FORECASTING THE FUTURE WATER HEIGHT OF A RESERVOIR WITH TEMPORAL DATA MINING

Mohamed Saudulla Aabid Rushdi

168265J

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

June 2018

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement, any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.

....

Mohamed Saudulla Aabid Rushdi Date

The above candidate has carried out research for the Masters thesis under my supervision.

....

Dr. Amal Shehan Perera Date

ABSTRACT

Throughout last few decades, flooding has been one of the most expensive natural disasters, which has had an increased impact on human causalities, property damage and rehabilitation. Thus, there is a crucial need in taking necessary actions to avoid or minimize the impact caused by floods to human lives as well as to the stability of the economy of the country. As to the most recent devastating flood experience faced in Sri Lanka in 2016, one of the root causes identified was the unpredictable decision making strategy, which was used to control and manage excess water in reservoirs due to heavy rain. Moreover, reservoirs are a key water storage source in water management, where they are utilized by various sectors for different purposes. Therefore, there is an essential need in taking the best decision in releasing water from reservoirs to minimize the damage caused by the floods and to optimize the utilization of water as a scarce resource.

Many researches have been carried out on the field of data science based on collected data from rivers, reservoirs & tanks to support decision making in water management. Those are mainly focused on classifying water release rules and ranges of a reservoir or tank. Research on forecasting the future water height of reservoirs, with both rainfall and uncertain water inflow due to human intervention, are very limited.

Hence, this research focuses on predicting the future water height of a reservoir, when the water inflow is uncertain and the reservoir receives a significant amount of rainfall. This would allow to minimize the risk of deadly floods by opening the sluice gates in time and to manage the water in an optimum manner for irrigation purposes. This research proposes the most effective set of features to forecast the water height of a reservoir on next three days. Furthermore, it presents the comparison of the performance of different regression models and the effectiveness of applying clustering techniques on top of the regression models. The result obtained from this research demonstrates that, K-Medoids clustering with feed forward artificial neural network model has the best performance in forecasting the reservoir water height when there is a significant amount of rainfall and the water inflow is uncertain.

ACKNOWLEDGEMENT

There is a tremendous amount of hard work and commitment behind the success of this research and I take this opportunity to convey my heartiest gratitude towards all the individuals who were a part of it. It is indeed my privilege to thank each one of them who gave me their valuable support, constant and necessary guidance while giving me the right level of motivation towards reaching this research's ultimate goal.

Firstly, I would like to express my heartfelt gratitude Dr. Amal Shehan Perera, Department of Computer Science and Engineering; as my project supervisor for his remarkable support, supervision and the inspiring advices given throughout, to make this research a success.

While my sincere thanks also go to Dr. Malaka Walpola, Dr. Charith Chitraranjan, Dr. Surangika Ranathunga and Dr. Indika Perera for their valuable feedback, encouragement and advices through time to time evaluations and providing me with the necessary advice and guidance in initiating and carrying out this research.

I also take this opportunity to thank Mr. Sanjaya Rathnayaka, and Mahaweli Authority for providing me the required data and helping me with the domain knowledge. I also should be grateful to darksky.net and timeanddate.com websites, from where I have collected some of the necessary data for the simulation.

Finally, I would like thank to each individual; especially my parents for their love & motivation and my loving wife for her support towards making this research a success during this entire period of time. I would also like to express my thanks to all my colleagues at MSc 16 batch and at Synopsys Lanka (PVT) Ltd.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

Abbreviation	**Description**
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
AR	Auto-Regressive
ARMA	Auto-Regressive Moving Average
CHRS	Center for Hydrometeorology and Remote Sensing
Day 0	The day from which the future water height is forecasted
Day 1	Following day to a given date (day 0), on which the water height is forecasted
Day 2	Second consecutive day from day 0, on which the water height is forecasted
Day 3	Third consecutive date from day 0, on which the water height is forecasted
GEP	Gene Expression Programming
GIS	Geographic Information System
IWMI	International Water Management Institute
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NN	Neural Network
PCA	Principal Component Analysis
RF-SVR	Reduced Feature - Support Vector Regression
RMSE	Root Mean Square Error
SPI	Standardized precipitation index
SVM	Support Vector Machine
SVR	Support Vector Regression
TDANN	Time Delay Artificial Neural Network

1 INTRODUCTION

1.1 Water Management

In the present, with the availability of low cost data storage solutions and high computational capacities, data science has become a rapidly growing and demanding field, where many practical problems are addressed or researched through data mining and machine learning. It is not only limited to electronic and computer fields, but also has been extended to medicine & health care, food & safety, irrigation, transportation as well as water management. There are many researches which have been carried on managing water as a scares resource, where new trends are heading towards addressing it through data science approaches.

Even though life cannot survive without water, it is crucial to manage water so that water does not harm us in return, where in the recent past, we have witnessed many natural disasters which stroke from water in the form of flood, tsunami...etc. It is also important to manage water, as it is the source of fishery industry and the life source of irrigation. With the modern technological advancements, electricity has become a basic need, where the main source of electricity generation in a tropical country like Sri Lanka is hydro-power stations, which is also involved with water.

Managing water in country level as well as in a global point of view is very important to mitigate natural disasters and to optimize other industries & needs that are mainly dependent on water. The problem becomes even more difficult to resolve, as different industries and different people have contradicting requirements in water management. For an example, hydro-power stations and fishery industry, which are dependent on reservoirs, need water to be stored in dams, while the department of irrigation needs it to be released to the farms and the department of disaster management needs it to be managed, to avoid floods. It is important to identify the tradeoff among different needs and demands to manage it in a proper way. One such study was conducted by Salim [1] on how to optimize such situations to gain the maximum benefit and to minimize disadvantages & losses.

The problem becomes more complex, when there are multiple reservoirs which are attached with a single river, such as in Mahaweli river [2] in Sri Lanka. In such cases, the problem cannot be addressed or optimized as an isolated reservoir, since the water outflow of one reservoir is the water inflow of another reservoir. It becomes even more complex, when addressing the problem in country level, where managing the electricity generation in hydro-power stations across the country becomes a crucial requirement.

Many researches are being carried out on addressing different aspects of water management. Most of such researches are based on big data technologies and data science, where one of the major sources of data is the sensors & monitoring system in reservoirs, which provides enormous amount of information for analytics, to get insights and build up prediction models.

1.2 Water Reservoirs and Tanks

Tanks and reservoirs play a vital role, not only in hydropower generation, but also in irrigation. For over 2000 years, thousands of irrigation tanks have been supporting the agricultural production, specially the paddy fields in the dry zone in Sri Lanka. These tanks are used to store the water coming from the upstream river as well as the water collected from the rainfall on and around the area of the tank. Afterwards, the water is released to downstream paddy fields according to the requirement of the crops.

Figure 1.1 shows relative positions of the sections of a small irrigation tank in Sri Lanka along with their traditional names. Usually, the tank is centered in a village where the tank bund is constructed crossing a river or multiple small streams to build a tank. The tank gets water from inflow streams or a river through a water hole, which is also called as ‘godawala’, where the sediment gets filtered [3]. The other main source of water inflow is the rainfall on the catchment area of the tank, where the water flows in to the tank from either side. The evaporation of the tank is reduced by the tree belt in either side, which blocks the wind flow across the tank as well as it reduces the temperature of the water in the tank. The tree belt area is also used to accommodate the excess water during spilling. Figure 1.2 shows the schematic of a tank and its main water balance components, where the tank receives water from the upper stream and through rain fall, while the water gets evaporated, seeped & released as spill flow and water issue.

Another advantage of tanks is prevention of floods, where it collects the water during heavy rainfalls and controls the water flow in the downstream. Subsequently, the dam prevents the flood by draining the water in a controlled steady manner.

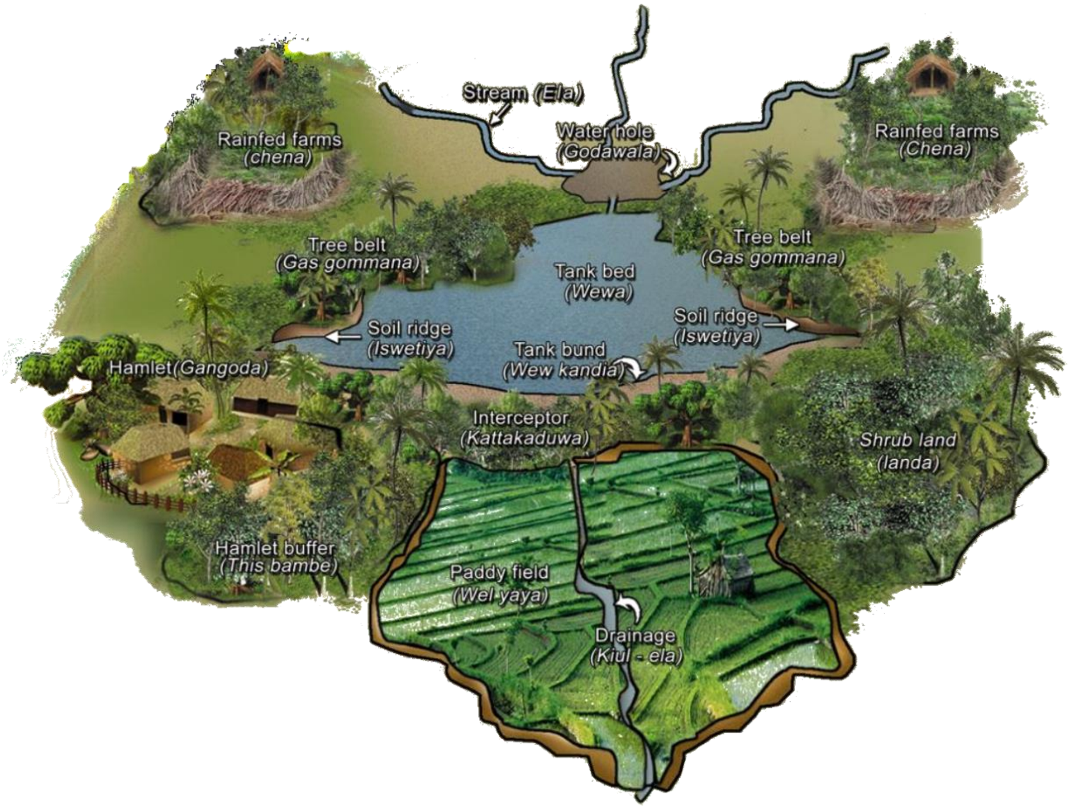


Fig.1.1 Components and their relative positions of a small tank in Sri Lanka [3]

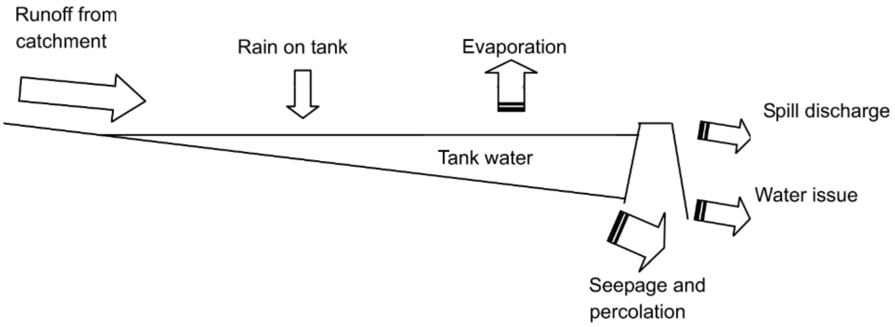


Fig.1.2 Schematic of a tank and the water balance components [7]

With the rapid development of several industries and technologies in the last century, there was a very high demand for electricity, which has led to construct large reservoirs to generate hydro power. These hydro power stations are popular in countries like Sri Lanka, because those countries are rich with hilly landscape with rivers flowing through it and the high availability of freely running water which in turn results in low cost electricity generation. But, creating a very large artificial dam across a natural river disturbs the natural water flow, which makes it essential to manage the water in the reservoir, so that consumers in the downstream banks, get it as per the requirement while not causing any flood condition by releasing a large quantity of water at once. Until recent past, these reservoirs and tanks were manually controlled as per the need without considering the probability of future water inflows.

1.3 Cascaded Tank System

In mountainous areas, tanks are built by constructing a dam across two mountains. If the elevation difference is sufficient, more than one tank can be built across a single river, which is called a cascaded system. Cascaded tanks are famous in Sri Lanka since ancient time, where the water outflow of several small tanks gets collected on a larger tank in the lower part of the river. Figure 1.3 is a map of Anuradhapura area in Sri Lanka, which shows a larger number of irrigation tanks connected in cascade. One such famous cascaded system is Thirappane cascaded system, which consists of 6 tanks within a distance of 8km [4].

Even though a cascaded system appears as getting the maximum benefit from a freely flowing river, managing the water in a cascaded system is a complicated problem, especially if there is a considerable amount of rainfall on catchment areas. If there are rainfalls on catchment areas, the water level of a tank may suddenly rise. But, due to the interdependency of tanks, spill flow or water issue gates cannot be opened at once, as it should be performed only after reducing water levels of downstream tanks without causing a flood. Reducing the water level of a tank cannot be done instantly, as it takes a considerable amount of time for a larger volume of water to get released through a spill flow in a controlled manner, without causing a flood in the downstream area.

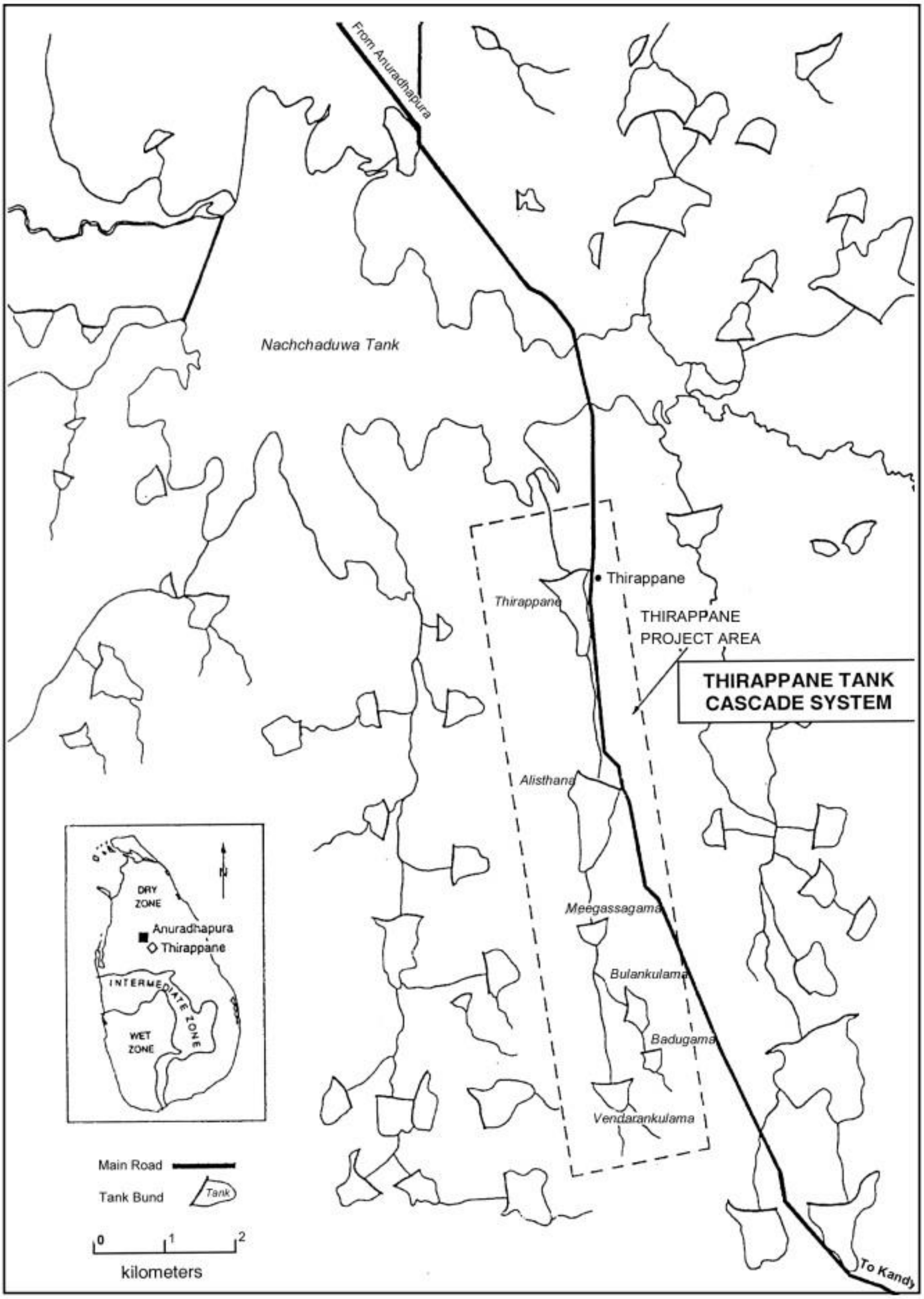


Fig.1.3 Cascaded irrigation tanks in Anuradhapura, Sri Lanka [7]

This delay affects the upstream tank, which may have already reached the spill flow water level. In most of the developing countries like Sri Lanka, this controlling is done manually as per the necessity of that moment. But it is unsafe, and may cause floods in downstream areas. One such similar incident occurred in the recent past in Sri Lanka in May, 2016, causing a massive flood condition in the downstream [5].

This problem can be appropriately solved through a proper prediction model, which could be used to predict future water levels in advance. Even though in early days, there wasn't any computational capacity to build prediction models, today with the present technology, this problem can be solved as a machine learning problem, with the availability of past data.

1.4 Knowledge Discovery in Data

Knowledge Discovery in data, data Mining or knowledge extraction is the process of extracting useful interesting knowledge from raw data. Reservoirs or tanks, provide raw data with time through data collection sensors, which consists of attributes such as water in flow & out flow volumes, water capacity and other environmental data such as precipitation, temperature, wind speed …etc. It is the process of data mining which converts them into useful knowledge and creates the potential future water level predicting ability. Typical knowledge extraction flow consists of following major steps, where most of the time is spent on data preprocessing steps. Once the preprocessed data is available, different data mining techniques can be applied on it to get insights and those can be evaluated for the accuracy and performance. Finally, the extracted knowledge can be presented as insights.

a) Data cleaning
b) Data integration
c) Data selection
d) Data mining
e) Pattern evaluation
f) Presenting knowledge

Data Mining can be either predictive or descriptive, where the descriptive approach gives direct insights from the current dataset, which is categorized under unsupervised learning such as clustering, association rule mining, anomaly detection…etc. The predictive approach provides a prediction model for the prediction of future data, such as classification models or regression models, which is considered as supervised learning.

The water level of a reservoir or a tank is a numeric value and it is continuous. So, future water levels can be predicted through building a prediction model from currently available data. But, if the water level of the reservoir or tank changes in a larger range, perhaps applying a direct regression model may not give correct results. It's because, most of the features which impact the water level may behave differently in different water levels, for which it can be considered as a classification problem to predict the future water level range. Moreover, it can be clustered in to different clusters based on the similarity of the features, where separate regression models can be trained within those clusters, which may simplify the regression model and may improve the overall prediction accuracy.

1.5 The Research Problem

To mitigate expensive disasters such as floods, it is important to understand the behavior of water, which requires a model. The model should be able to manage water in an optimum manner. There are many descriptive and predictive models that are already available, which are built based on different reservoirs, to address various research problems. But, most of these models are based on a single reservoir, addressing the problem as a temporal problem, which are mainly focused on classifying water release ranges and other water management rule classification problems. But, practically reservoirs are not isolated and most of the time there is more than one reservoir that are attached to one single river. Even in Sri Lanka, most of the reservoirs that are attached to a river are built in a sequence. Thus, there is a crucial need in building a model on the water flow, addressing it as a spatio-temporal problem, using the available data, which can be further extended to predict the behavior and to simulate the behavior of water across reservoirs which are attached with one single river. The problem is more challenging in Sri Lanka, since the country has a considerable amount of rainfall throughout the year, which was ignored in some of the previous researches, due to dry weather condition in the areas of studies. Moreover, the water inflow to the reservoir is inconsistent, when multiple reservoirs are connected in a sequence, which makes the problem more complicated, since the future water inflow is also unknown. Therefore, researches on predicting the exact water height of a reservoir, where there is an uncertain amount of water inflow and a considerable amount of rainfall in catchment area, are not available.

Thus, this research focuses on identifying the most effective set of features to predict the uncertainty of water inflow to the reservoir and the rainfall. Furthermore, it focuses on developing a simulation model, which could forecast the future water level of the reservoir for the next few days, irrespective of the uncertainty of water inflow and rainfall.

1.6 The Motivation

Currently in Sri Lanka, controlling sluice gates are done per the immediate situation, only to control the water level to protect the dam. But, opening the sluice gate at the very last moment to protect the dam may cause floods in the lower part of the river. Nevertheless, this model can be further extended to predict the future behavior of the reservoir water levels through which, necessary prevention mechanisms can be taken to mitigate any floods. This also can be extended to come up with a simulator to study the behavior of multiple reservoirs attached to one single river, for different inputs.

As a tropical country, Sri Lanka has lots of natural water resources, where those are utilized in irrigation since ancient times and in hydro-power generation. Even though, the state-of-the-art technology has been used during ancient times, currently the country is far behind in the modern technology in the field of irrigation and water resource management, whereas researches based on the water flow in reservoirs and rivers in Sri Lanka are also very limited. This research would be a turning point to re-establish the state-of-the-art technology in Sri Lankan irrigation and water resource management. Furthermore, this research can be extended to build water level prediction models on different main rivers and reservoirs chains and can be utilized to prevent floods like the recent flood condition in Colombo in May, 2016 [5], for which one of the major reasons was improper management of water in reservoirs. Thus, this research extends the existing researches on reservoir data and welcomes a new set of possible future work to resolve.

1.7 Reservoir in Study – Maduru Oya Reservoir

This research focuses on developing a simulation model based on the available past data of Maduru Oya reservoir in Sri Lanka. The Maduru Oya reservoir is located towards the center of the island, which is shown in the figure 1.4. It was reconstructed

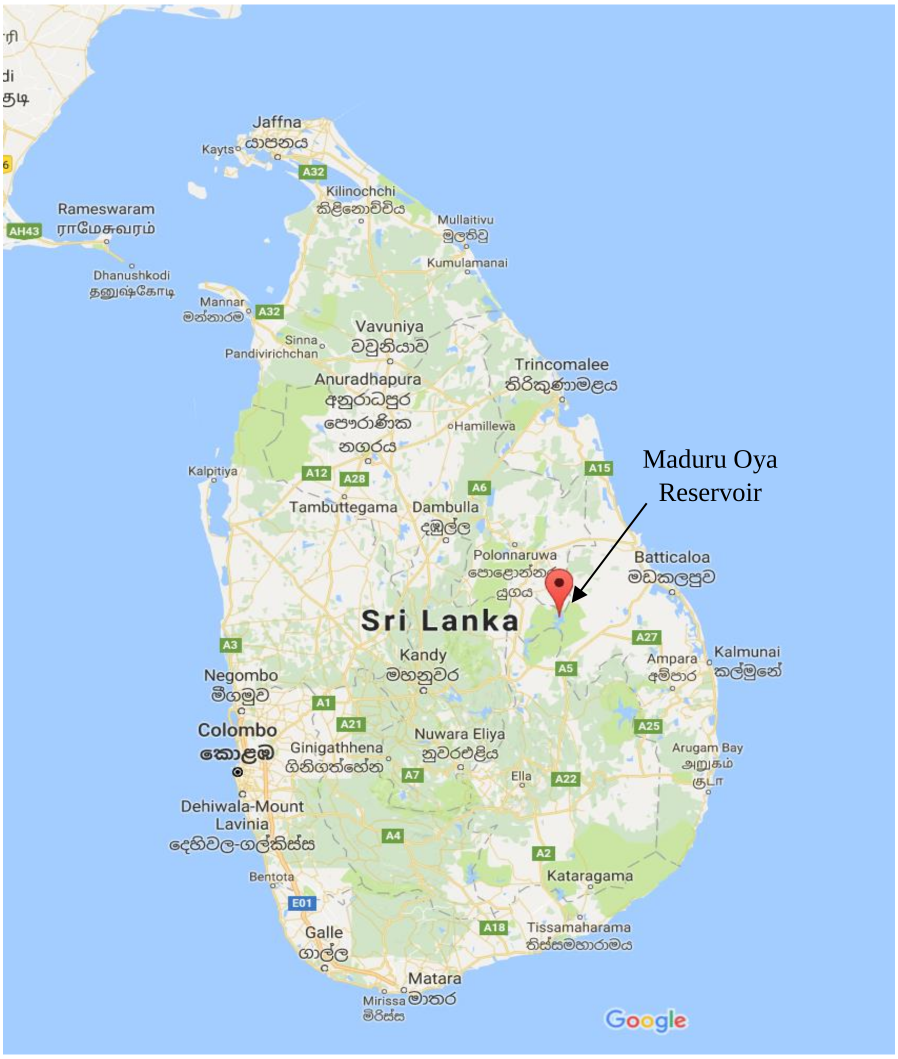

Maduru Oya Reservoir

Fig.1.4 Location of Maduru Oya Reservoir [35]

under Accelerated Mahaweli Development Project, which was originally constructed by an ancient king, of 1st century BC as per the carbon dating information [6]. Figure 1.5 presents, the ancient sluice structure which closely resembles the modern engineering design, which was constructed at the same location coincidently, which reflects the ancient hydraulic civilization in Sri Lanka.

As shown in the figure 1.6, this reservoir has a capacity of 596 million cubic meters, which can hold a water height of 96m from mean sea level. The main water inflow source of the reservoir is from a tunnel, which starts from Ulhitiya reservoir. In addition, it receives water from rainfall in the surrounded Maduru Oya national park, which acts as the catchment area of the reservoir. According to the available data, the reservoir has received an average rainfall of 2100mm annually. It can issue water from left bank and right bank sluices, while the excess water can be released as the spill flow.

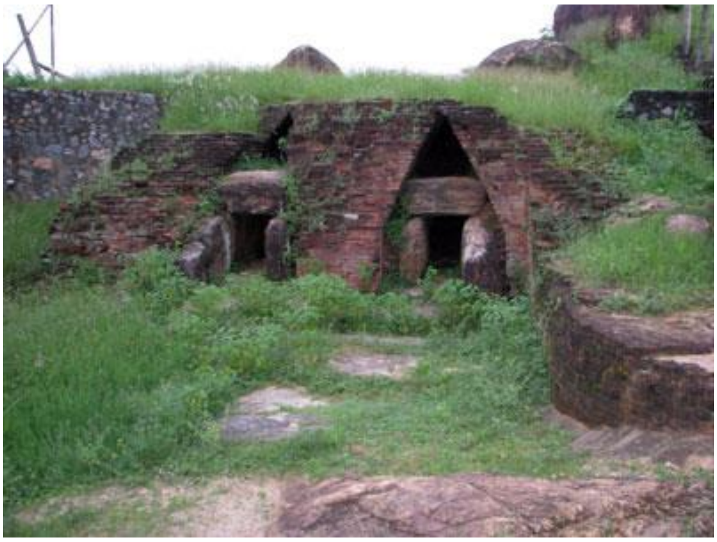
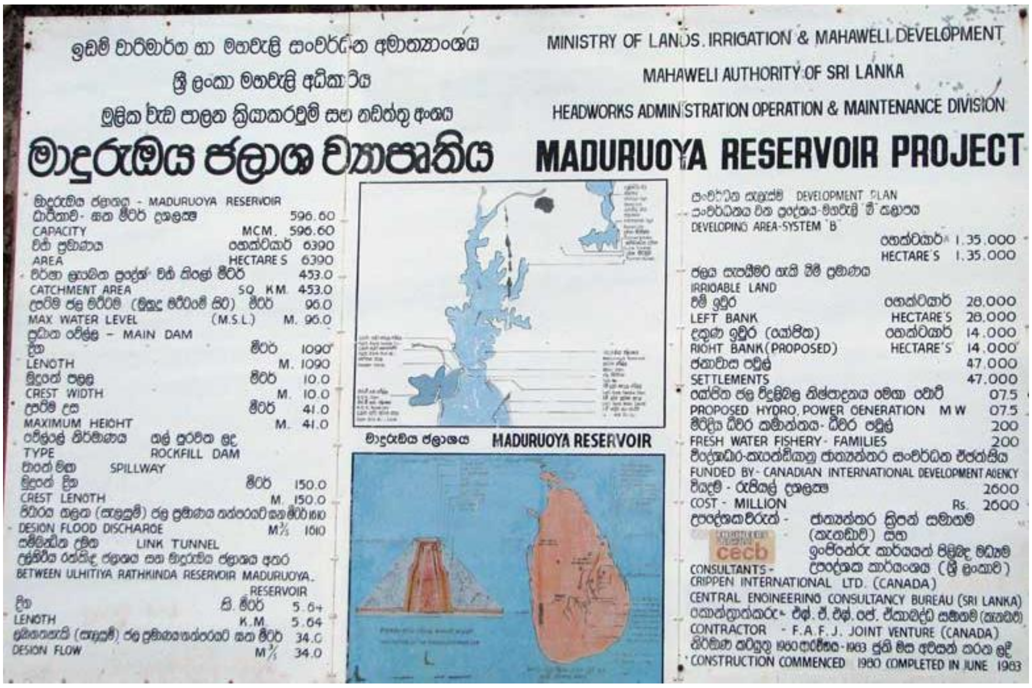


Fig.1.5 Ancient Maduru Oya Sluice of 1st century BC



ඉඩම් වාරිමාර්ග හා මහවැලි සංවර්ධන අමාත්‍යාංශය
ශ්‍රී ලංකා මහවැලි අධිකාරිය
මූලික වැඩ පාලන ක්‍රියාකරවුම් සහ නඩත්තු අංශය

MINISTRY OF LANDS, IRRIGATION & MAHAWELI DEVELOPMENT
MAHAWELI AUTHORITY OF SRI LANKA
HEADWORKS ADMINISTRATION OPERATION & MAINTENANCE DIVISION

මාදුරුඔය ජලාශ ව්‍යාපෘතිය

MADURUOYA RESERVOIR PROJECT

මාදුරුඔය ජලාශය - MADURUOYA RESERVOIR

ධාරිතාව - ඝන මීටර් දශලක්ෂ		596.60
CAPACITY	MCM.	596.60
වත් ප්‍රමාණය	හෙක්ටයාර්	6390
AREA	HECTARE S	6390
වර්ෂා ලැබෙන ප්‍රදේශ වර්ග කිලෝ මීටර්		453.0
CATCHMENT AREA	SQ. KM.	453.0
උපරිම ජල මට්ටම (මුහුදු මට්ටමේ සිට) මීටර්		96.0
MAX WATER LEVEL (M.S.L.)	M.	96.0

ප්‍රධාන වේල්ල - MAIN DAM

දිග	මීටර්	1090
LENGTH	M.	1090
මුදුනේ පළල	මීටර්	10.0
CREST WIDTH	M.	10.0
උපරිම උස	මීටර්	41.0
MAXIMUM HEIGHT	M.	41.0
වේල්ලේ නිර්මාණය	ගල් පුරවන ලද	
TYPE	ROCKFILL DAM	

වාන් මග SPILLWAY

මුදුනේ දිග	මීටර්	150.0
CREST LENGTH	M.	150.0
පිටාර ගලන (සැලසුම්) ජල ප්‍රමාණය තත්පරයට ඝන මීටර්		1610
DESIGN FLOOD DISCHARGE	M³/s	1610

සම්බන්ධක උමග LINK TUNNEL
උල්හිටිය රත්කිඳ ජලාශය සහ මාදුරුඔය ජලාශය අතර
BETWEEN ULHITIYA RATHKINDA RESERVOIR MADURUOYA RESERVOIR

දිග	කි. මීටර්	5.64
LENGTH	K.M.	5.64
ප්‍රවාහ ශක්ති (සැලසුම්) ජල ප්‍රමාණය තත්පරයට ඝන මීටර්		34.0
DESIGN FLOW	M³/s	34.0

මාදුරුඔය ජලාශය MADURUOYA RESERVOIR

සංවර්ධන සැලැස්ම DEVELOPMENT PLAN
සංවර්ධනය වන ප්‍රදේශය - මහවැලි "බී" කලාපය
DEVELOPING AREA-SYSTEM "B"

	හෙක්ටයාර්	1,35,000
	HECTARE'S	1,35,000

ජලය සැපයීමට හැකි බිම් ප්‍රමාණය
IRRIGABLE LAND

වම් ඉවුර	හෙක්ටයාර්	28,000
LEFT BANK	HECTARE'S	28,000
දකුණු ඉවුර (යෝජිත)	හෙක්ටයාර්	14,000
RIGHT BANK (PROPOSED)	HECTARE'S	14,000
ජනාවාස පවුල්		47,000
SETTLEMENTS		47,000
යෝජිත ජල විදුලිබල නිෂ්පාදනය මෙගා වොට්		07.5
PROPOSED HYDRO, POWER GENERATION	M W	07.5
මිරිදිය ධීවර කර්මාන්තය - ධීවර පවුල්		200
FRESH WATER FISHERY - FAMILIES		200

විදේශාධාර - කැනේඩියානු ජාත්‍යන්තර සංවර්ධන ඒජන්සිය
FUNDED BY - CANADIAN INTERNATIONAL DEVELOPMENT AGENCY

වියදම - රුපියල් දශලක්ෂ		2600
COST - MILLION	Rs.	2600

උපදේශකවරුන් - ජාත්‍යන්තර ක්‍රිපන් සමාගම (කැනඩාව) සහ ඉංජිනේරු කාර්යයන් පිළිබඳ මධ්‍යම උපදේශක කාර්යාංශය (ශ්‍රී ලංකාව)

cecb

CONSULTANTS - CRIPPEN INTERNATIONAL LTD. (CANADA)
CENTRAL ENGINEERING CONSULTANCY BUREAU (SRI LANKA)

කොන්ත්‍රාත්කරු - එෆ්. ඒ. එෆ්. ජේ. ඒකාබද්ධ සමාගම (කැනඩා)
CONTRACTOR - F.A.F.J. JOINT VENTURE (CANADA)

නිර්මාණ කටයුතු 1980 ආරම්භය - 1983 ජුනි මස අවසන් කරන ලදී
CONSTRUCTION COMMENCED 1980 COMPLETED IN JUNE 1983

Fig.1.6 Details of Maduru Oya Reservoir Project

2 LITERATURE REVIEW

2.1 Statistical Approach

The International Water Management Institute (IWMI) has developed a mathematical model for Thirappane cascaded tank system in Anuradhapura, Sri Lanka in 1992 [4]. This cascaded system consists of 6 tanks, where the distance from the most upstream tank to the most downstream tank is 8km. Researchers have considered the importance of managing the water level of all tanks as an interconnected problem and have developed a simulation model. Figure 2.1 shows a simplified model of the Thirappane cascaded tank system. Vendarankulama and Badugama tanks should be considered as starting tanks, where the water inflow from streams are unknown. In addition, these tanks have rainfall on tank surfaces and catchment areas. Water levels of these tanks are decreased by water release for irrigation, spill flow discharge, percolation through tank bed and through evaporation. Other tanks have adequately

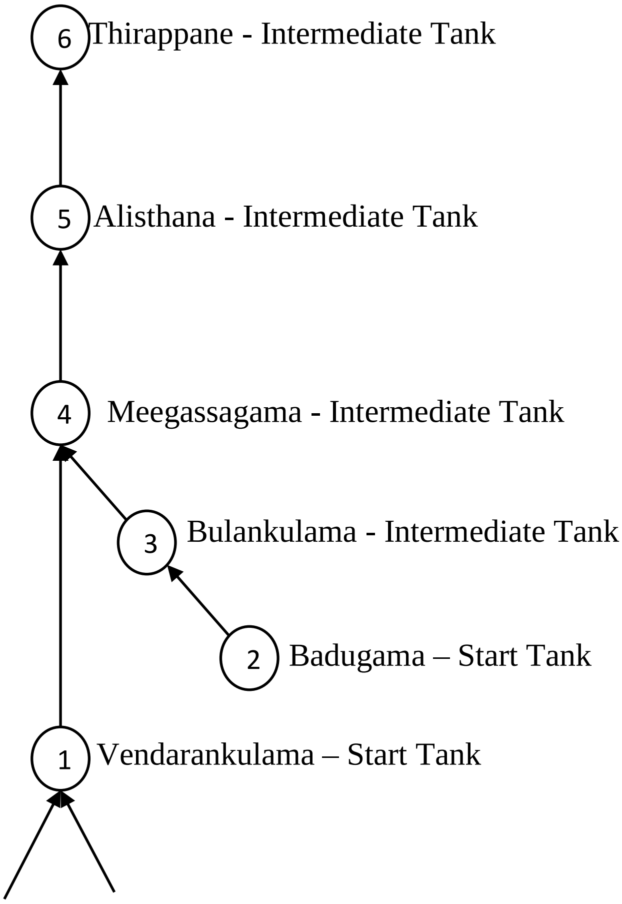

6 Thirappane - Intermediate Tank
5 Alisthana - Intermediate Tank
4 Meegassagama - Intermediate Tank
3 Bulankulama - Intermediate Tank
2 Badugama – Start Tank
1 Vendarankulama – Start Tank

Fig.2.1 Thirappane cascade system model

known water inflow components, where the water release of upstream tank is controlled manually. But, a percentage of released water from the upstream tank is utilized for irrigation and other purposes and percolated in the streams, which is an unknown fact. They have considered the previous tank storage, water inflow into the tank from upstream tanks & streams, water outflow from the tank as water issue, rainfall, evaporation on the tank surface and seepage & percolation as parameters, which affect the water balance in the tank. The evaporation is measured in a controlled environment in a bank and is calculated for the entire tank surface and the seepage & percolation is calculated as a function of the tank surface area, which is also dependent on the water height, where it is ignored. They have collected a set of training data and the coefficients are tuned with the training data to estimate the tank storage once in five days.

The IWMI has conducted another research on predicting water availability in irrigation tank cascaded systems more than a decade ago [7]. Even though they have not considered any data mining techniques, they have built a simulation model, which can be calibrated and used to predict the water availability in the tanks using data of one decade (i.e.1988-1997). They have also focused on the Thirappane cascaded tank system. Their cascaded model was formulated based on a simple structure, considering the major four tanks, which were fully functioning at the time of data collection. They have built mathematical expressions for the tank volume, water height and tank area of each tank based on survey data. Authors have used those mathematical expressions and constructed a simulation model for water balance in tanks, based on complex dynamic hydrologic processes associated with tanks. The water balance computation has been done starting from the most upstream tank to the most downstream tank on a process flow. They have used meteorological data and daily released amount of water for irrigation as input data for the model. For calibration, required water amount has been estimated based on the harvested area and the time of the crops, where crops require different amount of water in different stages. The water availability has been estimated based on the difference between estimated required water and the calculated balance water. This research was focused on conducting feasibility studies on the availability of water for irrigation.

2.2 Temporal Data Mining

2.2.1 Artificial Neural Network (ANN) models

With new trends in data science, there are many data based models and analytics that are built and researched on water resource management & planning. Even though classification algorithms, such as support vector machines or decision trees are widely used in classifying the data of one category, Peter Revesz and Thomas Triplet have focused on classifying temporal data, using linear classifiers [8]. In contrast to a standard classifier, which takes a set of inputs from one occurrence and provides a classification, their proposed temporal classifier takes feature values from the history. They have proven their approach through the data from meteorological database of the Texas Commission on Environmental Quality, by taking the historical values for past 15 days, where a higher prediction accuracy was achieved. They have also concluded that some features were more dependent on historical data than others.

The proposed method [8] was further extended on another research for mining temporal data of uncertain water reservoir data [9], where authors have built a model for a single reservoir in Cauvery River to predict the water release range on a monthly basis. They have compared the performance of Naïve Bayes classifier, decision tree classifier and multilayer perceptron classifier to classify water release ranges, where they have used storage, rainfall and inflow as feature variables and the release range as the classifying label. They have considered it as a monthly operational problem, where the data is collected on a monthly basis. Addressing the problem as a temporal problem, authors have considered the storage, inflow and rainfall values of past 2 months as features to predict the water release level range. Authors have obtained the minimum root mean square error of 0.2284 in multilayer perceptron classifier.

Hussain, Ruhana & Norita also have researched on a neural network based model to support decision making, which consists of predicting the water level and classifying the appropriate water release decision [10] based on Timah Tasoh reservoir in Malaysia. They have identified that the rainfall in far upstream and previous water levels as influencing parameters of the water level in the reservoir. Researchers have considered the sliding window effect of the water level of the reservoir and they have concluded that observing past 8 days of upstream rainfall data and water level data would give the minimum squared error in predicting the water level. For the test data, they have obtained an error of 0.416571, with ANN (artificial neural network) of 24-15-3. For the water release decision, authors have considered 7 classes, including six levels of gate opening states & a gate closing state and past & current water levels as parameters. In this process, they have obtained the best performance when considering 5 days of past data, which has given an error of 0.007085 against the test data set. It was obtained with ANN of 8-23-2 nodes. These reservoirs differ from the reservoir in our case study from the water inflow source. The main source of water inflow in these reservoirs is, through rainfall in catchment area while the reservoirs and lakes in Sri Lanka are mainly dependent on the water inflow from a river/stream or the water outflow from another reservoir, while it gets water from the rainfall as well. In addition, Athirah, Hussain & Ruhana have done a research on classifying the water level of Timah Tasoh reservoir in to 5 classes, on whether the water level would cause a flood [11]. They have also considered the past water level information as the main parameter, where they have trained an ANN of 4-21-1 considering water levels of past 4 days as features. For their classification model with 5 labels, they have obtained a mean square error of 0.0196 for the test data set.

When predicting the behavior of a reservoir, the main unknown parameter is the water inflow from the river. Diamantopoulou, Georgiou and Papamichail have researched on predicting the water inflow to the reservoir [12], for which they have proposed a cascade correlation Time Delay Artificial Neural Network (TDANN) model [13], which is a variation of multilayer feedforward ANN. The cascade part refers to the way the ANN is constructed, in adding hidden units once at a time connecting all previous units to the current unit. The correlation refers to the way those are trained, to maximize the correlation between the output of the hidden unit and the desired output of the ANN across training data sets. Authors have developed the model to forecast the one day ahead daily inflow into a planned reservoir in Almopeos river basin in Northern Greece. They have considered precipitation recorded in three precipitation stations and their statistical parameters of time series as features. Authors have obtained a root mean square error of 1.4035 (11.89%) and a correlation coefficient of 0.9946 in the test data.

2.2.2 Other prediction models

The Center for Hydrometeorology and Remote Sensing (CHRS) of University of California has conducted another research and they have built a simulation model of water outflow based on 12 reservoirs in California, using a tree-like graph model to classify and a regression model to predict the continuous target variable based on selected dependent variables and decision variables [14]. The model can be used to simulate the daily water outflow of reservoirs, based on precipitation, inflow, evaporation and few other parameters.

2.2.3 Reservoir water height forecasting models

Most of the water level prediction and decision making models are based on ANN, where most of the researchers have concluded that, ANN has given them better performance in prediction than other models, which may be due to the high tolerance of ANNs to noisy and nonlinear data. Shilpi and Falguni have also developed another ANN based model to forecast the water level of Sukhi reservoir, India [15]. which can predict the water level of next ten days based on the data of past ten days. They have used the data of past 23 years to build the model, where 16 years of data are used to train the model and 7 years of data are used to validate the model. Authors have considered the water inflow, water level and water release of past ten days as the set of features. The geographical area which they have considered is a significantly dry area and it has an annual rainfall of 700-1000mm, which has been ignored during the development of the prediction model. Authors have used testing patterns to evaluate the accuracy of the ANN trained model, where several patterns of input data have been employed to develop the optimum ANN model for the reservoir water level. They have experimented three different NN models namely Cascade, Elman and Feedforward back propagation, with 10 neurons in the hidden layer and one neuron in the output layer in each model. They have concluded that Feed Forward Backpropagation artificial neural network is the best model to forecast the water levels, out of the three models they have experimented. Authors have obtained a root mean square error of 0.82 and a correlation coefficient of 0.97 with the test data set using the model.

Fi-John and Ya-Ting have developed a novel approach named, adaptive neuro-fuzzy inference system (ANFIS) to predict the water height in next 1 to 3 hours, during floods [16]. They have validated the model with Shihmen reservoir in Taiwan, where they get frequent floods due to high mountains with steep slopes and typhoons. As per the authors, the water level changes suddenly during flood periods, which is not only due to meteorological effects, but also due to human operating decisions on reservoir controls. ANFIS is a multilayer feedforward neural network, where they use fuzzy reasoning to map an input space to output space. Researchers have developed two models, considering reservoir outflow feature as the user control variable in one model, and another model without considering reservoir outflow feature. Other features which they have considered are the water levels from five gauge stations in the upstream of the reservoir. Based on the distance to the gauges from the reservoir, different time shifted water heights of 0 to 5 hours have been considered to derive the 18 features of the first model and 17 features of the second model. Fi-John and Ya-Ting have observed RMSEs of 0.597, 1.007 and 1.186 for the predicted water height in next hour, in next two hours and in three hours respectively, in the first model, where they have considered the water outflow as a feature as well. Authors have obtained RMSEs of 0.720, 1.436 and 1.652 for the predicted water height in next hour, in next two hours and in three hours respectively, in the second model.

Nariman et al. have developed another two ANFIS models to predict the water height, which they have evaluated on Klang Gates Dam and Rantau Panjang station on the Johor river in Malaysia[17]. Authors have used the previous water height and rainfall to predict the current water height. They have obtained the lowest RMSEs of 0.2420263 and 0.193227682 for Klang Gates Dam and Rantau Panjang station respectively, where the corresponding maximum error percentages are 4.01054% and 4.3486%.

Fatih, Mustafa and Ozgur have researched on developing a water height prediction model based on the data of Millers Ferry Dam in United States[18]. They have researched on autoregressive model (AR), autoregressive moving average model (ARMA), multi-linear regression model and ANN model. Authors have developed models considering up to past 5 water heights to predict the following water height.

They have obtained the lowest mean squared error (MSE) of 0.0032 for the ANN model, when considering past 5 water heights, while the corresponding mean absolute error is 0.0415. The reservoir receives water from the Alabama river, which authors have not considered as a separate feature. Moreover, they have not considered the rainfall in the model, where the corresponding area receives an average annual rainfall of 1300-1400mm [19].

Ozgur has conducted another research with Jalal and Bagher on forecasting the water level of reservoirs, where they have forecasted the water height of Iznik Lake, in Bursa province of Turkey for following three days[20]. They have used the past water height data of 22 years on this research, where 11 years of data was used for training, another 6 years for testing and the remaining data of 5 years to validate the model. They have obtained the best results, when past three days of water heights are used to develop the model. Researchers have experimented with GEP, ANFIS ANN & ARMA models and have obtained lowest RMSE values for GEP model, which are 0.040, 0.070 and 0.102 for the forecasted water height on next three days respectively. Corresponding R^2 values are measured as 0.998, 0.994 and 0.988. Moreover, authors have observed RMSE values of 0.041, 0.073 and 0.104 for the forecasted values from ANFIS model and they have observed 0.042, 0.079 and 0.114 for the forecasted values from ANN. As per their experiments, the best performing GEP model was constructed with the function set of {+, -, ×, ÷, $\sqrt[3]{}$, √, ln, e^x, x^2, x^3, sin x, cos x, Arctgx}. The best performing ANN model was 3-10-1. The lake area receives an average annual rainfall of 600-800mm[21], which is lower than the average annual rainfall of the driest area in Sri Lanka, which is around 900mm[22]. Authors have not considered the uncertainty of rainfall separately, but it has given the best results with lowest RMSE values out of the past researches in predicting the water height of a reservoir or a lake. But this research may not be directly applicable to the reservoirs in Sri Lanka, since Sri Lanka receives an average annual rainfall of around 900mm in the driest area and over 5000mm in wettest regions[22]. Furthermore, it is not applicable in general as the amount of rainfall received varies based on the geographical area.

Manali and Megha built another model recently, combining K-Medoids clustering with the multiple regression technique, which is an extension of linear regression technique to forecast the water level in lakes, where they have used the data from lake Powell [23]. They have concluded that the regression model on top of K-Medoids clustering performs better than the regression model on top of K-Means clustering due to drawbacks of K-Means clustering compared to K-Medoids clustering such as, producing empty clusters, problems in handling outliers...etc. As the future work, authors have suggested on researching on various regression models to improve the forecasting accuracy further.

2.3 Spatio–temporal data mining

Mohan and Peter have come up with another approach named RF-SVR (Reduced Feature - Support Vector Regression) by combining Principal Component Analysis (PCA) with multi-class SVM (Support Vector Machine) to effectively model the non-linear spatial relationship between the dependent variable and a set of predictors in a spatial framework [24]. Authors have concluded that, RF-SVR model performs better than support vector regression and traditional ANN models with spatio-temporal data, where the data is distributed across time as well as space. They have used the data of multiple reservoirs, which are connected along North Platte river in United States, with larger number of attributes. But, they have used a very small dataset and have suggested further extending it using a larger dataset with more number of data sources. Venkateswara, Govardhan and Chalapati [25] have further studied issues and challenges, in handling spatio-temporal data in data mining problems. They have also presented different applications and researches done in spatio-temporal data and have suggested some future work for identified key issues.

2.4 Water disaster management with data mining

Bahram, Reza and Banafsheh have researched on building up a set of monthly operational rules, for a cascaded reservoir system [26], where they have considered the monthly inflows, reservoir storage at the beginning of the month and downstream water demands as the inputs and have built the rules set based on Naïve Bayesian classifier. The objective of this research was to minimize the damage of floods, which occurs due to poor management of water and to optimize the supply to fulfill irrigation demands. Another group of researchers from University of Minnesota, Crookston, have researched on building a model for flood prediction and risk assessment on Red Lake valley [27], where they have built a computer simulation platform to support flood prediction and risk assessment using advanced geo-visualization and data mining techniques. They have collected and analyzed observation data with geographic information system (GIS) data, to identify frequent patterns and to discover important relationships between the data variables using predictive models that they have integrated into a Geo-Simulation platform, which offers 3-D geo-visualization to help decision makers to predict potential floods and assess associated risks. In this research, authors have used temporal data to predict potential flood conditions, using Naïve Bayes classifier. Furthermore, they have tried to identify possible association rules and have tried to visualize possible flood affecting areas to assess the risk in flood.

Another research was done on predicting annual and seasonal droughts through handling it as a classification problem, where authors have used rainfall temporal data to identify droughts [28]. Standardized precipitation index (SPI) is usually used to identify problems such as drought, flood and crop yields, where it quantifies the precipitation deficit. Authors have constructed a cumulative rainfall series for the k-reference periods by using monthly rainfall stages to form sub-homogeneous regions for the regional frequency. They have used it to choose the best fit regional distribution for the cumulative rainfall series obtained from the stations in the sub-homogeneous regions and have normalized it. It is also used to find the SPI value and the decision tree is applied to classify seasonal and regional droughts.

Chia-Cheng, Mi-Cheng and Chih-Chiang have come up with a novel model using decision tree classifier combined with neural network based predictor for water stage forecasts in a river basin during typhoons [29]. As per the authors, the water level in Taipei Bridge station changes not only due to the rainfall over steep terrain in reservoir basin and water inflow from other reservoirs, but also due to vibrations of water body due to sea tides. The water level may rise to peak when the tidal effect & the rainfall are high and the water inflow from other reservoirs is also high, which occurs during typhoons. Authors have used a model, which combines both decision trees (classification and regression trees) with multilayer perceptron and radial basis function ANN techniques to get a higher accuracy. Researchers have considered the rainfall, released water capacity and current water height in several places as the set of features and they have obtained accuracies of 80.03%, 85.55% and 89.94% in classifying the water level after one hour, two hours and three hours respectively. Since this model is developed based on a single reservoir, authors have suggested extending the model to a reservoir network, with channel level routing as a future work.

2.5 Predicting the uncertainty of rainfall

Compared to previous researches, the reservoir in study, receives a substantial uncertain amount of rainfall throughout the year in addition to the unpredictable water inflow from the inflow stream. Therefore, it is crucial to predict the uncertainty of rainfall to forecast the future water levels. As per the previous researches[30], [31], satellite based predicted cloud cover is a good forecaster for the rainfall. Therefore, the could cover is a potential feature to mitigate the uncertainty of rainfall. In addition to that, tropical countries, such as Sri Lanka, receives the rainfall in various monsoons in different months of the Gregorian calendar, which occurs in a repetitive pattern [32]. Even though precipitation has a repetitive pattern, there are many occurrences, where sudden heavy rainfalls and thunderstorms occur due to various other facts. According to the Department of Atmospheric Sciences of University of Washington and Yamauchi, lunar phase and moon’s gravitational force are a couple of other factors, which affect the rainfall [33],[34].

3 METHODOLOGY

Predicting the water level of a tank/reservoir is a spatio-temporal problem, which spreads across the time and geographical location. Not only the uncertainty of water inflow to the reservoir from upstream but also the unpredictable weather condition in Sri Lanka makes it difficult to forecast the water level of a reservoir. This research focuses on identifying the potential features, which has a high correlation with the water level of the reservoir. Furthermore, it is necessary to come up with a suitable predictive model, which would forecast the future water levels of a tank for the next few days based on the historical data of past few days.

In addition to the historical data, the water level is also dependent on 'expecting volume of water issue' control variable.

3.1 Data collection

This Research is focused on developing a simulation model based on the available historical data of Maduru Oya reservoir in Sri Lanka. The dataset contains the following daily data of past 11.5 years, which dates from 1st January 2005 to 31st of May 2016.

Description	**Unit**
Amount of water received from Ulhitiya tunnel	Million cubic meter (mcm)
Volume of water issued through the left bank sluice	Million cubic meter (mcm)
Volume of water issued through the right bank sluice	Million cubic meter (mcm)
Spill flow volume	Million cubic meter (mcm)
Water head height from mean sea level	Meter (m)
Rainfall	Millimeter (mm)
Gross storage	Million cubic meter (mcm)

Table 3.1 Features in the dataset

For the development of machine learning model to predict the water level of tomorrow or day after tomorrow, it is possible to use the amount of water received and the rainfall in past few days as potential features. But, despite the historical data of volume of water received and rainfall in next 24 hours, which has the highest impact on next day's water level is available, it cannot be used to train the model, since practically the amount of water it is going to receive through the tunnel and the rainfall in next 24 hours is uncertain. Therefore, it is crucial to identify additional potential features, which may mimic those values. It is also important to identify the potential features, which may represent the seepage and evaporation of the reservoir.

For these information, predicted future weather forecast values, such as rainfall, temperature, cloud cover, humidity, wind speed…etc, can be considered as features. Other potential features which can be calculated are, the length of the day time, which is the difference between the sunset and sunrise time, lunar phase and the distance to the moon from the earth, which represents the gravitational force of moon [33], [34]. The predicted weather forecast data and information on sun rise & sunset time of Maduru Oya reservoir area are collected from www.darksky.net, which provides the predicted weather forecast data with a higher accuracy. These data contain a predicted categorical weather state, temperature, rainfall, cloud cover, wind speed, humidity, pressure, visibility and UV index. Moreover, calculated values of lunar phase and the distance from earth to moon of the 11.5 years are collected from 'www.timeanddate.com' web resource.

In addition to the above-mentioned features, it was observed that the month and the day of the year also can be potential features, which may have higher correlation values with the water height, where it reflects the repetitive pattern of monsoons in Sri Lanka. Figure 3.1 and 3.2 present the daily rainfall in 2009 and 2013. The graphs presented in the figures, illustrate the unpredictability of rainfall, where different annual rainfall patterns are visible in two different years. Despite this difference, both graphs mimic a similarity on seasons, where there is a considerable amount of rainfall, as a result of rainfall monsoon seasons.

3.2 Data preprocessing

When datasets are collected from various sources, those need to be carefully analyzed for any errors and mistakes. Reservoir data is collected by humans, which contains a considerable amount of human errors and negative values for sensor

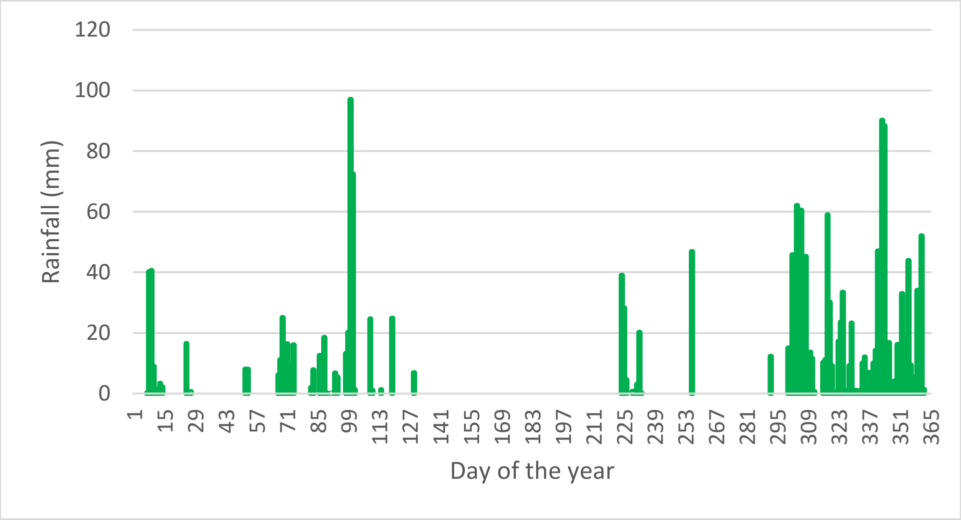


Fig. 3.1 Daily rainfall in year 2009

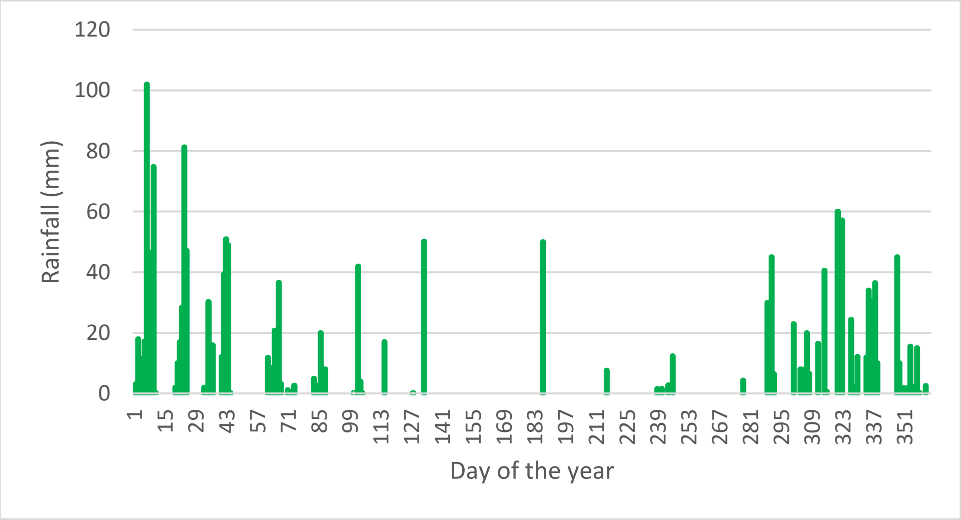


Fig. 3.2 Daily rainfall in year 2013

failures. As the preprocessing step, negative data points are removed from the data set, since they are very few compared to the dataset size. But, to rectify human errors, it is required to carefully analyze the dataset through necessary graphs, to identify the abnormal outlays. Human errors in the water head level are rectified by comparing with the neighbor data points and comparing with the corresponding gross volume values. Sudden spikes in the figure 3.3 are such anomalies.

Besides the errors found in the data, there are missing values in the predicted weather forecast values and calculated lunar phase & distance values, which are interpolated to find the missing values.

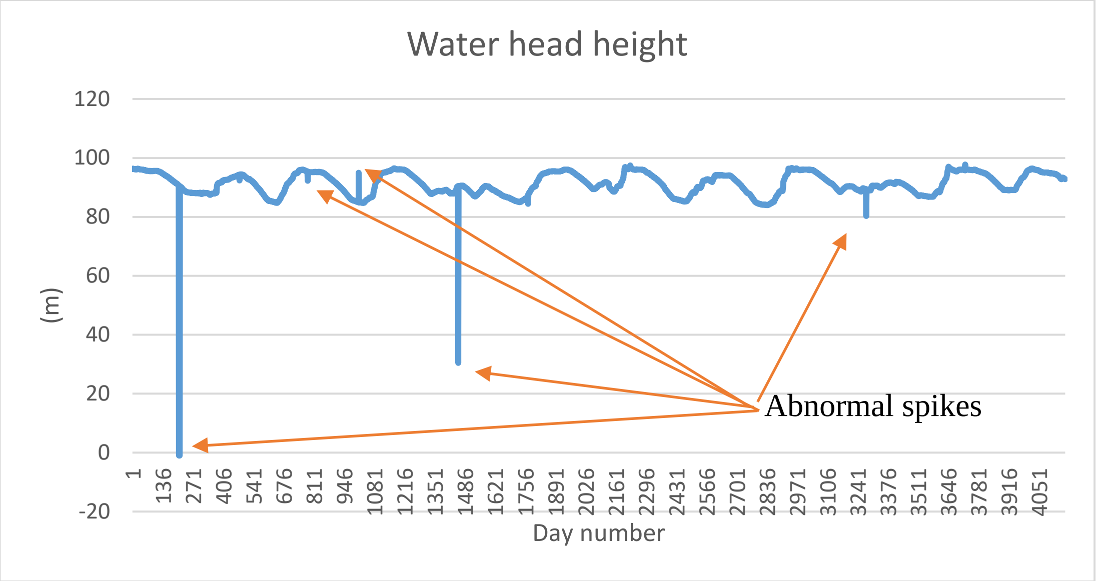


Fig. 3.3 Human errors in water head height

3.3 Building the machine learning model

Once the data set is preprocessed, out of the 11.5 years of data, 2.5 years of data is separated as the test data set, while the remaining data is used to train the model and to cross validate with 10-fold cross validation.

3.3.1 Statistics of the data

Once the data set is preprocessed to eliminate the errors occurred in data collection, following statistics are observed. Table 3.2 shows the observed mean, maximum, minimum values and the standard deviation of the training data set, which is of 9 years. It can be observed that; the rainfall has a very high variation throughout the period from 0 to 242 millimeters. Furthermore, it was observed from the data that Maduru Oya reservoir area has received an average annual rainfall of 2100mm, which is several times higher than the reservoirs considered in all the previous researches on predicting the water height. Moreover, as per the table 3.2, the length of the day time has varied throughout the year nearly by one hour. Even though the right bank sluice gate is opened as a secondary water source, as per the table, it has varied in a greater range, while having an average value closer to zero, which indicates the right skewed distribution of data.

	Maximum value	Minimum value	Average value	Standard deviation
Water height (m)	97.50	83.98	91.12	3.570
Rainfall (mm)	242	0	5.51	16.049
Volume of water received from Ulhitiya tunnel (mcm)	2.94	0	0.89	1.241
Volume of water issued from left bank sluice (mcm)	3.26	0	1.58	1.028
Volume of water issued from right bank sluice (mcm)	3.46	0	0.02	0.204
Spill flow volume (mcm)	30.41	0	0.18	1.396
Predicted cloud cover	1	0	0.60	0.205
Length of day time (minutes)	754	700	727.45	18.387

Table 3.2 Statistical parameters of the training data set

Table 3.3 shows statistical parameters for the test data set. It can be observed that, values of the test data set are much similar to the values in training data set.

	Maximum value	Minimum value	Average value	Standard deviation
Water height (m)	97.06	86.78	92.22	3.008
Rainfall (mm)	143	0	6.22	17.262
Volume of water received from Ulhitiya tunnel (mcm)	2.94	0	0.97	1.314
Volume of water issued from left bank sluice (mcm)	3.72	0	1.41	1.032
Volume of water issued from right bank sluice (mcm)	1.3	0	0.02	0.151
Spill flow volume (mcm)	51	0	0.67	4.523
Predicted cloud cover	1	0.31	0.58	0.191
Length of day time (minutes)	754	700	727.18	18.045

Table 3.3 Statistical parameters of the test data set

3.3.2 Developing the model

As per the researches performed previously, this research problem should be addressed as a temporal problem, with a sliding window effect, where the future water level is dependent on a set of previous features. Different features depend on different amount of historical data, where the optimum number of historical data points for each feature should be identified by calculating necessary mathematical measures and through cross validation. Table 3.4 shows a snapshot of the data points, which shows the necessity of sliding window effect. According to the data presented in table 3.4, even though there was a heavy rainfall, a spill flow could not be observed on 8th of January, while the rainfall on 9th of January has caused a spill flow on 10th of January. Not only the rainfall, but also the previous water heights and inflow volumes too has a sliding window effect on the past data.

According to the previous researches, a higher accuracy can be achieved by classifying the water level in to ranges or by clustering the data points into clusters and then predicting the water level within that range through different regression models for each range, instead of directly building a regression model for the entire range [23]. The main reason for this is that, the percolation of a tank is dependent on the water capacity, the evaporation is dependent on the surface area of the water, the water release is also dependent on the water head height of the tank and the accuracy of a regression model with a similar data points is higher than the accuracy of one large model, which covers all the data points.

Date	From Ulhitiya (mcm)	Spill Flow (mcm)	Head of water (m)	Rainfall (mm)	Gross storage (mcm)
8-Jan-11	0	0	95.6	242	565.22
9-Jan-11	0	5.4	96.15	106.5	600
10-Jan-11	0	15.4	96.75	18.8	639.61
11-Jan-11	0	17.3	96.9	0.4	649.78

Table 3.4 Necessity of sliding window technique

ANN, Naïve Bayesian classifier and SVM models have performed better in classifying water levels and classifying water release decisions. Furthermore, researchers have addressed it as a clustering problem, where the regression models are trained within those clusters, in which case K-Medoids clustering has performed better than the K-means clustering. ANFIS, GEP, ANN and linear regression models have been used to predict water levels in previous researches. Figure 3.4 shows the summary of the possible approaches, in which this problem can be addressed. Even though there are many possible combinations of approaches, this research focuses on experimenting on developing a solution, considering a single larger regression model and clustering the data points & training multiple smaller regression models for different clusters, which is illustrated by the left and the middle branch of the diagram in figure 3.4. Among various possible approaches and possible algorithms, the best performing feature set and the model should be identified through experimenting with different models and by cross validating the models.

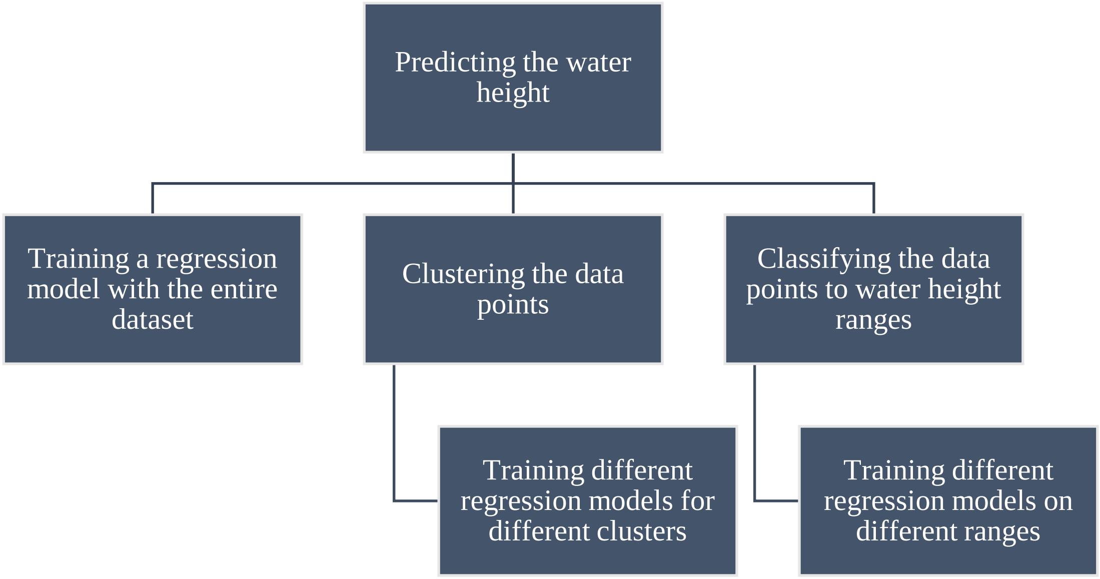

Predicting the water height
Training a regression model with the entire dataset
Clustering the data points
Classifying the data points to water height ranges
Training different regression models for different clusters
Training different regression models on different ranges

Fig. 3.4 Possible approaches

3.4 Evaluating the features and models

One of the basic and effective feature selection techniques is, analyzing the Pearson correlation coefficient values of the features and the corresponding target value, where it assumes that the two variables are in a linear relationship and provides a value in [-1,1] range. Equation 1, presents the equation of Pearson correlation coefficient, where x_i and y_i are corresponding instances of independent and dependent variables respectively. $\bar{x}$ and $\bar{y}$ in the equation denotes the mean of those variables and n denotes the sample size. The '-1' implies the inverse linear relationship between the independent variable and the dependent variable, such that when the independent variable increases, the dependent variable decreases linearly and vice versa. The '1' implies a proportional linear relationship between the independent and dependent variable, where as a value closer to ‘0’ implies that there isn’t a linear relationship among the variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{————————} \qquad (1)$$

Additionally, if the relationship between independent and dependent variable is non- linear, the Spearman correlation coefficient can be analyzed, which considers the non- linearity of the relationship as well.

The accuracy of a regression model can be evaluated with the root mean square error (root mean square deviation), which is abbreviated as RMSE or RMSD, where it gives an idea on the average error in predicting for the entire data set. As shown in the equation 2, it is calculated as the square root of the mean of the squared error. $\widehat{y_i}$ and y_i in the equation denotes the predicted and actual values accordingly for the ith instance. Since 10-fold cross-validation is done with the training data set, average RMSE of the 10-folds is considered as the RMSE value of cross-validation data set to evaluate the accuracy of the model. Moreover, the coefficient of determination (R^2) can be measured to identify how well the actual values are replicated with predicted values. It is the proportion of variance between the predicted values and the actual values, which is shown in the equation 3, where the same set of symbols of equation 2 is used.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}} \qquad \text{————————} \qquad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{————————} \qquad (3)$$

4 SOLUTION ARCHITECTURE

This research focuses towards building a simulation model, which could be used to predict the future water level of a reservoir. The system contains a pretrained machine learning model, which is capable of predicting the water level with a given set of features. The system would store the length of the day time feature internally to improve the usability.

4.1 User input

To provide the future predicted water levels, the system would request the following input values from the user.

1. Current date (dd/mm/yyyy)
2. Present water height (m)
3. Yesterday's water height (m)
4. Rainfall in last 24 hours (mm)
5. Water inflow in last 24 hours (mcm)

The main control variable of the simulation model is the amount of water, the user is planning to release. The user should provide the total water volume, which the user is planning to release through sluices and through spill flow during next 0-24 hours (day 1), 24-48 hours (day 2) and 48-72 hours (day 3).

4.2 Output

For the above explained user inputs, the system would provide the predicted water level in 24 hours (day 1), 48 hours (day 2) and in 72 hours (day 3).

4.3 Training flow of the system

Figure 4.1, provides an overview of the system, where the 'Machine Learning Model' is a pre-trained model to predict the water height for the given set of inputs. Figure 4.2 shows the flow illustrating the way the machine learning model is trained. As the initial step, temporal features such as, water height on previous day, volume of water issued on following day (day 1) and volume of water issued on second

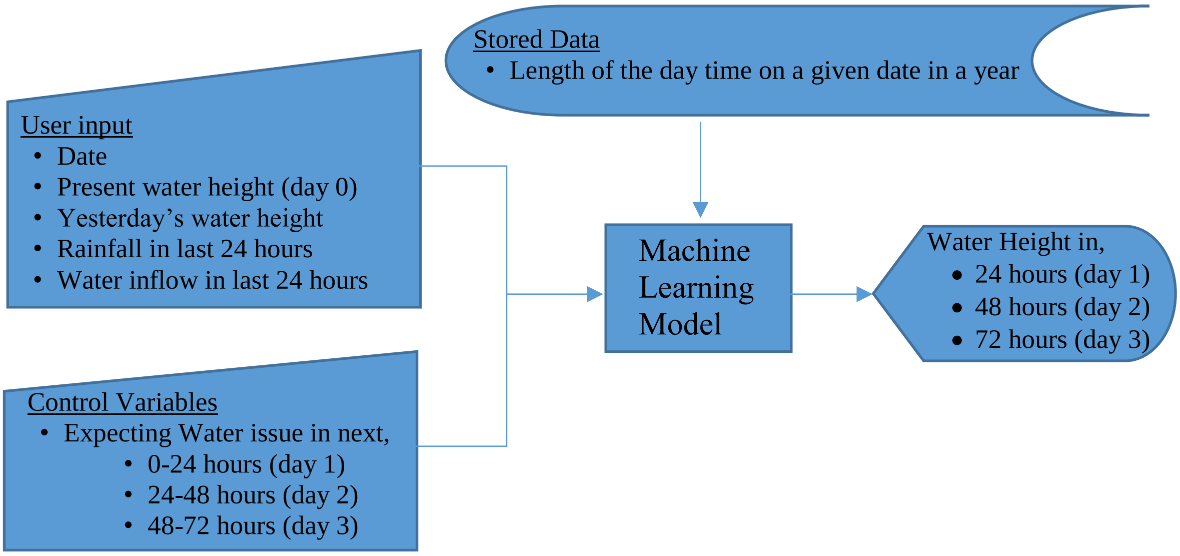


Fig.4.1 Overview of the system

consecutive day (day 2) …etc. are constructed from the data by sliding the datasets. Secondly, data instances with invalid values such as -1, which are reported due to sensor failures are eliminated. Next, the dataset is shuffled, before splitting to 10-fold cross validation to maintain the uniformity across the models in 10-fold cross validation.

The machine learning model of the system consists of two different phases, where at the first phase it would cluster the data instances into two different clusters based on the features with K-Medoids clustering. Clustered data instances are then passed through a regression model, where a separate regression model is trained for each cluster. Since the 10-fold cross validation creates ten different models, the average of those models is considered as the final value. Then the predicted water height on day 1 is fed back as a feature to train the model to predict water height on day 2. Subsequently, the predicted value from that model is fed back again, to train the model to predict the water height on day 3.

4.4 Prediction flow of the system

Figure 4.3 shows the prediction flow of the model. During the prediction process, the corresponding features to predict the water height on day 1, which are shown in the figure are fed into ten different models, which are constructed with 10-

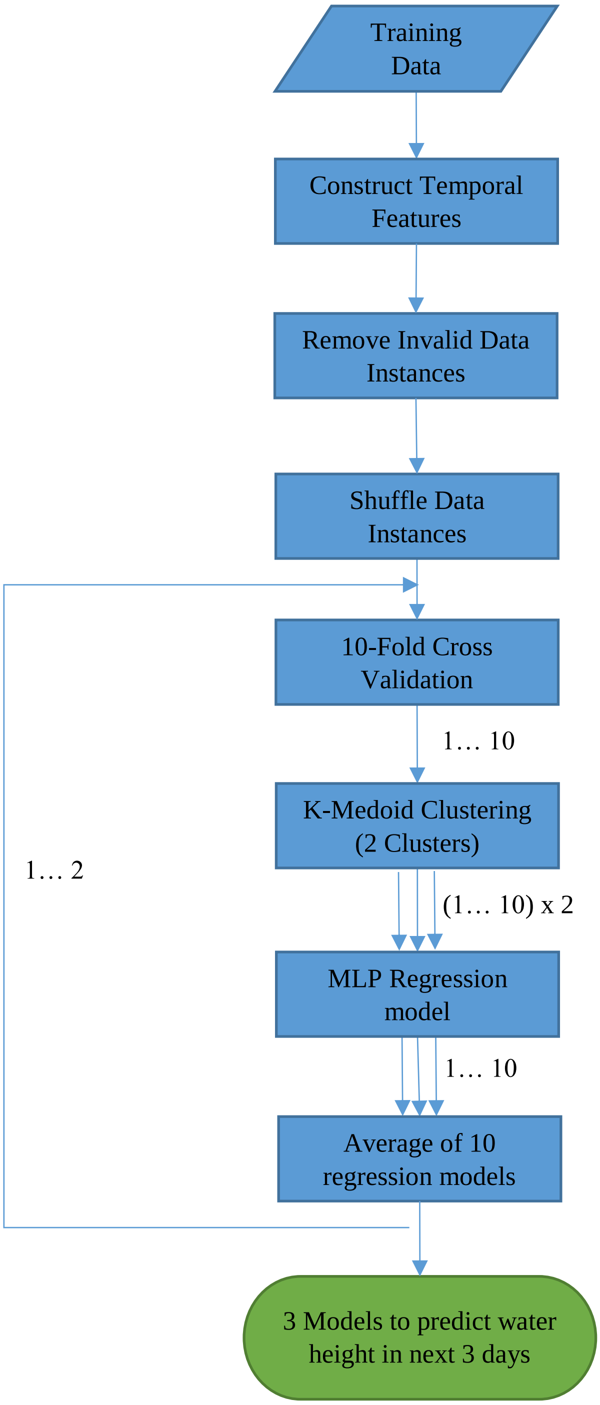

Training Data
Construct Temporal Features
Remove Invalid Data Instances
Shuffle Data Instances
10-Fold Cross Validation
1… 10
K-Medoid Clustering (2 Clusters)
1… 2
(1… 10) x 2
MLP Regression model
1… 10
Average of 10 regression models
3 Models to predict water height in next 3 days

Fig.4.2 Training the machine learning models

fold cross validation. In these models, firstly the closest cluster center is identified and the data instance is assigned to that cluster and the water height is predicted with the regression model in that cluster. Ten different models would predict similar values with minor variations, where the average of those ten values is taken as the final value.

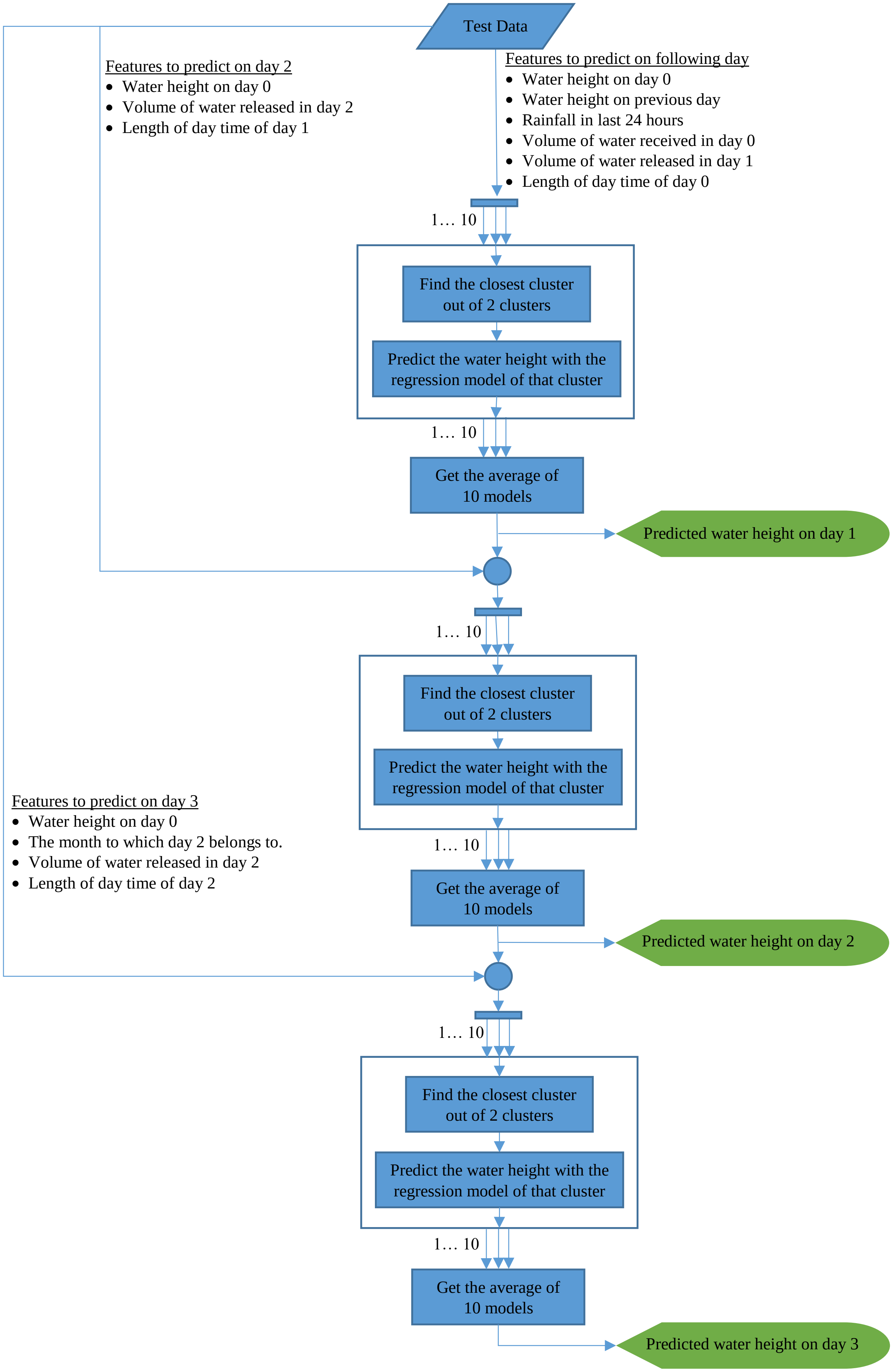

Test Data
Features to predict on day 2
• Water height on day 0
• Volume of water released in day 2
• Length of day time of day 1
Features to predict on following day
• Water height on day 0
• Water height on previous day
• Rainfall in last 24 hours
• Volume of water received in day 0
• Volume of water released in day 1
• Length of day time of day 0
1… 10
Find the closest cluster out of 2 clusters
Predict the water height with the regression model of that cluster
1… 10
Get the average of 10 models
Predicted water height on day 1
1… 10
Find the closest cluster out of 2 clusters
Predict the water height with the regression model of that cluster
Features to predict on day 3
• Water height on day 0
• The month to which day 2 belongs to.
• Volume of water released in day 2
• Length of day time of day 2
1… 10
Get the average of 10 models
Predicted water height on day 2
1… 10
Find the closest cluster out of 2 clusters
Predict the water height with the regression model of that cluster
1… 10
Get the average of 10 models
Predicted water height on day 3

Fig.4.3 Prediction flow of the machine learning model

Subsequently, to predict the water height on day 2, the predicted water height along with other corresponding features, which are shown in the figure 4.3 are fed into a similar model. The predicted output of that model along with another set of features from the original data instance is fed back again to another similar model to predict the water height on day 3.

5 RESULTS ANALYSIS AND DISCUSSION

5.1 Feature selection

The value of a machine learning model, primarily lies on the data used for it. If the data used is unclean, if it contains a significant amount of noise, or if the selected features are not effective, the developed model would not provide accurate predictions. Thus, the data and the features used in developing the model need to be selected carefully. Following are the three main types of data considered in this research.

1. Features derived from manually collected data, such as inflow water volume, released water volume, rainfall…etc.
2. Predicted feature data, which are mainly collected from a web resource (www.darksky.net), where they have provided the predicted weather data for a given geo location, which is mainly predicted from satellite data and based on historical data. Furthermore, when predicting the water height on two or three days following a particular day, the predicted water height on the previous day is also considered as a useful feature.
3. Mathematically calculated features, such as length of the day time, which is calculated as the difference between sunset and sunrise time, the distance to the moon on a given day and the phase of the moon on a given day.

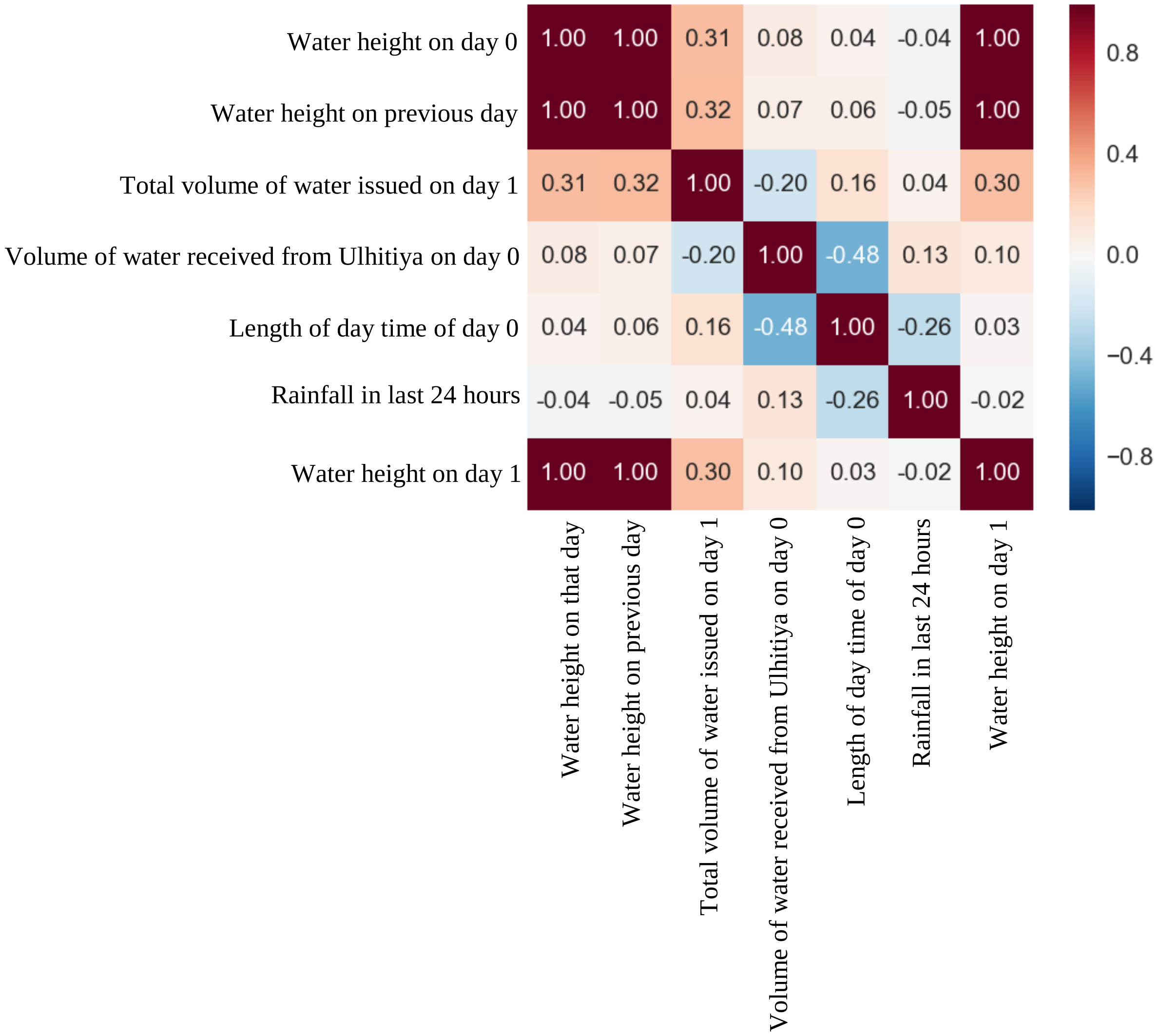
Table 5.1 shows the Pearson correlation coefficient of some of the features considered with the water height on the day following to a given date (day 1). Among the features listed in the table, the following list presents the only six features that had a considerable amount of impact on the regression model to predict the water height on day 1.

1. Water height on day 0
2. Water height on previous day
3. Total volume of water issued in day 1
4. Volume of water received from Ulhitiya in day 0
5. Length of day time of day 0
6. Rainfall on day 0

Feature	Pearson correlation coefficient with water height on day 1
Water height on the day	1.00
Water height on previous day	1.00
Total volume of water issued on day 1	0.30
Total volume of water issued in past 24hours	0.29
Spill flow volume	0.20
Volume of water issued from left bank sluice	0.19
Volume of water received from Ulhitiya	0.09
Volume of water issued from right bank sluice	0.06
Length of day time	0.04
Predicted maximum temperature	0.02
Year	0.01
Rainfall in last 24 hours	-0.02
Predicted maximum wind on the day	-0.06
Predicted cloud cover on the day	-0.11
Day of the year	-0.65
Month of the current date	-0.65
Month of day 1	-0.66

Table 5.1 Pearson correlation coefficient of different features with the height of water on day 1

Moreover, there were several features, which were seeming to be reducing the RMSE, but tending to overfit the model. Figure 5.1 shows a diagram of Pearson correlation coefficients of those six selected features. As per the figure 5.1, the highest impacting feature is the previous water heights, which has a temporal effect of past two days. Even though it seems features such as cloud cover, temperature and wind speed may cause an impact on the water height as they are factors of evaporation and precipitation, practically these values did not cause a considerable amount of impact, which may be because these values are predicted values and their accuracies may not be up to the required level. The web resource, which provides predicted weather information, predicts the weather information for certain geolocations, where they have a considerable amount of past information, and they interpolate it to other geolocations, which may not be correct always. Moreover, features such as ‘day of the year’, ‘month of the year’ seems to be having a higher correlation, but practically



	Water height on that day	Water height on previous day	Total volume of water issued on day 1	Volume of water received from Ulhitiya on day 0	Length of day time of day 0	Rainfall in last 24 hours	Water height on day 1
Water height on day 0	1.00	1.00	0.31	0.08	0.04	-0.04	1.00
Water height on previous day	1.00	1.00	0.32	0.07	0.06	-0.05	1.00
Total volume of water issued on day 1	0.31	0.32	1.00	-0.20	0.16	0.04	0.30
Volume of water received from Ulhitiya on day 0	0.08	0.07	-0.20	1.00	-0.48	0.13	0.10
Length of day time of day 0	0.04	0.06	0.16	-0.48	1.00	-0.26	0.03
Rainfall in last 24 hours	-0.04	-0.05	0.04	0.13	-0.26	1.00	-0.02
Water height on day 1	1.00	1.00	0.30	0.10	0.03	-0.02	1.00

Fig 5.1 Correlation coefficients of the features to predict the water height on day 1

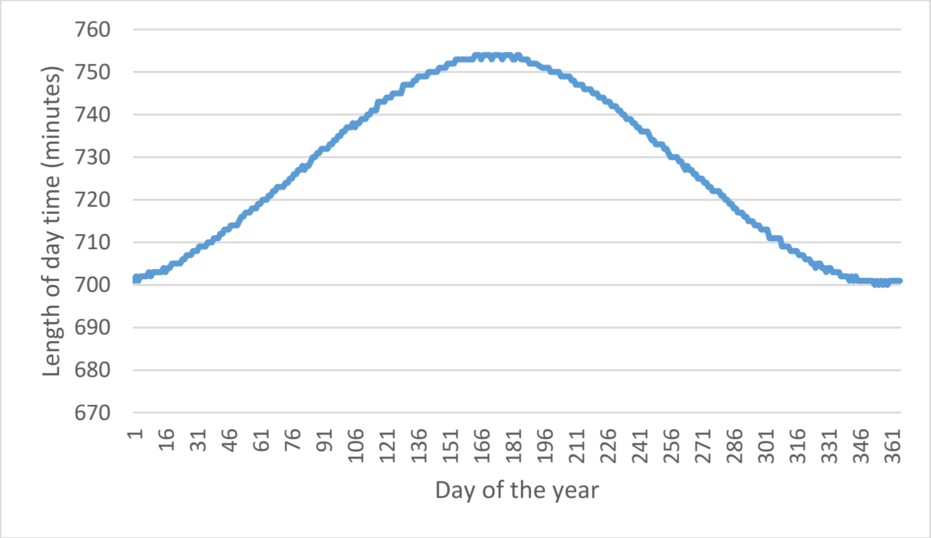


Fig 5.2 Variation of length of day time within a year

those are mimicked by the ‘length of the day time’ feature, which is a sinusoidal repetitive function and has a higher impact than the ‘day of the year’ feature, which is shown in figure 5.2.

Unlike the height of the water, the rainfall has a temporal effect of only one day. Even though usually it is assumed that the rainfall would increase the water height of the reservoir, it has a negative correlation with the water height on the following day. This may be because, whenever there is a considerable amount of water through the rainfall, the spill flow gates are opened. It can be confirmed by the rainfall on day 0 having a positive correlation of 0.21 with the volume of water issued through spill flow on day 1. In this system, the total volume of water issued, which is the summation of volume of water issued through left bank sluice, right bank sluice and spill flow gate is considered instead of considering them as individual features. Summing them to one feature does not have a considerable amount of impact on the accuracy of the model, but improves the usability of the model by reducing the number of inputs the user should enter.

Most of the features considered to predict the water height on second consecutive day (day 2) are shown in table 5.2 with their correlation coefficient values. Among those features, following features had the highest impact on the water height on day 2.

1. Water height on day 0
2. Predicted water height on day 1
3. Total volume of water issued on day 2
4. Length of day time of day 1

Correlation coefficients among these features and between the feature and the water height on day 2 is presented in figure 5.3. Similar to the prediction of water height on day 1, previous water heights have a temporal effect of two days, irrespective of one value being a predicted value. This may accumulate the prediction error introduced by the prediction model of day 1. Therefore, the model should have a lower RMSE value.

Feature	Pearson correlation coefficient with water height on day 2
Predicted water height on day 1	1.00
Water height on the day	1.00
Water height on previous day	1.00
Total volume of water issued on day 2	0.30
Total volume of water issued on day 1	0.29
Total volume of water issued on the day	0.29
Spill flow volume on day 1	0.20
Spill flow volume on the day	0.20
Spill flow volume on day 2	0.20
Volume of water issued from left bank sluice	0.18
Volume of water received from Ulhitiya	0.11
Volume of water issued from right bank sluice	0.05
Length of day time	0.03
Rainfall on the day	-0.01
Predicted cloud cover on the day	-0.10
Predicted cover on the day 1	-0.11
Predicted cover on the day 2	-0.12
Day of the year of day 2	-0.68
Month of day 2	-0.68

Table 5.2 Pearson correlation coefficient of different features with the height of water on day 2

Predicting the water height on the third consecutive day (day 3) is harder than other two days, since the rainfall and water inflow from Ulhitiya in between two days is unknown. This reduces the prediction accuracy of day 3. Various features are analyzed to mitigate the error caused by this uncertainty. Table 5.3 lists most of those features, along with the correlation coefficient of those features with the water height

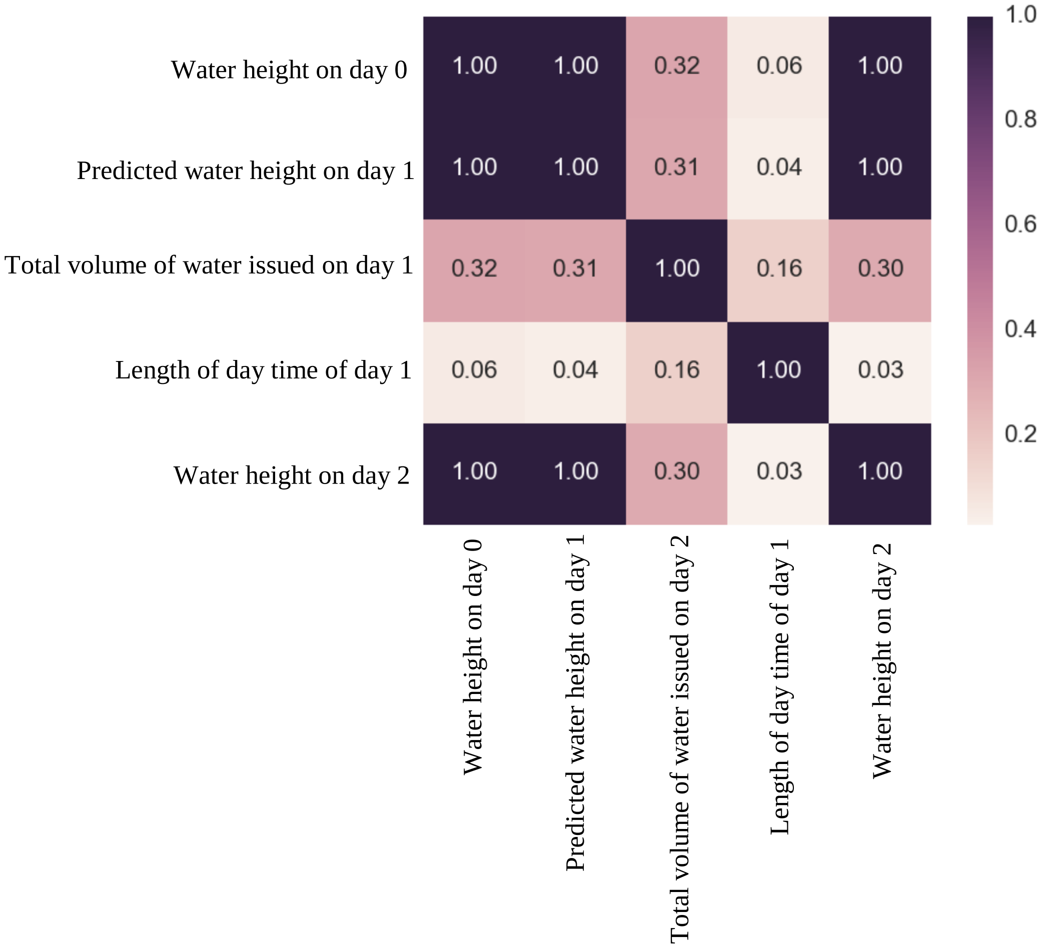


Fig 5.3 Correlation coefficients of the features to predict the water height on day 2

on day 3. Among these features, it was observed that the following list of features causes an impact on the prediction model.

1. Water height on day 0
2. Predicted water height on the day 2
3. The month to which day 3 belongs to.
4. Total volume of water issued on day 3
5. Length of day time of day 2

Previous two models had an impact from the rainfall, whereas this model doesn’t have an impact from the rainfall, because as per the observations from previous models, the rainfall has a temporal effect of 1-2 days. Additionally, the month of the year, to which the day belongs to is considered as a feature, which has a relationship with the rainfall in tropical countries like Sri Lanka, where mainly the rainfall occurs

Feature	Pearson correlation coefficient with water height on day 3
Predicted water height on day 2	1.00
Predicted water height on day 1	1.00
Water height on the day	1.00
Water height on previous day	1.00
Total volume of water issued on day 3	0.30
Total volume of water issued on day 2	0.29
Total volume of water issued on day 1	0.28
Total volume of water issued on the day	0.28
Spill flow volume on day 2	0.20
Spill flow volume on day 1	0.20
Spill flow volume on day 3	0.20
Spill flow volume on the day	0.20
Volume of water issued from left bank sluice	0.17
Volume of water received from Ulhitiya	0.12
Volume of water issued from right bank sluice	0.05
Length of day time of day 3	0.05
Length of day time of day 2	0.03
Length of day time of day 1	0.02
Length of day time of the day	0.00
Rainfall on the day	0.00
Predicted cover on the day 2	-0.11
Predicted cover on the day 3	-0.12
Day of the year of day 3	-0.68
Month of day 3	-0.68

Table 5.3 Pearson correlation coefficient of different features with the height of water on day 3

in certain annual monsoon periods. Correlation coefficients among these selected features and between each feature & the water height on day 3 is illustrated in figure 5.4.

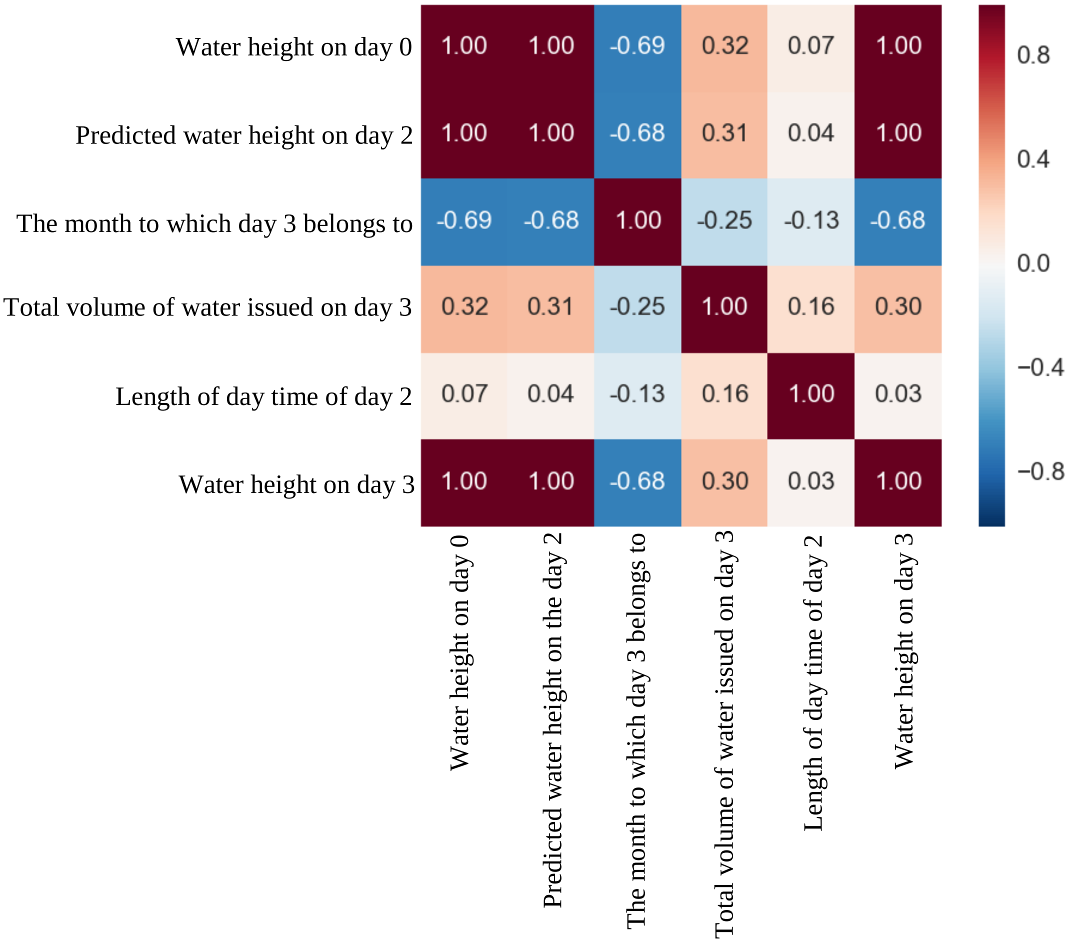


Fig 5.4 Correlation coefficients of the features to predict the water height on day 3

5.2 Clustering

As per the previous researches [23], clustering the data points prior to training the regression model should further reduce the prediction error. K-Means and K-Medoids clustering algorithms were taken into consideration and the results were analyzed. Figure 5.5, illustrates the variation of the average RMSE value of linear regression models to predict the water height on day 1, 2 and 3, where the accuracy is

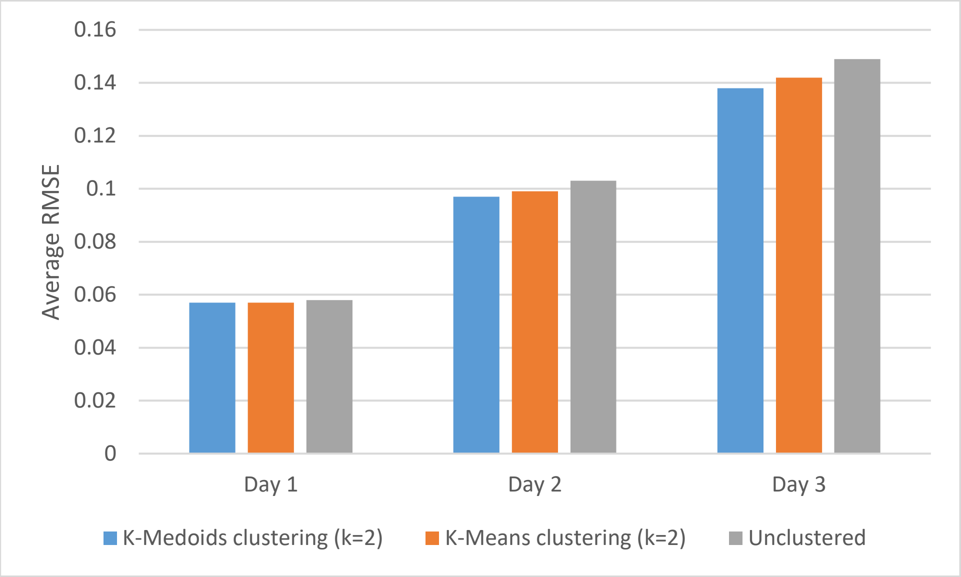


Fig 5.5 Variation of average RMSE value of linear regression model, with various clustering techniques

measured on cross validation data set. The graph presented in figure 5.5 shows that, clustering the data set with K-Medoids clustering and training separate regression models for different clusters, has reduced the average RMSE value than clustering with K-Means clustering and training a single regression model without clustering the data, which further validates previous researches[23]. Clustering the data points simplifies the problem by clustering similar data points into one cluster, by which it reduces the complexity of the regression model. K-Medoid clustering algorithm is more processor intensive than the K-Means clustering algorithm, but the cluster centers of K-Medoid clustering are real data points, while the point representing a cluster in K-Means clustering can be an unrealistic virtual point.

5.2.1 Number of clusters

The RMSE value of the models vary with the number of clusters. To identify the optimum number of clusters, K-Medoids clustering algorithm was experimented with K value 1 to 6, to cluster the water height predicting features of day 1, 2 and 3. The clustered data were then trained with separate linear and multilayer perceptron (MLP) regression models per cluster to predict the water height on day 1, 2 & 3. The variation of RMSE of linear regression models, to predict the water height on day 1 to 3 with the number of clusters are presented in the table 5.4.

Number of clusters	Day 1		Day 2		Day 3	
	Cross validation data set	Test data set	Cross validation data set	Test data set	Cross validation data set	Test data set
1	0.058	0.055	0.103	0.1	0.149	0.162
2	0.057	0.05	0.097	0.083	0.138	0.122
3	0.058	0.051	0.097	0.086	0.137	0.127
4	0.057	0.052	0.095	0.086	0.135	0.134
5	0.057	0.051	0.093	0.089	0.132	0.132
6	0.058	0.051	0.092	0.089	0.129	0.137

Table 5.4 Variation of RMSE value of predicted water height from linear regression model, with the number of clusters

Figure 5.6 and 5.7 present the variation of RMSE of predicted water height of day 1 for cross validation data set and test data set respectively. From figure 5.6, it can be observed that it has given low RMSE values, when the number of clusters is 2, 4 and 5. But, from the figure 5.7, it can be observed that with the higher number of clusters, the number of training samples per regression model reduces, which makes the model to overfit to training data. Moreover, higher number of clusters require to train more number of regression models, which costs time and processing power. Based on the observations, it can be concluded that the ideal number of clusters to predict the water height with linear regression model on day 1 is 2.

Graphs illustrated in figure 5.8 and figure 5.9 present the variation of RMSE of linear regression model, to predict the water height on day 2 against the number of clusters for cross validation data set and test data set respectively. As per the observations, the average RMSE on cross validation data set has continuously decreased with the number of clusters. But again, it has tended to overfit with higher number of clusters, in test data set. As per the observations, dividing the data set to 2 clusters would reduce the average RMSE error of cross validation set to 0.097, while having a relatively lower RMSE in test data set.

The variation of RMSEs of linear regression models to predict the water height on day 3 against the number of clusters is presented in figure 5.10 and figure 5.11, for cross validation data set and test data set respectively. Similar to the observation of day 2, the average RMSE of cross validation data set has dropped with the number of clusters, while the RMSE value has increased in the test data set. As per the graphs in figure 5.10 and figure 5.11, for the linear regression model to predict the water height on day 3, the optimum number of clusters can be considered as 2, since it has an average value for RMSE on cross validation data set, which is 0.138 and lowest RMSE value on test data set, which is 0.122.

Similar to the variation of RMSE of linear regression models to predict the water height on day 1 to 3, variations of RMSE values of MLP regression models to predict the water height on day 1 to 3 with the number of clusters is summarized in table 5.5 and illustrated in figure 5.12 to figure 5.17. From the figure 5.12 and figure 5.13, it can be observed that k=2 is the optimum number for K-Medoids clustering, when the MLP regression model is used to predict the water height on day 1. The MLP regression model was configured with a linear activation function for hidden layer, which consisted of one hidden layer with 3 perceptrons. The MLP regression model was fine-tuned with the cross-validation data set, which is explained in chapter 5.3.

Number of clusters	Day 1		Day 2		Day 3	
	Cross validation data set	Test data set	Cross validation data set	Test data set	Cross validation data set	Test data set
1	0.058	0.055	0.103	0.100	0.149	0.162
2	0.057	0.050	0.096	0.083	0.138	0.122
3	0.058	0.051	0.097	0.087	0.137	0.128
4	0.058	0.052	0.095	0.086	0.134	0.134
5	0.058	0.051	0.094	0.089	0.132	0.133
6	0.056	0.051	0.092	0.089	0.130	0.140

Table 5.5 Variation of RMSE value of predicted water height from MLP regression model, with the number of clusters

Figure 5.14 and figure 5.15 present the variation of RMSE values of MLP regression models to predict the water height on day 2 with the number of clusters. According to the graphs, configuring to cluster the data to 2 clusters would be the optimum number, which does not overfit the regression models and reduces the RMSE values. The MLP regression model developed to predict the water height on day 2, also consists of a linear activation function and a hidden layer with 8 perceptrons.

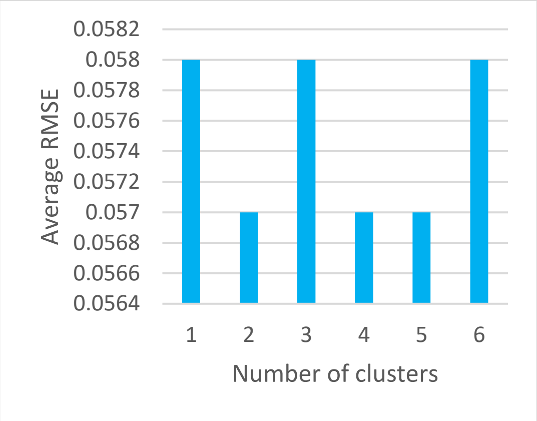


Fig 5.6 Variation of average RMSE of predicted water height of linear regression model on day 1 with number of clusters on cross validation data set

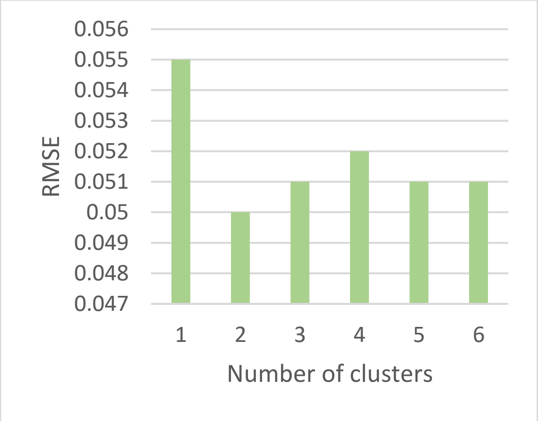


Fig 5.7 Variation of RMSE of predicted water height of linear regression model on day 1 with number of clusters on test data set

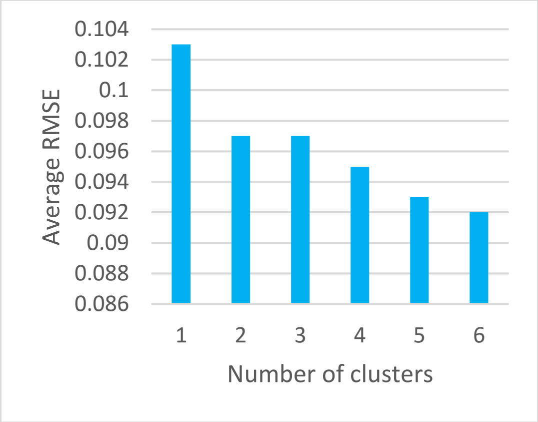


Fig 5.8 Variation of average RMSE of predicted water height of linear regression model on day 2 with number of clusters on cross validation data set

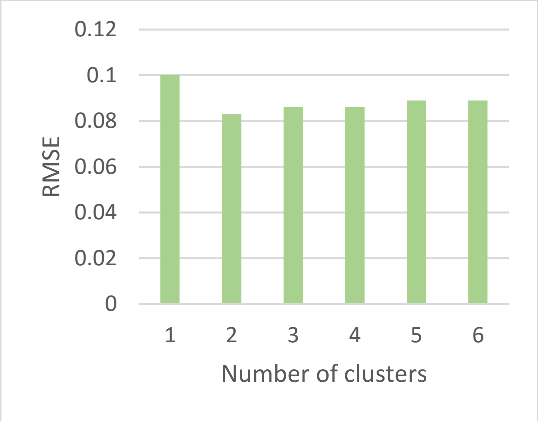


Fig 5.9 Variation of RMSE of predicted water height of linear regression model on day 2 with number of clusters on test data set

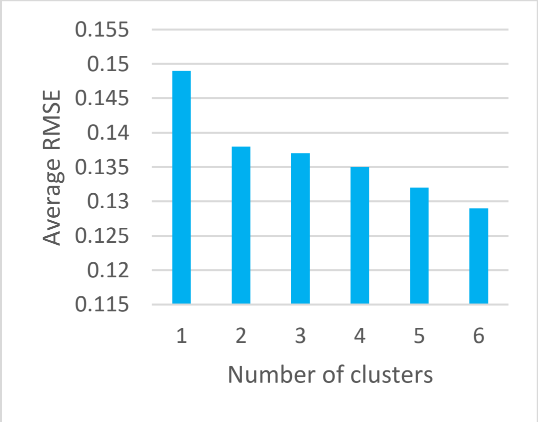


Fig 5.10 Variation of average RMSE of predicted water height of linear regression model on day 3 with number of clusters on cross validation data set

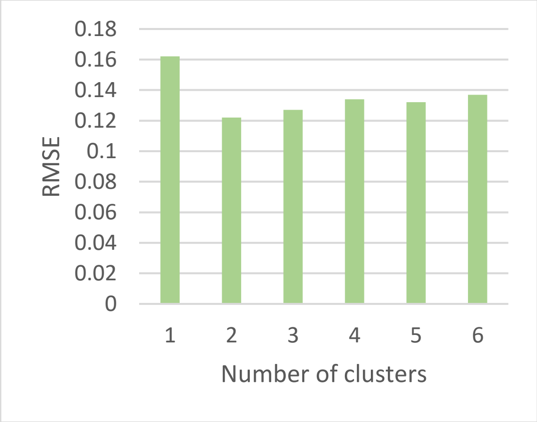


Fig 5.11 Variation of RMSE of predicted water height of linear regression model on day 3 with number of clusters on test data set

Figure 5.16 and figure 5.17 present the variation of RMSE values of MLP regression models, which were trained to predict the water height on day 3 with the number of clusters. Similar to day 1 and 2, this multilayer perceptron model was also trained with a linear activation function and a hidden layer. The hidden layer consisted of 5 perceptrons. As per the observations, the optimum number of clusters is 2, where the cross-validation data set has an average RMSE of 0.138 and test data set has an RMSE of 0.122.

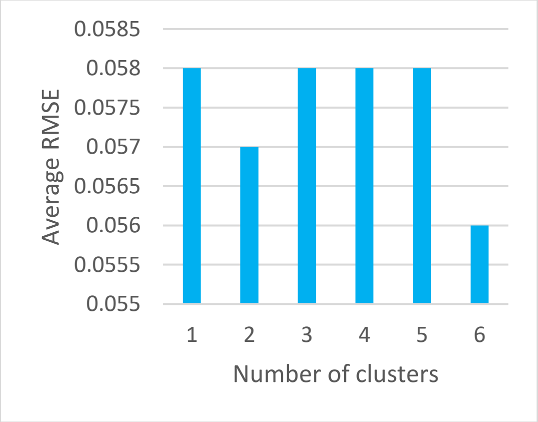


Fig 5.12 Variation of average RMSE of predicted water height of MLP regression model on day 1 with number of clusters on cross validation data set

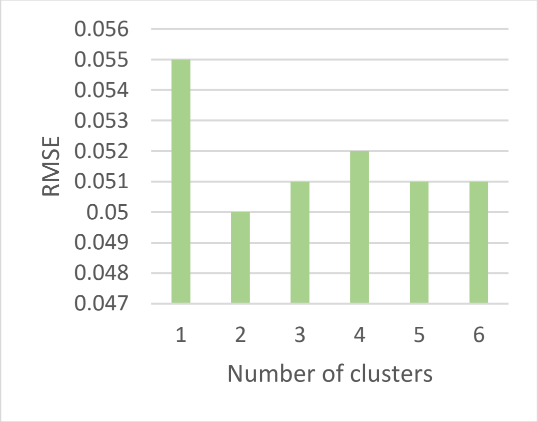


Fig 5.13 Variation of RMSE of predicted water height of MLP regression model on day 1 with number of clusters on test data set

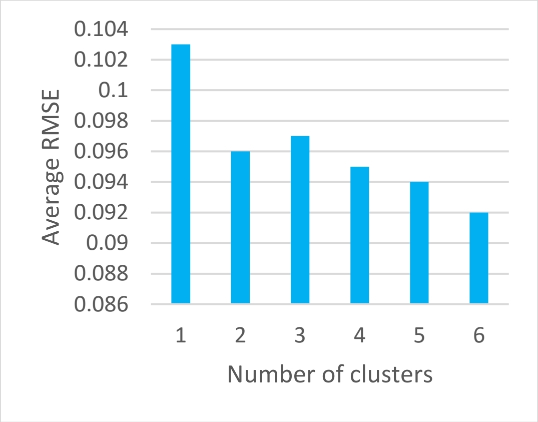


Fig 5.14 Variation of average RMSE of predicted water height of MLP regression model on day 2 with number of clusters on cross validation data set

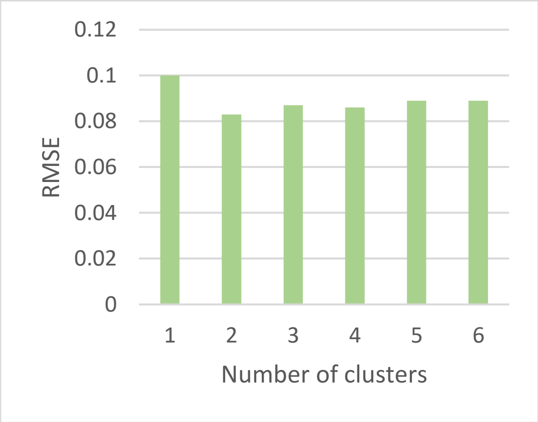


Fig 5.15 Variation of RMSE of predicted water height of MLP regression model on day 2 with number of clusters on test data set

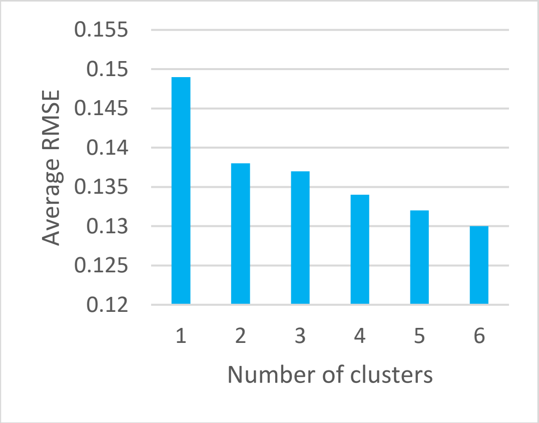


Fig 5.16 Variation of average RMSE of predicted water height of MLP regression model on day 3 with number of clusters on cross validation data set

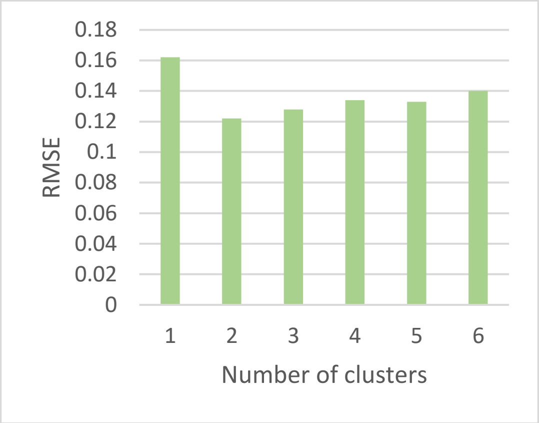


Fig 5.17 Variation of RMSE of predicted water height of MLP regression model on day 3 with number of clusters on test data set

5.3 Regression model

The training data set is shuffled and split in to 10 partitions to conduct the 10-fold cross validation, where at each fold, the training set is clustered to the configured number of clusters. In each cluster, a separate regression model is trained and water heights are predicted for the cross-validation data set, based on the cluster which the data point is assigned. But for a test data instance, to predict the water height, it is passed through the 10 different models, generated from the 10-fold cross validation, where in each model, it is assigned to the closest cluster and the water height is predicted. The average value of those 10 values is considered as the final value.

Among the available various regression models, the linear regression model and the multilayer perceptron regression model are trained with the data set.

5.3.1 Linear regression model

Table 5.6 presents the results, when the linear regression model is applied on the entire training data set with 10-fold cross-validation, without applying any clustering algorithm. It was observed that, this approach has given RMSE values of 0.055, 0.100 and 0.162 for the forecasted water height on next three days respectively on the test data set with the features mentioned in the chapter 5.1. But, as per the results of chapter 5.2, RMSE values should further reduce when data instances are clustered in to 2 clusters, which is verified by table 5.7. As per the table 5.7, the prediction models have predicted water heights very close to actual water heights on all three days, where those have obtained RMSE values of 0.50, 0.83 and 0.122 for predicting water heights on next three days respectively on the test data set. Moreover, it can be observed that the prediction error for the prediction model of day 1 is accumulated to the prediction model of day 2 and the error of the model of day 2 to the model of day 3.

	Day 1	Day 2	Day 3
Average RMSE in cross-validation data set	0.058	0.103	0.149
RMSE in test data set	0.055	0.100	0.162
R^2 value in cross-validation data set	1.000	0.999	0.998
R^2 value in test data set	1.000	0.999	0.997
Average correlation coefficient in cross-validation data set	1.000	1.000	0.999
Correlation coefficient in test data set	1.000	0.999	0.999
Maximum absolute error percentage in cross-validation data set	1.22%	1.96%	2.79%
Maximum absolute error percentage in test data set	0.73%	0.82%	1.48%

Table 5.6 Performance of the linear regression model on cross-validation and test data sets

	Day 1	Day 2	Day 3
Average RMSE in cross-validation data set	0.057	0.096	0.138
RMSE in test data set	0.050	0.083	0.122
R^2 value in cross-validation data set	1.000	0.999	0.998
R^2 value in test data set	1.000	0.999	0.998
Average correlation coefficient in cross-validation data set	1.000	1.000	0.999
Correlation coefficient in test data set	1.000	1.000	0.999
Maximum absolute error percentage in cross-validation data set	1.26%	1.97%	2.71%
Maximum absolute error percentage in test data set	0.69%	0.77%	1.05%

Table 5.7 Performance of the linear regression model on cross-validation and test data sets, when data sets are clustered

Figure 5.18 shows the predicted water height and the actual water height on day 1, where it can be observed that the prediction model has performed very well on the cross-validation data set by mimicking the actual water height as it is, except for a very few outlays. Figure 5.19 presents the variation of the actual water height and the predicted water height on day 1 on the test data set, which shows a better resemblance between the predicted water height and the actual water height. The actual water height and the predicted water height on day 2 is graphed in figure 5.20 and figure 5.21 for cross-validation and test data sets respectively. Even though the predicted water height shows a slight variation from the actual value in figure 5.20 and figure 5.21 compared to figure 5.18 and figure 5.19, it closely resembles the actual water height in both the graphs. Figure 5.22 and figure 5.23 illustrate the variation of the actual water height and the predicted water height on day 3 for cross-validation data set and test data set respectively. The predicted water height graphs have closely mimicked the pattern of the actual water height, but they have slight outlays with the actual graphs.

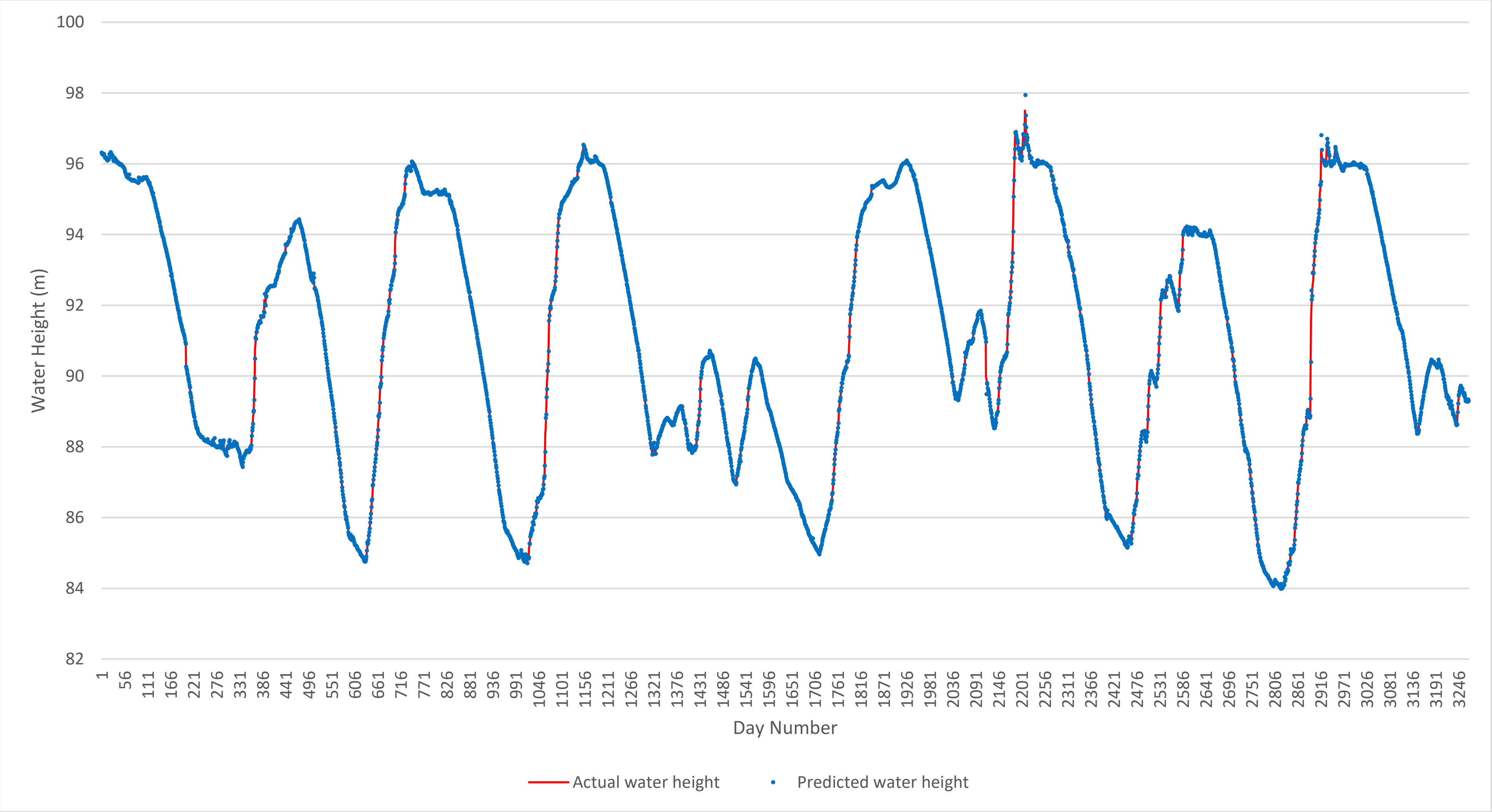


Fig 5.18 Variation of actual & predicted water height of cross-validation data set on day 1, with linear regression model

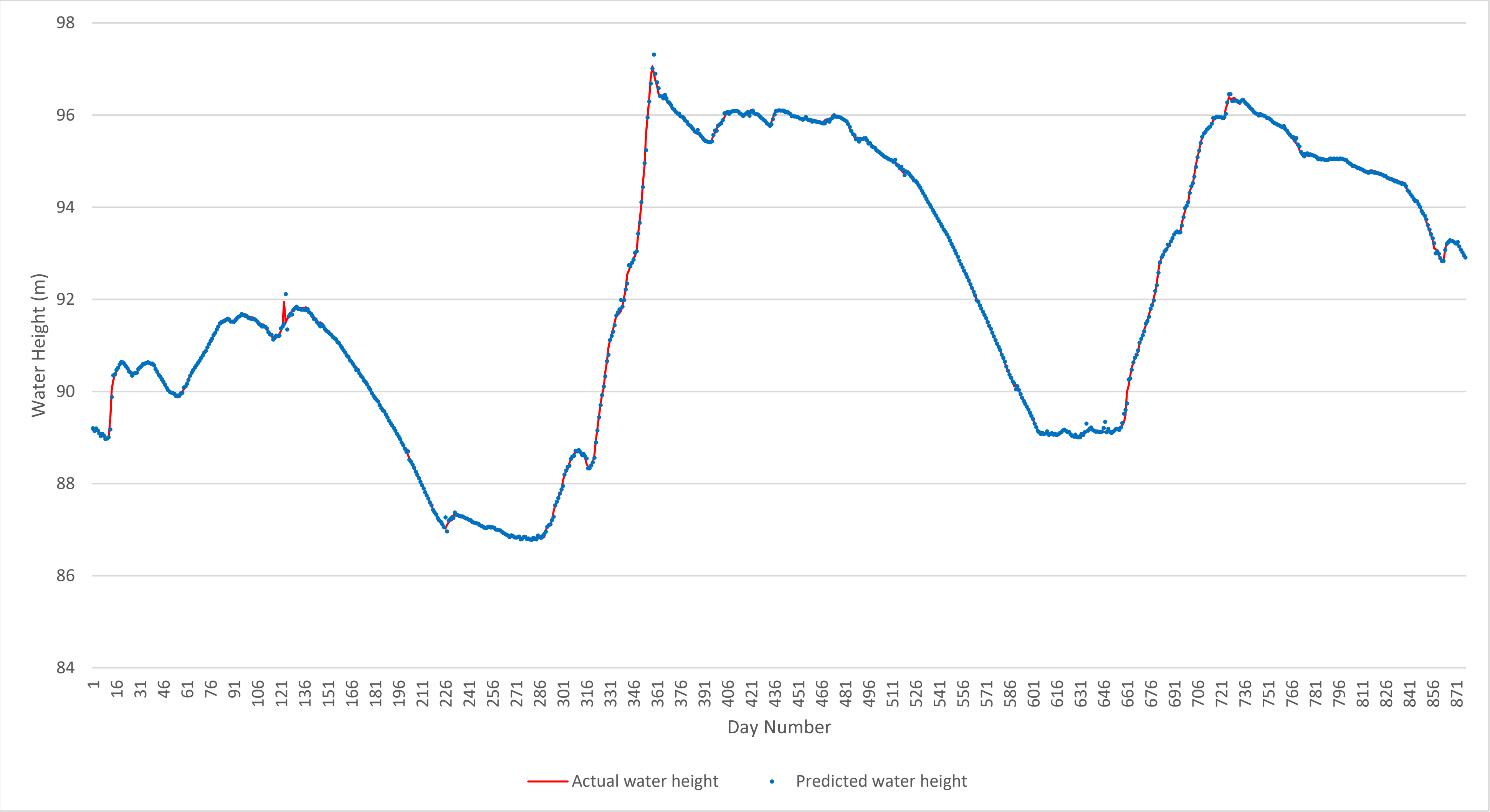


Fig 5.19 Variation of actual & predicted water height of test data set on day 1, with linear regression model

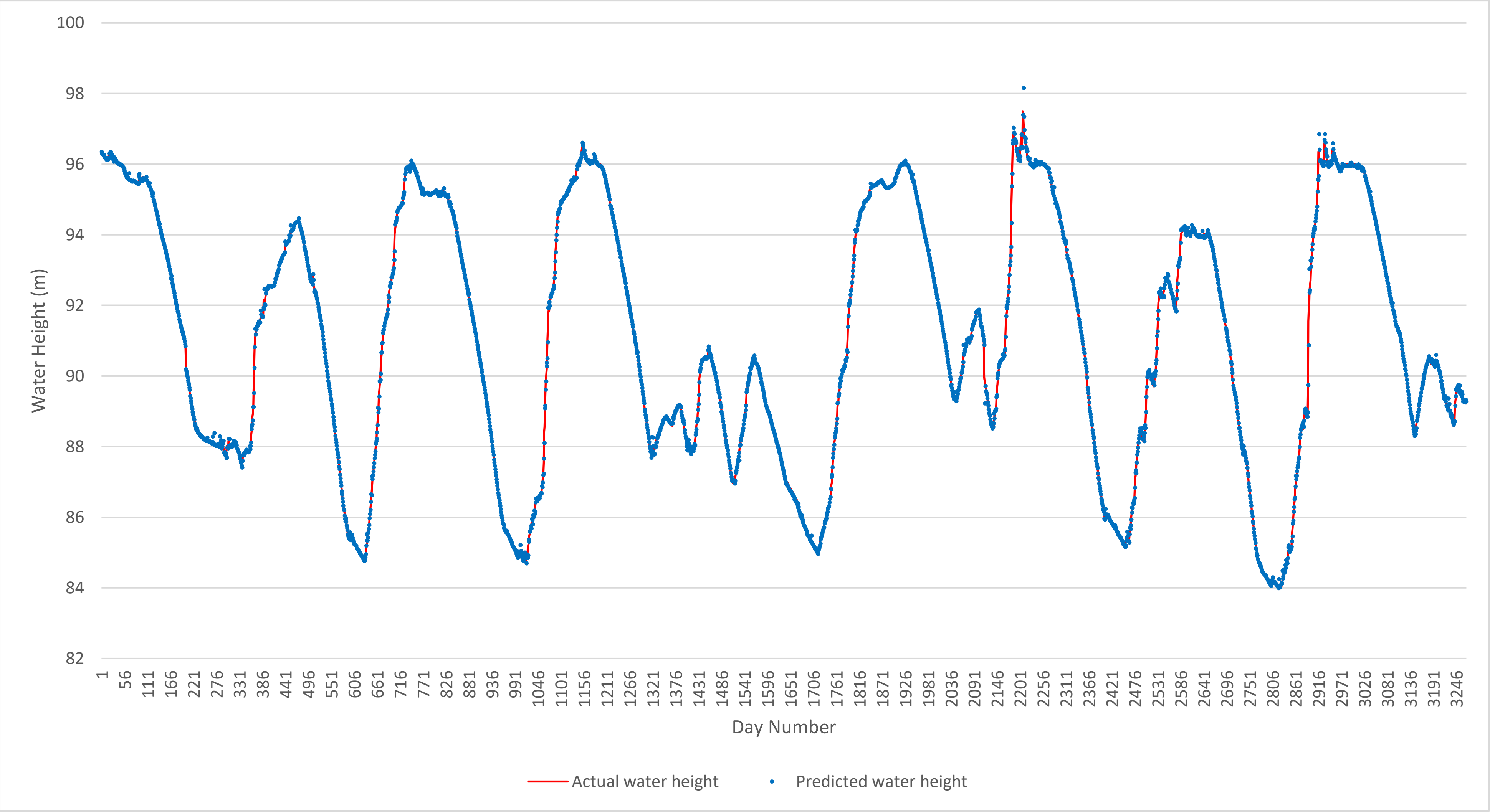


Fig 5.20 Variation of actual & predicted water height of cross-validation data set on day 2, with linear regression model

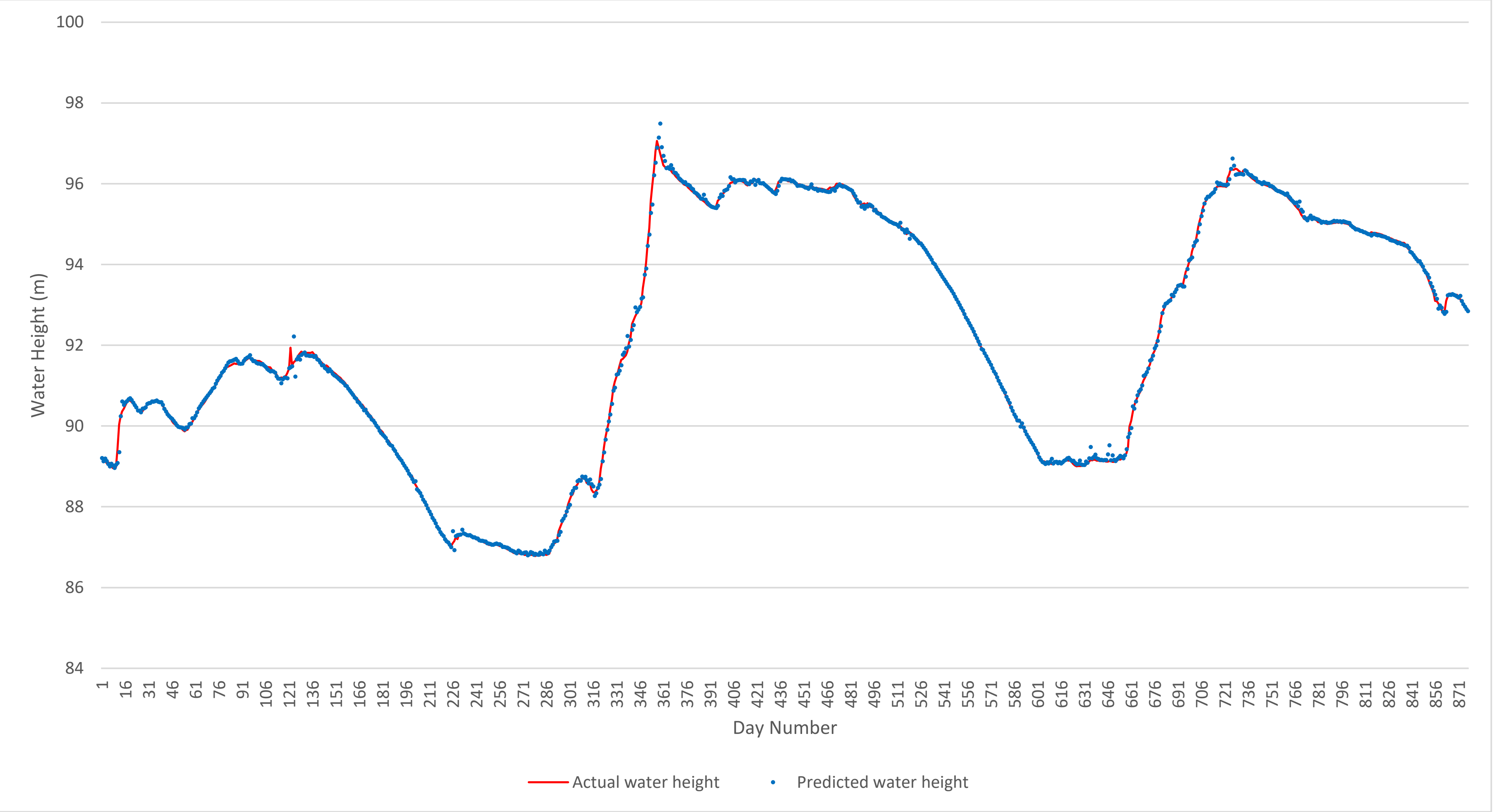


Fig 5.21 Variation of actual & predicted water height of test data set on day 2, with linear regression model

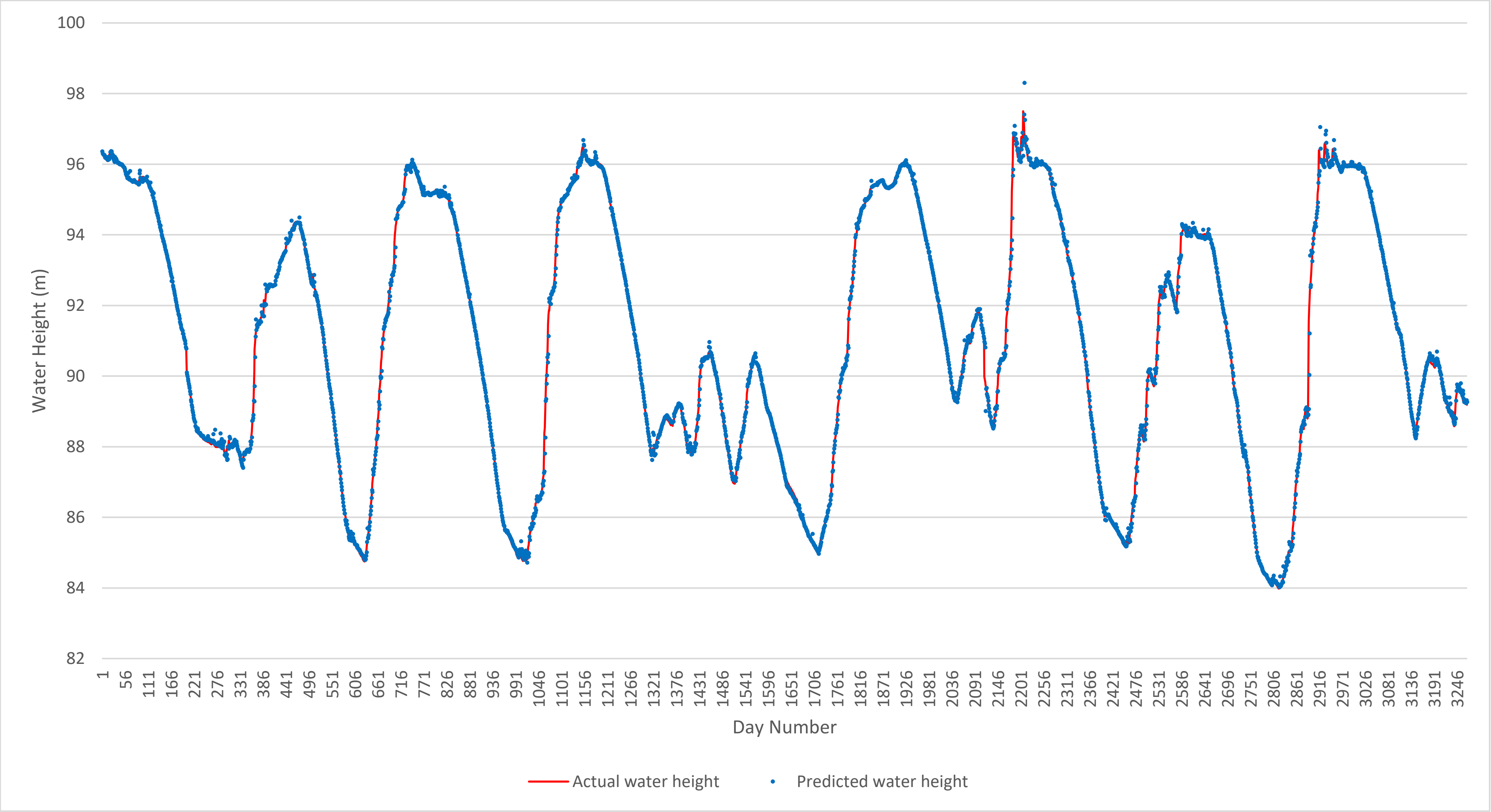


Fig 5.22 Variation of actual & predicted water height of cross-validation data set on day 3, with linear regression model

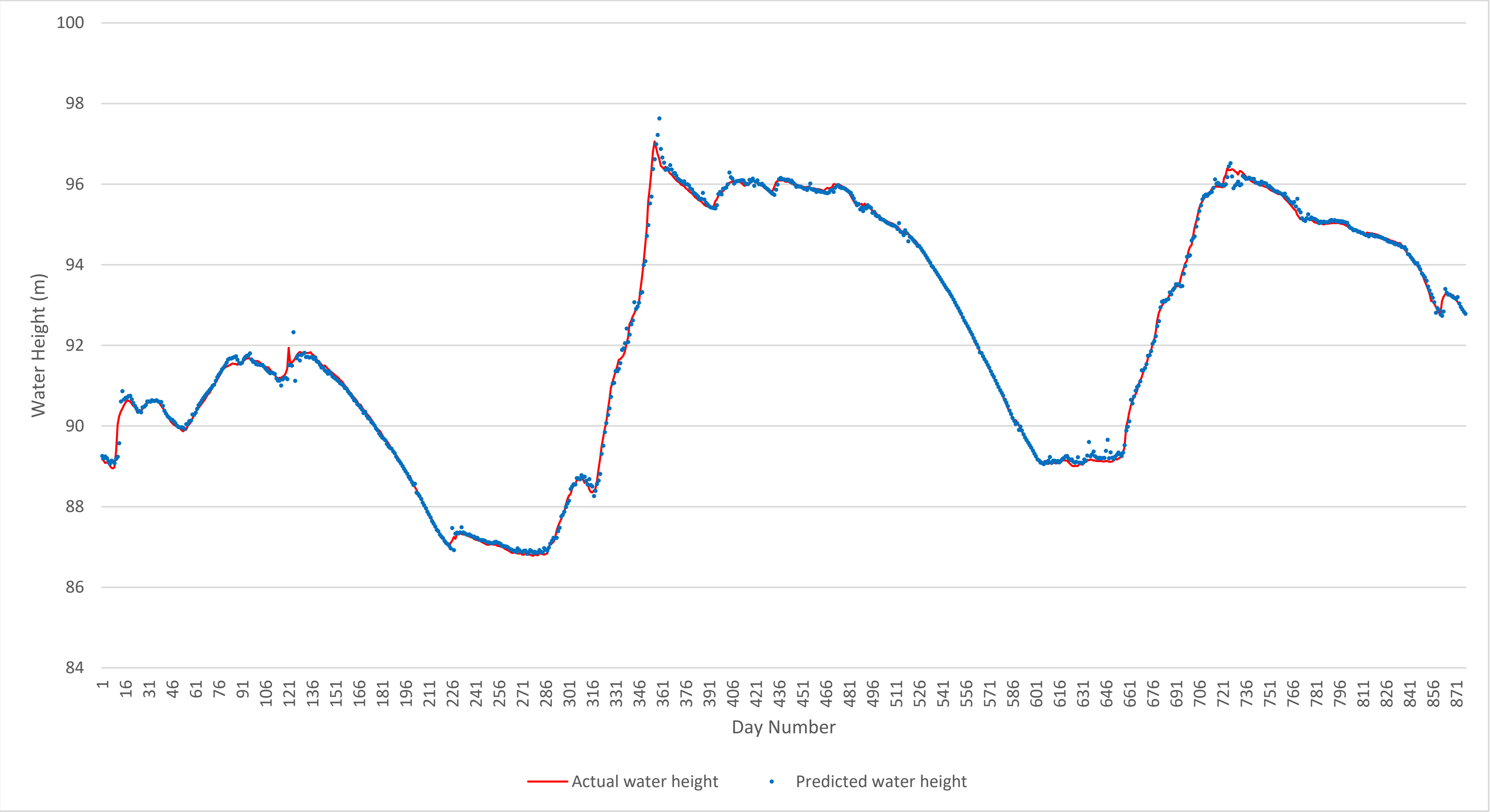


Fig 5.23 Variation of actual & predicted water height of test data set on day 3, with linear regression model

5.3.2 MLP regression model

In addition to the linear regression model, multilayer perceptron regression model (feed forward ANN) s are trained to predict the water height on day 1 to 3, where table 5.8 shows the performance of the MLP regression model, when the training data set is directly used to train a model with 10-fold cross validation without clustering the data. In this approach, it was observed that the values are very much similar to the values obtained from linear regression in table 5.6. According to chapter 5.2, it can be concluded that the regression model should become much simpler and perform better, if the training data set is clustered in to 2 clusters with K-Medoids clustering. The results of the models on prediction of water heights on next three days, when the training data is clustered prior to training the MLP regression model is shown in table 5.9. According to the table 5.9, it can be clearly observed that RMSE values and maximum absolute error percentages of test data set has decreased compared to table 5.8. Furthermore, it can be seen that the table 5.9 is almost similar to the table 5.7 of linear regression model. Moreover, it can be observed that the MLP regression model has a lesser maximum absolute error percentage in the test data set.

	Day 1	Day 2	Day 3
Average RMSE in cross-validation data set	0.058	0.103	0.149
RMSE in test data set	0.055	0.100	0.162
R^2 value in cross-validation data set	1.000	0.999	0.998
R^2 value in test data set	1.000	0.999	0.997
Average correlation coefficient in cross-validation data set	1.000	1.000	0.999
Correlation coefficient in test data set	1.000	0.999	0.999
Maximum absolute error percentage in cross-validation data set	1.22%	1.96%	2.79%
Maximum absolute error percentage in test data set	0.73%	0.82%	1.48%

Table 5.8 Performance of the MLP regression model on cross-validation and test data sets

The variation of corresponding predicted water heights against the actual water height is plotted in figure 5.24 and 5.25, which almost overlaps the identity line. Figure 5.26 and 5.27 illustrate the variation of predicted water heights on day 2 against their actual water heights for cross-validation and test data sets respectively. Even though they are slightly scattered compared to figure 5.24 and 5.25, they also closely resemble the identity line. Figure 5.28 and 5.29 present the variation of predicted heights against the actual height on day 3 for cross-validation and test data sets respectively, where the scatter plots resemble the identity line, with some amount of scattering.

	Day 1	Day 2	Day 3
Average RMSE in cross-validation data set	0.057	0.096	0.137
RMSE in test data set	0.050	0.083	0.122
R^2 value in cross-validation data set	1.000	0.999	0.998
R^2 value in test data set	1.000	0.999	0.999
Average correlation coefficient in cross-validation data set	1.000	1.000	0.999
Correlation coefficient in test data set	1.000	1.000	0.999
Maximum absolute error percentage in cross-validation data set	1.26%	1.97%	2.71%
Maximum absolute error percentage in test data set	0.69%	0.73%	0.93%

Table 5.9 Performance of the MLP regression model on cross-validation and test data sets, when data sets are clustered

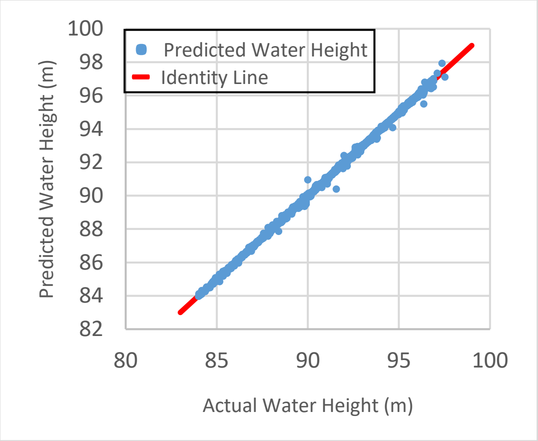


Fig 5.24 Variation of predicted water height against actual water height on day 1 of cross-validation data set with MLP regression model

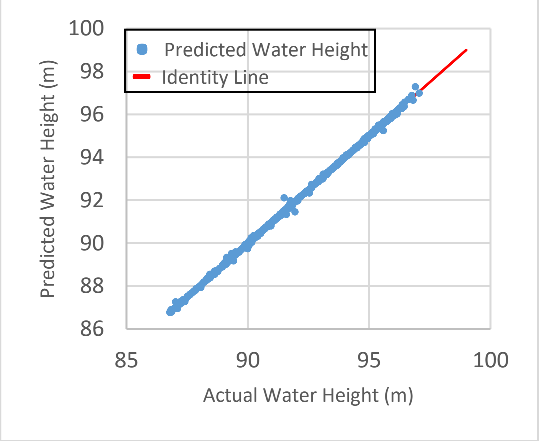


Fig 5.25 Variation of predicted water height against actual water height on day 1 of test data set with MLP regression model

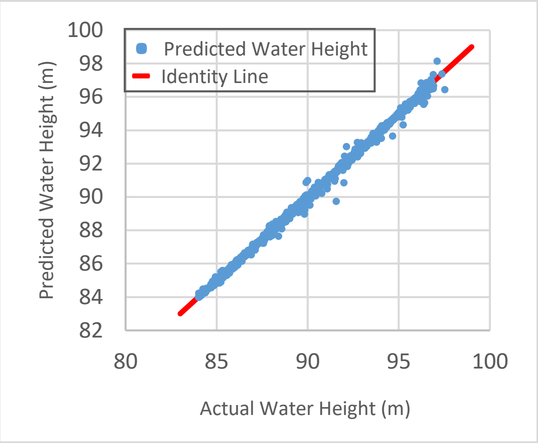


Fig 5.26 Variation of predicted water height against actual water height on day 2 of cross-validation data set with MLP regression model

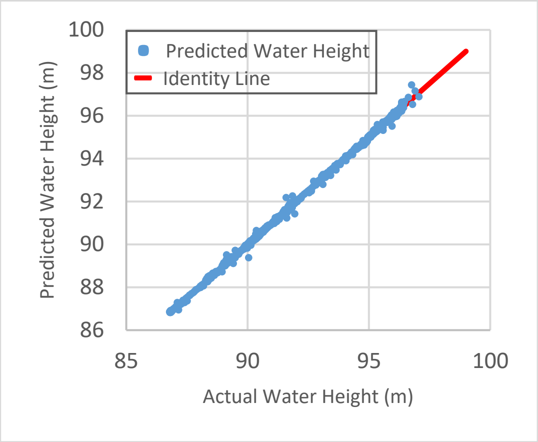


Fig 5.27 Variation of predicted water height against actual water height on day 2 of test data set with MLP regression model

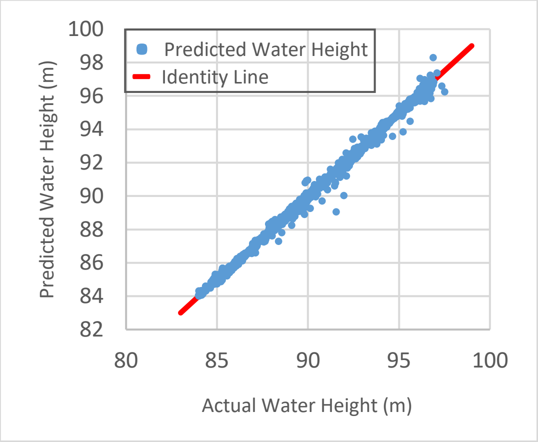


Fig 5.28 Variation of predicted water height against actual water height on day 3 of cross-validation data set with MLP regression model

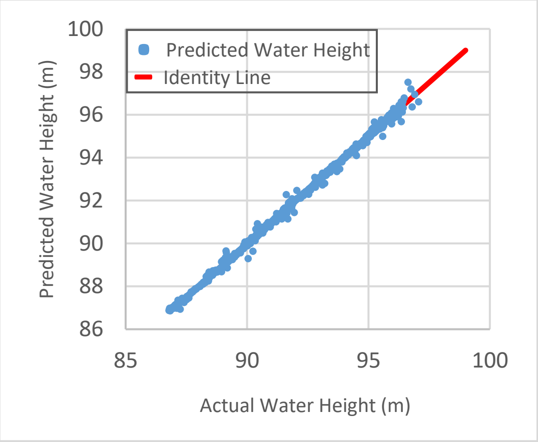


Fig 5.29 Variation of predicted water height against actual water height on day 3 of test data set with MLP regression model

Figure 5.30 and 5.31 present the variation of the predicted water height and the actual water height on day 1 of cross-validation and test data sets respectively for the MLP regression model. It can be observed that the variation of the graphs are similar to the graphs in 5.18 and 5.19 respectively. Figure 5.32 and 5.33 illustrate the variation of the predicted water height and actual water height on day 2, with the day for cross-validation and test data sets respectively, which are very much similar to the respective graphs of the linear regression model. The variation of the predicted water height and the actual water height on day 3 with the day is plotted in figure 5.34 and 5.35 for cross-validation and test data sets respectively. Even though the predicted graphs have some deviation compared to figure 5.30 and 5.31, which is on day 1, in both the graphs it can be observed that the predicted graph closely overlaps the actual graph.

All the MLP regression models consisted of a linear activation function and 3, 8 and 5 perceptrons in the hidden layer for day 1, 2 and 3 respectively. In this approach, the regularization parameter was set to 0.0001. Since the data set is already clustered with K-Medoids clustering, a simple model with one hidden layer tends to work better, whereas adding more hidden layers, reduces the average RMSE value of the cross-validation data set, but increases the RMSE value of the test data set, which means that it overfits the model. Similarly, adding complex activation layer functions such as sigmoid or tanh functions tend to overfit the model, where a linear activation function performs better. Table 5.10 displays the summary of results, when MLP regression models are trained with sigmoid activation function. In this approach, it has obtained low average RMSE values in cross-validation data set, but the RMSE value in test data set has significantly increased, which implies that the model has overfit the training data.

	Day 1	Day 2	Day 3
Average RMSE in cross-validation data set	0.057	0.092	0.125
RMSE in test data set	0.052	0.138	0.165
Average correlation coefficient in cross-validation data set	1.000	1.000	0.999
Correlation coefficient in test data set	1.000	0.999	0.999
Maximum absolute error percentage in cross-validation data set	1.27%	2.04%	2.64%
Maximum absolute error percentage in test data set	0.69%	1.31%	1.37%

Table 5.10 Performance of the MLP regression model on cross-validation and test data sets with sigmoid activation function

From table 5.7 and 5.9, it can be observed that, both MLP and linear regression models have performed better with the test data set than the cross-validation data set. One of the reasons for that is, the predicted values in the cross-validation data set are predicted with a single model out of the 10 models in 10-folds. But, the prediction for the test data set is done with all 10 models of 10-folds and then the average is taken as the predicted value. This was done to avoid overfitting the model by predicting for instance, from which the prediction model is trained.

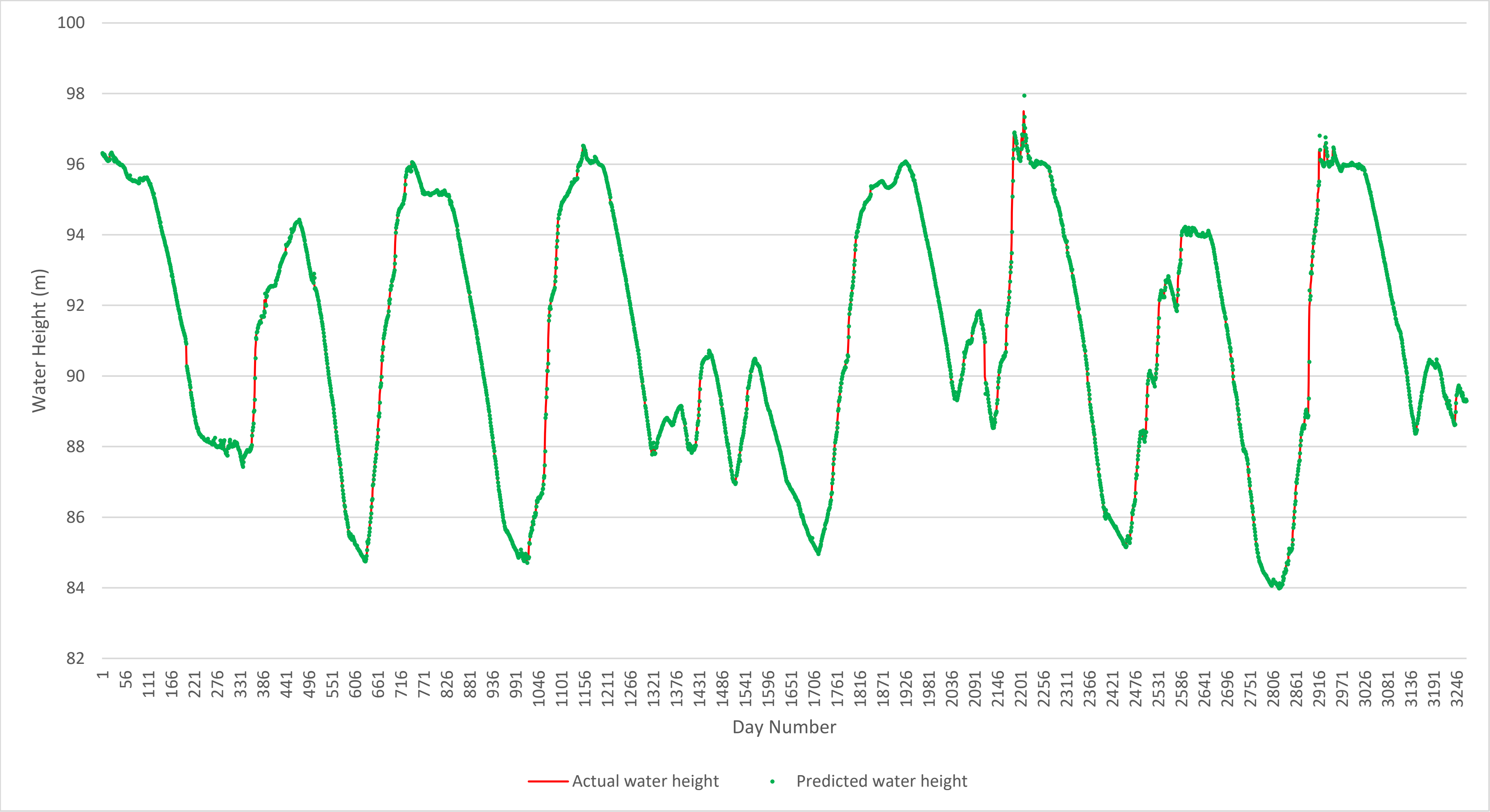


Fig 5.30 Variation of actual & predicted water height of cross-validation data set on day 1, with MLP regression model

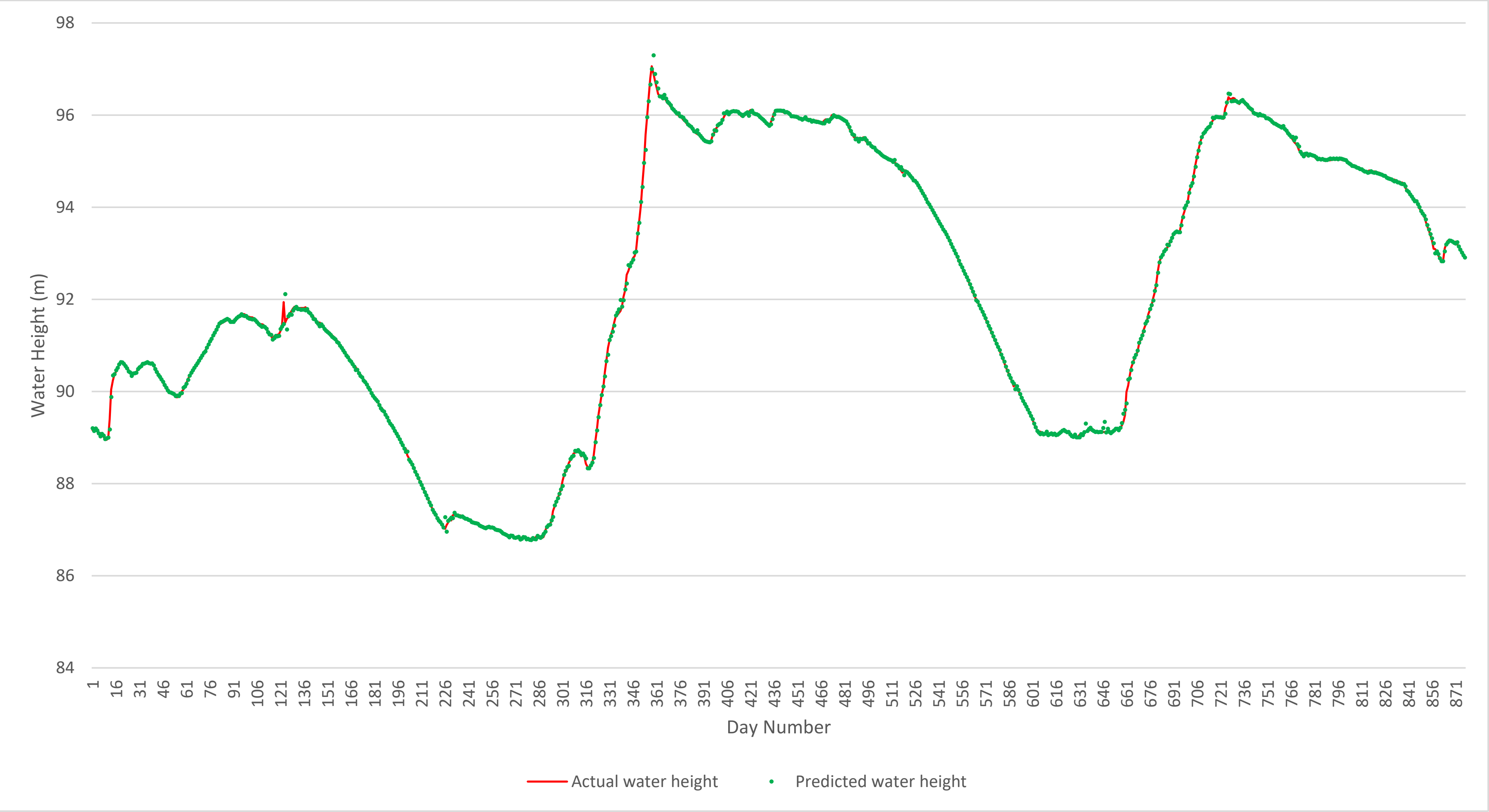


Fig 5.31 Variation of actual & predicted water height of test data set on day 1, with MLP regression model

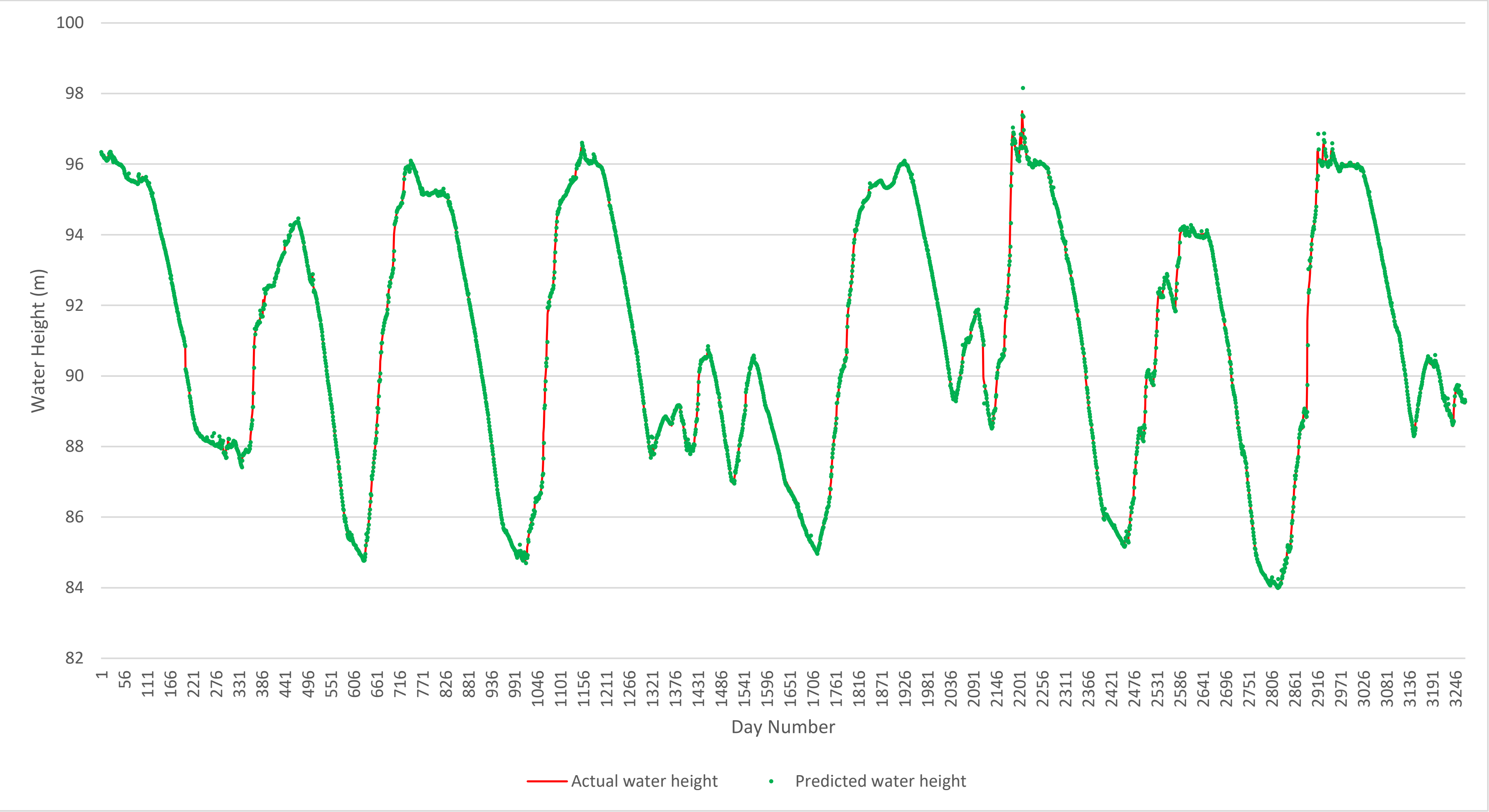


Fig 5.32 Variation of actual & predicted water height of cross-validation data set on day 2, with MLP regression model

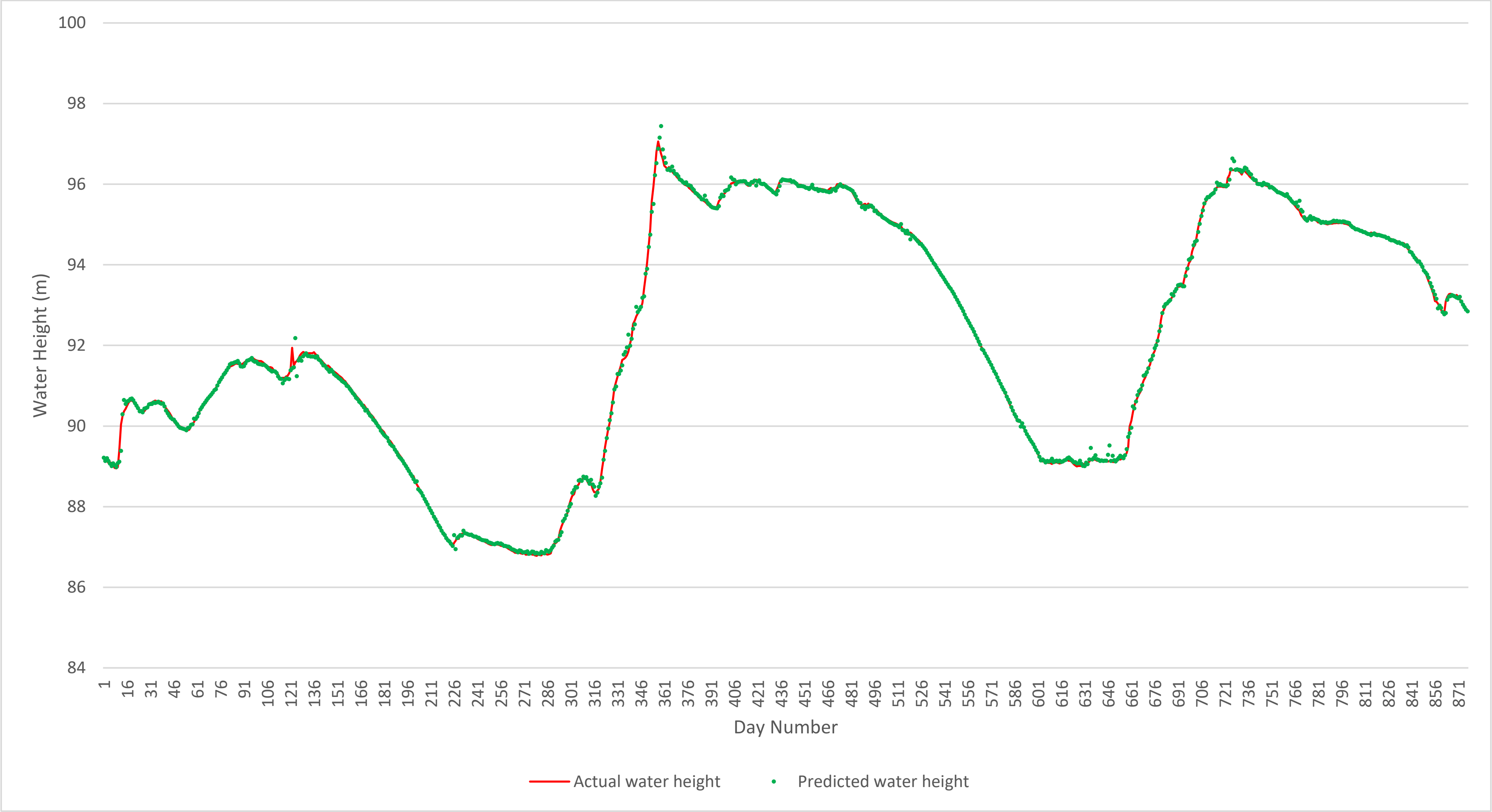


Fig 5.27 Variation of actual & predicted water height of test data set on day 2, with MLP regression model

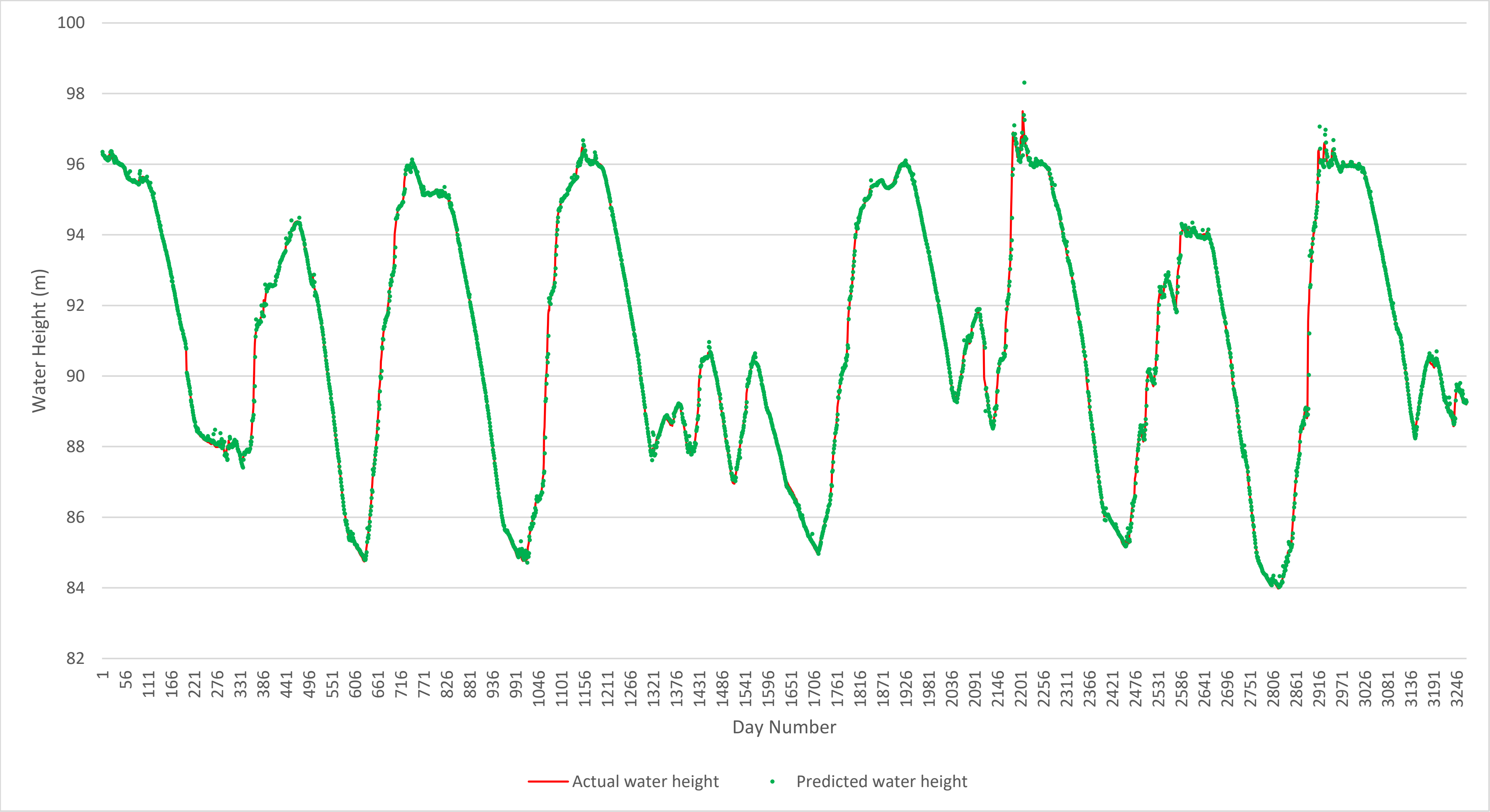


Fig 5.34 Variation of actual & predicted water height of cross-validation data set on day 3, with MLP regression model

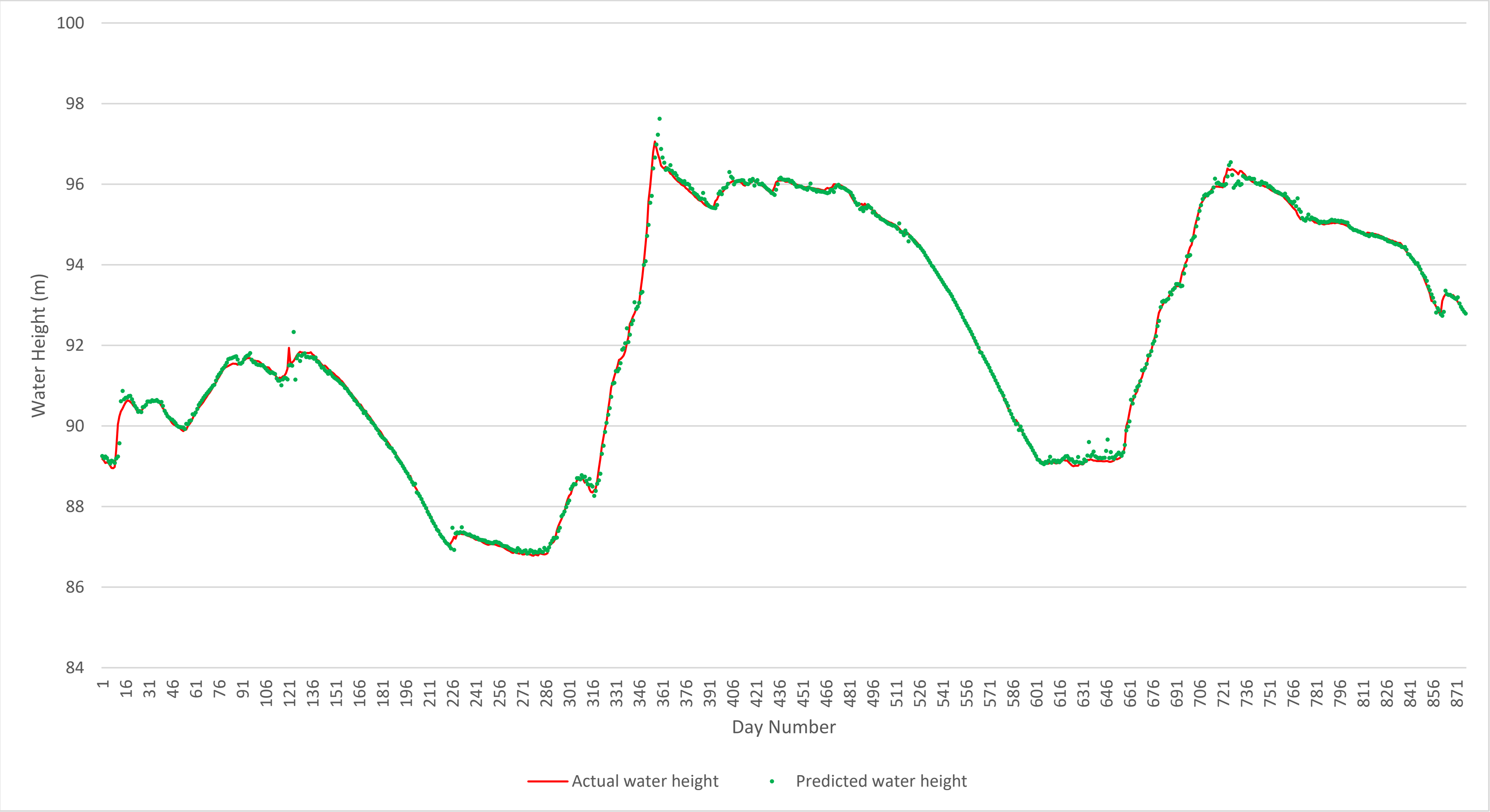


Fig 5.35 Variation of actual & predicted water height of test data set on day 3, with MLP regression model

5.4 Results comparison

Table 5.9 compares the performance of the proposed novel approach with the previous researches performed. As per the table 5.9, it is clear that the proposed novel approach has performed better than all other approaches except the methods proposed by Ozgur et al[20]. In this approach, authors have considered the Iznik Lake in Turkey as the case study, which receives an average annual rainfall of 600-800mm[21], which is far below than the average annual rainfall of Maduru Oya reservoir, which is 2100 mm. Thus, when the suggested method was replicated on Maduru Oya reservoir, the performance was poor than the mentioned values in the paper, which is summarized under model 9, in table 5.9. In addition, the approach was replicated considering rainfall as an additional feature, where similar values were observed. The proposed novel approach considers additional features to cover the uncertainty of rainfall & water inflow and applies K-Medoids clustering before applying the Feed Forward ANN (MLP regression). This approach has performed better than the model 9 and 10 presented in the table 5.9, which implies that it is the best approach to forecast the future water height of a reservoir, when there is a considerable amount of rainfall, and the water inflow is uncertain due to human intervention.

	Description	Case study reservoir	Model	RMSEs	Other measures
1	A model to forecast the water height on next 10 days from water height of past 10 days [15]	Sukhi reservoir, India	Feed Forward 10-10-1 ANN	0.82	Correlation coefficient: 0.97
2	A model to forecast the water height in next 1 to 3 hours, during floods [16]	Shihmen reservoir, Taiwan	Adaptive neuro-fuzzy inference system (ANFIS)	0.597, 1.007 & 1.186 for next 3 hours respectively	
3	A model to forecast the water height on following day, considering the current water height and rainfall as features [17]	Klang Gates Dam, Malaysiya	ANFIS	0.242	Maximum error percentage: 4.01%
4		Rantau Panjang station, Johor river, Malaysia	ANFIS	0.193	Maximum error percentage: 4.35%
5	A model to forecast the water height on following day, with past 5 water heights as features [18]	Millers Ferry Dam, Alabama, United States	ANN	0.057	MSE: 0.0032
6	A model to forecast the water height on next 3 days with the information of water height of past 3 days [20]. The lake region receives an average annual rainfall of 600-800mm [21], which was not considered separately	Iznik Lake, Bursa, Turkey	GEP	0.040, 0.070 & 0.102 for next 3 days respectively	Corresponding R^2 values: 0.998, 0.994 and 0.988
7			ANFIS	0.041, 0.073 & 0.104 for next 3 days respectively	Corresponding R^2 values: 0.998, 0.994 and 0.987
8			ANN	0.042, 0.079 & 0.114 for the next 3 days respectively	Corresponding R^2 values: 0.998, 0.993 and 0.985

9	Replicating the model number 6 in the table for Maduru Oya reservoir, which receives an average annual rainfall of 2100mm and a human controlled inflow from Ulhitiya reservoir	Maduru Oya reservoir, Sri Lanka	GEP	0.069, 0.132 & 0.187 for next 3 days respectively	Corresponding correlation coefficient values: 1.000, 0.999 & 0.999 R^2 values: 1.000, 0.999, 0.997
10	Replicating the model number 6 on Maduru Oya reservoir, with rainfall as an additional feature	Maduru Oya reservoir, Sri Lanka	GEP	0.069 for the prediction of following day	
11	Proposed novel approach	Maduru Oya reservoir, Sri Lanka	K-Medoids clustering + Feed Forward ANN	0.050, 0.083 & 0.122 for next 3 days respectively	Corresponding correlation coefficient values: 1.000, 1.000 & 0.999 R^2 values: 1.000, 0.999 & 0.998 Maximum absolute error percentages: 0.69%, 0.73% & 0.93%

Table 5.11 Comparison of water height prediction models

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

The objective of this research was to provide a simulation model to forecast the future water heights of a reservoir, irrespective of the uncertainty of rainfall and water inflow from upper stream and to identify the most effective set of features to forecast the future water level. Even though many features which may affect the water level of a reservoir are researched, it was identified from the case study performed on the Maduru Oya reservoir, that a few number of features is sufficient to build a prediction model to forecast the water level.

It was observed that the most impacting features were the previous water heights. Moreover, it has a temporal effect of 2 days, where it is sufficient to know the water heights of past two days to predict the water height on next day. In addition, the rainfall, volume of water received, volume of water issued and volume of water planned to issue have impacted on the future water height as expected. It was also observed that the month of the year has an impact on the future water level, when rainfall information is insufficient to predict three days ahead. It may be due to the fact that; the month of the year mimics monsoon seasons of a year. It was concluded that the length of day time, which is the difference between sunset and sunrise time also has an impact on the water level, which may be representing the evaporation and monsoon seasons, since it has a sinusoidal repetitive pattern across years.

Through this research, it is concluded that clustering data instances improves the prediction accuracy. Clustering the instances simplifies the problem, where it reduces the complexity of the regression models. Furthermore, it was concluded that K-Medoids clustering reduces the RMSE than the K-Means clustering. But, the model may lead to overfit, if the number of clusters is high, since the number of samples in a cluster will be less and the cluster separation will be fine-grained.

Once data instances are properly clustered, the regression model gets considerably simplified, where even the linear regression model can predict the future water height with a greater accuracy. The linear regression model had RMSE values of 0.050, 0.083 and 0.122 respectively for the predicted water height on day 1, day 2 and day 3 in the test data set. Subsequently, MLP regression model obtained same RMSE values for the test data set. But, maximum absolute error percentages of the MLP regression model were 0.69%, 0.73% and 0.93% for three days respectively, which is lower than the values of the linear regression model. Therefore, it can be concluded that both linear regression model and MLP regression model have performed well, since the required regression model has become simple due to the clustering prior to training the regression model. But, K-Medoids clustering with MLP regression model is the optimum solution to forecast the water heights of a reservoir when there is a significant amount of rainfall and uncertain water inflow.

This research concludes that the proposed novel approach has performed better than all other approaches except the methods proposed by Ozgur et al[20]. In their approach, authors have considered the Iznik Lake in Turkey as the case study, which receives an average annual rainfall of 600-800mm[21], which is far below than the average annual rainfall of Maduru Oya reservoir, which is 2100 mm. Thus, when the suggested method was replicated on Maduru Oya reservoir, the performance was poor than the mentioned values in the paper[20]. The proposed novel approach of considering additional features to cover the uncertainty of rainfall & water inflow and applying K-Medoids clustering before applying the MLP regression (feed forward ANN) model, has performed better than the presented method in the paper[20]. Thus, it can be concluded that the proposed approach is the best approach to forecast the future water height of a reservoir, when there is a significant amount of rainfall, and the water inflow is uncertain due to human intervention.

Finally, it can be concluded that predicting the water height of a reservoir is a temporal problem, where the water height mainly relies on previous water heights, and it can be accurately predicted for a greater extent, with a very limited set of features.

6.2 Future work

Most of the reservoirs in countries like Sri Lanka are interconnected, where there are cascaded reservoir networks. To manage the water in such networks, this research should be further extended on cascaded systems to simulate the entire system, addressing it as a spatio-temporal problem.

REFERENCES

[1] B. S. Salim, “Stochastic Optimization of Water Transfer Between Dams,” *Procedia - Soc. Behav. Sci.*, vol. 109, pp. 1035–1039, 2014.

[2] N.T.S. Wijesekera.(1999). *Reservoir System in River Mahaweli, Associated Hydrology and Flood Behaviour prior to and after Development Works*. Accessed on: Jan. 20, 2017.[Online]. Available: http://geoinfo.mrt.ac.lk/waterresources/publications/D028.pdf

[3] P. B. Dharmasena, “Essential Components of Traditional Village Tank Systems 1,” in *Proc. Irrigation Systems for Rural Sustainability*, Colombo, Sri Lanka, Dec. 9, 2010.

[4] J. Itakura. *Water Balance Model for Planning Rehabilitation of a Tank Cascade Irrigation System in Sri Lanka*. International Irrigation Management Institue,1995, pp. 4-5.

[5] “Food Insecurity follows Floods in Sri Lanka”, (2016). Accessed on: Aug. 12, 2017. [Online]. Available: http://floodlist.com/asia/food-insecurity-follows-floods-sri-lanka.

[6] “Maduru Oya Ancient Sluice,” (2017). Accessed on: Aug. 12, 2017. [Online]. Available: http://amazinglanka.com/wp/maduru-oya-ancient-sluice/.

[7] C. J. Jayatilaka, R. Sakthivadivel, Y. Shinogi, I. W. Makin, and P. Witharana. *Predicting water availability in irrigation tank cascade systems: the cascade water balance model*.International Water Management Institute,2001.

[8] P. Revesz and T. Triplet, “Temporal data classification using linear classifiers,” *Inf. Syst.*, vol. 36, no. 1, pp. 30–41, 2011.

[9] A. Mohan and P. Revesz, “Temporal data mining of uncertain water reservoir data,” in *Proc. Third ACM SIGSPATIAL Int. Work. Querying Min. Uncertain Spat. Data*, pp. 10–17, 2012.

[10] W. Hussain, W. Ishak, K. R. Ku-Mahamud, and N. Norwawi, “Neural Network Application in Reservoir Water Level Forecasting and Release Decision,” *Int. J. New Comput. Archit. Their Appl.*, vol. 1, no. 2, pp. 265–274, 2011.

[11] N. A. Ashaary, W. Hussain, W. Ishak, and K. R. Ku-, “Forecasting Model for the Change of Reservoir Water Level Stage Based on Temporal Pattern of Reservoir Water Level,” in *Proc. ICOCI 2015*, Istanbul, Turkey, Aug. 2015, pp. 692–697.

[12] M. J. Diamantopoulou, P. E. Georgiou, and D. M. Papamichail, “Daily Reservoir Inflow Forecasting Using Time Delay Artificial Neural Network Models,”in *Proc. IASME/WSEAS*, Chalkida, Greece, May 2006, pp.1-6.

[13] S. E. Fahlman and C. Lebiere, “The Cascade-Correlation Learning Architecture,” *Adv. Neural Inf. Process. Syst.*,vol. 2, pp. 524–532, 1990.

[14] T. Yang, X. Gao, and S. Sorooshian, “Simulation of California ’ s Major Reservoirs Outflow Using Data Mining Technique,” in *Proc. American Geophysical Union Fall Meeting,*United States, Dec.2014.

[15] S. Rani and F. Parekh, “Predicting Reservoir Water Level Using Artificial Neural Network,” *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 3, no. 7, pp. 14489–14496, 2014.

[16] F. Chang and Y. Chang, “Adaptive neuro-fuzzy inference system for prediction of water level in reservoir,” *Adv. Water Resour.*, vol. 29, pp. 1–10, Jul. 2006.

[17] N. Valizadeh, A. El-shafie, M. Mirzaei, H. Galavi, M. Mukhlisin, and O. Jaafar, “Accuracy Enhancement for Forecasting Water Levels of Reservoirs and River Streams Using a Multiple-Input-Pattern Fuzzification Approach,” *Sci. World J.*, vol. 2014, 2014.

[18] F. Üne, M. Demirci, Ö. Ki, and Ö. Ki, “Prediction of Millers Ferry Dam Reservoir Level in USA Using Artificial Neural Network,” *Period. Polytech. Civ. Eng.*, vol. 59, no. 3, pp. 309–318, 2015.

[19] “U.S. climate data,”.Accessed on: Dec. 27, 2017.[Online]. Available: https://www.usclimatedata.com/climate/birmingham/alabama/united-states/usal0054.

[20] O. Kisi, J. Shiri, and B. Nikoofar, “Computers & Geosciences Forecasting daily lake levels using artificial intelligence approaches,” *Comput. Geosci.*, vol. 41, pp. 169–180, 2012.

[21] S. Sensoy, M. Demircan, and Y. Ulupınar, “Climate of Turkey,” *Turkish State Meteorological Service*, 2008. [Online]. Available: https://www.mgm.gov.tr/files/en-US/climateofturkey.pdf.

[22] Department of Meteorology Sri Lanka (2017). *Climate of Sri Lanka*, Department of Meteorology,Sri Lanka. Accessed on: Dec. 27, 2017. [Online]. Available: http://www.meteo.gov.lk/index.php?option=com_content&view=article&id=94&Itemid=310

[23] M. Shukla and M. Seth, “Data Analysis in Forecasting Lakes Levels Using K-Medoid Clustering,” *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 5, no. 6, pp. 11067–11074, Jun. 2016.

[24] A. Mohan, “A New Spatio-Temporal Data Mining Method and Its Application to Reservoir System Operation,” Research Master Thesis, University of Nebraska, Nebraska, 2014. Accessed on: Dec., 28, 2016. Available: http://digitalcommons.unl.edu/computerscidiss/74/

[25] K. V Rao, A. Govardhan and K. C. Rao, “Spatiotemporal Data Mining: Issues, Tasks And Applications,” *Int. J. Comput. Sci. Eng. Surv.*, vol. 3, no. 1, pp. 39–52, Feb. 2012.

[26] B. Malekmohammadi, R. Kerachian, and B. Zahraie, “Developing monthly operating rules for a cascade system of reservoirs: Application of Bayesian Networks,” *Environ. Model. Softw.*, vol. 24, no. 12, pp. 1420–1432, 2009.

[27] O. Al-azzam, D. Sarsar, K. Seifu, and M. Mekni, “Flood Prediction and Risk Assessment Using Advanced Geo-Visualization and Data Mining Techniques : A Case Study in the Red-Lake Valley 1 Introduction 2 Literature Review,” in *Proc. ACACOS,*Malaysia,2014, pp. 18–27.

[28] K. Yurekli, M. Taghi Sattari, A. S. Anli, and M. A. Hinis, “Seasonal and annual regional drought prediction by using data-mining approach,” *Atmosfera*, vol. 25, no. 1, pp. 85–105, Jul. 2011.

[29] C. C. Tsai, M. C. Lu, and C. C. Wei, “Decision Tree-Based Classifier Combined with Neural-Based Predictor for Water-Stage Forecasts in a River Basin During Typhoons: A Case Study in Taiwan,” *Environ. Eng. Sci.*, vol. 29, no. 2, pp. 108–116, 2012.

[30] F. Richards and P. Arkin, “On the Relationship between Satellite-Observed Cloud Cover and Precipitation,” *Monthly Weather Review,* vol. 109, no. 5, pp. 1081–1093, 1981.

[31] P. Arkin and B. Meisner, “The relationship between large-scale convective rainfall and cold cloud over the western hemisphere during 1982-84,” *Monthly Weather Review*, vol. 115, pp. 51–74, 1987.

[32] R. Suppiah and M. M. Yoshino, “Rainfall Variations of Sri Lanka, Part 1: Spatial and Temporal Patterns,” *Arch. Meteorol. Geophys. Bioclimatol. Ser. B*, vol. 35, no. 1–2, pp. 81–92, 1984.

[33] T. Kohyama and J. M. Wallace, “Rainfall variations induced by the lunar gravitational atmospheric tide and their implications for the relationship between tropical rainfall and humidity,” *Geophys. Res. Lett.*, vol. 43, no. 2, pp. 918–923, 2016.

[34] K. Yamauchi, M. Ban, N. Kasahara, T. Izumi, H. Kojima, and T. Harako, “Physiological and behavioral changes occurring during smoltification in the masu salmon, Oncorhynchus masou,” *Aquaculture*, vol. 45, no. 1–4, pp. 227–235, 1985.

[35] “Google Maps.”. Accessed on: Aug 12,2017. [Online]. Available: https://www.google.lk/maps/place/Maduru+Oya+Reservoir/.