PREDICTION OF SEA SURFACE TEMPERATURE USING MACHINE LEARNING TECHNIQUES

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

In

Electronics and Communication Engineering

By

Geetali Saha

[129990911006]

under supervision of

Dr. Narendrasinh C. Chauhan

GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD

April – 2021

Annexure-I

PREDICTION OF SEA SURFACE TEMPERATURE USING MACHINE LEARNING TECHNIQUES

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

In

Electronics and Communication Engineering

By

Geetali Saha

[129990911006]

under supervision of

Dr. Narendrasinh C Chauhan

GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD

April – 2021

Annexure-III

DECLARATION

I declare that the thesis entitled **PREDICTION OF SEA SURFACE TEMPERATURE USING MACHINE LEARNING TECHNIQUES** submitted by me for the degree of Doctor of Philosophy is the record of research work carried out by me during the period from **2012** to **2019** under the supervision of **Dr. N. C. Chauhan** and this has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities, if noticed in the thesis.

Signature of the Research Scholar:  Date: **05/03/2021**

Name of Research Scholar: Geetali Saha

Place: **Anand, Gujarat**

Annexure-IV

CERTIFICATE

I certify that the work incorporated in the thesis **Prediction of Sea Surface Temperature using Machine Learning Techniques** submitted by Smt. **Geetali Saha** was carried out by the candidate under my supervision/guidance. To the best of my knowledge: (i) the candidate has not submitted the same research work to any other institution for any degree/diploma, Associateship, Fellowship or other similar titles (ii) the thesis submitted is a record of original research work done by the Research Scholar during the period of study under my supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

Signature of Supervisor: …… Date: **05/03/2021**

Name of Supervisor: **Dr. Narendrasinh C Chauhan**

Place: **Anand, Gujarat**

Annexure-IV (A)

Course-work Completion Certificate

This is to certify that **Geetali Saha,** enrollment no. **129990911006** is a PhD scholar enrolled for PhD program in the branch **Electronics and Communication Engineering** of Gujarat Technological University, Ahmedabad.

(Please tick the relevant option(s))

☐ She has been exempted from the course-work (successfully completed during M. Phil Course)

☐ She has been exempted from Research Methodology Course only (successfully completed during M. Phil Course)

☑ She has successfully completed the PhD course work for the partial requirement for the award of PhD Degree. His/ Her performance in the course work is as follows;

Grade obtained in Research Methodology (PH001)	Grade obtained in Self Study course (Core Subject) (PH002)
BB	AA

Supervisor's Signature

(Dr. Narendrasinh C Chauhan)

Annexure-V

Originality Report Certificate

It is certified that PhD Thesis titled **PREDICTION OF SEA SURFACE TEMPERATURE USING MACHINE LEARNING TECHNIQUES** has been examined by us. We undertake the following:

a. Thesis has significant new work / knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled / analysed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using **Turnitin** (copy of originality report attached) and found within limits as per GTU Plagiarism Policy and instructions issued from time to time (i.e. permitted similarity index <=10%).

Signature of the Research Scholar: ……  Date: 05/**03/2021**

Name of Research Scholar: **Geetali Saha**

Place: **Anand, Gujarat**

Signature of Supervisor: ……  ……. Date: 0**5/03/2021**

Name of Supervisor: **Dr. Narendrasinh C Chauhan**

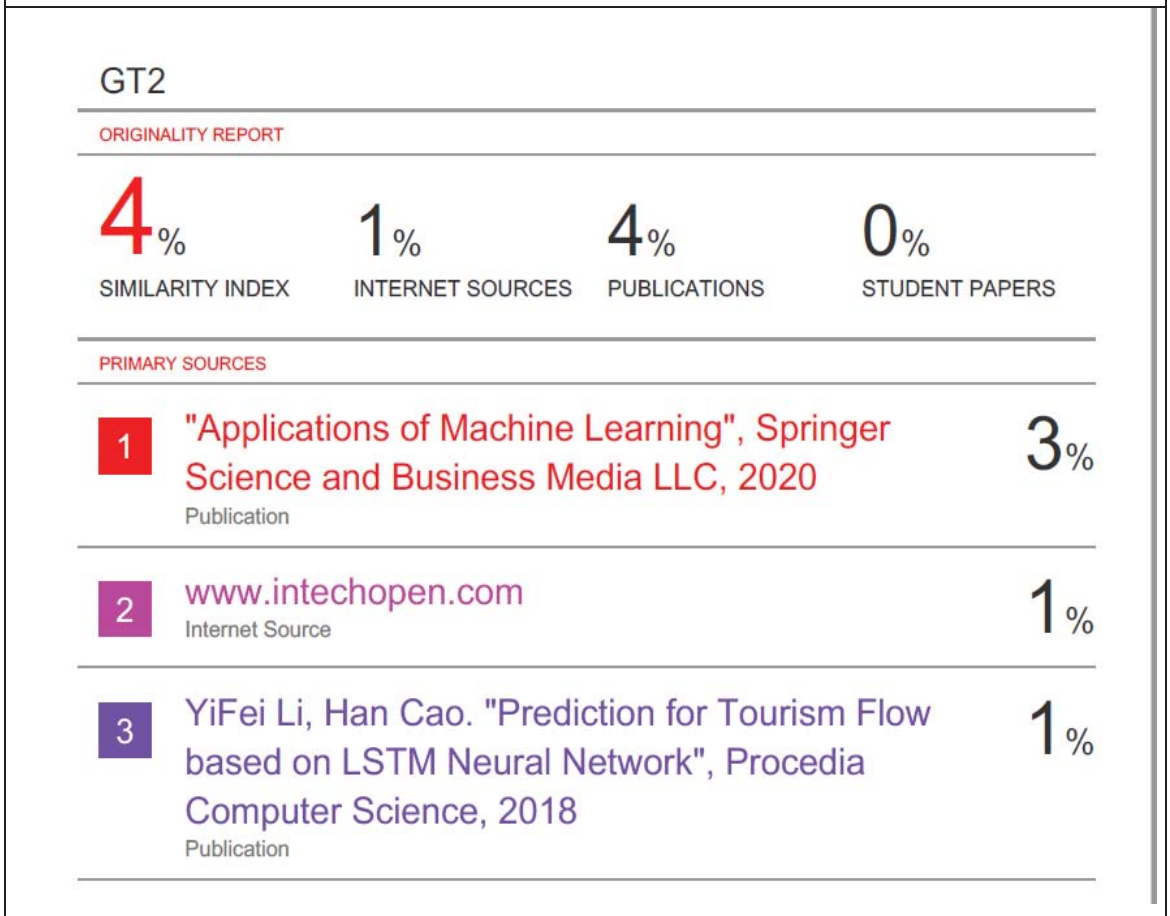
Place: **Anand, Gujarat**

Copy of Originality Report

Prediction of Sea Surface Temperature using machine learning techniques by

Geetali Saha

Submission date: 07-Mar-2021 09:02PM (UTC-0800)
Submission ID: 1527083085
File name: Revised_Thesis_6_3_Only.pdf (2.64M)
Word count: 33116
Character count: 16303



GT2

ORIGINALITY REPORT

4%	1%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Applications of Machine Learning", Springer Science and Business Media LLC, 2020 Publication	3%
2	www.intechopen.com Internet Source	1%
3	YiFei Li, Han Cao. "Prediction for Tourism Flow based on LSTM Neural Network", Procedia Computer Science, 2018 Publication	1%

PhD THESIS Non-Exclusive License

to

GUJARAT TECHNOLOGICAL UNIVERSITY

In consideration of being a PhD Research Scholar at GTU and in the interests of the facilitation of research at GTU and elsewhere, I, **Geetali Saha** having (**129990911006**) hereby grant a non-exclusive, royalty free and perpetual license to GTU on the following terms:

a) GTU is permitted to archive, reproduce and distribute my thesis, in whole or in part, and/or my abstract, in whole or in part (referred to collectively as the "Work") anywhere in the world, for non-commercial purposes, in all forms of media;

b) GTU is permitted to authorize, sub-lease, sub-contract or procure any of the acts mentioned in paragraph (a);

c) GTU is authorized to submit the Work at any National / International Library, under the authority of their "Thesis Non-Exclusive License";

d) The Universal Copyright Notice (©) shall appear on all copies made under the authority of this license;

e) I undertake to submit my thesis, through my University, to any Library and Archives. Any abstract submitted with the thesis will be considered to form part of the thesis.

f) I represent that my thesis is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this non-exclusive license.

g) If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required, I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection.

h) I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to my University in this non-exclusive license.

i) I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my University in this non-exclusive license.

j) I am aware of and agree to accept the conditions and regulations of PhD including all policy matters related to authorship and plagiarism.

Signature of the Research Scholar: ……

Name of Research Scholar: **Geetali Saha**

Date: **05/03/2021** Place: **Anand, Gujarat**

Signature of Supervisor: ……

Name of Supervisor: **Dr. Narendrasinh C Chauhan**

Date: 05/**03/2021** Place: **Anand, Gujarat**

Annexure-VII

Thesis Approval Form

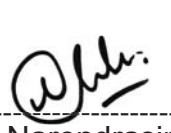
The viva-voce of the PhD Thesis submitted by **Geetali Saha** (Enrollment No. 129990911006) entitled **PREDICTION OF SEA SURFACE TEMPERATURE USING MACHINE LEARNING TECHNIQUES** was conducted on . . . 29/04/2021 (THURSDAY) . . . (day and date) at Gujarat Technological University.

(Please tick any one of the following option)

☑ The performance of the candidate was satisfactory. We recommend that he/she be awarded the PhD degree.

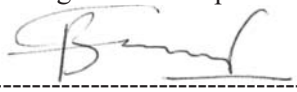
☐ Any further modifications in research work recommended by the panel after 3 months from the date of first viva-voce upon request of the Supervisor or request of Independent Research Scholar after which viva-voce can be re-conducted by the same panel again (briefly specify the modification suggested by the panel).

☐ The performance of the candidate was unsatisfactory. We recommend that he/she should not be awarded the PhD degree (The panel must give Justifications for rejecting the research work). .



Dr. Narendrasinh C Chauhan
Name and Signature of Supervisor with Seal

Dr. Fakhrul Hazman Yusoff
1) (External Examiner 1) Name and Signature



Dr. Brij Bhushan Gupta
2) (External Examiner 2) Name and Signature

3) (External Examiner 3) Name and Signature

ABSTRACT

The interests in seasonal climate prediction has attracted attention of scientific and economic communities in recent years. Some of important oceanic parameters such as sea surface temperature, sea surface salinity, sea bottom temperature, sea level pressure, the zonal and Meridional winds are known to reflect the dynamism involved in the global Biome. Out of all these parameters Sea Surface Temperature (SST) is the most fundamental parameter in understanding marine ecosystem, ocean conditions and climate dynamics. It is beneficial in the study of seasonal to interannual climate variability. SST prediction is a task of predicting future values of a given sequence using historical SST data. Initial methods of SST prediction model are driven by physics based phenomena of heat transfer across the atmosphere and the ocean over large spatial domain. Recent developments in the field of machine learning and computational intelligence have proved promising results in wide range of applications including climate predications. These models exhibit excellent performance for modeling specific parameters particularly used in fishing, sporting events, small marine ecosystems, aquatic flora and fauna population monitoring and observation. This research investigates and contributes in prediction of sea surface temperature by exploring existing and developing novel artificial intelligent methods based on concepts of machine learning and deep learning. Additionally, dependency amongst the Oceanic parameters is investigated and conclusions are derived.

In this thesis, we have investigated different models of neural networks for performance prediction of SST data and also presented machine learning based algorithms for SST prediction. In our experiments, we have used both monthly and daily SST datasets. For monthly datasets, exploration started with the Auto Regressive Integrated Moving Average (ARIMA) model by Box Jenkins methodology [73, 74]. Using the auto correlation function and the partial auto correlation function plots of the time series, ARIMA models are tested and optimized. Our research then investigates Nonlinear Auto Regressive (NAR) Neural Network model where a set of time delay elements are recognized to be the significant contributors in predicting the future values of SST time series data. This is experimented for single step prediction for daily data, multiple step prediction for weekly and monthly SST

values. The correlation coefficient reveals values in the range (0.87 to 1.00) and RMSE obtained is 0.08137 which are better compared to some of recent literature.

In a site specific approach from La Jolla, Western Californian Coast, readings recorded by contact based thermometers and using buoys, an archived dataset [78] comprising of SST, SSS and SBT is used for analysis to check the dependency of SSS and SBT on SST, one at a time using single step analysis and detecting the anomalous readings. Using a daily dataset from PMEL [77], we have attempted to recognize SST anomaly (Anomaly) by a modification of the NAR network using exogenous inputs and evaluating their influences over SST for weekly prediction. This research also investigated Long Short Term Model (LSTM) deep neural network for single step and multistep prediction of SST values and compared its performance with other approaches.

An important contribution in this thesis is development of a hybrid model, specifically to predict the daily SST dataset with multistep prediction. This approach models linear and nonlinear components (residuals) of SST dataset using different time series prediction techniques and finally hybridize them to overcome limitations of individual techniques and utilize their capabilities. This approach uses ARIMA for linear modeling, while NAR as well as LSTM for nonlinear modeling. The proposed hybrid approach is evaluated on various SST datasets and the approach is found better compared to individual machine/deep learning techniques. We have recorded NMSE in the range of (0 to 1.7711) and NRMSE (0.03 to 1.98) which is significantly low for daily SST value predictions.

In all, the thesis contributed in development of novel method and investigation of various machine learning and deep learning based approaches along with traditional ARIMA based models for prediction of the important parameter SST and also analyzed related anomalies.

Acknowledgement

The effective completion of this work could materialize due to the support of many individuals. I hereby take the opportunity to thank all who have contributed towards its successful completion.

To begin with, I bow down to Goddess Saraswati, the provider of all knowledge.

I express my deep gratitude to my Supervisor, Dr. Narendrasinh.C. Chauhan. I sincerely appreciate his suggestions and his keen interest in resolving queries. I shall remain ever thankful to him for enhancing presentation of minute details which reflects his deep understanding and provides clarity of perception to the reader. I am whole heartedly thankful to him for all his guidance, without which this compilation would not have been possible.

I am thankful to my Research Panel members- Dr. Himanshu Mazumdar and Dr. Narendra M Patel for their constant support and endless suggestions on improving my contribution.

I also convey my thanks to Dr. Sudhir Vegad, Prof. Anand Pandya, Mrs. Hemlata Patel, Rajendrabhai and Arjunbhai from my Supervisor's workplace-the Department of Information Technology, A. D Patel Institute of Technology (ADIT), New VV Nagar, Anand, Gujarat.

I express gratitude to the whole hearted support extended by – Dr. Himanshu Soni (Principal), Dr. Hitesh Shah (Head of the Department), Dr. Rahul Kher, Dr. Mehul Shah, Prof. Nilesh Desai, Dr. Falgun Thakkar, Dr. Samir Trapasiya, Prof. Pradeep Shah, Prof. Parthesh Mankodi and Prof. Rohit Parmar at my workplace, Department of Electronics and Communication, G H Patel College of Engineering and Technology (GCET), VV Nagar, Anand, Gujarat. I am also thankful to Dr. Yogesh Chauhan and Prof. Sneh Vyas from the Cultural Committee of GCET.

I am also thankful to Dr. M.C. Deo and Dr. K. Patil from the Civil Engineering Department, IIT Bombay for sharing their database with us for comparative performance.

I am extremely thankful to my family, my parents and my in laws who have blessed me to achieve this compilation.

My special thanks to my spouse-Sukanta, who has supported this cause beyond his priorities.

Annexure-IX

I convey my gratitude to my lovely daughter Sumona for her unbiased support at every stage.

Their blessings and goodwill have accentuated my journey to an altogether different level. Their constant support has empowered the contribution.

Last but not the least, I am thankful to all those who have been associated with me directly or indirectly on this remarkable journey.

Geetali Saha

Table of Content

List of Abbreviations

ANN	Artificial Neural Network
ANFIS	Adaptive Neuro-Fuzzy Inference System
ARIMA	Auto Regressive Integrated Moving Average
AVHRR	Advanced Very High Resolution Radiometer
CCA	Canonical Correlation Analysis
COAD	Comprehensive Ocean-Atmosphere Dataset
CSIL	Computational and System Information Lab
DNN	Deep Neural Network
DWT	Discrete Wavelet Transform
EEA	European Environment Agency
ESRL	Earth System Research Laboratory
ERSST	Extended Reconstruction Sea Surface Temperature
ENSO	El Nino Southern Oscillations
GHRSST	Group for High Resolution Sea Surface Temperature
GISST	Global Sea Ice and Sea Surface Temperature
HadISST	Hadley Centre Global Sea Ice and Sea Surface Temperature
ICOADS	International Comprehensive Ocean-Atmosphere Data Set
IOD	Indian Ocean Dipole
IO SST	Indian Ocean Sea Surface Temperature
LSTM	Long Short Term Model
LR	Linear Regression
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NLR	Non Linear Regression
NMC	National Meteorological Center
NOAA	National Ocean and Atmosphere Administration
OISST	Optimum Interpolation Sea Surface Temperature
PCA	Principal Component Analysis
PMEL	Pacific Marine Environmental Laboratory
PSD	Physical Science Division
SBT	Sea Bottom Temperature
SEIO	South East Indian Ocean
SLP	Sea Level Pressure
SSS	Sea Surface Salinity
SST	Sea Surface Temperature
SVM	Support Vector Machines
WNN	Wavelet Neural Network

List of Figures

List of Tables

List of Appendices

CHAPTER – 1

Introduction

Our residing planet, Earth has undergone many transformations, right from its inception many centuries back. As we all are aware that life began in water, hence the oceans that make up almost 60% to 70% of the Earth contribute towards monitoring such changes. Sea Surface Temperature (SST) monitoring using mathematical and physics based Radiative Transfer Modeling methods have evolved since the 19th Century [6]. Various parameters like the temperature at the surface of the sea, its salinity, humidity of the air in the surrounding, the speed and the directions of the winds, the temperature at the specific depth of the sea, the temperature of the air in the surrounding, the sea level pressure etc play a significant role in indicating the changes. Many researchers have studied one or more of such parameters using various techniques, for long or short durations at different locations [1, 2, 3, 4, 7, 9, 11]. Of all the different parameters under study so far, it is the Sea Surface Temperature (SST) that is most widely investigated [1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 13], firstly due to its easy availability and then followed by the development of the radiation transfer scattering models. Since the middle of the 19th century, the sea faring travelers initiated readings of the surface temperature and documented the same, giving rise to the Time Series. It is this surface that acts as the medium for exchange of gases, heat, radiation, momentum and moisture between air and water. The temperature of the water of the sea in the vicinity of the surface, typically in the range of a few mms to about 1 m is generally defined as the sea surface temperature [84].

Different communities seek the knowledge of SST for a variety of reasons ranging from mandatory activities like fishing, sample collection, prediction of events like El Niño, La Niña, tropical cyclones, climate analysis etc. [1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 28, 64] to leisure activities like sea travels, small marine ecosystems, aquatic flora and fauna population monitoring and observation, diving and adventure sports [17, 18].

Advancements in modern equipment, use of satellites and digitalization techniques have made the measurements and data collections possible for different sea parameters.

1.1 Time Series Analysis: an Overview

Time Series is a sequence of values measured for specified time duration at regular intervals for any phenomena. It is observed in many real time applications. Be it the share market, the electrical load consumption, the energy power generated by renewable or non-renewable sources, the emissions by industries, the height of the ocean tides, the solar radiation cycle, the efficiency of machineries, the performance of students in examination, the number of crashes recorded in the Air Transport Industry, the growth cycle of plants and plantations. All happenings that can be recorded are on the path of time series generation. The advent of new technological marvels has further increased the sphere of reach for generating time series sequences. And they penetrate our daily lives in a highly unpredictable manner thereby influencing not only the targeted parameters but also those that are correlated to it in a direct or indirect manner. There are; however certain intrinsic components that is linked to the existence and analysis of the time series sequences. Researchers have analyzed Time Series using various techniques. The important components of the time series include level, trend, seasonality and irregularity [70]. A time series that describes all four components is shown in Fig.1.1.And in most cases, more than one component is predominant in the time series.

The level of the series provides an overall average estimate value for the entire series. And is most often the most commonly calculated factor that provides a fair projection of the time series variation. This is followed by Trend calculation which senses the overall increment or decrement of the series. Any direct indication of gradual increment or decrement can actually trigger a set of correlated events, including alarms being generated. Trend calculation plays a significant role in projecting upward or downward growth of various industries, proposals under scrutiny and a fair perspective on newly introduced schemes. Seasonality is a mirror of repeatability that is reflected on a short term basis. Say, if I consider an annual representation of sales of certain goods, then it may be observed that specific goods tend to be purchased more often during certain festivals or during monsoon, or winter or summer. But the sale of

the same aforementioned commodities goes down as soon as the celebration is over or when the season changes and hence the name seasonality is associated with it. The same phenomenon repeated on a long term basis, say over years or decades, give a prognosis of the commodities and hence can play a very significant role at driving its production, packaging and availability at the counters. This is known as Cyclicity. Last but not the least is the Irregularity associated with random fluctuations in the series. These are, at most of the times, cause of a major concern. They include the outliers, their analysis, explained and unexplained sudden variations that can probably not be linked directly to any of the internal or external factors. However, a detailed analysis may reveal a remote linkage to some past occurrence or contributed by some sudden external development.

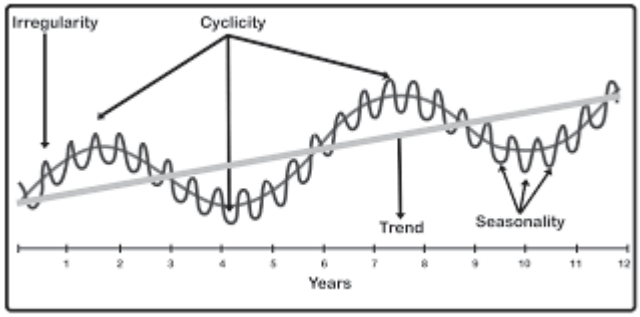


FIGURE 1.1 A time series showing all four components

The exploration of a time series involves its analysis in three possible dimensions. The first dimension being that of an *Insight based approach*; where the inherent characteristics are studied and possible linkages to certain statistical parameters are established based on the behavior of outliers or increasing or decreasing tail. The second dimension involves an *Outsight based approach* where a correlation is to be investigated amongst the various parameters that are under study. This could result into a series of dependency tests and the revelations could be quite intriguing at times, leading to conclusions that have often been overlooked. The third dimension is the dimension of foresight wherein depending upon the characteristics of the series, its future values can be predicted. This is also performed in a number of ways and involves statistical as well as non-statistical based approaches that could be done in terms of a single step or multi step analysis. Time Series Analysis involves identification of its characteristics and exploring the same to forecast future values. The

experimental dataset reflects the choice by researchers and is largely driven by market scenarios, prevailing conditions in the society and most importantly by the own interest of the researchers also.

Many models are reported to achieve a representation same as that of the actual data through various forecasting techniques which ranges from basic statistic based approaches to complex hybrid algorithms. In the beginning, there were more of the statistical based methodologies like the Box Jenkins Methodology [29] which is a popular choice even today after so many decades. Later on as the computation power offered by modern computers increased, these were replaced by algorithms supported largely by mathematical backgrounds and derivations. Models like the Auto Regressive Integrated Moving Average (ARIMA) [25, 34], Linear Regression (LR) [1, 4, 8, 13], Support Vector Machines (SVM) [32, 39], Wavelets [16, 26], Artificial Neural Network (ANN) [1, 2, 3, 4, 12, 13, 20, 67, 68], Deep Neural Network (DNN) [21, 29] and many more find wide popularity in this application domain.

1.2 Motivation

Consider the work by Hsieh and Tang [1], way back in 1997, who presented the use of Neural Network for the purpose of prediction and analysis of Meteorological as well as Oceanographic data along the equatorial Pacific belt, covering the regions, popularly known as the Niño3, Niño 3.4, Niño 4, P4 and P5, using Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and artificial neural networks (ANN). Parallelly, the duo along with Tangang [2], worked on the prediction of SST Anomalies (SSTA) in a seasonal perspective for a lead time of 9 to 12 months. They experimented with Rossby signature detection using Wind and SSTA data, separately and together, especially in the Niño3.4 region. Again in 1998, all three proceeded to explore and chart the prediction of SST using Sea Level Pressure (SLP) and Wind Stress as the dependent parameters including Niño3.5 and P2 regions along with the afore mentioned zones for lead time as high as 15 months [3]. In 2000, the trio along with Mohanan [4] performed an investigative study of linear and non-linear methods by means of RMSE and CC extending the lead time to 21 months. All of them used monthly datasets over the Pacific region during the years 1950 to 1997 and these datasets

are either reconstructed from Smith's SST dataset or Reynold's SST dataset [5] or a combination of both using various preprocessing steps. An improved version of these datasets using bias correction of the historical datasets, especially for high latitude locations using sea ice concentration is also available by the same authors for research [6]. The changes observed in the Indian Ocean are linked to variations observed in the Pacific Ocean [7] and also to the Asian Australian rainfall occurrence [8]. Hence, these are the preferred regions of interest.

Moreover, El Niño events are accompanied by the warming of the Indian Ocean. Utilizing the Equatorial Heat Content as one parameter and the Western Pacific Wind in combination to it, the occurrence of the El Niño and La Niña events [7, 11] can be forecasted to near precision and its peculiar relationship with the Indian Ocean SST (IO SST) is realized. Certain zones around the Indian Peninsula are recognized as potential locations that are linked to cause global variations. This motivated us to find locations of interest and the parameter we could most relate to is the sea surface temperature.

Based on the findings that revealed that the Indian Ocean SST could be a good indicator of El Niño and La Niña events, Tripathi et.al [12] had identified a small location in the Indian Ocean region, defined by 27º to 35ºS, 96º to 104ºE and proposed 12 different neural networks, one for each month of the year to predict monthly SST values. Additional locations - Central South Indian Ocean, Northwest of Australia, Southern Indian Ocean and Antarctic Circumpolar Current are identified by [13] and a relation between the SST and the Indian Summer Monsoon Rainfall is established. S.B. Mohongo [14] attempted the seasonal and monthly SST predictions using ARIMAX and the NARX model for the years 1981-2010. The sites selected were the Eastern coast of Africa (EAF)at 6-7ºS, 39-40ºE and Equator (EQT)located at 0-1ºS, 59-60ºE - both located within the western pole of the Indian Ocean Dipole. Later Patil.et.al [16] identified six locations- Arabian Sea (AS; 19ºN-20ºN, 68ºE), Bay of Bengal(BOB; 18ºN-19ºN, 90ºE), West of Indian Ocean (WEIO; 1ºS-1ºN, 65ºE), East of Indian Ocean (EEIO; 1ºS-1ºN, 90ºE), off the African coast (THERMO; 14ºS-16ºS, 56ºE-58ºE), and South of the Indian Ocean (SOUTHIO; 9ºS-11ºS, 95ºE-98ºE) and proposed a Wavelet based Neural Network for forecasting the SST values up to 5 time steps ahead in the future. Andreo [18] investigated the occurrence of bloom and its frequency based on Chlorophyll-a and Phytoplankton blooms using a 11 year database in the Argentinian region and correlated the same to the corresponding variation in the SST values. Using the Scripps Pier dataset, McGowan [28]

attempted to correlate the growth of Chlorophyll-a with various nutrient distribution along the vertical column of water, density, wind, surface temperature and bottom temperature anomaly. Ursu and Pereau [30] proposed a Periodic Auto Regressive Network using an Exogenous Variable to correlate the catch-per-unit-effort of shrimp using the SST as an exogenous variable for fishery data in the time period of study from Jan 1989- Dec 2012. They used a Genetic Algorithm based approach and clubbed the results using Bayesian Information Criterion. Salles [43] highlights the significance of the prediction limits and its relation with the dimension of the training datasets for the task of SST prediction. Using 21 buoys from the PIRATA weekly, monthly and daily datasets, SST predictions are performed in terms of single step only.

Based on the study, we realized that traditional methods like ARIMA and NAR are able to predict the SST values. However, the same is restricted to a single step only. Hence, we proposed optimization so as to enable multiple steps of prediction and we could achieve this successfully. The other concern we identified was that the Daily SST dataset still posed a challenge as far as SST multistep prediction was concerned.

1.3 Challenges involved in the study

Once certain key locations are identified as the region of interest, we tried to obtain the relevant datasets for analysis. However, as we all are aware that the Earth structure is such that the poles are mostly covered with ice and the water that results when this ice melts contributes to the seas and oceans eventually. And this in itself is a recurring cyclic process. The concentration of ice is always a challenge to the calculation of SST values. There is a point known by the name of Point Nemo, located in the South Pacific Ocean that is also popular by the name of the Oceanic Pole of inaccessibility. As it is farthest from land, it is also very sparsely populated. Apart from that there are parts of the oceans that were totally inaccessible in the earlier days when the measurement was dependent on manual techniques or localized sensors. However, with the advent of satellite communication and advanced

sensors this started reducing. But with satellite sensors, the major concern was cloud cover, as some of the earlier sensors failed to provide correct readings under various cloud cover and other precipitation conditions. Today the scenario is completely different with so many technical advancements and correction techniques; it is easy to obtain global coverage to even the most remote place. Various countries have undertaken missions to record the SST changes occurring not only at the seas and oceans, but also at various marine biomes, mangrove forests and aquatic observatories.

1.3.1 Data observation centers

The SST and relevant parameters' datasets are available through various Centers of observation distributed across various countries. These are governed by Government missions in the domain of Earth and Ocean Observatories. At an average, there must be around 3000 odd satellites contributing towards Earth Observation alone. These are dedicated towards various applications ranging from Aerosols monitoring, land formation and deformation, expansion of sea level, lowering of sea level, vegetation growth and green coverage. In fact, using high definition cameras installed outside satellites like ISS, people can nowadays view the Earth live as well. The SST and allied parameters are available through observation centers namely National Ocean and Atmosphere Administration (NOAA), Pacific Marine Environmental Laboratory (PMEL), Earth System Research Laboratory (ESRL) - Physical Science Division (PSD), National Centers for Environmental Prediction (NCEP), National Center for Atmospheric Research (NCAR)- Computational and System Information Lab (CISL), International Comprehensive Ocean-Atmosphere Data Set (ICOADS), European Environment Agency (EEA) and many more [15, 16].

Various popular platforms include but are not limited to, Group for High Resolution Sea Surface Temperature (GHRSST), Global Sea Ice and Sea Surface Temperature (GISST), Kaplan SST, Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST).

1.3.2 Missing values

Due to instrument failure or reading discontinuity contributed by manual/automatic processing of the time series, missing values creep into it. In spite of progress in the sensor specifications, their interfacing capabilities, and even advancements in this domain, the missing observations have been observed in many datasets. There then arises the need for interpolation techniques that can be used to identify the extent of such gaps and introduce the near correct readings to maintain continuity. There are different ways in which data interpolation can be achieved for such cases. In this thesis, missing data have been replaced with suitable interpolation techniques.

1.3.3 Addressing the daily SST dataset

In the initial phase of recording oceanic observations, it was mostly annual or monthly data that used to be available, and that too, the coverage was very poor at the near pole region. And hence the challenge mostly centered on these datasets. Due to ease of availability and based on requirements, researchers have addressed investigation and analysis of monthly dataset using various linear and non-linear techniques [12, 13, 14, 15, 38, 54, 55, 56, 57]. However, daily SST datasets have very few takers [16, 20, 21, 29]. Diurnal datasets show high variability and hence provide a fair projection of the changes that are occurring on a regular basis and present the true picture. Also, due to its high variability, it is challenging to understand and forecast the same. Their availability in terms of an uninterrupted chain of events is not so regular. The various satellite based data collection drives launched during joint Earth missions by different group of countries has changed this scenario to a very large extent. And this has resulted into time periods of uninterrupted data. Joint availability of temporal and spatial availability may get compromised in certain cases. This inspired us to work upon Daily SST datasets and attempt prediction algorithm for the same.

1.4 Research objectives, Problem statement, and Scope

The study is meant to address the prediction of SST parameters in a 3 dimensional manner involving the Insight, Outsight and Foresight based approach. This has led to the realization of a set of Research Objectives, Identification of the Problem Statements and Scope as applied to the specific domain of Research which in the present case is about the Forecasting of the Sea Surface Temperature. The Research Objectives, Problem Statements and Scope associated with this present Research work are stated as below:

1.4.1 Research Objectives:

1) The main objective of this work is investigation and development of machine learning and deep learning based methods for prediction of SST.
2) Another important objective is to develop hybrid method based on traditional time series prediction method with machine learning and deep learning methods and test it for prediction of SST using various dataset. The aim to develop hybrid method is to improve the prediction accuracy of diurnal SST as compared to literature methods.
3) One of the important objectives is the development and investigation of above methods to perform and evaluate single-step and multi-step prediction for SST time series dataset.
4) One of objectives is to perform comparison of methods developed and used for prediction of SST.
5) An objective in this work is identification of SST anomalies from the SST time series datasets. A minor objective is also to Investigate and understand dependency relationship of SST with other oceanic parameters like Sea Surface Salinity (SSS), Sea Bottom Temperature (SBT), Zonal winds, Meridional winds and Air temperature in a site specific approach.

1.4.2 Problem statements:

We propose the forecast of Sea Surface Temperature data using various existing models and seek to achieve the same with more precision.

- Traditional methods of time series analysis are used for linear prediction. Exploration of non-linear time series analytical methods is necessary for prediction of SST.
- Machine learning methods like Neural Network are used for SST time series prediction. Additional parameters may play significant role in increasing accuracy.
- Investigation of recently popular Deep Learning methods and generation of new methods can also improve the performance.
- Anomaly detection is another aspect that is very widely investigated by researchers. However, it is observed that anomaly detection is limited to monthly datasets only.

1.4.3 Scope:

- Our proposed research considers majorly time series dataset for SST prediction. The research mainly focuses on development and investigation of traditional, machine learning and deep learning based methods for SST time series prediction. The prediction of dependent attribute(s) with reference to set of independent attributes is considered to be outside the scope of this work.
- The present research investigates only artificial neural networks based methods for SST prediction. Other machine learning methods like decision tree, Bayesian methods, Support Vector Machine, etc. have not been considered for investigation. However, it also investigates role and use of LSTM based deep learning method for SST time series prediction.
- Of all the oceanic parameters, apart from SST, we have used SBT, SSS , Zonal winds, Meridional winds and Air Temperature to provide better prediction of SST values.

1.5 Organization of the Thesis

The thesis is organized as below:

Chapter 1 has introduced the problem of SST prediction, motivation for the work, objectives and scope of research.

Chapter 2 is split into two parts. The first part addresses the formation of the SST Climatology Maps by interpolation for missing data. The second part is about the development of techniques used for prediction of Ocean parameters and the progress is charted in terms of techniques used in Literature Review.

Chapter 3 highlights the approaches involved in justifying our contributions. This chapter highlights the parameters used for quantification and measurement of the error. It provides the framework that links the remaining chapters and highlights our Research contribution.

Chapter 4 is about the implementation of existing techniques on monthly SST datasets. We have also discussed how optimizing the parameters of existing technologies can lead towards minimizing the errors. It also provides a comparative of performance against the previously existing algorithms about the daily SST analysis using additional parameters like SSS and SBT. We have investigated the dependency of each parameter in identifying readings having large variation from the mean values successfully at many instances. It also addresses the SST Anomaly detection using Air Temperature and Winds- Zonal and Meridional, all, one at a time for a daily SST dataset.

Chapter 5 introduces Deep Neural Network technique- Long Short Term Model (LSTM) for SST analysis.

Chapter 6 addresses the successful prediction of daily SST using the proposed Hybrid model. The proposed Hybrid model tries to bridge the gap existing between actual and predicted values by exploring the characteristics of the time series sequence.

Chapter 7 is about the conclusion and future work.

CHAPTER – 2

Preliminaries and Reviews

2. Preliminaries and Review

2.1 Overview

The temperature of the water of the sea in the vicinity of the surface, typically in the range of a few mms to about 1 m is generally defined as the SST [84]. The measurements can be made using a variety of techniques ranging from the basic contact based instruments like thermometers or thermistors placed on ships, moored buoys, drifting buoys to contactless remote access using Satellite links. In the latter case, few specific wavelengths of the Electromagnetic Spectrum carry the signature of the SST and the same is retrieved by various techniques.

However, this entire dataset has inbuilt inconsistencies contributed due to various factors; most prominent amongst them being the formation of ice over the polar and near polar regions, the vastness of the Oceans giving rise to out of reach locations, the mismatch constituted by the different platforms of data collection, various political and economic changes (like formation of canals, gulfs, redefining territories, World war, etc.). While researchers were in the process of mapping these readings on the global grid by interpolation to fill in the gaps, some very significant drastic climatic changes were also recorded in the World Climate - the El Niño and La Niña [2]. Various machine learning techniques were devised to establish relationships amongst them [1, 3, 4].

SST is measured using various techniques - most popular amongst them being the satellite based measurements; the radiometers based measurement and the Infrared radiometers. Different techniques also provide measurements till variable depths. IR instruments are generally capable of reaching to a depth of 20 micrometers only whereas microwave

radiometers can penetrate up to a few millimeters [84]. Even if the readings are recorded using the same sensor set when it comes to forming a continuous dataset, the boundary conditions always create challenges. Other challenges arise when we need to merge readings recorded using various instruments or using different techniques as we expand along the spatial and/or temporal domain.

For ease of understanding, the entire discussion of this review is split into two sections- the first one about the SST climatologies and the latter one about the machine learning techniques used for analysis of these climatologies. The list of references used is non-exhaustive but it definitely provides a fair representation of the past advances, encompassing the present thereby providing a worthy picture of the future.
The organization of the chapter presented as below. Section 2.2 presents the Global SST climatologies highlighting its significance and characteristics. It also charts the formation of Global SST climatology maps. Section 2.3 represents the SST prediction using Machine Learning Techniques Section 2.4 is about the Hybrid Regression techniques for SST prediction.

2.2 Global SST climatology

In earlier days, Empirical Orthogonal Functions (EOF) was widely used for spatial analysis of climate variation over various time zones [19]. Principal Component Analysis (PCA) and Extended PCA (EPCA) are the most commonly used tool for regression and dimension reduction [1, 19]. Multichannel SST profiles are obtained from NOAA Advanced Very High Resolution Radiometer (AVHRR) data using linear methods [56] and using non-linear methods [57]. Statistical correction of Pacific SST to forecast the variations in the ENSO events using Canonical Correlation Analysis (CCA) to check for critical sequence identification gained popularity during 1990s to early 2000s [58, 59]. The purpose behind such climatological studies was largely to obtain improved, real time and global SST [5, 6, 46, 47, 48, 49, 52].

A near real time global SST analysis is carried out by Reynolds and Smith [5, 6, 46, 50], individually and with each other. Again, the duo with fellow researchers - Marsico [47] and Xue [48] at the National Meteorological Center (NMC) of NOAA further developed global SST climatology based on derivation from all available sources including in site (ships and buoys based) SST data, the retrievals of SST and the ice sea covered zone data and proposed a dataset at 2° resolution. Using Optimum Interpolation (OI) Analysis, Reynolds and Smith, later in 1994 introducing many error and bias corrections and proposed a 1° grid on weekly basis having better spatial and temporal resolution [5], also supported by AVHRR from 1982-1993. Using this data, they improved their previous dataset to 1° spatial resolution grid covering the time span from 1950-1979 [46]. In 2004, they improvised their previous work and introduced the Extended Reconstruction SST (ERSST) [6] dataset, this time enclosing a time range from the year 1857 to 1997. The latest dataset from NOAA OISST, Version 2 has weekly 1°grid and daily 0.25°grid blended analysis on daily SST and ice [72].

Trenberth undertook the study of global climate and meteorology and has worked upon a variety of topics ranging from ocean warming, to global water cycle, to precipitation, to El Niño Southern Oscillations (ENSO) and more linking the climate changes [9, 10, 11, 50, 51, 52]. With Shea, in 1992 [10, 52] and with Hurrell in 1994 [11, 52], he investigated the drastic changes that caused the major events of El Niño and La Niña(1976-1988). They proposed a monthly global SST climatology using data derived from Climate Analysis Center using Comprehensive Ocean-Atmosphere Dataset (COAD). They also established a strong connection linking the changes over the Pacific to the tropical variations.

Many researchers have investigated the inter relationships using either dataset or datasets derived from one and supported by other climatological maps. As a result, it is found that the Indian Ocean SST (IOSST) causes variation in the Pacific [7] as well as in the monsoon on the Asian and the Australian [8] fronts. A detailed analysis of the Indian Ocean Dipole (IOD) and its impact on the Ocean Atmosphere coupling especially on the climate of Asia, Africa and Australia is well charted by Vinayachandran, Francis and Rao [69].

These climatology maps helped in the progress of Time Series based models further leading towards Regression of the ocean parameters. As the various numerical approaches improved,

there was significant improvement in the modeling of such Time series SST datasets across the globe. However, the data before 1880 and the data during the World Wars are not so consistent and hence the need for interpolating the missing grid values becomes vital. Also the presence of ice, especially in the vicinity of the poles poses a major challenge to the continuity of such SST datasets. Applications that are affected by SST variation are the Fishery Catch, Water Tourism, Pollution monitoring, Oceanographic research and so on.

2.3 SST prediction using machine learning techniques

Data mining of various temporal SST data helps in the identification of useful patterns at short, mid and long term time scales. There are broadly two manners in which this can be achieved. The first one uses statistic based techniques that involve identification of trend, seasonality, mean, variance, deviation, etc. using Box Jenkins models and/or other linear algorithms. However, the second approach uses the ANN thereby capturing the non-linearity of the temporal SST data. There are few other methods also like the use of exponential smoothing, fuzzy logic systems, evolutionary algorithms and the support vector machines that finds few applications in this domain.

2.3.1 Linear Prediction Techniques

Linear prediction techniques generally find applications where the parameters tend to have an intrinsic linear relationship. Geetha and Nasira [34] utilized a time series of 6 years data from the years 2007-2012 consisting of 14 attributes. They trained the ARIMA model using the training dataset of 2007-2011 years and the last year, 2012 for testing. The software they opted for is Statistical Package for Social Studies (SPSS) 2.0. Reynolds [50] had performed SST calculations using Satellite and In situ data together and could very correctly detect the ENSO signals at the Tropical Pacific region in the 1982-83 year span. He also successfully tracked the SSTA in the Niño-3 region. The Tropical Oceans and Global Atmosphere (TOGA)

program initiated by the US National Weather Service provided the in-situ (ship and buoy) data and the Satellite data. [50]. Using South East Indian Ocean SST as a predictor, Dominiak and Terray [54] had improved their prediction of ENSO using linear regression techniques. The same is tested at Niño3.4 location also. Gaffen et.al [55] using Radiosonde data had established a relationship between water vapor and SST data. The calculations involved the mean monthly and annual data. They analyzed as many as 35 locations spread over the years 1973-1990 in a dynamic time zone. Barnstone and Ropelewski [58], together had predicted ENSO episodes using CCA. Tippet et.al [59] performed statistical correction of the SST forecasts in the Pacific Tropical regions using multichannel CCA analysis.

Tang et al performed a comparison of the performance of the Box Jenkins method versus the NN topologies, in 1991 stressing upon mapping the underlying characteristics of the Time series. They analyzed the Box Jenkins methodology using the basic steps that involves- model identification, parameter estimation, choosing the right ARIMA model and finally forecasting. They investigated the NN networks in terms of a number of factors like the topology of the neural network structure, the training algorithm, learning method, hidden layers for the NN [62]. Based on the comparative by Hsieh and Tang [1], they realized that the Neural Network could outperform the rest of all the Linear regression methods for the purpose of prediction and analysis of Meteorological as well as Oceanographic data along the equatorial Pacific belt, covering the regions, popularly known as the Niño 3, Niño 3.4, Niño 4, P4 and P5, using Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Artificial Neural Networks (ANN). Again with Tangang [2], they worked on seasonal prediction of SST Anomalies. The highlight of their work was the Rossby signature detection using Wind and SSTA data, separately and together, especially in the Niño3.4 region. They also explored and charted the prediction of SST using Sea Level Pressure (SLP) and Wind Stress as the dependent parameters including Niño3.5 and P2 regions along with the afore mentioned zones for lead time as high as 15 months [3]. In 2000, the trio along with Mohanan [4] performed an investigative study of linear and non-linear methods in terms of CC and RMSE extending the lead time to 21 months. As time progressed, more researchers inclined towards investigating the power of Neural Networks in the domain of regression analysis leading to a paradigm shift.

2.3.2 Nonlinear Regression Techniques

This paradigm shift is attributed to various inherent characteristics of the Neural Network right from its inception. It was in 1943 that a physiologist of the neuron system Warren McCulloch with a mathematics expert Walter Pitts together contributed an article on the working of the neurons and represented the same using circuits from the Electrical perspective [85]. Three decades later Kohonen and Anderson, both working on such circuits; but independently, proposed an ADALINE (ADAptiveLINear Elements) circuit. This rekindled research interests and thereby initiated significant contribution towards the development of multilayer neural networks. A decade and a half later, in 1986, Rumelhart, Hinton and Williams [61], proposed back propagation of errors. Although non-linear regression techniques were proposed a couple of decades back, with time, there has been continuous upgradation in the functions used and the performance delivered. Hence, it's popularity still continues. Typically, Neural Networks that initially had a single hidden layer can now be formed using multiple hidden layers and are popular by the name of Deep Neural Networks. They possess capability to extract features inherently leading to better performances and predictions to many real-time complex problems including time-series analysis.

Tripathi, et al. [12], identified such an area in the Indian region (27S-35S, 96E-104E) that they claim to be having a high potential influence over the global climate.

Using 52 years (1950-2001) of data derived from Reynold's dataset of reconstructed SST, they have used twelve neural networks, one corresponding to each month of SST Anomaly data and with the help of corresponding time series presented a comparative with linear regression models using CC and RMSE as the prediction parameters.

Zhang [21] used the Long Short Term Model (LSTM) for prediction of SST on coastal seas of China SST Dataset. They have used the Adagrad Optimization Method over Stochastic Gradient Descent (SGD) method for robust LSTM with a fully connected layer. Further comparing it with the Support Vector Regression (SVR) technique, it is exhibited that the proposed LSTM method is better at predicting the SST values. Athira [22] utilizes the six indices on Air Quality (PM_{10}, $PM_{2.5}$, NO_2, CO, O_3, SO_2, AQI) from the China National

Environmental Monitoring Center (CNEMC) and the weather data provided by NOAA at 6 hours interval. Using the LSTM network, they could generate ten steps ahead prediction of the pollutant content in Air, PM10.

Many researchers have represented a comparative of linear versus nonlinear regression in SST analysis [1, 2, 3, 4, 11, 12, 62]. Many others have addressed the prediction of SST and associated parameters [12, 14, 15, 19, 20, 37, 38, 42]. One remarkable comparative of few existing neural network techniques using various datasets, topologies, transfer functions and training algorithms is projected by Zhang, et al. [60]. Gooijer [63] has offered a detailed comparative of various non-linearity tests of time series models. Apart from that, he has shared methods to test the feasibility of non-linear networks. He has also investigated the need to address multistep prediction. A case of direct linkage of the effects of SST variation to the population of short beaked common dolphins based on sightings at the Alboran West, Alboran East, Gulf of Vera and the Global area (all four regions together) is investigated by Canadas and Vazquez [90]. Using 20 years of sighting dataset and a time series of anthropogenic effects, they developed two approaches to predict changes in their population over the coming century. They had used the projection obtained from a linear line of regression and also the HadCM3 climate model showing temporal variation of anthropogenic effects. Their findings yielded that an increment in SST values could cause a reduction in their habitat occupancy from East to West. Picone et al. [89] have undertaken a comprehensive study of SST across the Adriatic Sea and the Tyrrhenian Sea under the influence of variable quality of the measurement collection techniques of SST in Italian Seas. The study tries to project the effect of various measurement techniques and how their temporal distribution has influenced the prediction performance in terms of monthly mean SST Residuals. The datasets included are

- International Comprehensive Ocean Atmosphere Data Set (ICOADS)
- Italian Institute for the Environmental Protection and Research (ISPRA)
- The Italian Data buoy network (RON)
- The Group for High Resolution SST (GHRSST)

Within the boundaries of Italian Seas (35ºN- 46ºN, 7º E- 20ºE) from the years 1900 to 2016, a comparison of long-term SST variations to Short Time SST variation is demonstrated using scatter plots and annual boxplots.

2.4 Hybrid Regression for SST prediction

Each individual regression technique has its own advantage and drawback associated it. The traditional linear regression methods have shown good performance when it comes to linear events, but most daily life events tend to possess non-linear characteristics. Linear regression analysis of 35 radiosonde stations for short term trend recognition provides a weak relationship between Water and SST whereas the same is strong for long term trends [55]. However this method fails to address the same specifically in the Tropical region. Another study presents a comparative of three ENSO predictors- Equatorial Pacific Ocean Heat Content, the zonal equatorial wind stress and the South east IO SSTA specifically during the late boreal winter [54] using linear regression analysis.

At times it is observed that individual non-linear regression techniques like NN also possess certain shortcomings-instability related with fewer observations, limitations to address a larger spatial extent, and the uncertainty associated with the extent of the hidden layer [1]. Most of the relevant literature in the domain of oceanography involves the study of the Pacific regions and the IOD and both the regions are very vast in the spatial domain. Hence reaching out to these regions and predicting the values of the oceanic parameters there is always a constant challenge. For example, the Equatorial Heat Content and the Western Pacific Wind [7] are the major contributors to ENSO events and driven by the SST variations in the Indian Ocean, which itself is asymmetrical. The significance of the northern and southern IO over the Asian Australian monsoon [8] is addressed where EOF leads towards the identification of leading IOSST and SVD is used to establish a relation between IOSST and the Asian Australian Monsoon.

Further, this is significantly evident in the time series datasets that are tested by Tang et al. [62]. They had analyzed the same for long, short and irregular memory patterned datasets. As per their analysis, Box Jenkins and NN provide reliably good agreement for long memory trend datasets. However, for short term forecasting, Box Jenkins outperforms NN, and for long term forecasting, it is NN that outperforms Box Jenkins. For short memory time series, NN is better than Box Jenkins method based on their analysis of the three different datasets. The aim for development of hybrid techniques is to merge their individual capabilities and minimize the shortcomings.

In 2003, Zhang [24] proposed a hybrid model consisting of two sections-first ARIMA model and then the ANN method to predict step ahead values using Canadian lynx dataset, Sunspot data and weekly BP/USD Exchange rates. This came to be recognized as a trend setter. Later in 2014, Babu [45] had proposed another hybridization of the ARIMA with the ANN method based on Moving Average filter tested on standard real life as well as on a simulated dataset. In 2015, Khandelwal, Adhikari and Verma [26] proposed a hybridization of the ARIMA and the ANN methods based on Discrete Wavelet Transformation (DWT) and successfully tested it on Lynx, Indian mining, Exchange rate and US temperature datasets. In 2016, Patil and Deo [16] proposed a Wavelet Neural Network (WNN) – a hybridization of Wavelet and ANN for daily SST prediction. In 2018, Sanghani, et al. [71] have presented a novel hybrid method by combining ARIMA and SVM technique by testing on Sunspot data, Canadian lynx dataset and US monthly electricity dataset using R tool. In 2018, Khashei and Hajirahimi [41] have proposed a hybrid ARIMA/MLP model for stock market price forecasting. Significant hybrid models are suggested by Santos et.al [96], Oliveira et.al [97] and Mattos et.al [98] specially to address different time series. The same is addressed in details in Chapter 6.

2.5 Conclusion

In this chapter, we have discussed the various Global SST climatologies and their significance in Climate analysis. We also discussed how the results of linear and non-linear prediction techniques differ from each other. The review of various machine learning techniques suggests further investigation of artificial neural network as one of most prominent machine

learning technique for prediction of time series data of SST. With the development of different techniques, many researchers have tried to merge both these methods and have shown better performance with minimum error [16, 24, 26, 41, 45, 71, 96, 97, 98]. The study and review of such hybrid methods used for prediction applications in different domains motivates us that using a hybrid model for prediction of SST is a worthy exploration.

CHAPTER 3:

Research Methodology

This chapter addresses the contents that we have studied as a part of the background study. It is vital to understand how the entire learning evolved towards study and research gap identification. This finally led to the proposed algorithms to address the forecasting of SST values. The organization of the Chapter is as follows

Section 3.1 is about the Introduction and the significance of the SST in the real world. It highlights the contribution made by various researchers in the past. In Section 3.2, we have the Approach Discussion. This is performed in a completely holistic manner so that all aspects of evaluation are well covered. It projects a three-dimensional approach of the study of the research problem. In Section 3.3, we have addressed the Datasets used. It also justifies the necessity of selecting 10% of our dataset as the testing dataset. It provides an overview of the various dataset collection methods and their variability. Section 3.4 is about the set of error parameters used for evaluation. For regression techniques, the accuracy is calculated in terms of error parameters and the corresponding time series plots. Section 3.5 provides prediction of monthly SST Satellite derived dataset. A direct comparative with past literature is also highlighted in a phased manner. The first four problem statements are addressed in the last sub sections and a comparative of the error parameters with respective time plots is also presented for justification. The non-linear optimized algorithm provides appreciable results.

3.1 Introduction

Various methodologies have been identified by different researchers in the domain of forecasting of different sea parameters and SST happens to be one of the foremost parameters to be addressed. SST alone and at times in combination with various other sea parameters like wind, relative humidity, air temperature, salinity have played a vital role in establishing a connection with events that initially felt unrelated but later revealed correlation. A variety of applications ranging from the very basic observation of the large oceans to marine ecosystem monitoring have evolved over the years. There has also been significant development in the sensors and their deployments. At times, occurrence of sudden events has been also correlated to the gradual variation in the values of SST in a site specific approach. Nowadays, the site specific approach is gathering more momentum as the capabilities of the computating systems are increasing at a rapid rate. The algorithms involved in computational front have also undergone multifold progress in terms of implementation support/ease, resource requirements and reduction in execution time. Various linear and non-linear prediction techniques have made their way into analysis of time series data. Remarkable contributions are made by researchers in this domain. A remarkable compilation of various time series by Adhikari and Agrawal [70] gives a detailed basic understanding of various Stochastic and SVM based approaches. A detailed review of soft computing techniques for analysis of time series is presented by by Sanghani, Bhatt, and Chauhan [71]. They reviewed methods like ARIMA, ANN, SVM, and hybrid techniques like ARIMA with ANN, and Adaptive Neuro-Fuzzy Inference System (ANFIS) with their advantages and drawbacks.

Notable in this case are various variations of the regression techniques related to all basic proposed algorithms. To begin with the most initial Box Jenkins methodology proposed by Box and Jenkins [29], which is more popularly known as the ARIMA models that deals with the detection of global minima using Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF). Later it was observed that the ARIMA models fail to capture the non linearities associated with the time series [1, 2, 4, 12, 14, 19, 37, 38, 42], and this directed our efforts towards machine learning as an alternative popular category of time series prediction methods. A literature survey of machine learning methods revealed that the Neural Networks is a popular method that can better project the variations in the time series. Hence,

on the machine learning front, the thesis focused on investigation and development of machine learning techniques (specifically ANN), its variants and hybrid method for the prediction of SST. The thesis also investigated the role and use of deep learning methods for the same purpose. The major research tasks identified and performed in this research is shown in Fig 3.1.

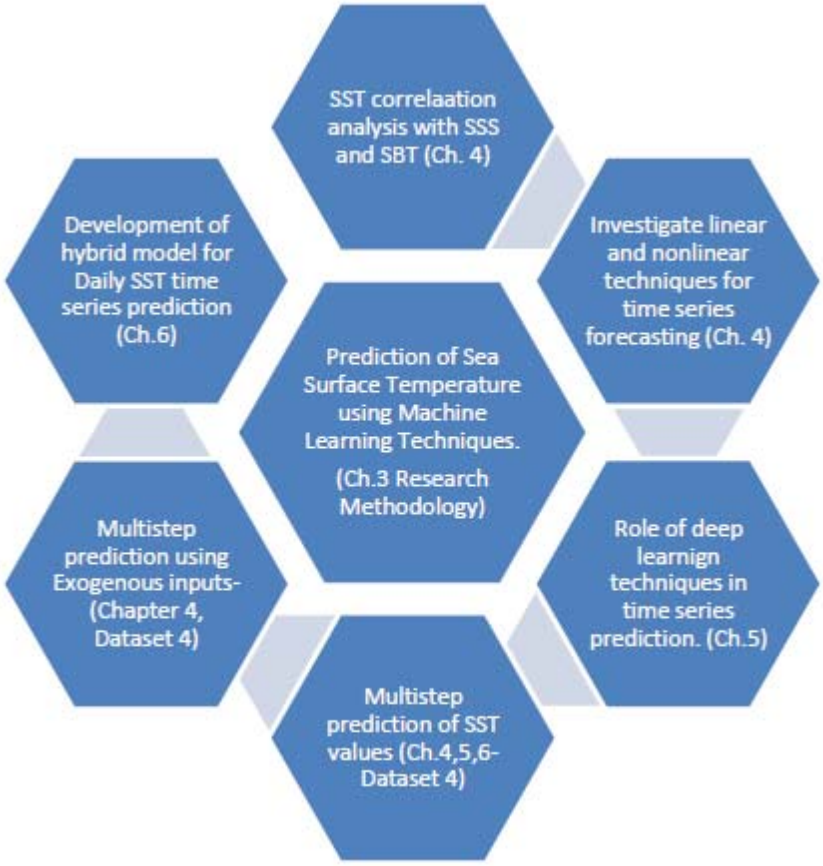


Fig 3.1 Major research tasks targeted in the proposed Research

3.2 Approach discussion:

When we started the study, with very obvious reasons, the very first part of analysis was to understand how the physics based processes can be linked to SST and its forecasting. We studied the basics of the Radiative Transfer Models and learnt about the derivation of the Brightness Temperature from the Energy of the Atmosphere. Based on the Brightness temperature, various relevant parameters like Wind, Water Vapour, Cloud Coverage,

Precipitation and SST products could be derived. Satellite based sensors can record the responses at signature wavelengths of radiation in several parts of the electromagnetic spectrum based on emitted light and energy. A pictorial view that follows explains this better.

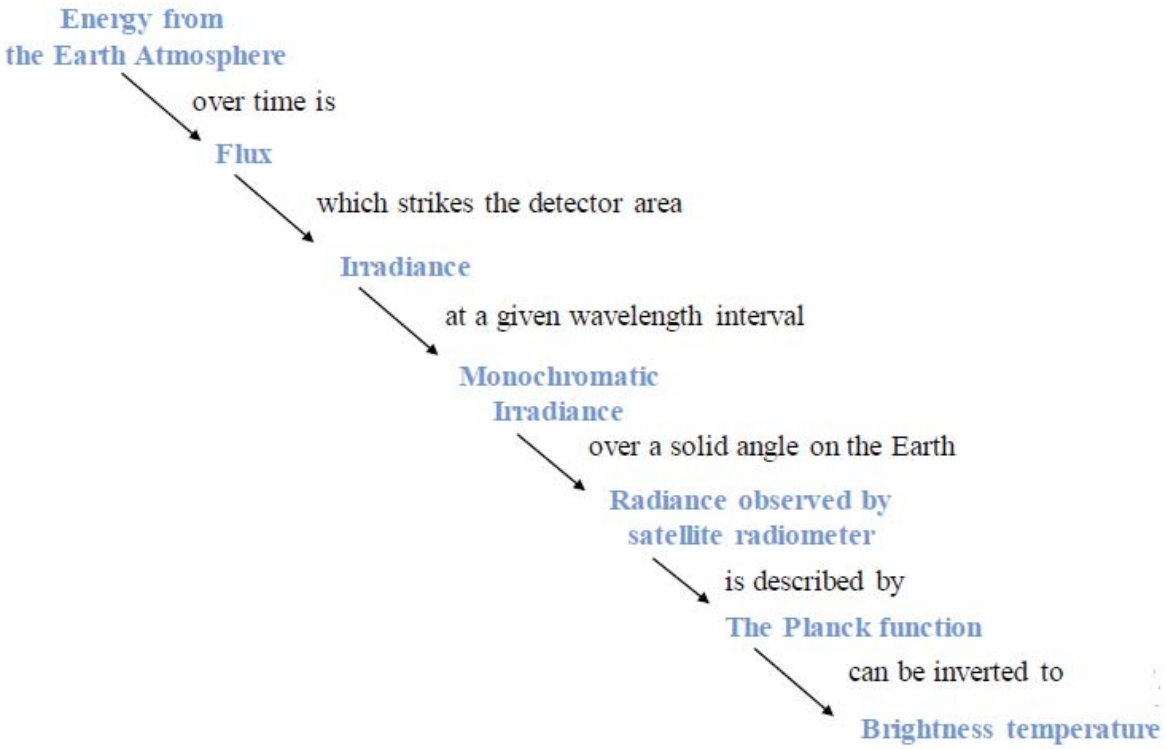


Fig 3.2 The transformation of Radiant Energy into Brightness Temperature

As the study intensified, we realized that a 3D approach would lead to better understanding. Hence, we divided the task into three dimensions- Insight, Outsight and Foresight as stated below.

3.2.1 Insight based approach

It essentially involves the understanding of the hidden characteristics of a time series.

Every time series has dynamism inherent to it and the initial part of the study is basically directed towards learning these characteristics. A few basic techniques involve basic statistical analysis like Trend analysis, Seasonality, and Cyclicity checks. Visualization of the time series for a significant duration of time provides insight into its inherent characteristics. A

change in its level or value over a period of time indicates an increasing or decreasing trend. Such trend identification is a measure of its pattern and is most often the first and most dominant criterion of long term variation. It is also a strong indicator of how future values may align in consequence.

This is classified into three major categories-Short term trends, Intermediate trends, Long term trends. Trend analysis generally tries to achieve a linear fit to an existing time series. Trends necessarily indicate progress; these could be either positive or negative or neutral (that is absence of trend). The easiest way to identify trend is to connect a series of highs and lows in a time series using a line. As little as three points are sufficient to plot a trend. The trend may or may not be monotonous. A Chi square test is most commonly used for detecting such kind of trend.

A plot that is repetitive regularly over a period of time (less than a year) tends to express a seasonal characteristic. In fact the duration could vary from a few minutes to a few months. For SST considerations, the causes for it could vary from natural ones like the solar day, its variation and the inclination angle of the sun, the rate of evapotranspiration, the ebbs and tides to manmade reasons like industrialization, global warming and pollutions. The length of the seasonality period provides leads to the design and optimization of various algorithms that can be used for forecasting future values.

If this kind of repeatability is detected over a number of years or says over decades, mostly about the trend line, then this leads to cyclic behavior. Cyclic fluctuations are generally not of a fixed duration. However, they tend to recur and this is generally revealed when the time series is studied for a long duration. It provides an estimate about upcoming challenges based on the study of the past.

If the time series doesn’t possess trend and/or seasonality, then we may say that the time series is a stationary time series. Unit root or the Dickey Fuller test is the most commonly used test to check this. This is followed by the Augmented Dickey Fuller (ADF) test and the Phillips Pheron (PP) test. A stationary time series has statistical values like mean and variance that are time invariant. It does not imply that the time series does not change its values. It simply means that the variation in these values doesn’t vary over time. A strictly stationary process

is termed as stochastic process where the unconditional joint probability distribution is time invariant. Differencing a time series is the most convenient method to convert a non-stationary time series into a stationary time series.

In the past, many researchers have attempted to recognize the traits in various existing time series from variable domains like atmospheric, economic, sales, market analysis, stock exchange and many more. Recognizing such traits helps in optimizing the existing algorithms and extends their performance.

3.2.2 Outsight based approach

Outsight based approach is the approach that helps us in correlating the study parameter to various other relevant parameters and provides a kind of dependency check. We all know that when we target to study about a parameter, then there may be various other parameters that may or may not be associated with it. Such kind of study always provides a knowledge that links them. Typical neural networks having characteristics of the regression type that is input-hidden-output feed forward neural network (FFNN) topologies are popularly investigated by researchers in the past. These are typically one input neuron, a sigmoidal function in the hidden layer and a single output neuron, controlled by a linear activation mapping function. However, various exogenous inputs can be fed in parallel. And the system response can be generated under these additional input conditions. This has two possible approaches depending upon the step size- the first one being a single step of prediction, the other being multistep. Single step prediction is very common but multistep prediction is challenging as the obtained results are to be carry forwarded to enable further predictions.

Auto Regressive models with eXternal inputs (ARX) models are known to operate using Exogenous inputs- inputs other than the parameter under consideration that support better prediction of the parameter under study. Alippi [66] has built a predictive motor control of a six joint robotic arm controller using the Hysteresis and the Coulomb friction as non-linear parameters. Angular velocity or position of his sixth joint is most optimally modeled by the ARX (3, 3) model using exogenous inputs. Non-linear Auto Regressive with Exogenous

inputs (NARX) Recurrent Neural Network owes its popularity to the feedback structure of the model. Various challenges involving real and simulated datasets are very popular for testing the capabilities of the proposed prediction algorithm. Some such popular challenges are the NN3 and NN5 challenges. Consider the NN3 database that contains an array of time series sequences for analysis and is a part of a competition database. Safeveih [33] has done a comparative of Non-linear regression with and without exogenous inputs for 11 such sequences and claims to have obtained a set of remarkable results.

With respect to the Oceanic studies, the Wind stress and Sea Level Pressure (SLP) are used as predictors for estimating the value of the Sea Surface Temperature [3]. The formation of Cyclones in the various water bodies have now become a very common occurrence. Ali et. al have used a NN based approach to build the vertical temperature profile of the Oceans with the help of SST, Wind Stress, net radiation, Sea Surface height and net heat flux [65]. Later, in 2013, Ali et.al have established a relationship between Cyclonic Intensities (obtained from Joint Typhoon Warning Centre) and the SST (obtained from Tropical Rainfall Measuring Mission Microwave Imager) [64].

As a part of study of SST and its association to other existing Sea parameters, it is planned the use of NARX technique to correlate them. The use of NARX technique is proposed in two manners- The first part is the single step or One Step Ahead Prediction.

One Step Ahead Prediction or Single Step Ahead Prediction

This involves the method of taking the help of previous inputs to predict the next SST value. Now in order to predict the next upcoming value, we again take the help of previous inputs and this continues for a set of readings. However, it is to be noted that all the previous values are original values. This is a very common approach undertaken by many researchers. And the results reveal good agreement between the original and predicted values. We have attempted single step predictions using various datasets

The Shore Stations Program at Scripps Pier is involved in collection of historical SST and SSS measurements providing coverage across the western shores of the United States [78]. [90] SST dataset is obtained via ships and buoys from the Scripps Pier at the coast of the

Western California from August 1916- October 2015 [78]. Using the site of La Jolla as a study zone, the effect of SSS and SBT on SST is explored in our work. A significant reduction is observed in the values of the error parameters- MSE, NMSE, RMSE and NRMSE when the assistive predictors (SSS and SBT) are used [90]. Even the prediction of extreme points becomes more viable under the support of such predictors. This task is covered in details in Section 4.3.1 and the results are discussed in 4.3.2.

The second part is the multistep prediction.

Multi Step Ahead Prediction or Multiple Step Ahead Prediction

This involves the method of taking the help of previous inputs to predict the next SST value. Once this value is calculated, it is stored and then again, the next value is predicted. For the prediction of this next value, the predicted answer is utilized and not the original value. This continues for a set of readings.

To achieve the target of multistep prediction, we chose the Elnino dataset. The International TOGA program developed the Tropical Atmosphere Ocean (TAO) array. This array consists of around 70 moored buoys distributed across the Equatorial belt around the Pacific Ocean providing coverage to vital El Niño, La Niña and ENSO sites. They measure various surface and subsurface oceanographic parameters including surface temperature, subsurface temperature (up to 0.5km beyond the surface), relative humidity, surface winds and air temperature [78]. This dataset [78] consists of - location of the buoy, date of measurement, Zonal Winds (consider for East>0; for West<0), air temperature, Meridional Winds (consider for North>0; for South<0), SST and relative humidity. The fluctuations in wind data is + 10m/s and the same with respect to relative humidity is 70%-90%. Variations in the values of the SST and Air Temperature are both restricted in the limit of 20 to 30 degree Celsius. It is ensured that all the readings are recorded at the same HH:MM:SS in the day. Using this daily SST dataset, we attempted to obtain the values of SST Anomaly (SSTA).

El Niño (warm) and La Niña (cool) events in the tropics of the Pacific are categorized using the standards by NOAA [77], known by the name of The Oceanic Niño Index (ONI). For a total of 3 months and more in continuation, a variation of $\pm$ 0.5°C in the values of SST

categorizes it either as a warm or cold event. Intensity is further classified based on the magnitude of the change in SST values. For Weak events, this is 0.5 °C to 0.9 °C. For moderate it is 1.0°C to 1.4°C. For Strong and Very Strong, the same is 1.5 °C to 1.9 °C and beyond 2°C respectively [77]. A part of this dataset is so identified that it has minimum missing values and yet is larger than a span of 10 years, which we thought is a duration correct enough to be considered for Time Series study. The location that could meet the above mentioned criteria is (0°N , -110 E). The duration covers the dates from 10th May, 1985 to 20th July 1995. We distributed our dataset into two sections- training (from 10th May, 1985 to 10th May, 1995) and testing (11th May, 1995 to 20th July, 1995) thus making the test dataset independent of the training dataset. Now we have a timeseries each of SST, Meridional Winds, Zonal Winds and Air temperature having the common timelines. Using a Non-linear Auto Regressive Network with Exogenous inputs, we have predicted the SST values and then computed SSTA with the help of different parameters and finally compared all the cases.

This task is covered in details in Section 4.4.1 and the results are discussed in 4.4.2

However, as we have opted for prediction in both the cases, this also involves a foresight based approach, that is discussed in the next section. Hence, although we investigate the influence of external parameters on SST values, it probably has mixed approaches involved in its implementation.

3.2.3 Foresight based approach

This is the most popular approach out of all the approaches. The prediction of the next value using the trends and patterns recognized in the time series is possible using various technologies.

The most basic technique is based on the Box Jenkins methodology and uses ACF and PACF plots to derive the nearest fit and describes the Auto Regressive Integrated Moving Average (ARIMA) [34, 43] model that is more specifically represented as ARIMA (p, d, q) model. It is a statistical approach based on Stochastic Correlation.

This is covered in details in section 4.1

The ARIMA algorithm is a statistic based method that explores the linear relationship amongst the parameters. However, they tend to fail to capture the non-linear characteristics of the sequence of events in time. As discussed in the previous chapter, where we have explored the capabilities and drawbacks of ARIMA in Section 2.3.1, it can depict the linear characteristics in a highly efficient manner. However, whenever the time series exhibit nonlinearities, the Neural Networks are known to provide better performance. The ANN [35, 36, 37, 38, 42] algorithms is a data driven approach and can gel with the time series characteristics provided it is trained optimally. However, it suffers from two issues: overfitting and underfitting

Overfitting is a phenomenon in which the network is so well trained to the training dataset that it fails to adapt to the test dataset (i.e. the model is not able to generalize its learning). To avoid overfitting, two very common techniques are suggested: (1) To segregate the dataset into training dataset and the testing dataset at the very beginning of the Experimental setup and perform train & test on different subsets of the data; hence building up a good estimate of network parameters. (2) To introduce a validating dataset, wherein the parameters learnt and weights acquired by the network during training remain unchanged. Once your network is ready with its learned parameters, it can be used for the task of validation.

Underfitting happens when the model is neither able to adapt to the training data nor can it model the test data. This is very easy to detect. The remedy involves in (1) Alter the parameter values and train and test till desired level of accuracy is achieved. (2) Move on with some other machine learning algorithms. This is covered in details in Section 4.2.

Long Short Term Networks are special type of deep learning networks. An LSTM network is an extension of the RNN with certain different characteristics, the most vital amongst them being its capability to memorize the dependencies with respect to a set of sequential data in between the predictions. This is particularly useful when part of the series is unknown to us. It also helps when we need to predict multiple number of steps, specifically on a long series. The constituent parts of an LSTM network are made up of a sequence input layer and an LSTM layer. A *sequence input layer* is the input node and connects the set of input terms or time series data into the network. An *LSTM layer* is the brain that trains itself to realize the

interdependency between sequential data time steps. The prediction of SST using LSTM networks is covered in Section 5.2.

As will be evident later, the methods discussed here can predict the monthly SST values with a fair degree of precision. A comprehensive analysis of the error components highlights that pure numerical value is not sufficient to judge the acceptability of the proposed methods. The analysis of the time series plot is equally essential to have a complete view of this. It is also noted that the monthly datasets can be predicted with a greater degree of ease. However, when it comes to displaying the forecasting capabilities for the Daily SST datasets, these models fall short of delivering the best [12,14]. Hybrid methods [16, 24, 25, 26, 45, 67] strike a balance between the linear and non-linear algorithms. A proposed hybrid method and its performance on prediction of various time series data is presented in Chapter 6.

3.3 Datasets

There are innumerable websites, archives and data portals that facilitate the availability of data for study and analysis to the researchers. The UCI Machine Learning Repository [79] is one very popular such website that has abundant datasets meant to meet the purpose of variable tasks like classification, regression, time series analysis etc. It is a vibrant and well structured repository. This is well acknowledged amongst the Researchers from the Machine Learning Community.

Next to it, there is the Time Series Data Library (TSDL) that is available at DataMarket.com [80] and created by Hyndman. It stores a set of various datasets based on different application sections. It consists of real and simulated datasets. Two of these datasets is taken as a basic dataset. Both of them belong to the category of Meteorology [75]. Of all the datasets that we are utilizing, one dataset is derived from Satellite data. There is also one dataset that is taken from the on-site data. SST datasets can be obtained by the different methods shown in Table 3.1.

Table 3.1 Sources of Ocean Observations.

	Type	Variables observed
In Situ	Global surface drifting buoy array	Sea Surface Temperature, Sea Surface Pressure
	Volunteer Observing Ship (VOS) fleet	Essential Climate Variables
Satellite based	Infrared (IR)	SST, Sea Ice
	Microwave radiance	SST, Wind Speed, Sea Ice
Sub Surface based	Ship based hydrography	All feasible variables
	Cabled ocean observatories	Chlorophyll, CO_2, Nitrates, Irradiance, Methane

Data buoys located near sea shores are used to collect data on wind, weather and waves. Drifters attached to 4 small floats are suspended below the water surface at variable heights to collect vital information like current, wind, pressure, temperature, ocean color, salinity and plankton population measurement.

The SST measurements are performed in the following manners. And the most common sensor being used is the reversing thermometer. Reversing Thermometers are attached to each bottle and the temperature is measured at different depths. Unlike conventional mercury thermometers, a reversing thermometer is able to record a given temperature to be viewed at a later time. If the thermometer is flipped upside down, the current temperature will be shown until it is turned upright again. Thermometers connected to the body of the Niskin and Nansen bottles and seal the value of the temperature measured. A Rosette composing of Conductivity, Temperature and Depth (CTD) and Nansen/Niskin bottles is used to sample sea water at different depths. Niskin and Nansen Bottles are the standard bottles used in measurement of the sample value. A CTD measures conductivity, temperature and depth that lead to determining essential physical properties. Conductivity is a measure of how easily an electric current pass through a sample of water. The higher the salinity of the sample, the easier it is to conduct electricity through the sample. Using CTD, pressure is calculated and then the Salinity based on this is further evaluated. The arrangement involves the collection of water samples at different depths. On pulling the thread, these bottles placed at different heights slams shut and collects the water at that depth. Some CTDs are known to have very high

efficiency by measuring conductivity, temperature and pressure as many as 24 times per second. The full instrument (CTD and Niskin bottles) is called a rosette. A CTD is typically deployed on a frame with several Niskin bottles for water sampling at different depths. A setup of Niskin bottle and a rosette containing a set of Niskin/Nansen bottles is shown in Fig. 3.1. A sonde looks similar to a CTD but is more complex and can also be used to measure pH, depth, temperature and turbidity.

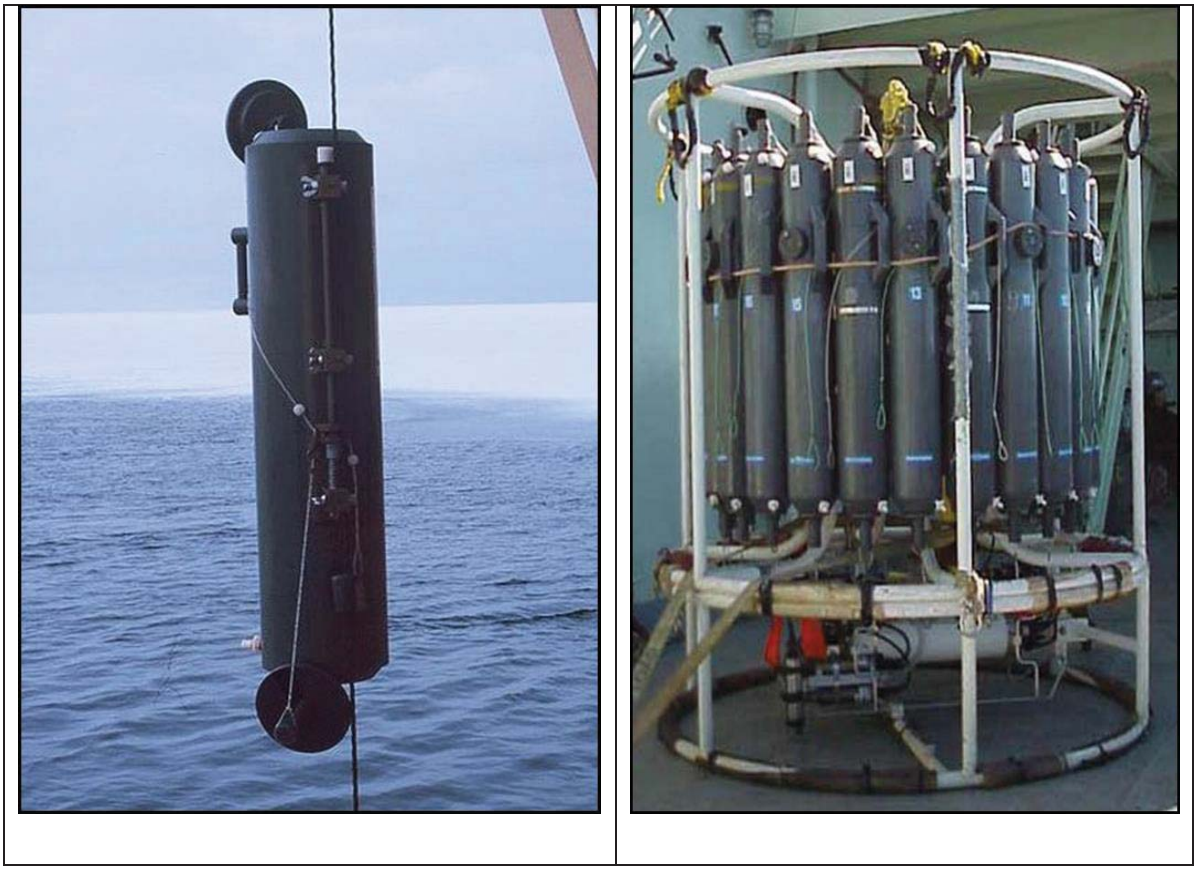


Fig 3.1 A Niskin bottle and a Rossete containing a set of Niskin/Nansen bottles [77].

The datasets used in the present research are mentioned as below.

Dataset 1: Melbourne Mean Temperature; R J Hyndman Library Dataset [75]
Parameters: Mean Temperature (Deg. F); Melbourne; 240 timestamps
Provider: Australian Bureau of Meteorology
Time range: Jan 1971 – Dec 1990

Dataset 2: Nottingham Castle Monthly Temperature; RJ Hyndman Library Dataset [75]
Parameters: Mean Temperature (Deg. F); Nottingham; 240 timestamps
Provider: O.D. Anderson
Time range: Jan 1920 – Dec 1939

The above mentioned two datasets are readily available with the UCI Machine Learning Repository and although it is not exactly SST values but this being a temperature dataset, it would suffice the initial requirements for training and testing the various algorithms as we explored various other test datasets. It is to be noted here that both the datasets are taken from different continents and from different time frames also.

Dataset 3: HadISST Mean Monthly SST; 6 locations around the Indian Ocean [76]
Parameters: Mean SST (Deg. C); Around Indian Ocean; 1791 timestamps
Provider: Marine Data Bank and ICOADS -International Comprehensive Ocean-Atmosphere Data Set merged with satellite-derived SSTs
Time range: January 1870 – April 2019

This dataset is obtained on request from a fellow Researcher (K. Patil) with due permission taken from his guide (Dr. M.C. Deo) of Doctoral studies. This dataset contains monthly observations. However, with this dataset we could demonstrate our tweaking skills, and show the optimization that can be done to the presently existing algorithms and how the same can be upgraded to provide better prediction.

Dataset 4: Elnino Daily SST; Equatorial Pacific Ocean [92]
Parameters: SST, Humidity, Air Temperature, Zonal winds, Meridional winds;
Equatorial Pacific Ocean; 5479 timestamps
Provider: Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric (NOAA) Administration, US Department of Commerce
Time range: March 1980 – July 1995

This is a Daily SST dataset and is obtained from Kaggle [92]. Although this dataset is relatively older, it projects the extreme happenings of El Niño, La Niña at critical locations in the Niño regions in and around the Pacific Ocean and other Regions near the Indian Subcontinents and hence forecasting the SST values for this dataset is always challenging. Predictions using Exogenous inputs were very feasible for this dataset and is depicted in more details in the chapters 4, 5 and 6.

Dataset 5: Scripps Pier Daily SST; Western Pacific Coast [78]
Parameters: SST, SSS, SBT;
Western Pacific Ocean; 36230 timestamps
Provider: University of California, San Diego
Time range: August 1916 – October 2015

This is an insitu database and provides a large variability in terms of location of occurrence. However, the major challenge in this database, as the occurrence of missing values. We have substituted all missing values with nearest values, in most cases the preceding or the following values.

For every prediction of datasets shown, we have taken the following points in consideration:

1] Algorithm/model is used for Single step prediction first with optimized parameters. Then the algorithm is modified to perform multistep prediction. Multistep prediction in the form of sliding window of twelve steps is calculated in an iterative manner. There are two reasons for selecting the multistep of 12. For the monthly dataset, a multistep of 12 months is a year long prediction and for the daily dataset, 12 is a number which is close to two weeks.

2] The size of the test dataset is approximately 10% of the size of the total dataset as per the general norms for Machine learning based prediction performances. All our datasets are of variable sizes and hence for testing results to be comparable, it is mandatory to have a fixed ratio. For example, consider the Daily SST dataset 3: the HadISST dataset that is derived by satellite links. This dataset is composed of 1791 timestamps. Hence to have a correct visual of the time series response, we performed 15 iterations of multistep predictions in

steps of 12, which accounts to the value of 180, which is almost 10 % of the original dataset size.

3] In order to evaluate the true representation of the prediction technique, average of each individual error parameter is calculated and then it is this average value that is quoted for final comparison.

3.4 Evaluation metrics

Once the model outputs are available, the next challenge is in recognizing the differences in the values in terms of error. The original value is known as true or actual value and the numerical value obtained by the process of prediction is designates as the predicted value. This comparison between the true value and the predicted value is called Error. There are many variants of Error available to address the visualization of the different interpretations. And hence we end up with a set of error metrics, each capable of having the capacity to uniquely address some statistical aspect of mathematical interpretation. Following error metrics are evaluated as a part of our study and they play a very significant role in interpretation of the success of the proposed algorithm.

Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_t - f_t)^2 \quad (1)$$

Normalized Mean Square Error (NMSE)

$$NMSE = \frac{MSE}{y_{max} - y_{min}} \quad (2)$$

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_t - f_t)^2} \quad (3)$$

Normalized Root Mean Square Error (NRMSE)

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (4)$$

Maximum Absolute Error (MAE)

$$MAX_{AE} = \max\big(abs(y_t - f_t)\big) \quad (5)$$

Mean Absolute Error (MeanAE)

$$Mean_{AE} = \text{mean}\ \big(abs(y_t - f_t)\big) \quad (6)$$

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}(abs((y_t - f_t)/y_t)) \quad (7)$$

Standard deviation of data (SDD),

$$SDD = std\ dev(DATA) \quad (8)$$

Standard deviation of error (SDE),

$$SDE = std\ dev\ (y_t - f_t) \quad (9)$$

Residual Standard deviation (Sres)

$$SRes = \sqrt{\sum_{i=1}^{m}(y_t - f_t)^2 \frac{1}{m-2}} \quad (10)$$

Nash Sutcliffe Efficiency (NSE)

$$NSE = 1 - \frac{\sum_{i=1}^{m}((y_t - f_t)^2)}{\sum_{i=1}^{m}((y_t - \mu)^2)} \quad (11)$$

Correlation Coefficient

$$CC = 100 - \frac{\sum_{i=1}^{m}\left((y_t - \mu).(y_f - \mu)\right)}{sqrt(\sum_{i=1}^{m}((y_t - \mu)^2))} \quad (12)$$

where y_t = actual data; f_t = forecasted data ; μ=mean of the data

y_{max} = maximum value of the dataset; y_{min} = minimum value of the dataset

n = total count of dataset and m = count of multistep; for one week m=7

Unlike classification techniques, where accuracy is the measure of correctness or near correct values, in regression/prediction models the metrics shown in Eq. (1)-Eq. (12) are generally used to provide an efficiency check of the proposed algorithms. In a way, they contribute to various possible dimensions of error visualization.

MSE and RMSE are the error parameters that are generally used for all kinds of error computations. They are respectively the quadratic and root quadratic that judges the average magnitude of error. RMSE due to its square component has the unique characteristics that it provides higher weight to large errors. Apparently, the same square factor ensures that the polarity of the difference is always additive. However, both these values are scale dependent. And hence their normalized counterparts, NMSE and NRMSE are also calculated to provide a range independent idea about the errors. They are a better representation of the goodness of fit.

MeanAE or MAE gives an average magnitude of the difference in actual and predicted values in a set of predictions, irrespective of their sign. The Maximum Absolute Error (MaxAE) is the maximum of the original difference between the actual and predicted values.

SDD and SDE are a statistical representation of the variation/spread of the data and its calculated error. In general, it is emphasized that if RMSE < SDD, then the functionality of the model in prediction is better than the mean value.

The Residual Standard Deviation (SRes) is the standard deviation of the residual values. It provides a numerical value to the distribution of the data points across the regression line/plane.

Nash Sutcliffe Efficiency (NSE) is a normalized statistic expression, specifically for hydrological models. It was proposed by Nash and Sutcliffe in the year 1970. The variance in terms of residue is compared to the variance in terms of data using this measure. It provides an indication of how well the predicted value versus the actual value fits the 1:1 line. For NSE = 0, the fit is as good as the mean. For NSE = 1, it indicates perfect fit of the predicted values to the observed values.

Correlation Coefficient (CC) is an indication of the strength of the relationship between the actual and the predicted values. A value near to 1 is a strongest match.

3.5 Comparison with Literatures

The detail result comparison (in tabular and description) of our proposed methods with existing literature methods cannot be represented here in full details as the subsequent chapters contains all relevant theory and optimization steps introduced. Hence, we have introduced the same in short.

3.5.1 Case 1: Comparison of Performance of Proposed Model with Tripathi, et al. [12]

The Indian Ocean Dipole in literature is termed to be Indian Ocean counterpart of the Pacific El Niño and La Niña. Different SSTs are reported in the eastern pole - somewhere in the south Indian Ocean and the western pole - that is in the Arabian Sea.
Tripathi, et al. [12], identified such an area in the Indian region (27S-35S, 96E-104E) that they claim to be having a high potential influence over the global climate.

Using 52 years (1950-2001) of data derived from Reynold's dataset of reconstructed SST, they have used twelve neural networks, one corresponding to each month of SST Anomaly data and with the help of corresponding time series presented a comparative with linear regression models using CC and RMSE as the prediction parameters

The Testing data is for the years 1997-2001.

The validating data is for 7 randomly selected years from the remaining pool of 45 years, which itself is a matter of concern as it is difficult to identify and segregate the training and the validating datasets. The Training data was for the remaining 38 years, again randomly distributed in the span from 1950-1997. Using a maximum lag of 2 years, a multivariate regression model was proposed.

Tripathi, et al. [13], in 2008, also tried to map the occurrence of Indian Summer Monsoon with the quarterly mean SST variations with details from four locations namely Central Southern Indian Ocean (CSIO; 22S-24S, 79E-81E) , Northwest of Australia (NWA; 14S-17S, 114E-116E), Southern Indian Ocean (SIO; 40S-41S, 82E-85E), Antarctic Circumpolar Current (ACC; 38S-42S, 64E-68E) using Artificial Neural network.
Based on the data, that we had collected, we could pick a location nearest to the sites as (9S-11S, 95E-98E).
Tripathi with Das and Sahai [12] have opted for twelve neural networks one for each month of the year.

We took the same dataset for analysis. We also selected the same years for training and testing. Instead of using 12 individual neural networks, we opted for a single Non linear Auto Regressive NAR network, which could be optimized using its inherent characteristics. Different sites require different tuning of the two major parameters- the number of neurons in the hidden layer and the number of delay elements so as to make the most accurate predictions. This is further addressed in Chapter 4.

3.5.2 Case 2: Comparison of Performance of Proposed Model with Mohongo and Deo [14]

Mohongo and Deo [14] identified a coastal site (EAF; 6S-7S, 39E-40E) located on the East African shore and another site (EQT; 0-1S, 59E-60E) that is located in the Indian Ocean. Using the monthly SST data derived from HadISST datasets from January 1870 to December 2011 (142 years), they have compared the performance of NARX, FFNN, RBFN, GRNN and the ARIMAX model for predicting the SST Anomalies using monthly and seasonal approach. We have identified the site (1S-1N, 65E). As suggested by Mohongo, Deo [14], the training dataset is data of the years 1874-1979; validating dataset for 1980 and the testing dataset is from 1981-2010 (30 years) and we used a NAR network NAR (12,18)
The proposed optimized network predicted values are very close to the actual values time series. To understand the influence of the error terms, we have utilized the same vital error indicators as used by them - Correlation Efficiency (CE), RMSE and MAPE.

Inspired by the good results, we attempted monthly multistep prediction (12 months-1st year, 24 months-2nd year, …..360 months- 20 years). Remember these are all monthly data, and once the Neural Network is trained with the optimum values of hidden neurons and delay/lag factor, it very efficiently mimics the pattern. This signifies the potential of the optimized multilayer perceptron model. Such models are pre-existent, but trained to their optimized configurations so as to provide the best predicted values. This is further addressed in Chapter 4.

3.5.3 Case 3: Comparison of Performance of Proposed Model with Patil, et al [15]

In 2013, Patil, et al. [15] have expanded this work into six different locations in the Indian Ocean vicinity. Their attempt is about Sea Surface Temperature SST value forecast using Neural Networks using 61 year data (January 1945 to December 2005) at six different locations around India over 1 to 12 months in advance. And the locations are

1] Arabian Sea with latitude and longitude values
(AS; 19N-20N, 68E),

2] Bay of Bengal with latitude and longitude values
(BOB; 18N-19N, 90E),

3] East of Indian Ocean with latitude and longitude values
(EEIO; 1S-1N, 90E),

4] West of Indian Ocean with latitude and longitude values
(WEIO; 1S-1N, 65E),

5] South of the Indian Ocean with latitude and longitude values
(SOUTHIO; 9S-11S, 95E-98E) and

6] Off the African Coast with latitude and longitude values
(THERMO; 14S-16S, 56E-58E).

61 years of data (January 1945 to December 2005) at six different locations around India over 1 to 12 months in advance is derived from HadISST. The time series is composed of 61 times 12 = 732 values. They had used previous 24 values (ie D=24) and H is not specified in the publication to predict the single step output.
To emphasize the goodness of fit of the proposed optimized algorithm, error analysis in terms of the parameters specified by them, using the same number of years for training, validating and testing dataset.

Hence, we may conclude that our proposed algorithms are optimized and have attained a near perfect fit for all SST values calculation at various different locations that are investigated by fellow researchers in the past.

3.6 Conclusion

This chapter addresses the theory and techniques involved in learning the background related to the SST datasets and how they are derived. The physics behind it is also depicted in details for understanding. This also depicts the approaches taken to study the SST time series in a holistic manner. A set of error metrics are identified and defined that are used to judge the efficiency of the study methods. This chapter also introduces the datasets used for analysis. Finally, we share the introductory challenges present in the existing literature study. Subsequent chapters shall be addressing them.

CHAPTER 4

Linear and Non linear Regression of Time Series and Its Application to SST Prediction

4.1 Introduction

SST is one of the most popular ocean parameters that is investigated by researchers all over the world due to its obvious linkage to other components that include the flora as well as fauna life of the world. In fact the entire global biome is connected to the oceans in a direct or indirect manner. Apart from SST, there are other ocean parameters that are recorded by transducers as well as by Satellite links like the Air Temperature at the interface, the Sea Level Pressure, the winds- the easterlies, the Westerlies, the zonal, the meridional, the salinity of the sea at the surface and also at the bottom of the sea. All these parameters are dependent on each other and contain signature of SST. As our study continued, it was realized that there are several aspects that could be addressed significantly. However the major constraint has been in obtaining the data, parallelly for all the parameters from the same location. This lead to the implementation of future predictions in the values of SST readings in terms of One Step Ahead (OSA) and Multi Step Ahead (MSA).

The chapter is divided into different sections as follows:
Section 4.1 is about the Introduction. Section 4.2 is about the Auto Regressive Integrated Moving Average (ARIMA) algorithm. Section 4.3 deals with the basics of the Nonlinear Auto Regressive Network (NAR) inputs. Section 4.4 address the single step analysis of the Daily SST dataset from Dataset 5 (using NARX-OSA) and an attempt is made to diagnose outlier or anomalous readings with the help of SSS and SBT readings. Section 4.5 addresses another Daily SST dataset from Dataset 4 (using NARX-MSA) that is using multistep prediction with the help of Air Temperature, Zonal Winds and Meridional Winds one at a time. Section 4.6 addresses the comparative with past literature and Section 4.7 summarizes the Chapter.

4.2 The Auto Regressive Integrated Moving Average (ARIMA) algorithms – A linear approach

The ARIMA is a statistical approach based on stochastic correlation. Every ARIMA model is represented as ARIMA (p, d, q), where p is the auto regressive (AR) component, d is the order of differentiation and q is the moving average (MA) component. The model components are analyzed using Box- Jenkins strategy. Here the information at any denoted time t is y_t, is taken under consideration as a function of the earlier p information values, say y_{t-1}, y_{t-2},… y_{t-p} and let, the errors at time t, t-1,….t-q say n_t, n_{t-1}…..n_{t-q}.

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots\dots\dots a_p y_{t-p} + n_t + b_1 n_{t-1} + b_2 n_{t-2} + \dots\dots\dots b_q n_{t-q} \quad [13]$$

where a_1 to a_p are the AR coefficients; evaluated using the PACF plots

b_1 to b_q are the MA coefficients; evaluated using the ACF plots.

Every correlation is defined as the measure of the resemblance of the time series to itself or to any other series. The numerical values of this measurement at variable lags give rise to the ACF and the PACF plots. For our understanding we have plotted this for the last 40 lags. Generally the first 20 lags are used to understand the response generated by the coefficients and how they affect the selection of the parameters p, q and d of the ARIMA model (p, q, d) we extended this range to an analysis of 40 lags to investigate the influence of the previous inputs, upto 40 steps beyond. We shall now see how predictions are made using these plots. For ease of understanding, all steps are implemented on standard datasets. Later, the same is propagated to all other datasets in accordance. These are broadly classified into Optimum parameter detection of the characterization of ARIMA *(p,d,q)* models, single step ahead implementation and multiple step implementations. The process of multistep prediction using AR and MA terms is presented in Algorithm 4.1 along with a calculation on example dataset-1.

Algorithm 4.1: Single step Prediction and Multistep prediction using ARIMA model (Auto Regressive (AR) and Moving Average (MA) terms):

Step 1: Consider the time series we want to predict.

Here, let we consider the Melbourne Mean Temperature (Dataset 1) as shown in Fig.4.1.

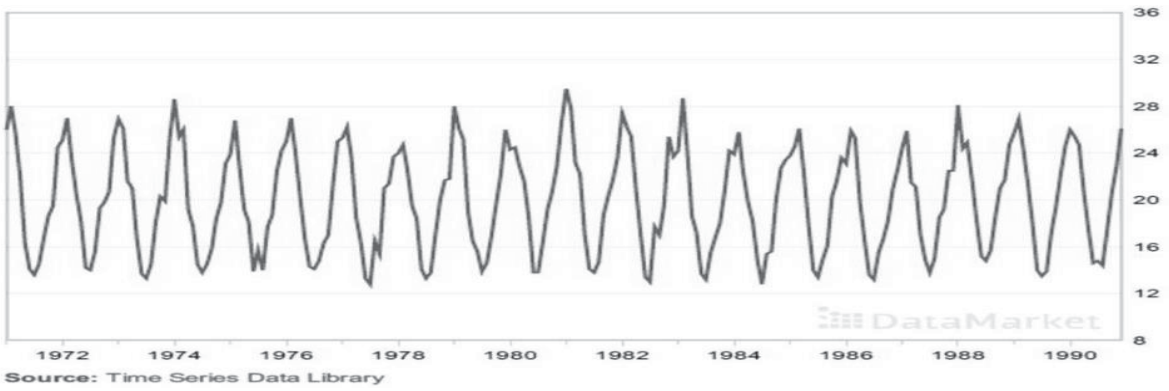


Fig 4.1 Time series plot of the Melbourne dataset [75]

Step 2: Identify if the given time series is stationary or not by observing the variations around its mean and trend assessment. A stationary time series does not require to be differenced.

Step 3: Plot the ACF and PACF components using either the inbuilt functions or by computing the same till the desired number of lags. Autocorrelation [81] is the correlations of the series with its own prior deviations from the mean and the plot represents these coefficients. For a stationary time series it is constant over time

Step 4: (a) Observe the ACF plot. Identify the first zero crossing. This is the MA (*q*) term.

For our dataset, ACF plot is shown in Fig. 4.2 in light color. The first zero crossing – i.e. MA*(q)* term comes at 3 which is indicated using pointed arrow from label ACF in the same figure.

(b) Observe the PACF plot. Identify the zero coincident point. This is the AR (*p*) term. For the present dataset, PACF plot is shown in Fig. 4.2 in light color. The first zero crossing – i.e.

MA(q) term comes at is 21 which is indicated in Fig.4.2 using pointed arrow from label PACF. Thus the identified ARIMA model is (21, 0, 3)

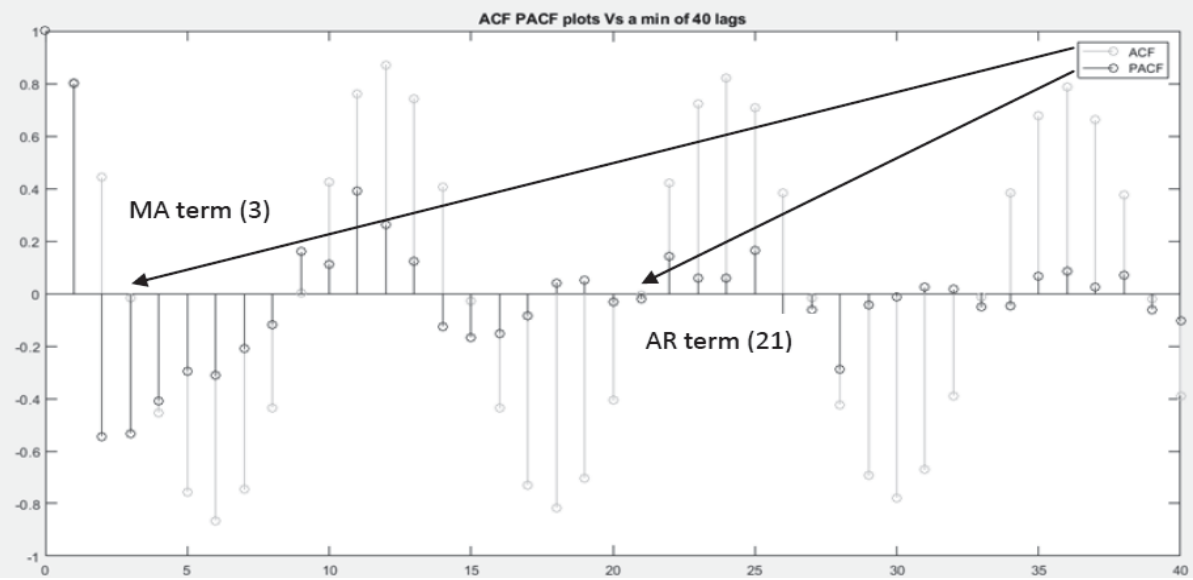


Fig 4.2 ACF and the PACF plots for the Melbourne city mean Monthly Temperature

Step 5: For the given time series, compute the step ahead values using ARIMA (21, 0, 3). The step ahead values can be calculated by using the equation 13 shown above.

This is One Step Ahead computation and as the name suggests, it is capable of predicting only the next value. However, for multistep prediction to be possible, it is vital to retain the values calculated and then opt for the next predicted using the predicted values and not the actual values. This is facilitated with the help of a closed loop network, wherein the predicted values are feedback to the input node and thus multistep calculations in the real sense gets implemented.

Step 6: Devise a closed loop to retain the forecasted values and store these to provide multistep prediction (as described in previous step-5).

In the present case it is done in steps of 12 each. For monthly SST datasets, 12 steps would make it a yearly forecast and for daily SST datasets, this is for a period of 12 days that is considered to approximate two weeks. The time frame of 12 steps is kept constant for all

monthly and daily databases. The result of multistep prediction on dataset-1 is shown in Fig.4.3.

In our experiment, SST test dataset is considered to be 10% of the complete dataset. As per algorithm 4.1, multistep prediction is performed on all other datasets (dataset-2 to dataset-5). The prediction result on test dataset for dataset-2 to dataset-5 is shown in Fig.4.4 to Fig.4.7 respectively. The AR(p) and MA(q) terms i.e. ARIMA(p,d,q) model obtained for dataset-2 to dataset-5 are shown in the captions of Fig.4.4 to Fig.4.7 respectively.

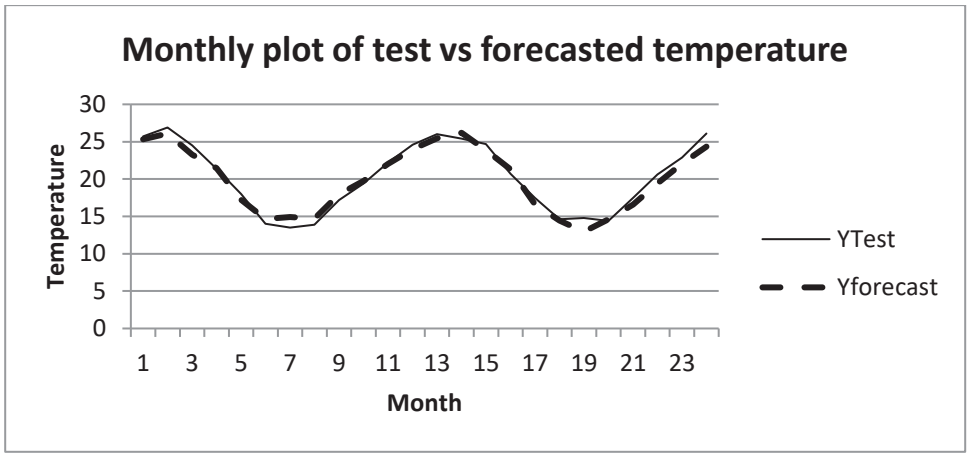


Fig 4.3 Time response using ARIMA (21, 0, 3) for Dataset 1

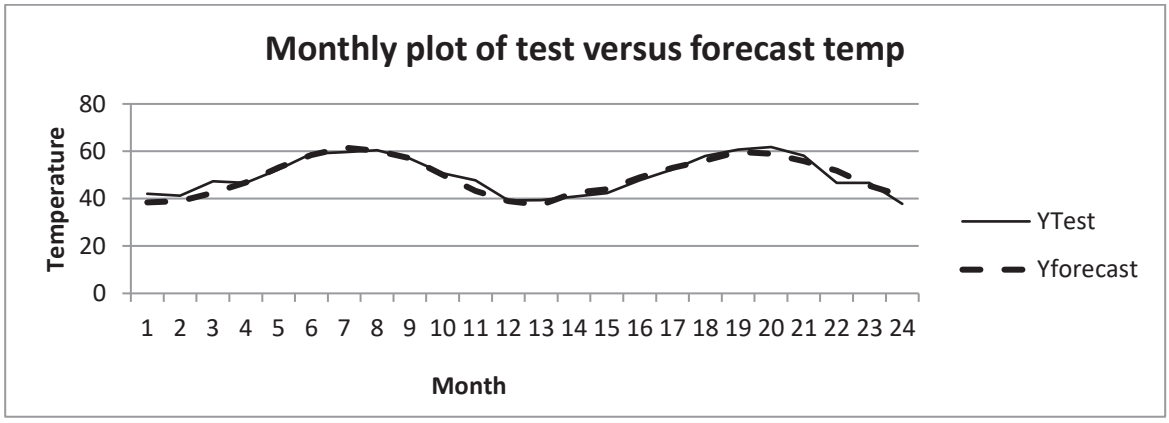


Fig 4.4 Time response using ARIMA (17, 0, 2) for Dataset 2

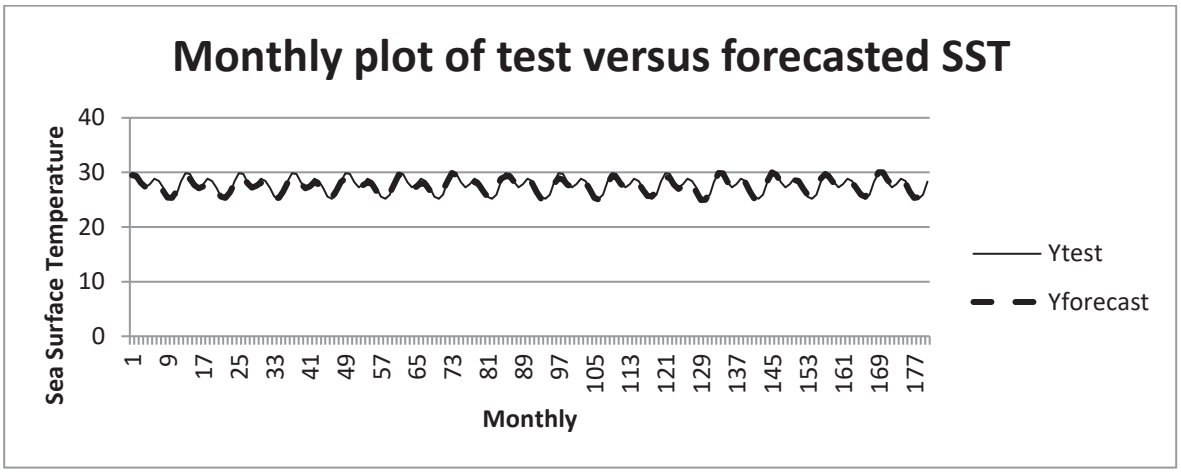


Fig 4.5 Time response using ARIMA (10, 0, 3) for Dataset 3

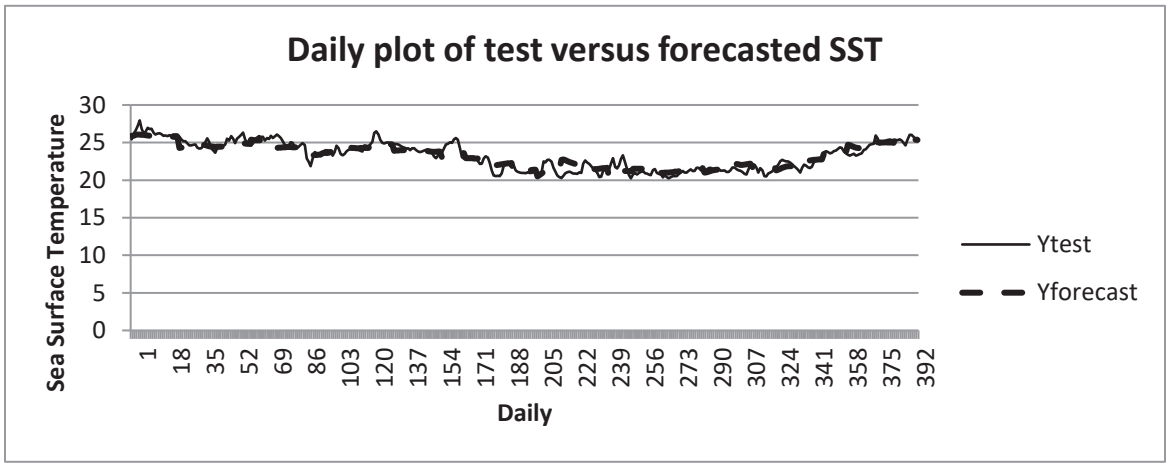


Fig 4.6 Time response using ARIMA (10, 0, 2) for Dataset 4

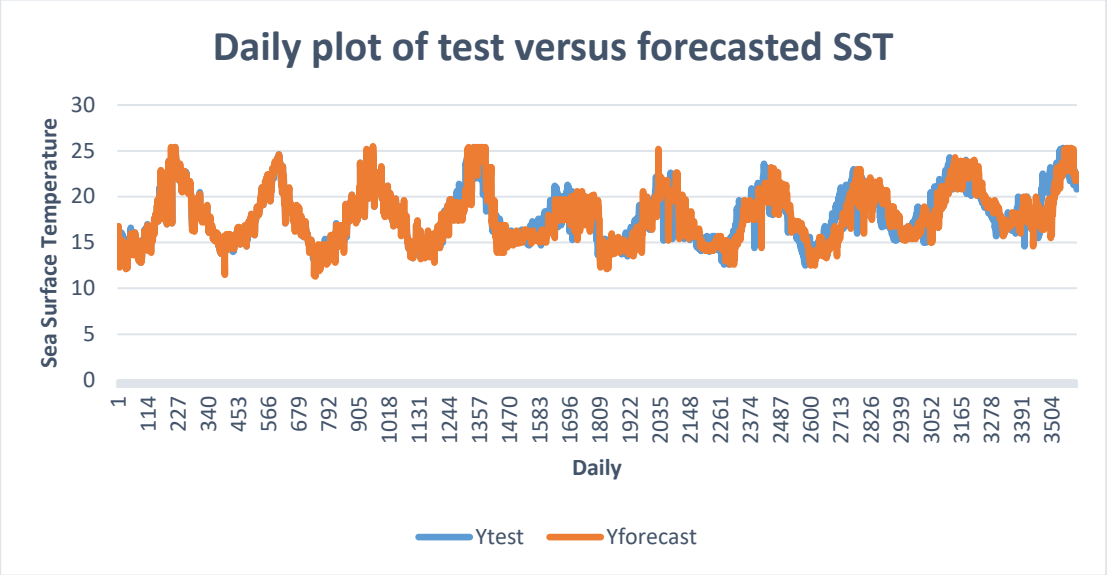


Fig 4.7 Time response using ARIMA (12, 0, 1) for Dataset 5

Table 4.1 Error performance comparative of ARIMA on all datasets

Dataset	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	Sres	CC	NSE
Dataset 1	**0.374**	**0.0224**	**0.6115**	**0.038**	1.6399	0.7713	0.0405	4.4224	0.8192	0.2414	**97.201**	**0.9601**
Dataset 2	0.702	0.0422	0.8378	0.050	4.9962	1.8562	0.0394	8.6198	2.3207	0.5276	93.303	0.9037
Dataset 3	0.636	0.0489	0.7974	0.062	0.8708	**0.3881**	**0.0140**	1.3930	**0.4077**	0.1642	93.538	0.9089
Dataset 4	1.294	0.1273	1.1375	0.1136	1.1823	0.5494	0.1500	**0.9930**	0.4837	**0.0997**	87.175	0.7472
Dataset 5	1.927	0.163	1.3882	0.1174	2.456	1.3891	0.0668	**0.9928**	0.8078	0.3666	84.883	0.7461

Various error measures (metrics) described in chapter-3 (section-3.4) are calculated for all five datasets (section-3.3) and the values of these error measures on test dataset (for each dataset) are shown in Table 4.1. As evident from the Table 4.1, all monthly datasets are predicted in a highly effective manner. This includes both single step and multistep ahead prediction for the test database. The best (minimum) RMSE value obtained is 0.374 °C for the 1st dataset (i.e. Melbourne Temperature dataset records) which shows the best performance among all datasets. Correspondingly notice the MSE also being registered the lowest. The Absolute error represented in terms of the Maximum Absolute Error and the Mean Absolute Error is lowest for the Dataset 3 as compared to the rest of the datasets. The standard deviation of the data (SDD) is the deviation of the values from the mean and the Standard deviation of the error (SDE) reflects the variation of the error of each reading. The size of test dataset is

kept same for all other methods investigated further in subsequent chapters. The SDE is always smaller than the SDD. The Residual Standard Deviation (SRes) plays a significant role in projecting the variation of the predicted values about the regression line. The Correlation Coefficient (CC) provides a statistical correlation between the actual and predicted values, and the more the value is near to 100, the better the prediction. Nash Sutcliffe Efficiency is most commonly associated with Hydrological studies and provides a fair estimation on the prediction skills. A value near to 1 indicates minimum error or a near match case. Notice that the overall values are larger than 0.9 for monthly data, and values near to 0.7 for daily data. This implies that investigation of other better methods is necessary for prediction of daily SST datasets.

4.3 The Non-linear Auto Regression (NAR) for SST Prediction

Non-linear Auto Regressive (NAR) algorithm is a time delayed feed forward neural network [61]. In this type of structure is useful for time series prediction. The future value of a time series y (t+1) can be predicted based on past values of the same series. This can be represented as follows:

$$y(t+1) = F(u(t), u(t-1), u(t-2) \dots \dots \dots u(t-D)) \quad (14)$$

where *u(t)*, *u(t-1)*, *u(t-2)*,….,*u(t-D)* are the values of time series at time *t*, *(t-1)*, *(t-2)*,…..,*(t-D)* respectively. *F* is a function that describes the nonlinear correlation between the model output *y* (t+1) and the input u(t) based on previous states from time *t*, *t-1* to *t-D* only. The errr weights of this network are tuned as governed by the Levenberg-Marquardt rule. This is shown in Fig 4.8.

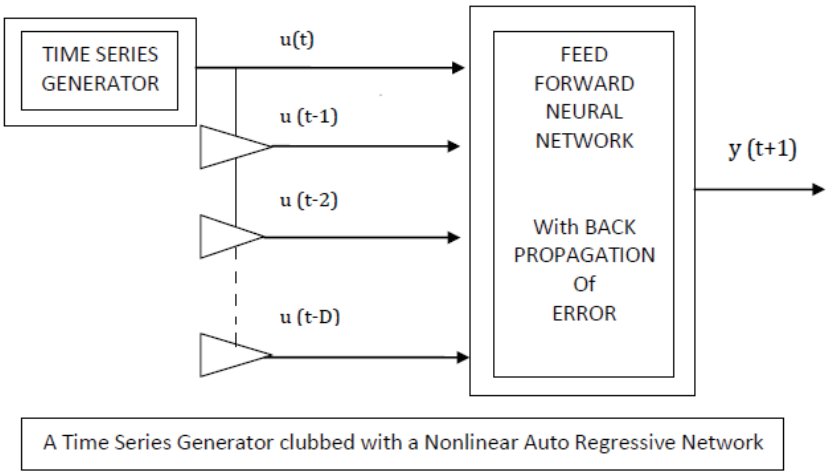


Fig. 4.8 Block diagram of the Time delayed Nonlinear Auto Regressive Neural Network for time series analysis

We attempted the Nonlinear Auto Regressive (NAR) NN model where a set of *D* time delay elements and the *H* number of hidden neurons are important parameters in predicting the future values. The model was used for single step prediction. Later, this is extended to multiple steps prediction for weekly (7 days) predictions and yearly (12 months) predictions. The structure of NAR for Single and multistep predictions is shown in Fig.4.10. The steps of SST prediction using NAR model for Single and multistep predictions is presented in Algorithm 4.2.

Algorithm 4.2: Single and Multistep Prediction using NAR for SST Prediction

Step 1: Consider the dataset for SST Prediction.

Here, the Melbourne Mean Temperature (Dataset 1) that we want to predict is considered and shown in Fig 4.1.

Step 2: Perform single step prediction using coarse values of *D* delay neurons and *H* hidden neurons. Initially coarse random values of Hidden neurons *(H*) and Delay neurons (*D)* are chosen and the algorithm is tested for error performance results based on the most common error parameters, RMSE, and its normalized value NRMSE (scale independent) measurements.

Table 4.2 Error comparative for Dataset 1 using coarse combinations of *D* and *H*

H= hidden neurons	D= delay neurons	RMSE	NRMSE
20	4	7.1891	0.4305
20	**10**	**4.1941**	**0.2511**
20	20	7.1095	0.4257
10	4	7.2708	0.4354
10	**10**	**4.3953**	**0.2632**
10	20	6.8675	0.4112

Step 3: Identify the optimum value for parameter *D*

Here ***D*=10** corresponds to minimum error and keeping it constant, vary *H.*

Table 4.3 Error comparative for Dataset 1 using fine combinations of *D* and *H*.

Evidently ***D*=10 (bold)** seems to be giving least error; keeping ***D*=10, vary *H***			
H	D	RMSE	NRMSE
1	10	4.1063	0.2459
2	10	4.2672	0.2555
3	10	4.2312	0.2534
4	10	4.3295	0.2593
5	10	4.3739	0.2619
6	10	4.1855	0.2506
7	10	4.2717	0.2558
8	10	4.1260	0.2471
9	10	4.1591	0.2490
10	10	4.3953	0.2632
11	10	4.4597	0.2670
12	10	4.3312	0.2594
13	10	4.2323	0.2534
14	10	4.5123	0.2702
15	10	4.2704	0.2557
16	10	4.2658	0.2554
17	10	4.3959	0.2632
~~**18**~~	~~**10**~~	~~**4.5532**~~	~~**0.2726**~~
19	**10**	**4.0792**	**0.2443**
20	10	4.1941	0.2511

NOTE: Maximum error is strikethrough and minimum error is bold ***H*=19** is minimum error

Step: 4 Optimize the second parameter while keeping the first constant.

Hence, keeping ***H*=19, vary *D***

Table 4.4 Error comparative for Dataset 1 for fine combinations of *D* and *H*, keep ***H*=19.**

Evidently ***D*=10 (bold)** seems to be giving least error; keeping ***D*=10, vary *H***			
H	D	RMSE	NRMSE
1	10	4.1063	0.2459
2	10	4.2672	0.2555
3	10	4.2312	0.2534
4	10	4.3295	0.2593
5	10	4.3739	0.2619
6	10	4.1855	0.2506
7	10	4.2717	0.2558
8	10	4.1260	0.2471
9	10	4.1591	0.2490
10	10	4.3953	0.2632
11	10	4.4597	0.2670
12	10	4.3312	0.2594
13	10	4.2323	0.2534
14	10	4.5123	0.2702
15	10	4.2704	0.2557
16	10	4.2658	0.2554
17	10	4.3959	0.2632
~~18~~	~~10~~	~~4.5532~~	~~0.2726~~
19	**10**	**4.0792**	**0.2443**
20	10	4.1941	0.2511

NOTE: Maximum error is strikethrough and **minimum error is bold.**

Finally, optimum values of lag (*D*) is 12 and Hidden (*H*) neurons is 19 for dataset 1;

NAR (*D, H*) as **NAR (12, 19).**

Step 5: Using the optimized parameters compute NAR (*D, H*), predict the step ahead values.

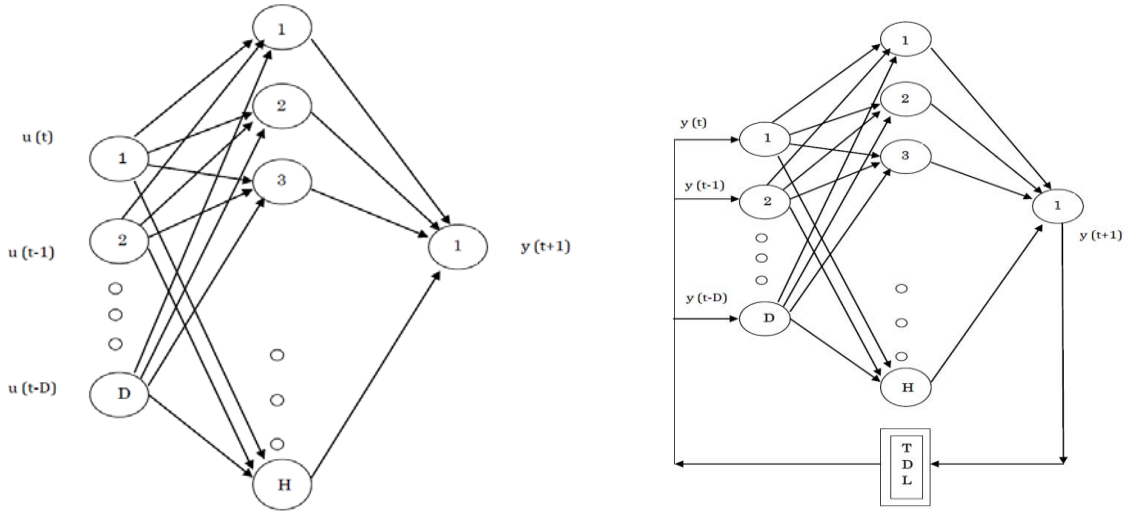
Here, the NAR(12,19) is used for step ahead prediction.

This is One Step Ahead computation and as the name suggests, it is capable of predicting only the next value. However, for multistep prediction to be possible, it is vital to retain the values calculated and then opt for the next predicted using the predicted values and not the actual values. This is facilitated with the help of a closed loop network, wherein the predicted values are feedback to the input node and thus multistep calculations in the real sense gets implemented.

Step 6: Devise a closed loop to retain the forecasted values and store these to provide multistep prediction. In the present case it is done in steps of 12 months.

In the present case it is done in steps of 12 each. For monthly SST datasets, 12 steps would make it a yearly forecast and for daily SST datasets, this is for a period of 12 days that is considered to approximate two weeks. The time frame of 12 steps is kept constant for all monthly and daily databases. The result of multistep prediction on dataset-1 is shown in Fig.4.12.

In our experiment, SST test dataset is considered to be 10% of the complete dataset. As per algorithm 4.2, multistep prediction is performed on all other datasets (dataset-2 to dataset-5). The prediction result on test dataset for dataset-2 to dataset-5 is shown in Fig.4.11 to Fig.4.14 respectively. The *D* and *H* terms i.e. NAR (*D, H*) model obtained for dataset-2 to dataset-5 are shown in the captions of Fig.4.11 to Fig.4.14 respectively. The optimized network is without feedback (in the left) and with feedback (in the right) is represented in the figure 4.9 where TDL is the time delay layer.



(a) (b)

Fig 4.9 NAR (D, H) networks for (a) Single step prediction (b) Multi step prediction

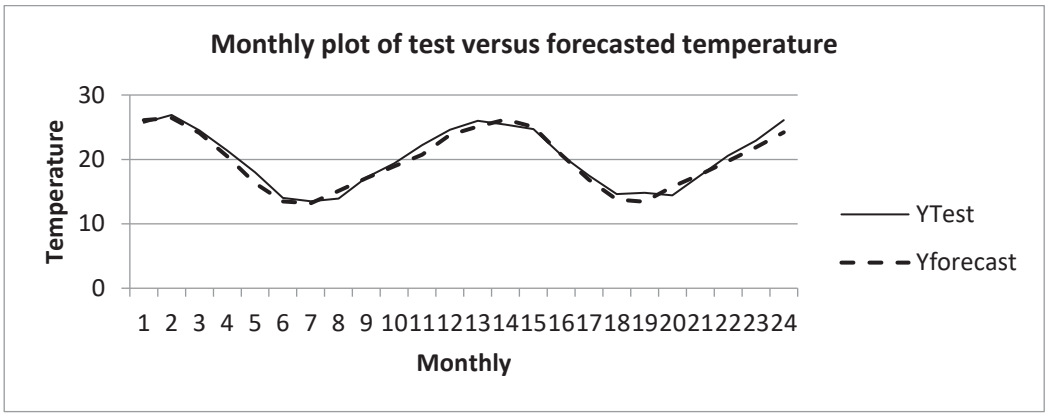


Fig 4.10 Time response using NAR (12, 19) for Dataset 1

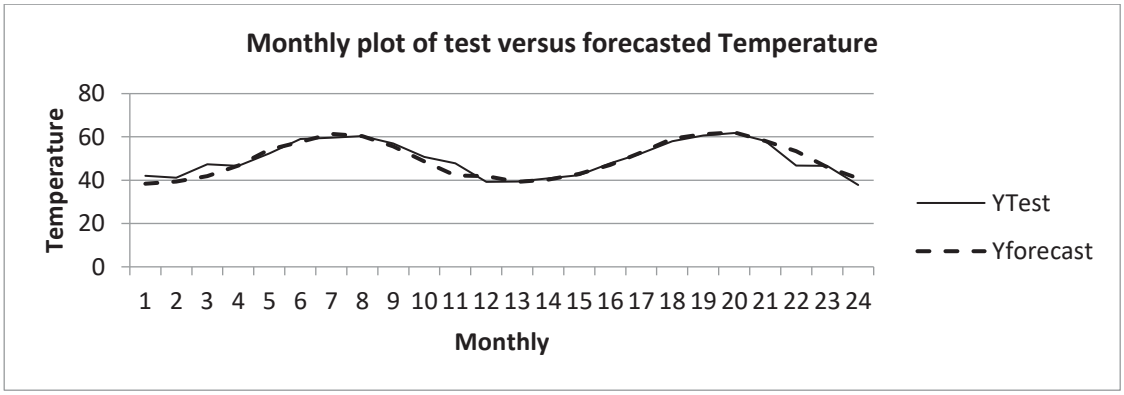


Fig 4.11 Time response using NAR (12, 5) for Dataset 2

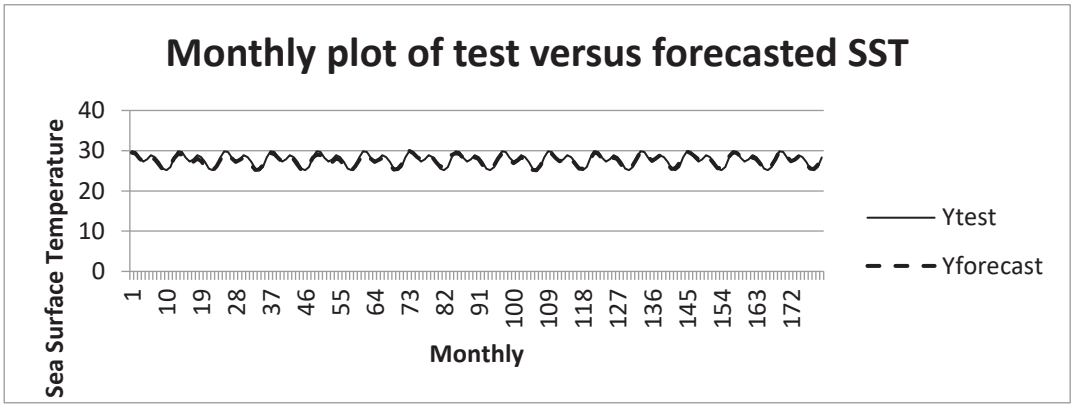


Fig 4.12 Time response using NAR (12, 18) for Dataset 3

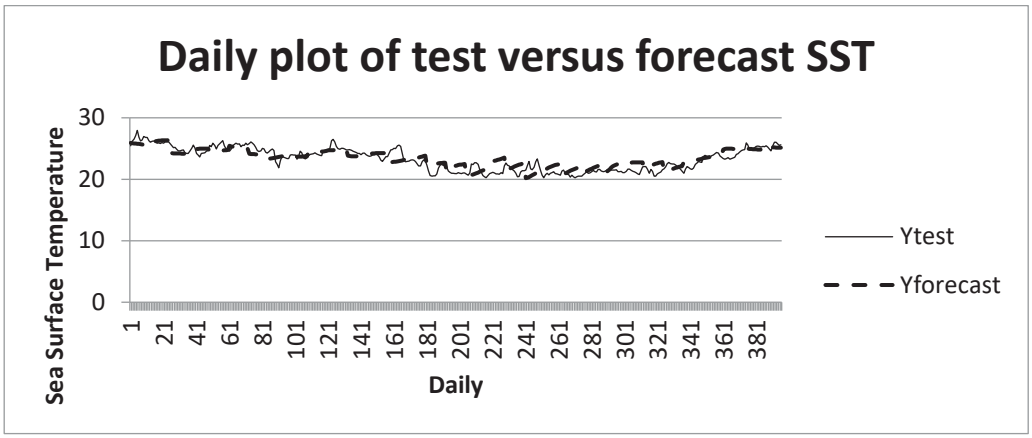


Fig 4.13 Time response using NAR (2, 20) for Dataset 4

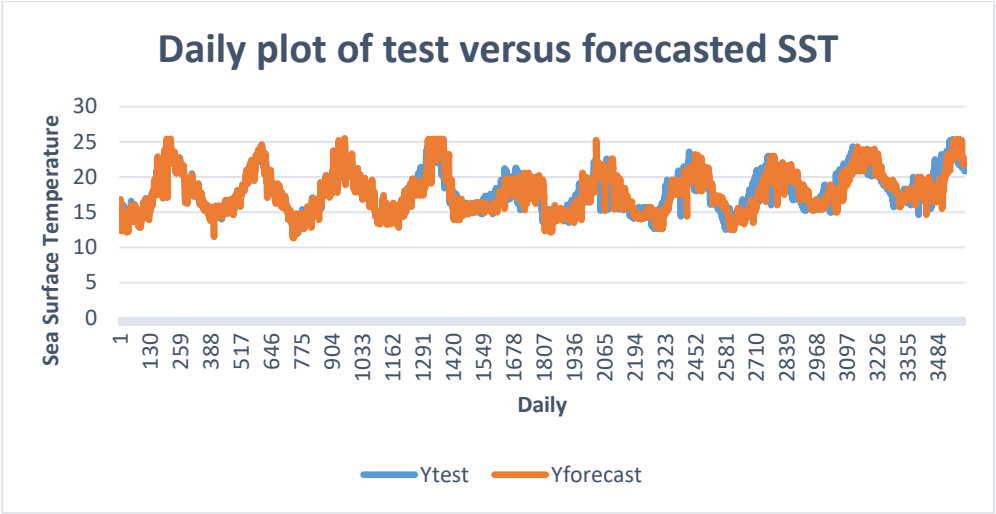


Fig 4.14 Time response using NAR (5, 5) for Dataset 5

Table 4.5 Error performance comparative of NAR on all datasets

Dataset	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	Sres	CC	NSE
Dataset 1	**0.289**	**0.0174**	**0.537**	**0.0323**	1.79	0.7848	0.0410	4.4224	0.8564	0.3321	**97.2292**	**0.9575**
Dataset 2	0.544	0.038	0.7375	0.0546	6.1923	1.7654	0.0379	8.6198	2.40155	0.8262	92.3156	0.8871
Dataset 3	0.455	0.035	0.6745	0.0518	**0.5956**	**0.3092**	**0.0113**	**1.3171**	**0.22896**	**0.0663**	96.5537	0.9135
Dataset 4	1.185	0.1187	1.0862	0.1081	1.7983	0.9787	0.1413	2.0190	0.5096	0.1848	86.7585	0.6985
Dataset 5	1.714	0.3709	1.3017	0.304	0.8898	0.4605	0.0206	2.029	0.7651	0.3771	84.8019	0.6911

Various error measures (metrics) described in chapter-3 (Section-3.4) are calculated for all five datasets (Section-3.3) and the values of these error measures on test dataset (for each dataset) are shown in Table 4.5. As evident from the Table 4.5, all monthly datasets are predicted in a highly effective manner. This includes both single step and multistep ahead prediction for the test database. RMSE values as low as 0.537 for the 1st dataset, ie for the Melbourne Time Series dataset records the best performance. Correspondingly notice the MSE also being registered the lowest. The Absolute error represented in terms of the Maximum Absolute Error and the Mean Absolute Error is lowest for the HadISST dataset. The standard deviation of the data (SDD) is the deviation of the values from the mean and the Standard deviation of the error (SDE) reflects the variation of the error of each reading. As the sample size gets larger the SDE would further decrease. And hence this is one more reason why the sample size of the test dataset is kept at a fixed 10% of the total dataset so as to avoid ambiguity in interpretation. Moreover, it is also kept of the same size for all the methods investigated. The SDE is always smaller than the SDD. The Residual Standard Deviation (SRes) plays a significant role in projecting the variation of the predicted values about the regression line. The Correlation Coefficient (CC) provides a statistical correlation between the actual and predicted values, and the more the value is near to 100, the better the prediction. Nash Sutcliffe Efficiency is most commonly associated with Hydrological studies and provides a fair estimation on the prediction skills. A value near to 1 indicates minimum error or a near match case. Notice the values near to 0.69 for dataset 4 that is a daily SST dataset encompassing many major climatic events in the vast expanse of the Pacific Ocean. The presence of sudden events is presently not so well captured using the Auto regressive network, and hence this can be further improved.

4.4 Non-linear Auto Regression with Exogenous Inputs (NARX)

Various techniques are used to analyze the different parameters. Many researchers have been incorporating different parameters into their study right from Winds to Pressure to Air Temperature for analysis. Consider the NN3 database that contains an array of time series sequences for analysis. Safeveih [33] has done a comparative of Non-linear regression with

and without exogenous inputs for 11 such sequences and claims to have obtained a set of remarkable results. Allippi and Piuri [66] have used the NARX model to identify time series in motor control and believe that nonlinearity involved in the system is better captured. The predictions of different ocean parameters like Sea Level Prediction (SLP) and temperature using Neural networks is found in [3] and [39].

A block diagram of an NARX network with an exogenous input is shown in Fig.4.15.

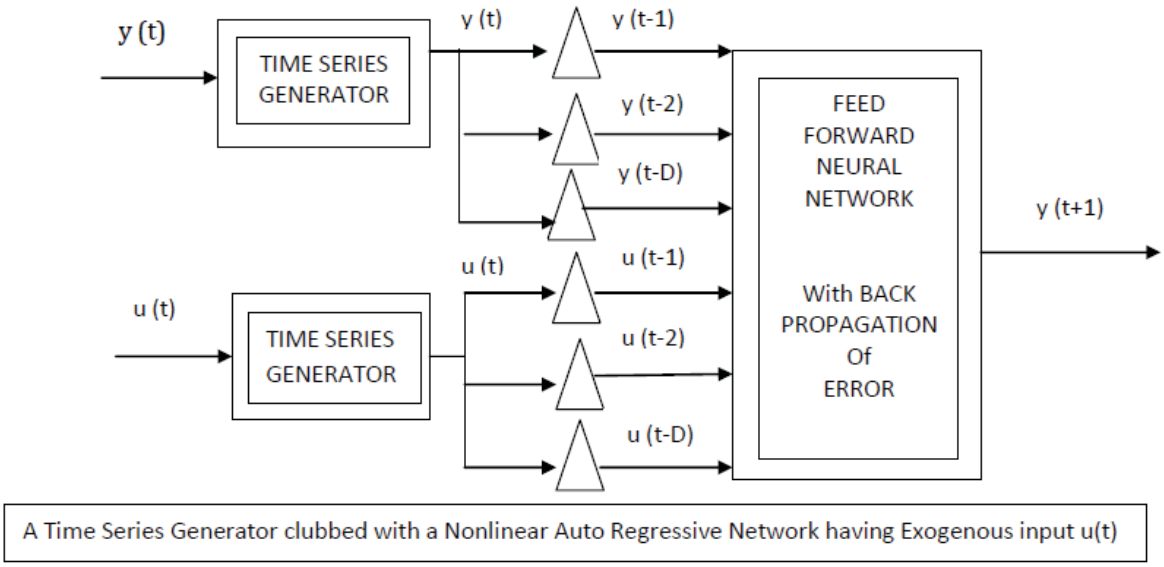


Fig 4.15 Block diagram of the NARX model for time series prediction with Exogenous Inputs

The nonlinear NARX can be mathematically represented as

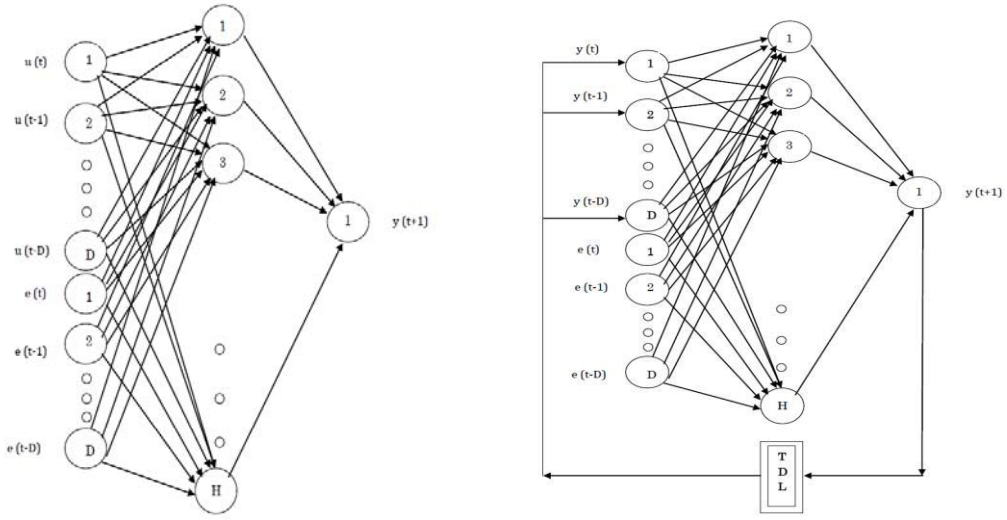
$$y(t+1) = G[u(t); e(t)] \qquad (15).$$

where the vectors *u(t)* and *e(t)* denote the input and exogenous time series, respectively. *G* function describes the nonlinear correlation between the model output *y(t+1)* and the inputs (*u(t)* and *e(t)*) based on previous states from time *t*, *t*-1 to *t-D*.

One step ahead computation as the name suggests, is capable of predicting only the next value. Many researchers have very successfully attempted and achieved notable success in this domain [4, 12, 14, 15, 24, 45]. However, for multistep prediction to be possible, it is vital to

retain the values calculated and then opt for the next predicted values using previous predicted values and not based on the actual values. This is facilitated with the help of a closed loop network, wherein the predicted values are feedback to the input node and thus multistep calculations in the real sense gets implemented. The NAR network when assisted by other supportive time series sequences is found to be providing a better prediction value as compared to the case where no exogenous input is provided. The following two experimental setups will further accentuate the same.

The NARX networks, without feedback and with feedback are shown in the Figure 4.16. Here *D* is the number of delay neurons and *H* is the number of Hidden neurons similar to NAR network. The component *D* is an indicator of the number of lag components that is needed so as to predict the next SST value and the component *H* is an indicator of the number of neurons needed in the hidden layer. The network essentially has two input nodes (each fed by a time series of SST and an assistive parameter- in this case it is the SBT and the SST Bottom, a hidden layer composed of *H* number of neurons, and an output node providing the predicted SST value.



(a) (b)

Figure 4.16 NARX (H, D) network for (a) Single step prediction, and (b) Multistep Prediction

For many years, researchers have been using the data collected by buoys for analysis [77]. Such data is used for analysis/study of flora and fauna, marine ecosystem and monitoring. We attempted the single step or one step ahead (OSA) daily SST prediction using NARX network

4.4.1 Datasets Used

The Shore Stations Program at Scripps Pier is involved in collection of historical SST and SSS measurements observed at the western coastline spread across the United States as shown in Fig.4.17. It provides access to current and previous data, subject to authentication as a Researcher [78].

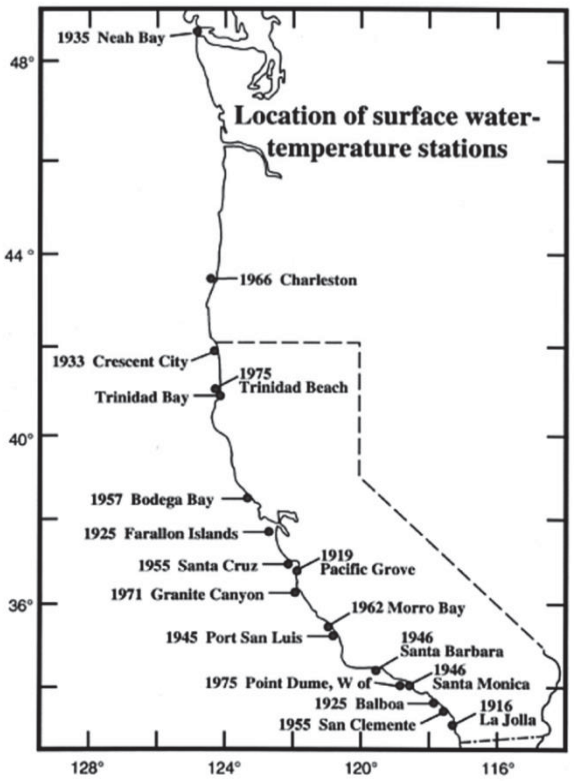


Fig 4.17 The stations on the west shore of California [78]

The SST dataset is obtained via ships and buoys from the Scripps Pier at the coast of the Western California from August 1916- October 2015. Temperature data are recorded using Glass mercury and analog thermometers and since 2008, digital thermometers are used. Inductive salinometer is used for salinity tests. It is calibrated before and after every use. The

surface measurements are done at a depth of 0.5 m. A digital thermometer is kept in a sampling insulated bucket which is inserted to the desired depth. For salinity, rinsing of the Niskin and Nansen bottles are done with sample water thrice and then actual measurement is performed. For near bottom measurements the depth is 5 m. The procedure remains the same however the depth is well around 5 m this time.

The California Parks and Recreation, Dept. of Boating and Waterways is handling the task of data collection. The Table 4.6 contains details about the dataset used for analysis. The measurements are made at variable depths-5m for all Bottom readings, ie the Sea Bottom Temperature (SBT). For all measurements of the surface readings, 0.5m is the depth at which both SST and SSS is calculated.

Table 4.6 Parameter Details of the dataset used [78]

Parameter	Depth	Dates
Salt at the Surface	Salt @ 0.5m Surface	08/22/1916- 10/31/2014
Temperature at the Surface	Temp @ 0.5m Surface	08/22/1916- 10/31/2015
Temperature at Bottom of the Sea	Temp @ 5m Bottom	07/21/1926- 10/31/2015

4.4.2 SST Prediction with Exogenous Inputs

Algorithm 4.3: SST Prediction using NARX with Exogenous Inputs

Step 1: Perform Single step prediction of this Daily SST data for parameter optimization Find the optimum values of delay neuron *D* and hidden neuron *H* using the NAR neural network discussed before.

Step 2: Using the SST dataset (with the optimized values of *D=5; H=5*), train the network with previous values and test the dataset. (Refer Fig. 4.17(a))

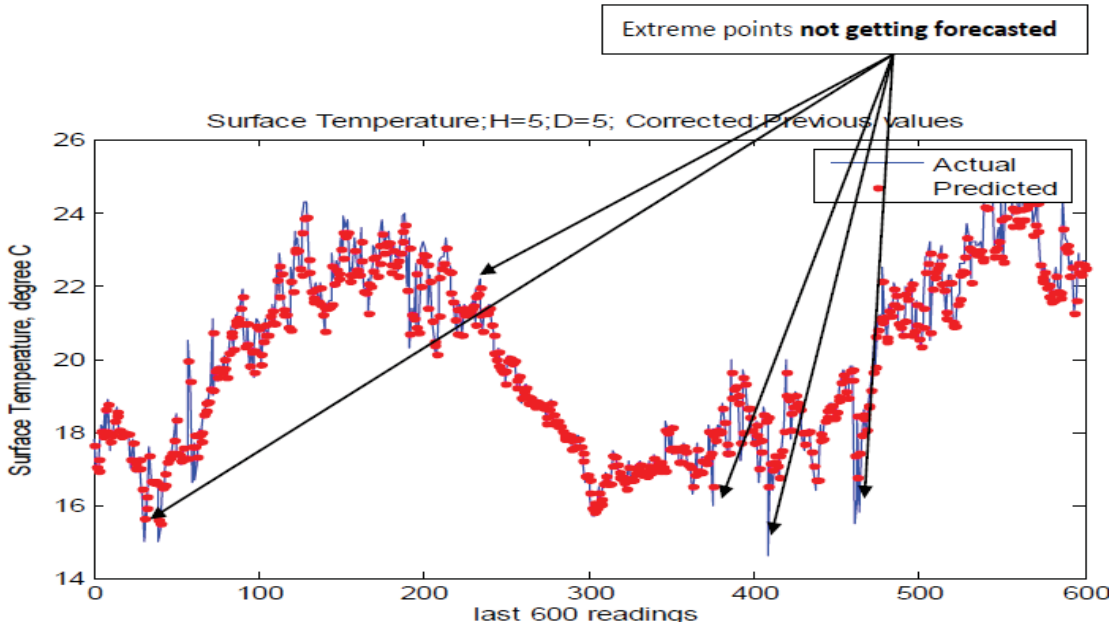


Fig. 4.18 Single step prediction of the SST dataset using NAR

Step 3: To this existing network, we now provide the additional time series as an exogenous input, Feed the additional time series having temperature readings from the bottom of the sea, SBT time series.

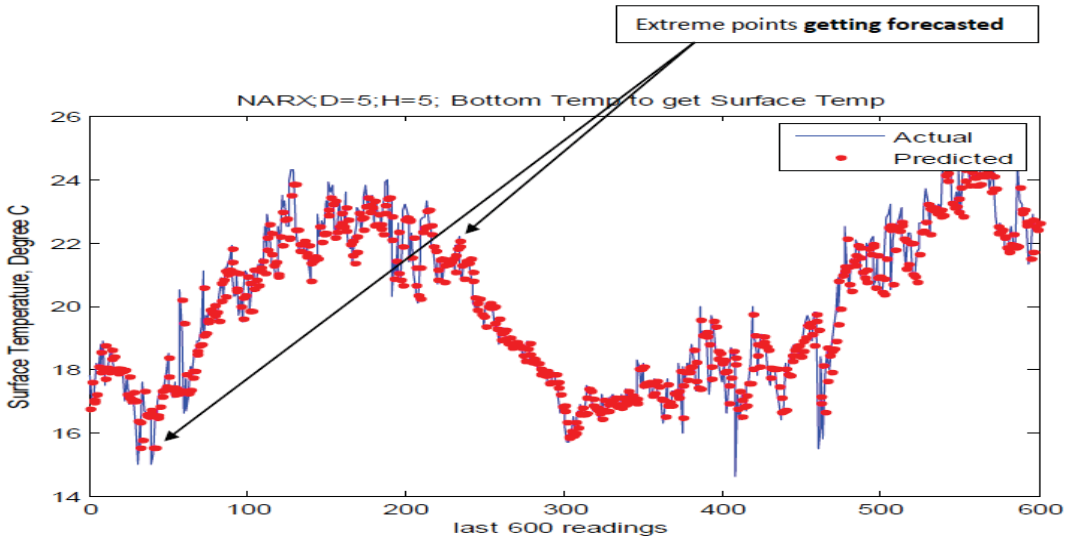


Fig 4.19 Few anomalous points getting detected

Comparing figure 4.18 and 4.19, it is visually observed that few points towards the extremes are getting detected. All such points generally correspond to those readings which have deviated from the regular trend and hence indicate the irregular patterns.

Step 4: To this existing network, we now provide the additional time series as an exogenous input, feed the additional timeseries having salinity readings from the surface-SSS time series.

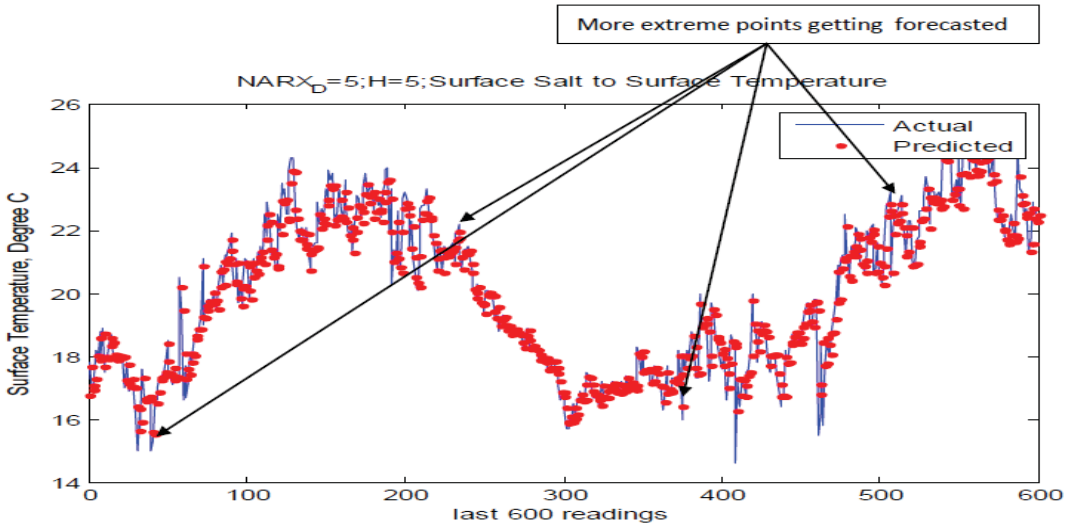


Fig 4.20 Few more anomalous points getting detected.

Table 4.7 Comparison of Error metrics for Prediction of SST with and without Exogenous Inputs – SBT and SSS for Dataset-5[78]

	SST Prediction using NAR	SST Prediction using NARX with SBT	SST Prediction using NARX with SSS
MSE	1.4315	0.3459	0.3462
NMSE	1.1964	0.5882	0.5884
RMSE	0.0912	0.0220	0.0221
NRMSE	0.0762	0.0375	0.0377

values, it is found that the additional parameter SST at the bottom (SBT) is supportive to the prediction of SST time series This is evident in the numerical values of MSE, RMSE, NMSE and NRMSE in the Table 4.7. However, SSS time series at the surface as an additional parameter towards predicting the value of SST timeseries is able to detect more extreme points

as compared to the NAR model and also better than the NARX model using SST Bottom Temperature as an assistive time series as evident from the time series plots. Notice the values of MSE and RMSE- we observe that without any additional input time series, the values are 1.4315°C and 1.1964 °C respectively. However, with either of the supportive time series, there is a drastic reduction in the MSE value; it drops down to 0.3459 °C and 0.3462°C using temperature at 5m and surface salinity as reference respectively. Similarly, the RMSE value is 0.0912°C. This is less than one half the same at 0.0220°C and 0.0221°C with SBT and SSS respectively. Their range independent counterparts NMSE and NRMSE also show identical responses. Hence time series plots portray the best picture of point to point variation. In order to avoid ambiguity of interpretation, we have thus ensured that for every analysis, the time series plot is accompanied by an error comparative for every implementation. This study is published as an IEEE publication [90]

4.5 Multistep SST Anomalies Prediction using NARX Network

Once we were able to utilize the single step analysis to conclude some meaningful content, we came across the Elnino dataset that was located around the Equatorial Pacific, a zone of immense interest to the Research community because of the famous El Niño and La Niña occurrences. Apart from SST, SSTA is the next significant parameter in the literature study [12, 14, 16, 28, 43, 86, 89, 90].

4.5.1 Datasets Used

Since decades, researchers have been studying this region and its vicinities to detect the correlation between significant Global events and the physical climate here. All the regions are shown in terms of their latitude and longitude locations in the Pacific Ocean in Fig 4.22 El Niño (warm) and La Niña (cool) events in the tropics of the Pacific are categorized using the standards by NOAA [77], known by the name of The Oceanic Niño Index (ONI). For a total of 3 months and more in continuation, a variation of $\pm$ 0.5°C in the values of SST

categorizes it either as a warm or cold event. Intensity is further classified based on the magnitude of the change in SST values. For Weak events, this is 0.5 °C to 0.9 °C. For moderate it is 1.0°C to 1.4°C. For Strong and Very Strong, the same is 1.5 °C to 1.9 °C and beyond 2°C respectively [77] (shown in the figure 4.21 in terms of increasing colour intensities). Table 4.8 provides a complete insight into the details of the different Niño sites, their locations, their characteristics and the most prominent feature of that region. This is based on the study conducted in the Pacific Ocean and its nearby vicinity.

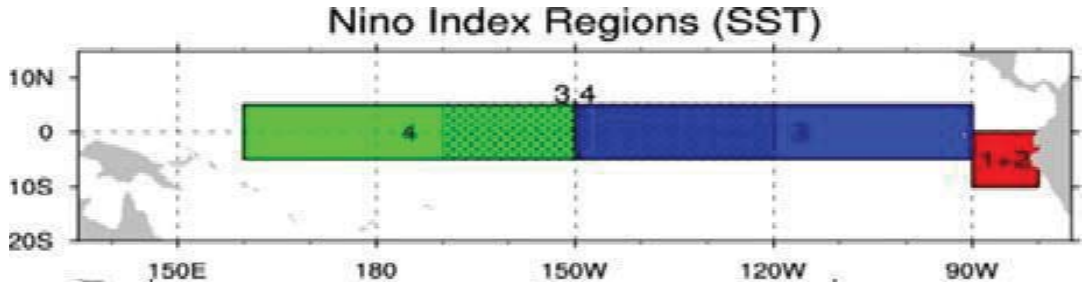


Fig 4.21 The figure shows the region outline on the Equatorial Pacific. [10]

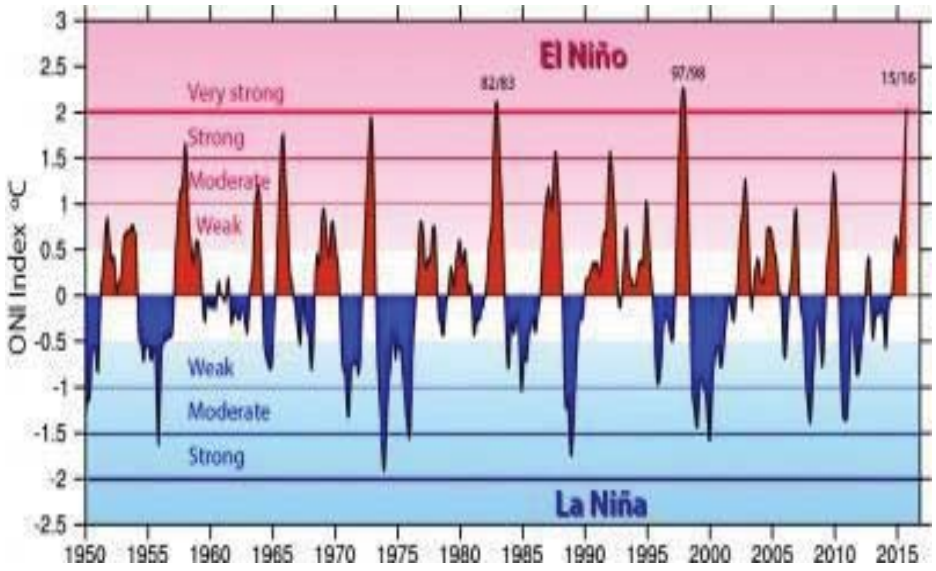


Fig 4.22 The Oceanic Niño Index (ONI) shows warm (red) and cold (blue) phases of abnormal sea surface temperatures in the tropical Pacific Ocean [11]

Table 4.8 The Niño indices with their locations and characteristics [10]

Type	Location	Characteristics	Remarks
Niño 1+2	0-10S, 90W-80W	Smallest region Most Eastern Niño SST regions	Largest variance of Niño SST
Niño 3	5N-5S, 150W-90W	Was primary focus for monitoring and predicting El Niño	Trenberth (1997) proved hat the ENSO lies further west
Niño 3.4	5N-5S, 170W- 120W	Represents the average Equatorial Pacific SSTs	5-month running mean, SSTAs ≥+ 0.4C for six months
ONI	5N-5S, 170W-120W	Operational definition used by NOAA; widely accepted.	3-month running mean, SSTAs ≥+ 0.5C for five months
Niño 4	5N-5S, 160E-150W	Captures SSTAs in the Central Equatorial Pacific	Less variance than the rest of the Niño regions

This is the dataset-4[92] from Section 3.3. The ownership of this data lies with PMEL laboratory of NOAA. A set of buoys, stationed by International TOGA's TAO program are installed for recording.

Donated by Dr. Di Cook, Department of Statistics, Iowa State University in June 1999, this dataset consists of 178080 instances and 12 attributes. It also contains some missing values.

This array consists of around 70 moored buoys distributed across the Equatorial belt around the Pacific Ocean providing coverage to vital El Niño, La Niña and ENSO sites. They measure various surface and subsurface oceanographic parameters including surface temperature, subsurface temperature (up to 0.5km beyond the surface), relative humidity, surface winds and air temperature [78]. This dataset [78] consists of - location of the buoy, date of measurement, Zonal Winds (consider for East>0; for West<0), air temperature, Meridional Winds (consider for North>0; for South<0), SST and relative humidity. The fluctuations in wind data is + 10m/s and the same with respect to relative humidity is 70%-90%. Variations in the values of the SST and Air Temperature are both restricted in the limit of 20 to 30 degree Celsius. It is ensured that all the readings are recorded at the same HH:MM:SS in the day. Using this daily SST dataset, we attempted to obtain the values of SST Anomaly (SSTA).

El Niño (warm) and La Niña (cool) events in the tropics of the Pacific are categorized using the standards by NOAA [77], known by the name of The Oceanic Niño Index (ONI). For a

total of 3 months and more in continuation, a variation of $\pm$ 0.5°C in the values of SST categorizes it either as a warm or cold event. Intensity is further classified based on the magnitude of the change in SST values. For Weak events, this is 0.5 °C to 0.9 °C. For moderate it is 1.0°C to 1.4°C. For Strong and Very Strong, the same is 1.5 °C to 1.9 °C and beyond 2°C respectively [77]. A part of this dataset is so identified that it has minimum missing values and yet is larger than a span of 10 years, which we thought is a duration correct enough to be considered for Time Series study. The location that could meet the above mentioned criteria is (0°N , -110 E). The duration covers the dates from 10th May, 1985 to 20th July 1995. We distributed our dataset into two sections- training (from 10th May, 1985 to 10th May, 1995) and testing (11th May, 1995 to 20th July, 1995) thus making the test dataset independent of the training dataset. Now we have a timeseries each of SST, Meridional Winds, Zonal Winds and Air temperature having the common timelines. Using a Non-linear Auto Regressive Network with Exogenous inputs, we have predicted the SST values and then computed SSTA with the help of different parameters and finally compared all the cases.

4.5.2 SSTA Prediction with Exogenous Inputs

A set of time stamped SST data is converted to time series data. The following steps are involved in the analysis.

Algorithm 4.3: SSTA Prediction using NARX with Exogenous Inputs

Step 1: The sea surface temperature time series is analyzed against different possible values of Delay and Hidden neurons in an open loop configuration using a nonlinear auto regressive network for single step prediction. (Refer Fig 4.16 (a))

The optimized values of *D* (number of delay neurons) and *H* (number of hidden neurons) obtained during this prediction are 2 and 20 respectively.

Step 2: Using these optimum values of D and H, closed loop architecture is implemented (Refer fig 4.16 (b)) to retain predicted values and continue the forecasting for 7 days of SSTA.

In the present case, last ten years of SST data is fed to the Neural Network as the training dataset. The testing dataset is a total of 70 days, one week each at a time. Results show 10 iterations compiled one after the other forming a time series representation of 70 timestamps of SST data.

Step 3: Using the optimum values of D and H, closed loop architecture is implemented (Refer fig 4.16 (b)) to retain predicted values and continue the forecasting for 7 days of SSTA. In the present case, last ten years of SST data and Air Temperature data is fed to the Neural Network as the training dataset. The testing dataset is a total of 70 days, one week each at a time. Results show 10 iterations compiled one after the other forming a time series representation of 70 timestamps of SST data.

Step 4: Using the optimum values of D and H, closed loop architecture is implemented (Refer fig 4.16 (b)) to retain predicted values and continue the forecasting for 7 days of SSTA. In the present case, last ten years of SST data and Zonal wind data is fed to the Neural Network as the training dataset. The testing dataset is a total of 70 days, one week each at a time. Results show 10 iterations compiled one after the other forming a time series representation of 70 timestamps of SST data.

Step 5: Using the optimum values of D and H, closed loop architecture is implemented (Refer fig 4.16 right) to retain predicted values and continue the forecasting for 7 days of SSTA. In the present case, last ten years of SST data and Meridional wind data is fed to the Neural Network as the training dataset. The testing dataset is a total of 70 days, one week each at a time. Results show 10 iterations compiled one after the other forming a time series representation of 70 timestamps of SST data.

Using a 10 year dataset of the equatorial Pacific region (dataset-X) in a site specific approach, the proposed NARX model is used to forecast the next week SST values. The prediction of SSTA independently as well as with different exogenous inputs is abbreviated as below. SSTA is calculated by taking the difference of mean SST values from the actual SST values [1, 2, 4, 12, 13, 14].

SSTA_NAR is the SSTA prediction using NAR. SSTA_AirT - SSTA prediction using the Air Temperature data as an exogenous input. SSTA_Z - SSTA prediction using the Zonal Winds as an exogenous input. SSTA_M - SSTA prediction using the Meridional winds as the exogenous input.

The multistep (a week ahead) prediction of SSTA using NAR and SSTA using NARX with Zonal winds, Meridional wind and Air Temperature are shown in common Fig.4.24. The performance of SSTA prediction with different error metrics – MSE, NMSE, RMSE, NRMSE, MaxAE, MeanAe, MAPE, SDD, SDE, SRes (as defined in Section 3.4) are calculated for SSTA using NAR, SSTA using NARX with Air Temperature, Zonal winds and Meridional winds for 10 weeks as test data are presented in Table 4.9, Table 4.10, Table 4.11, Table 4.12 respectively.

Table 4.9 The week (W) wise details of error components using NAR multistep-SSTA_NAR

Week	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	SRes
1	0.1388	0.0117	0.3726	0.0315	0.7518	0.2872	0.012	2.0136	0.389	0.1972
2	0.1502	0.0127	0.3876	0.0328	0.6352	0.3331	0.0141	2.0135	0.3801	0.2051
3	0.0647	0.0055	0.2544	0.0215	0.4976	0.2023	0.0084	2.0135	0.2745	0.1346
4	0.8155	0.069	0.903	0.0764	1.3796	0.8216	0.0328	2.0134	0.8626	0.4778
5	0.0487	0.0041	0.2206	0.0187	0.3495	0.2069	0.0083	2.0146	0.1763	0.1167
6	0.0508	0.0043	0.2254	0.0191	0.418	0.1943	0.0079	2.015	0.2311	0.1193
7	0.0369	0.0031	0.1921	0.0163	0.3084	0.1744	0.0073	2.017	0.207	0.1016
8	0.1009	0.0085	0.3176	0.0269	0.6354	0.2663	0.0113	2.0137	0.3392	0.168
9	0.6445	0.0545	0.8028	0.0679	1.3429	0.6661	0.0275	2.0128	0.7299	0.4248
10	0.846	0.0716	0.9198	0.0778	2.1496	0.6136	0.0257	2.0127	0.9011	0.4867

Table 4.10 Error component using NARX multistep prediction Air temperature-SSTA_AirT

Week	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	SRes
1	0.3829	0.032	0.6188	0.0521	1.1558	0.545	0.0231	2.0136	0.3918	0.3274
2	0.5146	0.0435	0.7174	0.0607	1.1055	0.6191	0.0263	2.0135	0.3914	0.3796
3	0.1422	0.0119	0.377	0.0319	0.6115	0.3349	0.0139	2.0135	0.2767	0.1995
4	1.58	0.1337	1.257	0.1063	2.0783	0.9847	0.0378	2.0134	0.8439	0.6651
5	0.3701	0.0313	0.6083	0.0515	0.9968	0.5774	0.0231	2.0146	0.2069	0.3219
6	0.0777	0.0066	0.2788	0.0236	0.4216	0.2385	0.0097	2.015	0.2557	0.1475
7	0.1851	0.0157	0.4302	0.0364	0.6978	0.3774	0.0158	2.017	0.2232	0.2277
8	0.4986	0.0422	0.7061	0.0597	1.3287	0.6291	0.0266	2.0137	0.3464	0.3737
9	0.6732	0.057	0.8205	0.0694	1.4532	0.7027	0.0299	2.0128	0.702	0.4342
10	0.8724	0.0738	0.934	0.079	1.3896	0.8616	0.0348	2.0127	0.9129	0.4942

Table 4.11 Error components using NARX multistep prediction Zonal winds -SSTA_Z

Week	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	SRes
1	0.3521	0.0298	0.5934	0.0502	1.1439	0.5011	0.0212	2.0136	0.3833	0.314
2	0.5009	0.0424	0.7078	0.0599	1.0422	0.628	0.0266	2.0135	0.3526	0.3745
3	0.1428	0.0121	0.3779	0.032	0.6782	0.3042	0.0127	2.0135	0.2765	0.2
4	1.4944	0.1264	1.2512	0.1034	2.0118	0.9835	0.0379	2.0134	0.7842	0.6469
5	0.412	0.0349	0.6419	0.0543	1.0528	0.6026	0.0241	2.0146	0.2389	0.3396
6	0.1113	0.0094	0.3336	0.0282	0.5034	0.2897	0.0118	2.015	0.3084	0.1765
7	0.1862	0.0158	0.4315	0.0365	0.7618	0.3518	0.0147	2.0147	0.2699	0.2283
8	0.497	0.042	0.705	0.0596	1.391	0.6028	0.0256	2.0137	0.3949	0.373
9	0.592	0.0501	0.7694	0.0651	1.3971	0.6516	0.0277	2.0128	0.6548	0.4071
10	0.9673	0.0818	0.9835	0.0832	1.4504	0.9128	0.0369	2.0127	0.9581	0.5204

Table 4.12 Error components using NARX multistep prediction Meridional winds -SSTA_M

Week	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	SRes
1	0.5189	0.0439	0.7203	0.0609	1.2998	0.6207	0.0262	2.0136	0.3948	0.3812
2	0.556	0.047	0.7456	0.0631	1.1213	0.6599	0.028	2.0135	0.375	0.3945
3	0.1605	0.0136	0.4006	0.0339	0.651	0.3547	0.0148	2.0135	0.2744	0.212
4	1.5094	0.1277	1.2286	0.1039	2.0399	0.9521	0.0366	2.0134	0.8387	0.6501
5	0.2624	0.0222	0.5122	0.0433	0.963	0.4577	0.0183	2.0146	0.2484	0.2711
6	0.0767	0.0065	0.2769	0.0234	0.3962	0.2464	0.01	2.015	0.2976	0.1465
7	0.3037	0.0257	0.5511	0.0466	0.8668	0.4967	0.0207	2.0147	0.2578	0.2916
8	0.6785	0.0574	0.8237	0.0697	1.4975	0.7485	0.0317	2.0137	0.3715	0.4359
9	0.7569	0.064	0.87	0.0736	1.487	0.7292	0.0311	2.0128	0.6592	0.4604
10	0.8288	0.0701	0.9104	0.077	1.5577	0.835	0.0339	2.0127	0.936	0.4817

Consider Table 4.9, 4.10, 4.11 and 4.12. The term Week indicates the week number and this multistep prediction is continued for ten iterations with the objective to establish the correlation of the other factors (like Air Temperature, Zonal Wind, Meridional wind) with SSTA. This table contains the magnitude of error parameters computed for the case when the SST time series itself is fed to the Non-linear Auto Regressive network NAR Network without any supportive time series first. Observe the fact that for all cases, the RMSE<SDD and hence the proposed algorithm is a better predictor than mean. The RMSE actually signifies the concentration of data points in the vicinity of the regression line. The smaller its value the better fit it is. However, point to point comparison between the actual and predicted SSTA is equally needed to ensure that the prediction is correct. NRMSE is in the range of (0.0163^0C - 0.0778^0C), which indicates that the predicted average value over the week is almost a near match to the actual value. Max AE and Mean AE are indicative of spurious errors averaged over a week with respect to the actual value with the help of the predicting algorithm. The value of the standard deviation of error is 0.1075^0C - 0.4842^0C. With the Air Temperature as an additive time series, the RMSE value is 0.2788^0C -0.934^0C. The same with Zonal winds is 0.3336^0C – $1.2535\ ^0$C and under the effect of Meridional winds is 0.2769^0C to 1.2286^0C. It

is to be noted that the time series plot shows max agreement of the actual SSTA with the support of both Zonal and Meridional winds.

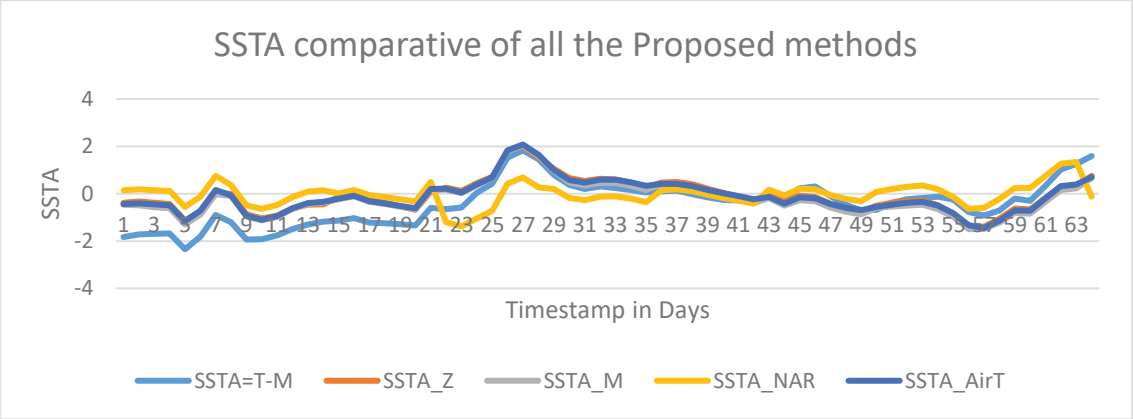


Fig. 4.23 SSTA comparative for 10 iterations, each of multistep value of 7; total of 70 time step index.

Fig 4.23 shows a comparative of all the proposed techniques. We initiated with the calculation of the values of SSTA for our test dataset by subtracting the mean SST value from the actual SST value. This is indicated in the figure as SSTA. This is the actual SSTA. Then we performed time series prediction using NAR algorithm with optimized parameters; and does not include the influence of any other additional time series sequence. This is depicted as SSTA_NAR in the figure 4.25. Next, we evaluate the value of SSTA under the influence of the Zonal winds by providing the Zonal wind time series as an additional input to the previously optimized NAR network. We call this SSTA_Z. Similarly, we have obtained SSTA_M with the Meridional winds time series as the exogenous input for SST calculations. Similarly, we have obtained SSTA_AirT with the Air Temperature time series as the exogenous input for SST calculations.

This work is contributed as a Book chapter [91] by the authors in the Springer Nature Series.

4.6 Comparison with Literatures

4.6.1 Case 1: Comparison of Performance of Proposed Model with Tripathi, et al. [12]

The Indian Ocean Dipole in literature is termed to be Indian Ocean counterpart of the Pacific El Niño and La Niña. Different SSTs are reported in the eastern pole - somewhere in the south Indian Ocean and the western pole - that is in the Arabian Sea.

Tripathi, et al. [12], identified such an area in the Indian region (27S-35S, 96E-104E) that they claim to be having a high potential influence over the global climate.

Using 52 years (1950-2001) of data derived from Reynold's dataset of reconstructed SST, they have used twelve neural networks, one corresponding to each month of SST Anomaly data and with the help of corresponding time series presented a comparative with linear regression models using CC and RMSE is the prediction parameters (shown in Table 4.13).

The Testing data is for the years 1997-2001.

The validating data is for 7 randomly selected years from the remaining pool of 45 years, which itself is a matter of concern as it is difficult to identify and segregate the training and the validating datasets. The Training data was for the remaining 38 years, again randomly distributed in the span from 1950-1997. Using a maximum lag of 2 years, a multivariate regression model was proposed.

Tripathi, et al. [13], in 2008, also tried to map the occurrence of Indian Summer Monsoon with the quarterly mean SST variations with details from four locations namely Central Southern Indian Ocean (CSIO; 22S-24S, 79E-81E) , Northwest of Australia (NWA; 14S-17S, 114E-116E), Southern Indian Ocean (SIO; 40S-41S, 82E-85E), Antarctic Circumpolar Current (ACC; 38S-42S, 64E-68E) using Artificial Neural network. Based on the data, that we had collected, we could pick a location nearest to the sites as (9S-11S, 95E-98E).

Tripathi with Das and Sahai [12] have opted for twelve neural networks one for each month of the year and an error comparative of the testing year is available in Table 4.13. They had distributed their data in the following manner.

Table 4.13: ANN error measures for test cases K.C. Tripathi, I.M.L. Das, A. K. Sahai [12]

Test cases of the years 1997, 1998, 1999, 2000 and 2001 [12]						
Method proposed by K.C. Tripathi [12]				**Our proposed method**		
Month	**Year**	**CC**	**RMSE**	**MM:YYYY**	**CC**	**RMSE**
Jan	2001	**0.97**	0.32	01:2001	0.9618	**0.0597**
Feb	2001	0.95	0.16	02:2001	**0.961**	**0.0677**
Mar	2001	0.59	**0.25**	03:2001	**0.9601**	0.2656
Apr	2001	0.76	0.13	04:2001	**0.959**	**0.104**
May	2001	0.90	0.26	05:2001	**0.9601**	**0.2334**
Jun	2001	0.92	0.10	06:2001	**0.959**	**0.0269**
Jul	2001	0.77	0.18	07:2001	**0.9358**	**0.0366**
Aug	2001	0.89	**0.08**	08:2001	0.8701	0.1029
Sep	2001	0.87	0.10	09:2001	**0.8927**	**0.1163**
Oct	2001	0.69	0.31	10:2001	**0.9969**	**0.1594**
Nov	2001	0.76	0.20	11:2001	**0.9929**	**0.0152**
Dec	2001	0.99	0.31	12:2001	**1**	**0.0719**

As per Table 4.13, the value of Correlation Coefficient (CC) is minimum, 0.59 for the month of March as per their algorithm and based on our proposed the same is 0.9601. We obtained minimum CC for the month of August at a value of 0.8701 which is far better than their minimum. Also observe that all values of CC are in the range of (0.87 to 1.00) as per our proposed algorithm. For RMSE values, they are getting a maximum error of 0.32ºC for the month of January and a minimum of 0.10ºC for the months of June and September. The minimum RMSE we are getting is at 0.0152ºC for the month of November and the maximum is at 0.2656ºC for the month of March. However, all variations in the values of RMSE for our proposed algorithm are restricted in the range of (0.0152ºC – 0.2656ºC) against their range that is (0.10ºC – 0.32ºC). Surprisingly, they [12] have not shared a time series plot of SST readings which could provide a real picture. On the contrary, observed anomaly is compared to the output anomaly and plotted for various months using linear regression and ANN techniques.

Table 4.14 ANN performance measures using our proposed model NAR (12,18) results

Months	CC	RMSE
12 steps - 1st year Jan1997-Dec1997	0.97	0.2562
24 steps-2nd year_Jan1997-Dec1998	0.96	0.2722
36 steps- 3rd year_ Jan1997-Dec1999	0.96	0.2564
48 steps- 4th year_ Jan1997-Dec2000	0.95	0.2578
60 steps- 5th year_ Jan1997-Dec2001	0.95	0.2571

Inspired by the results in the preceding table for monthly SST prediction, we then attempted multistep predictions using monthly SST data (12 months- 1st year, 24 months- 2nd year….. 360 months-20th year). Results are displayed in the Table 4.14. Evidently, the network is able to predict the values to a large degree as the numerical values of CC are in the 0.95 to 0.97 range. RMSE provides an estimate on the deviation of the residuals form the regression line data points. These are in the range of 0.2562ºC – 0.2722ºC for multistep regression.

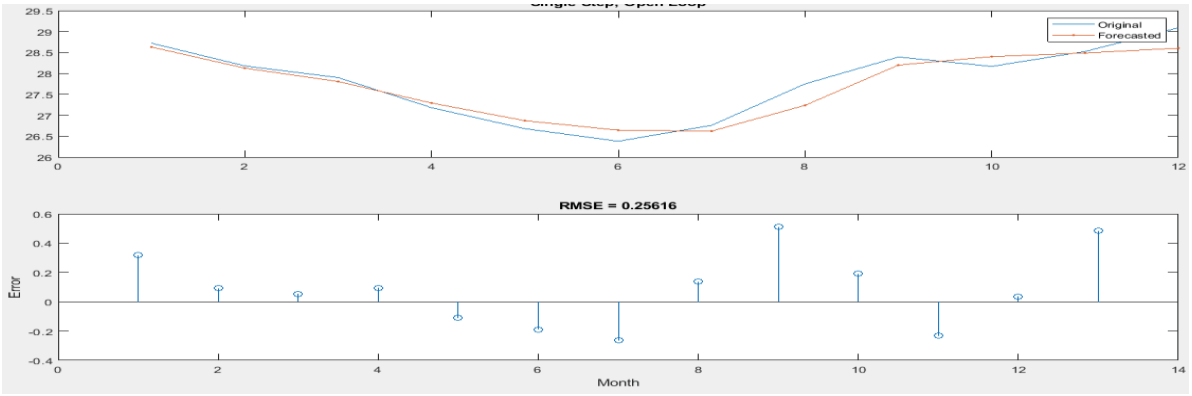


Fig 4.24 The time series plot of 1 year using NAR (12, 18)

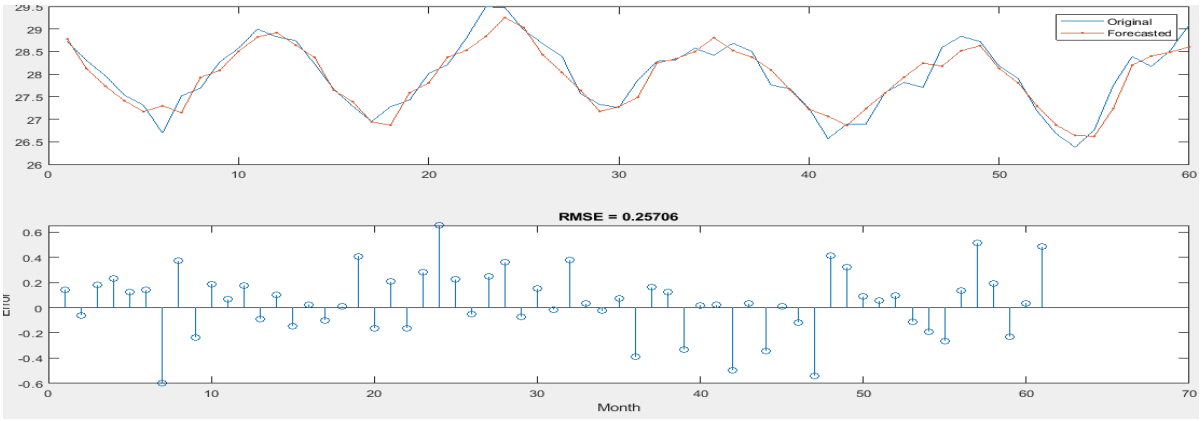


Fig 4.25 The time series plot for 5 years, ie, 60 months, using NAR (12, 18)

4.6.2 Case 2: Comparison of Performance of Proposed Model withMohongo and Deo [14]

Mohongo and Deo [14] identified a coastal site (EAF; 6S-7S, 39E-40E) located on the East African shore and another site (EQT; 0-1S, 59E-60E) that is located in the Indian Ocean (Fig 4.26). Using the monthly SST data derived from HadISST datasets from January 1870 to December 2011 (142 years), they have compared the performance of NARX, FFNN, RBFN, GRNN and the ARIMAX model for predicting the SST Anomalies using monthly and seasonal approach.

We have identified the site (1S-1N, 65E). As suggested by Mohongo, Deo [14], the training dataset is data of the years 1874-1979; validating dataset for 1980 and the testing dataset is from 1981-2010 (30 years) and we used a NAR network NAR (12,18)

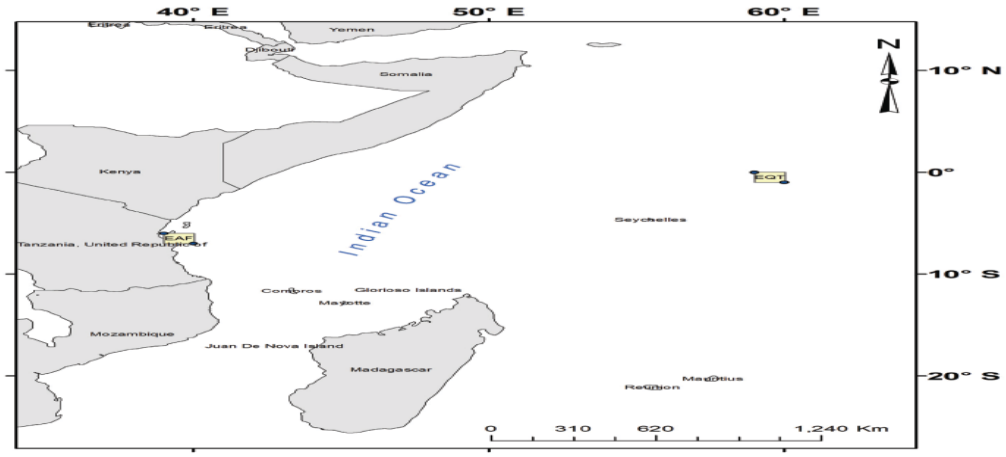


Fig 4.26 Locating the EAF and EQT on the map

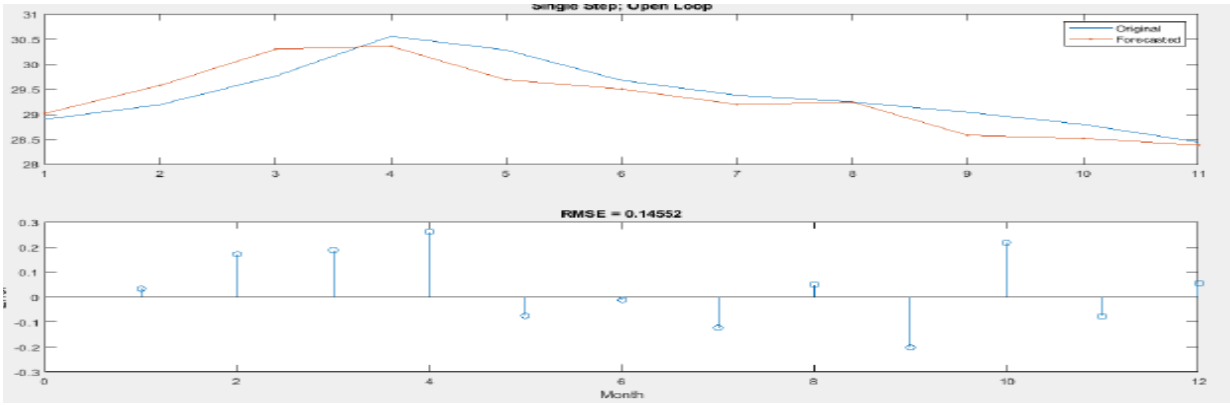


Fig 4.27 Time series plot of (1S-1N, 65E) near to EQT for 1 year using NAR (12, 18)

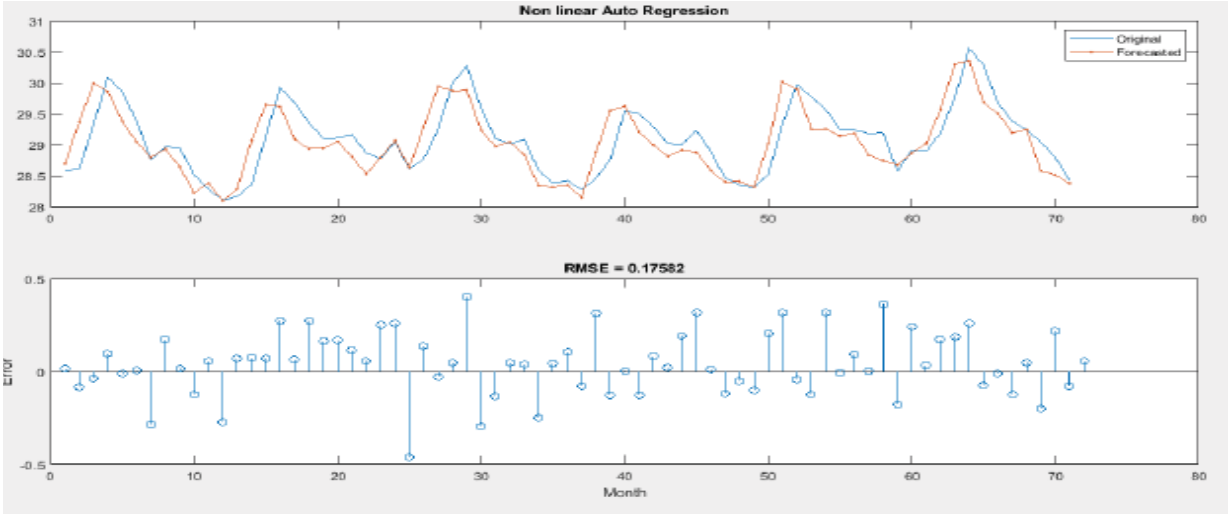


Fig 4.28 Time series plot of (1S-1N, 65E) near to EQT for 6 years using NAR (12, 18)

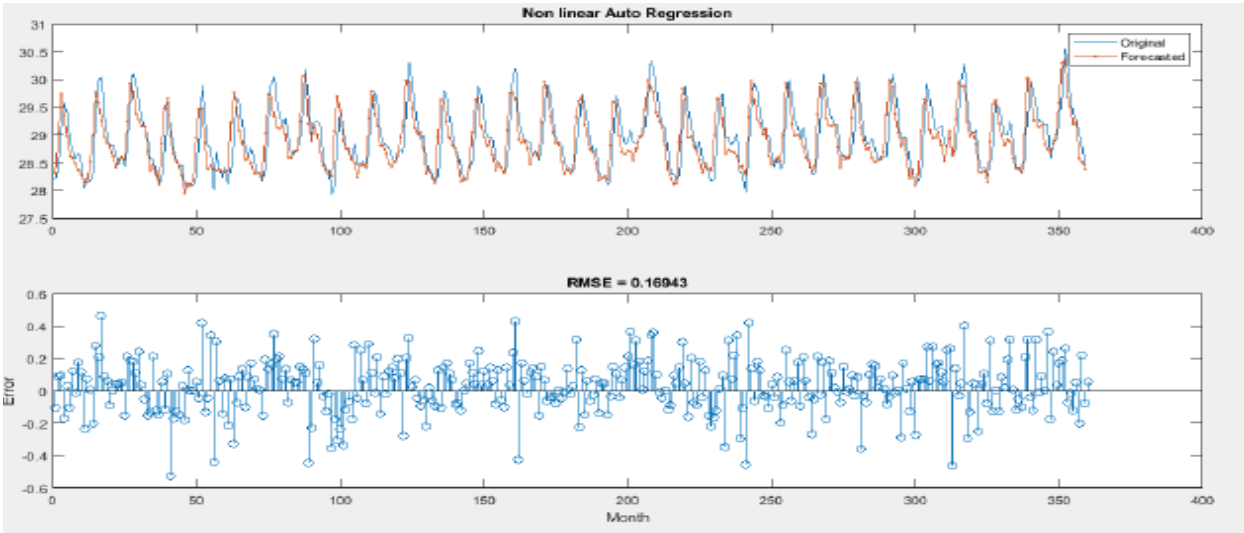


Fig 4.29 Time series plot of (1S-1N, 65E) near to EQT for 30 years using NAR (12, 18)

As evident from Fig 4.27, 4.28, 4.29 the proposed optimized network predicted values are very close to the actual values time series. To understand the influence of the error terms, consider Table 4.15 having vital parameters- Correlation Efficiency (CE), RMSE and MAPE.

Table 4.15 Error comparative at location (1S-1N, 65E) by S.B. Mohongo, M.C.Deo [14]

	S B Mohongo's method			**Proposed method**		
Month	CE	RMSE	MAPE	CE	RMSE	MAPE
Jan	63	0.18	127.37	**100**	**0.1229**	**0.03**
Feb	87	0.12	176.72	**100**	**0.0904**	**1.19**
Mar	79	0.13	134.07	**94.4**	**0.1223**	**2.43**
Apr	84	0.10	63.73	**96**	**0.1897**	**1.99**
May	74	0.14	192.31	**97**	**0.0765**	**1.74**
June	67	0.19	124.28	**93**	**0.0530**	**2.21**
July	64	0.16	190.09	**93**	**0.0736**	**1.86**
Aug	62	0.16	213.50	**95**	**0.0960**	**1.26**
Sep	61	0.15	77.51	**100**	**0.0406**	**0.2**
Oct	72	0.13	97.14	**100**	**0.0732**	**1.44**
Nov	66	0.13	99.44	**100**	**0.0357**	**2.31**
Dec	81	0.10	107.81	**100**	**0.0025**	**0.8**
Avg	72	0.14	133.66	97.36	0.08137	1.45

A study of Table 4.15 shows that our proposed method outperforms in terms of all the error parameters. To begin with, let us see the value of Correlation Efficiency (CE). For the month of January, June, July, August, September and November, the proposed neural network provides 100, 93, 93, 95, 100 and 100 percent correlation efficiency as compared to 63, 67, 64, 61 and 66 percent respectively. Their best performing network provides CE of 87% for the month of February against 100% obtained by the proposed method. A comparative of RMSE values which is the most common error indicator shows that the least square error is also significantly less in the proposed algorithm. The average value calculated over the entire year is 0.08137 ºC in our case against their algorithm is 0.14ºC. The value of MAPE is minimum 63.73 % for the month of April and for our calculation it is 0.2 % for the month of September. Average value of MAPE for their proposed model exceeds 100% whereas for our algorithm, the same is restricted at 1.45%.

Table 4.16 Error comparative at location (1S-1N, 65E) using NAR (12, 18)

multistep	CE	MaxAE	SDD	RMSE	SDE	MAE	MAPE
12	97.76	0.2595	0.6193	0.1455	0.1459	0.1220	0.45
24	96.89	0.3635	0.6193	0.1775	0.1684	0.1438	0.51
36	96.40	0.3635	0.6171	0.1713	0.1641	0.1360	0.48
48	95.80	0.4610	0.6158	0.1845	0.1831	0.1433	0.50
60	95.31	0.4610	0.6144	0.1828	0.1746	0.1452	0.51
72	94.90	0.4610	0.6138	0.1758	0.1718	0.1373	0.48
84	94.56	0.4610	0.6117	0.1707	0.1672	0.1320	0.46
96	94.12	0.4610	0.6104	0.1655	0.1616	0.1277	0.44
108	93.78	0.4610	0.6087	0.1341	0.1609	0.1280	0.44
120	93.42	0.4610	0.6084	0.1680	0.1652	0.1295	0.45
360	89.14	0.5285	0.5794	0.1694	0.1669	0.1336	0.46

Inspired by the good results, we attempted monthly multistep prediction (12 months-1st year, 24 months-2nd year, …..360 months- 20 years). The results of our proposed model are shown in Table 4.16. Remember these are all monthly data, and once the Neural Network is trained with the optimum values of hidden neurons and delay/lag factor, it very efficiently mimics the pattern. This is evident in the correlation efficiency (range 89.14% to 97.76%). The max Absolute Error which is the maximum possible point to point error never exceeds 0.5285 º C. Even the RMSE which represents the least mean square error is restricted in the range of 0.1455 º C to 0.1694º C. The maximum absolute error is in the range of 0.1220 º C to 0.1452º C. The value of MAPE is also restricted in the range of 0.44% to 0.51% that is far less than even 1%. This signifies the potential of the optimized multilayer perceptron model. Such models are pre-existent, but trained to their optimized configurations so as to provide the best predicted values.

4.6.3 Case 3: Comparison of Performance with the Model of Patil, et al [15]

In 2013, Patil, et al. [15] have expanded this work into six different locations in the Indian Ocean vicinity. Their attempt is about Sea Surface Temperature SST value forecast using

Neural Networks using 61 year data (January 1945 to December 2005) at regions of interest distributed across six different locations (fig 4.30) at a time 1 month to 12 month in ahead.

1] Arabian Sea with latitude and longitude values
(AS; 19N-20N, 68E),

2] Bay of Bengal with latitude and longitude values
(BOB; 18N-19N, 90E),

3] East of Indian Ocean with latitude and longitude values
(EEIO; 1S-1N, 90E),

4] West of Indian Ocean with latitude and longitude values
(WEIO; 1S-1N, 65E),

5] South of the Indian Ocean with latitude and longitude values
(SOUTHIO; 9S-11S, 95E-98E) and

6] Off the African Coast with latitude and longitude values
(THERMO; 14S-16S, 56E-58E).

61 years of data (January 1945 to December 2005) at regions of interest distributed across six different locations (fig 4.30) at a time 1 month to 12 month in ahead.
The time series is composed of 61 times 12 = 732 values. They had used previous 24 values (ie D=24) and H is not specified in the publication to predict the single step output.

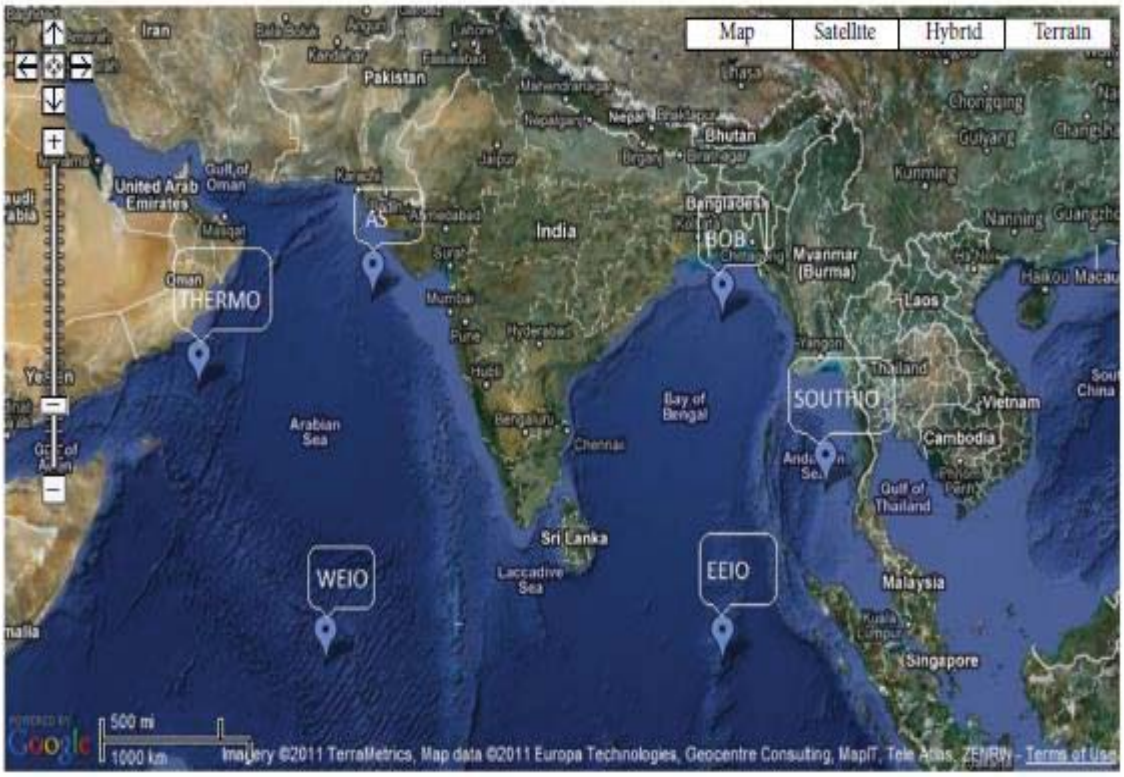


Fig 4.30 The 6 locations around the Indian Ocean [15]

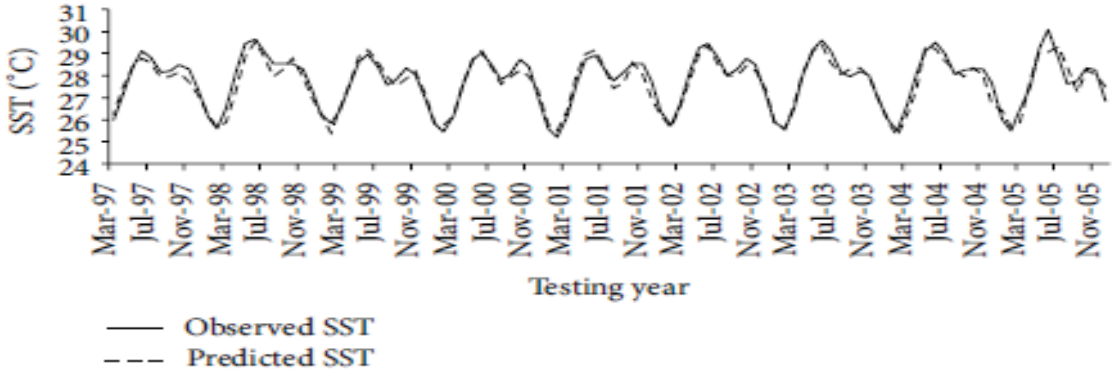


Fig 4.31 The time series at site AS of the proposed algorithm by K.Patil et.al [15]

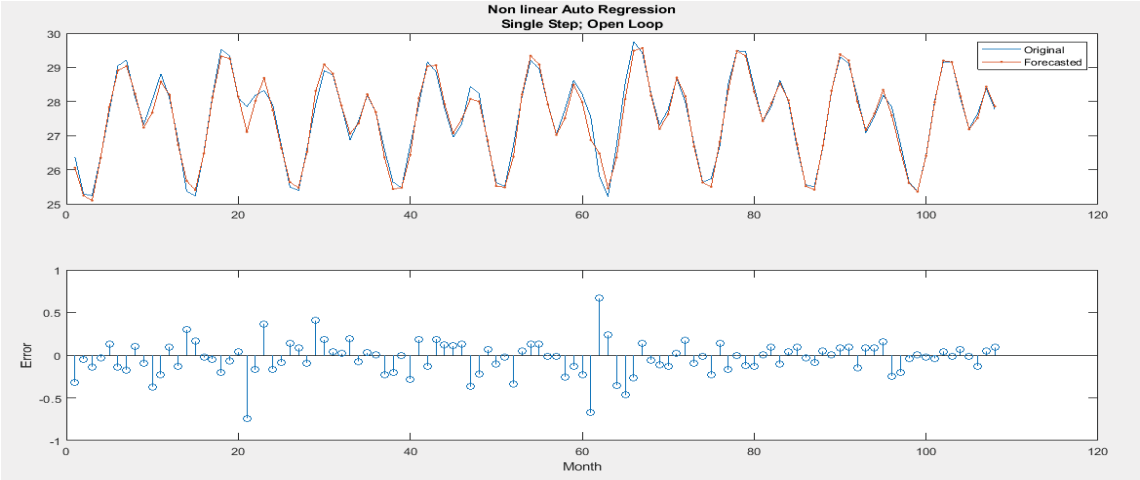


Fig 4.32 The time series of the proposed algorithm at site AS using NAR (12, 18)

To emphasize the goodness of fit of the proposed optimized algorithm, error analysis in terms of the parameters specified by them, using the same number of years for training, validating and testing dataset. Table .4.17 and 4.18 compares K.Patil et.al [15] and our results according to the respective proposed algorithms contains the details of the error comparative for the location Arabian Sea (AS).

Table 4.17 The comparative of the error parameters at site AS-Arabian Sea [15]

Prediction horizon in months	MAE	MSE	NSE
SST $_{(t+1)}$	0.15	0.04	96.90
SST $_{(t+2)}$	0.28	0.13	90.70
SST $_{(t+3)}$	0.27	0.12	91.20
SST $_{(t+4)}$	0.29	0.15	89.50
SST $_{(t+5)}$	0.28	0.14	90.10
SST $_{(t+6)}$	0.28	0.14	90.20
SST $_{(t+7)}$	0.28	0.14	90.30
SST $_{(t+8)}$	0.30	0.14	89.90
SST $_{(t+9)}$	0.28	0.13	90.10
SST $_{(t+10)}$	0.30	0.15	89.20
SST $_{(t+11)}$	0.30	0.15	89.30
SST $_{(t+12)}$	0.28	0.12	91.00
SST $_{(t+i)}$: SST at ith time(month) ahead that the present time (month) "t"			

The Mean Absolute Error (MAE) is reported to be minimum (0.15ºC) for the first month and the same is valid for Mean Square Error (MSE) 0.04ºC. The value of the NSE is highest at

96.90% for the month of January. The Mean Absolute Error is max at 0.1987ºC for our calculations and is almost constant at 0.05ºC from the 5th month onwards, till the month of December. Similarly, the MSE is comparatively very low in the range of (0.0056ºC – 0.0395ºC), minimum 0.006ºC almost for the months of July, August, September, October, November and December. The value of NSE is in the range of 92.14% to 99.78% and again from the month of June to December, the NSE is almost 99% for every monthly SST reading.

Table 4.18 The error comparative at the site Arabian Sea (AS) using NAR (12, 18)

Prediction horizon in months	MAE	MSE	NSE
$SST_{(t+1)}$	0.1987	0.0395	92.14
$SST_{(t+2)}$	0.1196	0.0206	98.26
$SST_{(t+3)}$	0.0801	0.0137	99.19
$SST_{(t+4)}$	0.0651	0.0104	99.14
$SST_{(t+5)}$	0.0596	0.0086	98.93
$SST_{(t+6)}$	0.0572	0.0075	99.47
$SST_{(t+7)}$	0.0509	0.0065	99.55
$SST_{(t+8)}$	0.0536	0.0063	99.77
$SST_{(t+9)}$	0.0491	0.0056	99.74
$SST_{(t+10)}$	0.0575	0.0068	99.29
$SST_{(t+11)}$	0.0569	0.0064	99.53
$SST_{(t+12)}$	0.0598	0.0066	99.78
$SST_{(t+i)}$ is the SST when seen t months ahead with respect to the ith time(month)			

Hence, we may conclude that our proposed algorithms are optimized and have attained a near perfect fit for all SST values calculation at various different locations that are investigated by fellow researchers in the past.

4.7 Conclusion

Chapter 4 summarizes our proposed algorithms using the linear and nonlinear techniques. It introduces the various linear and nonlinear algorithms. It shows the importance of ACF and PACF plots to identify the optimum machine learning parameters associated with the ARIMA models. The Non-linear auto regressive models are also optimized. Later with these optimized parameters, we obtain single step ahead prediction SST values. Using Closed loop configuration, the prediction is extended to multiple step SST values. Also, comparison with existing literature provides better performance on the same dataset using ARIMA, NAR and

NARX models with and without Exogenous inputs. For Dataset 2, NRMSE obtained by ARIMA is 0.0193, and that for the NAR method is 0.0302. The correlation coefficient is found to be 97.5% for both ARIMA and NAR model. The values of NSE is found to be 0.9601 (for ARIMA) and 0.9575 (for NAR).

CHAPTER 5

Deep Neural Network and Its Application to Sea Surface Temperature Prediction

5.1 Deep Neural Network in Time Series Analysis: An Overview

Time series occupy such a common place today in our lives. Researchers realize the potential of exploring the hidden features in these series using Deep learning technologies [21, 22, 27, 41, 83]. The Deep Neural Networks as the name suggests, have multiple hidden layers and these hidden layers themselves also have a large number of neurons which increases the capability of these networks to judge the variations in the incoming data stream. Recurrent Neural Networks (RNNs) are recognized as potentially strong tool to explore the nonlinearities of the input and hence are good at predicting the upcoming value to a large extent. The Long Short Term Memory (LSTM) is basically a Recurrent Neural Network that can learn from the dependency of the order of the sequence from the items in sequence. If the sequence is sufficiently long, then the neural network learns from the series of occurrence and retains required information that is carry forwarded onto the next element of the sequence. Many applications of the deep neural network are found in the domain of time series analysis. Apart from that, there are many innumerable case studies finding a linkage between different parameters.

5.2 Related Work

Researchers in the past have investigated the use of LSTM networks for various applications. Dong [21] used the LSTM model for prediction of SST on Bohai SST Dataset. They have used the Adagrad optimization method over Stochastic Gradient Descent (SGD) method for robust LSTM with a fully connected layer. Further comparing it with the Support Vector Regression (SVR) technique, it is exhibited that their proposed LSTM method is better at

predicting the SST values. Athira [22] utilizes the six indices on Air Quality (PM_{10}, $PM_{2.5}$, NO_2, CO, O_3, SO_2, AQI) from the China National Environmental Monitoring Center (CNEMC) and the weather data provided by NOAA at 6 hours interval. Using their LSTM network, they could generate ten steps ahead prediction of the pollutant content in Air, PM10. Zhao and Chen [27] have used the LSTM network to predict short term traffic using data collected from more than 500 observation points at a time gap of 5 minutes each. The model proposed by them is able to predict traffic in a time interval of 15, 30, 45 and 60 minutes using 2, 3, 5 and 6 layers respectively in the LSTM network. Li and Cao [83] have compared the output of a Simplex and a Stacked LSTM network for the purpose of Tourism Flow Prediction. They have implemented a one step ahead prediction using the last 7 readings of tourism flow with the help of 64 neurons in the hidden layer for the location Small Wild Goose Pagoda in X'ian, China.

5.3 The LSTM network

Long Short Term Networks [93] are a special unit of Deep Learning networks. An LSTM network is a type of recurrent neural network (RNN) that can learn long-term dependencies between time steps of sequence data. LSTM networks can remember the state of the network between predictions. The network state is useful when we do not have the complete time series in advance, or if we want to make multiple predictions on a long time series.

The core components of an LSTM network are a sequence input layer and an LSTM layer. A *sequence input layer* inputs sequence or time series data into the network. An *LSTM layer* learns long-term dependencies between time steps of sequence data.

These networks find large scale application in:

1] Sequence classification using Deep Learning.

2] Sequence Regression using Deep Learning.

For the LSTM layer, choose an output size, and specify the output mode ‘sequence’. The property of gradient flow implemented over consequent sequences during the training period strengthens the network. This is further accentuated by the fact, that the same is not identical to a regular neuron [94] as

It controls the time of arrival of the input neuron.

It controls whether to recall the computations performed in the previous time stamps.

It also decides, the instant at which the output is available at the end of the time stamp

And the best part of this is that the LSTM is capable of deciding based on current input.

LSTM units are a variation on the classic artificial neuron as it has:
1] Connections from the previous time-step (outputs of those units)
2] Connections from the previous layer

The memory cell in an LSTM network is the central concept that allows the network to maintain state over time. LSTM was initially proposed in 1997 by Hochreiter and Schmidhuber [93] under the influence of the process of flow of error in the existing RNN wherein effect of long time lags couldn’t be accessible to the present architectures. They performed well for predicting interval and delay based functions in time series. A typical RNN is an improved version of a multilayer perceptron having an input, hidden and output layer.

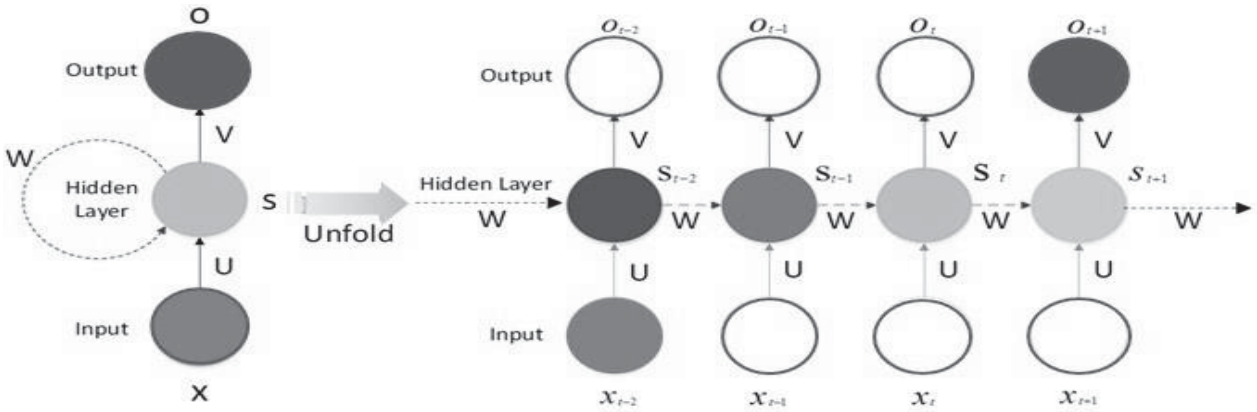


Fig 5.1 A typical RNN network unfolded

An unfolded view of a typical RNN structure is shown in Fig. 5.1.

Here, X_t is the input at time step t, O_t is the output at time step t. S_t is the hidden state at time step, and it is the 'memory' of the network. *W, U, V* are parameters in different network layers. Unlike a traditional deep neural network, which uses different parameters at each layer, an RNN (fig 5.1) shares the same parameters (*U, V, W)* across all iterations.

In LSTM networks, as shown in the fig 5.2 the issue of vanishing gradients that is observed in RNN, can be resolved using three internal cell gates and memory cells to store information.

There are four gates-

Input gate – to take a new input from outside; process the newly arrived data.

Memory gate- to collect input from the output of the LSTM NN cell in its last iteration.

Forget gate- when to forget the set of output results and

Output gate- collects results based on calculation and generates the output.

A typical LSTM Network cell structure is shown in Fig. 5.2.

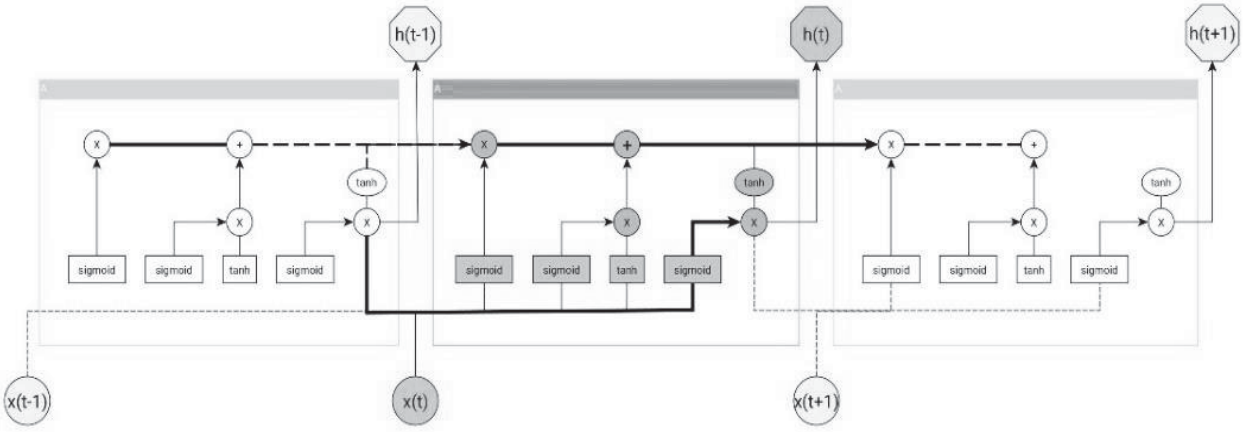


Fig 5.2 A typical LSTM NN cell

In the SST prediction model, following are the sequences of the LSTM cell.

Let t=1 ,……,$T$are different time stamps of the time series. Let the input time series be denoted as X =(x_1, x_2, … ,x_T), hidden state cells as H = (h_1, h_2, … h_T), and output sequence as Y = (y_1, y_2, … ,y_T).The LSTM makes computation based on Eq.(16) and Eq.(17),

$$h_t = H(W_{hy}x_t + W_{hh}h_{t-1} + b_h) \tag{16}$$

$$y_t = W_{hy}h_t + b_y \tag{17}$$

The implementation of the LSTM is presented as Eq.(18)-Eq.(23).

$$i_t = \sigma\,(W_{xi}\,x_t + W_{hi}\,h_{t-1} + W_{ci}\,c_{t-1} + b_i) \tag{18}$$

$$f_t = \sigma\,(W_{xf}\,x_t + W_{hf}\,h_{t-1} + W_{cf}\,c_{t-1} + b_f) \tag{19}$$

$$c_t = f_t * c_{t-1} + i_t * tanh\left(W_{xc}\,x_t + W_{hc}\,h_{t-1} + b_f\right) \tag{20}$$

$$o_t = \sigma(W_{xo}\,x_t + W_{ho}\,h_{t-1} + W_{co}\,c_t + b_o) \tag{21}$$

$$h_t = o_t \tanh\,(c_t) \tag{22}$$

$$\sigma_x = \frac{1}{1+e^x} \tag{23}$$

Here, σ_x and tan h are the activation functions of LSTM.

σ is the standard sigmoid function.

This layer outputs values in the range of 0 to 1, thereby deciding the part of the component to pass through. 0 means “nothing goes”, 1 indicates “pass all through”, i, o, f and c indicate the inner cell gates, respectively the input gate, forget gate, output gate and cell activation vectors, c should be same as h. W indicates the weight matrices. Input gate decides the variation on the memory cells. Output gate permits memory to reflect on the outputs. The forget gate decides what to forget or remember from the previous states.

5.4 LSTM model for SST prediction

Following steps are used to obtain the SST prediction

Experimental setup:

We use the window method for LSTM regression and the size of the window is D and the input is of the order of $\{(x_{t-D}, x_{t-D-1}, \ldots\ldots x_t) \rightarrow x_{t+1}\}$. We first try to evaluate the size of the input layer, the number of hidden layers and the number of hidden units in each hidden layer. The input dimension is unity. For the hidden layer, there are 150 hidden neurons and an output layer with a single node. Default sigmoid activation function is used for the LSTM units. A dropout of 0.2 is selected to avoid the issue of over-fitting. The entire network is trained for 250 max epochs with an initial learning rate of 0.005 using root mean square- 'rmsprop' activation function.

Algorithm 5.1 Single step Prediction and Multistep prediction using LSTM model:

Step 1: Consider the Dataset 1-5 (Section 3.3) Construct a LSTM network using the specifications mentioned above. We have opted for the use of 150 hidden layers.

Step 2: Perform single step prediction by using individual time series dataset to the model. Using the parameters specified, the Error metrics are verified so as to ensure minimum error.

Step 3: Now with these optimized values of the parameters, apply the same to multiple (12) step ahead prediction. The LSTM algorithm is applied to each dataset 1-5 and the results are tabulated as follows. Refer Fig. 5. 3 (a, b, c) - 5.7 (a, b, c).

LSTM algorithm is applied to dataset 1-5 (as mentioned in Section 3.3) and its time response is available in the fig 5. 3 (a, b, c) - 5.7 (a, b, c). Evaluation based on the metrics (defined in Section 3.4) is available in Table 5.2. Every figure 5. 3 (a, b, c) - 5.7 (a, b, c) is composed of two sub parts (a, b) and (c). The figure 5.3 (a) represents the multistep prediction of 12 steps using the LSTM network with the mentioned parameters for dataset 1[75]. Figure 5.3 (b)

represents the error components individually and provides the value of RMSE. We have mentioned the distribution of our datasets into training and testing in the ratio of 90%-10% (Section 3.3) and in order to show the correct representation, fig 5.3 (c) is also available for analysis. Same is applicable for fig 5.4 (a, b, c) - 5.7 (a, b, c)

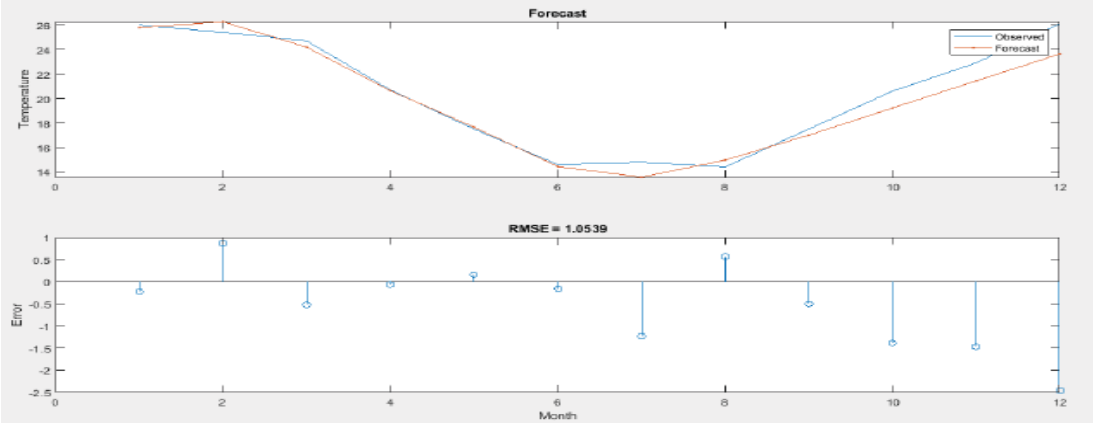


Fig 5.3 The predicted time series response (Multistep prediction with step size 12) for Dataset-1 (a) Plot of test data vs. predicted response (b) Error plot for the predicted response (RMSE=1.0539)

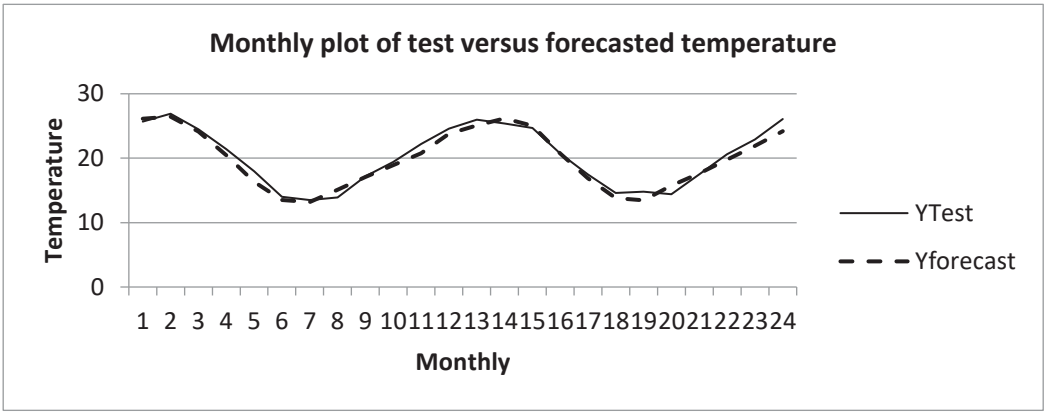


Fig 5.3 c The predicted time series response (Multistep prediction with step size 12 for two iterations) using for Dataset 1

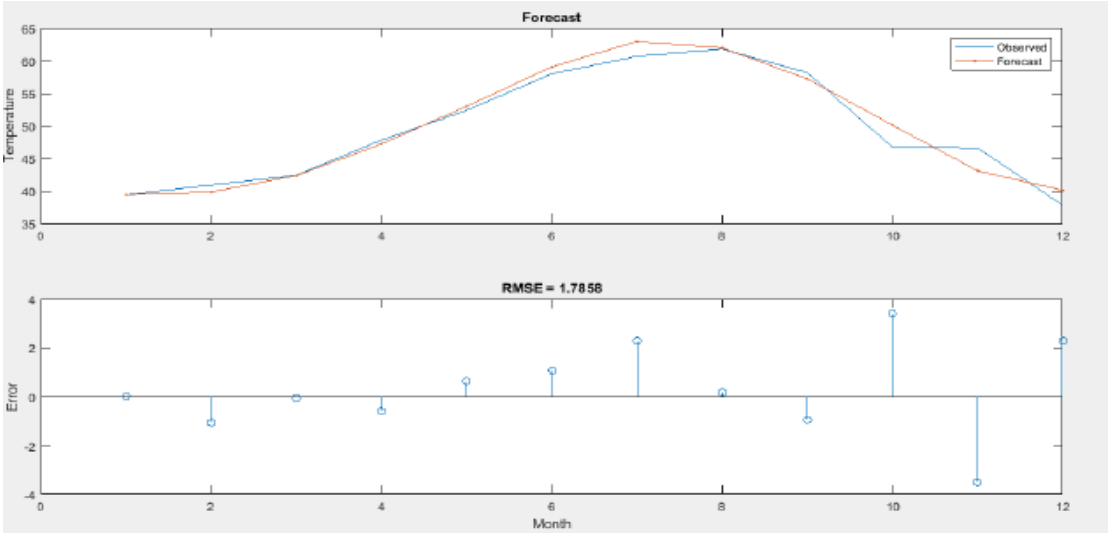


Fig 5.4 The predicted time series response (Multistep prediction with step size 12) for Dataset- 2 (a) Plot of test data vs. predicted response (b) Error plot for the predicted response (RMSE=1.7050)

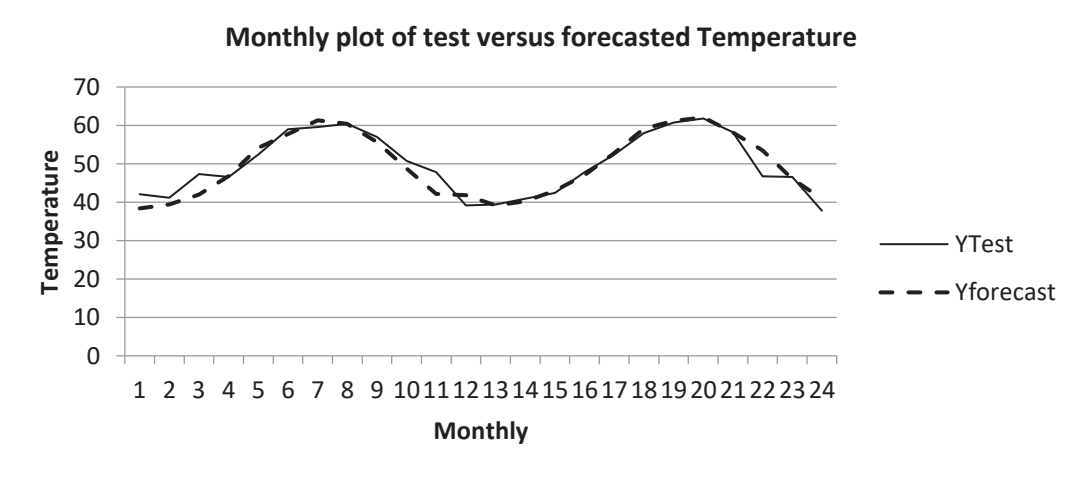


Fig 5.4 c The predicted time series response (Multistep prediction with step size 12 for two iterations) using for Dataset 2

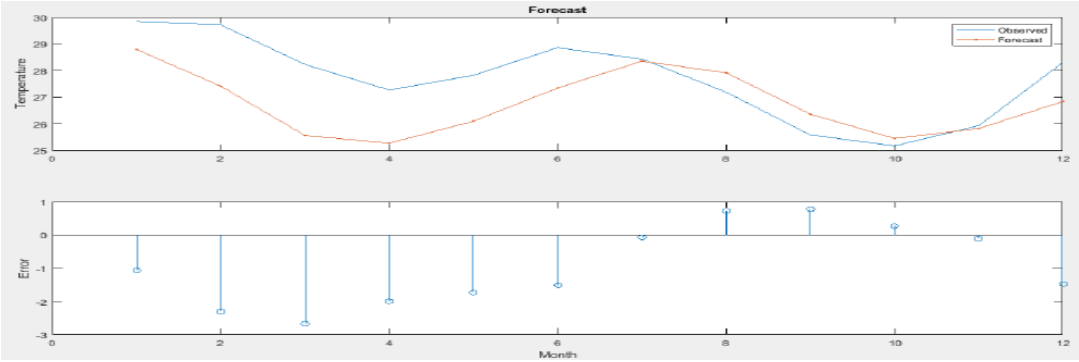


Fig 5.5 The predicted time series response (Multistep prediction with step size 12) for Dataset- 3 (a) Plot of test data vs. predicted response (b) Error plot for the predicted response (RMSE=0.929)

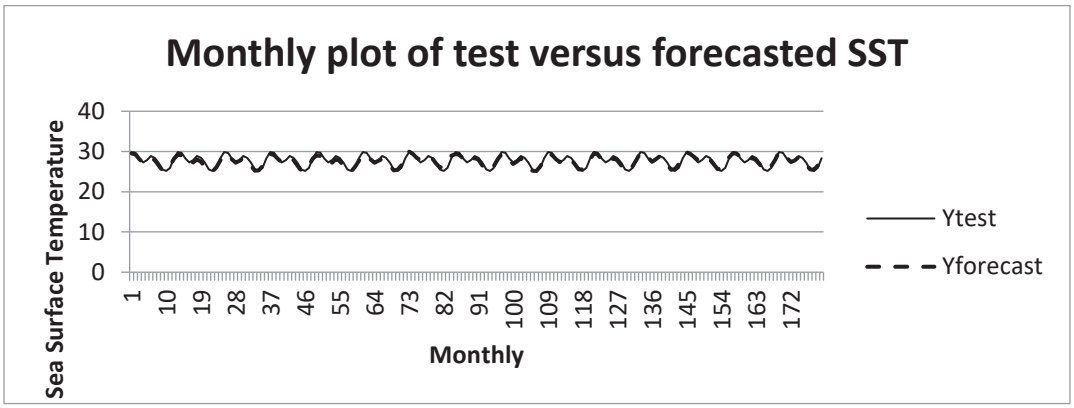


Fig 5.5 c The predicted time series response (Multistep prediction with step size 12 for fifteen iterations) using for Dataset 3

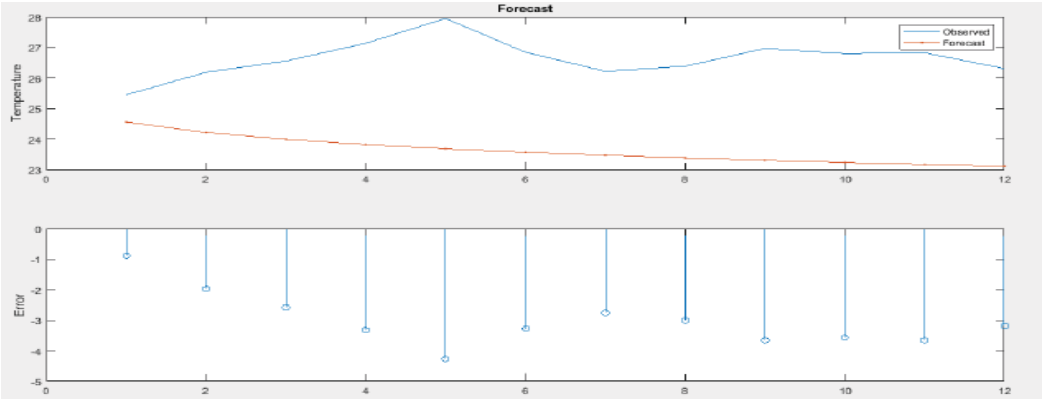


Fig 5.6 The predicted time series response (Multistep prediction with step size 12) for Dataset- 4 (a) Plot of test data vs. predicted response (b) Error plot for the predicted response (RMSE=3.0012)

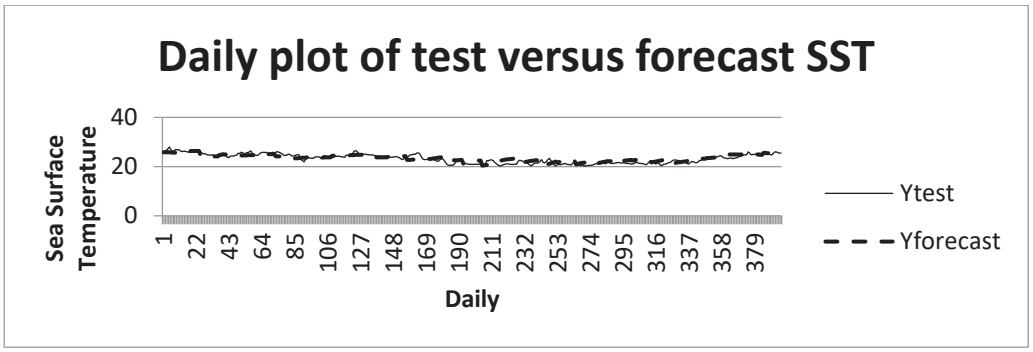


Fig 5.6 c The predicted time series response (Multistep prediction with step size 12 for thirty three iterations) using for Dataset 4

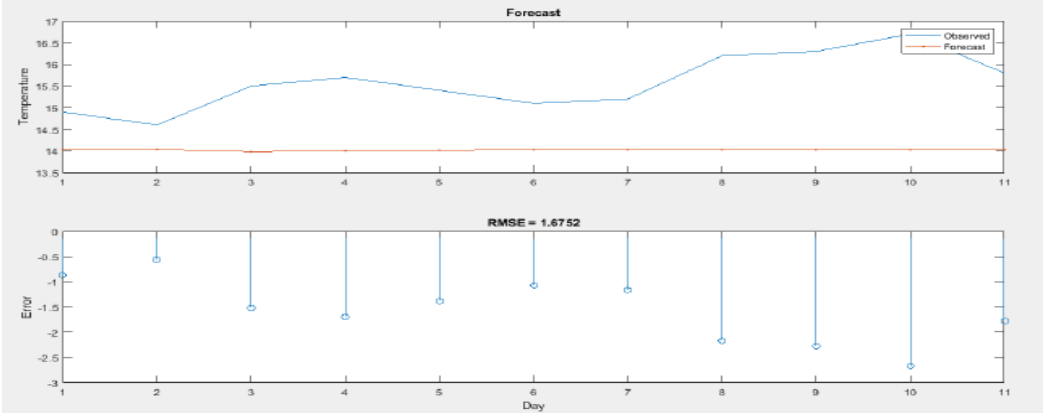


Fig 5.7 The predicted time series response (Multistep prediction with step size 12) for Dataset- 5 (a) Plot of test data vs. predicted response (b) Error plot for the predicted response (RMSE=1.6752)

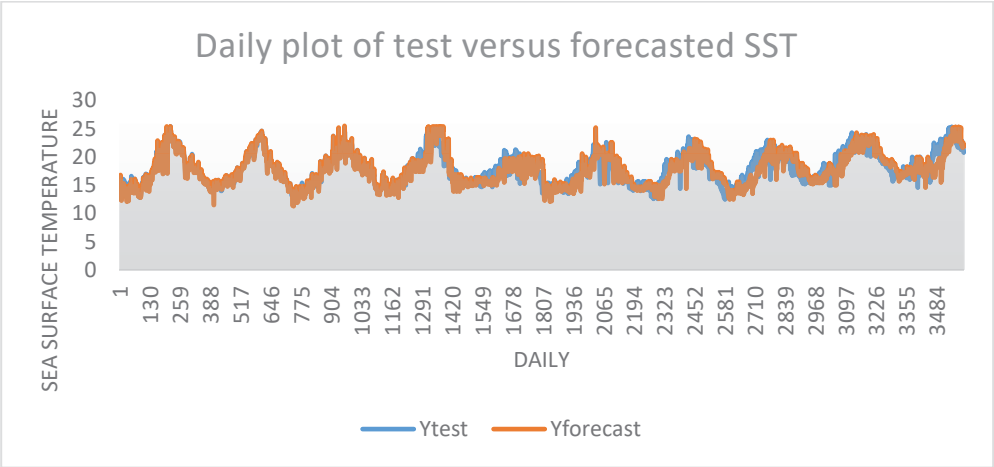


Fig 5.7 c The predicted time series response (Multistep prediction with step size 12 for three hundred and two iterations) using for Dataset 5

5.5 Results and Discussion

Initially, it is observed that for single step ahead calculations, the outputs generated by the LSTM network shows good agreement in the temporal domain. While evaluating the values of the Error measures also, the errors are observed to be low enough and can be acceptable. Hence, the experiment of multistep prediction is extended for all time series with a step size of 12 in each iteration. The monthly datasets- Melbourne Mean Temperature dataset [75], Nottingham Castle Mean Temperature dataset [75] and the HadISST datasets dataset [76] have the values of RMSE error in the range of 0.9290 to 1.7858 which is distributed over a size of 12 steps and hence can be categorized as low error as evident from Table 5.2. The Normalized RMSE is also comparatively low. Observe that when we try to address the Daily SST datasets, the same errors show a higher value- RMSE in the range of 1.6752 to 3.0122. As we know the Elnino dataset and La Jolla, California, West Coast dataset are daily SST datasets and reflects more variation on a day to day basis. To capture and tune to these variations, either better tuning of the model parameters is desired or the model capability needs to be strengthened.

Table 5.2 Error performance of LSTM on all datasets (single iteration of 12 steps)

Dataset	Type	Multistep	RMSE	NRMSE
Melbourne Mean Temperature	Monthly	12	1.0539	0.0631
Nottingham Castle Mean Temperature	Monthly	12	1.7858	0.1069
HadISST	Monthly	12	0.9290	0.1984
Elnino dataset	Daily	12	3.0122	1.2049
La Jolla, California, West Coast	Daily	12	1.6752	0.2043

The test dataset in every case is considered to be 10% of the total dataset. This is to ensure point to point comparisons of various techniques in a uniform manner. So, multiple iterations consisting of multiple steps (12 steps each) are implemented and the results are analyzed in time series plots and their error values are also evaluated for consideration. The choice of a step size is meant to ensure that we can rightly address the multistep prediction for monthly and daily datasets. A step size of 12 (consider Table 5.2) means it is annual prediction for monthly SST datasets and the same is little less than two weeks for a daily SST dataset. The implementation of the simple LSTM network over the test data reveals the following observations as evident in Table 5.3. Evidently for time series prediction, it is always advisable to simultaneously analyze the time series plot and the error parameters. Observe the values of Correlation Coefficient for example. The values are in the range more than 95% for all datasets, distributed over a bulk of data steps. However, analysis of the figure 5.4 a, 5.4 b, 5.4 c, 5.5 a, 5.5 band 5.5 c provides a very clear indication that the LSTM system is able to predict the SST values, but the same can be definitely worked upon and made better.

Table 5.3 Error performance of LSTM on all datasets (Average over multiple iterations)

Dataset	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	Sres	CC	NSE
Dataset 1	**0.2048**	**0.0123**	0.4526	0.0271	1.6744	0.7246	0.0383	4.4169	0.7695	0.252	97.2949	0.965
Dataset 2	0.3104	0.023	0.9002	0.066	1.7643	1.2521	0.0276	8.5862	2.1168	0.7002	97.3198	0.9278
Dataset 3	0.444	0.044	**0.0664**	**0.0061**	**0.3542**	**0.1647**	**0.006**	**1.3154**	**0.1906**	**0.0188**	**97.5413**	**0.9746**
Dataset 4	1.792	0.1792	1.3386	0.1338	2.622	1.3615	0.1506	2.0195	0.6861	0.2571	96.6179	0.7104
Dataset 5	1.656	0.3094	1.9123	0.3578	2.7761	1.9123	0.0902	1.995	0.6741	0.3672	82.0919	0.6861

Table 5.4 Comparison of NRMSE, CC and NSE using all the three methods for all data sets

Dataset	NRMSE_ARIMA	NRMSE_NAR	NRMSE_LSTM	CC_ARIMA	CC_NAR	CC_LSTM	NSE_ARIMA	NSE_NAR	NSE_LSTM
Dataset 1	**0.038**	**0.0323**	0.0271	**97.201**	**97.2292**	97.2949	**0.9601**	**0.9575**	0.965
Dataset 2	0.050	0.0546	0.066	93.303	92.3156	97.3198	0.9037	0.8871	0.9278
Dataset 3	0.062	0.0518	**0.0061**	93.538	96.5537	**97.5413**	0.9089	0.9135	**0.9746**
Dataset 4	0.1136	0.1081	0.1338	87.175	86.7585	96.6179	0.7472	0.6985	0.7104
Dataset 5	0.1174	0.304	0.3578	84.883	84.8019	82.0919	0.7461	0.6911	0.6861

Let us now take up another analysis.

A comparative of some specific error components namely NRMSE, CC and NSE is done for all the monthly and daily SST datasets and their numerical values are compared in the same backdrop. Remember the test dataset size is restricted to 10% of the complete dataset. As the dataset length is variable, we can still provide uniform comparison in a comprehensive manner by ensuring that the ratio of the training and test dataset is kept constant throughout the entire experimental setup. The values of NRMSE noticeably quite low, even for daily SST datasets do not reflect agreement in the temporal representation. The values of NRMSE using ARIMA, NAR and LSTM techniques are fairly low for monthly datasets and when their time series plot is studied, it shows a good agreement. Similarly, when we observe the Correlation Coefficient (CC), although the values are very near to 100%, still for daily SST datasets, the time series responses differ at various instances. Nash Sutcliffe Efficiency (NSE), that is a very strong indicator of error calculations related to hydrological studies reflects NSE nearest to 1 is at 0.9746 (dataset 3, LSTM method) and the worst NSE is located at 0.6911 (dataset 5, NAR). The CC value is highest at 97 (Dataset 1, ARIMA method; Dataset 1. NAR and Dataset 3, LSTM). The NRMSE is lowest at 0.0061 (dataset 3, LSTM).

5.6 Conclusion

This chapter provides inferences based on the implementation of the LSTM network for the various monthly and daily SST datasets. For monthly SST datasets, the LSTM network provides very appreciative results in terms of time series agreement of the real and the predicted values. Even the same is supported by average values of error metrics. For daily SST datasets, there is still scope of improvement as far as the time series response is

concerned. Of all the error metrics used for LSTM network, NAR network and ARIMA, it is observed that NSE provides a good estimate numerically and aligns with the time series response simultaneously. Nash Sutcliffe Efficiency (NSE), that is a very strong indicator of error calculations related to hydrological studies reflects Nash Sutcliffe Efficiency (NSE), that is a very strong indicator of error calculations related to hydrological studies reflects NSE nearest to 1 is at 0.9746 (dataset 3, LSTM method) and the worst NSE is located at 0.6911 (dataset 5, NAR). The CC value is highest at 97 (Dataset 1, ARIMA method; Dataset 1. NAR and Dataset 3, LSTM). The NRMSE is lowest at 0.0061 (dataset 3, LSTM).

CHAPTER 6

Hybrid Algorithm for Prediction of Sea Surface Temperature

6.1 Introduction

In the past few chapters, we have investigated the use of linear and non-linear methods for prediction of sea surface temperature and tested the same over variable datasets having variable length and also variably spaced in the time domain. As observed so far, the monthly SST dataset is investigated by many researchers [1, 2, 3, 4, 12, 13, 14, 15, 40] however, as the Earth is subject to dynamic changes every day of the year, hence it is a challenge to predict the daily SST. To make things adverse there are couple of other factors that play a very vital role like the inclination angle of the sun, the climatic variations contributed by natural and manmade reasons contributing towards more dynamism.

It has also been suggested by many researchers that as the seas and oceans and all other water bodies are linked across the surface of Earth, variation in any one of them can actually bring repercussions across not only some other water body but also over other land forms. The formation of extreme events like cyclones, Tsunami is also linked and contributed by a combined emphasis of this and various other parameters. There are groups of researchers who realized that seasonal and non-seasonal variations could probably be explored with the help of some unique approaches. For SST dataset, Tripathi, et al. [12, 13, 14] had proposed the formation of 12 networks, one corresponding to each month of the year. From the existing time series, he had reconstructed 12 parallel time series and using Neural Networks, he had concluded in his study that specific months; in this case the months of June, September, October and November provided a higher match with the actual values as compared to linear regression models. They had proposed the formation of a neural network having multiple inputs, 2 hidden neurons and 1 output using data from the Indian Ocean vicinity. The sites

identified were 27 ºS to 35 ºS and 96 ºE to 104 ºE using Reynold's reconstructed SST dataset ranging from the year 1950-2001.

Aparna, et al. [20] came up with a novel idea of calculating SSTA using neural networks. They devised an image based approach where the SST data of the previous day is fed as input to the network to generate the present day's SST using OCEAN FINDER. OCEAN FINDER is a PSC (0105) programmer of CSIR-NIO to study the development of SST fronts in the ocean and the contribution of these SST fronts towards ecological dynamism. The 8 outer locations of a 3×3 grid is used as the input layer corresponding to 4 km of resolution. The number of nodes in hidden layer is 5 (ie two thirds of the input) and in the output is 1. The mean value is calculated by dividing the sum of SSTs at each of the 8 grid positions. This is then subtracted from the SST of that position, giving rise to the SSTA. After normalization, they are trained and tested. Using around 25,000 patterns per day, the next day's SST value is predicted and the same is compared with the next day SST reading. can be calculated.

Using a combined structure of a LSTM Cell and fully connected layers on the Bohai Sea in China, Zhang, et al. [21] have tried to predict the SST values for a testing dataset of '*t*' time slot that is one fourth of the training dataset length. Adagrad optimization technique is used in Stochastic Gradient Descent (SGD) for frequent updates on smaller variations. From the Scripps Pier Institute of Oceanography, a study correlating algal blooms and its prediction is undertaken by McGowan, et al. [28] to understand the effect of biological stochastic chaos with respect to the algae bloom. Two methods Separate Training Method (STM) and Direct Training Method (DTM) using MLP for predicting the SST values are proposed by Wie [42]. With Mean SST and SST anomaly separately used to train the MLP layers, especially to chart the areas with significantly large variations in SST values, the model is able to predict the SST values to a large extent. These all are certain instances where soft computing techniques have supported the forecasting of the SST values in one or some other form. However, while we referred different literature, we realized that there are many instances where the dynamism can be explored with hybridization. Some of the hybrid algorithms had gained popularity in other domains of applications like the Economic timeseries forecasting, Traffic management, Pollution forecasting and so on.

6.2 Related work

Zhang [24] in the year 2001, had proposed a hybrid model for prediction using standard datasets. Our proposed model presented in this chapter is inspired from Zhang's technique and has been tested for various combinations so as to provide the nearest match to the SST time series data. The results evidently show a very appreciable forecast for week ahead predictions. A remarkable compilation of various time series by Adhikari and Agrawal [70] gives a detailed basic understanding of various stochastic processing and SVM. A detailed review of soft computing techniques for time series forecasting on various standard datasets is presented by Sanghani, et al. [71] which includes methods like ARIMA, ANN, SVM, and hybridization of ARIMA with ANN.

Smyl [95] addresses the challenges of M4 forecasting competition and is the winner of this challenge. It consists of a hybrid model that has a statistical component contributed by Exponential Smoothing and the Deep NN component in terms of LSTM network. It consists of two parts- Prediction Intervals (PIs), that is predefined for the competition and is represented in terms of Mean Scaled Interval Score (MSIS) and Pinball Functions (PFs) that are asymmetric and are used with the intention to overcome the effect of bias. There are separate models to address the prediction on the various types of timeseries models-hourly, daily, weekly, monthly and annual datasets. Santos Júnior et.al [96] have proposed a hybrid model that consists in the first phase of the ARIMA model and in the second phase it uses either a MLP or a SVR model. Therefore, the resulting hybrid models are – a) ARIMA + MLP and b) ARIMA + SVR. 6 real time series are tested using this hybrid model and the same are compared using error metrics- MSE, MAE and MAPE. All the cases are one step ahead prediction. The same is projected up to variable points (ie testing dataset) depending upon the original data set dimension. Olivera [97] have shortlisted 13 timeseries datasets from the Timeseries library and the ARIMA part is implemented using R software and the remaining portion is done using MATLAB. This is also done in a step ahead manner. Friedman hypothesis test is used for general comparison and the Nemenyi test performs comparison in pairs through Critical Distance calculation using hybrid methods. Neto [98] have proposed a nonlinear combination hybrid prediction model for one step ahead prediction is proposed that

has higher accuracy than other state of the art methods. It performs this in a two-stage pattern. The first stage involves the prediction of the model parameters and also their residuals. And the second stage tries to find the best function that combines the models using MLP. They proposed three different variations of the NoLiC model and six metrics are used to check the efficiency for PM concentrations at two different locations of Helsinki, Finland.

6.3 A Proposed Hybrid Model for SST Prediction

Looking at the time series responses in the previous chapters, we observed that traditional methods (like ARIMA), machine learning methods (like neural networks) and deep learning methods (LSTM) still keeps scope for the improvement of the Daily SST dataset. This motivated us to develop more effective hybrid model exploiting the advantages of individual methods and overcoming the drawbacks of them. We experimented the proposed method on Elnino daily dataset.

This array consists of around 70 moored buoys distributed across the Equatorial belt around the Pacific Ocean providing coverage to vital El Niño, La Niña and ENSO sites. They measure various surface and subsurface oceanographic parameters including surface temperature, subsurface temperature (up to 0.5km beyond the surface), relative humidity, surface winds and air temperature [78]. This dataset [78] consists of - location of the buoy, date of measurement, Zonal Winds (consider for East>0; for West<0), air temperature, Meridional Winds (consider for North>0; for South<0), SST and relative humidity. The fluctuations in wind data is + 10m/s and the same with respect to relative humidity is 70%-90%. Variations in the values of the SST and Air Temperature are both restricted in the limit of 20 to 30 degree Celsius. It is ensured that all the readings are recorded at the same HH:MM:SS in the day. Using this daily SST dataset, we attempted to obtain the values of SST Anomaly (SSTA).

El Niño (warm) and La Niña (cool) events in the tropics of the Pacific are categorized using the standards by NOAA [77], known by the name of The Oceanic Niño Index (ONI). For a total of 3 months and more in continuation, a variation of $\pm$ 0.5°C in the values of SST categorizes it either as a warm or cold event. Intensity is further classified based on the magnitude of the change in SST values. For Weak events, this is 0.5 °C to 0.9 °C. For moderate

it is 1.0°C to 1.4°C. For Strong and Very Strong, the same is 1.5 °C to 1.9 °C and beyond 2°C respectively [77]. A part of this dataset is so identified that it has minimum missing values and yet is larger than a span of 10 years, which we thought is a duration correct enough to be considered for Time Series study. The location that could meet the above mentioned criteria is (0°N , -110 E). The duration covers the dates from 10th May, 1985 to 20th July 1995. We distributed our dataset into two sections- training (from 10th May, 1985 to 10th May, 1995) and testing (11th May, 1995 to 20th July, 1995) thus making the test dataset independent of the training dataset. Now we have a timeseries each of SST, Meridional Winds, Zonal Winds and Air temperature having the common timelines. Using a Non-linear Auto Regressive Network with Exogenous inputs, we have predicted the SST values and then computed SSTA with the help of different parameters and finally compared all the cases.

Using this Elnino daily dataset obtained from NOAA's PMEL, it is observed that for daily dataset, the above stated models, even with optimized values, are not much efficient in predicting the future values efficiently for multiple steps.

In this research work, an efficient hybrid model by combining traditional and machine learning/deep leaning methods is proposed in this chapter. A block diagram of the proposed model is shown in Fig. 6.1.

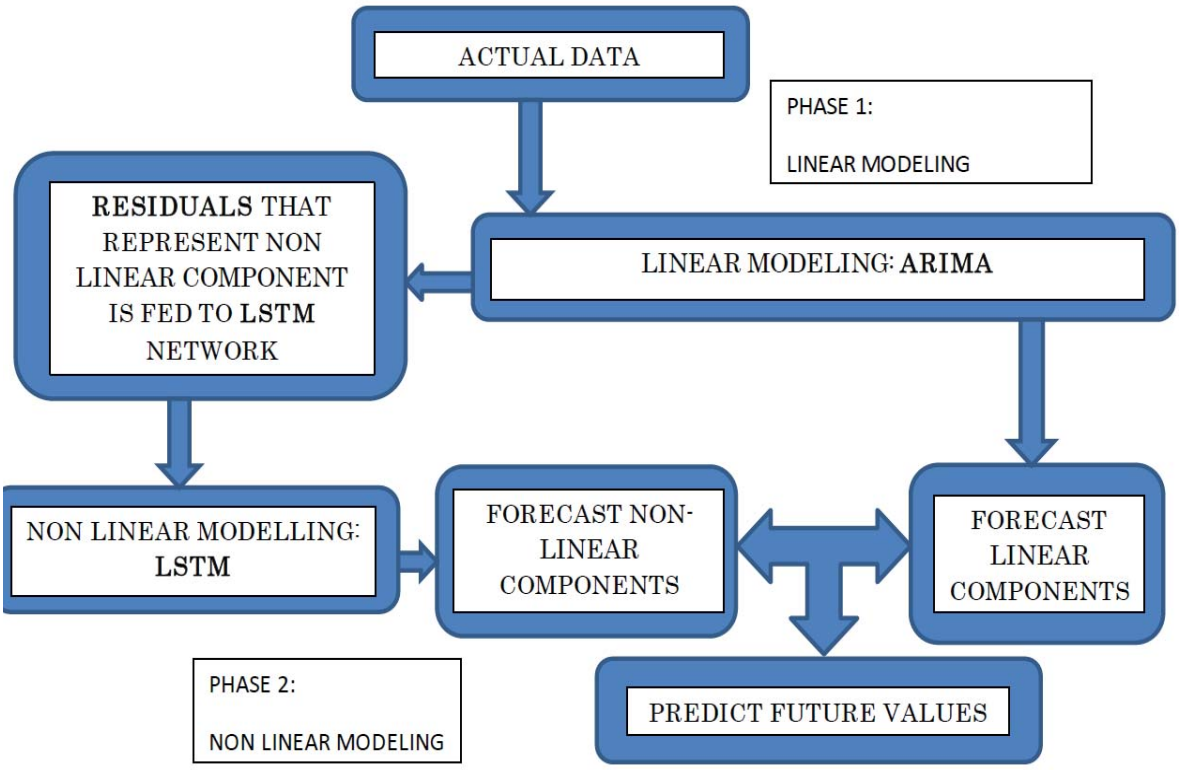


Fig 6.1 Block diagram representation of the proposed model

We attempted to explore the individual capabilities of the Linear model (ARIMA) and then separately we also did the same for the Non-linear component (NAR/LSTM). It was observed from the last two chapters, that the NAR had performed fairly strong for monthly SST datasets even in the extended task of multistep prediction. Singularly LSTM models have also provided excellent results. Here the actual data is individually modeled using ARIMA model first and a set of residuals are obtained. These residuals are derived from past data. Based on these residuals, a set of linearly forecasted residuals are now obtained. Again, the same actual SST readings are modeled using NAR model. This will also provide a set of residuals, thereby again forecasting a set of Non-linear components. It is here now that the linearly forecasted components are algebraically added to the Non-linearly forecasted components to predict the future values. Such predicted values show a larger degree of agreement with the actual time series dataset.

In our proposed model, the same is repeated by replacing the NAR model by the LSTM (a deep learning) model.

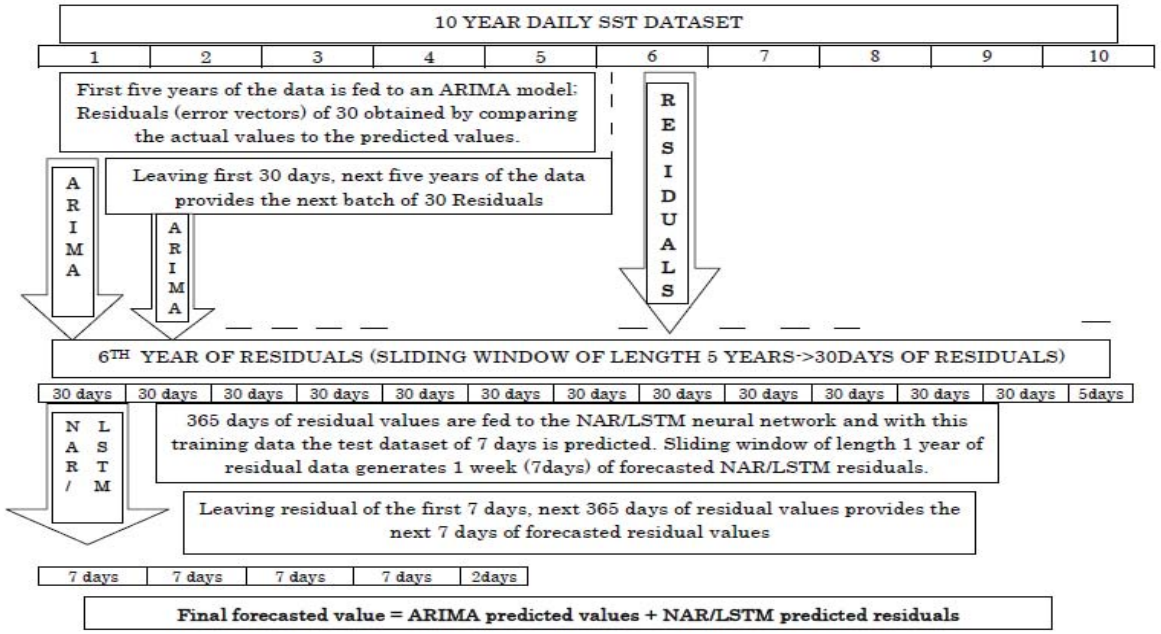


Fig 6.2 Representation of the proposed model in terms of numerical components

As shown in Figure 6.2, a 10 year dataset is first split into a year data each. Using five years of data, a set of 30 days of value is predicted using ARIMA model and Residuals are calculated. This is now a sliding window kind of approach, where the window shifts by 30 days every time. Using 365 days of residual values as input to a NAR network, a week ahead (7 days) of data is predicted. To these 7 values, now the previously predicted ARIMA residuals are added thereby providing the final outcome of the proposed Hybrid model. This is how we proceeded taking the results and plotted the same in a time series plot. Also simultaneously, all ten error parameters mentioned under Section 3.4 were also evaluated. Fig. 6.3 shows the time series plot that we obtained when implemented for a dozen (?) iterations. In the results of the prediction, a significant mismatch is observed in the values of the actual SST and the predicted SST values during the period of 85 days to 91 days approximately.

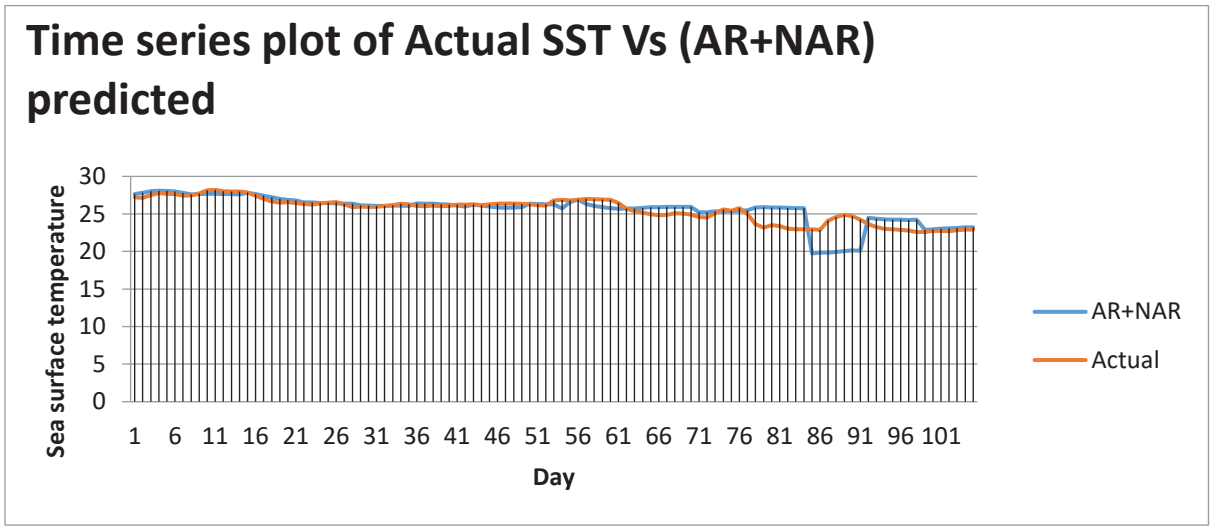


Fig 6.3 SST Time series prediction using hybrid AR+NAR approach

Continuing the observation from the same figure (Fig.6.3), there is complete agreement of the actual and predicted data in the initial part of prediction. This continued for approximately 10 iterations of multistep prediction worth 7 days each. Around the 12^{th} iteration, a very significant deviation in the prediction identifies observed. In our experiments, the same deviation was observed for different step size too. And this was the considered to be challenging part in the data set as it might have been reflected as vital climatic changes in the Pacific Ocean. We were not in much favor of changing the step size because 7 days (or one week) of multistep prediction was targeted. Moreover, 7 is a prime number.

So, we attempted to provide some diversity into the proposed algorithm. And this brings about the novelty in it. We did some trial and realized that if we take dynamic step size in a patterned manner, we could probably combat this deviation in performance. In the experimental observations, it was realized that whenever the ARIMA and the NAR window coincided, there was this repetitive performance deviation that was turning up throughout the entire time series. In near vicinity of the window, the errors tend to propagate and hence non coinciding windows for the linear and nonlinear part are mandatory to ensure that the errors are restricted. So, a pattern was selected where both the windows do not coincide. A set of patterns were experimented and after a set of attempts; it was identified that the pattern in steps of +9,+9,+7,+7,+7,+7,+7,+7,+9,+9……..

Ensured that the ARIMA and the NAR prediction windows do not coincide. The change in the SST prediction after adapting this pattern of varied step size is shown in Fig. 6.4.

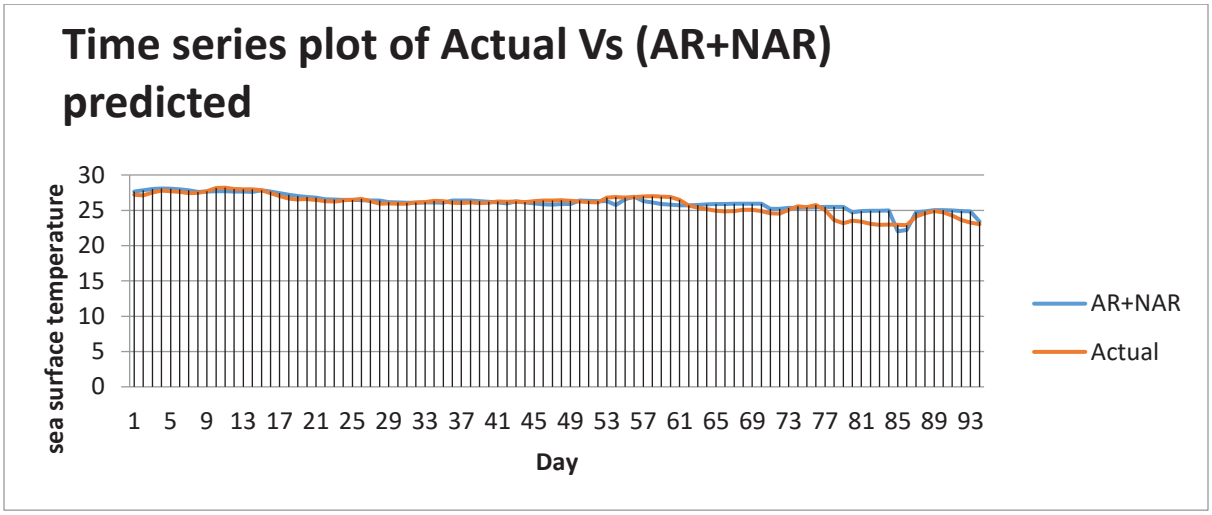


Fig 6.4 SST Prediction using hybrid AR+NAR approach after adopting varied stepsize

The experiment was conducted iteratively (for 234 iterations) till the coverage of the entire dataset. And throughout the entire time series, the same pattern was applied.

The same experiment was performed for another hybrid AR+LSTM model also. A similar deviation of performance was observed in the prediction for this hybrid approach too. This kind of deviation was observed in the vicinity of the series, where both the window boundaries coincided. And the similar approach of adopting dynamic stepsize was used to avoid performance deviations. Hence, in the entire experiments of both hybrid models (ARIMA+NAR and ARIMA+LSTM), a predefined pattern was used as dynamic step adoption.

The final proposed hybrid method is shown in the Figure 6.5. It is an iterative approach and the entire dataset is shown split in the desired manner so as to predict SST values very near to the actual values. The ARIMA model gives rise to a set of linear residuals by linear predictions and the NAR and the LSTM models are used to predict a set of non-linear residuals. The final predicted value is the algebraic addition of the ARIMA predicted values to the Non-linearly (using NAR/LSTM) predicted residuals. The algorithm for Prediction of SST using proposed hybrid method (shown in Fig. 6.5) is presented as Algorithm 6.1. The specific values of

variables *n, p* and *k* used in our experiment are also presented as additional elaboration in appropriate steps for effective understanding of the algorithm.

Algorithm 6.1: Hybrid Method for SST Prediction

1] Select the daily SST time series

Here, Elnino dataset is considered for experiment.

2] Let *n* be the number of time stamps of SST time series, *p* is the number of time stamps considered for initial linear predictions using ARIMA model, (*k–p)* indicate prediction step size for hybrid model for each iteration

Example: Here n = 10 years of SST data, p = 5 years of SST data; k = 1 year of SST data.

3] Perform Linear predictions of the SST time series data using ARIMA model and compute residuals.

Example: Here, ARIMA makes predictions of last 5 years of SST data and 30 residuals are calculated. This leads to a multistep prediction in steps of 30 each distributed over the remaining time series of 5 + years.

4] Provide residuals obtained in step-3 for creating nonlinear prediction using NAR and LSTM individually. Perform a week ahead predictions of residuals using NAR and LSTM. Also identify and adopt dynamic optimized time stamp pattern.

A set of 1 year of such residuals is fed to the NAR/LSTM network to predict 7 residuals. So, basically this leads to a multistep prediction of 7 residuals each.

5] Perform algebric addition of ARIMA predictions with NAR/LSTM predictions for matching time stamps.

The predicted values using ARIMA (linear) model are preserved. These (non-linear) residual values are algebraically added to the ARIMA predicted values to give the final forecasted values.

6] Forward the predictions of ARIMA and NAR as well as LSTM individually with a specified prediction window size.

7] Repeat steps 3-7 till prediction of remaining SST test time series data.

8] Plot the predicted responses for both hybrid models ARIMA+NAR and ARIMA+LSTM. Also compute the error measures (as mentioned in Section 3.4) for the predicted responses.

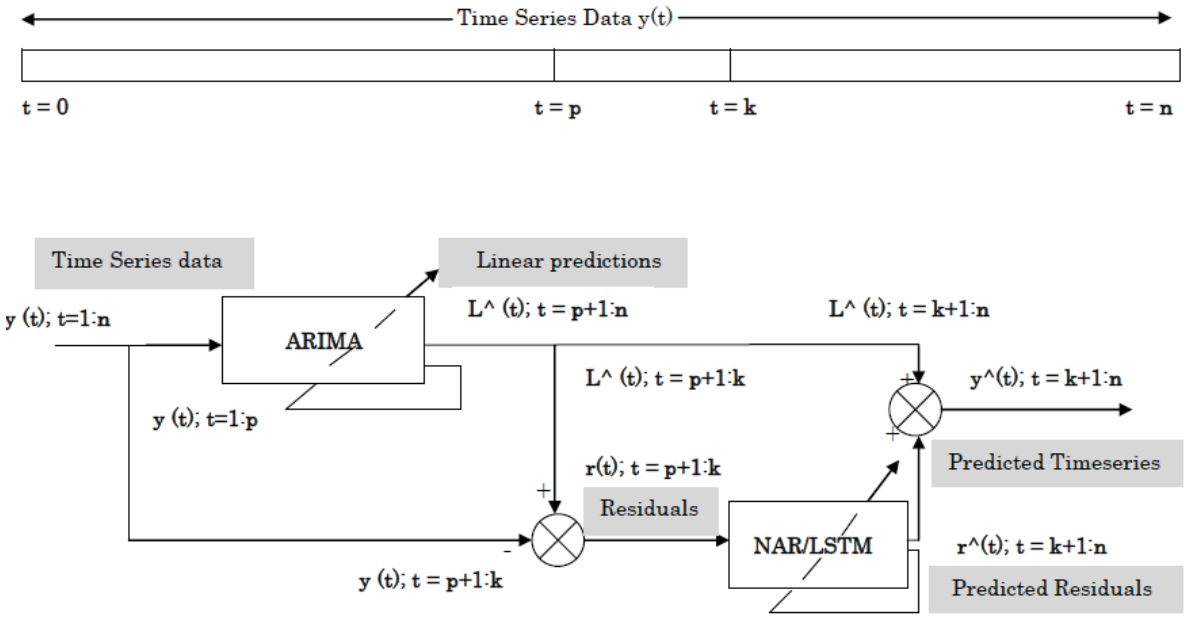


Fig.6.5 SST time series prediction by the proposed hybrid method

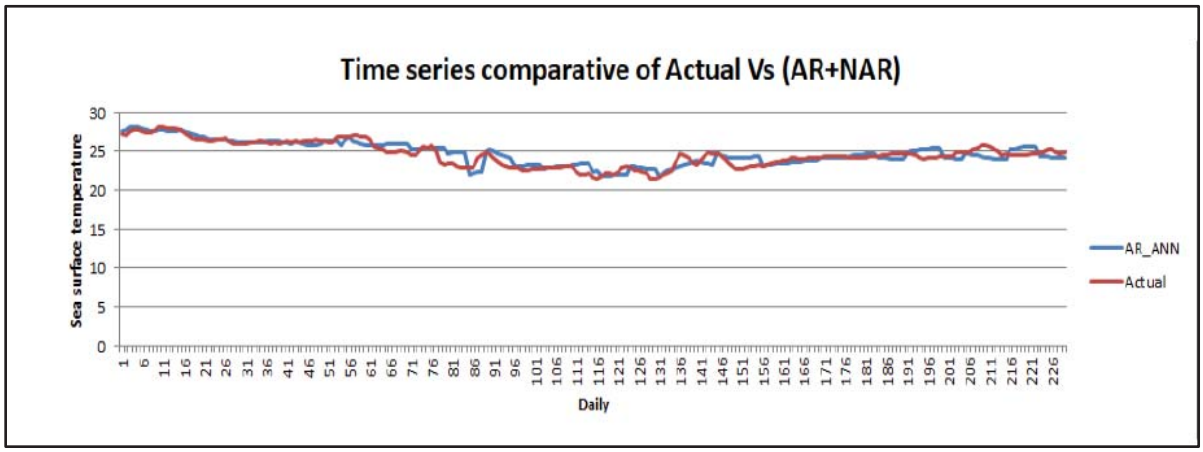


Fig 6.6 Time response of the Actual vs (ARIMA+NAR) using optimized hybrid model

In the figures, Fig 6.6 and Fig 6.7, the prediction responses of individual hybrid models (ARIMA+NAR and ARIMA+LSTM) are shown.

Fig 6.6 depicts the response of the optimized hybrid model using ARIMA and NAR models in cohesion using the optimized pattern of step size. It can be observed from the figure that the prediction shown good match with the actual SST time series. Fig 6.7 depicts the response of the optimized hybrid model using ARIMA+LSTM models in cohesion using the optimized pattern of step size. The results are observed to be appreciative except for a few points where marginal error can be also be noted.

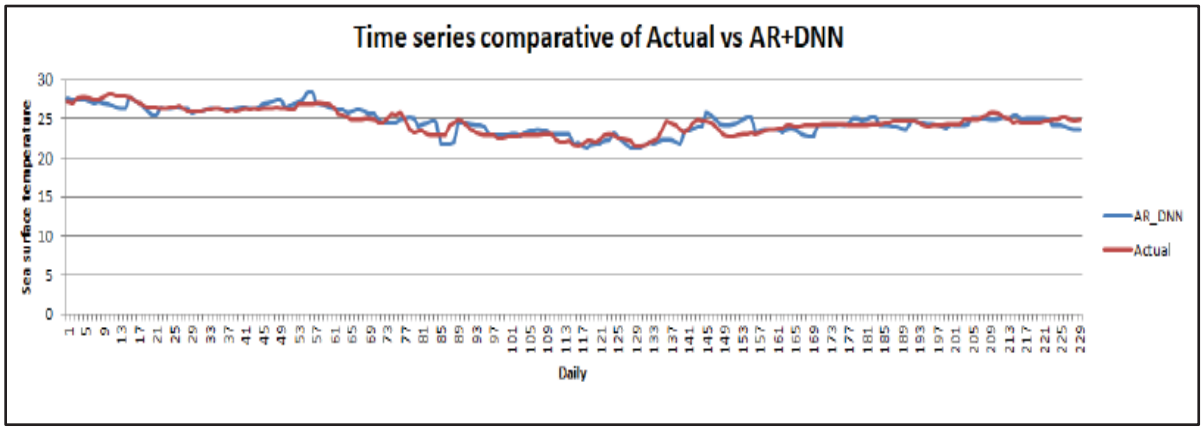


Fig 6.7 Time response of the Actual vs (ARIMA+LSTM) using optimized hybrid model

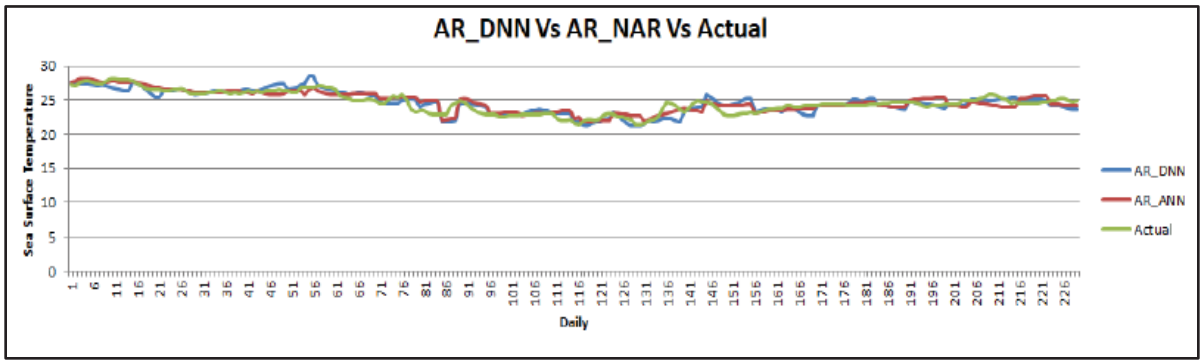


Fig 6.8 SST prediction responses of optimized hybrid models AR+NAR and AR+LSTM.

The figure 6.8 shows the SST time series responses of the hybrid model (AR+NAR) as compared to the hybrid model (AR+LSTM) with the actual data over more than 30 iterations, amounting to more than 200 timestamps. It is observed that the predicted and actual time series are in agreement to each other.

Table 6.1 The error values using (AR+NAR) as the proposed hybrid model

Week	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDD	SDE	Sres
9	0.7191	0.1406	0.848	0.1658	1.1354	0.7716	2.6791	0.9016	0.4254	0.4487
10	0.7762	0.1997	0.881	0.2266	1.0626	0.8604	1.5922	0.8627	0.2045	0.4662
11(9)	0.5044	0.1203	0.7102	0.1694	1.8744	0.4796	2.0455	0.8701	0.6593	0.3044
12(9)	2.5589	0.4883	1.5996	0.3053	2.0279	1.5335	2.645	0.8881	1.4807	0.6856
13	0.5444	0.0928	0.7379	0.1258	1.1446	0.6568	2.2784	0.974	0.363	0.3904
14	0.1994	0.034	0.4466	0.0762	0.6141	0.4081	1.1161	0.9665	0.1958	0.2363
15	0.0172	0.0029	0.1311	0.0224	0.2609	0.1085	1.4233	0.9603	0.0794	0.0693
16	1.1341	0.1934	1.0649	0.1816	1.4542	0.9334	1.1812	0.9664	0.5517	0.5635
17	0.167	0.0285	0.4086	0.0697	0.9181	0.3011	2.0181	0.976	0.3563	0.2162
18	0.3463	0.0591	0.5885	0.1004	1.166	0.496	1.3796	0.9825	0.37	0.3114
19(9)	0.4902	0.0836	0.7002	0.1194	1.4311	0.6158	1.1043	0.9845	0.7332	0.3001
20(9)	0.7739	0.132	0.8797	0.15	1.6627	0.6251	1.0671	0.9929	0.8227	0.377
21	1.455	0.2481	1.2062	0.2057	1.6781	1.0201	1.1592	1.0025	0.8103	0.6383
22	0.0167	0.0028	0.1292	0.022	0.2248	0.1007	1.9468	1.0138	0.139	0.0683
23	0.0751	0.0128	0.2741	0.0468	0.4087	0.2393	2.078	1.0117	0.1443	0.1451
24	0.0218	0.0037	0.1476	0.0252	0.248	0.1317	2.5254	1.0033	0.072	0.0781
25	0.1348	0.023	0.3671	0.0626	0.5364	0.3359	1.0384	1.0029	0.16	0.1943
26	0.1546	0.0264	0.3932	0.0671	0.658	0.347	1.6229	1.0019	0.1996	0.2081
27(9)	0.8618	0.147	0.9283	0.1583	1.2781	0.8348	0.5909	0.9951	0.4306	0.3979
28(9)	0.2158	0.0368	0.4645	0.0792	0.7878	0.4016	0.8517	0.9886	0.2477	0.1991
29	1.6742	0.2855	1.2939	0.2207	1.6739	1.2184	1.3713	0.9829	0.4705	0.6847
30	0.88	0.1501	0.9381	0.16	1.2258	0.8851	2.4696	0.9809	0.3358	0.4964
31	0.4567	0.0779	0.6758	0.1153	1.0453	0.6045	0.9971	0.9739	0.3264	0.3576
32	0.128	0.0218	0.3578	0.061	0.5809	0.3076	1.0344	0.9727	0.2423	0.1894

The error measures (described in Section 3.4) calculated33 based on predictions of proposed hybrid method ARIMA+NAR are shown in Table 6.1. The prediction begins with the 9th week of observation and extends to the 32nd week. In between, there is multistep prediction of (9 days, twice) after every span of six 1 week time. Out of the entire dataset, which is slightly larger than 10 years, it is observed that as per the proposed model, the linear predicted values are algebraically added to the non-linearly (NAR) predicted residuals. The table contains a randomly selected set of details taken from a complete set of 234 multistep iterations that covers the entire test dataset. The multistep pattern that is followed for this is (in days)

+7, +7, +7, +7, +7, +7, +9, +9, +7,……

Notice that the RMSE and the NRMSE values are in the range of (0.1292 to 1.2939) and (0.022 to 0.2266) respectively which is very low as compared to previous chapter methods (ARIMA, NAR and LSTM models. Throughout all observations, it is observed that the value of Standard Deviation of Errors (SDE) < Root Mean Square Error (RMSE). This confirms that the proposed hybrid method is stronger than the mean in predicting the SST values. Also the value of Residual Standard Deviation (SRes) that represents the deviation of the residual value about the regression line is always found to be less than 0.68. The MAPE value is also limited to 2.64.

Table 6.2 The error parameters using (AR+LSTM) as the proposed hybrid model

Week	D_MSE	D_NMSE	D_RMSE	D_NRMSE	D_MaxAE	D_MeanAE	D_MAPE	D_SDD	D_SDE	D_Sres
9	0.0012	0.0011	0.0353	0.03	0.727	0.3248	0.7775	0.9016	0.4292	0.0495
10	0.7979	1.7711	0.8932	1.9828	1.33	0.8932	0.6665	0.8627	0.3223	1.2505
11(9)	0.0081	0.0039	0.09	0.0434	1.9261	0.917	0.951	0.8701	0.1133	0.1157
12(9)	0.0102	0.0026	0.1012	0.0258	2.5971	1.4745	2.1002	0.8881	0.1697	0.1301
13	0.2492	0.1407	0.4992	0.2819	1.1098	0.6596	2.3876	0.974	0.6165	0.6989
14	0.0431	0.1349	0.2075	0.65	0.3515	0.2075	0.2762	0.9665	0.1116	0.2905
15	0.1617	0.4489	0.4021	1.1166	0.7069	0.4588	0.8279	0.9603	0.2916	0.5629
16	0.4315	0.4252	0.6569	0.6474	1.1389	0.6654	0.5364	0.9664	0.4429	0.9196
17	0.3431	0.2694	0.5858	0.4599	0.9311	0.5858	1.5673	0.976	0.2642	0.8201
18	0.0547	0.0352	0.234	0.1505	0.7914	0.449	1.0609	0.9825	0.4833	0.3276
19(9)	1.3773	0.4259	1.1736	0.3629	2.35	1.1736	1.2335	0.9845	0.9478	1.5089
20(9)	0.0179	0.0076	0.1339	0.0571	1.0309	0.733	2.5031	0.9929	0.8192	0.1722
21	2.4029	1.3805	1.5501	0.8906	2.4304	1.5501	1.2677	1.0025	0.917	2.1702
22	0.0052	0.0096	0.0722	0.1325	0.435	0.2151	0.2712	1.0138	0.2743	0.1011
23	0.6103	1.268	0.7812	1.6231	1.39	0.7812	0.78	1.0117	0.5335	1.0937
24	0.0095	0.0414	0.0974	0.4246	0.1849	0.0974	2.9902	1.0033	0.048	0.1364
25	0.528	0.3946	0.7267	0.5431	1.0255	0.7267	2.3522	1.0029	0.2631	1.0173
26	0.2179	0.5443	0.4668	1.1661	0.9357	0.4668	1.6825	1.0019	0.3206	0.6535
27(9)	0	0	0.0173	0.0266	0.4416	0.187	1.0592	0.9951	0.2468	0.0222
28(9)	0.0434	0.0698	0.2082	0.335	0.5511	0.2594	0.836	0.9886	0.2624	0.2677
29	0.0296	0.0289	0.1721	0.1681	0.589	0.3582	0.4347	0.9829	0.3873	0.2409
30	0.2483	0.4522	0.4983	0.9076	0.593	0.4983	0.6937	0.9809	0.0909	0.6976
31	0.9219	1.2096	0.9602	1.2598	1.5424	0.9602	1.8382	0.9739	0.4639	1.3443
32	0	0	0.0097	0.0187	0.4921	0.1942	0.2915	0.9727	0.2628	0.0135

The error measures (described in Section 3.3) calculated based on predictions of proposed hybrid method ARIMA+NAR are shown in Table 6.2. The prediction begins with the 9th week of observation and extends to the 32nd week. In between, there is multistep prediction of (9 days, twice) after every span of six 1 week time. The dataset continues to be the same; slightly larger than 10 years, it is observed that as per the proposed model, the linear predicted values are algebraically added to the non-linearly predicted (LSTM) residuals. The table contains a randomly selected set of details taken from a complete set of 234 multistep iterations that covers the entire test dataset. The multistep pattern that is followed for this is (in days) also the same.

+7, +7, +7, +7, +7, +7, +9, +9, +7,……

A little discussion on the various metrics that evaluate the performance of prediction error parameters in the preceding table is shared. The notation used is same as before with the letter D (for example MSE is replaced by D_MSE for Deep Neural based approach and so on for all error metrics) representing the Deep NN based approach. The Normalized MSE and RMSE ensure that the metrics will stay unbiased while estimating the predicted values. MAPE measures the magnitude of the error in percentage units. MAPE in the range of 0.29 to 2.9 is the observed values measured. Standard deviation of data provides an idea on the spread/distribution of the individual SST values from the mean of the series. And standard deviation of error provides the same for error. It is observed that the standard deviation of the data is near to 1.00 and at the same time the standard deviation of the error is in the range of 0.090 to 0.900 which ensures that the SDE < SDD for practically all readings. The agreement in the numerical values is alone not sufficient to serve as a base for the goodness of fit. Simultaneous study of the time series plots is available in fig 6.6, 6.7, 6.8 for hybrid of (AR+NAR), (AR+LSTM) and (Actual vs (AR+NAR) and (AR+LSTM)) respectively. As evident from the time series plots in the figures (Fig. 6.6, Fig. 6.7, Fig. 6.8), the actual SST values are predicted to a large extent using the proposed hybrid models.

Table 6.3 Comparative of all methods against the proposed hybrid model

Algorithm	MSE	NMSE	RMSE	NRMSE	MaxAE	MeanAE	MAPE	SDE	Sres	CC	NSE
ARIMA	1.294	0.1273	1.1375	0.1136	1.1823	0.5494	0.1500	0.4837	0.0997	97.1756	0.7472
NAR	1.185	0.1187	1.0862	0.1081	1.7983	0.9787	0.1413	0.5096	0.1848	96.7585	0.6985
LSTM	1.792	0.1792	1.3386	0.1338	2.622	1.3615	0.1506	0.6861	0.2571	96.6179	0.7104
AR + NAR	0.596	0.0596	0.7728	0.0772	**1.0457**	0.5123	**0.0592**	**0.4091**	0.3344	**99.6454**	**0.9904**
AR + LSTM	**0.354**	**0.0357**	**0.5949**	**0.0594**	1.0667	**0.4582**	0.1243	**0.3784**	**0.0685**	99.0681	**0.9905**

As evident from Table 6.3, the minimum values of the errors are recorded for both the proposed hybrid model with CC being as high as 99%. The same is accompanied by the value of NSE in the range of 0.9904-0.9905. Also observe that the Standard Deviation of Error is also minimum for the Hybrid model (AR+LSTM) and the value is as low as 0.3784. The MAPE is lowest for the (AR+ NAR) model and is evaluated at a value of 0.0592. The residual standard deviation is 0.0685 and is recorded for the (AR + LSTM) model. The MeanAE and the MaxAE values are respectively 0.4582 (using AR + LSTM) and 1.0457 (using AR + NAR). The minimum values of MSE, NMSE, RMSE and NRMSE are recorded as 0.354, 0.0357, 0.5949 and 0.0594 respectively using the (AR + LSTM) model of implementation,

6.4 Chapter Conclusion

In this chapter, a hybrid method with two versions of a hybrid models – ARIMA+NAR and ARIMA+LSTM are presented, and it was used for prediction of Daily SST data. An algorithm representing steps of the proposed method is also presented.

The presented hybrid model explores the linear characteristics of the SST dataset using ARIMA model Box Jenkin's methodology and non-linear characteristics of the NAR/LSTM model. During experiment, it was observed that at some specific length intervals, the multistep prediction tends to offer deviation in prediction performance. A variable length of multistep prediction having a predefined pattern is experimentally identified and adopted. The final method and algorithm were tested on Elnino daily dataset. Comparing the results of proposed hybrid method with previous chapter methods indicate improvements in SST prediction. The

minimum values of the errors are recorded for both the proposed hybrid model with CC being as high as 99%. The same is accompanied by the value of NSE in the range of 0.9904-0.9905. Also observe that the Standard Deviation of Error is also minimum for the Hybrid model (AR+LSTM) and the value is as low as 0.3784. The MAPE is lowest for the (AR+ NAR) model and is evaluated at a value of 0.0592. The residual standard deviation is 0.0685 and is recorded for the (AR + LSTM) model. The MeanAE and the MaxAE values are respectively 0.4582 (using AR + LSTM) and 1.0457 (using AR + NAR). The minimum values of MSE, NMSE, RMSE and NRMSE are recorded as 0.354, 0.0357, 0.5949 and 0.0594 respectively using the (AR + LSTM) model of implementation.

CHAPTER 7

Conclusion and Future Work

7.1 Conclusion

In this work, we have improved the performance of existing algorithms by optimizing the parameters in a site specific approach. The algorithms addressed are the basic linear algorithm- the ARIMA model and the most commonly implemented nonlinear algorithm – the Neural Network model, specifically the Feed Forward Neural Network using back propagation of error. Apart from that, a Deep Neural Network algorithm, specifically the Long Short Term Model is implemented whose performance is comparable to the regular Neural Network.

While working with these datasets, it is observed that the monthly datasets show good performance with pure basic models even for multistep prediction. Few of the algorithms proposed in previous literature review are also optimized so as to achieve better precision not only for short term forecasting but also in the long run. However, for daily SST datasets, the error increases and a very evident mismatch is observed in the time series plots.

1) The main objective of this work is investigation and development of machine learning and deep learning based methods for prediction of Sea Surface Temperature. This is achieved in Chapter 4 to maximum extent where the machine learning methods are compared and evaluated in the backdrop of monthly and diurnal SST databases. The proposed methods involve the study of the ARIMA and the NAR models. Both of these Machine Learning techniques are able to generate the forecasted values with minimum errors. The Correlation Coefficient (CC) is found to be near to 97.5% for both the ARIMA and the NAR method. The values of NSE is 0.96 (ARIMA) and 0.95 (NAR). Chapter 5 addresses the deep learning based, LSTM method of prediction that shows a better performance.

2) Another important objective is to develop hybrid method based on traditional time series prediction method with machine learning and deep learning methods and test it for prediction of SST using various dataset. The aim to develop hybrid method is to improve the prediction accuracy of diurnal SST as compared to literature methods. This is achieved in Chapter 6 where SST values are predicted using a pattern of predefined multi steps. It is to be noted that the ARIMA and NAR/LSTM windows should not overlap at the window periphery else the error tends to propagate and amplify. The same is justified for a daily SST dataset. The minimum values of the errors are recorded for both the proposed hybrid model with CC being as high as 99%. The same is accompanied by the value of NSE in the range of 0.9904-0.9905. Also observe that the Standard Deviation of Error is also minimum for the Hybrid model (AR+LSTM) and the value is as low as 0.3784. The MAPE is lowest for the (AR+ NAR) model and is evaluated at a value of 0.0592. The residual standard deviation is 0.0685 and is recorded for the (AR + LSTM) model. The MeanAE and the MaxAE values are respectively 0.4582 (using AR + LSTM) and 1.0457 (using AR + NAR). The minimum values of MSE, NMSE, RMSE and NRMSE are recorded as 0.354, 0.0357, 0.5949 and 0.0594 respectively using the (AR + LSTM) model of implementation,

3) One of the important objectives is the development and investigation of above methods to perform and evaluate single-step and multi-step prediction for SST time series dataset. As far as single step prediction of monthly SST data is concerned, almost all machine learning and deep learning methods provide desired outputs. However, for multi-step prediction of daily SST values, it is observed from Chapter 4 and 5 that there is still some scope for improvement and hence the hybrid method is proposed in Chapter 6.

4) One of objectives is to perform comparison of methods developed and used for prediction of SST. Chapter 5 ensures that the same is addressed. It also highlights the significance of various error metrics that are used for SST timeseries prediction. Parallelly the time series plots are also studied. The time series plots show deviation when evaluated on a point to point basis although the error parameters are not significantly high. Nash Sutcliffe Efficiency (NSE), that is a very strong indicator of error calculations related to hydrological studies reflects NSE nearest to 1 is at 0.9746 (dataset 3, LSTM method) and the worst NSE is located

at 0.6911 (dataset 5, NAR). The CC value is highest at 97 (Dataset 1, ARIMA method; Dataset 1. NAR and Dataset 3, LSTM). The NRMSE is lowest at 0.0061 (dataset 3, LSTM).

5) An objective in this work is identification of SST anomalies from the SST time series datasets. A minor objective is also to Investigate and understand dependency relationship of SST with other oceanic parameters like Sea Surface Salinity (SSS), Sea Bottom Temperature (SBT), Zonal winds, Meridional winds and Air temperature in a site specific approach. The same is achieved in Chapter 4. For the single step calculations, the investigation of SSS and SBT is performed. Notice the values of MSE and RMSE- we observe that without any additional input time series, the values are 1.4315°C and 1.1964 °C respectively. However, with either of the supportive time series, there is a drastic reduction in the MSE value; it drops down to 0.3459 °C and 0.3462°C using temperature at 5m and surface salinity as reference respectively. Similarly, the RMSE value is 0.0912°C. The same is calculated to be 0.0220°C and 0.0221°C with SBT and SSS respectively as additional time series. For multistep prediction, the outcome reveals that the SSTA values computed are maximally matched when the Zonal and Meridional Winds are used as assistive time series along with the SST time series. With the Air Temperature as an additive time series, the RMSE value is 0.2788^0C - 0.934^0C. The same with Zonal winds is 0.3336^0C – 1.2535 ^{0}C and under the effect of Meridional winds is 0.2769^0C to 1.2286^0C. It is to be noted that the time series plot shows max agreement of the actual SSTA with the support of both Zonal and Meridional winds.

7.2 Future Work

There are a couple of aspects that is worth investigation.

ENSO events, the Asian Australian Monsoon, the IOD and the Pacific Niño sites are all correlated. It is observed during literature survey that they influence the occurrence of events in a highly dynamic manner. At times, this is evaluated to be distributed across multiple locations. The study of the SSTA and the ONI indications can help in harnessing the dynamic variation in the occurrence of sudden events and their adverse effects can also be restrained.

- The next study could involve study of the SSTA across various prominent locations.

- Looking at the success of the proposed hybrid model, we may target prediction of other parameters like Financial Market forecasting and Industrial manufacturing forecasting.
- Also we can target forecasting of other time series that tend to be dynamic in nature.

REFERENCES:

1] W.W. Hsieh and B. Tang, 1998, 'Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography', *Bulletin of the American Meteorological Society*, Vol. 79(9), pg.1855-1870.

2] F.T. Tangang, W.W. Hsieh and B. Tang, 1998, 'Forecasting regional sea surface temperatures in the tropical Pacific by neural network models, with wind stress and sea level pressure as predictors', *Journal Of Geophysical Research*, Vol. 103(C4), pg 7511-7522,

3] F.T. Tangang, W.W. Hsieh & B. Tang, 1997, 'Forecasting the equatorial Pacific sea surface temperatures by neural network models', *Climate Dynamics*, Vol. 13(2), pg 135–147.

4] B. Tang, W.W. Hsieh, A.H. Monahan and F.T. Tangang, 2000, 'Skill Comparisons between Neural Networks and Canonical Correlation Analysis in Predicting the Equatorial Pacific Sea Surface Temperatures'**,** *Bulletin of the American Meteorological Society*, pg 287-293.

5] R.W. Reynolds and T.M. Smith, 1994, 'Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation', *Journal of Climate*, Vol. 7, pg 929-948.

6] T.M. Smith and R.W. Reynolds, 2004 'Improved Extended Reconstruction of SST (1854–1997)', *Journal of Climate*, Vol. 17, pg 2466-2477.

7] J.S. Kug, S.I. An, F.F. Jin and I.S. Kang, 2005, 'Preconditions for El Nino and La Nina onsets and their relation to the Indian Ocean', *Geophysical Research Letters*, Vol. 32(5), pg 1-5

8] S.H. Yoo, S. Yang and C.H. Ho, 2006, 'Variability of the Indian Ocean Sea surface temperature and its impacts on Asian-Australian monsoon climate', *Journal of Geophysical Research*, Vol. 111 (D03) pg 108-125.

9] K.E. Trenberth, 1990, 'Recent Observed lnterdecadal Climate Changes in the Northern Hemisphere', *American Meteorological Society*, Vol. 71(7), pp 983-993.

10] D.J. Shea, K.E. Trenberth, 1992, 'A global monthly Sea Surface Temperature Climatology', *American Meteorological Society*, Vol. 5, pp 987-1001.

11] K.E. Trenberth, J.W. Hurrell, 1994, 'Decadal atmosphere-ocean variations in the Pacific', *Climate Dynamics*, Vol. 9, pp 303-319.

12] K.C. Tripathi, I.M.L. Das, A.K Sahai, 2006, 'Predicatability of sea surface temperature anomalies in the Indian Ocean using Artificial Neural Networks', *Indian Journal of Geo Marine Sciences*, Vol. 35(3), pg 210-220.

13] K.C. Tripathi, S. Rai, A.C. Pandey & I.M.L. Das, 2008, 'Southern Indian Ocean SST Indices as early predictors of Indian Summer Monsoon', *Indian Journal of Marine Sciences*, Vol. 37 (1), pg 70-76.

14] S. B. Mohongo, M.C. Deo, 2013, 'Using Artificial Neural Networks to forecast monthly and seasonal sea surface temperature anomalies in the western Indian Ocean', *International Journal of Ocean and Climate Systems*, Vol. 4(2), pg 133-150.

15] K. Patil, M.C. Deo, S. Ghosh and M. Ravichandran, 2013, 'Predicting Sea Surface Temperatures in the North Indian Ocean with Nonlinear Autoregressive Neural Networks', *International Journal of Oceanography*, Vol. 2013, 11 pages.

16] K. Patil, M.C. Deo and M. Ravichandran, 2016, 'Prediction of Sea Surface Temperatures by combining Numerical and Neural Techniques', *Journal of Atmospheric and Oceanic Technology*, Vol. 33, pg 1715-1726.

17] L. Hu, M. HE, 2014, 'Impacts of Sea Surface Temperature Anomaly to the coverage area and early appearance time of green tide in the Yellow Sea', *Conference: IGARSS IEEE International Geoscience and Remote Sensing*, pg 4465-4468.

18] V.C. Andreo, A.I. Dogliotti and C.B. Tauro, 2016, '*IEEE Journal of elected Topics in Applied Earth Observations and Remote Sensing*', Vol. 9(12), pg 5315-5324.

19] A.Wu, W.W. Hseih, B.Tang, 2006, '*Neural network forecasts of the tropical Pacific sea surface temperatures*', Neural Networks Vol. 19, pg 145–154.

20] S. G. Aparna, S. D'Souza & N. B. Arjun, 2018, 'Prediction of daily sea surface temperature using artificial neural networks' *International Journal of Remote Sensing,* Vol. 39(12), pg 4214–4231.

21] Q. Zhang, H. Wang, J. Dong, G. Zhong, X. Sun, 2017, 'Prediction of Sea Surface Temperature Using Long Short-Term Memory', Submitted to *IEEE Geoscience and Remote Sensing Letters*, pg 1-6.

22] V. Athira, P. Geetha, R. Vinaykumar, K. P.Soman, 2018, 'DeepAirNet: Applying Recurrent Networks for Air Quality Prediction', *Procedia Computer Science*, Vol. 132 pg 1394-1403.

23] R.J. Hyndman, A.B. Kohler, 2005, 'Another look at measures of forecast accuracy', *International Journal of Forecasting*, Vol. 24(6), pg 389-402

24] G.P. Zhang, 2003, 'Time series forecasting using a hybrid ARIMA and neural network model' *Neurocomputing*, Vol. 50, pg 159-175.

25] A.K. Fard, Md. R. A Zadeh, 2014, 'A hybrid method based on wavelet, ANN and ARIMA model for short-term load forecasting', *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 26(2) pg 167-182.

26] I. Khandelwal, R. Adhikari, G. Verma, 2015, 'Time Series Forecasting using Hybrid ARIMA and ANN Models based on DWT Decomposition', *Procedia Computer Science* Vol. 48, pg 173 – 179.

27] Z. Zhao, W. Chen, X. Wu, Peter. C. Y. Chen, J. Liu, 2017, 'LSTM network: a deep learning approach for short-term traffic forecast', *IET Intelligent Transport Systems*, Vol. 11(2), pp. 68-75.

28] J.A. McGowan, E. R. Deyle, H. Ye, M. L. Carter, C. T. Perretti, K.D. Seger, A.D. Verneil, G. Sugihara, 2018, 'Predicting coastal algal blooms in southern California' *Ecology*, Vol. 98(5), pg 1419-1433.

29] G.E.P. Box & G.M. Jenkins 1989, 'An Unexpected Route to Time Series', *Time series analysis: forecasting and control*, Vol. 44, pg 575.

30] E. Ursu, J.C. Pereau, 2017, 'Estimation and identification of periodic autoregressive models with one exogenous variable' *Journal of the Korean Statistical Society*, Vol.46(4), pg 629-640.

31] G. Mahalakshmi, S. Sridevi, S. Rajaram, 2016, 'A Survey on Forecasting of Time Series Data', A Survey on Forecasting of Time Series Data, *IEEE International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, pg 1-8.

32] Nicholas I. Sapankevych, Ravi Sankar, 2009, 'Time Series Prediction Using Support Vector Machines: A Survey', *IEEE Computational Intelligence Magazine*, pg 24-38.

33] E. Safavieh, S. Andalib, and A. Andalib, 2007, 'Forecasting the Unknown Dynamics in NN3 Database Using a Nonlinear Autoregressive Recurrent Neural Network' *IEEE Proceedings of International Joint Conference on Neural Networks*, USA, pg 1-5.

34] A. Geetha, G.M. Nasira, 2016, 'Time series modeling and forecasting: Tropical Cyclone prediction using ARIMA model', *3rd IEEE International Conference on Computing for Sustainable Global Development (INDIACom),* pg 3080-3086.

35] I. M.Yassin, A. Zabidi, M. K. M. Salleh, N. E. Abdul Khalid, 2013, 'Malaysian Tourism Interest Forecasting using Nonlinear Auto-Regressive (NAR) Model', *IEEE 3rd International Conference on System Engineering and Technology,* Malaysia, pg 32-36.

36] S. N. Kadir, N.M. Tahir, I. M. Yassin, A. Zabidi, 2014, Kota Kinabalu, Malaysia, 'Malaysian Tourism Interest Forecasting using Nonlinear Auto-Regressive Moving Average (NARMA) Model', *IEEE Symposium on Wireless Technology and Applications (ISWTA)*, pg 193-198.

37] S.M. Gupta, B.A. Malmgren, 2009, 'Comparison of the accuracy of SST estimates by artificial neural networks (ANN) and other quantitative methods using radiolarian data from the Antarctic and Pacific Oceans', *Earth Science India* Vol.2 (II), pg 52 -75.

38] E. G. Gorriz, J.G.Sanchez, 2007, 'Prediction of sea surface temperatures in the western Mediterranean Sea by neural networks using satellite observations', *Geophysical Research Letters*, Vol. 34 L11603, pg 1-6.

39] J. Wang, F. Wei, J. Feng, 2011, 'Impact Of Sea Surface Temperature Anomaly on two Global Warming Periods in the 20th Century over the Arid Central Asia', *IEEE International Geoscience and Remote Sensing–IGARSS'11*, pg 3249-3252.

40] Y.I. Xue, A. Leetrnaa, 2000,'Forecasts of tropical Pacific SST and sea level using a Markov Model', *Geophysical Research Letters*, Vol. 27(17), pg 2701-2704.

41] M. Khashei, Z. Hajirahimi, 2018, 'A comparative study of series ARIMA/MLP hybrid models for stock price forecasting', *Communications in Statistics - Simulation and Computation*, Vol. 48(9), pg 1-16.

42] L. Wei, L. Guan, L. Qu, L. Li, 2019, 'Prediction of Sea Surface Temperature in the South Sea by Artificial Neural Networks', *IEEE International Geoscience and Remote Sensing–IGARSS'19,* pg 8158-8161.

43] R. Salles, P. Mattos, A. M. D. Iorgulescu, E. Bezerra, L. Lima and E. Ogasawara, 2016 'Evaluating Temporal Aggregation for Predicting the Sea Surface Temperature of the Atlantic Ocean', *Ecological Informatics*, Vol. 36, pg 94-105.

44] Spyros Makridakis, 1993, 'Accuracy measures: theoretical and practical concerns', *International Journal of Forecasting* Vol. 9, pg 527-529.

45] C. N. Babu, B. E. Reddy, 2014, 'A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data', *Applied Soft Computing*, Vol.23, pg 27-34.

46] R.W. Reynolds, T.M.Smith, 1995, 'A high resolution Global Sea Surface Temperature Climatology', *Journal of Climate,* Vol. 8, pg 1571-1583.

47] R.W. Reynolds, D.C. Marsico, 1993, 'An Improved real time Global Sea Surface Temperature Analysis', Journal of Climate, Vol. 6, pg 114-119.

48] Y.Xue, R.W. Reynolds, T.M.Smith, 2001, 'A New SST Climatology for the 1971-2000 Base Period and Interdecadal Changes of 30-Year SST Normal', *Proceedings of the Twenty-Sixth Annual Climate Diagnostics and Prediction Workshop, San Diego*, pg 1-4.

49] D.J. Shea, K.E. Trenberth, R.W. Reynolds, 1990, 'A Global Monthly Sea Surface Temperature Climatology', *NCAR TECHNICAL NOTE* November 1990.

50] R.W. Reynolds, 1988, 'A real time Global Sea Surface Temperature Analysis', *Journal of Climate*, Vol. 1 pg 75-86.

51] K. E. Trenberth, 2012, 'Framing the way to relate climate extremes to Climate change', *Climatic Change*, Vol. 115, pg 283–290.

52] J. W. Hurrel, J. J. Hack, D. Shea, J. M. Caron, J. Rosinski, 2008, 'A New Sea Surface Temperature and Sea Ice Boundary Dataset for the Community Atmosphere Model', *Journal of Climate*, Vol. 21, pg 5145-5153.

53] K. E. Trenberth, 2011, 'Changes in precipitation with climate change', *Climate Research*, Vol. 47, pg 123-138.

54] S.Dominiak, P. Terray, 2005, 'Improvement of ENSO prediction using a linear regression model with a southern Indian Ocean sea surface temperature predictor', *Geophysical Research Letters*, Vol. 32, pg 1-4.

55] D. J. Gaffen, W.P. Elliot, A. Robock, 1992, 'Relationships between Tropospheric Water Vapor and Surface Temperature as observed by Radiosondes', *Geophysical Research Letters*, Vol. 19 (18), pg 1839-1842.

56] W.G. Pichel, 1991, 'Operational production of multichannel sea surface temperature from NOAA polar satellite AVHRR data', *Palaeogeography, Palaeoclimatology, Palaeoecology (Global and Planetary Change Section*), Vol. 90 pg 173-177.

57] C.C. Walton, 1988, 'Nonlinear Multi Channel Algorithms for Estimating Sea Surface Temperature with AVHRR Satellite data', *Journal of Applied Meteorology*, Vol. 27, pg 115-124.

58] A.G. Barnston, C.F. Ropelewski, 1991, 'Prediction of ENSO episodes using Canonical Correlation Analysis', *Journal of Climate*, Vol. 5, pg 1316-1435.

59] M.K. Tippett, A.G. Barnston, D.G. Dewitt, 2005, 'Statistical Correction of Tropical Pacific Sea Surface Temperature Forecasts', *Journal Of Climate*, Vol. 18, pg 5141-5162.

60] G. Zhang, B. E. Patuwo, M.Y. Hu, 1998, 'The state of the art: Forecasting with Artificial Neural Networks', *International Journal of Forecasting* Vol. 14, pg 35–62.

61] D.E. Rumelhart, G.E. Hinton and R.J Williams, 1986, 'Learning representations by back-propagating errors', *Nature*, Vol. 323, pg 533-536.

62] Z. Tang, C. Almeida, P.A. Fishwick, 2013, 'Time series forecasting using neural networks vs. Box- Jenkins methodology', *Simulation*, Vol.57 (303)., pg 303-310.

63] J.G.D. Gooijer, K. Kumar, 1992, 'Some recent developments in non-linear time series modelling, testing, and forecasting', *International Journal of Forecasting* Vol.8 pg 135-156

64] M. M. Ali, D. Swain, T. Kashyap, J. P. McCreary, P. V. Nagamani, 2013, 'Relationship Between Cyclone Intensities and Sea Surface Temperature in the Tropical Indian Ocean', *IEEE Geoscience And Remote Sensing Letters*, Vol. 10 (4), pg 841-844.

65] M. M. Ali, D. Swain, R.A Weller, 2004, 'Estimation of ocean subsurface thermal structure from surface parameters: A neural network approach' *Geophysical Research Letters,* Vol. 31, pg 1-4.

66] C. Alippi, V. Piuri, 1996, 'Experimental Neural Networks for Prediction and Identification', IEEE Transactions On Instrumentation And Measurement, Vol. 45(2), pg 670-676.

67] A. Sanghani, N. Bhatt and N. C. Chauhan, 2019, 'A Novel Hybrid Method for Time Series Forecasting Using Soft Computing Approach' *Proceedings of the International Conference on ISMACin Computational Vision and Bio-Engineering 2018 (ISMAC-CVB)*, Lecture Notes in Computational Vision and Biomechanics 30, pg. 1123-1134.

68] Z. Tang, P.A. Fishwick, 2013, 'Feed forward Neural nets as models for time series forecasting' Computer Science, University of Florida.

69] P. N. Vinayachandran, P. A. Francis and S. A. Rao, 2015, 'Indian Ocean Dipole: Processes and Impacts', Current Trends in Science, N. Mukunda (ed), Indian Academy of Sciences, Bangalore.

70] R. Adhikari, R.K. Agrawal, 2013, 'An Introductory Study on Time Series Modeling and Forecasting', LAP Lambert Academic Publishing.

71] A. Sanghani, N. Bhatt, N.C. Chauhan, 2016, 'A Review of Soft Computing Techniques for Time Series Forecasting', *Indian Journal of Science and Technology*, Vol 9 (S1), pg 1-6.

72] https://www.nhc.noaa.gov/sst/-used for SST Reconstruction, last assessed on 28/12/2019

73] G.E.P. Box, G.M. Jenkins, 1976, 'Time Series Analysis: Forecasting and Control, San Francisco, Holden Day, 1976.

74] Box and Jenkins: Time Series Analysis, Forecasting and Control T. C. Mills, A Very British Affair © Terence C. Mills 2013

75] R. J. Hyndman, *Time Series Data Library*, https://datamarket.com/data/list/?q=provider:tsdl last accessed on 26/09/2019.

76] https://climatedataguide.ucar.edu/climate-data/sst-data-hadisst-v11-about HadISST dataset; last accessed on 26/09/2019

77] http://www.pmel.noaa.gov/-data related to the 4th Dataset, last accessed on 26/09/2019

78] https://scripps.ucsd.edu/programs/shorestations/shore-stations-data/data-sio/-real time details of the data about Dataset 5, its collection details also, last accessed on 26/09/2019

79] https://archive.ics.uci.edu/ml/index.php-UCI machine learning database; last accessed on 30/08/19.

80] https://datamarket.com-Time series data library; last accessed 01/03/2019

81] https://people.duke.edu-to study ACF and PACF functions and how to evaluate the best possible model for an existing time series; last accessed on 25/11/2019

82] J.E. Nash, J.V. Sutcliffe, 1970, 'River flow forecasting through conceptual models part I —A discussion of principles', Journal of Hydrology. Vol.10 (3), pg 282–290.

83] Y.Li, H.Cao, 2018, 'Prediction for Tourism Flow based on LSTM Neural Network', Procedia Computer Science, Vol. 129 pg. 277–283.

84] http://www.remss.com/measurements/sea-surface-temperature/-remotesensing measurement of SST data, last accessed on 03/03/2021.

85] W.S. McCulloch, W.H. Pitts, 1943, 'A logical calculus of the ideas immanent in nervous activity', Bulletin of Mathematical Biophysics, Vol.5, pg 115-133.

86] C. A. Stock, K. Pegion, G. A. Veechi, M. A. Alexander, D. Tommasi, N. A. Bond, P. S. Fratantoni, R. G. Gudgel, T. Kristiansen, T. D. O'Brien, Y. Xue, X. Yang, 2015, 'Seasonal Sea Surface Temperature Anomaly prediction for Coastal Ecosystems', *Progress in Oceanography*, 137, pg 219-236.

87] A. Tealab, H. Hefny, A. Badr, (2017), 'Forecasting of nonlinear time series using ANN', *Future Computing and Informatics Journal*, Vol. 2, pg 39-47.

88] S.B. Mustapha, Pierre Larouche, Jean-Marie Dubois, 2016 'Spatial and temporal variability of sea-surface temperature fronts in the coastal Beaufort Sea', *Continental Shelf Research* 124, pg 134-141.

89] M. Picone, A. Orasi and G. Nardone, 2018 'Sea Surface Temperature monitoring in Italian Seas: Analysis of long term trends and short-term dynamics', *Measurement,* Vol.129, pg 260-267.

90] G. Saha, N.C. Chauhan, 2019, 'Dependency Investigation of Sea Surface Temperature on Sea Bottom Temperature and Sea Surface Salinity', IEEE International Conference on Innovations in Power and Advanced Computing Technologies, iPact2019, VIT University, Vellore, 22nd-23rd April.

91] G. Saha, N.C. Chauhan, 'Week ahead Time series Prediction of Sea Surface Temperature using Non Linear Auto Regressive Network with and without Exogenous Inputs', Book chapter for book titled "Elements of Statistical Learning" in Springer Nature (Singapore) Book Series "Algorithms for Intelligent Systems (AIS)". Editors: Dr. Prashant Johri, Galgotias University, Dr. J.K. Verma, Galgotias University, Dr. Sudip Paul, NEHU, Shillong.

92] https://www.kaggle.com/uciml/el-nino-dataset - for downloading the Elnino dataset; last assessed on 28th February, 2021.

93] S. Hochreiter, J. Schmidhuber, 1997, 'Long Short-Term Memory' Neural Computation 9(8), pg: 1735-1780.

94] mathworks.com-for understanding the basics of internal functions and default parameter list against function called; last assessed on 28th February, 2021.

95] S.Smyl, 2020, ‘A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting’, *International Journal of Forecasting*, Vol 36, pg 75-85.

96] D.S. de O. Santos Júnior, J.F.L. de Oliveira, P.S.G. de Mattos Neto, 2019 ‘An intelligent hybridization of ARIMA with machine learning models for time series forecasting’, Knowledge-Based Systems, Vol 175, pg 72-86.

97] J.F.L. de Oliveira, L.D.S. Pacífico, P.S.G. de Mattos Neto et al., 2019, ‘A hybrid optimized error correction system for time series forecasting’, Applied Soft Computing Journal.

98] P.S.G. de Mattos Neto, G.D.C. Cavalcanti, F. Madeiro, 2017, ‘Nonlinear Combination Method of Forecasters applied to PM Time Series’, *Pattern Recognition Letter.*

LIST OF PUBLICATIONS

Journal/Book Chapter:

1] G. Saha, N. C Chauhan "Week ahead Time series Prediction of Sea Surface Temperature using Non Linear Auto Regressive Network with and without Exogenous Inputs", Book chapter for book titled "Applications of Machine Learning" in Springer Nature (Singapore) Book Series "Algorithms for Intelligent Systems (AIS)".

Conference:

1] G. Saha, N.C, Chauhan, "Numerical Weather Prediction using Nonlinear Auto Regressive Network for the Manaus Region, Brazil", in *IEEE International Conference on Innovations in Power and Advanced Computing Technologies*, iPact2017, VIT University, Vellore, 21st-22nd April, 2017.

2] G.Saha, N.C. Chauhan, "Dependency Investigation of Sea Surface Temperature on Sea Bottom Temperature and Sea Surface Salinity", in *IEEE International Conference on Innovations in Power and Advanced Computing Technologies*, iPact2019, VIT University, Vellore,22nd-23rdMarch,2019.

APPENDIX A

GLOSSARY OF TERMS

El Nino-El Niño is a climate cycle in the Pacific Ocean with a global impact on weather patterns. The cycle begins when warm water in the western tropical Pacific Ocean shifting Eastwards along the Equator toward the coast of South America. Normally, this warm water pools near Indonesia and the Philippines.

Climate Analysis- Study of historical, current and forecasted climate conditions across the globe covering all aspects including sea, ice, land and forest cover.

Extended Reconstruction SST - The Extended Reconstructed Sea Surface Temperature (ERSST) dataset is a global monthly sea surface temperature dataset derived from the International Comprehensive Ocean–Atmosphere Dataset (ICOADS). Production of the ERSST is on a 2° × 2° grid with spatial completeness enhanced using statistical methods. This monthly analysis begins in January 1854 continuing to the present and includes anomalies computed with respect to a 1971–2000 monthly climatology. The newest version of ERSST, version 5, uses new data sets from ICOADS Release 3.0 (Sea Surface Temperatures) SST; SST comes from Argo floats above 5 meters, Hadley Centre Ice-SST version 2 (HadISST2) ice concentrations.

Irregularity - The irregular component of a time series is the residual time series after the trend-cycle and the seasonal components (including calendar effects) have been removed. It corresponds to the high frequency fluctuations of the series.

Indian Ocean Dipole - The Indian Ocean Dipole, also known as the Indian Niño, is an irregular oscillation of sea-surface temperatures in which the western Indian Ocean becomes alternately warmer and then colder than the eastern part of the ocean

La Nina- La Niña is a coupled ocean-atmosphere phenomenon that is the colder counterpart of El Niño, as part of the broader El Niño–Southern Oscillation climate pattern. The name La Niña originates from Spanish, meaning "the little girl", analogous to El Niño meaning "the little boy"

Level- A given time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise. Level is the average value in the series.

Seasonality - A given time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise. Seasonality is the influence of seasons on an annual data

Trend - A given time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise. Trend is the overall increment or decrement of the sequence.

Single Step Prediction – A prediction in the value of the variables for the next consecutive time stamp only is known as the Single Step Prediction.

Multistep Prediction – A prediction in the value of the variables for the next consecutive set of time stamps is known as the Multistep Prediction

Box and Jenkins method – It is a statistics based model approach proposed by G. E. P. Box and Gwilym Jenkins that uses auto regressive moving average (ARMA) or with differencing (ARIMA) models using past values of a time series to find its best fit.

Auto Regression (AR) –For a stationary time series, an auto regression models sees the value of a variable at time ‘t’ as a linear function of values ‘p’ time steps preceding it. Mathematically it can be written as −. y t = C + ϕ 1 y t − 1 + ϕ 2 Y t − 2 +... + ϕ p y t − p + ϵ t. Where, ‘p’ is the auto-regressive trend parameter.

Moving Average (MA) – In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set.

Insight based approach –Insight is the understanding of a specific cause and effect within a particular context. It generally involves the study of the inherent characteristics of a sequence.

Outsight based approach – Outsight is the understanding of supporting causes and effects beyond a particular context. It generally involves the study of the external factors that affects the characteristics of a sequence.

Foresight based approach – Foresight includes understanding the relevant recent past; scanning to collect insight about present, visualizing the future including trend research; environment research to explore possible trend breaks from developments on the fringe and other divergencies that may lead to alternative futures.