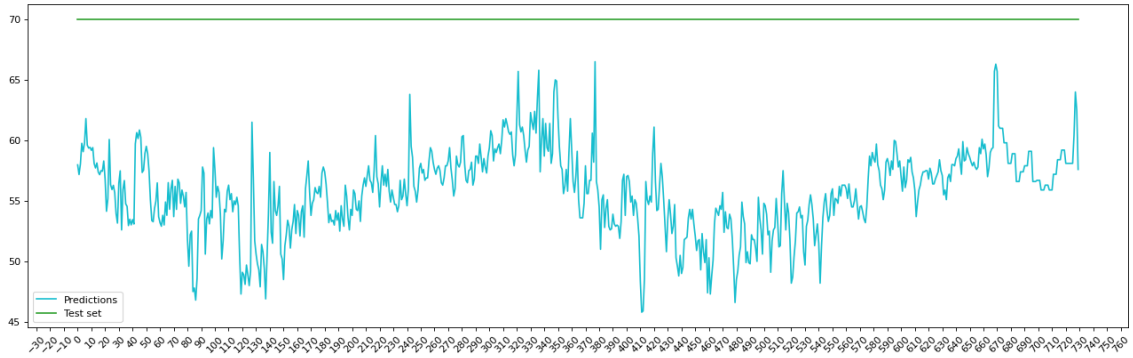


Figure 30: Favourable Temperature vs. Predicted Temperature of Pacific Grove

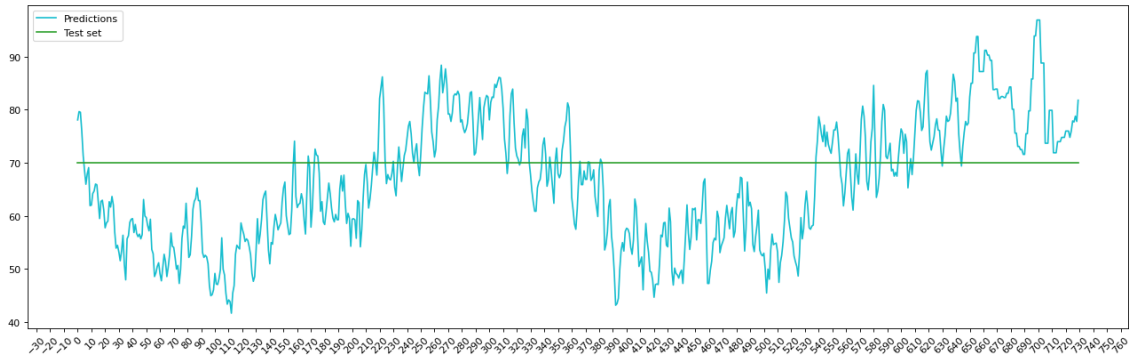


Figure 31: Favourable Temperature vs. Predicted Temperature of Santa Clarita

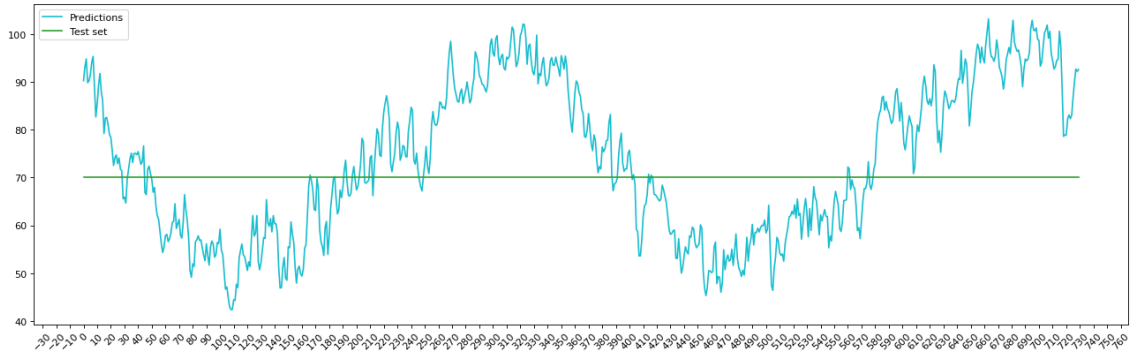


Figure 32: Favourable Temperature vs. Predicted Temperature of Seeley

Appendix A

Appendix B : Table

Appendix B : Tables

Bennett Valley		
Time Period	**Days**	**Temperature**
September 2020	1	49.51032241
	2	49.37236506
	3	45.61008629
	4	41.03238187
	5	51.81055072
March 2021	180	69.22431747
	181	68.9309192
	182	63.62363986
	183	63.62088309
	184	66.3277387
	185	63.32873771
August - September 2021	350	46.12521724
	351	41.02520545
	352	44.1252042
	353	45.82521585
	354	43.72521273
	355	36.52520329
April - May 2022	570	63.82520978
	571	63.72520978
	572	61.42520978
	573	60.82520978
	574	62.82520978
September 2022	726	42.22520978
	727	44.52520978
	728	41.32520978
	729	43.82520978
	730	42.92520978

Table 1: Predicted Temperature of Bennett Valley

Bishop		
Time Period	**Days**	**Temperature**
September 2020	1	39.7787591
	2	26.61278688
	3	31.68158043
	4	36.0627704
	5	36.54758144
March 2021	180	79.25298502
	181	79.46794683
	182	76.67574936
	183	75.28789659
	184	77.99836925
	185	78.02354142
August - September 2021	350	50.185328
	351	48.3169755
	352	49.9484404
	353	46.27982028
	354	43.91119037
	355	50.14263031
April - May 2022	570	81.34626448
	571	77.87790798
	572	77.9095518
	573	75.24119561
	574	75.27283912
September 2022	726	53.78256039
	727	53.81420329
	728	57.24584619
	729	50.87748909
	730	50.409132

Table 2: Predicted Temperature of Bishop

Brentwood		
Time Period	**Days**	**Temperature**
September 2020	1	54.38711857
	2	51.84643971
	3	52.78255792
	4	47.23720049
	5	47.83849696
March 2021	180	76.13084041
	181	79.43084041
	182	78.93084041
	183	75.83084041
	184	75.33084041
	185	79.03084041
August - September 2021	350	52.73084041
	351	49.43084041
	352	48.43084041
	353	48.23084041
	354	45.33084041
	355	44.83084041
April - May 2022	570	76.13084041
	571	78.23084041
	572	76.33084041
	573	75.63084041
	574	78.23084041
September 2022	726	52.83084041
	727	47.53084041
	728	40.63084041
	729	45.43084041
	730	49.93084041

Table 3: Predicted Temperature of Brentwood

Buntingville		
Time Period	**Days**	**Temperature**
September 2020	1	51.93250343
	2	52.8132788
	3	46.70621267
	4	41.70036069
	5	48.63488957
March 2021	180	74.62139622
	181	72.53453282
	182	73.50392507
	183	73.74255077
	184	73.75547032
	185	73.73797829
August - September 2021	350	66.50681643
	351	62.96389495
	352	64.21422112
	353	69.20038519
	354	76.67698639
	355	75.86348856
April - May 2022	570	109.5536076
	571	97.05095639
	572	87.78452015
	573	91.59099115
	574	94.81070029
September 2022	726	111.8701975
	727	111.9590409
	728	105.6400042
	729	99.07630644
	730	98.65826086

Table 4: Predicted Temperature of Buntingville

Five Points		
Time Period	**Days**	**Temperature**
September 2020	1	44.26650089
	2	44.55664775
	3	47.10878049
	4	46.47369366
	5	47.23804551
March 2021	180	87.11087371
	181	86.61087372
	182	86.21087372
	183	86.81087372
	184	86.91087372
	185	87.01087372
August - September 2021	350	44.01087372
	351	39.41087372
	352	39.81087372
	353	40.91087372
	354	46.21087372
	355	50.41087372
April - May 2022	570	77.31087372
	571	75.41087372
	572	77.01087372
	573	75.21087372
	574	77.71087372
September 2022	726	44.81087372
	727	43.41087372
	728	44.11087372
	729	44.91087372
	730	44.71087372

Table 5: Predicted Temperature of Five Points

Gerber South		
Time Period	**Days**	**Temperature**
September 2020	1	54.27499365
	2	47.26347552
	3	53.48042029
	4	54.37253257
	5	46.75355255
March 2021	180	78.22319768
	181	77.84794508
	182	79.37269441
	183	80.39744233
	184	83.72219019
	185	81.84693865
August - September 2021	350	52.73022446
	351	61.05497166
	352	58.57971886
	353	61.20446606
	354	61.52921326
	355	57.75396046
April - May 2022	570	87.67460879
	571	89.599356
	572	89.8241032
	573	87.0488504
	574	88.8735976
September 2022	726	72.83517223
	727	65.15991944
	728	71.08466664
	729	76.50941384
	730	72.13416104

Table 6: Predicted Temperature of Gerber South

Lake Arrowhead		
Time Period	**Days**	**Temperature**
September 2020	1	29.62496124
	2	29.27399414
	3	30.56724709
	4	33.19390319
	5	39.5462403
March 2021	180	68.87687533
	181	67.57687533
	182	67.47687533
	183	67.47687533
	184	70.67687533
	185	72.97687533
August - September 2021	350	38.07687533
	351	34.27687533
	352	36.47687533
	353	39.77687533
	354	39.17687533
	355	35.57687533
April - May 2022	570	62.27687533
	571	61.07687533
	572	60.77687533
	573	61.27687533
	574	61.27687533
September 2022	726	35.27687533
	727	40.87687533
	728	35.17687533
	729	37.77687533
	730	39.67687533

Table 7: Predicted Temperature of Lake Arrowhead

Miramar		
Time Period	**Days**	**Temperature**
September 2020	1	54.97982417
	2	61.48732661
	3	61.10757144
	4	57.77293699
	5	58.91603856
March 2021	180	71.4935934
	181	70.1935934
	182	70.5935934
	183	70.1935934
	184	71.6935934
	185	70.4935934
August - September 2021	350	46.7935934
	351	46.4935934
	352	48.2935934
	353	49.1935934
	354	51.6935934
	355	52.5935934
April - May 2022	570	68.9935934
	571	68.4935934
	572	67.3935934
	573	66.4935934
	574	67.5935934
September 2022	726	51.7935934
	727	51.0935934
	728	49.7935934
	729	51.1935934
	730	52.0935934

Table 8: Predicted Temperature of Miramar

Nipomo		
Time Period	**Days**	**Temperature**
September 2020	1	51.1
	2	53.9
	3	51.3
	4	52.2
	5	47.7
March 2021	180	52
	181	50.5
	182	50.1
	183	48.9
	184	50.3
	185	52.8
August - September 2021	350	41.6
	351	40.9
	352	37.2
	353	40.1
	354	40.6
	355	42.2
April - May 2022	570	50.1
	571	51.1
	572	51.6
	573	50.1
	574	48.2
September 2022	726	40.4
	727	40.3
	728	40.2
	729	39.6
	730	38

Table 9: Predicted Temperature of Nipomo

Pacific Grove		
Time Period	**Days**	**Temperature**
September 2020	1	57.32381057
	2	61.71619738
	3	56.17332786
	4	55.23014781
	5	58.02217555
March 2021	180	57.31881711
	181	56.11881711
	182	56.01881711
	183	55.51881711
	184	55.11881711
	185	55.41881711
August - September 2021	350	53.41881711
	351	54.51881711
	352	53.91881711
	353	59.71881711
	354	57.91881711
	355	55.61881711
April - May 2022	570	61.01881711
	571	59.11881711
	572	58.21881711
	573	59.11881711
	574	62.21881711
September 2022	726	54.41881711
	727	53.11881711
	728	53.01881711
	729	54.21881711
	730	53.81881711

Table 10: Predicted Temperature of Pacific Grove

Santa Clarita		
Time Period	**Days**	**Temperature**
September 2020	1	56.7293083
	2	56.97745194
	3	57.51241398
	4	59.45394841
	5	58.89376531
March 2021	180	78.2984139
	181	79.4984139
	182	80.8984139
	183	81.2984139
	184	80.4984139
	185	81.1984139
August - September 2021	350	53.3984139
	351	52.4984139
	352	52.8984139
	353	52.5984139
	354	51.2984139
	355	47.1984139
April - May 2022	570	74.7984139
	571	72.1984139
	572	68.2984139
	573	71.0984139
	574	79.1984139
September 2022	726	59.9984139
	727	53.3984139
	728	54.6984139
	729	55.3984139
	730	56.1984139

Table 11: Predicted Temperature of Santa Clarita

Seeley		
Time Period	**Days**	**Temperature**
September 2020	1	53.12510316
	2	52.59788442
	3	47.61171993
	4	52.10821692
	5	59.10722949
March 2021	180	92.34723439
	181	86.94751592
	182	84.04808817
	183	86.1487955
	184	84.84946555
	185	89.04994977
August - September 2021	350	48.04890951
	351	44.54890977
	352	40.84890994
	353	41.34890999
	354	39.14890991
	355	37.14890975
April - May 2022	570	96.24890947
	571	96.14890947
	572	93.64890947
	573	87.84890947
	574	91.64890947
September 2022	726	59.9984139
	727	53.3984139
	728	54.6984139
	729	55.3984139
	730	56.1984139

Table 12: Predicted Temperature of Seeley