WATER LEVEL PREDICTION AND FORECASTING

A Project Report

Submitted by:

IKBIR SINGH (1800855)

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

BABA BANDA SINGH BAHADUR ENGINEERING COLLEGE

FATEHGARH SAHIB, PUNJAB (INDIA) - 140406

(AFFILIATED TO I.K.G. PUNJAB TECHNICAL UNIVERSITY, KAPURTHALA, PUNJAB (INDIA) DECEMBER 2021

CANDIDATE'S DECLARATION

I hereby certify that the project entitled "**WATER LEVEL PREDICTION AND FORECASTING** " submitted by **IKBIR SINGH (1800855)** in partial fulfillment of the requirement for the award of degree of the B.Tech. (Computer Science & Engineering) submitted in I.K. Gujral Punjab Technical University, Kapurthala at Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib is an authentic record of my own work carried out during a period from August, 2021 to December, 2021 under the guidance of Prof. **Balpreet Kaur** (ASSISTANT PROFESSOR, Department of Computer Science & Engineering). The matter presented in this project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

Signature of the Student

IKBIR SINGH

(University Roll No: 1800855, College Roll No: 181089)

Project In-charge

Prof. BALPREET KAUR

BABA BANDA SINGH BAHADUR ENGINEERING COLLEGE

Approved by AICTE, Govt. Of Punjab, Affiliated to IKGPTU

(Courses Accredited by NBA (AICTE))

Ref. No. ..... **Date 25 November 2021**

CERTIFICATE

This is to certify that the project titled "**WATER LEVEL PREDICTION AND FORECASTING**" is the bonafide work carried out **IKBIR SINGH (1800855)** in partial fulfillment of the requirement for the award of degree of the B. Tech. (Computer Science & Engineering) submitted in I.K. Gujral Punjab Technical University, Kapurthala at Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib is an authentic record of work carried out during a period from August, 2021 to December, 2021 under the guidance of Prof. **Balpreet Kaur** (ASSISTANT PROFESSOR, Department of Computer Science & Engineering)The Major Project Viva-Voce Examination has been held on (01/12/2021)

Signature of the Guide

Signature of the HoD,
Department of CSE.

Signature of the Principal
BBSBEC, Fatehgarh Sahib

CHANDIGARH ROAD, FATEHGARH SAHIB – 140407 PUNJAB (INDIA)
Ph. : 01763 503056, 503143, 503141 Fax: 01763 503139

ABSTRACT

According to World Resources Report (WRI), cities in the Global South are facing unreliable, inadequate, and polluted supply of freshwater.[1] Bengaluru is one such city included in this list, which is on the verge of imminent water crisis. Bengaluru sources water from the five Cauvery reservoirs, so it is important to monitor and analyze the decreasing water levels in these reservoirs. Since Machine Learning [11] algorithms are in hype in predicting and forecasting future probabilities based on past data by using statistical method as its basis these techniques can be applied to train the machine on weather reports and dam flow control and capacity data so as to provide efficient control over reservoir water level management and create forecasts to get better insights of the future. Also, we use the ensembling/ bagging technique on the results of the machine learning algorithms to improve the accuracy of the model.

Marks to be filled by Guide	**Marks Obtained**
Regularity (4)	
Self-Motivation and Determination (4)	
Working within Team (4)	
Total (12)	
Signature of the Guide	

ACKNOWLEDGEMENT

I express my sincere gratitude to the I.K. Gujral Punjab Technical University, Kapurthala for giving me the opportunity to work on the Major Project during my final year of B.Tech. (CSE) is an important aspect in the field of engineering.

I would like to thank **Dr. Lakhvir Singh, Principal** and **Dr. Kanwalvir Singh Dhindsa, Head of Department**, CSE at Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib for their kind support.

I also owe my sincerest gratitude towards Prof. **Balpreet Kaur** for her valuable advice and healthy criticism throughout my project which helped our team immensely to complete my work successfully.

I would also like to thank everyone who has knowingly and unknowingly helped me throughout the project. Last but not least, a word of thanks for the authors of all those books and papers which I have consulted during my project work as well as for preparing the report.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Water is an essential element for human survival and plays a major role in producing food supplies and in daily activities. But if uncontrolled or if not handled properly it can turn fatal to human life. Hundreds have died, and thousands have lost their houses due to recent monsoon shifts and climatic changes. High rainfall in unexpected seasons have caused disastrous incidents in many regions of India.

Water level plays an important part in the community's well-being and economic livelihoods.

For example, water level changes can impact physical processes in lakes, such as circulation, resulting in changes in water mixing and bottom sediment resuspension, and thus could further affect water quality and aquatic ecosystems. Hence, water level prediction and forecasting attracts more and more attention. For example, the International Joint Commission (IJC) suggests more efforts should be implemented to improve the methods of monitoring and predicting water level [21].

Reservoirs are large natural or artificial lake-like structures used as a source of water supply. It can be built in many ways but is usually built by building a wall-like structure across a watercourse that drains the existing body of water, interrupting the watercourse to form an embayment within it. A dam constructed in a valley relies on the natural topography to provide most of the basin of the reservoir. Dams are typically located at a narrow part of a valley downstream of a natural basin. The valley sides act as natural walls, with the dam located at the narrowest practical point to provide strength and the lowest cost of construction.

Cities in the Global South face unreliable, inadequate, and polluted supply of freshwater. About 1 billion people do not have access to safe and continuous (24/7) water supply. Rapidly growing urban populations and increased competition for water across sectors, coupled with climate change, pose increasing risks to water supplies. [1]

There is an urgent need for transparent data to inform water supply risk management

policymaking, especially during periods of water stress and increasing water insecurity in India. The data should include access to near real-time water risk information as well as short-term forecasting (1-3 months in advance) of reservoir water availability.

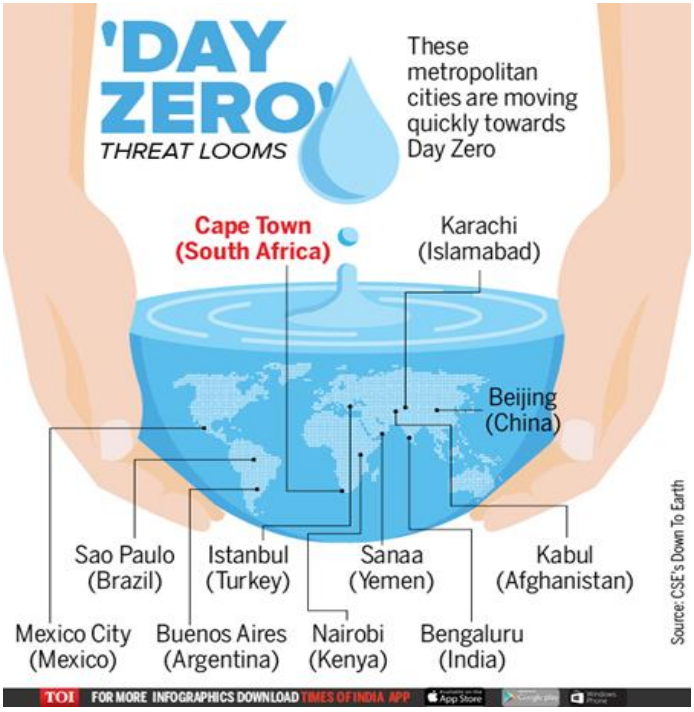

Figure 1. 'Day zero' Water threat in the cities of Global South

Most of the research tries to predict rainfall anomalies and flood prediction as in Supriya et al. [20] using various regression models. Although some of the models do predicts yearly and monthly reservoir Inflow such as Somchit et al. [13] which predicts monthly inflow estimate for a reservoir in Thailand using wavelet artificial neural networks and Tiantian et al. [14] uses ANN, random forest, and Support vector machine(SVMs) to predict the monthly Inflow, very few have attempted daily rainfall predictions. One of the reasons being that on a daily basis it is extremely hard to understand rain patterns and correlate it to reservoir inflow for the model. Also, Inflow or in the broad aspect, the river water flow depends on various aspects like precipitation, atmospheric pressures, temperatures, regional aspects of river basins, and also population using those waters. Even then finding the source of the river's regional climate requires a lot of groundwork and analysis. This experiment is an attempt to predict and forecast the reservoir water levels using only

past reservoir data and rainfall patterns in the respective catchment areas.

During the last decade, the Artificial Neural Network (ANNs)model has become popular in hydrological modeling and forecasting. Humans are able to do complex tasks like perception, pattern recognition, or reasoning much more efficiently and also are able to learn from examples and neural systems of the human brain are to some extent fault tolerant. Research and development in Artificial Neural Networks (ANNs) started with an attempt to model the bio-physiology of the human brain, creating models which would be capable of mimicking processes characteristics of humans on a computational level. Neural networks are widely regarded as a potentially effective approach for handling large amounts of dynamic, non-linear and noisy data, especially in situations where the underlying physical relationships are not fully understood.

1.1. Problem Statement

The overuse and wastage of water has led to the decline in the water levels of reservoirs sourcing water to the Bengaluru city. Uneven rainfall and climate shifts have adverse effects in reservoir outflow management and have caused disasters. An attempt is made to use deep learning techniques and algorithms to forecast and predict water levels at the Krishnaraj Sagar, Kabini, Harangi and Hemavathi reservoirs. Karnataka can hence work towards managing the water outsourcing from these reservoirs.

1.2. Objective of the Project

To build a time series analysis model using Deep Learning Techniques like Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN), Facebook Prophet etc. with near real time reservoir level data for better temporal water level forecasting.

1.3. Hardware Specification

- Processor: 1.8 GHz or faster processor. Dual-core(minimum)
- RAM: 4 GB (for development) 2GB or above
- Hard Disk: 20-25 GB of free space (for development)
- Display: 1024x768 minimum screen resolution.

1.4. Software Specification

- Operating System: Windows 10/8.1 or Linux or Linux server
- Coding Language: Python
- Tools and frameworks
    - ❖ Jupyter Notebook
    - ❖ PyCharm
    - ❖ Google Colab
    - ❖ ScikitLearn
    - ❖ GeoJSON
    - ❖ TensorFlow
    - ❖ Plotly Dash

2. LITERATURE SURVEY

Water-level change is a complex hydrological phenomenon due to its various controlled factors, including meteorological conditions, as well as water exchange between the lake and its watersheds. Thus, many tools used to forecast water levels, while considering influencing factors, have been developed, such as process-based models. For example, Gronewold et al. showed that the advanced hydrologic prediction system (AHPS) can be used to capture seasonal and inter-annual patterns of Lake Erie water level from 1997 to 2009 [2]. However, the effectiveness of process-based models mainly depends on the accuracy of the models to represent the aquatic conditions and the abilities of the models in describing the variabilities in the observations. In addition, process-based models are often time-consuming, so numerous studies have proposed using statistical models to predict water level, e.g., autoregressive integrated moving average model, artificial neural network, genetic programming, and support vector machine. These studies mainly focused on leveraging historical water level without considering the physical process driven by meteorological conditions.

2.1. Existing System

- Basically dam/reservoir gates are opened to release water when dam water level is at its full capacity. In other scenarios, opening release valves will be different based on the type of the dam.

- Only a limited number of dams exist solely for flood control which are known as attenuation or balancing reservoirs. The rest are solely meant for the purpose of storing water.

- Water inflow is acknowledged by calculating water level rise within a time interval, precalculated reservoir area and elevation values. In other cases, monthly predictions are made based on the hydrological model of the reservoir.

2.2. Drawbacks

- No prediction methods are used to predict water inflow on a daily basis which is critical in current climatic situations.
- Varying rainfall and climatic changes can result in inflow which might exceed the water level capacities but releasing large amounts of water in a flash can result in floods.
- Difficult to vacate flood prone areas in the event of aforementioned scenarios.

2.3. Proposed System

- Water inflow can be predicted well in advance using various Machine learning methodologies.
- Various machine learning models can be trained to propose a near accurate model based on rainfall and other related data available on WRI website.
- The data will be used to train the model to predict water inflow, water level that can be expected in the future.
- The system will suggest the amount of water to be released and at what increments in case of flood alerts so as to minimize the loss.

2.4. Feasibility

All systems are feasible when provided with unlimited resources and infinite time. But unfortunately, this condition does not prevail in the practical world. So it is both necessary and prudent to evaluate the feasibility of the system at the earliest possible time.

- **Economic Feasibility:** This study is carried out to check the economic impact that the system will have on the organization. Since the project is Machine learning based, the cost spent in effectuating this project would not demand cost for softwares and related products,

as most of the products are open source and free to use. Hence the project would consume minimal cost and is economically feasible.

- **Technical Feasibility**: This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Since machine learning algorithms are based on pure math there is very little requirement for any professional software. And also, most of the tools are open source. The best part is that we can run this software in any system without any software requirements which makes them highly portable. Also, most of the documentation and tutorials make it easy to learn the technology.

- **Social Feasibility**: The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The main purpose of this project which is based on the dam inflow is to prevent flooding by predicting the inflow and thereby fluctuating the outflow according to the needs necessary at that moment. Thus, this is a noble cause for the sake of the society, a small step taken to achieve a secure future.

3. THEORETICAL OVERVIEW

3.1. Overview of Machine Learning:

Machine learning is an application of artificial intelligence (AI) that gives systems the ability to automatically learn and evolve from experience without being specially programmed by the programmer[10]. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The main aim of machine learning is to allow computers to learn automatically and adjust their actions to improve the accuracy and usefulness of the program, without any human intervention or assistance. Traditional writing of programs for a computer can be defined as automating the procedures to be performed on input data in order to create output artifacts. Almost always, they are linear, procedural and logical. A traditional program is written in a programming language to some specification, and it has properties like:

- We know or can control the inputs to the program.
- We can specify how the program will achieve its goal.
- We can map out what decisions the program will make and under what conditions it makes them.
- Since we know the inputs as well as the expected outputs, we can be confident that the program will achieve its goal

Traditional programming works on the premise that, as long as we can define what a program needs to do, we are confident we can define how a program can achieve that goal. This is not always the case as sometimes, however, there are problems that you can represent in a computer that you cannot write a traditional program to solve.

Such problems resist a procedural and logical solution[10]. They have properties such as:

- The scope of all possible inputs is not known beforehand.
- You cannot specify how to achieve the goal of the program, only what that goal is.
- You cannot map out all the decisions the program will need to make to achieve its goal.

- You can collect only sample input data but not all possible input data for the program.

3.1.1. Supervised and Unsupervised Learning

Supervised learning: “Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values.”,[12].

Unsupervised learning: “Unsupervised learning is a type of machine learning algorithm used to draw inferences from data sets consisting of input data without labeled responses.”, [11]. In data mining and machine learning an abundance of models and algorithms can be found, but most fundamentally these are divided into supervised and unsupervised learning. One fundamental example has been mentioned in the foregoing section, the clustering of iris-species. Former is a supervised process where data points are labeled (“species A”, “species B” or “species C”) and labels are calculated for new data points. Comparing calculated labels according to the trained model with the original label gives the model’s accuracy, hence supervised.

Unsupervised learning on the other hand does not require any labeling since the algorithm is searching for a pattern in the data. This might be useful when categorizing customers into different groups without a priori knowledge of which groups they belong to.

3.1.2. Types of Machine Learning Algorithm

For machine learning many different algorithms can be found. For simplicity these can be subdivided into four categories, where each category is good for different kinds of problems. Anomaly detection algorithms are good for finding unusual data points. Trained classification algorithms can be used to categorize unseen data. As an example, it could be used to take in data from a phone on movement to categorize what activity is being performed. Clustering algorithms group data into clusters and look for the greatest similarities. This can be used to find unknown connections on huge sets of data. Regression algorithms are used to find patterns and build models to predict numerical values from datasets. These will take multiple inputs and determine how much

each input affects the output.

Within the regression category there are many different basic algorithms. Linear Regression is the most classic type which solves linear relationships between inputs and outputs. Neural network regression is most common in deep learning and adaptable to regression problems, but might be too complex for simple regression problems and requires thorough training.

Machine learning algorithms are tools to automatically make decisions from data in order to achieve some overarching goal or requirement. The promise of machine learning is that it can solve complex problems automatically, faster and more accurately than a manually specified solution, and at a larger scale. Over the past few decades, many machine learning algorithms have been developed by researchers, and new ones continue to emerge and old ones modified. In this project, we have focused on only supervised learning methods since our dataset contains labels.

3.2. Machine Learning Tools

Tools are a big part of machine learning and choosing the right tool can be as important as working with the best algorithms. Machine learning tools are not just implementations of machine learning algorithms. They can be, but they can also provide capabilities that you can use at any step in the process of working through a machine learning problem.

There are many different software tools available to build machine learning models and to apply these models to new, unseen data. These tools typically contain libraries implementing some of the most popular machine learning algorithms. Following are

the major tools used in the project-

- SciKitLearn
- Plotly Dash

3.2.1. SciKit learn

SciKit learn is an open source machine learning library built for python. Since its release in 2007,

Scikit-learn[24] has become one of the most popular open source machine learning libraries. Scikit-learn (also called sklearn) provides algorithms for many machine learning tasks including classification, regression, dimensionality reduction and clustering. It also provides utilities for extracting features, processing data and evaluating models. It provides in-built code for many of the popular machine learning algorithms. The documentation for scikit-learn is comprehensive, popular and well maintained. Sklearn is built on mature Python Libraries such as NumPy, SciPy, and matplotlib. It has a very active development community with regular update releases of the library.

3.2.2. Plotly Dash

Dash is a python framework created by plotly for creating interactive web applications. Dash is written on the top of Flask, Plotly.js and React.js. With Dash, you don't have to learn HTML, CSS and Javascript in order to create interactive dashboards, you only need python. Dash is open source and the applications built using this framework are viewed on the web browser. [10]
Dash applications are made up of 2 building blocks :

1. Layout
2. Callbacks

Layout describes the look and feel of the app, it defines the elements such as graphs, dropdowns etc and the placement, size, color etc of these elements. Dash contains Dash HTML components using which we can create and style HTML content such as headings, paragraphs, images etc using python. Elements such as graphs, dropdowns, and sliders are created using Dash Core components. Callbacks are used to bring interactivity to the dash applications. These are the functions using which, for example, we can define the activity that would happen on clicking a button or a dropdown.

3.3. Artificial Neural Network

Artificial neural networks (ANN), a type of biologically inspired computational models, is being successfully used in the area of hydrology and water resources. ANN model is a data driven model as it performs an input-output mapping via a series of simple processing nodes or neurons and in hydrology it is considered as black box model. For the first time McCullochand Pitts presented a basic artificial neural network model. With the development of computational techniques, various

researchers have suggested different ANN structures to model various real-life problems. The task of each individual neuron is in ANN model structure twofold (i) it integrates information from an external source or from other neurons, often via a linear function, and (ii) it outputs this value via a transfer function, such as the sigmoid. The ability to map a function is derived via the configuration of these neurons into a set of weighted, interconnected layers. Between the two external layers are one or more interconnected, hidden layers, which is the key to learning the relationships in the data. Hydrological processes are generally nonlinear in nature and the ability of ANN in modeling nonlinear processes [4] advocates their use in hydrology and water resources. There are many ANN architectures and algorithms developed. Out of them most common are multilayer feed forward, Hopfield networks, Radial basis function network, recurrent network, Self-organization feature maps, counter propagation networks. [22] Coppola et al., showed that ANN has potential in predicting groundwater level fluctuations in an unsteady state of an aquifer influenced by pump and different weather conditions. They noted that predicted results of ANN are more accurate than quantitative models and also showed that ANN models are good at simulating karstic and leaky aquifers where other numerical models are weak in such cases. Taiyuan et al., simulated the effects of hydrological, weather and humidity conditions on groundwater level by ANNs in the lower part of Shenyang river basin, North West of china [23]. The ANN model developed by them was able to predict groundwater levels with the average error of 0.37 m or lower with high accuracy.

3.4. Convolutional Neural Network

Convolutional Neural Network (CNN) is a Deep Learning algorithm which takes in an input image and assigns importance (learnable weights and biases[15]) to various aspects/objects in the image, which helps it differentiate one image from the other.

One of the most popular applications of this architecture is image classification. The neural network consists of several convolutional layers mixed with nonlinear and pooling layers. When the image is passed through one convolution layer, the output of the first layer becomes the input for the second layer. This process continues for all subsequent layers. [6][8]

After a series of convolutional, nonlinear and pooling layers, it is necessary to attach a fully connected layer. This layer takes the output information from convolutional networks. Attaching

a fully connected layer to the end of the network results in an N dimensional vector, where N is the number of classes from which the model selects the desired class.

3.5. LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) capable of learning order dependence in sequence prediction problems. This is most commonly used in complex problems like Machine Translation, Speech Recognition, and many more.

The reason behind developing LSTM was, when we go deeper into a neural network if the gradients are very small or zero, then little to no training can take place, leading to poor predictive performance and this problem was encountered when training traditional RNNs. LSTM networks are well-suited for classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series. [5]

LSTM is way more effective and better compared to the traditional RNN as it overcomes the short-term memory limitations of the RNN. LSTM can carry out relevant information throughout the processing of inputs and discards non-relevant information with a forget gate.

3.6. CNN-LSTM Architecture

The CNN-LSTM architecture involves using CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction**.** This model is specifically designed for sequence prediction problems with spatial inputs, like images or videos. They are widely used in Activity Recognition, Image Description, Video Description and many more. [8]

The general architecture of the CNN-LSTM Model is as follows:

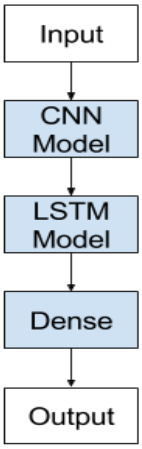

Figure 2 : CNN LSTM Combined architecture

CNN-LSTMs are generally used when their inputs have spatial structure, such as the 2D structure or pixels in an image or the 1D structure of words in a sentence, paragraph, or document and also have a temporal structure in their input such as the order of images in a video or words in text or require the generation of output with temporal structure such as words in a textual description.

4. SYSTEM ANALYSIS AND DESIGN

4.1. System Requirements Specification

A system requirements specification (SRS) is a description of a software system to be developed. It lays out functional and nonfunctional requirements and may include a set of use cases that describe user interactions that the software must provide.

In order to fully understand one's project, it is very important that they come up with an SRS listing out their requirements, how are they going to meet it and how will they complete the project. SRS also functions as a blueprint for completing a project with as little cost growth as possible. SRS is often referred to as the parent document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it.

Requirement is a condition or capability to which the system must conform. Requirement Management is a systematic approach towards eliciting, organizing and documenting the requirements of the system clearly along with the applicable attributes. The elusive difficulties of requirements are not always obvious and can come from any number of sources.

4.1.1. Functional Requirements

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality. Following are the functional requirements on the system:

- Dynamic Dashboard to view the results of the predictions
- Visualizations of forecasted data
- Predicting the water level using other features such as Inflow, Outflow, Rainfall, Soil Moisture, Humidity etc.
- 1-3 months forecasts of water level
- Correlation of the features crucial of water level prediction
- Reservoir-wise forecasting
- Scalable design to add new reservoir forecasting

4.1.2. Non-Functional Requirements

Non functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviours. They may relate to emergent system properties such as reliability, response time and store occupancy. Non functional requirements arise through the user needs, because of budget constraints, organizational policies and the need for interoperability with other software and hardware systems.

4.1.2.1. Product Requirements:

- **Correctness:** Project must follow a well defined set of procedures and rules to engage a conversation with the user and a pre-trained machine learning model to compute, also rigorous testing is performed to confirm the correctness of the data.
- **Modularity:** The complete product is broken up into many modules and well defined interfaces are developed to explore the benefit of flexibility of the product.
- **Robustness:** This software is being developed in such a way that the overall performance is optimized and the user can expect the results within a limited time with utmost relevance and correctness.

Non functional requirements are also called the qualities of a system. These qualities can be divided into execution quality and evolution quality. Execution qualities are security and

usability of the system which are observed during run time, whereas evolution quality involves testability,maintainability, extensibility or scalability.

4.1.2.2. Basic Operational Requirements

The customers are those that perform the primary functions of system engineering, with special emphasis on the operator as the key customer. Operational requirements will define the basic need and, at a minimum, will be related to these following points:

- **Mission profile or scenario :** It describes the procedures used to accomplish mission objectives. It also finds out the effectiveness or efficiency of the system. In our case a web portal with all functional requirements satisfied.
- **Performance and related parameters :** It points out the critical system parameters to

accomplish the mission. The response time of the predictions, data visualization, effectiveness in how to convey the alerts and suggestions, along with a responsive user interface.

- **Utilization environments :** It gives a brief outline of system usage. Finds out appropriate environments for effective system operation. The required tools and frameworks like keras, facebook prophet, flask and also web frameworks like bootstrap, jquery and selenium(for web scraping).
- **Operational Life Cycle :** It defines the system lifetime. In our case the life cycle continues until the performance of the model is not degraded. In such cases the models must be retrained for better performance.

4.2. Analysis

4.2.1. Performance Analysis

Most of the software we use is open source and free. The models which we use in this software,learn only once ,i.e once they are trained they need not be again fed in for the training phase. One can directly predict values, hence time-complexity is very less. Therefore this model is temporally sound.

4.2.2. Technical Analysis

As mentioned earlier,the tools used in building this software are open source. Each tool contains simple methods and the required methods are overridden to tackle the problem. The presentation layer helps one to use the software with ease.

4.2.3. Economical Analysis

The completion of this project is not free of cost in its entirety. Even though the software used in building the model is free of cost, there are a few expenditures which are spent. Some of them include travelling to the WRI organization in Bangalore, the server for storing the data, hosting web applications etc.

4.3. System Development Methodology

System Development methodology is the development of a system or method for a unique situation. Having a proper methodology helps us in bridging the gap between the problem statement and turning it into a feasible solution. It is usually marked by converting the System Requirements Specifications (SRS) into a real world solution. System design takes the following inputs: Statement of work, Requirement determination plan, Current situation analysis. Proposed system requirements including a conceptual data model and metadata (data about data). The development method followed in this project is the waterfall model. Although the waterfall model is used as a project model, a spiral model is used for ML model selection and development implementation.

4.3.1. Model Phases

The waterfall model is a sequential software development process, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of Requirement initiation, Analysis, Design, Implementation, Testing and maintenance.

- **Requirement Analysis**: This phase is concerned about collection of requirements of the system. This process involves generating document and requirement review.
- **System Design**: Keeping the requirements in mind the system specifications are translated into a software representation. In this phase the designer emphasizes on:- algorithm, data structure, software architecture etc.
- **Coding**: In this phase the programmer starts his coding in order to give a full sketch of the product. In other words system specifications are only converted into machine readable compute code.
- **Implementation**: The implementation phase involves the actual coding or programming of the software. The output of this phase is typically the library, executables, user manuals and additional software documentation.

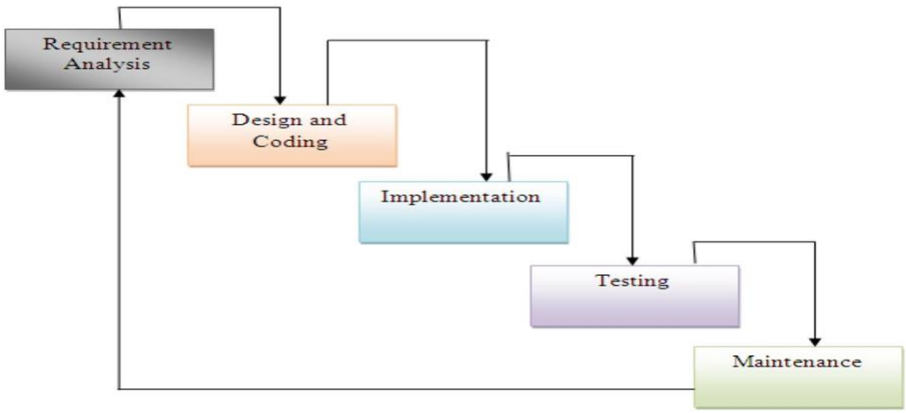

Figure 3: Waterfall Model

- **Testing**: In this phase all programs (models) are integrated and tested to ensure that the complete system meets the software requirements. The testing is concerned with verification and validation.
- **Maintenance**: The maintenance phase is the longest phase in which the software is updated to fulfill the changing customer needs, adapt to accommodate changes in the external environment, correct errors and oversights previously undetected in the testing phase, and enhance the efficiency of the software.

4.3.2. Advantages of Waterfall model

- Clear project objective
- Stable project requirements
- Progress of the system is measurable.
- Logic of software development is clearly understood.

4.4. Flow charts for Model Building Phase and Sequence Diagram

A Sequence diagram is an interaction diagram that shows how processes operate with one another and what is their order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations

in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios.

Sequence diagrams are an easy and intuitive way of describing the behavior of a system by viewing the interaction between the system and the environment. A sequence diagram shows an interaction arranged in a time sequence. A sequence diagram has two dimensions: vertical dimension represents time, the horizontal dimension repre-

sents the object's existence during the interaction.

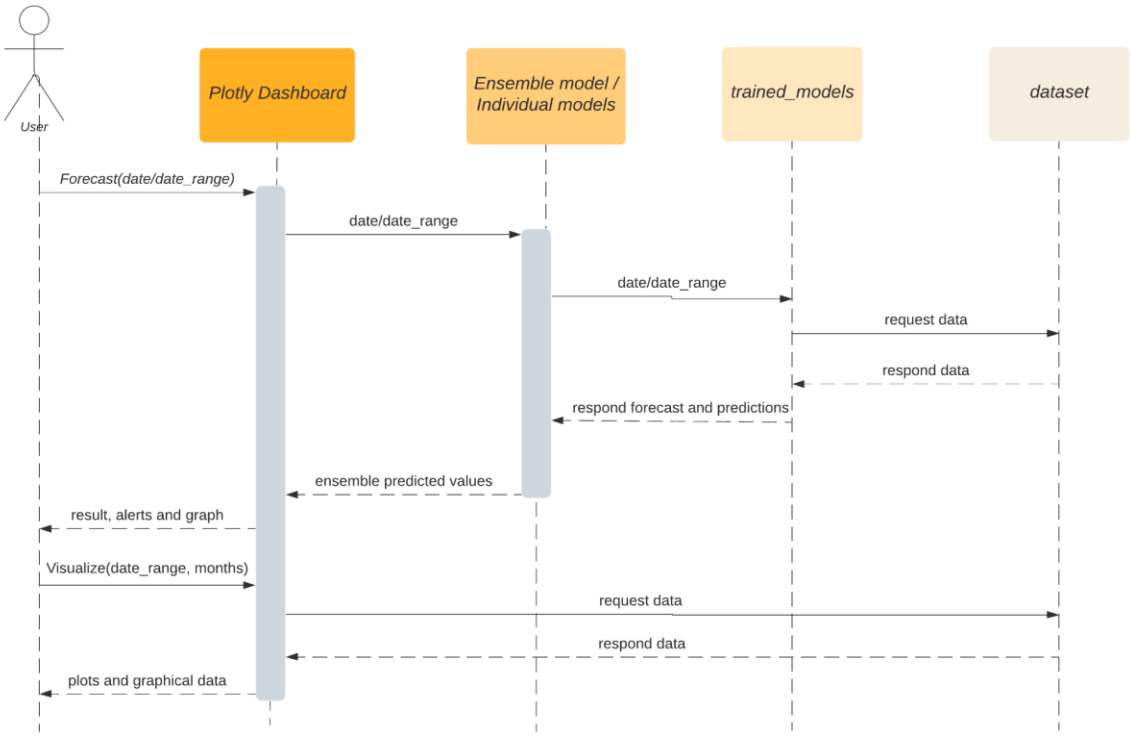

Figure 4 : Sequence Diagram

The sequence diagram depicted in the above figure 4 shows sequential processes for two different transactions where the user interacts with the plotly dashboard. One of them being prediction for a given date or for a date range. As seen in the gure, request is sent to an instance of ensemble model which invokes all the trained base models. These base models collect the dataset for respective dates and forecast the reservoir water level. Based on the predictions ensemble model instance predicts the reservoir parameters which is used by the dashboard to display results, graphs and alerts. The second transaction is data visualization for user specified duration(year and month). The dashboard directly requests the relevant data and generates plots (such as box plots) and

graphical representation based on the user's decision.

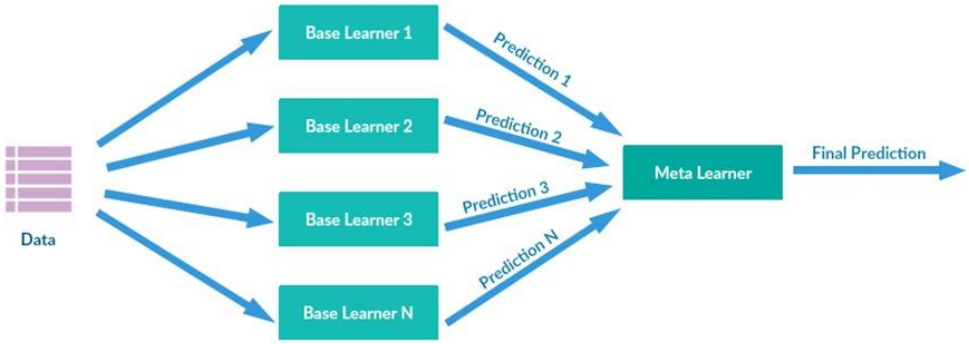

Figure 5 : Model Structure

4.5. Design and Test Steps

Design of a system is a conceptual model using which we can define the structure and behaviour of that system. It is a formal representation of a system. Depending on the context, design can be used to refer to either a model to describe the system or a method used to build the system [18]. Building a proper system design helps in analysis of the project, especially in the early stages.

- **Datasets**: The dataset is obtained from the official WRI website and India WRIS datasets [24][25]. It includes the geospatial data of four cauvery river basin reservoirs Hemavathi, Harangi, Kabini and KRS.

Dataset Source	Dataset Name	Date Range	Original Row Count	Missing Rows	Final Date Range	Final Row Count
WRI	Harangi	30-09-2010 to 16-12-2020	3321	332	01-01-2011 to 31-12-2020	3653
WRI	Hemavathi	30-09-2010 to 16-12-2020	3314	339	01-01-2011 to 31-12-2020	3653
WRI	KRS	30-09-2010 to 16-12-2020	3313	340	01-01-2011 to 31-12-2020	3653
WRI	Kabini	30-09-2010 to 16-12-2020	3314	339	01-01-2011 to 31-12-2020	3653

Figure 6: WRI Datasets Description

- **Database**: The dataset and files are of limited size, considering the fact that there is no

transactional overhead or necessity for complex storage architecture. Unix file system is used as a database or in other words a Storage system and all data is just stored as files.

- **Data Preprocessing and Vectorization:** Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent and is likely to contain many errors. Data preprocessing is a method used to resolve such issues. DataSet will be passed to further modules in the form of vectors for increasing computational efficiency.
- **ML algorithm selection module:** Multiple ML models are compared and reviewed to provide best results. Although it is run iteratively to get best results, once a model is trained it is not changed unless the dataset has grown at least 200% its size.
- **ML prediction module:** Daily data of weather and Dam parameters will be fed to the module for prediction of dam parameters and suggestions.
- **Suggestions and alert System:** The module collects the predicted information and generates suitable alerts and suggestions based on the received information.
- **Data Visualization module:** The module will contain visual comparison of dam parameters with various weather parameters over the past years. Visualization frameworks like MatplotLib will be used.

4.5.1. Implementation

Implementation of aforementioned system architecture is nothing but implementing all the modules, but in real scenario many modules overlap with each other. For example, the Suggestion and alert module was integrated with prediction yet holding its modularity. Following are the tasks performed to realize the project.

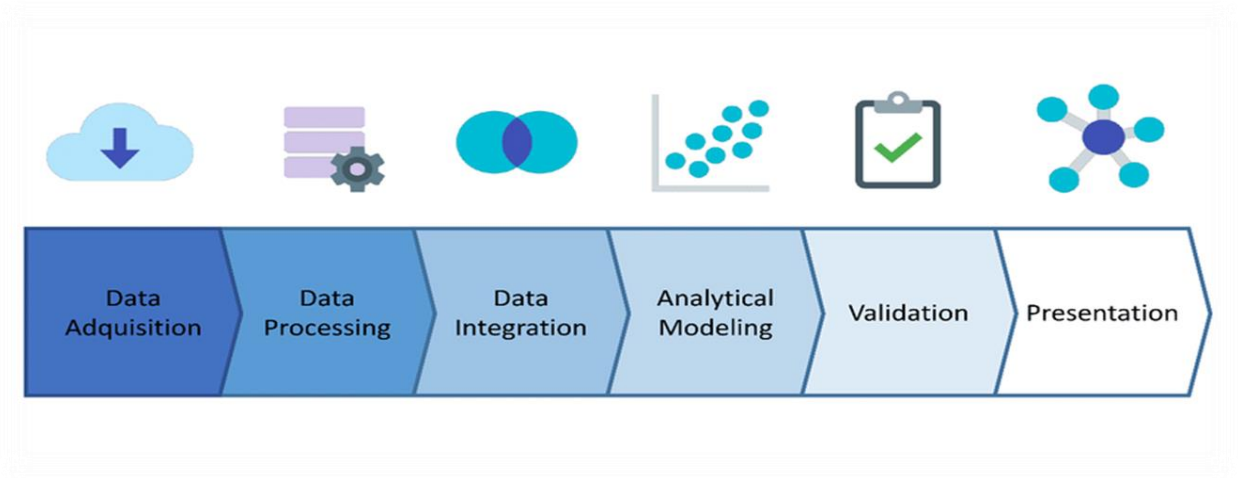

Figure 7: Data modelling pipeline

Data Collection and Preprocessing

- **Data Collection:**

  The data collection is collected in the form of datasets from WRI sites and India WRIS data sets directly. The reservoir details were collected by web scraping techniques using Selenium web driver. The program performs some automated tasks which involve launching a chrome web browser, reaching WRI website, applying required parameters and finally recording the required details for the respective reservoir.

- **Preprocessing:**

  Data preprocessing is an important step as it helps in cleaning the data and making it suitable for use in machine learning algorithms. Most of the focus in preprocessing is to remove any outliers or erroneous data, as well as handling any missing values. Missing data can be dealt with in two ways. The first method is to simply remove the entire row which contains the missing or erroneous value. While this an easy to execute method, it is better to use only on large datasets. Using this method on small datasets can reduce the dataset size too much, especially if there are a lot of missing values. This can severely affect the accuracy of the result. Since ours is a relatively small dataset, we will not be using this method. One more reason for not dropping rows is the fact that the project depends on time series data and hence removing rows will affect the results. Interpolation

methods are used in cases where it is relevant and in other cases the missing data is replaced with zeros.

Converting the dataset into time series data is the most important step in the project. Every data feature is converted into an event based where each event is every day. The rain dataset is received with dates as columns, which are converted to dated rows. At the end of this step, the dataset will be date indexed.

Dataset Source	Dataset Name	Date Range	Original Row Count	Missing Rows	Final Date Range	Final Row Count
IndiaWRIS	Harangi	01-01-2001 to 19-05-2021	7324	120	01-01-2001 to 19-05-2021	7444
IndiaWRIS	Hemavathi	01-01-2001 to 19-05-2021	7328	116	01-01-2001 to 19-05-2021	7444
IndiaWRIS	KRS	01-01-2001 to 19-05-2021	7324	120	01-01-2001 to 19-05-2021	7444
IndiaWRIS	Kabini	01-01-2001 to 19-05-2021	7328	116	01-01-2001 to 19-05-2021	7444

Figure 8: IndiaWRIS datasets description

- **Data Analysis:**

  One of the important steps in building Machine learning models is analysing the data. This helps in identifying relationships between the various attributes present in the dataset and also to gain more knowledge on data which is essential in building efficient models. We used graphs to visualize the relationships between the data attributes.

- **Ensembling the models:**

  Before going to ensembling it must be understood that the trained models are dumped as pickle files using the Pickle library provided by python as well as the h5 module(by keras) for LSTM.

  In statistics and machine learning, ensembling is the way in which we use multiple learning algorithms to induce higher second-sighted performance than is also obtained from any of the constituent learning algorithms alone. In our case choosing one best model among the aforementioned algorithms weighed less than ensembling all models so as to get the best

of all models. Even though it doesn't work in many cases, ensembling the models did improve the performance of the model. Gradient boosted regression is used as ensemble method.[19] The prediction result is given along with upper and lower bounds to provide a prediction interval with 95% confidence level instead of providing with single valued prediction.

4.6. Testing

The program comprises several algorithms which are tested individually for accuracy. we check for the correctness of the program as a whole and how it performs.

4.6.1. Unit Testing

Unit tests focus on ensuring that the correct changes to the world-state take place when a transaction is processed. The business logic in transaction processor functions should have unit tests, ideally with 100 percent code coverage. This will ensure that you do not have typos or logic errors in the business logic. The various modules can be individually run from a command line and tested for correctness. The tester can pass various values, to check the answer returned and verify it with the values given to him/her. The other work around is to write a script, and run all the tests using it and write the output to a log file and use that to verify the results. We tested each of the algorithms individually and made changes in pre-processing accordingly to increase the accuracy.

4.6.2. System Testing

System Testing is a level of software testing where a complete and integrated software is tested. The purpose of this test is to evaluate the system's compliance with the specified requirements. System Testing is the testing of a complete and fully integrated software product. Usually, software is only one element of a larger computer-based system. Ultimately, software is interfaced with other software/hardware systems. System Testing is actually a series of different tests whose sole purpose is to exercise the full computer-based system.Two Categories of Software Testing are Black Box Testing and White Box Testing. System test falls under the black box testing category of software testing.

Different Types of System Testing:

- **Usability Testing** - Usability Testing mainly focuses on the users ease to use the application, flexibility in handling controls and ability of the system to meet its objectives.
- **Load Testing** - Load Testing is necessary to know that a software solution will perform under real-life loads.
- **Regression Testing** - Regression Testing involves testing done to make sure none of the changes made over the course of the development process have caused new bugs. It also makes sure no old bugs appear from the addition of new software modules over time.
- **Recovery Testing** - Recovery testing is done to demonstrate a software solution is reliable, trustworthy and can successfully recoup from possible crashes.
- **Migration Testing** - Migration testing is done to ensure that the software can be moved from older system infrastructures to current system infrastructures without any issues.

4.6.3. Functional Testing

Functional Testing is also known as functional completeness testing, Functional Testing involves trying to think of any possible missing functions. Testers might make a list of additional functionalities that a product could have to improve during functional testing. In addition, business process flows like data fields, predefined processes and successive processes must be considered for testing.

5. RESULTS

Although it may seem illogical to predict inflow only with rainfall patterns as the independent factor, during the monsoon season of every year the inflow is evidently dependent mostly on rainfall even though there exist other affecting factors. To provide better results, the predicted values from ensembled regression are classified.

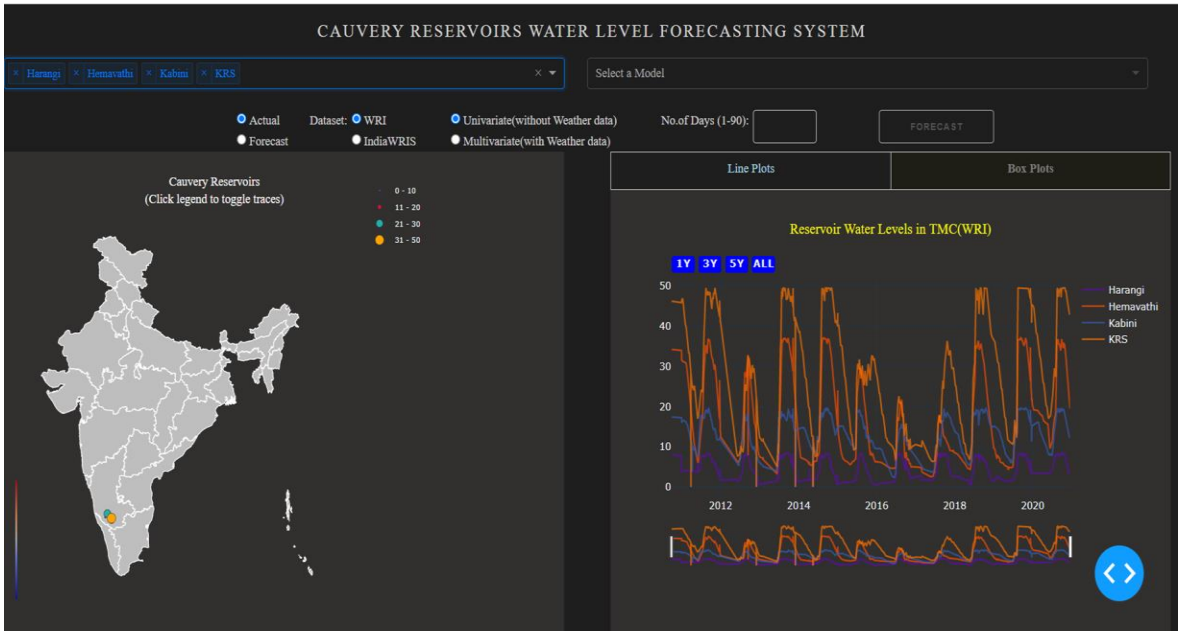

Figure 9: Dashboard interface

The dashboard consists of two modes

1. Actual
2. Forecast.

Actual mode enables selection of multiple reservoirs from the dropdown list. You can see line and box plots of the selected reservoirs with actual data. The data can come from the **WRI dataset or IndiaWRIS dataset.**

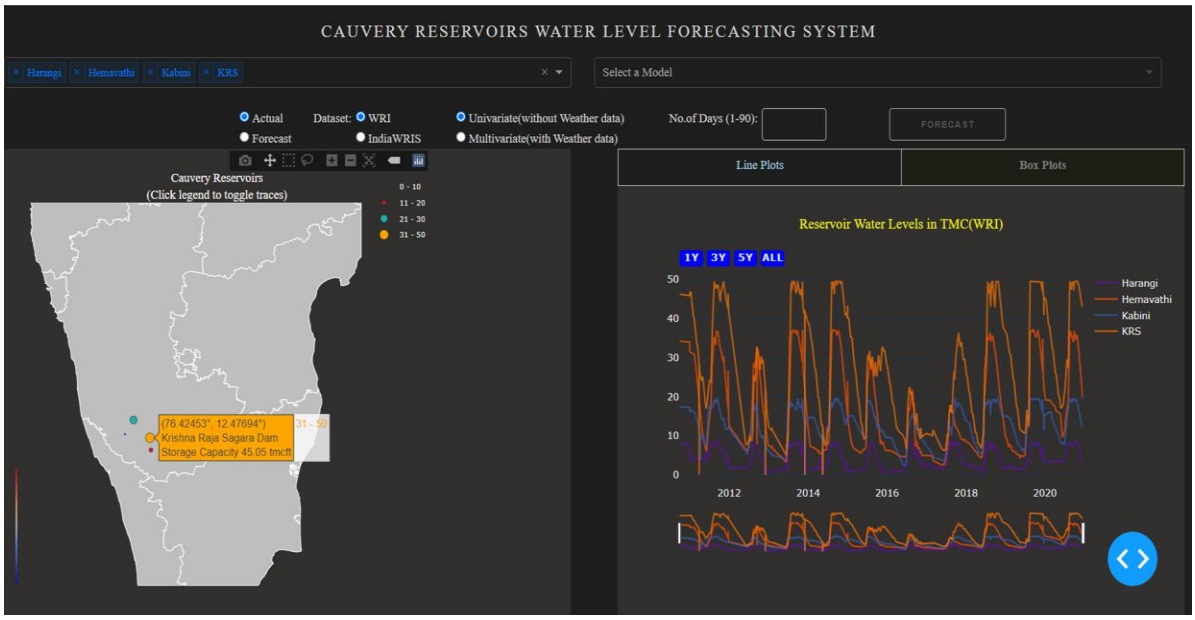

Figure 10: Dashboard features - Map

Dashboard Features - Map : You can zoom and see the details of each reservoir on the map. The map is generated from a geoJSON file.

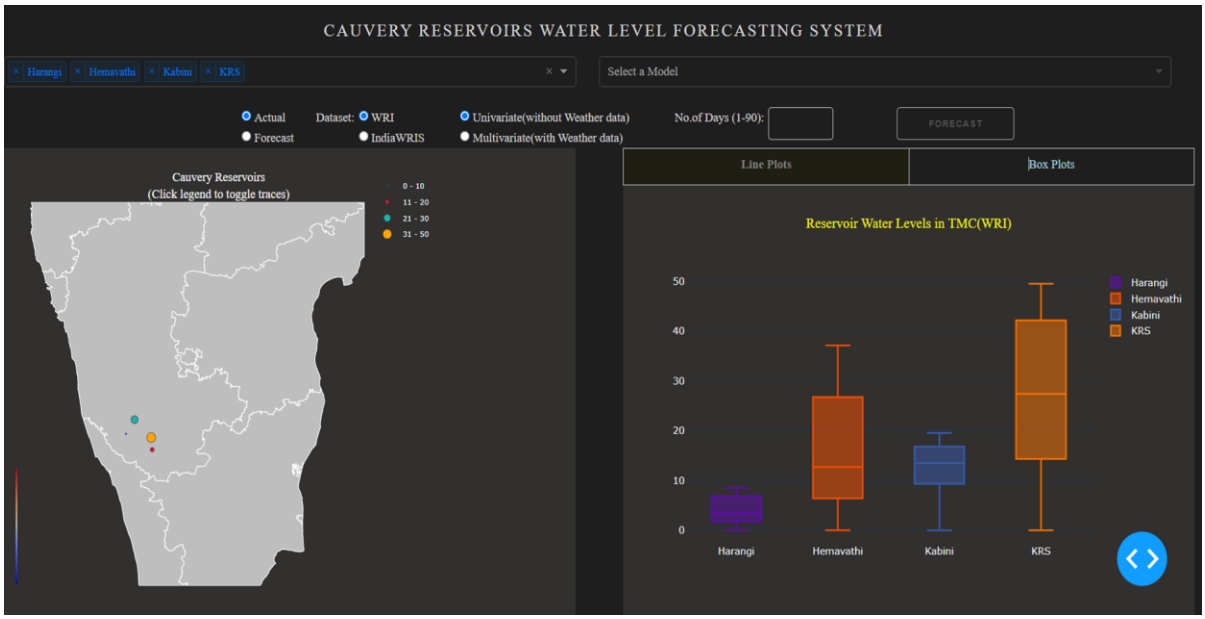

Figure 11: Dashboard features - Box plots

Box plots of target variables for 4 reservoirs are shown for both WRI and IndiaWRIS datasets.

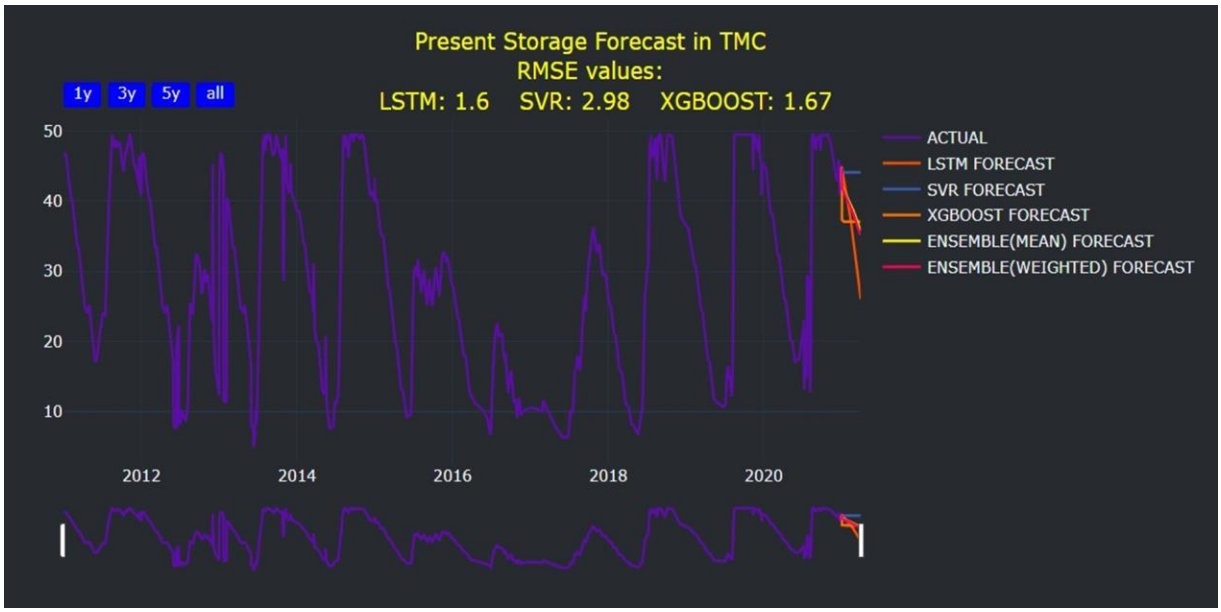

Figure 12: Forecasting plots

Present Storage Forecasting plots:

The models were trained till 2019 and tested in the year 2020. The Present Storage (TMC) values were forecasted for the year 2021.

6. CONCLUSION AND FUTURE SCOPE

Although it takes a lot of ground work and analysis to identify all the dependent factors, collect relevant information, convert, extract and pre-process the data, so as to develop model with higher accuracy and precision, most of them affects the water level of the reservoir only to a small extent making it hard to identify the factors that are actually crucial for prediction. This project used three univariate models namely Support Vector Machine (SVM), Long Short Term Memory(LSTM) and XGBoost with a limited number of dependent factors available to provide a dependable prediction model for forecasting the water level in the four reservoirs of Karnataka i.e. KRS, Kabini, Harangi, and Hemavathi. Despite the excellent performance of models, operations are not well understood. To get deeper insight into their internal working we followed and discussed the architectures and workings of the models used.

It was a great experience learning both deep learning and machine learning techniques as well as gaining the understanding on using the machine learning tools and frameworks. Nevertheless, Machine learning in general is an active area of research. Even though confusing, once the dimensionalities and vectorization concepts are well digested, developing the ML models could provide best results along with reasonable training time.

Although it may include complex computation as multiple reservoirs join and separate causing increased complexity, given enough time and resources the prediction models can be trained to give better results. Also various other deciding factors can be identified, collected and used for training to improve accuracy even though it takes a lot of ground work and research.

In future versions the model could be trained to further more reservoirs to provide management to further more reservoirs. Features such as rainfall, soil moisture, temperature, humidity etc. can be included in the prediction of the water level by implementing multivariate models.

7. REFERENCES

[1] Unaffordable and Undrinkable: Rethinking Urban Water Access in the Global South, "https://www.wri.org/wri-citiesforall/publication/unaffordable-and-undrinkable-rethinking-urban-water-access-global-south" accessed on August 2021.

[2] Gronewold, Andrew & Clites, Anne & Hunter, Timothy & Stow, Craig. (2011). An appraisal of the Great Lakes advanced hydrologic prediction system. Lancet. 37. 577-583. 10.1016/j.jglr.2011.06.010.

[3] Lohani AK, Goel NK, Bhatia KKS (2007) Deriving stage–discharge–sediment concentration relationships using fuzzy logic. Hydrological Sciences Journal 52: 793-807.

[4] Kar AK, Lohani AK, Goel NK, Roy GP (2011) Development of Flood Forecasting System Using Statistical and ANN Techniques in the Downstream Catchment of Mahanadi Basin, India. Journal of Water Resource and Protection 2: 880- 887.

[5] Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition". arXiv:1410.4281 [cs.CL].

[6] CNN – LSTM Architecture, "https://machinelearningmastery.com/cnn-long-short-term-memory-networks/" accessed on August 2021.

[7] Scikit Learn, "https://www.tutorialspoint.com/scikit_learn/index.htm" accessed on September 2021.

[8] Plotly Dash, "https://towardsdatascience.com/building-dashboards-in-dash-591a6223efd3" accessed on September 2021

[9] Tom M. Mitchell, **Machine Learning**, India Edition 2013, McGraw Hill Education.

[10] Unsupervised learning, "https://www.ibm.com/cloud/learn/unsupervised-learning" accessed on October 2021.

[11] Supervised learning, “https://www.javatpoint.com/supervised-machine-learning” accessed on October 2021.

[12] Somchit Amnatsan, Sayaka Yoshikawa and Shinjiro Kanae, Improved Forecasting of Extreme Monthly Reservoir Inflow Using an Analogue-Based Forecasting Method: A Case Study of the Sirikit Dam in Thailand, 9 November 2018, Water 2018, 10, 1614; doi:10.3390/w10111614

[13] Tiantian Yang, Ata Akbari Asanjan, Edwin Welles, Xiaogang Gao, Soroosh Sorooshian and Xiaomang Liu, Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information, May 2017, 10.1002/2017WR020482

[14] Sergey, Christian Szegedy (2015). ”Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. ArXiv: 1502.03167 [cs.LG].

[15] Leo Breiman (1996). ”BIAS, VARIANCE, AND ARCING CLASSIFIERS”, Retrieved 19 January 2015.”Arcing [Boosting] is more successful than bagging in variance reduction”

[16] Soft computing approach for rainfall-runoff modelling: A review Vinay Chandwani*, Sunil Kumar Vyas, Vinay Agrawal, Gunwant Sharma (ICWRCOE 2015) Aquatic Procedia 4 (2015) 957 – 963

[17] R. J. Abrahart, L. M. See. Neural network modelling of nonlinear hydrological relationships. Hydrology and Earth System Sciences Discussions, European Geosciences Union, 2007, 11 (5), pp.1563-1579. ffhal-00305094f

[18] Long-Term Rainfall Analysis and Runoff Estimation in Mountainous Watershed: A Case Study from Mhadei River Basin, Goa and Karnataka Manoj Ibrampurkar1* and A. G. Chachadi2 Gond. Geol. Mag., V. 27(2), December, 2012. pp. 153-158

[19] Regression Analysis of Annual Maximum Daily Rainfall and Stream Flow for Flood Forecasting in Vellar River Basin P.Supriya a* , M.Krishnaveni a , M.Subbulakshmi . Aquatic Procedia 4 (2015) 1054 { 1061

[20] International Joint Commission. Levels Reference Study: Great Lakes-St. Lawrence River

Basin; The Board: Windsor, ON, Canada, 1993.

[21] Coppola Jr E, Szidarovszky F, Poulton M, Charles E (2003) Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions. Journal of Hydrologic Engineering 8: 348-360.

[22] Taiyuan F, Shaozhong K, Zailin H, Shaqiun C, Xiaomin M (2007) Neural Networks to Simulate Regional Ground Water Levels Affected by Human Activities. Groundwater 46: 80-90.

[23] Scikit Learn, "https://scikit-learn.org/stable/user_guide.html", accessed on October 2021.

[24] WRI resevoirs' datasets, "https://github.com/wri/ReservoirWatchHack", accessed on August 2021.

[25] IndiaWRIS reservoirs' datasets, https://indiawris.gov.in/wris/#/Reservoirs, accessed on September 2021.