1 Introduction

‘Weather’ is a blanket term under which there are many components to consider. Be it the temperature, humidity, rainfall, air pressure, wind speed, wind direction, and a multitude of factors that play a role in the overall atmospheric conditions of a region, taking this into consideration; it is a safe assumption to say that it is improbable that the weather of a particular region is the same two days in a row. This makes it a challenge to predict the conditions one might face when actively taking part in an activity. It could be argued that the weather of a region is one of the essential features of the place’s development in terms of infrastructure and the foundations of society. For example, if a site is prone to extremes in weather conditions that make it inhabitable by most communities, then obviously, the prosperity of that community living there is also severely affected. Weather is something that has an influence on all living beings on a daily basis and cannot be avoided or ignored under any circumstances. Thus the underlying motive of weather forecasting could be considered one of the ways we can ensure the safety, evolution, and growth of the human species.

Weather forecasting is the science of predicting the weather of a particular region with the help of statistical principles and models that are based on significant and historical data about a place. This, and the principles of physics that we know influence the weather combined are the foundational principles of weather forecasting. In most prediction models in weather forecasting, we must leave some room for error as it isn’t feasible to assume an accurate prediction with no flaws, primarily because of the numerous factors that play a role in the weather. The data that we use tends to have significant factors, but there are multiple factors that could be directly correlated to the weather, such as geographical location, wind speed, moisture, air pressure, seasonality, and a wide range of subsequent features. Now there is a spectrum of variables that indirectly affect the weather, such as calamities like earthquakes, landslides, tsunamis, and other rare occurrences or phenomena that happen so rarely that they may be considered outliers and not taken into account whilst doing our analysis and predictions, which leads to less accurate results. This being said, science and technology have made a substantial improvement in this field, and so what we can expect is nothing short of sufficient in terms of predictions.

Climate change is a phenomenon that is in direct correlation to weather and may be considered as the periodic modification of Earth’s climate brought about as a result of changes in the atmosphere as well as interactions between the atmosphere and various other geologic, chemical, biological, and geographic factors within the Earth system (Jackson, S. T., 2017). Thus the main aim of this report is to accurately interpret and predict the change in the average temperature of various regions in the world to determine a trend and estimate if the temperature of the planet is increasing or decreasing over time. This is significant and plays a vital role in the well-being of not just the human species but all the living beings on the planet. Ecosystems around the world, both marine as well as land-dwelling creatures as well as plants and microorganisms, live in a delicate balance with each other, and a tiny change in this ecosystem could cause detrimental damage to it. Thus climate change growing at a rapid pace is dangerous and has already led to the extinction of multiple species that have been recorded. It is hard to say the damage it has caused as we do know the possible undiscovered species and environments that have been damaged or destroyed. One this is for certain, and that is that climate change growing at such a rapid pace does not have any positive impact on our planet, and many studies have shown how climate change can directly lead to natural calamities such as the melting of the polar caps in the north pole which leads to the increase in the water levels of the planet which subsequently has led to many communities being displaced from their homes as the land has been captured by the ocean. This is just one of the infinite use cases that prove that climate change is real, and we use real-life data to come to our own conclusions as well.

We use the average temperature of different cities across the world from the time period 1995 to 2019 and asses to see if there is an increase in the overall temperature or not; we do this using the average temperature of the place. We implement various statistical data modelling techniques, such as time series analysis, long-term short memory, which is an application of artificial intelligence, and a few other models, after which we pass them through a few agnostic models that are specifically designed to test the interpretability of each model and check to see which gives us the best results. The dissertation may be broken down as follows.

Chapter 2 consists of a literature review and all the ways, methods, and techniques that have been used in the past to determine the weather forecasting and climate change calculations, followed by Chapter 3, which includes a description of the data we use as well as the methodologies implemented by us to determine our results and finally in chapter 4 we discuss the results and the conclusion we have come to and the significance it beholds.

2 Literature Review

As discussed in the introduction we can see that weather forecasting is an essentiality not just for convenience but rather for the wellbeing and safety of a region. Climate change can be accurately measured with respect to change in the temperature of a region over a period of time. There are various statistical techniques that have been implemented to determine the change in temperature as well as predict it. In this literature review we are going to delve deep into the ways that it has been done in the past. We may also compare to see the most common methods used and their effectiveness. Using these techniques we are able to compare and see the interpretability of each model as well as the statistical significance and accuracy they tend to produce. Some of the methodologies used are as follows.

2.1 General Analysis and Pre processing

This is usually the first stage of most project where we analyse the data to search for missing values, outliers and any trends that may occur over a period of time. In the case where we use variables such as temperature, snowfall in cms, amount of rainfall, which are numerical we usually clean the data by either eliminating the missing values if they are insignificant or we may also use the mean or median of the particular column to fill in the missing data from the previous and next day. There are various libraries in R and Python which are designed to combat this, such as MICE, tidyr, Pandas and many more. As temperature is primarily the variable that we used to determine this climate change we see that in the analysis of climate change in Switzerland by M. Beniston et al. The evolution of daily minimum temperatures at the four stations from the beginning of the century to the end of 1992. Based on the daily temperature values, mean annual statistics have been established; a five-year running mean has been applied in order to filter out some of the high-frequency modes inherent to the inter annual variability. (Beniston, M., Rebetez, M., Giorgi, F. , 1994). The temperature has a pattern which shows that it has increased by 2K over the year it has been observed. This is along with several other variables individually is what the analysis was conducted on. Similarly we see that M.S Shekhar et al in the paper Climate-change studies in western Himalaya used seasonal temperature over a significant period of time where they noted that the maximum temperature as well as the minimum of the region has increased by 2.8 degrees celcius and 1 degree celcius respectively. This inturn lead them to believe there will be anomalies in the other climatic conditions of the region such as at the snowfall in the region and after careful analysis they were able to see a significantly less amount of snowfall over the same period of time. It was observed that there was around 280 cms of snowfall less than the previous years. (Shekhar, M. S et al. (2010)). Another important feature of general analysis and preprocessing is to determine if two or more variables are correlated with each other in any way and try to spot any trends if possible. There are different ways that the correlation can be measured such as Pearson’s Correlation coefficient which is used to quantify the linear relationship between two variables that are random in nature. (W. Xie, M. He et al, ‘2020’) uses Pearson’s correlation coefficient as well as Spearman’s Rank coefficient to understand the type of relationship that exists between wildfire and drought severity, we let the random variable X denote the number of forest wildfire, and the random variable Y denote the Palmer Modified Drought Index (PMDI) which can be used to measure the drought severity. Using the formulas of each correlation coefficient they observed that the Pearson’s correlation coefficient is 0.712. The Spearman’s correlation coefficient has a numerical value of -0.714.These numerical results indicate that these two variables are highly correlated in this case study. The major drawback in this case is that sometimes we may find that the variables are not correlated with each other which is why we need to conduct this analysis multiple times with different variables taken into consideration. The other major setback is that correlation does not necessarily mean causation. This means even if they are very much correlated we cannot assume that they are dependent on each other in any way. Thus further analysis must be done and this can be considered as the initial step of an analysis in the climate change forecasting but definitely not the final step of the process.

2.2 Time Series Analysis (ARIMA)

Time series analysis from its name is a statistical modelling technique that is used for analysis and prediction over a period of time. From the term climate change we can see that change is a process that occurs over a period of time, thus this technique would seem to the most apt for climate analysis. According to (Kaufmann, R.K. et al, ‘2016’) the evidence of the effect of human activity on the climate is mainly evident from two sources: The experiments run by climate models and also the statistical analysis of historical data. In general there are various ways to analyse and predict data over a period of time but the most common way to go about it is using the principle of Regression, the three most popular models that are used as AR, MA and ARIMA models. These represent Auto-Regression, Moving Average and Autoregressive Integrated Moving Average. These are of different orders which have different kappa values that are used to measure which is the best fitted model for a particular dataset with respect to the timeseries. The auto arima function is used to determine which is a good model to fit after which we can use the appropriate model to generate predictions. (Dmritri, T. Ahmad, S. et al ‘2020’) used the precipitation and the maximum as well as minimum temperature from the years 1901-2000 by monthly means to generate a timeseries dataset and conduct the analysis. They also fit a separate SARIMA model on the precipitation and temperature timeseries. They then did the following steps to make sure that they had the best model to fit on the timeseries data to find an accurate and reliable prediction.

Initially the first and foremost step was to determine which order of the series is best suited to stationarise the series. Different ARIMA models were fitted with different orders but having a constant coefficient. Now that the differenced series exists it is considered to be stationary although it may still have auto-correlated errors.

The next step undertaken was to identify the AR(p) and MR(q) components. This can be done by generating an Auto correlation function (ACF) as well as a Partial Auto Correlation function (PACF) which can show us how well the present value of the series is related to that of the past values. By seeing if there is a sharp cutoff of the differenced series on the PACF graph we can observe that AR needs to be added to the model and if the same occurs on the ACF graph we know that MA needs to be added to the model.

Following this they have made some estimations using appropriate p,d and q values which are fitted on the ARIMA models with appropriate residuals. Then the seasonality is removed for the models and the best SARIMA model needs to be selected. After which the forecasting is done and the results are tested under the Akaike Information Criterion which is an estimate of a constant and the relative distance between the unknown true likelihood and function of the data fitted on the model. Thus we can see that the lower the AIC value the closer it is to the truth. Lower values indicate that the less information the model loses the more higher the quality of the model. The Bayesian Information Criterion (BIC) is also a criterion used with the same principle of lower value the better and is based on the likelihood function. Using this method they were able to observe the maximum and minimum temperature of the regions respectively but also were able to forecast predictions for the precipitation which showed a constant trend in values.

2.3 LSTM

Long term short memory is a type of Recurral Neural Network which as a long term dependency which means that it can retain information for a long period of time due to the internal memory it possesses. It is diagrammatically represented in figure 2.2 as shown below

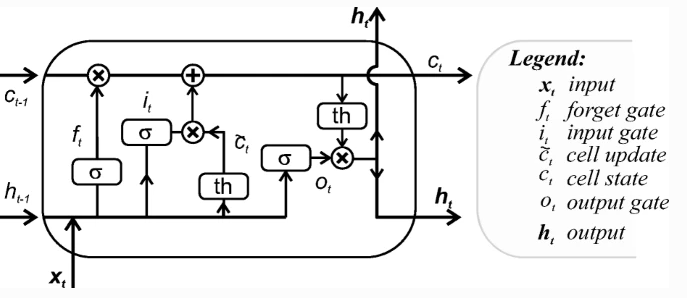
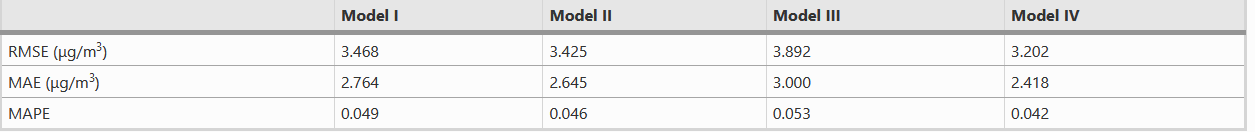


Figure 2.2 The internal structure of LSTM (Hrnjica, B. et al ‘2019’)

As we can see the mechanics revolve around gates primarily these gates are comprised of gates, which are structures through which information is added or removed in the cell state; and also has weights Wj , where j signifies the state name. It contains sigmoid neural, σ, net layer and a point wise multiplication operation. The sigmoid layer is called a forget gate denoted by ft . This element decides what information will be thrown away from the cell state. Its decision is made by looking at xt and ht−1. It consists of two property values, one is the hidden state H(t) which is primarily responsible for the long term memory and the forget gate F(t) adjusts the connection of the input with that of the previous hidden state to the cell state, which then allows it to forget when needed. (Pak, U. et all ‘2018) in the prediction of ozone concentration uses LSTM and CNN models to determine the concentration by implementing the following methods. In the project undertaken by Pak, U. et al they combine CNN and LTSM network in different ways to create four model which has a simple combination of the convolution layer and the layer of LSTM in different combinations such that Model I is a simple combination of the convolutional layer with one layer of LSTM. Model II is a combination of the complete convolutional and pooling layers with one layer of LSTM. Model III is a combination of the convolutional layer with two layers of LSTM. Model IV is a combination of the complete convolutional and pooling layers with two layers of LSTM. They then run all the four models with respect to RMSE, MAE and MAPE and we find that Model IV has the lowest values amongst the four and thus we can figure that it is the best fit model for predicting the ozone concentration. As shown in the table below we can see the values of Model 4 is much better than the rest on all accounts.



They then go on to compare the proposed ozone predictor with some other prediction methods and see that MLP which is an application of artificial neural networks proved to be able to perform that the Conventional time series predictions using statistics. Thus finally they have come to the conclusion that the CNN has efficiently extracted the inherent features and the LSTM has sufficiently retained the long term process of the input time series data. Which makes this a really good model for its prediction performance and also the satisfactory seasonal stability. Although we can see in this case the LSTM is not enough and needs to be combined with CNN in order to get a respectable and satisfactory model for predictions.

2.4 ANN

There are three types of neural networks: recurrent neural networks, multilayer neural networks, and single layer neural networks.

Feed-forward networks with a single layer: The information only moves forward in the network, where the neurons are stacked in layers. The output layers of computational neurons are referred to as the single layers.

Multilayer Feed-forward network : This is different from one layer in that it has one or more hidden layers Due to the fact that these layers don't directly communicate with each other, they are referred to as "hidden" the network's perimeter because their values are not tracked during training.

Recurrent network: Its existence of feedback loops sets it apart from feed-forward neural networks. An example of a recurrent neural network would be neurons that send back their output signal to the input.

Artificial Neural Networks are based on the human brain. In an ANN, the nodes are mostly structured in layers. These layers come in three different types: input, hidden, and output. They are designed to take in a collection of inputs, carry out intricate calculations, and provide an output. The weight of each synapses makes up the ANN's collection of synapses. These weights (wjm) specify how an input will affect a neuron. An adder (additive junction) is a component of the ANN that contributes to adding the weighted signals. A selection criterion may be imposed by the adder based on the architecture. Among these conditions are minimum, maximum, average, and so forth (Kubat, M. , 1999). The basic structural element of ANN is called perceptron, and the transfer function for neuron m is given by yj = ϕ(vj ) = ϕ Xm i=1 wjixj + bj ! with vj = z = Xm i=1 wjixj + bj where {x1, x2, .., xm ∈ R} represents the inputs, w1m, ..., wjm ∈ R are the respective weights of the m neuron. bj ∈ R is the bias that has the effect of increasing or decreasing the net input of the activation function. There exist different kinds of activation functions in machine learning namely sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) (Kubat, M. , 1999). This is diagramtically represented below in figure 2.1.

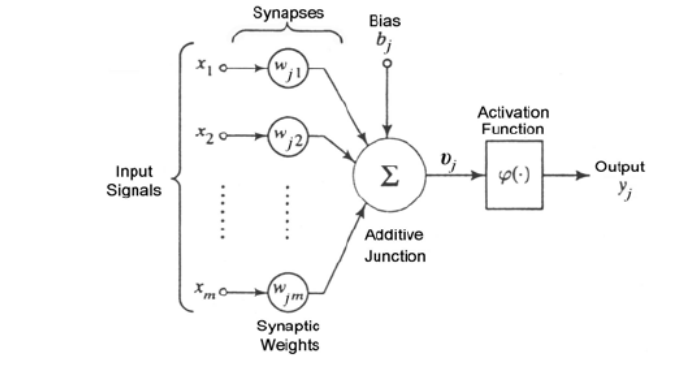


Figure 2.1 Artificial Neural Network Mechanics ((Kubat, M. , 1999)

(Chithra, N.R., Thampi et al. ‘2015’) used ANN-based models were developed for obtaining projections of monthly mean maximum and minimum temperatures at station scale. The capability of the model was assessed by applying it to the Chaliyar river basin in Kerala, India. In the case of prediction of *T* max, data pertaining to the predictors and the predict and were divided into three seasons, viz. dry period (January–May), wet period (June–November) and the month of December. Although in the case of prediction of *T* min, they divided it into 2 seasons (wet and dry) and the networks were split and trained separately for each season. They used the correlation coefficient between the predictors and the predicted values. They found that the ANN model is feasible to downscale the climate data and generate appropriate results which in this case was they found both the maximum as well as the minimum value have both increased. The main drawback in this model and approach is that there is a certain level of uncertainty that may be present in the result. This could be combatted using ensemble models.

2.5 RNN

2.6 Statistical Significance Testing

2.7 Linear Regression

3 Data and Methodology

3.1 Description of Data

Since the primary objective of this project is determine and predict a change in the surface temperature of the Earth, we use a felicitous dataset which has been obtained by the university of Dayton in the united states of America. This dataset has data from the year 1995 upto present day and is regularly updated by the university. The National Climatic Data Center's Global Summary of the Day (GSOD) database serves as the files' primary source of information (NCDC). The Global Summary of the Day (GSOD) data's 24 hourly temperature observations are used to calculate the daily average temperatures. The dataset consists of eight columns namely Region, Country, State, City, Month, Day, Year, and Average Temperature. We have the data of over 324 cities which consist of 157 cities in the US and 167 cities from across the world. Since we are looking at data from 1995 we have 2885424 rows of information under these columns, this is a particularly large dataset and thus we filter and sort out different regions and cities to conduct our analysis as an effective way to measure the average temperature and determine if it is changing over a period of time or not. After some basic exploration we can see that the data that denotes the respective area in which we use as the basis of our analysis is found to be in the object format and the month, day and year are integers where the average temperature is in float format as it has decimal values. There are very few missing values that can be easily rectified with using the mean of the previous and next day temperatures as it is an accurate assumption to make that the missing value is not too varied from the previous and next day in a particular season. There are a few outliers which were easily determined with the help of basic boxplots. Although the data is not perfect and there may be a few indescrepencies we are able to clean it to the point where the dataset is suitable to carry out multiple different statistical calculations using machine learning and artificial intelligence algorithms to make predictions and forecasts.

Conclusion

References

Jackson, S. T.. "climate change." Encyclopedia Britannica, April 27, 2021. Available at <https://www.britannica.com/science/climate-change>.

Beniston, M., Rebetez, M., Giorgi, F. *et al.* An analysis of regional climate change in Switzerland. *Theor Appl Climatol* **49**, 135–159 (1994). <https://doi.org/10.1007/BF00865530>

Shekhar, M. S., Chand, H., Kumar, S., Srinivasan, K. and Ganju, A. (2010) “Climate-change studies in the western Himalaya,” *Annals of Glaciology*. Cambridge University Press, 51(54), pp. 105–112. doi: 10.3189/172756410791386508.

Kubat, M. (1999). Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. The Knowledge Engineering Review, 13(4):409–412.

Chithra, N.R., Thampi, S.G., Surapaneni, S. *et al.* Prediction of the likely impact of climate change on monthly mean maximum and minimum temperature in the Chaliyar river basin, India, using ANN-based models. *Theor Appl Climatol* **121**, 581–590 (2015). <https://doi.org/10.1007/s00704-014-1257-1>

W. Xie, M. He and B. Tang, "Data-Enabled Correlation Analysis between Wildfire and Climate using GIS," 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 31-35, doi: 10.1109/ICICT50521.2020.00013.

Dimri, T., Ahmad, S. & Sharif, M. Time series analysis of climate variables using seasonal ARIMA approach. *J Earth Syst Sci* **129**, 149 (2020). <https://doi.org/10.1007/s12040-020-01408-x>

Hrnjica, B., Bonacci, O. Lake Level Prediction using Feed Forward and Recurrent Neural Networks. *Water Resour Manage* **33**, 2471–2484 (2019). <https://doi.org/10.1007/s11269-019-02255-2>

Pak, U.-J. et al. (2018) “A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction,” Air Quality, Atmosphere & Health, 11(8), p. 883. doi:10.1007/s11869-018-0585-1.