Machine learning based Intelligent weather monitoring and prediction system

By

Semere Gebretinsae

A Final Research Report Submitted to Adama Science and Technology University

Adama, Ethiopia
December, 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The problem of weather prediction for agricultural domain is of prime importance for the agriculture experts, farmers and the research institutions across Ethiopia. This research proffered time series and machine learning modeling techniques for the design, development and implementation of light weight, easy to deploy models for the meteorology centers and researchers in the field of weather forecasting for Ethiopia. Totally five important weather parameters named TEMPERATURE, PRECIPITATION, SUNSHINE HOURS, RELATIVE HUMIDITY and RAINFALL were selected for this research in Adama region. The data was collected from 44 meteorology stations in the region. Past ten years data for 33 variables was obtained from Adama Meteorology center and Addis Ababa meteorology center. Before the machine learning models were actually proposed, a thorough study of existing models for the weather prediction task was performed by the research team. Major limitation of existing, traditional models used by Adama meteorology center and nearby weather stations was the requirement of high performance computing resources, which is not available in every center in Ethiopia. These models need continuous sequence of observations for doing model correction and perturbations. High Performance Computing environment need highly trained staff and continuous training and development activity for the task of weather modeling and prediction, which is difficult to provide in case of rural areas of Ethiopia. In this research, machine learning based weather prediction models were proposed as an alternative way of doing this task. Unlike traditional models which are based on grid based finite element method, and simulate the weather phenomenon across the large (full) span of the geographical area, Machine learning models work on the principle of learning the patterns in the observed data from the recent past. The problem of weather forecasting was formulated in terms of uni-variate and multivariate time series models which were ensembled across yearly observation data to obtain greater confidence in the prediction results. The Block Bootstrap sampling technique

with averaging based ensemble of base models was proposed. Different models were developed to predict five parameters in short, medium and long term settings. The base models used in this research included moving average model, exponential smoothing model, auto-regressive models, autorgeressive neural network model and vector autoregression model. On experiments, for the short term predictions up to 15-20 days, uni-variate ensemble models were found to be more effective, particularly Exponential Smoothing Model and Auto Regressive Integrated Moving Average (ARIMA) model were able to learn the past behavior of selected weather parameters with least mean absolute error. The mean absolute error reported by the Exponential Smoothing model and Auto-regressive integrated models on the short term weather prediction was 0.97 and 0.93 for maximum temperature. Similarly above two models have shown better performance for the remaining weather parameters in short term as well as medium term case, which suggest that the proposed models can be used in weather prediction and creating advisories for agriculture sector as agriculture is more sensitive towards short and medium term changes in weather parameters. Also, auto-regressive multivariate form of neural network was found to be useful and shown a forecast accuracy of 94% which is comparable to existing models like Weather Research and Forecasting (WRF). This research has provided a direction for the Ethiopian meteorology centers to adopt the machine learning in weather prediction tasks as compared with the traditional models. The comparative results and accuracy in the prediction of short term and medium term weather forecasts with less number of resources and simple script based execution of prediction task have encouraged meteorology personal to learn and use techniques proposed under this research. Introducing machine learning models in the prediction task was a new topic for meteorology center in Adama, as the previous system was based on manual data processing and traditional software like Leap and Excel sheet were used for regression based predictions. The current research have successfully introduced and involved the professionals from Adama meteorology center in the process of model

building, validation and testing. In the future, meteorology centers across Ethiopia have a plan to implement High Performance Computing environment, which is essential for the implementation of traditional finite elements based weather models as well as distributed machine learning models.

CHAPTER 1

INTRODUCTION

This research is focused on the problem of weather prediction for Adama and nearby areas like Asella, Ethiopia. From the point of view of farmers and agriculture, this area is very important as a variety of crops, vegetables and fruits are produced in this belt. The quantity and the quality of agricultural products are highly dependent on the weather parameters like temperature, rainfall, wind, soil moister and sunshine across the season. Farmers are always eager to know that what kind of weather is expected in the next quarter, so that they can select a crop which is better suited for the predicted weather. On daily basis, farmers are interested to know the variations in sunshine, wind speed and direction, humidity, rainfall patterns and soil properties, so that they can decide on irrigation requirements of the crops, requirement of insecticides and monitoring of the crop health. Like other countries, this task of agricultural weather prediction is done by meteorology centers in Ethiopia. There are 1200 conventional meteorology centers in Ethiopia, 25 automatic weather stations distributed among 25 directorates and 11 regions, with more than 800 professionals, 400 contractual observations staff and 1200 employees including one center in Adama and a central Meteorology station in Addis Ababa [40]. The job of a meteorology center is to observe, collect and process the data used to predict different weather parameters. Meteorology centers also study that how the weather of a particular place is defined in terms of globally affecting factors like Ocean, global climate and seasonal changes. The meteorology centers periodically observe and record the data across all the regions assigned to them, including all the important attributes affecting the weather of a region, convert the analogous and

manually collected data into the digital form and make this data available for the researchers and the farmers. For researchers, this data is provided as a data set while for farmers, the interpretation of the observed and the predicted data are provided in the form of Agro-advisories.

In addition to the basic exploratory data analysis, meteorology centers uses finite elements based grid models for the prediction of weather under daily, monthly and yearly settings. However, the major problem of traditional weather models like Weather research and Forecasting Model (WRF) is the requirement of High Performance Computing (HPC) infrastructure for model training and forecasting tasks [37]. Providing HPC infrastructure, trained professionals and continuous power supply to maintain 24X7X365 computation of weather parameters is not possible in every part of Ethiopia. Also, traditional models are difficult to customize and update as most of the component scripts are complex and written in FORTRAN and C++, models uses MPI-library for communication over HPC infrastructure, it require a multidisciplinary experts for doing updates, enhancements and extensions in the model. On the other hand, unless these models are tuned with the local terrain, geography and other local parameters, the quality of model output remains low and unacceptable.

Most of the times, farmers are dependent on heuristic knowledge for the prediction of rain, cloud, sunshine and seasonal variations in important weather parameters. Heuristic knowledge have no scientific basis, except it has a mindset among the farmers of being very useful. But, it is observed that the heuristic knowledge can't predict a range of accurate numeric values of 33 weather parameters across the wide region of Ethiopia, round the year. Therefore, the interest in scientific models for weather prediction is felt by various researchers in Ethiopia [39]. In Context of Ethiopia various research on indigenous methods of weather prediction are performed [37] [38] [39]. In [39] a survey of farmers and local participants in Oromia region for understanding the traditional ecological indicators (TEK) including timings and amount of rainfall in the region is performed. It is identified that the

farmers and the local communities heavily relies on traditional methods of rainfall prediction. The color of sunrise, wind direction, stacking and shape of clouds, positions of sun, moon and starts, color of water in the lake, behavior of animals, behavior of bees, bird songs and other atmospheric observations forms the basic indicators, which are available at the hands of farmers for doing rainfall prediction. The rainfall is expected between the June-September and there is no rain between the months of April-May. Author identified that sometimes indigenous methods are unable to predict the delay or early start of rainfall which causes different problems to the farmers regarding sowing of the seeds, maintaining water level in sown crops. The need to investigate the appropriateness of the indigenous methods and empirical comparison with the existing scientific rainfall prediction mechanism is expressed by many researchers in such scenario.

Traditional scientific weather prediction models and their predictive capabilities are limited unless regional parameters and the local factors are integrated into the models such as Weather Research and Forecasting (WRF), Global Forecast System (GFS) etc. During initial survey we have found that the models which are used by the meteorology centers in Adama are based on finite elements method, which is a well-established method in mathematics. There was no use of any machine learning technique or model either in predicting the components of weather, or post processing of the results obtained from the existing models. Existing models were found to be inadequate to explain why certain parameters like rain, temperature and humidity are modified as we travel in various regions and what can be learned from the patterns hidden in the daily observations of these variables. With only global data provided by various international organizations, the traditional models have limited capability to predict and generate true recommendations. This problem causes significant loss in the crop yield due to unmanaged circumstances arsing with sudden weather changes, this limitation of traditional models also poses management issues for the farmers and the government. Due to down-scaling of the global

results in the smaller geographical areas, the results are inaccurate for the local predictions. There is a narrow scope available for meteorology professionals to generate an accurate Agro-advisory based on this data. Another major problem of the traditional models is that their output is composed of complex output types, file formats and Heat-maps, which require significant post processing efforts and technical skills to understand and analyze the output of these models in context of agriculture sector.

In the light of above problems, this research has tried to identify and fill the gaps of existing system and decided to provide better prediction and recommendation models. The pre-contextual study performed by the research team was followed by an in depth literature review, concluded with the proposed alternative direction of modeling the variations in the weather by means of machine learning models. Various researchers have used machine learning models to predict important weather parameters and reported comparative performance with traditional models in terms of accuracy of the results [33] [34] [35]. In this research, the temporal dimension of forecast have been fixed to short term (15-20 days), medium term (2-3 months or seasonal) and long term (a year ahead) weather prediction task. However, this research is more focused on the models for short term and medium term prediction tasks, because these forecasts plays important role in agriculture sector. Initially, the study of existing models and tools used by Ethiopian meteorology centers was performed, the process of observing, collecting, processing, analyzing and regenerating the weather data was studied during various field visits . We also studied the geographical factors and terrain of the study area to understand the current state of the weather prediction capabilities and their limitations in the context of Ethiopia. The main limitation which is found in the current scenario is that the farmers are unable to get the accurate results of the sudden weather events due to knowledge gap between the understanding of concept of weather to the farmers and what the professionals working in the agricultural meteorology centers understands. Farmers are less intended to study and understand the complex outputs of the finite

elements methods, Bitmap Images, raw Graphs, Hot maps, Heat maps and other scientific publications on the meteorology websites not only in Ethiopia but across the world. Instead, they want to grasp the complexities of the weather in the form of simple rules and alarms. As per our study, the weather of Adama (region of study) is highly affected by the global as well as local geographical parameters, which are still not integrated with global models at the national level.

By providing an alternative direction to weather prediction by means of advanced machine learning models, this research has a significant importance for meteorology centers, farmers and various other stake holders in weather related fields. This research identifies the time spans of interest, corresponding models for each time span, forecasting scripts which lead to the development of Agro-advisories and the estimation of adverse events.

PROBLEM STATEMENT

The current system available in the meteorology centers are based on 5^{th} generation numerical methods of weather prediction and does not have any machine learning component, Currently Arial based forecasting is available but this method has limited accuracy because of absence of point and grid based observations. The range and coverage of the models available is minimal and almost non-functional for the short range. Data analysis, post processing of model results and meaningful interpretation are done in the few centers, which is often manual in nature, no support of HPC infrastructure was available for the stations outside capital area, also cluster computing environments were not implemented using available hardware in rural meteorology centers. This research have identified the problems from the point of view of weather experts, farmers and the meteorology centers. The main problem which we have identified for this research is that, the rural area meteorology centers in absence of HPC infrastructure and trained staff, require alternative set of models which can be used

to learn from the past data and easily retrained with the newly available data. In performance these models should be comparable to the existing weather forecasting models, but should run on existing hardware and software infrastructure. The proposed models should be lightweight and can be executed as a standalone script as well as integrated to the web based system. The important weather parameters for the agriculture sector shall be identified and for each weather parameter, the predictions of models shall be segregated in appropriate time interval. The output of the models shall be convertible to Agro-advisories with minimal post processing efforts.

The proposed models shall help experts in the estimation of impact of various events like expected variations in the temperature, rainfall, humidity and other environmental conditions on the various seasonal crops (in the selected regions of the country). In the light of above discussions following sections formulate the objectives and research questions for this research in order to address the problems we have identified as worth addressing.

OBJECTIVES

The objective of this research at the time of proposal was to come up with the suitable Machine Learning models and help in efficient use of existing hardware and sensor devices, satellite data, radar and other type of field specific data available in meteorology centers. This requirement is addressed in this research work in terms of following main and specific objectives:

MAIN OBJECTIVE

The main objective of this research is to develop a set of machine learning models for the prediction of daily, monthly and yearly weather events in order to generate the recommendations for the farmers and other stakeholders.

We formulated the following specific objectives:

SPECIFIC OBJECTIVES

- To identify important weather parameters for crops in the Adama region.
- To study the impact of past observations on the future values for each important weather variable.
- To provide a suitable model for each selected variable.
- To design and develop time series based uni-variate and multivariate forecasting models using the real time data provided by meteorology department.
- To develop ensemble models from the base models.
- To test the models developed for weather prediction by using test data provided by meteorology center.
- To compare the performance of the proposed models.
- To generate the seasonal results with short term, medium term and long term prediction intervals.
- To provide scripts of the proposed algorithms and models along with their parameters in a programming language which can be used by the rural meteorology centers via online/offline platform of meteorology centers.
- To prepare an Agro-Advisory in standard format for dissemination to the farmers.

RESEARCH QUESTIONS

The research questions of this work are as follows

1. Whether the selected machine learning models like ensemble of EMA(q), ARIMA(p, d, q) AR-ANN etc. as proposed in this research are able to significantly modify the existing forecasting capabilities for short, medium and long term weather prediction problem in context of Ethiopia.

2. What is the best machine learning based weather ensemble method for the current study area.
3. How the output of different weather prediction models developed by proposed research can be converted into easy to understand Agro-advisories.

Side Note: In order to achieve our objectives and conduct the research in collaboration with the meteorology center of Adama, we have established a formal communication with the professionals in this organization. The research activity has proceeded with the initial data and necessary information provided by the meteorology center Adama.

Rest of the paper is organized as follows: in chapter 2, thorough literature survey is provided with references to the recent and classical papers. Chapter 3 provide a brief overview of the research methodology, data collection and data pre-processing methods used. Chapter 4 describe the mathematical structure of the models and the proposed solution. Chapter 5 provide the experiments and results with different models proposed in chapter 4. The discussions on the results of various classes of models is done in chapter 6. Chapter 7 provide conclusion, future work which our team has planned, followed by the valuable recommendations.

CHAPTER 2

LITERATURE REVIEW

This section survey important work in the weather prediction field in traditional and recent time, as well as identifies the important gaps and opportunities of research in this field.

WEATHER AND CLIMATE

Study of average daily and seasonal weather observations like temperature, humidity and precipitation estimated over a long term data comes under the climate research. Scope of climate research spans over 25-30 years and the amount of data required for climate prediction is of magnitude of decades. On the other hand weather prediction is a task to predict daily, monthly or yearly weather conditions. The weather predictions are made in terms of important variables like Temperature, Wind Speed, Pressure, Humidity and many more. In both cases, scientists have tried various approaches to model the atmosphere as a physical system and developed the models based on probabilistic, algebraic and finite elements equations approaches. In most of the cases, a weather prediction model require initial atmospheric conditions to be collected across the geographical grid points. A system of partial differential equations is used to simulate the behavior of atmospheric flow, implemented in the form of a model or workflow to represent the global weather phenomenon [35]. Equations are solved using numerical methods because a close form solution does not exist or it is difficult to compute. In the recent days, time series and machine learning based models are proposed by many researchers to model short term prediction task with high accuracy, and similarly deep learning models like Recurrent neural networks (RNN), Long Short term Memory (LSTM) and Convolution Neural Networks (CNN) have shown promising results in the prediction task. Machine learning models are based on learning the

patterns from the past data and are less susceptible to initial conditions and missing observations. Also, the advances in deep learning and ensemble learning methods coupled with the easily available data are expected to bring the field into mainstream machine learning research [35].

IMPORTANCE OF WEATHER AND CLIMATE

Research on climate and weather is very important for Ethiopia in context of drought like situation experienced in the past. The climate and weather influences agricultural yields of various crops. Research in weather forecasting enables farmers to choose right kind of crop for a season and helps monitoring overall health of the crop during the period of cultivation and harvesting. In the rural areas, farmers are mostly dependent on heuristic methods of predicting weather in a season. Based on these traditional methods and human memory for calculations, they select the crop for that season. But our survey has found that this kind of methods are inefficient in predicting the season wide variations in the important weather parameters, and therefore results in underestimation of various weather events. The current situation of lack of scientific information regarding weather conditions is observed across various parts of Ethiopia.

WEATHER PREDICTION: HISTORICAL BACKGROUND

Early weather related research and transmission of important observation data started by 1843 and continued developments between 1870-193 areas are reported in [42]. Initially, weather information was transmitted in the form of weather maps using telegraphy and iconography. This phase involved sharing long term information across various weather monitoring and prediction centers. From 1870 to 1900 the forecasts were based on empirical knowledge. Basic rules of physics started to influence the forecasts starting from 1903, when Vilhelm Bjerknes of Norway put forward the idea of physical models of atmosphere in weather forecasting [41][42]. His research paper introduced seven variables to completely determine the weather of a place, as well as formulated the problem as initial value

problem [41]. The seven basic variables included in his research were air temperature, pressure, air density, moisture content, and the three components of the wind. The seven equations were formulated from the basic physical laws for representing the prediction of the weather conditions, namely, the three hydrodynamic equations of motion, the continuity equation, the equation of state and the equations expressing the first and second laws of thermodynamics. The solution to these equations using numerical method of finite differences was given be Lewis Fry Richardson in year 1920 [42]. This method evaluates the equations on every point in the vertical plane as well as in horizontal plane, therefore, it require huge number of computational steps. The predictions obtained based on only initial values and approximate models were found to be inaccurate for the local predictions, this limitation led to the development of numerical weather prediction (NWP) models with specialized workflow for global atmosphere simulation.

NUMERICAL WEATHER PREDICTION (NWP)

In 1950, John Von Neumann and his colleague, Jule Charney successfully made use of the ENIAC computer system for creating the very first weather forecasts through computer using a simplified atmospheric dynamic model for a 24 hours weather prediction. The computer based implementation of models containing set of physical equations was implemented. The research on numerical techniques to solve these equations were started in early 20^{th} century with the integration of chemical, topological and other details into the models [4]. NWP models advanced with the advancement of computational power of machines. Today, a number of countries uses NWP models to predict their daily weather forecasts on the top of high performance machines and sometimes specialized supercomputers. The internal components of any NWP workflow can be broadly classified into three categories: (i) Governing Equations of the atmosphere [1]; (ii) Numerical Techniques [2]; and (iii) Parameterization of Physical Processes [3].

LIMITATIONS OF THE NWP MODELS

The approximation of actual physical process as implemented in NWP models is not equivalent to the original relation of variables in the atmosphere [5]. Sensitivity of the model towards initial values, non-linearity of parameters are two major problems in NWP models [2] [4]. NWP model need complete observation data regarding the initial condition of the atmosphere at every grid point to be able to predict correctly. It is generally not possible to take observations from every grid point and supply as input to the model while considering the cost of installation of sensor devices in every grid point, especially in inhabited areas [35]. One implementation of NWP modeling technique is weather research and forecast (WRF) model, which is currently used by Ethiopian meteorology centers [30] [32]. In context of Ethiopia, the main problems our team have identified are the lack of available computation power for NWP based weather modeling, unavailability of sensors and automatic devises at all the weather stations in rural areas, absence of customized, third party modules for the integration of local information with global information and unavailability of observation data at certain locations. NWP models if operated in constrained environment, gives limited accuracy in prediction results as we have witnessed in case of Adama.

MACHINE LEARNING MODELS

Due to uncertainties, complexity of computation and other issues like difficulty of integration with the local information, NWP models have limited area of application in context of Ethiopia [14] [32]. The full-functioning implementation of these models require a supercomputer set up and it is a costly operation [15]. Some of the Ethiopian meteorology centers have recently upgraded to the HPC (high performance computing) platforms and special interest groups are planned in various universities for the HPC research. However, the need of lightweight, quickly trainable and scalable forecasting models is being felt at priority. Alternative methods are being explored across the world for the prediction of

weather parameters as well as for post-processing of the results of NWP models. In recent times, Machine learning has evolved as a De-facto alternative for almost all kind of prediction and forecast related problems. Machine learning is a specialization of artificial intelligence in which predictive models are developed using past experience. In weather prediction case, ML models employ past data in the form of time series of observations. ML models for the prediction of weather related events have been used by many recent researchers like [21] [17] [18] [19]. It is found that the limitations of traditional NWP models can be handled using ML and time series type models at various stages. Superiority of the machine learning models over traditional models is established by the fact that they need very less time to train, they can be trained on ordinary computers, as well as they are easy to update and extend. Therefore many researchers have been attracted to apply ML in weather prediction, in Ethiopian context few research reports of using machine learning models for weather forecasting are found [30] [31].

However, it has been a topic of debate that whether the new models based on machine learning and deep learning will be able to replace the numerical weather prediction models or not [35], but we have obtained promising results with short term weather prediction tasks using Auto-Regressive model [17], Moving Average and ARIMA model [18] and Artificial Neural Network based models like Deep Neural Networks such as Recurrent Neural Networks [19].

In the next section we have provided a comprehensive review of the field with pointers to recent work. A feed forward neural network based model is developed by [34] to predict one and two month rainfall in advance using the past two and three time step data as input to the model. Author has used Levenberg-Marquardt (LM) as training function and different experiments were performed with three layer ANN with number of hidden layers between 5 to 40. The evaluation criteria for this research was mean square error and the best r-square value achieved by this model was 0.906 while the model was

trained on the data obtained from various meteorology center of Pune, India. However, Feedforward ANN is very basic model and it may not be suitable for multivariate prediction and considering other constraints like terrain, topography and other geographical parameters. Major limitation of the feed forward neural network is that the training and the test data are completely independent of each other, while in case of rain fall and other weather parameters this assumption do not hold. Also, a feed forward neural network cannot learn the patterns which are globally deciding the intensity of rainfall in a region, this is because a fixed number of historical steps are used to train the network in which patterns of dependency of rainfall on global factors is not given to the network.

An enhanced approach for predicting rainfall generated stream-flow is addressed in [36] using LSTM (long short term memory) networks. In this paper, sequence-to sequence LSTM model is developed for predicting rainfall runoff events in flood like situation in Hostun, Texas. When compared with the grid based models, the LSTM model was able to predict better than ridded Surface Subsurface Hydrologic Analysis GSSA models and improved the efficiency of existing models from .666 to 0.942. For fitting and testing the model author has used the rainfall data from 153 gauges and 10 years data was used to prepare training set for the proposed LSTM model. Author have pointed out that the data driven models are superior than the process driven models for flood prediction as an aftermath of rainfall because there are many places where rainfall gauge are not installed, while ANN and Deep neural network based models have excellent capability to extrapolate the data on these locations. In general, LSTM and RNN models suffers from lacking the methods to select the appropriate features for the training, like how many number of rain gauges are required, they also degrade in performance after certain time of prediction.

The [35] provide an excellent survey over different ANN and Deep learning based alternatives to the Numerical Weather Prediction models. Main limitations of the NWP models are described and

motivations to employ deep learning solutions to NWP workflow are proposed. Main limitations off DL models are also described with citations to the recent publications in the field. In summary, a NWP model require exact state of atmosphere at every grid point in order to be able to optimize the loss function. The atmospheric initial conditions may not be correctly available due to missing data, corrupt senor, far reach areas and loss of signals from various sensors. In this situation, NWP model will provide incorrect results. In addition to data issues NWP model also need huge computation power which is not available in most of the country. The simulation of physical system involve partial differential equations in terms of multiple variables. This can result in a huge computation. Also, since the grid are defined in many square kilometers, the internal changes in the weather profile of specific area need some kind of re- parameterization or model calibration necessary. On the other hand ANN and Deep learning based models are robust for missing values, we can simply drop the corrupted records from training data set. These models have excellent capability to eliminate outliers and their effects in the training phase itself. However, author has pointers to most of the recent work in the field of DL based weather prediction, but some of the serious limitations while training neural network with seasonal and periodic data have been pointed out by the author. ANN are not very good at learning periodic data, and if the NN are trained on long term data composed of many seasons of varying lengths, it becomes difficult to sample the data from each season equally while training the NN.

Based on the literature review, our research have selected various uni-variate time series models to capture the variations in individual weather variables. Multivariate time series models such as variable auto-regression (VAR) and Multivariate, Auto-regressive Neural Network model are selected for the purpose of mapping the effects of various weather parameters on each other.

CHAPTER 3

RESEARCH METHODOLOGY

METHODOLOGY

This Research is based on a hybrid of longitudinal, cross sectional and experimental methodology since it has a span over time across the various weather observation stations, as well as observations were taken before the consideration of effects and after the propagation network was formed. The interest in machine learning based models was raised because; in Ethiopia there is scarcity of advanced hardware and the sensors required to observe the weather parameters at every grid point. The change in representation off information and introduction of new models into the picture have resulted in devising appropriate change in methodology of data collection, pre-processing and selection of samples from available data. While NWP models are strongly affected by initial conditions the sampling at grid level observations do not impact the quality of predictions in case of NWP models, while machine learning models are not based on initial conditions, therefore, sampling of the data is very important to capture maximum possible trend and the seasons. Therefore, in this research we have used block sampling with randomized sample selection technique. The block length of sample is decided as per the requirement of prediction interval. For a season long prediction, we should be able to sample complete sample i.e. 3 months data from training. Missing a season in partial or in complete may result in under-fitting of the models. Similarly for the short term prediction the models are trained on 20 days to 1 month block samples up-to one year data in the past. In ensemble settings we have used bootstrap block sampling, which involve random sampling by replacement of individual blocks. Formation of blocks during training ensure that relevant sequences of the observations are not missed.

Following paragraphs discuss data collection methods used, encoding and data pre-processing steps, (data sets for individual models have been described in the model fitting section in this document), followed by actual description of models and the important definitions. Subsequently we have included the tools and software used and the evaluation criteria.

DATA COLLECTION

Data for this research was collected from Adama meteorology center, Addis Ababa Meteorology center and some parts of the Asella region. There are totally 33 variables which are observed across 44 meteorology stations in this region. We have obtained five year data starting from year 2010. The data was organized in separate excel sheets, each year data was organized into months and days, tagged with the place of observation and the time of observation. The data was encoded and organized in a different way as required by our models. Therefore, our team has to create different data formats by encoding and normalization process for each proposed model separately. The description for each variable and its abbreviation is given in the appendices of this report. Model specific requirements of the data encoding, format, normalization and feature selection task is given in place while each model is defined. Following table provide the details of each abbreviation, name of the variable, description and time of observation which is made available for us from the list of meteorology stations given in appendix.

Abbreviation	Name	Description	Time of Observations
CLDCOV	Cloud cover	Cloud cover, total cloud cover	06:00,09:00,12:00,18:00
CLDTPH	Cloud type high	Cloud type high	06:00,09:00,12:00,18:00
CLDTPL	Cloud type low	Cloud type low	06:00,09:00,12:00,18:00

CLDTPM	Cloud type medium	Cloud type medium	06:00,09:00,12:00,18:00
DRYBUB	Temp, dry bub	Temperature. Dry Bulb	06:00,09:00,12:00,18:00
EVAPND	Evap, pan dly	Evaporation pan, daily total	9:00
EVAPNH	Evap, pan hly	Evaporation pan, hourly	06:00,09:00,12:00,18:00
GNBELD	Radiation, solar dly	Radiation, Solar daily	06:00,09:00,12:00,18:00
GNBELH	Radiation, solar hly	Radiation, Solar hourly	06:00,09:00,12:00,18:00
GRSMIN	Temp, Grass min	Temperature, Grass minimum	06:00,09:00,12:00,18:00
PERMSL	Pres, sea level	Pressure, Mean Sea Level	06:00,09:00,12:00,18:00
PERSTL	Press, stn level	Pressure, Corrected to Station level	06:00,09:00,12:00,18:00
PITCHE	Pitche, hly	Pitche, evaporation hourly	9:00
PRECIP	Precipitation	Precipitation	9:00
RADDIF	Radiation, diffused	Diffused radiation	06:00,09:00,12:00,18:00
RADDIR	Radiation, direct	Direct radiation	06:00,09:00,12:00,18:00
RADGLO	Radiation, global	Glaobal radiation	06:00,09:00,12:00,18:00
RANINT	Rainfall int	1 hour SUM	06:00,09:00,12:00,18:00
SUNHRS	Sunhrs, dly	Sunshine, Daily total Amount	06:00,09:00,12:00,18:00
SUNINT	Sunshn, intensity	Sunshine, Intensity	06:00,09:00,12:00,18:00
TMPMAX	Temp, dly max	Temperature, Daily Maximum	18:00
TMPMIN	Temp, dly min	Temperature, Daily Minimum	9:00
TSL005	Soil temp, 5cm	Soil temperature at 5cm	06:00,09:00,12:00,18:00
TSL010	Soil temp, 10cm	Soil temperature at 10cm	06:00,09:00,12:00,18:00
TSL020	Soil temp, 20cm	Soil temperature at 20cm	06:00,09:00,12:00,18:00
TSL050	Soil temp, 50cm	Soil temperature at 50cm	06:00,09:00,12:00,18:00
TSL100	Soil temp, 100cm	Soil temperature at 100cm	06:00,09:00,12:00,18:00
VISBLY	Visblty	Visibilty, horizontal	06:00,09:00,12:00,18:00
WETBUB	Temp, wet bub	Temperature. Wet Bulb	06:00,09:00,12:00,18:00
WINDLY	Daily wind run	Daily wind run at 2 mts height	18:00
WINSPD	Wind spd, 10m	Wind Speed at 10 mts	06:00,09:00,12:00,18:00
WINDIR	Wind Dir	Wind Direction at 10 mts	06:00,09:00,12:00,18:00

Table 1: List of all Weather parameters under observation

DATA PROCESSING

The meteorology department of Adama has provided us the data from various surrounding stations which is captured by means of sensors and communicated to the central meteorology station. The data is manually recorded into excel files, therefore it is prone to human errors, some sensors missed the observations at certain time due to fault or data was not transmitted because of electricity problem, we had to realign the data and remove unobserved records.

DATA ENCODING

In order to learn the meaningful machine learning models from the data, we encoded each target variable on numeric scale and normalized it, the variables on which a particular weather parameter depends were also coded and normalized into excel sheet and related metadata like day and time of observation, place etc. was also tagged along-with each record. The data was encoded on numeric scale and each identifier was written as abbreviation of corresponding physical quantity as well as a corresponding metadata for the station id for each recorded variable. There are almost 40 stations in which data is manually fed into the excel sheets or printed formats, therefore our team had to work hard to bring the data into a format which is suitable for learning the weather models.

DATA CLEANING

Data cleaning was performed on observation data of all the 33 parameters, due to manual encoding, the data contained a number of missing values, some of the values were outliers and some values were totally out of the domain. We have removed outlier records and for missing values mean values of the column were used as they appeared to be most likely.

FEATURE ENGINEERING

The statistical and machine learning models learn the parameters from the important features. Therefore, we have performed the task of feature engineering such as deletion of duplicate features, elimination of correlated features and selection of appropriate features for each model.

The uni-variate time series models as described in the subsequent sections require a single feature with start date and end date, time of observation, value of observed data. A time series object is constructed from the sequence of observations which have well defined period of observations and all values are present in the data column. For each variable under question, time series object is developed using these features. Features were normalized as per the requirement of forecasting model. The feature set were created for three type of models:

SHORT TERM PREDICTION MODEL (STPM):

For short term, model features were constructed with a time lag of 3:00 hours in observation data. For each member of selected feature subset, we have found daily average of the various physical quantities, which directly or indirectly affect the weather on daily basis. The daily data set is used for learning and forecasting of daily events using appropriate models.

MEDIUM TERM PREDICTION MODELS (MTPM):

The Medium term models tries to predict the monthly averages of weather parameters, therefore a feature set was constructed by taking point average across the various realizations of the observed data to construct features for the across cross-sectional properties of the modeled time series. This type of data is used to learn models for monthly average temperature, humidity, rainfall, and other quantities of interest.

LONG TERM PREDICTION MODEL (LTPM)

This type of models are developed for yearly prediction, therefore the feature set for these models was constructed after selecting important features which causes the weather to behave differently in all the seasons across whole year.

MODELS AND ALGORITHM

The features developed in the form of raw series in the feature construction and feature selection step are subsequently used to develop various time series models like AR model, Moving Average Model, Moving Average Model, Exponential Smoothing Model, ARIMA Model, VAR Model and Neural Network Model. For each model data set is divided into training, validation and test data in its required format. We have selected target weather parameters based on the expert advice and farmers' requirement. The target variables of interest are **TEMPERATURE, PRECIPITATION, SUNSHINE HOURS, RELATIVE HUMIDITY and RAINFALL**, but because of similarity in modeling procedure we have included only selected models in this report. Ensemble models are developed for the respective time span and average values of the predicted results are reported. An algorithm to learn ensemble of the models with block bootstrap sampling is proposed.

TOOLS AND SOFTWARE SUED

Following tools have been used in this research

Weather Models: WRF Model Pre-Parameterized by IGSSA professionals and Pre-trained in 4th kilo Addis Ababa IGSSA Campus.

Machine learning Models and Time series Model from Scratch

Cluster Manager/ HPC: Torque Cluster manager for WRF Model Simulation

Programming and Scripting: R and Python

Data Processing: MS-Excel

Plotting Results: R

EVALUATION CRITERIA

Each class of models are tested under short term, medium term and long term settings using hold out data set. The test data set was created for each model after the main data set was divided in the ratio of 60:20:20. The testing was performed on different ratio of training and test data. The fivefold cross validation was applied to select individual models. The training, validation and test errors were reported and analyzed along-with accuracy and mean absolute error and mean square error.

CHAPTER 4

DESIGN AND DEVELOPMENT

Weather prediction is a complex problem because at any point in time there are hundreds of observable variables, directly and indirectly affecting factors in the each spatiotemporal coordinate in the atmosphere. There is a large number of interaction patterns between cause-effect variables and all the effects cannot be written in the functional form. For example, a place where temperature is high, pressure tend to lower down, inviting winds and therefore clouds movement from high pressure areas to low pressure areas, that may lead to decrease in temperature after certain time period. This is a classic example of one possible cause-effect type interaction between weather predictors. The movement and rotation of earth creates periodicity in certain weather parameters. For example day time temperature and night time temperature has a difference. This difference is also not same in the summer and winter. However, being infinite such interactions possible, it is almost impossible to formulate the mathematical function which can deterministically compute exact physical state of the atmosphere at a particular place and at a particular instant.

Therefore, this research tries to implement the problem in the form of stochastic process in the time. The trend of developing stochastic models as opposed to the physical equation based weather models i.e WRF and GFS is growing, especially time series models and machine learning based models such

as Artificial Neural Network (ANN) have received attention of researchers as they can learn the patterns hidden in the past data, and produce quality forecasts.

We define a particular sequence of observations of a weather predictor as a realization. There, can be infinite possible realizations, each giving rise to a time series data for that variable. The observed values for the particular variable can be analyzed in two dimensions. The first dimension we decide to analyze is at a particular instance of the time across different realizations. For example, we take an average on monthly basis across various realizations of ***TEMP, RH, PRECIP, RANINT, SUNHRS, and WINSPD*** on particular instance of time ***(i.e. 6:00 AM, 9:00 AM, 12:00 PM, 3:00 PM and 6:00 PM)*** across many years to determine the general characteristics of the time series of each individual variable. This kind of models gives us an idea about average temperature, humidity, rainfall, sunshine, and wind speed which prevails (on an average) in Adama city in past many years and therefore in the coming years predicted values are expected to lie around these values. This kind of modeling with certain assumptions of stationarity give us a fair idea of mean and variance of each variable in each season.

However, this model can describe the estimates of mean and variance, but in this way, it is not possible to learn the parameters of the stochastic process across time that is defined in terms of other variables. Therefore, we have decided to model each selected weather predictor as a multivariate time series, which has been generated by a stochastic process under the influence of multiple random variables in the form of a joint probability distribution. The series of observations for each variable can be decomposed into its constituent components, which are, defined as trend component, stationary component (deterministic part), periodic component and the stochastic component.

In both cases, we have used block of samples to retain complete season information and the data was sampled using bootstrap method which perform sampling from dataset by replacement. Next section describe the mathematical structure of different models used under this research.

THE BEST PREDICTION MODEL APPROACH

The best prediction model is that, which computes conditional expectation of a variable on all the other cause variables observed at various time points. Let the variables are coded as follows:

$$Y = \{PRECIP\}$$
$$X = \{TEMP, RH, WINSPD, SUNHRS\}$$

The expected value of Y given that the X variables are observed at an instant of time, will be computed by following integral

$$E(Y \mid X) = \int yF(y \mid x)dy$$

The limitation of this approach is that, the computation of marginal distribution of Y i.e. $F(Y \mid X)$ requires the knowledge of joint probability density function $F(Y, X)$ which is not possible to know in advance at every point in time as well as across all the 33 variables in this research. Since functional form of weather equation fails to describe the cause effect relation among the observed variables in close form, therefore we have decided to approximate models which learn from the past sequence of observations (and at each significant point in time). As an enhancement to the existing models, an ensemble of base models is created in order to predict the value of a particular variable.

PROPOSED CLASSES OF MODELS

The problem of learning to predict weather parameters as a function of observed predictor variables is an approximation problem. There are many ways to model the prediction of future values of ***TEMP, RH, WINSPD, SUNHRS, PRECIP*** taking into account other predictor variables like univariate time series models, multivariate time series models and the models based on Artificial Neural Networks.

In this research, we have proposed two types of ensemble models: Ensemble of uni-variate time series models and second type is pure machine learning based multivariate ensemble models. The purpose of univariate time series model is to understand the behavior of a single weather predictor across the realization of different time spans. While ensemble across the time span at different proposed spatial and temporal resolution is expected to give a better results as compared to the individual models (and comparative results to the WRF and GFS models). In proposed 30KM spatial resolution of observations, the WRF model is parameterized by a suitable weighting mechanism which reflect the topological irregularities in the area of observation i.e. Adama city. This kind of regional details are not easily integrated in the traditional models like WRF and GFS. But in our proposed models there is no dependency on the topography and other regional parameters for short and medium term forecasting (unless a major cycle of weather variation is skipped in the training phase).

The next sections define each model used in this research in their mathematical form and also explains how these models have been used in context of Ethiopia.

AUTO-REGRESSIVE MODELS

Each variable to be predicted is modeled as a univariate time series under a realization. A realization is a yearlong set of observations, we have multiple temporal resolutions for which models are fitted. The daily, monthly and yearly univariate time series models in each year are considered as individual realization of the stochastic data generating process for that variable. The different years of observations are modeled as separate realizations. In very basic form, an Autoregressive model of each realization is learned in order to estimate the statistical properties and the parameters of the model.

The stochastic process which generated the observations for ***TEMP, RH, WINSPD, SUNHRS, PRECIP*** in the form of a sequence of observed random variables at different time instances is defined as follows:

$$Fz_{t1}.....Fz_{tn}(x_1.....x_n) = P(\omega : z_{t1} < x_1, ...z_{tn} < x_n)$$

Where, $Fz_{ti}.....Fz_{tn}$ represent a ***n***-dimensional joint probability distribution of ***n-indexed*** random variables each given by $F(\omega, t)$ where, ω belongs to the sample space. At a particular value of t, $F(\omega, t)$ reduces into a real value called as the realization of the random variable, when computed at different time instances give rise to a time series for that variable. The model we developed, tries to approximately learn the above distribution function, which have generated this time series data, so that the model can be used to predict the future values of the variable.

DEFINITIONS AND ASSUMPTIONS

For autoregressive models following definitions are used:

AUTO COVARIANCE FUNCTION

An ACVF (auto-covariance function) is defined as the covariance between the two observations of a series. Let $v(k_1)$ and $v(k_2)$ be the two observations on the respective time points, then ACVF is computed by following equation:

$$Cov(v_{k1}, v_{k2}) = \sigma_{vv}(k_1, k_2) = E((v[k_1] - \mu_1)(v[k_2] - \mu_2))$$

Where μ_1 and μ_2 are the means of the time series at two instances k_1 and k_2 , for a stationary process ,

mean is a constant so we can replace it by μ (under the assumption of stationarity). Therefore, for stationary process ACVF is a function of time lag between the observations and not the time instances

at which the observations are taken. ACVF is not a time average property of the process, but it is an ensemble property.

AUTOCORRELATION FUNCTION

An ACF (autocorrelation function) is used to compute the effects of the previous observations of a variable in the future values based on the current and the past observations with a given time lag but independent of intermediate observations. ACF between two observations separated by a lag of *i*, is given by

$$\rho = Cov(v_k, v_{k+i})$$

$$= E(v_{k1}, v_{k+i})$$

PARTIAL AUTOCORRELATION FUNCTION (PACF)

A PACF (autocorrelation function) is used to compute the effects of the previous observations of a variable on the future values, based on the covariance between the current and the past observations separated at a particular time lag, conditioned on all the intermediate observations.

$$\rho = Cov(v_k, v_{k+i} \mid v_{k+1} \ldots v_{k+i-1})$$

In this research, we have used a sample ACVF and sample ACF on different values of time lags in order to compute fit for the models. The time lag values were chosen based on the visual inspection in all the three types of forecasting models i.e. daily, monthly and yearly.

STATIONARITY ASSUMPTION

In an integrating process, stationarity of a time series can be maintained by differencing operation. In case of non-integrating process, there may be trend type quadratic non-stationarity, it can be handled by second order differencing operation.

MATHEMATICAL DESCRIPTION OF THE MODELS

This section describe the structure of models, mathematical equations and the formulation we have used for our research.

AUTOREGRESSIVE MODELS (AR)

In an AR process of lag *p,* the prediction of the value at time instance ***t*** is made by consulting previous *p* values in the series. In addition to the past values, the prediction is also affected by a normally distributed white noise given by ε_t . The specification for an ***AR(P)*** model is given by following equation,

$$y_t = c + \phi_1 y_{t-1} + \phi 2 y_{t-2} + \quad _p + \varepsilon_t$$

Where *p*, $\phi_1, \phi 2 \ldots\ldots \phi_{t-p}$ are the parameters of the AR model.

In this research the problem of learning the parameters of the AR model like *p* and $\phi_1, \phi 2 \ldots\ldots \phi_{t-p}$ is handled at three levels of models named as Short Term Prediction Model (STPM), Medium Term Prediction Models (MTPM) and Long term prediction Model (LTPM). Each of the AR model is developed in order to predict appropriate future values of the selected variables.

MOVING AVERAGE MODEL (MA)

A moving average model of order *q* uses the past error values from *q* steps to compute the forecast value at time *t*. The mathematical representation of ***MA(Q)*** model is given by following equation:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

Where,

c = Deterministic component,
$\theta_1, \theta_2 \ldots\ldots$ are the model parameters,
ε_i denotes error in the ith lag prediction

Each set of parameters identifies a unique MA model. For example a MA(1) model is that whose characteristic equation is given by following function:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

MA(1) model learn to forecast from only one past error to do prediction. The meteorology center of Adama uses a software known as Leap which implements MA model to predict three days future temperature values from the observations of past few days. We have implemented above model using R programming language for all other variables of interest and with various time lags identified with the help of ACF function of each variable. All the forecasts were made after learning the parameters of the MA(q) model in ensemble settings. In this report experiments of long term prediction model using MA(q) are included.

EXPONENTIAL MOVING AVERAGE MODEL

An exponential Moving Average model is a weighted model in which the weight of older data points decreases exponentially, but never become zero. For example, given the observed value

as *val*, smoothing factor *sm* and the number of days as *d*, the EMA value of the current day based on the EMA value of one day before is given by the following equation:

$$EMA_{yt}=(Val_t^*(Sm)/(1+d)+EMA_{t-1}*(1-(Sm/(1+d))$$

The recursive definition of exponentially smoothed MA model is given by following characteristic equation:

$$S^t=\begin{cases} Y_1, & t=1 \\ \alpha\cdot Y_t+(1-\alpha)\cdot S_{t-1}, & t>1 \end{cases}$$

Where,

α Control the degree of decay and is valued between 0 and 1,

y_t is the model response at time *t*

S_t is EMA computed at time t.

ARIMA MODEL

ARIMA (Auto Regressive Integrated Moving Average) combines the strengths of past models, it is capable to work on non-stationary time series and have an integrating effect on the past errors. An instance of ARIMA model is written as ***ARIMA(p,d,q)*** where, p is called lag order, it represents the number of time lags used in the model estimation, d is the degree of differencing performed on the raw observations, for example for a quadratic non-stationary series twice the differencing operator will be enough to make it stationary series, ***q*** is the size of moving average

... $\phi_p y_{t-p}$ window. The values of these parameters can be easily learned from ACF plots. The characteristic equation of the ARIMA model is given as follows:

$$y'_t = c + \phi_1 y'_{t-1} +$$

Where,

$c, \phi_1 \ldots \phi_p, \theta_1 \cdots \theta_q$ are the model parameters

ε_t represents the error in step t prediction

The above equation in terms of lag shift operator can be written as

$$(1 - \phi_p B_p)(1 - B)d\, y_t = c + (1 + \theta_1 B + \theta_q B_q)$$

We have used R programming language to implement this model. R has following lag shifted form, with a new parameter added, this model can be shifted for non-zero mean time series:

$$(1 - \phi_p B p)(y_t - \mu) = (1 + \theta_1 B (1 + \theta_q B_q))\varepsilon_t,$$

Where μ is the mean of y_t.

The model parameters in this research are learned for each time span in a separate time series object. Each model has its own ***p, d*** and ***q*** parameters and each model has a forecasting capabilities based on the learned parameters. The ensemble of ARIMA models is designed by our team in order to improve the limitation of a single window model in Ethiopian context as single window model fell short to capture the variations across topography and short seasonal variations. An ARIMA model equivalently reduces into the following specific model based on the values of the parameters *p, d* and *q*:

White noise	ARIMA(0,0,0)

Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Auto regression	ARIMA(pp,0,0)
Moving average	ARIMA(0,0,qq)

Table 1 Models types in ARIMA

REGRESSION MODELS

LEAST SQUARE REGRESSION

The regression model is used to fit a straight line to the time series data, the characteristic function of linear regression model is given as follows,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

In multivariate settings, following regression model has been used with ordinary least square method of parameter estimation:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \quad + \varepsilon_t,$$

$\beta_k x_{k,t}$

Weather variables depends on a number of environmental and atmospheric observations, as seen in this equation used to predict the value of a variable at a particular instant of time as a function of other weather variables.

The quadratic regression is also used to model nonlinear dependency among the observed variables. Regression model with exponential smoothing is also used to improve the prediction capability of the forecasts.

MATHEMATICAL DEFINITIONS OF MULTIVARIATE MODELS

The univariate time series models consider that the effects of all other variables are either negligible or already included in the past values of time series in the form of trend, stationary and stochastic components. But a multivariate model tries to capture the effects of the all the observed variables on the certain variable of interest. In weather prediction, it can be understood as if there are various features like TEMP, PRESSURE, WIND, RH etc. which describe the amount of rainfall in a particular interval. This kind of relation can be modeled with lagged multivariate regression model or using a machine learning model like simple feedforward neural network. In more advanced settings recurrent neural network (RNN) and Long Short Term Memory (LSTM) are used. In multivariate settings, the predicted variable such as rainfall can also affect the TEMP, RH etc. (in the reverse order) as we see in practice that when rain falls, temperature immediately comes down and wind speed increases. This kind of reverse relation can easily be modeled across the RNN and LSTM. But in this report our implementation of multivariate models is limited to vector autoregressive (VAR) model and autoregressive Neural Network models.

Note: We have studied two models for multivariate nonlinear forecasting problem of our selected weather variables, however a number of machine learning models like support vector regression, tree based models, bootstrapping based models like random forest and latest deep learning models are included in the future work of this research. Seems to be promising in near future.

VECTOR AUTO REGRESSION MODEL

The VAR model is defined in terms of two model parameters: the number of variables, which have an effect on the measured property, and the time lag of each variable affecting the prediction values. In weather prediction it can be understood as if PRECIP is a variable whose value for

the next two days is to be predicted. For every next day prediction, we must consider 4 variables in one instance RH, WINDSPD, MINTEMP, MAXTEMP, and each of them affects PRECIP with a lag of 3 days. In this case, we shall call the process model as VAR (4, 3), w here 4 and 3 are the parameters of the model itself. Next set of parameters comes from model equation itself, which relates the variables with each other.

In mathematical form, a vector auto regression model with two variable and one lag can be written with following equation:

$$y_{1,t} = c_1 + \phi_{11,1} y_{1,t-1} + \phi_{12,1} y_{2,t-1} + e_{1,t}$$
$$y_{2,t} = c_2 + \phi_{21,1} y_{1,t-1} + \phi_{22,1} y_{2,t-1} + e_{2,t,}$$

In the characteristic equation of the VAR model, a large number of parameters is required to compute the dependency of each variable on all the other variables, at each possible lag value. Therefore, it is difficult to compute for the problems like weather prediction where large number of variables simultaneously affect each other. In comparison to this model, we have used ANN model, in which there is no need to generate explicit functional form of the multivariate relationships, rather neural network has its own structure (whether be feedforward or feedback network) to learn the parameters form data. At the time of forecast from VAR model, estimated values of parameters shall be plugged into the model equation along with the lags for which the particular variable is to be predicted.

ANN BASED MODELS: AUTOREGRESSIVE FEEDFORWARD NEURAL NETWORKS

The general equation of ANN is given by a weighted sum of all the inputs, fed to an activation function at each node, which generates the output. The summation function is given as follows:

$$z_j = b_j + \sum_i^k w_{i,\,j} * x_i$$

And, the activation function is written as follows:

$$s(z) = 1/1 + e^{-z},$$

In simplest form, there can be a number of hidden layers before a final output is constructed, and there can be only forward direction flow of signals (feedforward NN) or feedback connections can also be made to provide the corrective measures to the previous layers as the signal propagates in forward direction.

The weather prediction problem involves, time series for each variable, therefore not only the selection and number of predictor variables is important, but also the number of lags and how much lag shifted error is used in learning process is a defining factor in the performance of ANN. In this, research we have used lag shifted ANN with different values of lags taken from ACF function of the individual time series and the number of variables is selected based on expert advice. However, the feature selection can also be performed based on PCA or SVD type mathematical procedure, but we have decided to model based on the heuristic knowledge instead. In autoregressive form, ANN learn from different time lagged variables and the number of hidden layers is determined after looking into the suitable time lag value, for each variables under question.

PROPOSED ENSEMBLE MODEL

Same as the characteristic equation, each model has a forecast equation, which tells us how much time lags in the future can be forecasted by the certain learned model, with certain number of parameters of fixed cardinality. Instead of defining each forecasting method, we have implicitly

used previously explained models with the selected hyper-parameters and designed individual fit with required number of parameters, depending upon the number of variables and time lags taken into consideration. The proposed ensemble model has been developed on a sample of model space. The purpose of proposed ensemble model is to average out the errors in prediction of a single model. The architecture of the proposed framework under which models are learned is given in the following figure.

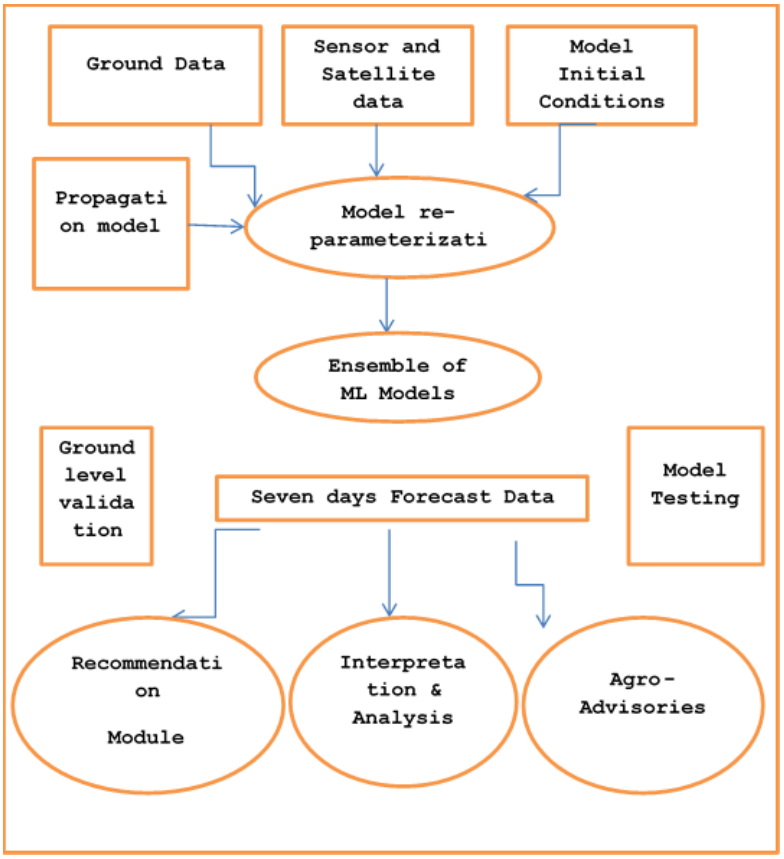

Figure 1: Modeling and Prediction Framework

In this research, we have proposed three types of models to predict selected weather variables. The Short Term Prediction Models focus on learning model parameters from daily data, and this type of models can be used to predict three to five days weather using fitted parameters. The second type of

models are based on monthly data, and this type of models when learned across five years to eight years data can be used to predict three to five months weather in one go. Similarly, the long-term prediction models are used to learn the parameters from yearlong data, and they can be used to predict if there will be some significant change in the weather variables in coming years. Except rainfall i.e. ***PRECIP,*** all other variables have daily non-zero observations round the year. Rain fall has only seasonal existence, therefore its records across many months include zero values, and therefore it is a sparse data. For rainfall, we have decided to perform the modeling and analysis work in the monsoon season of Ethiopia (May-Sep)**.**

The proposed model tries to implement the concept of making ensemble using block bootstrap sampling on observed time series of individual weather parameters. For each type of modeling and analysis work, a set of models are learned on appropriate past data in univariate and multivariate settings, and the parameters are stored in an array. The models learned on the same data but different samples are also stored in the corresponding index. The series is decomposed into its constituent components like trend, stationary component and the stochastic component. A block of sample is taken from the stochastic part of the learned series. At the time of prediction this bootstrap block of sample is convoluted with the predicted value of the output in its scaled form. The sample is convoluted with the output of the model in order to give similar stochastic behavior as seen in the original time series. The stochastic sample is additive in nature and scaled with some random scaling factor (it is convoluted on the same time span to maintain similarity in the seasonal variations).

The output of each model is combined using averaging to generate the result for a given time interval. The objective of making ensemble of individually learned models is to control the unnecessary variance in the output and to average out the effect of bad prediction from one model. This kind of ensemble models are more suitable for univariate time series modeling of the weather

variables, because making ensembles of multivariate neural networks is difficult and computationally not scalable due to the large number of parameters involved. The bootstrap block sampling is a technique used to capture the full season and random components of the weather, such as sudden rain of two or three days in a season of winter. If the training sample size is less than the seasonal block, the important events can be missed and never predicted. Next Figure shows the architecture of the proposed ensemble model and subsequently we also show the exact algorithm which we have developed to facilitate the ensemble learning.

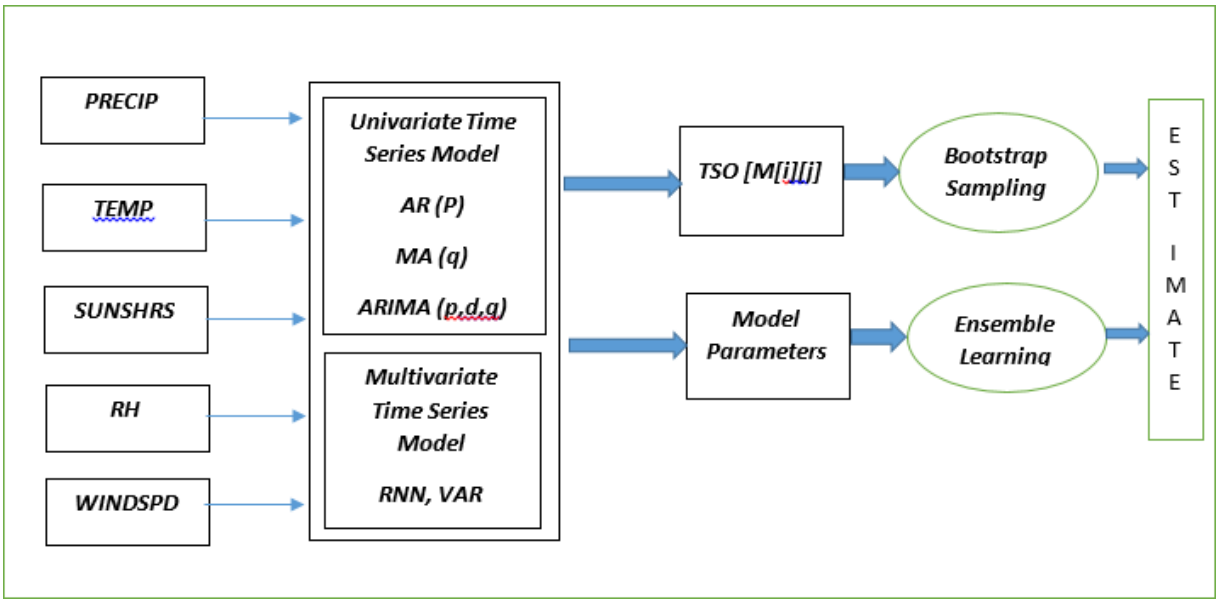

Figure 2: Architecture of proposed System

PROPOSED LEARNING ALGORITHM

The algorithm tries to combine individual models with a scaled stochastic sample taken from the original time series after decomposition. First we learned an individual time series object i.e.(TSO) corresponding to each variable of interest, for a given forecasting period. We have used the block sampling method called bootstrap sampling, which is used to take a sample of certain given number of days in proportion to the time span of model under development. The block bootstrap sampling is done on the stochastic component of a weather variable and its scaled version is convoluted before a forecast is generated. Model outputs are aggregated using averaging method. In fact, bootstrap aggregation is a strong technique to average out the individual errors of the models. Therefore, above algorithm, when used to train the models across various yearly series of observations is expected to give better results as compared to individual models.

$Algorithm\_Ensemble\_forecast(time\_ahead)$

For each season $i = 0$ *to* n

Define TSO $[\,]$ $= \{TEMP,\ RH,\ PRECIP,\ RANINT,\ SUNHRS,\ WINSPD\ \ldots.m\ features\}$

For iterator. $TSO[i] = 0$ *to* m

$M[i][j] = Generate(Model[i][year])$ // save the i^{th} model for j^{th} year

$\hat{\theta}_i = \{\theta_1 \theta_2 \ldots\ldots \theta_k\}$ // save the parameters of the model learned

$Sample(i, j) = Bootstarp(tso, start, end, year)$ // take bootrstap sample for ith tso, jth year between
// start and end dates in particular year

$Enseble\_Model[i] = Bagging((forecast(Model[i][j] \ldots\ldots [m][n]), timestep)$ // m mod *elds*, n *years*
$*scalefactor[i][\delta t] * Sample[startdate][enddate])$ // it can be convolution or additive

$Forecast[timeahead] = predict(Ensemble\_\ \text{mod}\ el[i], timestep)$

repeat

The scripts of the algorithm for individual models are developed using R programming language. Next chapter describe the details of training, validation and testing of the models and forecasting experiments performed on the proposed models.

CHAPTER 5

EXPERIMENTS AND RESULTS

IMPLEMENTATION, EXPERIMENTS AND THE RESULTS

A machine learning model need training before it can be used to do actual predictions. In current scenario, each weather parameter is observed at a particular place in Adama, an equivalent time series is constructed for each parameter, spanning over different time intervals. This section include the information related to training, testing and forecasting of the models in ensemble settings. The corresponding results on short term, medium term and long term forecasting windows are shown. In the end a method to convert a forecast into an Agro-advisory is suggested.

Note: Due to large number of experiments performed on the models described in the previous chapter, we have selected only few models to be included in this report with minimal technical details due to space and redundancy issues. Details of ensemble making process are kept implicit, because as per the weather variable, data set, topography choice of optimal hyper-parameters for ensemble model also vary.

DECOMPOSITION OF TIME SERIES FOR WEATHER FORECASTING

Following components of the time series are extracted by a process of decomposition in our experiments:

TREND

A *trend* exists when there is a long-term increase or decrease in the data. It need not have to be linear. Sometimes we will refer to a trend as "changing direction", when it go from an increasing to decreasing direction and vice versa.

SEASONAL

A *seasonal* pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency.

CYCLIC

A *cycle* occurs when the data exhibit rises and falls that are not of a fixed frequency but repeated.

RANDOM

This component is used to depict the random or stochastic variations on the same dates across years in a time series data. We use this component to add stochastic behavior in ensemble of the learned models with certain scale factor while doing ensemble learning.

EXPERIMENTAL SETTINGS

Each model require specific data preparation step in order to learn the parameters from training data. In addition to the general data processing steps as described in the methodology section of this research, we have included additional data formatting steps along-with the model training steps in the next sections.

DATA SET PREPARATION

During the data preparation task, all the required formatting, standardization, and the conversion of daily records in to monthly record then yearly is performed in specific format required by the model. Each Model is implemented using R programming language and require a data in .csv file to perform the model fitting and subsequent forecasting task. We describe the process for humidity and max temperature in this section, which is repeated for each variable of interest. The similar data sets are

prepared for each model at daily, monthly and yearly resolution and divided into training, test and validation data sets as per the standards.

In the case of humidity, 6:00, 9:00, 12:00, 15:00 and 18:00 hours forecasting data of six year from 2010 to 2015 has been collected. The size our data was six years * 31 values * 5 different hours (2190 days * 31 values * 5 hrs = 339450 instance) has been collected. For the purposes of this study, we have selected 9:00 hours for forecasting purpose then the dimension of the data reduced to 2190 days* 31 values * 1 hour = 67,890. This data have been used build the time series forecasting models with the different lag values. This records was converted to monthly average data. Similar procedure has been followed for all the weather parameters. Regarding the daily forecasting we considered 12 month * 31 value totally: 372 data points have been used as a sequence of row data for training univariate time series models. The following tables summarized the monthly average data for relative humidity and max-temp respectively.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010	42	54	50.77	48.23	52	52	66	68	61	36.41	40.48	40.29
2011	43.9	36.67	36.25	35	43	50	61.7	79.9	62.9	32.25	51.03	40.8
2012	43.38	35.34	29.67	47	36.9	50.43	71.8	70.51	62	36	38	44.58
2013	48.51	37.42	43	45.66	52.48	51.46	71.25	65.22	55.93	45.8	45.6	39.41
2014	42.93	47.1	47.03	41	45.9	43.53	62	66	62	49	46	40
2015	44	33.89	34.06	30	51.38	52.93	58	61	54	37	47	51

Table 1: AVERAGE MONTHLY RELATIVE HUMIDITY

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	27.3	29.4	29.4	31.7	30.8	30	27.4	26.2	26.7	27.9	26.9	25.8
2012	27.4	29	31	30.3	31.7	30.7	25.8	25.9	27	27.9	28	27.2
2013	27.3	29.41	31	31.19	31.14	30.4	26.08	26.31	27.99	28.08	27.5	26.34
2014	28.11	29.5	29.8	30.7	31.1	31.6	28.3	26.7	27.3	26.9	27.5	26.2

2015	27	30.5	31.3	31.5	30.7	30.5	29.5	27.9	28.9	29.9	28	26.9

Table 2: AVERAGE MONTHLY MAX-TEMPERATURE

EXPERIMENTS WITH UNIVARIATE BLOCK BOOTSTRAP ENSEMBLE MODELS

EMA AND ARIMA MODELS

In this section, we report the implementation, training, testing and other details of the univariate block bootstrap based ensemble models such as EMA and ARIMA. We also discuss their results in this section. Similar kind of experimental setup is required for other models, which are not included in this report due to the redundancy of the process, but described earlier in their mathematical form.

As we described in the model definition section, an exponential smoothing model uses weight for the past observations just like a moving average model. But unlike a moving average model, in EMA, more weight is given to the more recent observations. There are three possible smoothing parameters to estimate: the overall smoothing parameter, a trend parameter, and smoothing parameter. If no trend or seasonality is present, then the parameter become null. In this section we describe Short term Prediction Models developed using various component models and depict their results. The hyper-parameter block size used for short term prediction has been varied between 5 to 20 days window size and final value used was consistent with the ACF and PACF plots of the time series.

TEMPERATURE MODELS

We fitted EMA and ARIMA models on temperature time series object using R-script written by our team for training and testing with our ensemble learning algorithm. Totally, five year temperature data is taken for creating the block ensemble of each model in a loop structure. The output of individual models was combined using averaging method. Training time output distribution of maximum temperature for daily and yearly forecasting using EMA algorithm is plotted in the following figures:

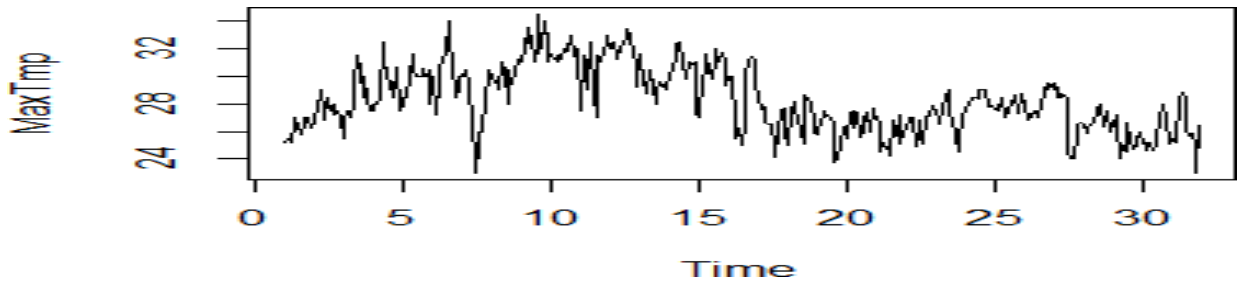

Figure 3: EMA model on Max Temp for Monthly data

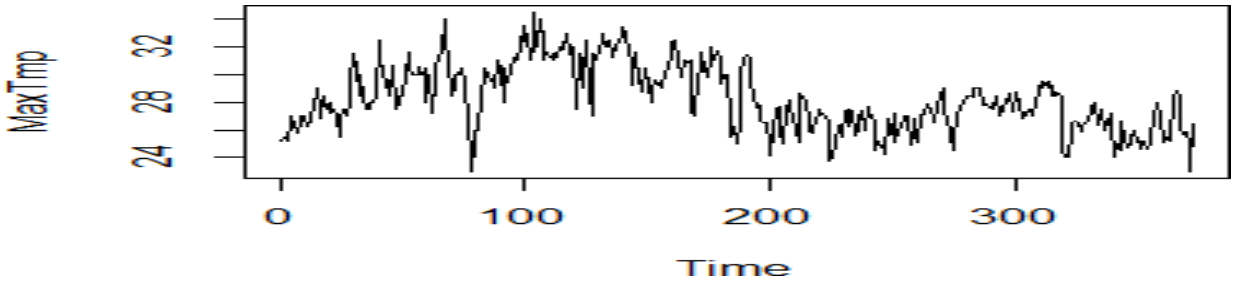

Figure 4: EMA model on Max Temp for Yearly data

From the above figures 3 and 4 the first figure shows yearly forecasting using the daily records and later figure show daily output of the time series.

Holt-Winters exponential smoothing estimates the level, slope and the seasonal component at the current time point. Smoothing is controlled by three **parameters**: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively (at the current time point of prediction).

The adjusted parameters of the EMA model Holt-Winters function details are described as follows:

tempForecast.ts_hw

Holt-Winters exponential smoothing with trend and without seasonal component.

Call:

HoltWinters(x = tempForecast.ts, gamma = F)

Smoothing parameters:

alpha: 0.7499153
beta : 0.0120326
gamma: FALSE

Coefficients:

[,1]
a 25.67825232
b -0.01211893

In this study, we have 372 sequential time series data points for the daily forecasting purpose. Based on the training data we have the next 29 days maximum temperature forecasting using the proposed model. The experimental results of the predicted output is depicted using following data frame

tempForecast.ts_hw_fcst # using EMA model

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
373	25.66613	24.01400	27.31827	23.13941	28.19285
374	25.65401	23.57995	27.72808	22.48201	28.82602
375	25.64190	23.21053	28.07326	21.92345	29.36034
376	25.62978	22.88035	28.37921	21.42488	29.83467
377	25.61766	22.57698	28.65833	20.96734	30.26797
378	25.60554	22.29329	28.91779	20.53989	30.67119
379	25.59342	22.02471	29.16213	20.13555	31.05128
380	25.58130	21.76814	29.39446	19.74958	31.41302
381	25.56918	21.52135	29.61701	19.37856	31.75980
382	25.55706	21.28267	29.83145	19.01995	32.09418
383	25.54494	21.05083	30.03906	18.67179	32.41810
384	25.53283	20.82482	30.24083	18.33255	32.73310
385	25.52071	20.60383	30.43758	18.00100	33.04042
386	25.50859	20.38721	30.62996	17.67612	33.34105
387	25.49647	20.17442	30.81852	17.35710	33.63584
388	25.48435	19.96499	31.00371	17.04321	33.92548
389	25.47223	19.75853	31.18593	16.73389	34.21057
390	25.46011	19.55473	31.36550	16.42861	34.49162
391	25.44799	19.35328	31.54271	16.12693	34.76905
392	25.43587	19.15394	31.71780	15.82849	35.04326

temp_model_forecast # using ARIMA model

	Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
373		26.31909	24.72830	27.90989	23.88618	28.75200
374		26.18163	24.21245	28.15081	23.17004	29.19323
375		26.59271	24.41506	28.77036	23.26228	29.92314
376		26.58004	24.21329	28.94680	22.96040	30.19968
377		26.84465	24.35374	29.33555	23.03513	30.65416
378		26.88573	24.28530	29.48617	22.90871	30.86276
379		27.06523	24.38636	29.74411	22.96825	31.16222
380		27.12537	24.37905	29.87169	22.92523	31.32550
381		27.25315	24.45603	30.05027	22.97533	31.53097
382		27.31602	24.47600	30.15603	22.97259	31.65944
383		27.41072	24.53739	30.28404	23.01635	31.80508
384		27.46919	24.56807	30.37031	23.03231	31.90607
385		27.54159	24.61847	30.46472	23.07106	32.01212
386		27.59306	24.65171	30.53441	23.09466	32.09146
387		27.64967	24.69372	30.60562	23.12894	32.17040
388		27.69365	24.72567	30.66162	23.15451	32.23278
389		27.73860	24.76091	30.71630	23.18462	32.29259
390		27.77554	24.78987	30.76122	23.20935	32.34173
391		27.81162	24.81948	30.80377	23.23553	32.38772
392		27.84234	24.84488	30.83979	23.25812	32.42655

The above prediction result have been plotted in the following figures for ease of understanding:

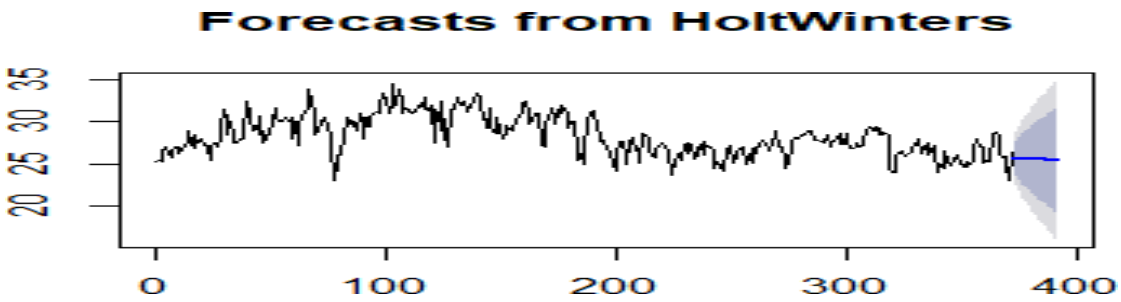

Figure 5: EMA 20 days forecasting

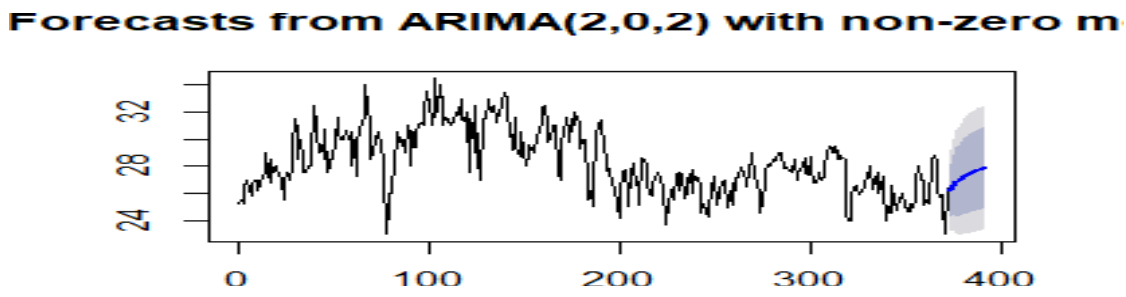

Figure 6: ARIMA 20 days forecasting output

The same data when modeled with the ARIMA model, shows a better fit to the training and test data. One of the potential of ARIMA model is to automatically obtaining the PDQ values to enhance the prediction performance of the model. From 20 days ahead daily forecast, we have realized that the test error produced by the ARIMA model is less than the test error produced by EMA model in individual and ensemble settings using averaging criteria and a block bootstrap aggregation of the stochastic components between 5 to 20 days block length. Therefore, the ARIMA model was selected for further enhancement and analysis for the temperature variable. The final script was prepared after model tuning, analysis of the residuals and the errors on the test set of 20 days. (Shown in the next sub sections).

MODEL TUNING After performing the model tuning by selection of hyper parameters of EMA model for short-term prediction task, final results on the test data are obtained. But this kind of manual hyper-parameter selection during fitting indicate a possible over fit for EMA model. Therefore, ARIMA model was preferred over EMA under auto-fit configuration. Auto-fit ARIMA model has least MSE (mean square error) for the test data upto 20 days.

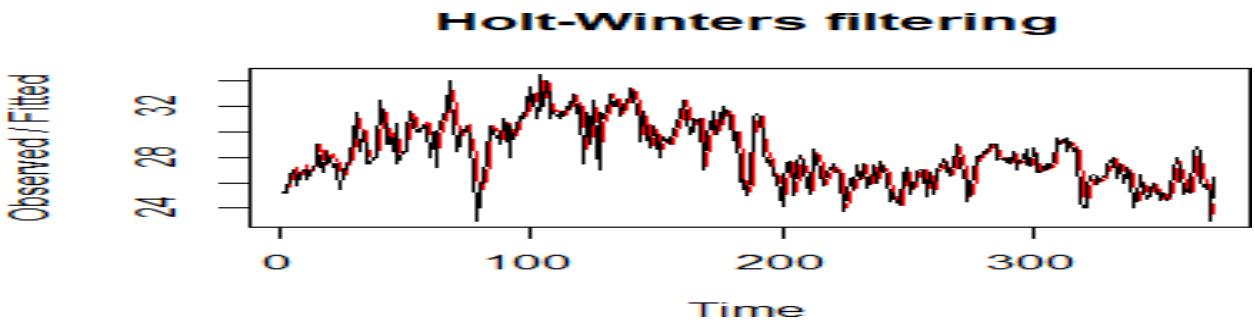

Figure 7: EMA model with Manual Parameter tuning

In the above figure 7, we show the comparison between actual observation values in the test dataset of 20 days and the model predictions for EMA model after forced model tuning which is also possible in case of ARIMA model with auto-fit mode of learning the model parameters.

In the next section, we present the comparison of two models based on the evaluation metric like mean square error, mean average square error and mean of the error.

RESULTS ON TEMPERATURE DATA: EMA VS. ARIMA

The following table summarizes the results of evaluation and prediction performance of the above two time series forecasting models using maximum temperature of Adama region.

Temp Dataset	TS model	Accuracy measurement			
		MAE	MASE	ME	ACF1
Max-temperature	EMA	0.97	0.99	0.09348	-0.0199
	ARIMA	0.937	0.958	0.0139	-4.283

Figure 8: Summary of models performance on temperature forecasting

From the above table, we can see that based on accuracy measurement, ARIMA model outperforms the prediction result obtained by EMA model for the given datasets. This can be explained in terms of

capability of ARIMA model to combine the effects of random components, error components and autoregressive components in one model as compared to EMA model.

RESIDUAL ANALYSIS

Evaluating the statistical distribution of residuals on the training and test data help us to manage if there is irregularity, skewness, and outliers in our output. From the histogram, we can see that the distribution of residuals is Gaussian in nature, which indicate proper fit for ARIMA at the training and the test time.

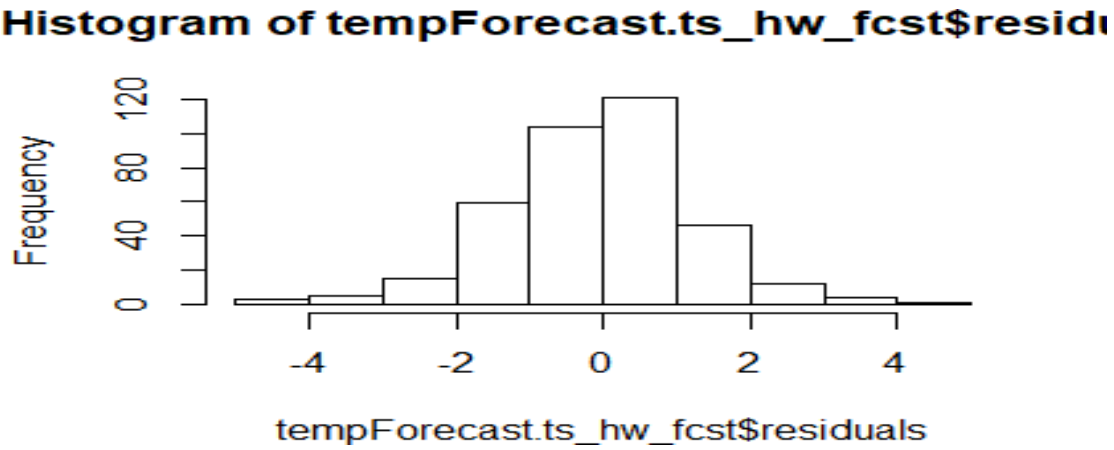

Figure 9: Distribution of Residual errors

Smaller residuals error has been obtained from the proposed ensemble of ARIMA model as compared to the exponential moving average model in manual settings. From this setting of the experiment, we have found that ensemble of ARIMA models with appropriate post-processing, and model tuning is a better candidate for the prediction of daily temperature in context of Ethiopia. (The similar experiment scripts are prepared for other variables for the purpose of deliverable both in individual and ensemble settings with manual and auto-fit options both for short and medium term temperature prediction tasks).

PRECIPITATION FORECASTING

The precipitation forecasting is performed for four months (from month 7 to 9 of from every year's data), the remaining months have the precipitation values almost zero. In this study, five years daily precipitation records have been used to develop the ensemble of yearly forecasting models with block bootstrapping and the additive stochastic component to capture the variability in the training data. The following figure show a sample of input time series for precipitation forecasting.

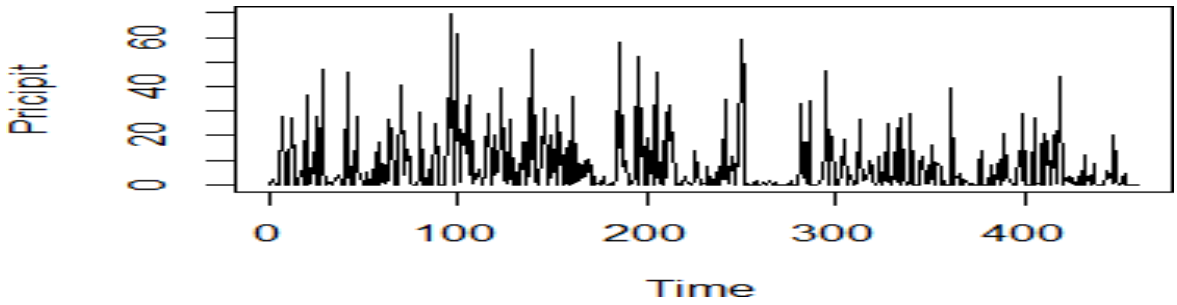

Figure 10: Observed PRICP data

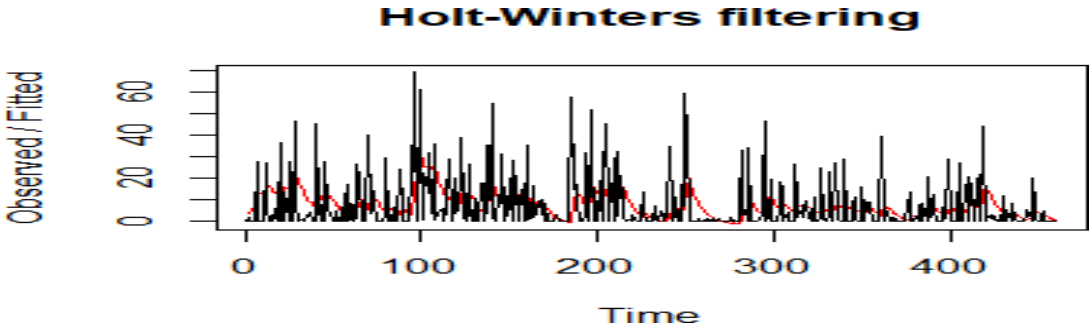

Figure 11: Observed data and fitted the model(red)

From the above figure, we can see that how closely the EMA model is fitted with the observed data. In the parameters of the model there is a slight variation possible due to irregularities in the statistical distribution of the training data, based on the year to year differences in the global conditions related to rainfall.

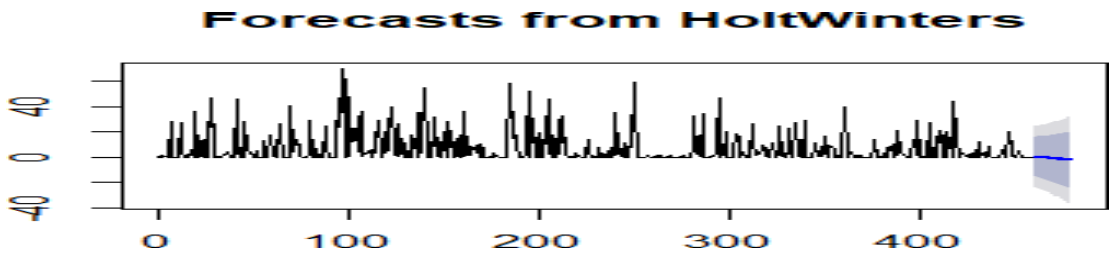

Figure 12: EMA 20 days forecasting output

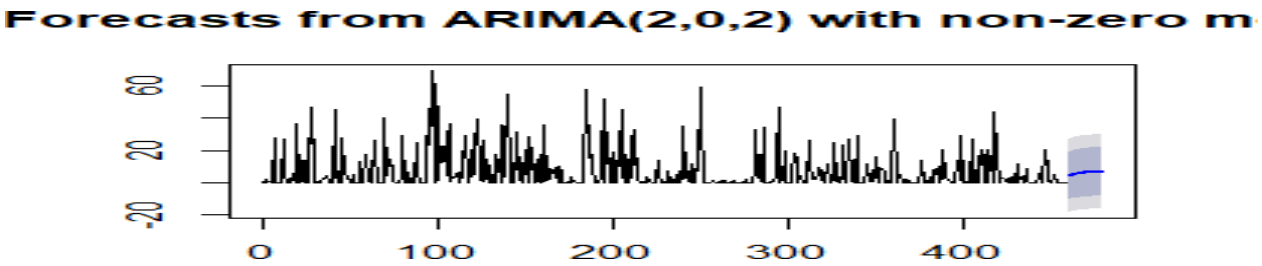

Figure 13: ARIMA model 20 days forecasting

In the case of ARIMA model, we have utilized the auto-ARIMA function to obtain the PDQ values automatically. But, an experienced expert can determine the suitable lag values by seeing PACF.

In the output window of EMA and ARIMA Models, Accuracy measure of precipitation using EMA and ARIMA models are given as follows:

Summary(pricipForecast.ts_hw_fcst) # EMA Algorithm

Forecast method: HoltWinters

Model Information:

Holt-Winters exponential smoothing with trend and without seasonal component.

HoltWinters(x = pricipForecast.ts, gamma = F)

Smoothing parameters:

alpha: 0.1614538
beta : 0.04145294
gamma: FALSE

Coefficients:

```
      [,1]
a  0.5429788
b -0.1373947
```

Error measures:

```
                    ME     RMSE      MAE MPE MAPE      MASE       ACF1
Training set -0.6972935 12.21271 8.642007 NaN  Inf 0.8831332 -0.01709878
```

Forecasts:

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
461	0.405584085	-15.23719	16.04836	-23.51797	24.32914
462	0.268189400	-15.59418	16.13056	-23.99121	24.52759
463	0.130794715	-15.96563	16.22722	-24.48656	24.74815
464	-0.006599971	-16.35158	16.33838	-25.00409	24.99089
465	-0.143994656	-16.75206	16.46407	-25.54383	25.25584
466	-0.281389341	-17.16702	16.60424	-26.10573	25.54295
467	-0.418784027	-17.59640	16.75883	-26.68968	25.85211
468	-0.556178712	-18.04011	16.92775	-27.29555	26.18319
469	-0.693573397	-18.49803	17.11088	-27.92313	26.53599
470	-0.830968083	-18.97000	17.30806	-28.57222	26.91028
471	-0.968362768	-19.45585	17.51912	-29.24253	27.30581
472	-1.105757454	-19.95540	17.74389	-29.93380	27.72228
473	-1.243152139	-20.46844	17.98214	-30.64570	28.15939
474	-1.380546824	-20.99476	18.23367	-31.37790	28.61681
475	-1.517941510	-21.53413	18.49825	-32.13006	29.09418
476	-1.655336195	-22.08631	18.77564	-32.90182	29.59115
477	-1.792730880	-22.65107	19.06561	-33.69281	30.10735
478	-1.930125566	-23.22817	19.36792	-34.50267	30.64242
479	-2.067520251	-23.81735	19.68231	-35.33102	31.19598
480	-2.204914936	-24.41838	20.00855	-36.17748	31.76765

```
pricip_model_forecast # ARIMA Algorithm
Point     Forecast      Lo 80    Hi 80     Lo 95    Hi 95
461       4.466360 -10.400369 19.33309 -18.27034 27.20306
462       4.667186 -10.222977 19.55735 -18.10535 27.43972
463       5.026443  -9.923002 19.97589 -17.83676 27.88965
464       5.297790  -9.686324 20.28190 -17.61843 28.21401
465       5.534774  -9.475558 20.54511 -17.42155 28.49110
466       5.735202  -9.293882 20.76429 -17.24980 28.72020
```

	467	5.905868	-9.136793	20.94853	-17.09990	28.91163	
	468	6.050982	-9.001488	21.10345	-16.96979	29.07175	
	469	6.174407	-8.885156	21.23397	-16.85721	29.20602	
	470	6.279378	-8.785313	21.34407	-16.76008	29.31883	
	471	6.368655	-8.699743	21.43705	-16.67647	29.41378	
	472	6.444584	-8.626496	21.51566	-16.60464	29.49381	
	473	6.509162	-8.563857	21.58218	-16.54303	29.56136	
	474	6.564085	-8.510337	21.63851	-16.49025	29.61842	
	475	6.610796	-8.464641	21.68623	-16.44509	29.66669	
	476	6.650524	-8.425647	21.72669	-16.40649	29.70754	
RESULTS	477	6.684312	-8.392389	21.76101	-16.37351	29.74214	**ON**
	478	6.713048	-8.364037	21.79013	-16.34536	29.77146	
	479	6.737489	-8.339874	21.81485	-16.32135	29.79632	
	480	6.758275	-8.319289	21.83584	-16.30087	29.81742	

PRECIPITATION DATA: EMA VS ARIMA

In the above experiment, outputs of the EMA and ARIMA Models while trained with our algorithm shows the forecasting result with the degree of confidence at 80 percent and 90 percent respectively. The forecasting performance of the proposed models on the precipitation data is summarized using the following table.

Temp Dataset	TS model	Accuracy measurement			
		MAE	MASE	ME	ACF1
Precipitation	EMA	8.6%	0.88%	0.69	-0.017
	ARIMA	8.1%	0.83%	0.022	0.00

Table 3: Performance of the proposed models on the precipitation data

From the above table 3, the univariate time series forecasting models achieves different but almost comparable performance in terms of prediction accuracy on the given precipitation data. But in this

case also ARIMA model is better in terms of MAE (mean absolute error) as well as Mean Absolute standard error.

RELATIVE HUMIDITY MODELS

From the given hourly data-set (6:00 hour and 12:00 hour), data of one year is used to train the univariate time series model for RH and finally averaging ensemble with block bootstrapping is taken across the models developed for different year data. Relative Humidity forecasting give us insight into the water contents of the environment.

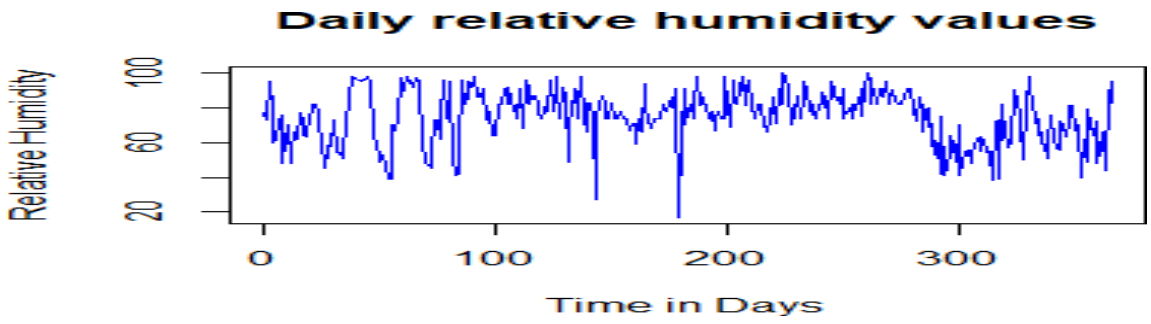

Table 4: Observed data for RH (year)

The above figure 4 shows the distribution of relative humidity data used to train the proposed model.

EMA MODEL FOR RELATIVE HUMIDITY

EMA model with the adjusted parameters have been used to train the model using above 365 records of daily relative humidity observations**.**

```
humidtyForecast.ts_hw # EMA
```

```
Holt-Winters exponential smoothing with trend and without seasonal component.
```

```
Call:
```

```
HoltWinters(x = humidtyForecast.ts, gamma = F)
```

Smoothing parameters:

alpha: 0.5174968
beta : 0.02463962
gamma: FALSE

Coefficients:

[,1]
a 83.1382082
b 0.3901859

Once the model parameters are adjusted to manage the irregularities and outliers in the data, the model response on the training data is plotted. Black line pattern shows the actual data and the red pattern shows the fitted data. From the figure below, we can see that the proposed EMA model is able to capture the patterns in the training data.

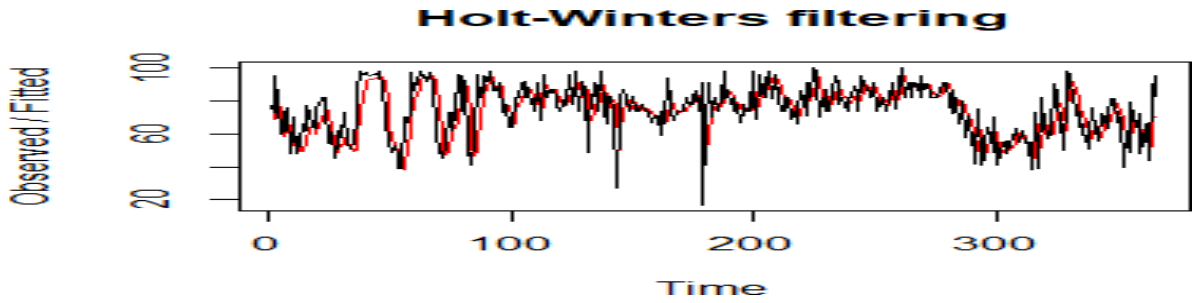

Figure 14: EMA model for RH

After successfully training the model with the help of ensemble learning algorithm, we have demonstrated the forecast for relative humidity for 20 days in advance using EMA model. The model's forecasting output is plotted in the figure 15, followed by the output of training process.

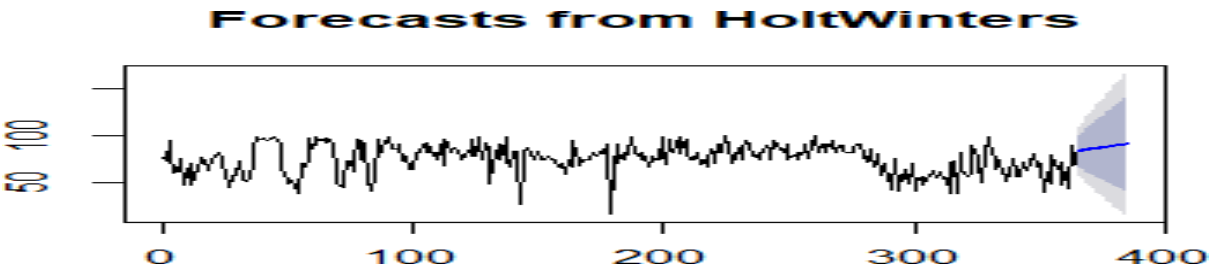

Figure 15: The 20 days RH forecast from EMA

Figure: Relative humidity forecasted output

summary(humidtyForecast.ts_hw_fcst)# EMA Model

Forecast method: HoltWinters

Model Information:

Holt-Winters exponential smoothing with trend and without seasonal component.

Call:

HoltWinters(x = humidtyForecast.ts, gamma = F)

Smoothing parameters:

alpha: 0.5174968
beta : 0.02463962
gamma: FALSE

Coefficients:

	[,1]
a	83.1382082
b	0.3901859

Error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.9484943	12.99643	9.511787	-1.946417	14.74242	0.9297235	0.01552934

Forecasts:

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
366	83.52839	66.89429	100.1625	58.08872	108.9681
367	83.91858	65.09069	102.7465	55.12381	112.7134
368	84.30877	63.42643	105.1911	52.37199	116.2455
369	84.69895	61.86191	107.5360	49.77270	119.6252
370	85.08914	60.37161	109.8067	47.28695	122.8913
371	85.47932	58.93808	112.0206	44.88799	126.0707
372	85.86951	57.54874	114.1903	42.55663	129.1824
373	86.25970	56.19426	116.3251	40.27858	132.2408
374	86.64988	54.86747	118.4323	38.04288	135.2569
375	87.04007	53.56278	120.5174	35.84097	138.2392
376	87.43025	52.27570	122.5848	33.66601	141.1945
377	87.82044	51.00262	124.6383	31.51244	144.1284
378	88.21062	49.74053	126.6807	29.37569	147.0456
379	88.60081	48.48697	128.7147	27.25198	149.9496
380	88.99100	47.23984	130.7422	25.13811	152.8439
381	89.38118	45.99739	132.7650	23.03139	155.7310

382	89.77137	44.75810	134.7846	20.92951	158.6132
383	90.16155	43.52068	136.8024	18.83049	161.4926
384	90.55174	42.28400	138.8195	16.73261	164.3709
385	90.94193	41.04710	140.8368	14.63438	167.2495

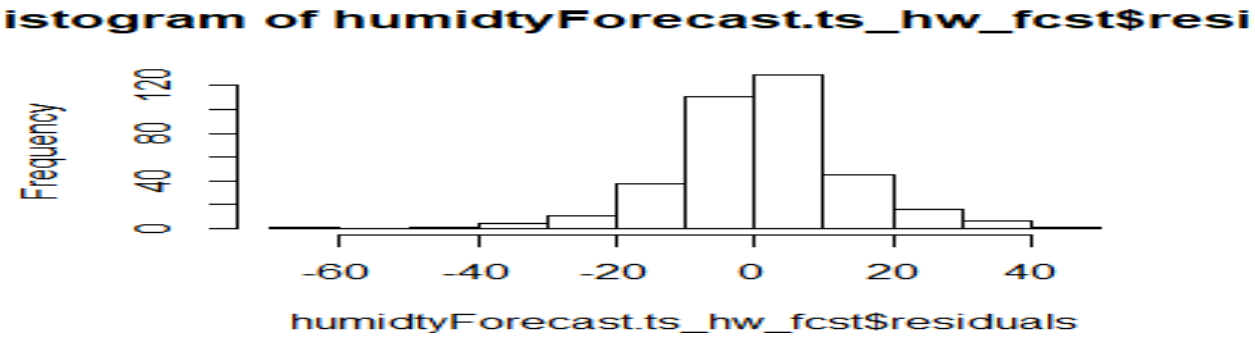

Figure 16: Distribution of Residuals under EMA

The above figure shows statistical distribution of the residuals for the forecast data. As we can see there is no irregularity, skewness and significant outliers in the distribution of the test data, therefore the model is properly fit and the parameters are able to capture the past patterns.

ARIMA MODEL FOR RELATIVE HUMIDITY

ARIMA model is based on the assumption that over a period of time the current values are related or correlated with their immediate previous or *n* previous values. ARIMA encloses the parameters as ***(p, d, q)*** where,

P = values from partial autocorrelation

D = lagged difference between current and previous values

Q = values from autocorrelation

Similar procedure is followed to forecast relative humidity using ARIMA model trained with our algorithm. We utilized auto-arima to get the optimal values of PDQ while training the model. ACF and PACF techniques are used to control the outliers using the lag values. The figures next shows the plots for ACF and PACF values on the same data set as used in EMA model, followed by the training time and testing time outputs of the model.

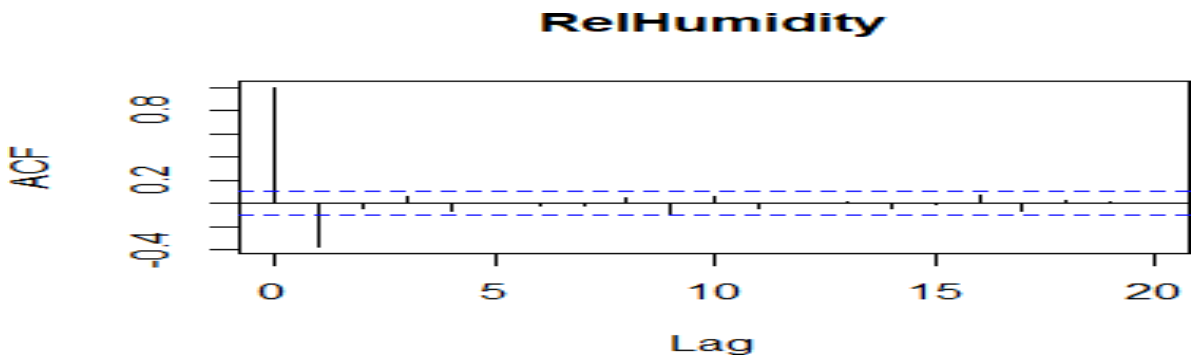

Figure 17:ACF function Plot for ARIMA model

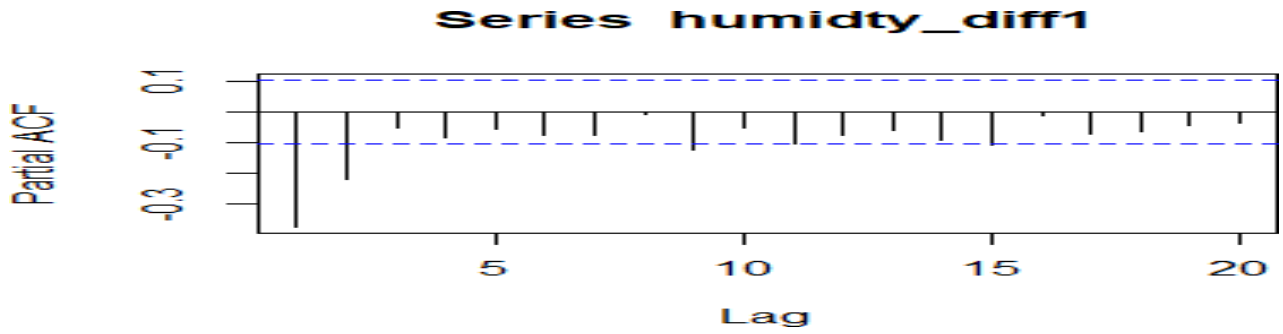

Figure 18: PACF function Plot for ARIMA model

```
Box.test(humidty_arima$residuals, lag = 20, type = "Ljung-Box")
```

Box-Ljung test

data: humidty_arima$residuals

X-squared = 12.721, df = 20, p-value = 0.889

Prediction accuracy of the ARIMA model has been summarized as follows:

```
summary(humidty_model_forecast)
```

Forecast method: ARIMA(2,0,2)

Model Information:

Call:

arima(x = humidtyForecast.ts, order = c(2, 1, 1))

Coefficients:

	ar1	ar2	ma1
	0.4168	0.1748	-0.9702
s.e.	0.0562	0.0555	0.0196

sigma^2 estimated as 150.7: log likelihood = -1430, aic = 2868

Error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.1053576	12.25806	9.146702	-4.057979	14.49081	0.8940385	-0.01609437

Forecasts:

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
366	79.78519	64.05429	95.51609	55.72685	103.84353
367	77.98919	60.76075	95.21763	51.64056	104.33781
368	74.58032	56.28801	92.87263	46.60464	102.55600
369	72.84538	54.06376	91.62700	44.12137	101.56939
370	71.52618	52.45363	90.59874	42.35722	100.69514
371	70.67295	51.42448	89.92143	41.23496	100.11095
372	70.08665	50.72059	89.45271	40.46881	99.70449
373	69.69308	50.24256	89.14359	39.94607	99.44008
374	69.42651	49.91065	88.94237	39.57957	99.27345
375	69.24658	49.67692	88.81625	39.31736	99.17580
376	69.12498	49.50867	88.74128	39.12443	99.12552
377	69.04283	49.38445	88.70120	38.97794	99.10772
378	68.98732	49.28984	88.68480	38.86263	99.11202
379	68.94982	49.21520	88.68445	38.76832	99.13133
380	68.92449	49.15403	88.69495	38.68818	99.16080
381	68.90737	49.10197	88.71278	38.61762	99.19712
382	68.89581	49.05607	88.73554	38.55355	99.23806

383	68.88799	49.01436	88.76163	38.49389	99.28210
384	68.88271	48.97549	88.78994	38.43723	99.32819
385	68.87915	48.93855	88.81975	38.38263	99.37566

RESULTS ON RH WITH EMA AND ARIMA

The experimental results for the forecasting data using EMA and ARIMA models are shown with the degree of confidence at 80% and 90% respectively. The ARIMA model is found to be better than EMA model on the training as well on the test data. The forecasting performance of the proposed models on the relative humidity data has been summarized using the following table 6:

Temp Dataset	TS model	Accuracy measurement			
		MAE	MASE	ME	ACF1
Precipitation					
	EMA	9.51	0.92%	0.94	0.015
	ARIMA	9.14	0.89	-0.10	-0.016

Table 5: Results of ARIMA and EMA on RH

From the above table, is clear that ARIMA model with our algorithm has outperformed the EMA Model on the test data for relative humidity. In case of EMA model, further tuning of the model parameters by re-evaluating the lag values, preferably manual setting of the model parameters is required in order to enhance the prediction accuracy. PDQ values for the ARIMA were selected best in the 20 legs.

Form the comparative experiments, we have found that the ARIMA model in ensemble form is far more suitable to model the weather in terms of temperature, rainfall, precipitation and Relative humidity in univariate time series settings and gives us reliable forecasts up to 20 days in advance. Similarly the models can be trained for medium term forecasting task. The other parameters have been omitted from the report on behalf of being similar in workflow.

LONG TERM PREDICTION MODEL

The only long term prediction model reported here is MA (Q) model. This is a nonlinear model in the parameters and can depend on the various lags of the same univariate or multivariate time series data. Generally, yearlong predictions are not in the scope of our work (i.e. weather prediction for the crops), because the longest season in which a crop survive is four months, also year-wise predictions mostly fall in the area of climate research. For the sake of completeness we give MA model applied on RH data in the next section. Similar workflow can be followed in case of fitting other variables on the Moving Average model.

MOVING AVERAGE MODEL: MA (Q) ALGORITHM

For the long term prediction task, we have used multivariate moving average model, with auto selected time lag for input features which find the optimum set of hyper parameters for the model in question.

DATA PREPARATION FOR MA MODEL

The following figure shows six years relative humidity data visualization in R. From the output of the R script we can examine the patterns, in terms of their distribution, skewness, the distribution of outliers and the regularity in the humidity.

humidt.ts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	42.00	54.00	50.77	48.23	52.00	52.00	66.00	68.00	61.00	36.41	40.48	40.29
2	43.90	36.67	36.25	35.00	43.00	50.00	61.70	79.90	62.90	32.25	51.03	40.80
3	43.38	35.34	29.67	47.00	36.90	50.43	71.80	70.51	62.00	36.00	38.00	44.58
4	48.51	37.42	43.00	45.66	52.48	51.46	71.25	65.22	55.93	45.80	45.60	39.41
5	42.93	47.10	47.03	41.00	45.90	43.53	62.00	66.00	62.00	49.00	46.00	40.00
6	44.00	33.89	34.06	30.00	51.38	52.93	58.00	61.00	54.00	37.00	47.00	51.00

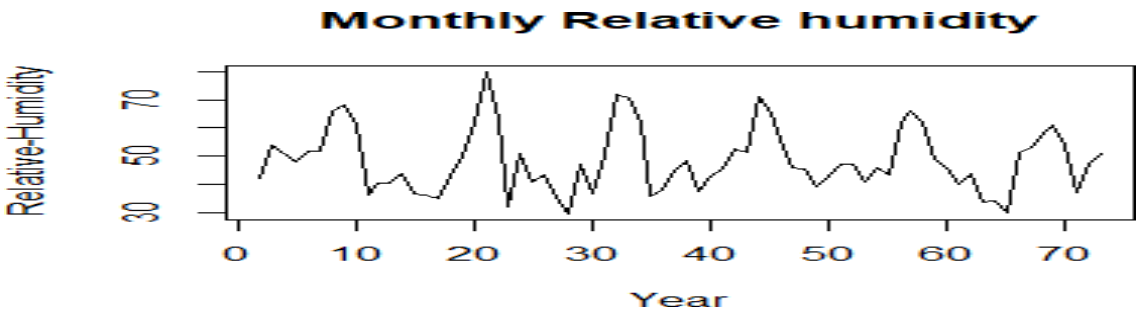

Figure 19: MA Training data set for 72 months

LAG SELECTION: ACF AND PACF

MA model with different lag values are tried, also in auto mode MA is trained in order to perform best fit. Plots of the input time series along with its ACF and its PACF is shown, lagged scatterplot of spectrum is plotted to show the effect of lag values to manage the outliers in the data.

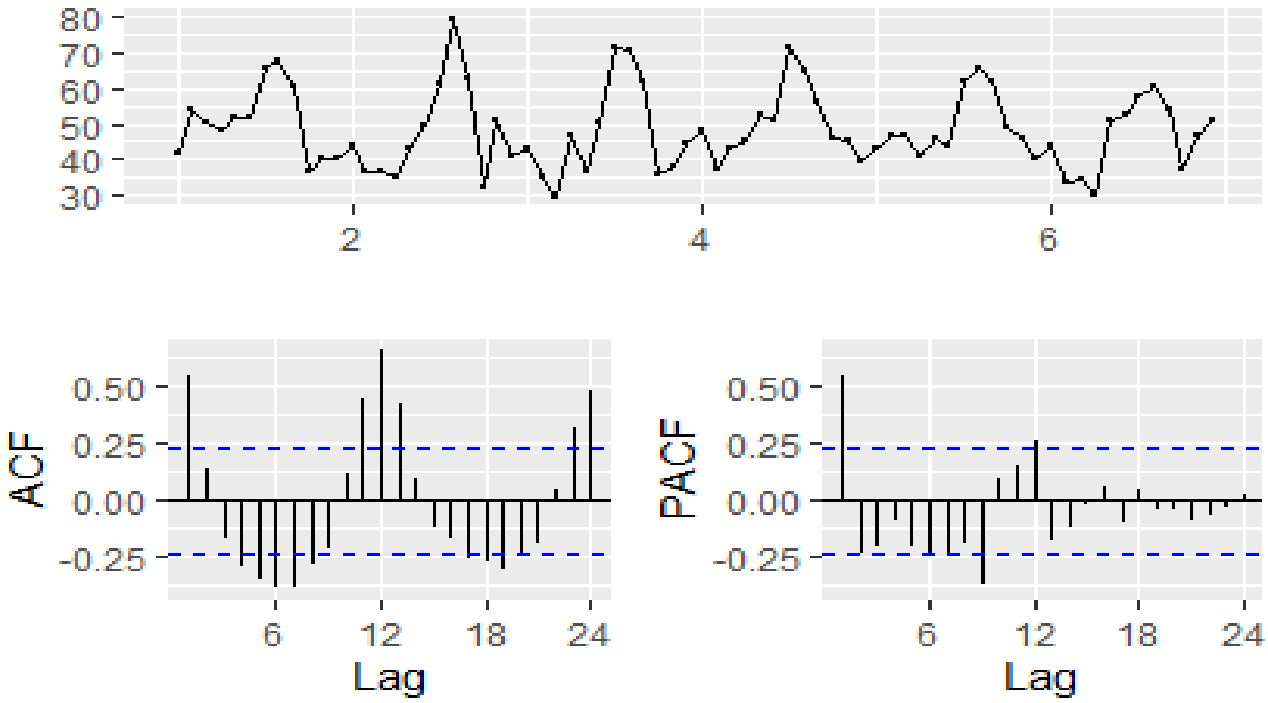

Figure 20: Data, ACF function, PACF Plot

The ACF and PACF function has been used to control the outliers in the distribution of training data. The lag values have correlation on the performance of the proposed model.

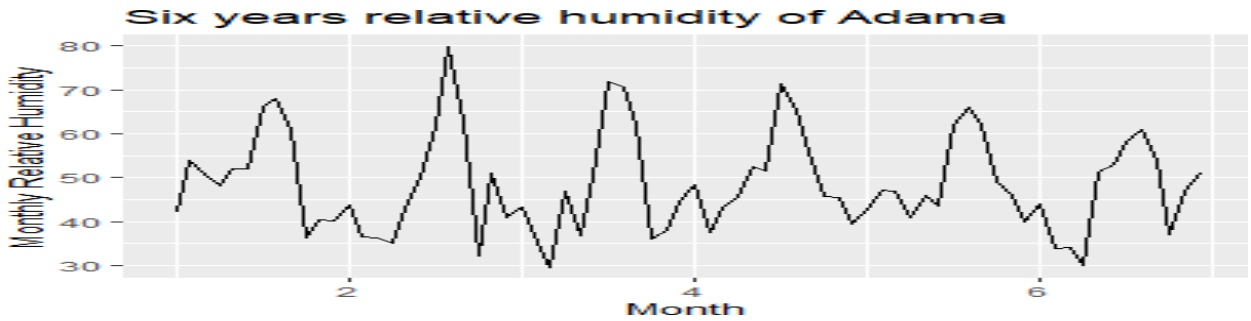

Figure 21: Yearly plot of RH

The above figure 21 shows the relative humidity of training data. We have evaluated how the models are fitted with the training data.

The first step in the classical time series decomposition is to use a moving average method to estimate the trend-cycle. The averaging operation eliminates some of the randomness in the data, leaving a smooth trend-cycle component. We call this an **MA (m)**, i.e. moving average of order (**m**).

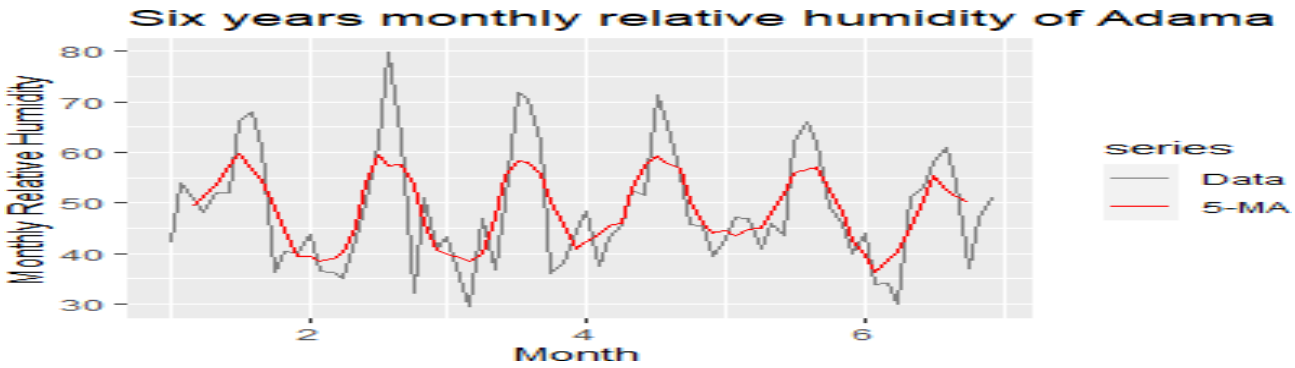

Figure 22: RH forecast with MA-5

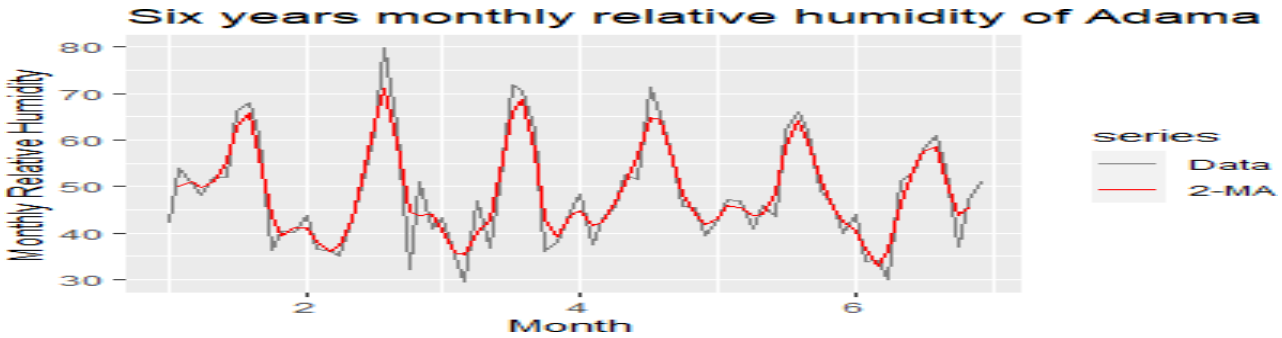

Figure 23: RH forecast with MA-2

The above figure 23 shows the experimental output of moving average model fitted with the given data. In this experiment, scaling the values of **m** will have significant impact on the model fitting. Each value in the 5-MA column is the average of the observations in the six year window centered on the corresponding year. The following figure also display the output after changing the **m** value from 2 to 3. It is possible to visualize the difference in the plot when the scale values are changed.

MULTIVARIATE TIME SERIES FORECASTING MODELS

In the previous univariate sub-section, we have seen the implementation of EMA and ARIMA time series models and their ensembles. In this subsection, we have implemented the multivariate models like Autoregressive Neural Network (modified form of ANN) and the VAR (vector auto regression) models. The multivariate models are designed to learn complex relationship among the predictor variables with the target variables. Unlike other domains, in weather prediction there is a tight coupling between the variables, as each variable can play the role of target variable, all the remaining variables being the predictor variables.

The neural network based models like Feedforward neural network model, AR-NN, RNN, CNN and LSTM etc. tries to learn the relationship among the variables into the weight vector as the parameters. VAR model learns each relation as a separate equation and the coefficients are reported as the parameters. The current research work is limited to AR-NN model and VAR, because mostly we have concentrated on autoregressive models in this research.

In the next sections, we describe the model fitting and testing experiments on each type of model on the train and test data of selected features.

AUTO-REGRESSIVE NEURAL NETWORK MODEL

AR-Neural Network model is developed for monthly relative humidity data of one year, separate model for each year data are developed and ensemble with the block bootstrapping are taken by using averaging of individual output. The model parameter are described as follows:

AR-NN

Series: humidt.ts

Model: NNAR(4,2)

Call: nnetar(y = humidt.ts, p = 4, P = 0, size = 2)

Average of 20 networks, each of which is
a 4-2-1 network with 13 weights
options were - linear output units

sigma^2 estimated as 40.68

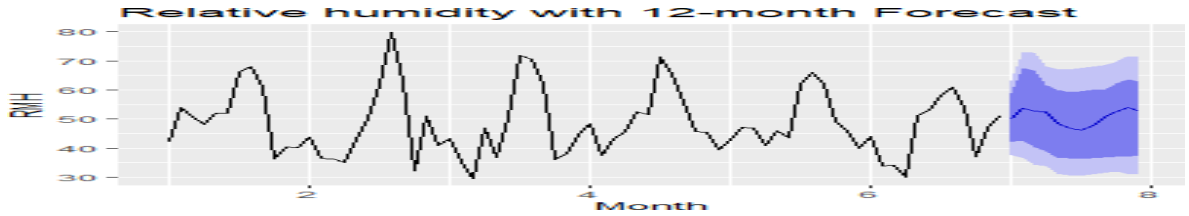

Figure 24: The 12 months forecasting output

The above figure demonstrate the neural network auto-regressive forecast experimental output with the confidence interval of 80% and 90% respectively. We can change the confidence interval to analyze the pattern.

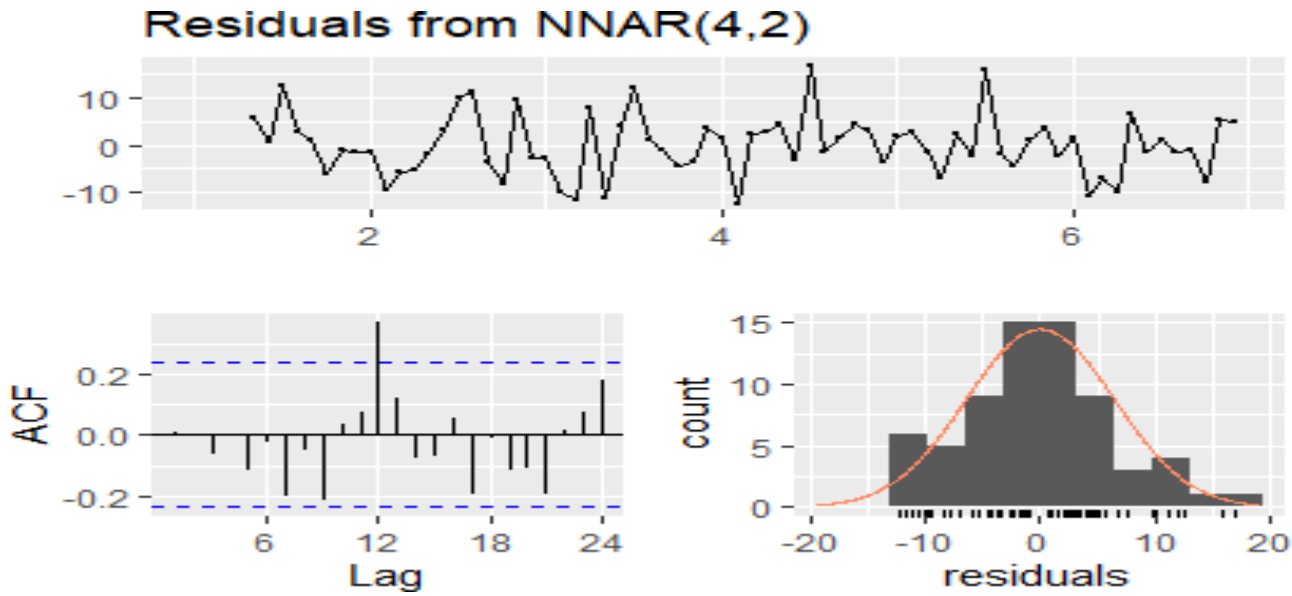

Figure 25: Residuals from NNETAR Model

The above figure shows the residual result of 12 month prediction however for a medium term prediction model, we expect 3-4 months advance prediction. The outlier observed in the dataset have been adjusted using ACF lag values. After the adjustment the final histogram of the distribution show normal statistical distribution of the residuals on this data.

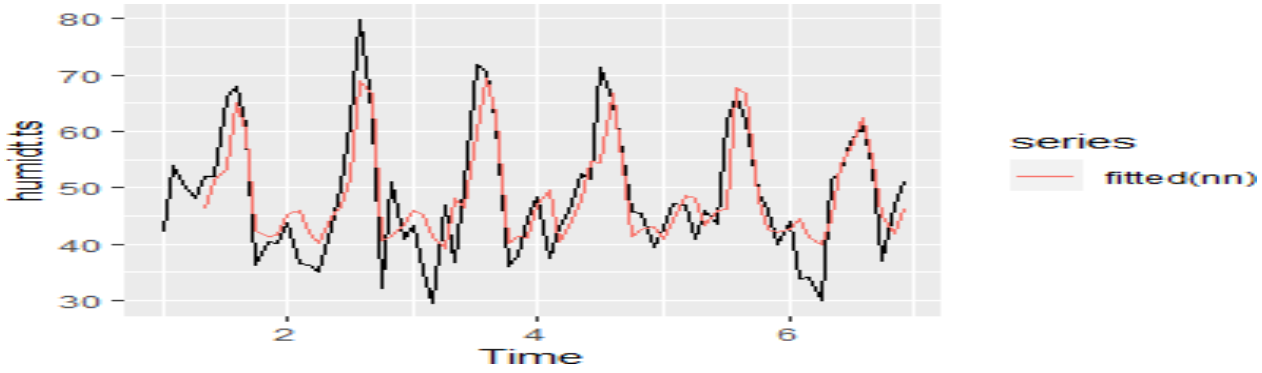

Figure 26: Fitting NNETAR model with training data

As we can see from the above figure, the model's forecasting output are almost nearly aligned with the actual training data. From this experiment, it possible to analyze the performance of the implemented model using a forecast for 3-4 months in advance. The next section describe the test time performance and the accuracy measures for AR-NN model.

Forecast method: NNAR(4,2)

Model Information:

Average of 20 networks, each of which is
a 4-2-1 network with 13 weights
options were - linear output units

Forecasts:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
50.13	53.60	52.85	52.51	48.51	46.89	46.15	47.64	50.56	52.56	54.045	52.85

Error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1

```
Training set -0.00225 6.3784    4.9908   -2.2812   11.113   0.7776    0.0058
accuracy(fc1.PI)
                ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set -0.002259348 6.378468 4.990855 -2.281263 11.11329 0.7776744 0.005877228
```

This is the experimental output of relative monthly humidity forecast of 12 months. As we can see the that the above model shows almost 94% correct predictions with a confidence of 90%, it seems to be suitable model for us, the reason behind this lack of performance is the fact that we have not included all the 31 input variables for this demo. It is expected that if remaining variables are used or some kind of feature selection mechanism is used with AR-NN model, it will be able to predict the medium term values in more precise way.

VECTOR AUTO REGRESSION MODEL

VAR model is used to learn the relationship among the variables in both the forms: as a target variable and as a member of the predictor variables. We have used up to five selected variables for the model fitting and created output equations with learned parameters.

DATA PREPARATION FOR VAR

Analyzing and Understanding the nature of our data will help us to determine the numbers of parameters, frequency of time series and the type of VAR models to be built. In this subsection, some part of the data has been plotted as follows:

VAR.temp.ts

month	Temp	Humidity	MinTemp	Pricipt
Jan 1	1 27.30	43.90	31.10	0.7
Feb 1	2 29.40	36.67	14.20	0.0
Mar 1	3 29.40	36.25	14.90	0.3
Apr 1	4 31.70	35.00	16.60	0.6
May 1	5 30.80	43.00	16.30	2.2
Jun 1	6 30.00	50.00	16.80	1.8
Jul 1	7 27.40	61.70	16.60	6.9

Aug 1	8	26.20	79.90	16.70	5.0
Sep 1	9	26.70	62.90	15.60	6.4
Oct 1	10	27.90	32.25	14.20	0.0
Nov 1	11	26.90	51.03	14.90	0.0
Dec 1	12	25.80	40.80	14.80	0.0
Jan 2	13	27.40	43.38	14.10	1.8
Feb 2	14	29.00	35.34	14.10	1.0
Mar 2	15	31.00	29.67	16.40	1.3
Apr 2	16	30.30	47.00	16.90	2.7
May 2	17	31.70	36.90	17.00	1.6
Jun 2	18	30.70	50.43	18.00	3.1
Jul 2	19	25.80	71.80	15.90	17.2
Aug 2	20	25.90	70.51	16.00	10.6
Sep 2	21	27.00	62.00	15.90	4.3
Oct 2	22	27.90	36.00	13.70	0.0
Nov 2	23	28.00	38.00	13.90	0.0
Dec 2	24	27.20	44.58	14.40	0.1
Jan 3	25	27.30	48.51	14.30	0.1
Feb 3	26	29.41	37.42	15.12	0.0
Mar 3	27	31.00	43.00	17.36	2.2
Apr 3	28	31.19	45.66	17.36	1.2
May 3	29	31.14	52.48	18.30	1.2
Jun 3	30	30.40	51.46	18.22	1.3
Jul 3	31	26.08	71.25	15.60	14.3
Aug 3	32	26.31	65.22	16.68	3.4
Sep 3	33	27.99	55.93	15.87	4.4
Oct 3	34	28.08	45.80	14.82	1.1
Nov 3	35	27.50	45.60	14.70	0.2
Dec 3	36	26.34	39.41	11.90	0.0
Jan 4	37	28.11	42.93	13.90	0.0
Feb 4	38	29.50	47.10	16.40	0.2
Mar 4	39	29.80	47.03	16.70	4.0
Apr 4	40	30.70	41.00	17.40	0.2
May 4	41	31.10	45.90	17.30	2.0
Jun 4	42	31.60	43.53	18.10	0.3
Jul 4	43	28.30	62.00	17.10	6.8
Aug 4	44	26.70	66.00	16.50	5.8
Sep 4	45	27.30	62.00	16.00	5.0

Oct 4	46	26.90	49.00	14.70	3.0
Nov 4	47	27.50	46.00	15.00	0.3
Dec 4	48	26.20	40.00	13.30	0.0
Jan 5	49	27.00	44.00	12.60	0.0
Feb 5	50	30.50	33.89	15.80	0.0
Mar 5	51	31.30	34.06	17.10	0.4
Apr 5	52	31.50	30.00	16.40	0.1
May 5	53	30.70	51.38	17.20	5.8
Jun 5	54	30.50	52.93	18.10	2.5
Jul 5	55	29.50	58.00	17.40	4.2
Aug 5	56	27.90	61.00	16.60	6.7
Sep 5	57	28.90	54.00	16.30	2.6
Oct 5	58	29.90	37.00	17.10	0.1
Nov 5	59	28.00	47.00	15.90	1.6
Dec 5	60	26.90	51.00	15.50	0.0

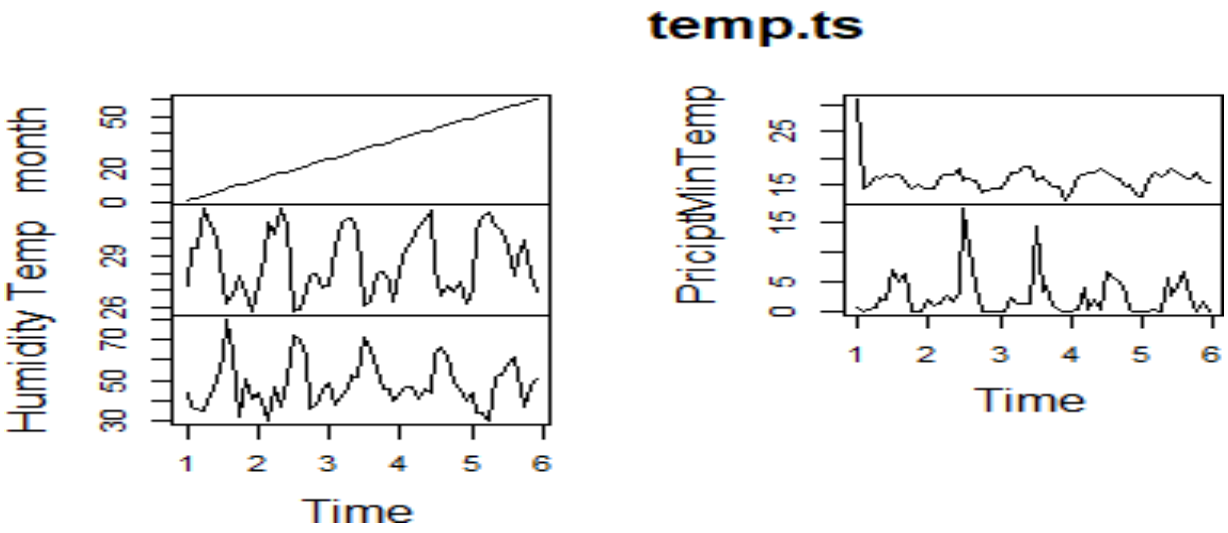

Figure 27: Data plot for features

The other important insight in the domain of machine learning was evaluating the summary of training data before proceeding to build a model. We need to answer some questions such as is the mean, STD and median are constant. In the process of building VAR based time series forecasting these values are

importance in determining the stationarity. The next sections provide the results of summary statistics and the correlation plots between the various pairs of variables.

```
summary(temp)
    month           Temp         Humidity       MinTemp         Pricipt
 Min.: 1.00   Min.   :25.80    Min.:29.67   Min.   :11.90      Min. : 0.000
 1st Qu.:15.75  1st Qu.:27.15   1st Qu.:39.85   1st Qu.:14.81 1st Qu.: 0.100
 Median :30.50  Median :28.09   Median :45.95   Median :16.15   Median:1.250
 Mean  :30.50  Mean  :28.67    Mean   :48.12   Mean   :16.13   Mean: 2.477
 3rd Qu.:45.25  3rd Qu.:30.50  3rd Qu.:53.20  3rd Qu.:17.02  3rd Qu.:3.550
 Max.  :60.00  Max.  :31.70 Max.  :79.90  Max.  :31.10  Max.:17.200
```

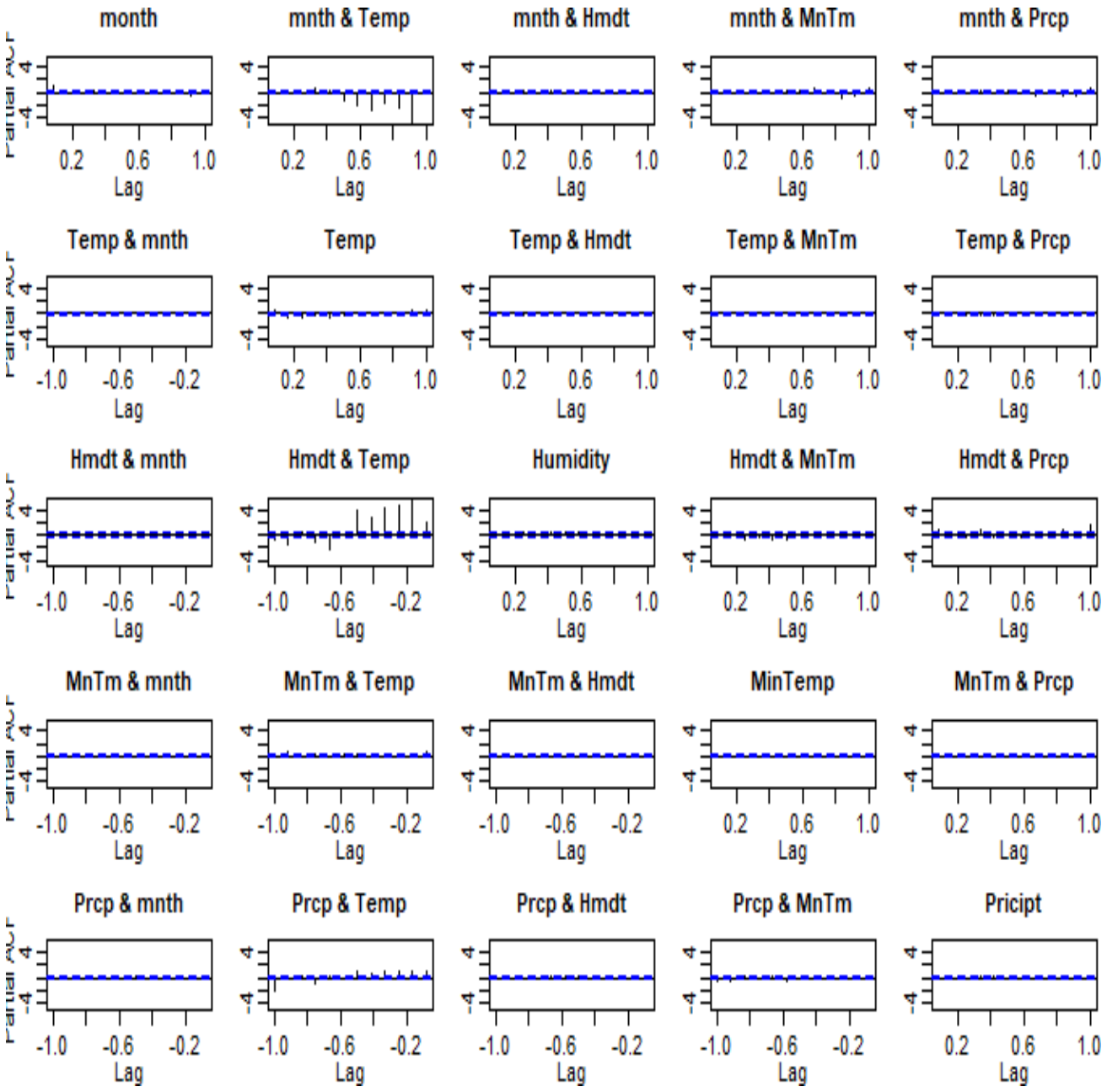

Figure 28: The correlation between different variables using the lag values

The correlation of lag values between different variable has been depicted using the above figure.

From the multivariate VAR model training step, the maximum numbers of lags used for error correction and the unit root regression type has been described as follows:

```
apply(temp.ts, 2, adfTest,
+       lags=10, #maximum number of lags used for error term correction
+       type="c", #type of unit root regression
+       title = "ADF Test for weather Data") #title of the project
$month

Title:
 ADF Test for weather Data

Test Results:
  PARAMETER:
    Lag Order: 10
  STATISTIC:
    Dickey-Fuller: -1.7321
  P VALUE:
    0.4153

Description:
 Mon Mar 23 17:28:55 2020 by user: Tagel


$Temp

Title:
 ADF Test for weather Data

Test Results:
  PARAMETER:
    Lag Order: 10
  STATISTIC:
    Dickey-Fuller: -1.215
  P VALUE:
    0.604

Description:
 Mon Mar 23 17:28:55 2020 by user: Tagel


$Humidity

Title:
 ADF Test for weather Data

Test Results:
  PARAMETER:
    Lag Order: 10
  STATISTIC:
    Dickey-Fuller: -2.6072
```

P VALUE:
0.09835

$MinTemp

Title:
ADF Test for weather Data

Test Results:
PARAMETER:
Lag Order: 10
STATISTIC:
Dickey-Fuller: -2.6998
P VALUE:
0.08416

$Pricipt

Title:
ADF Test for weather Data

Test Results:
PARAMETER:
Lag Order: 10
STATISTIC:
Dickey-Fuller: -2.831
P VALUE:
0.06406

A time series is said to be stationary if it holds the following conditions

1. The mean value of time-series is constant over time, which implies, the trend component is nullified.
2. The variance does not increase over time.
3. Seasonality effect is minimal.

This means it is devoid of trend or seasonal patterns, which makes it looks like a random *white noise* irrespective of the observed time interval. The stationarity of the dependent variables in this multivariate time series forecasting has been demonstrate using the following figure:

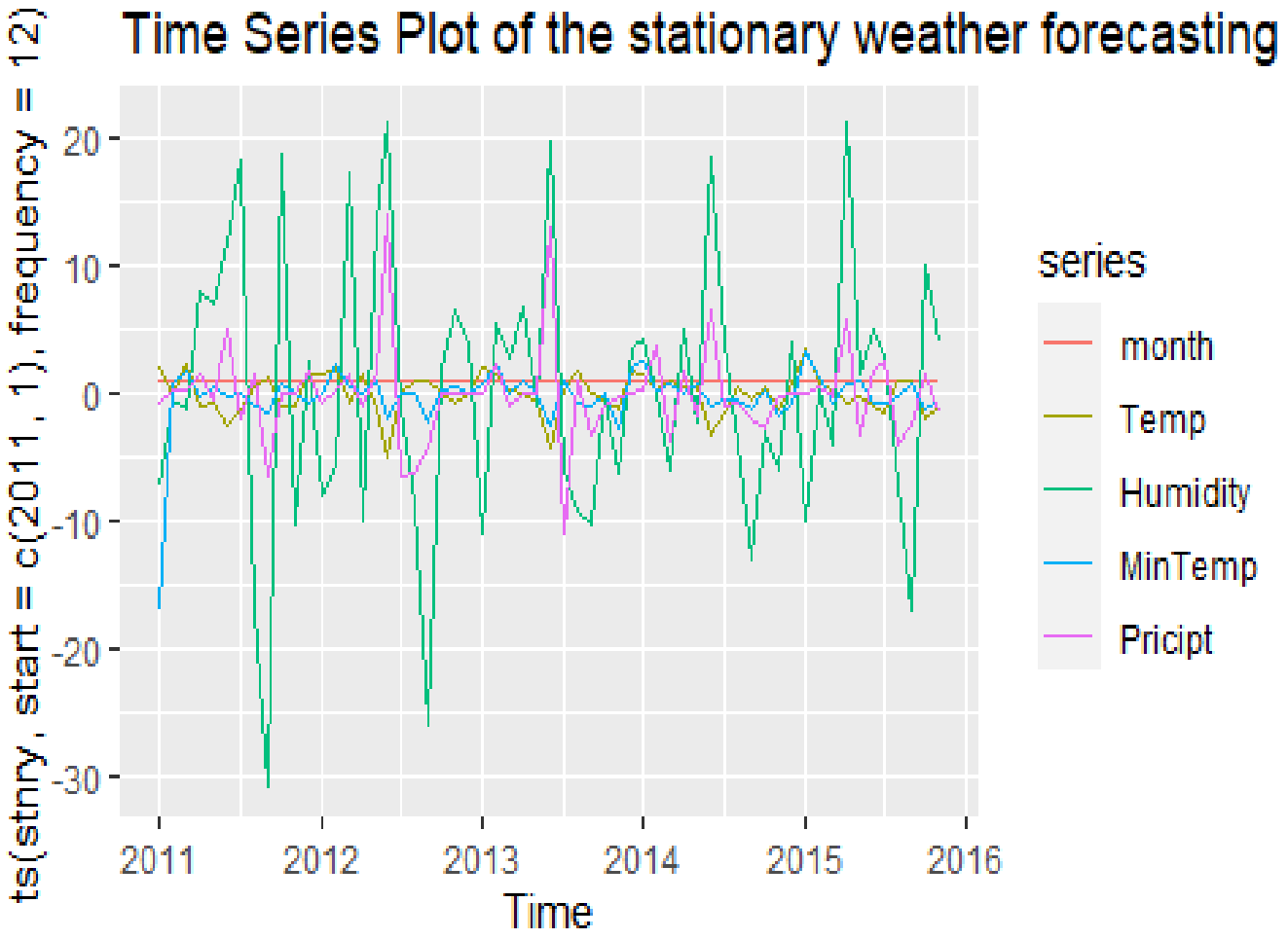

Figure 29: First order stationarity test

The multivariate weather forecast on training data has been plotted using the next figure. In this figure, we see the patterns of multiple dimensions over a period of time. Some variables shows irregularity over the given period of time. Irregularities can be removed using differencing before VAR model was fitted.

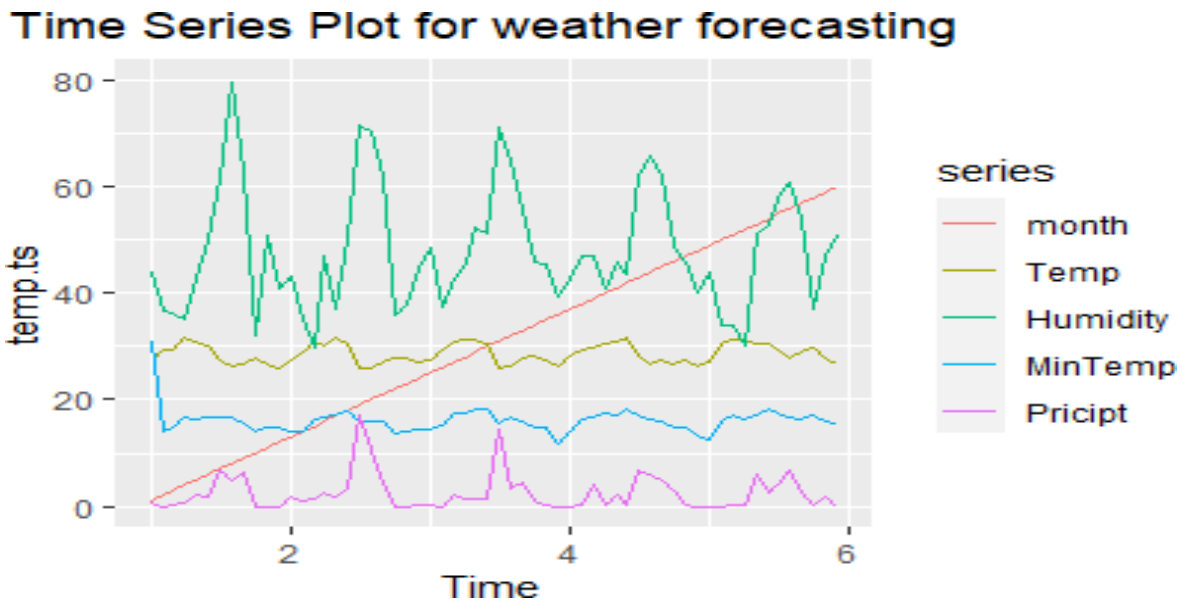

Figure 30: Individual weather parameter plot

```
VARselect(stnry,
+         type = "none", #
+         lag.max = 12) #highest lag order
temp.ts.VAR.const
$selection
AIC(n) HQ(n) SC(n) FPE(n)
    10    10     1    10
```

For vector auto-regressive multivariate time series forecasting model, the experimental results contain the values of parameters found in the equations that represents the relation of a weather variable with other variables. The estimated values of the coefficients are summarized as follows:

```
VAR Estimation Results:
========================

Estimated coefficients for equation month:
==========================================
Call:

month = month.l1 + Temp.l1 + Humidity.l1 + MinTemp.l1 + Pricipt.l1 + month.l2 + Temp.l2 + Humidity.l2 +
MinTemp.l2 + Pricipt.l2 + month.l3 + Temp.l3 + Humidity.l3 + MinTemp.l3 + Pricipt.l3 + month.l4 + Temp.l4 +
Humidity.l4 + MinTemp.l4 + Pricipt.l4 + month.l5 + Temp.l5 + Humidity.l5 + MinTemp.l5 + Pricipt.l5 + month.l6 +
Temp.l6 + Humidity.l6 + MinTemp.l6 + Pricipt.l6 + month.l7 + Temp.l7 + Humidity.l7 + MinTemp.l7 + Pricipt.l7 +
month.l8 + Temp.l8 + Humidity.l8 + MinTemp.l8 + Pricipt.l8 + month.l9 + Temp.l9 + Humidity.l9 + MinTemp.l9 +
Pricipt.l9 + month.l10 + Temp.l10 + Humidity.l10 + MinTemp.l10 + Pricipt.l10
```

```
   month.l1       Temp.l1   Humidity.l1    MinTemp.l1    Pricipt.l1      month.l2       Temp.l2
2.000000e+00 -4.235418e-15 -1.154791e-15  1.361639e-16  7.607202e-16 -1.000000e+00  6.886125e-15
 Humidity.l2    MinTemp.l2    Pricipt.l2      month.l3       Temp.l3   Humidity.l3    MinTemp.l3
2.779715e-16 -3.791376e-15  2.716856e-16            NA  2.210217e-15  3.746780e-16 -5.456157e-15
  Pricipt.l3      month.l4       Temp.l4   Humidity.l4    MinTemp.l4    Pricipt.l4      month.l5
-1.103959e-15            NA  3.031583e-15 -2.693280e-16  4.137953e-15 -3.661806e-16            NA
     Temp.l5   Humidity.l5    MinTemp.l5    Pricipt.l5      month.l6       Temp.l6 Humidity.l6
1.898572e-15  2.562196e-16 -4.393563e-15  8.152769e-17            NA  7.943025e-15  8.222059e-16
  MinTemp.l6    Pricipt.l6      month.l7       Temp.l7   Humidity.l7    MinTemp.l7    Pricipt.l7
5.338921e-16 -6.017594e-16            NA  1.986992e-15  9.264268e-16 -7.517953e-15 -5.129973e-16
   month.l8       Temp.l8   Humidity.l8    MinTemp.l8    Pricipt.l8      month.l9       Temp.l9
         NA  5.104375e-15  9.675249e-17  5.382746e-16  7.837031e-16            NA -5.026950e-16
 Humidity.l9    MinTemp.l9    Pricipt.l9     month.l10      Temp.l10 Humidity.l10  MinTemp.l10
3.531191e-16 -2.596056e-15  7.910875e-17            NA -9.072854e-15 -6.537764e-16  8.722788e-15
  Pricipt.l10
-5.817124e-16


Estimated coefficients for equation Temp:
==========================================
Call:
Temp = month.l1 + Temp.l1 + Humidity.l1 + MinTemp.l1 + Pricipt.l1 + month.l2 + Temp.l2 + Humidity.l2 +
MinTemp.l2 + Pricipt.l2 + month.l3 + Temp.l3 + Humidity.l3 + MinTemp.l3 + Pricipt.l3 + month.l4 + Temp.l4 +
Humidity.l4 + MinTemp.l4 + Pricipt.l4 + month.l5 + Temp.l5 + Humidity.l5 + MinTemp.l5 + Pricipt.l5 + month.l6 +
Temp.l6 + Humidity.l6 + MinTemp.l6 + Pricipt.l6 + month.l7 + Temp.l7 + Humidity.l7 + MinTemp.l7 + Pricipt.l7 +
month.l8 + Temp.l8 + Humidity.l8 + MinTemp.l8 + Pricipt.l8 + month.l9 + Temp.l9 + Humidity.l9 + MinTemp.l9 +
Pricipt.l9 + month.l10 + Temp.l10 + Humidity.l10 + MinTemp.l10 + Pricipt.l10

   month.l1       Temp.l1   Humidity.l1    MinTemp.l1    Pricipt.l1      month.l2       Temp.l2
51.528529307   1.088436898   0.021507302  -0.742228533  0.003446497 -51.502808008  -0.505408915
 Humidity.l2    MinTemp.l2    Pricipt.l2      month.l3       Temp.l3   Humidity.l3    MinTemp.l3
-0.031422203  -0.132819318   0.204673283            NA -0.251397662  -0.003831512  -0.181958549
  Pricipt.l3      month.l4       Temp.l4   Humidity.l4    MinTemp.l4    Pricipt.l4      month.l5
-0.263437094            NA   0.505286241   0.042320044  -0.978528078  -0.035985317            NA
     Temp.l5   Humidity.l5    MinTemp.l5    Pricipt.l5      month.l6       Temp.l6  Humidity.l6
-0.646345731  -0.057634280   0.803646349  -0.074727598            NA  0.506004523   0.027483370
  MinTemp.l6    Pricipt.l6      month.l7       Temp.l7   Humidity.l7    MinTemp.l7    Pricipt.l7
-1.301494188  -0.134907312            NA  0.046048108  -0.010945584   0.785507714   0.186359425
   month.l8       Temp.l8   Humidity.l8    MinTemp.l8    Pricipt.l8      month.l9       Temp.l9
         NA   0.233545860   0.018858105  -1.472844007  -0.116033476            NA  0.630289564
 Humidity.l9    MinTemp.l9    Pricipt.l9     month.l10      Temp.l10 Humidity.l10 MinTemp.l10
 0.073647874   0.142103249   0.059337592            NA  -0.596923840  -0.070906714  -0.252982597
 Pricipt.l10
 0.048990784


Estimated coefficients for equation Humidity:
==============================================
Call:
Humidity = month.l1 + Temp.l1 + Humidity.l1 + MinTemp.l1 + Pricipt.l1 + month.l2 + Temp.l2 + Humidity.l2 +
MinTemp.l2 + Pricipt.l2 + month.l3 + Temp.l3 + Humidity.l3 + MinTemp.l3 + Pricipt.l3 + month.l4 + Temp.l4 +
Humidity.l4 + MinTemp.l4 + Pricipt.l4 + month.l5 + Temp.l5 + Humidity.l5 + MinTemp.l5 + Pricipt.l5 + month.l6 +
Temp.l6 + Humidity.l6 + MinTemp.l6 + Pricipt.l6 + month.l7 + Temp.l7 + Humidity.l7 + MinTemp.l7 + Pricipt.l7 +
month.l8 + Temp.l8 + Humidity.l8 + MinTemp.l8 + Pricipt.l8 + month.l9 + Temp.l9 + Humidity.l9 + MinTemp.l9 +
Pricipt.l9 + month.l10 + Temp.l10 + Humidity.l10 + MinTemp.l10 + Pricipt.l10
```

```
   month.l1      Temp.l1  Humidity.l1   MinTemp.l1   Pricipt.l1      month.l2      Temp.l2
-87.32527943  -0.68002744 -0.20782478   1.68145097   0.14498131  87.14722306  -0.31426935
 Humidity.l2   MinTemp.l2   Pricipt.l2      month.l3      Temp.l3 Humidity.l3   MinTemp.l3
 -0.02189227  -1.30879100  -1.98734668           NA  7.50356484   0.15700918  -0.58990243
  Pricipt.l3     month.l4      Temp.l4  Humidity.l4   MinTemp.l4   Pricipt.l4     month.l5
  0.67432967           NA   0.17128349   0.19091390   2.97481030   0.44153939           NA
    Temp.l5  Humidity.l5   MinTemp.l5   Pricipt.l5     month.l6      Temp.l6 Humidity.l6
  6.59826006   0.63705616  -1.64239439  -0.11291223           NA  -1.94735163   0.38405111
  MinTemp.l6   Pricipt.l6     month.l7      Temp.l7 Humidity.l7 MinTemp.l7 Pricipt.l7
  4.26951523   0.33371614           NA  -1.21097032   0.20559704  -1.00231211  -0.82277852
    month.l8      Temp.l8  Humidity.l8   MinTemp.l8   Pricipt.l8      month.l9      Temp.l9
          NA  -3.36704557   0.01202990   5.68124127   0.62917960           NA -10.83431089
 Humidity.l9   MinTemp.l9   Pricipt.l9   month.l10     Temp.l10 Humidity.l10  MinTemp.l10
 -1.03289338   6.81076002   0.13084871           NA  -1.75008418  -0.06172194   1.80750773
 Pricipt.l10
  0.33895547
```

Estimated coefficients for equation MinTemp:

==========

Call:

MinTemp = month.l1 + Temp.l1 + Humidity.l1 + MinTemp.l1 + Pricipt.l1 + month.l2 + Temp.l2 + Humidity.l2 + MinTemp.l2 + Pricipt.l2 + month.l3 + Temp.l3 + Humidity.l3 + MinTemp.l3 + Pricipt.l3 + month.l4 + Temp.l4 + Humidity.l4 + MinTemp.l4 + Pricipt.l4 + month.l5 + Temp.l5 + Humidity.l5 + MinTemp.l5 + Pricipt.l5 + month.l6 + Temp.l6 + Humidity.l6 + MinTemp.l6 + Pricipt.l6 + month.l7 + Temp.l7 + Humidity.l7 + MinTemp.l7 + Pricipt.l7 + month.l8 + Temp.l8 + Humidity.l8 + MinTemp.l8 + Pricipt.l8 + month.l9 + Temp.l9 + Humidity.l9 + MinTemp.l9 + Pricipt.l9 + month.l10 + Temp.l10 + Humidity.l10 + MinTemp.l10 + Pricipt.l10

```
month.l1      Temp.l1  Humidity.l1   MinTemp.l1   Pricipt.l1      month.l2      Temp.l2
-10.93075022   1.14487100   0.02408142  -0.71921162  0.06083166  10.91404987  -0.07483811
 Humidity.l2   MinTemp.l2   Pricipt.l2      month.l3      Temp.l3 Humidity.l3   MinTemp.l3
 -0.04834853  -0.68407197   0.07577221           NA  0.38090029   0.01889276  -0.28118488
  Pricipt.l3     month.l4      Temp.l4  Humidity.l4   MinTemp.l4   Pricipt.l4     month.l5
 -0.25810167           NA   0.69126278   0.08240488  -1.09278189  -0.16799958           NA
    Temp.l5  Humidity.l5   MinTemp.l5   Pricipt.l5     month.l6      Temp.l6 Humidity.l6
  0.28950765   0.03196802   0.45792957  -0.17036995           NA  0.61998922  0.04770999
  MinTemp.l6   Pricipt.l6     month.l7      Temp.l7  Humidity.l7  MinTemp.l7   Pricipt.l7
 -1.01531613  -0.12363055           NA  0.28582034   0.01148636   0.46774515   0.09149170
    month.l8      Temp.l8  Humidity.l8   MinTemp.l8   Pricipt.l8      month.l9      Temp.l9
          NA  -0.12979412  -0.02846814  -0.77945849  -0.01619911           NA  0.13013163
 Humidity.l9   MinTemp.l9   Pricipt.l9   month.l10     Temp.l10 Humidity.l10  MinTemp.l10
 -0.01083224   0.31368339   0.03055463           NA  -0.53119090  -0.05472270  -0.17297305
 Pricipt.l10
  0.06922467
```

Estimated coefficients for equation **Pricipt**:

==========

Call:

Pricipt = month.l1 + Temp.l1 + Humidity.l1 + MinTemp.l1 + Pricipt.l1 + month.l2 + Temp.l2 + Humidity.l2 + MinTemp.l2 + Pricipt.l2 + month.l3 + Temp.l3 + Humidity.l3 + MinTemp.l3 + Pricipt.l3 + month.l4 + Temp.l4 + Humidity.l4 + MinTemp.l4 + Pricipt.l4 + month.l5 + Temp.l5 + Humidity.l5 + MinTemp.l5 + Pricipt.l5 + month.l6 + Temp.l6 + Humidity.l6 + MinTemp.l6 + Pricipt.l6 + month.l7 + Temp.l7 + Humidity.l7 + MinTemp.l7 + Pricipt.l7 +

```
month.l8 + Temp.l8 + Humidity.l8 + MinTemp.l8 + Pricipt.l8 + month.l9 + Temp.l9 + Humidity.l9 + MinTemp.l9 +
Pricipt.l9 + month.l10 + Temp.l10 + Humidity.l10 + MinTemp.l10 + Pricipt.l10

month.l1       Temp.l1   Humidity.l1    MinTemp.l1    Pricipt.l1      month.l2       Temp.l2
 2.555200e+02 -7.699548e-02  2.035872e-02 -5.103131e-01 -5.289365e-01 -2.555450e+02 -4.369217e-01
Humidity.l2    MinTemp.l2    Pricipt.l2      month.l3       Temp.l3   Humidity.l3    MinTemp.l3
 8.135482e-02  9.442635e-01 -8.026395e-01            NA  1.699409e+00  1.013032e-01 -1.318602e+00
Pricipt.l3      month.l4       Temp.l4   Humidity.l4    MinTemp.l4    Pricipt.l4      month.l5
-1.453809e-01            NA -1.576706e+00 -4.886199e-02  3.424668e+00  6.115144e-02            NA
     Temp.l5   Humidity.l5    MinTemp.l5    Pricipt.l5      month.l6       Temp.l6 Humidity.l6
 4.664979e-02  1.190950e-01 -2.095698e+00 -5.589419e-01            NA -5.626578e-01  8.865048e-02
   MinTemp.l6    Pricipt.l6      month.l7       Temp.l7   Humidity.l7    MinTemp.l7    Pricipt.l7
 3.018179e+00  2.030535e-01            NA -3.477876e+00 -2.406524e-01  6.883444e-01 -2.925892e-01
month.l8       Temp.l8   Humidity.l8    MinTemp.l8    Pricipt.l8      month.l9       Temp.l9
         NA -2.977688e+00 -2.960382e-01  3.978875e+00 -2.314567e-03            NA -4.734016e+00
  Humidity.l9    MinTemp.l9    Pricipt.l9     month.l10      Temp.l10  Humidity.l10  MinTemp.l10
-4.868060e-01  2.332818e+00  7.564954e-02            NA -1.120486e+00 -1.257929e-01  2.951081e-01
  Pricipt.l10
-4.399360e-01
```

Finally, the performance of the model in prediction of target (first) variable in terms of the given variables over the test data is reported as follow:

```
$criteria
         1           2           3           4           5           6
AIC(n) -5.882271e+01 -5.759853e+01 -5.839147e+01 -5.822621e+01 -5.856815e+01 -5.811714e+01
HQ(n) -5.838584e+01 -5.679761e+01 -5.722649e+01 -5.669718e+01 -5.667507e+01 -5.586001e+01
SC(n) -5.767550e+01 -5.549531e+01 -5.533223e+01 -5.421096e+01 -5.359689e+01 -5.218987e+01
FPE(n) 2.858546e-26 1.002657e-25 4.914623e-26  6.806817e-26 6.439564e-26  1.644660e-25
              7           8           9          10
AIC(n) -5.801974e+01 -5.832954e+01 -5.980656e+01 -6.507548e+01
HQ(n) -5.539854e+01 -5.534429e+01 -5.645725e+01 -6.136213e+01
SC(n) -5.113646e+01 -5.049025e+01 -5.101125e+01 -5.532417e+01
FPE(n)  4.142665e-25  1.350771e-24  8.548485e-24 -2.137372e-23
```

We can quantify the effect of selected variables and the lag on the predicted variables in terms of AIC value. We select the model with the higher AIC value. The AIC value is changed if the number of predictor variables and the lag is changed.

The model here obtained has given us good fit on lag of 10 for the five selected variables, while in the actual implementation, we have a provision to increase the number of variables in order to include the all effects into the equation. This type of models are open to overfitting, therefore, we prefer ANN based models for multivariate weather prediction problem as compared to the VAR model.

PREPARATION OF AGRO-ADVISORIES FOR CROPS

An Agro-advisory is a report which can be easily understood by the farmers, it should be printed in the local language and contain information regarding important weather variables for next 7-20 days or a full crop season. The Agricultural meteorology centers of Ethiopia in association with Agricultural research institutes can publish the results of our models in the form of Agro-advisories and issue the normal and warning signs upto 20 days with almost 94% of accuracy in case of multivariate models and 98% accuracy for short term predictions using ensembles of univariate models. This research has developed a script for collecting the model outputs for 7 days and make a tabular representation of various weather parameters. In the each row, one crop is listed in the table. The weather variable under question is represented in a cell of the table, with three level color coding for easy understanding, where the range of predicted variable is assigned orange color for the values below expectation, green color for the values in the favorable range of the crop and red color in case if the value of the variable is above the expected value for that crop. Color intensity of the particular cell works as alarm for the farmers to take necessary actions. We have collected the questionnaire from the field workers in agricultural meteorology department regarding the favorable values of the parameters under study for different crops in Adama region. However, the translation of the agro-advisories in local languages has been done by meteorology center in offline mode and this service is on-demand basis.

CHAPTER 6

DISCUSSIONS

From the experimental results over the proposed models in short term and medium term prediction settings, it is clear that EMA model gives a comparable performance to ARIMA based model for short term prediction. From the results of short term prediction models it is understood that the univariate ensemble models are capable to learn the past behavior of the weather conditions individually in the form of a time series given that they are fitted with proper values of hyper parameters and the data provided to them follow the stationarity of the first and second order.

Exponential smoothing model is good for short term forecasting. It is easy to learn as it need three two to five observations from the past which significantly decide the predicted values. In this research, EMA when applied on short term prediction models gives a comparable performance to ARIMA on same training set (as well as on the test set). Upon testing it has been found that forecasting results of EMA model are influenced by the most recent data points in the temporal dimension, as we move away from the current time, the difference between the predicted value and the observed value is increasing, which is expected from the EMA model. The disadvantage of EMA model can be seen in Medium Term and Long term prediction models where the "smoothing" effect completely flattened the large variations in the values, especially those which were far behind the current prediction point. It means EMA model is unable to learn the variations in the weather parameters beyond its smoothing limits. The variables like wind speed, rainfall and humidity may have larger variations as compared to the temperature and pressure which remains almost stable in medium term. Therefore, EMA model is expected to under-perform in medium term forecasts (3 to 5 months) in case of high variance parameters, which is indeed a case as we have seen in the experiment section.

ARIMA model is found to be superior than the EMA model in short term as well as in Medium term settings. This behavior can be explained in terms of the parameters of the ARIMA model which have been learned in autofit mode. In R when this mode is instructed, the parameters P, D, Q of the model are found using AIC criteria. While EMA model is fitted with manually selected parameter values, on the other hand ARIMA model is based on the optimal parameters. However the ARIMA models are also rigid for outlier data points as they lie outside the domain of learned model. ARIMA model is not suitable for those weather parameters which shows high variance and can have extreme values. This behavior is seen in case of medium term prediction model where moving average component of the ARIMA model tries to fix the average values of predicted variable, but the peaks found in the regions outside the MA parameter are simply clipped by the model during prediction. This property is not suitable for high variance weather parameters like rainfall, because if the model is unable to capture the trends and extreme events properly, the forecast of the model may not be useful for many stakeholders including farmers. These limitations have inspired many researchers to develop extensions to basic ARIMA model, which may be our focus area in the future.

VAR model is complex set of equations, since in this research we have taken only limited number of weather parameters, it was easy to represent the equations and finding the values for each set of coefficients. But in case of production grade multivariate forecasting model, there may be 2^{31} possibilities of combining variables, which will be difficult to account for. In this case solving such a large number of equations is also difficult task.

Most promising models for the forecasting of weather events are based on neural networks, AR-NN is composed of only two parameters, i.e. the number of lags and the number of hidden neurons. Here we have developed single hidden layer network with four variables as input and one variable at output. As per the results, 94% is the accuracy of AR-NN model for the multivariate short term prediction. We

have found that the NN model is free from the restrictions of stationarity and can learn nonlinear relation among the input and output variables. But simple Neural Network models are unable to capture periodicity in the input, as well as they consider the output at two successive time instants are independent. This can be seen in very structure of the AR-NN which only takes into account the lagged values of the input variables, while predicting the current value of output variable. But in weather prediction, surely it is a limitation, because two successive outputs of the series data are not independent. Therefore, we believe that extension of this research into deep learning based neural network models like RNN, CNN and LSTM will be able to solve this problem.

The models proposed in this research are suitable for Adama meteorology center as they are basic models and easy to execute and gives comparable results to the traditional models in short term. In fact, it was the first research up to our knowledge which have used auto-regressive models for prediction of TEMP, PRECIP, RH etc. in Adama region.

This research addressed the question of whether the proposed models like EMA, ARIMA, VAR and AR-NN will be able to significantly improve the accuracy of prediction for short, medium and long term prediction task. In this case, we have seen that most of the models are able to prove their importance for the short term weather prediction task. This property can be extended for medium term prediction tasks for relatively stable weather variables. However, if the weather parameter changes abruptly, or data has outliers, which can happen in the form of occasional rain in the non-rainy seasons, the models we proposed are unable to capture this event.

The second research question is to find out the best model for the weather prediction task, as we have seen that ARIMA is best for short term prediction task. We have omitted the experiments of relatively weaker models like MA and Linear Regression in favor of EMA on short term prediction task, as these models have not shown a comparable performance. As the prediction window and lag increases, it is

expected to use improved versions of ARIMA with the ability to handle periodicity, trends, and the outliers. The extension of ARIMA models shall improve the prediction of models in medium and long term cases.

Similarly, AR-NN is good for short term prediction as long as the complexity of network is limited by the number of lags, but it is difficult to ensemble if the lag value for learning the parameters is very high. For medium term weather prediction, we have started to explore the deep learning based neural networks. The cascaded and stacked architectures of LSTM and CNN models are our future work.

Third research question we address by providing a script to convert the output to an Agro-advisory with color coding of the events. However, upon field survey we have found that farmers are more likely to use the Agro-advisories, if they are in local languages.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

This research was conducted to address the problems of the meteorology centers in Ethiopia, and to provide alternative models to the currently available traditional NWP models. The work done in this research has been able to identify the main problems of traditional models, we have tried our best to survey the state of machine learning applications in context of weather prediction task in Ethiopia. We found that very limited practical work has been done in this area. This research has identified the important set of models like autoregressive models, moving average, exponential smoothing, ARIMA, VAR and autoregressive neural network model for modeling the weather parameters in short, medium and long term time span. The important weather parameters selected for the modeling are: TEMP, PRECIP, RHUM, RAINFALL, SUNHRS. Parameters were selected based on the requirement of the farmers and stakeholders in the agriculture field. Most of the models reported in this report are developed in R programming language with manual and automatic tuning of hyper-parameters. The ensemble of model output is done using averaging method and usability of our algorithm is empirically established. The ensemble setting of the proposed model is able to handle variability in the input in more stable way as compared to the individual models. However, exhaustive comparison of the performance of proposed ensemble models with the traditional models is not done, but we have obtained acceptable level of accuracy in most of the short term prediction models.

Currently, EMC is using WRF model, this model is also used by the Adama meteorology center to predict only temperature variable. Our initiative in this regards have immense impact on the predictive capabilities of the meteorology centers especially in the rural areas where there is a problem of electricity, internet and computational power is limited to the desktop computers. The models proposed in this research do not require HPC infrastructure to train and test. However, our model require some degree of manual data pre-processing. The models and the algorithm developed is found to be efficient in learning the weather prediction task. The scripts of these models have been provided to EMC center and integration of this workflow in their prediction task is expected for the purpose of internal assessment and development of Agro-advisories.

The performance of the models proposed is comparable to the existing WRF model and other prediction tools used by the EMC professionals. One of such tool regularly being used to predict day's ahead temperature was leap software which is based on the moving average model, but since we have found this model weak for short term predictions, we have dropped it from analysis, we have used better time series models like EMA and ARIMA.

This research has tried to address most of the research questions and objectives of the research proposal. In the next section we give an idea of the future work which our team shall be engaged as an extension to the work done in this research.

FUTURE WORK

The current research have started from the study of traditional finite element models and implemented these models in our SIG along-with Torque cluster management software. But using NWP models to compute meaningful results on selected variables we required up to 48 hour computation on 30 computers connected with the cluster. But this was not possible due to power fluctuation and breaking of Torque due to limited RAM in the PC. Therefore, the results were collected from IGSSA, Addis

Ababa, HPC Lab. In the future, we plan to set up HPC infrastructure in ASTU so this experiment when extended with the newer machine learning models, can be compared with the traditional models in different runs.

Another important feedback we have received from meteorology professionals is to provide deep learning based sophisticated pre-trained models to those centers which already have HPC installation. We are committed to design and develop deep learning based advanced models and variants of ARIMA model for the purpose of weather prediction task.

Another direction is to implement distributed machine learning algorithms for handling the problem of the data collection and preparation, this can be done if we provide a Hadoop based pipeline for data processing at the remote server, where observations can be directly saved in the distributed file system as well machine learning algorithms can be parallelized.

RECOMMENDATIONS

This research was focused on the design and development of the machine learning models for weather prediction task in Adama region. The important variables of interest were selected based on the need of farmers and agriculture sector in general. As we went through the literature of the field and visited the several meteorology centers in the region, we have identified that there was a significant gap between the current state of the art and the weather prediction capabilities in case of rural weather stations. The models selected in this research for the implementation are consistent with the current state of the meteorology centers. However, more advanced models with few percent gain in the accuracy are possible at the cost of hiring expert machine learner and data science professionals in the meteorology department.

Application of new machine learning techniques in weather prediction task is relatively new field in context of Ethiopia. There exists a number of opportunities for the researchers who want to work in this field. Our research have tried to identify the gaps and opportunities for the researchers, throughout the design and development work. We strongly recommend following points for the future researchers and meteorology professionals:

1. There are several issues in the data collection process at the rural meteorology centers, such as manual coding of the data is done which is error prone, we recommended the officials that since the data is coming from analog and digital sensors it is possible to directly read the data into the computer, rather than manually visiting every sensor led screen.
2. The data collected across the days, months and years should be complete in the sense that if the rural station itself provide the solution to the missing observations, it will be more effective than the researchers using some averaging techniques or removing the records. Sometimes we have observed a number of records has to be removed in a row.
3. We recommend the routine training of meteorology professionals should be performed on the machine learning and data science frameworks as they are becoming more and more important for understanding the data and finding the trends in the data for every filed of science and technology.
4. We recommend ASTU to provide HPC infrastructure through ICT center in which interested researchers are provided with their individual accounts. In each account there should be minimal set of the resources like processor cores, memory, disk space and necessary permission to install libraries and required programming environments should be provided. In addition to this, a professional set up is expected in which a team of administrators is available to help users to set up the connection to main HPC server.

Acknowledgments

Our team sincerely acknowledges ASTU for providing this opportunity to undertake this research and supporting the research through various facilities. We also acknowledge the professionals from Adama meteorology center for their multiple visits in the research lab, participating in the design and development process, also helping us in understanding the details of local issues. The support provided by Addis Ababa meteorology center and IGSSA professionals with an excellent workshop is also appreciated.

REFERENCES

1. David J. Stensrud, “Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models”-CUP (2009)
2. Thomas Tomkins Warner, “Numerical Weather and Climate Prediction”, Cambridge University Press (2011)
3. Austin Woods-Medium-Range, “Weather Prediction: The European Approach”, Springer (2005)
4. Jean Coiffier, ”Fundamentals of Numerical Weather Prediction”, CUP (2012)
5. G Marchuk., ”Numerical methods in weather prediction” Elsevier Publications2012.
6. Ling Chen, Xu Lai., “Comparison between ARIMA and ANN models used in short-term wind speed forecasting”, Power and Energy Engineering Conference (APPEEC), 2011 Asia-Pacific. IEEE, 1--4.
7. Felix A Gers, Douglas Eck, Jürgen Schmidhuber, “Applying LSTM to time series predictable through time-window approaches”, Neural Nets WIRN Vietri-01. Springer, 193—2011.
8. Tilmann Gneiting, Adrian E Raftery, “Weather forecasting with ensemble methods”, Science, Vol. 310, 5746 (2005), 248—249
9. Emilcy Hernández, Victor Sanchez-Anguix, Vicente Julian, Javier Palanca, Néstor Duque, “Rainfall prediction: A deep learning approach”, International Conference on Hybrid Artificial Intelligence Systems, Springer, 151--162. 2016.
10. Allan H Murphy, “What is a good forecast? An essay on the nature of goodness in weather forecasting”, Weather and Forecasting, Vol. 8, 2 (1993), 281--293.
11. Lewis Fry Richardson, “Weather prediction by numerical process”, Cambridge University Press, 2005
12. Navin Sharma, Pranshu Sharma, David Irwin, Prashant Shenoy, “Predicting solar generation from weather forecasts using machine learning”, International Conference on Smart Grid Communications, IEEE, 528--533.
13. Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, Wang-chun Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”, Neural Information Processing Systems, 802—810-2015.
14. MA Tolstykh, AV Frolov, “Some current problems in numerical weather prediction”, Izvestiya Atmospheric and Oceanic Physics, Vol. 41, 3 (2005), 285--295.
15. Cyril Voyant, Marc Muselli, Christophe Paoli, Marie-Laure Nivet, “Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation”, Energy, Vol. 39, 1 (2012), 341--355.

16. US EPA, “Climate Change Indicators: U.S. and Global Temperature”, US EPA, 27-Jun-2016. [Online]. Available at: https://www.epa.gov/climate-indicators/climate-change-indicators-us-and-global-temperature. [Accessed: 18-Sep-2018].
17. Jonathan D. Cryer, Kung-Sik Chan, “Time Series Analysis With Applications in R“ , Springer Texts in Statistics, Springer (2008)
18. Cyril Voyant, Marc Muselli, Christophe Paoli, Marie-Laure Nivet, “Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation”, Energy 39 (2012) 341-355
19. Aditya Grover, Ashish Kapoor, Eric Horvitz, “A deep hybrid model for weather forecasting”, International Conference on Knowledge Discovery and Data Mining, 2015.
20. Priya Narayanan, Ashoke Basistha, Sumana Sarkar, Kamna Sachdeva, “Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India”, C. R. Geoscience 345 (2013) 22–27
21. Datta A., Si S., Biswas S., “Complete Statistical Analysis to Weather Forecasting” , Computational Intelligence in Pattern Recognition, Advances in Intelligent Systems and Computing, vol 999. Springer, Singapore, 2020
22. David John Gagne, Hannah M. Christensen, “Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz ’96 Model”, Journal of Advances in Modeling Earth Systems (JAMES)-2019
23. K. Puri, G. Dietachmayer, et. al., “Implementation of the initial ACCESS numerical weather prediction system”, Australian Meteorological and Oceanographic Journal 63 (2013) 265–284
24. Shao, M., Smith, W. L., “Impact of atmospheric retrievals on Hurricane Florence/Michael forecasts in a regional NWP model”. Journal of Geophysical Research: Atmosphere (2019), 124, 8544-8562. .
25. Weyn, J. A., Durran, D. R., Caruana, R., “Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data”, Journal of Advances in Modeling Earth Systems (2019), 11, 2680–2693. https://doi.org/10.1029/2019MS001705-(2019).
26. Guerra, J. A., Murray, S. A., Doornbos, E., “The use of ensembles in Space Weather Forecasting”, Space Weather, 18, https://doi.org/ 10.1029/2020SW002443-(2020).
27. Nikita Shivhare et. al., “ARIMA based daily weather forecasting tool: A case study for Varanasi”, MAUSAM, (January 2019), vol 70-1 , pp133-140
28. Ricardo Aguasca-Colomo, Dagoberto Castellanos Nieves et. al, “ Comparative Analysis of Rainfall Prediction Models Using Machine Learning in Islands with Complex Orography: Tenerife Island”, Appl. Sci. 2019, 9, *4931; doi:10.3390/app9224931*

29. Daniel Eni, Fola J. Adeyeye, "Seasonal ARIMA Modeling and Forecasting of Rainfall in Warri Town, Nigeria, Journal of Geoscience and Environment Protection", 2015, 3, 91-98 Published Online August 2015 in SciRes. http://www.scirp.org/journal/gep http://dx.doi.org/10.4236/gep.2015.36015
30. Degefu, M.A., Rowell, D.P., Bewket, W., "Teleconnections between Ethiopian rainfall variability and global SSTs: observations and methods for model evaluation". Meteorol Atmos Phys **129,** 173–186 (2017). https://doi.org/10.1007/s00703-016-0466-9
31. Simane B, Beyene H, Deressa W, Kumie A, Berhane K, Samet J., "Review of Climate Change and Health in Ethiopia: Status and Gap Analysis", Ethiop J Health Dev. 2016;30(1 Spec Iss):28–41.
32. Declan Conway, "Over one Century of Rainfall and Temperature Observations in Addis Ababa, Ethiopia", International Journal of Climatology 24 (1):77 – 91, January 2004
33. Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", December 2015, Pp 802–810 NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1,December 2015, Pp 802–810
34. Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, A K Upadhyay, "Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data", International Journal of Intelligent Systems and Applications (2018), Vol.10, No.1, pp.16-23, DOI: 10.5815/ijisa.2018.01.0
35. M.G. Schultz, C. Betancourt, B.Gong et.al. "Can deep learning beat numerical weather prediction?", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol 379,(2194), 2021
36. Wei Li, Amin Kiaghadi, Clint Dawson, "High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks", Neural Computing and Applications (2021) 33:1261–1278
37. W. Adefires, "Management of Dry Forests for Social-ecological Resilience of the Pastoral and Agro-pastoral Communities in the Dry Zone of Ethiopia. Dissertation:" Technology University of Dresden, Faculty of Environmental Sciences, Germany.
38. Abera, Z. Mohammed, and M. Bekele, "Local People Perception on the Role of Area Exclosure in the Central Rift Valley of Ethiopia : a Case Study at," Int. J. Sci. Res. Publ. (2016), vol. 6, no. 10, pp. 583–594,.
39. Martha Kidemu, Martha Gebreyesus, "Traditional Ecological Knowledge for Climate Change Assessment and Rainfall Prediction: A Case of Adami Tulu Jido Kombolcha District, Oromia Region, Ethiopia", International Journal of Natural Resource Ecology and Management, 2020; 5(2): 43-48
40. http://www.ethiometmaprooms.gov.et/about/facilities, Accessed on June 30, 2021
41. Gabriele Gramelsberger, "Conceiving Meteorology as the exact science of theatmosphere: Vilhelm Bjerknes ´s paper of 1904 as a milestone", Meteorol. Z.,18, 2009

42. Monmonier, Mark, “Telegraphy, Iconography, and the Weather Map: Cartographic Weather Reports by the United States Weather Bureau, 1870-1935.”, Imago Mundi, vol. 40, 1988, pp. 15–31. JSTOR, www.jstor.org/stable/1151009. Accessed 30 June 2021.
43. Hunt, J.C.R, "Lewis Fry Richardson and His Contributions to Mathematics, Meteorology, and Models of Conflict" (PDF). Annual Review of Fluid Mechanics. doi:10.1146/annurev.fluid.30.1.0. 2008.