Prediction of Irrigation Water Supply Using Supervised Machine Learning Models in Koga Irrigation Scheme, Ethiopia

Menwagaw T. Damtie[1(✉)], Seifu A. Tilahun[2], Fasikaw A. Zimale[2], and Petra Schmitter[3]

[1] Department of Hydraulic and Water Resources Engineering, Debre Tabor University, 272 Debre Tabor, Ethiopia
[2] Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, 26 Bahir Dar, Ethiopia
[3] International Water Management Institute (IWMI), Addis Ababa, Ethiopia

Abstract. Estimating water supply through irrigation canal distribution systems is a crucial process for better water management in the irrigation schemes. This study aimed to develop an approach to predict the discharge delivered to unregulated irrigation canals (such as quaternary canals) from geometric and hydraulic information of regulated section of the system (such as tertiary and others) using machine learning approaches. The prediction performance of four Caret-based Supervised Machine Learning Models, namely; Multivariate Adaptive Regression Splines (MARS), Artificial Neural Networks (ANN), Random Forest (RF), and Radial Basis Support Vector Machines (SVM), were developed in the R programming environment, followed by variability assessment among canal outlets at Koga irrigation Scheme. Water delivery performance at quaternary canals showed a significant flow variation among the canal outlets. The comparative study of model prediction results showed identified MARS as the optimal model, both at the training stage (RMSE = 0.074 & R^2 = 0.86 with normalized data) and testing stage (RMSE = 3.89 & R^2 = 0.85 with rescaled data). Furthermore, the model building process and output equations of MARS were relatively interpretable compared to neuro and tree-based models, such as Artificial Neural Network and Random Forest. Thus, the MARS model was recommended to estimate the water supply to ungated irrigation canals as a function of flow rate information at gated distributary canal and other field data at lower components of irrigation schemes.

Keywords: Canal outlet · Machine learning · R environment · MARS

1 Introduction

Most large-scale irrigation schemes developed in Africa have serious water management problems caused by the absence of flow controlling gates in canal distribution networks, infrequent maintenance of the systems' infrastructures and poor operation of the system relying on traditional basis, which all are greatly reducing the sustainability of irrigation schemes [1]. The study of Adhakari [4] also explained that many of the

M. L. Berihun (Ed.): ICAST 2021, LNICST 412, pp. 68–81, 2022.
https://doi.org/10.1007/978-3-030-93712-6_5

modern irrigation systems are equipped with water control gates only at the upper canal networks whereas, the irrigation water flow across low level canals is kept either at full supply level or at no flow condition for some days on a rotational basis. These low-level canals are mostly built with non-adjustable outlet structures, to provide proportional distribution of irrigation supply to the watercourses [5]. However, the assessment result of Tariq et al. [5] clearly indicates significant flow variability across canal outlets with head outlets are withdrawing more than planned discharge, while tail outlets suffer the most [5].

Koga irrigation project, one of the modern schemes in Ethiopia, currently faces water share disputes of users where flow is not adequately controlled and regulated [6]. Since flow controlling mechanisms are absent at quaternary canal levels of the scheme, quantifying the actual amount of irrigation water at these outlets is very difficult, which leads to inequitable, unreliable and inadequate supply of water to the irrigation fields within and across irrigation blocks [27].

Associating irrigation water supply at ungated irrigation canal outlets with available operational and field information is a crucial step towards improving scheme governance and efficiency.

Since irrigation water supply is reliant on field and operational conditions, machine learning models are required to predict the discharge response delivered to small irrigation water courses at data scarce environments.

Divya et al. [10] defined that machine learning is simply training a model with data and then using the model to predict any new data. After the training, the machine can perform automatically and can also learn to fine-tune. In the learning system, it comprises of four design choices; namely, choosing the training data, the target function, the representation and the learning algorithm [11, 12].

The study of Mohri et al. [13] classifies machine learning into three broad categories: namely; supervised learning, unsupervised learning and reinforcement learning. In supervised machine learning technique, the machine is fed with paired sets of inputs and outputs, which are called labeled datasets whereas, in case of unsupervised machine learning, the machine is given a dataset without the output sets [14].

The application of machine learning models is currently emerging to create new opportunities to understand data intensive processes in agricultural activities [15]. Prediction of various water resource problems were investigated by different scholars. Gu et al. [16] used neural network and genetic algorithm machine learning algorithms to develop the yield-irrigation water model for predicting the corn yield for different irrigation systems under subsurface drip irrigation. The hybrid model of these algorithms gave accurate predictions of the yield with the average error of 0.71%.

Parsaie et al. [17] had applied supervised machine leaning models such as, Multivariate Adaptive Regression Splines (MARS), Artificial Neural Network (ANN), and Support Vector Machine (SVM) for prediction of discharge coefficient (Cd) of lateral intakes, irrigation and drainage networks. The result of the study indicated that MARS model outperforms the ANN and SVM models. The tangent sigmoid and radial basic functions were found to be the most efficient transfer and kernel functions for ANN and SVM respectively.

It was found that the performance of MARS model was high compared with neuro based and fuzzy inference systems, for prediction of monthly stream flow in a

mountainous basin [18]. It was also studied that the MARS and SVM-Radial Basis Function models generally performed better than gene expression programming (GEP) and SVM-Polynomial models in estimating monthly mean reference evapotranspiration [19]. A study of groundwater potential mapping was made using multivariate adaptive regression spline (MARS) and random forest (RF) with the aid of GIS tool [2]. The prediction performance of RF and MARS are relatively good in estimating groundwater spring potential, the later slightly performed higher from the two models [2].

In general, MARS was first presented by Friedman [20], and recently it is used in most fields of water resources engineering. It is a nonparametric statistical method which simply develops interpretable functional relationships between a set of input variables and the target dependent variables [21]. The nature of the MARS feature breaks the predictor into two groups and models the relationships between the predictor and the outcome in each group [22]. Among the many options, machine learning algorithms are chosen on the basis of the input data and the learning task.

R provides support for machine learning in input data processing and learning tasks [23]. It is a free open-source programming language that has several packages needed for machine learning [24]. Caret (short, for Classification and regression training) is a unified interface package available for download in Cran into R, to simplify the analysis and interpretation of black box machine learning models [24]. Caret includes several other packages and also have methods for pre-processing training data, model training, tuning, calculating variable importance, and model visualizations [24, 25].

In this study, the performance of four machine learning models (multivariate adaptive regression spline artificial neural network, support vector machine, and random forest) were compared for irrigation water supply prediction, to provide insights by applying such models in data scarce irrigation systems.

2 Materials and Methods

2.1 Description of the Study Area

Koga Irrigation scheme is found in the upper source of the Blue Nile, specifically located in Koga Watershed near Merawi town, West Gojjam Zone, Ethiopia [26]. The reservoir of Koga has a capacity to store about 83 mm^3 of water, which can irrigate 12 irrigation blocks covering a total of 7,000 ha reaching more than 10000 beneficiaries [26]. The canal network comprises of 19.7 km of lined main canal, 52 km of lined secondary canals, 156 km of tertiary canals, 905 km of unlined quaternary canals and 11 Night Storage Reservoirs [27]. The head of quaternary canals are built without controlling gates and quantifying the water delivered through these units is a difficult task. Canal flow through quaternary units is operated by group leaders so called “water fathers” in a rotational basis among farmers.

2.2 Field Investigation and Data Collection

This study was carried out at Koga Irrigation Scheme through the support of a collaborative partnership between the International Water Management Institute (IWMI) and Bahir Dar University in a FAO funded project entitled 'Closing the Water Productivity Gap' during 2018 and 2019 irrigation seasons. The field setup was appropriately planned at different reaches of the scheme, to observe the spatial and temporal variability of water supply through canal outlets. Six blocks out of twelve available irrigation blocks (two blocks from each head, middle and tail reaches of the scheme), six tertiary canals (one tertiary canal from each block), and 18 quaternary canal outlets (three outlets from each tertiary canal) were selected. Water level data were recorded in a weekly basis by using a standard thin plate 90^0-notch weir. This type of weir is used as temporal flow measuring device at minor irrigation canals [6]. According to Erikson [6] and Halefom et al. [7], a 90° v-notch thin-plate weir is often preferred because of its greater accuracy at low flows, and its lesser sensitivity to approach channel geometry and velocity distribution.

The weirs were installed at selected irrigation blocks just near to the exit of quaternary canals. The water level of quaternary canals was first recorded from the graded notches and then converted to discharge using an equation stated by USBR's water measurement manual [8] as follows,

$$Q = 1.38H^{2.5} \tag{1}$$

where, Q = discharge over the weir and H = water head above the weir notch. Alongside, water level data were collected at regulated tertiary canal heads. The existing rectangular concrete weirs at tertiary canal heads variable dimensions and hence a dimensionless number (h), the ratio of water level to weir width was used for further analysis. The data of dimensionless parameter (a = the ratio of irrigated area under a quaternary canal to tertiary canal), was collected from Koga Agricultural Bureau. The manning's coefficient for canal beds were used from design information and previous soil analysis results whereas, the distance of quaternary canal outlets from tertiary canal heads (l) were directly measured using GPS tracks. The diameter of canal outlet pipes (d) was measured using tapes, and the rank of operated outlets (r) was recorded during flow measuring events.

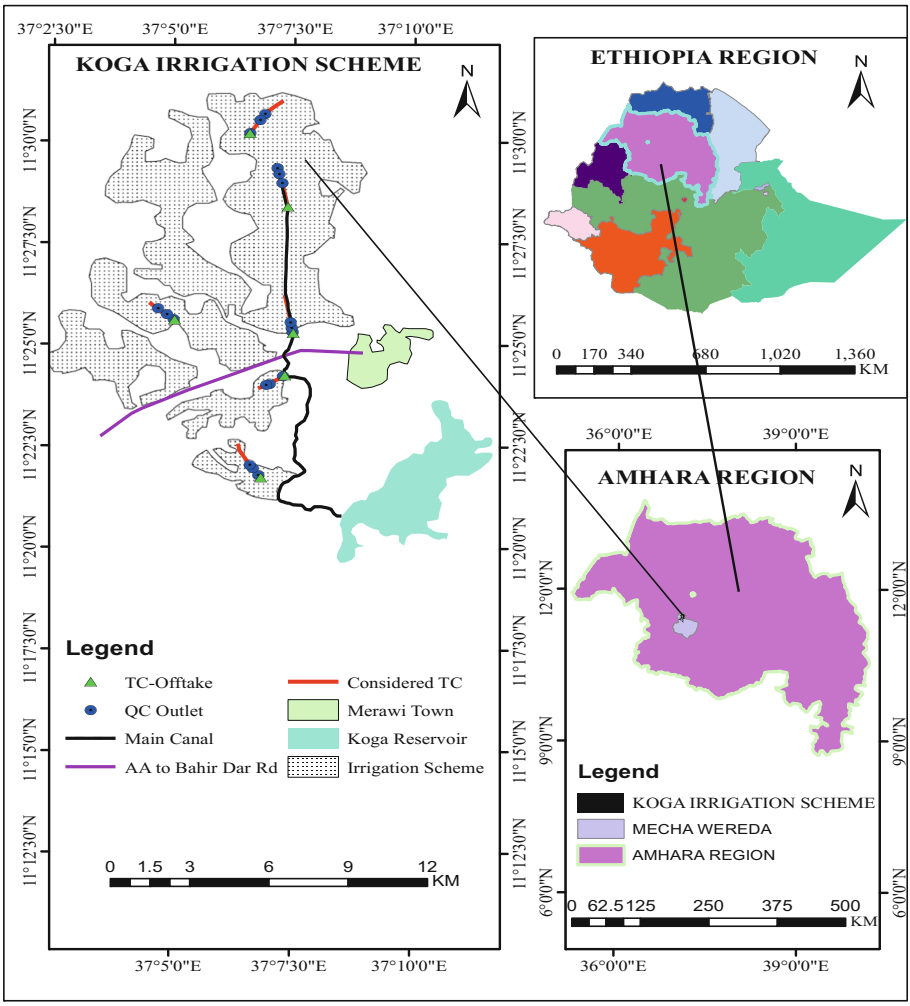

Fig. 1. Location of the study area

2.3 Developing Models

Model Inputs: The primary goal of this study was to develop a predictive model for estimating the irrigation water supply at a quaternary canal outlet, located at certain distance from the regulated distributary head canal. Several factors such as, geometric of the outlet structures, hydraulic characteristics, location, and operational management factors would affect the delivery performance of the outlets. Finally, six parameters were involved for estimating outlet discharge as follows;

$$Q \approx f(h, a, l, n, r, d) \tag{2}$$

where, Q = estimated outlet discharge at quaternary canal (l/s), h = water level per unit width at tertiary canal head weir($h = \frac{H}{B}$, dimension less), a = ratio of irrigated area (A) under quaternary canal outlet to total irrigated area(At) under tertiary canal weir ($a = \frac{A}{At}$, dimensionless), H = water level at tertiary canal off-taking weir (m), B = width of off-taking weir structure (m), l = distance of the outlet from the TC head (m), n = Manning roughness coefficient at tertiary canal (dimensionless), d = diameter of the outlet structure, and r = the ranking order of the operated outlets in ascending order along a tertiary canal (dimensionless) to describe the cumulative flow errors at the outlets u/s of the considered outlet. Table 1 shows number of pairs of observations and the statistical description of model input variables.

Table 1. Statistical summary of variables

Variable	Unit	Dataset	Minimum	Maximum	Mean	Std. Dev
Q	l/s	453	2.35	47.57	27.05	
h	–	453	0.41	0.98	0.71	0.13
a	–	453	0.07	0.20	0.15	0.03
l	m	453	20.00	1734.64	665.19	517.26
n	–	453	0.02	0.03	0.02	0.00
r	–	453	1.00	7.00	3.51	1.85
d	m	453	0.15	0.15	0.15	0.00

Data Preprocessing: The goal of transforming a set of data in machine learning is to facilitate faster learning of the algorithms. Unless the values of all variables have the same ranges, some machine learning algorithms such as, ANN would show syntax error to display the model performance criterion. A maximum- minimum normalization technique was used to transform data into the required format. Data partitioning into training and testing sets were performed many times in different ratios to get models best fit of the data.

Machine Learning Models: In this study, the motive behind applying machine learning models was their strong ability to map and learn the input data to predict the desired output where other methods such as multiple linear regression could not perform well. Several machine learning models were reviewed. The research was a multivariate supervised machine learning problem which required a quantitative modeling approach. The R environment has an immense number of packages for machine learning. Since selection of models from several individually named machine learning algorithms is difficult and time consuming, the Caret package was selected to train and compare many algorithms at a time, using a resampling technique. The initial selection of machine learning models was mostly literature based using recent water science studies with a similar scope.

Based on the research target, accessibility of algorithms, and the reviewed prediction performance, four classic machine learning models were selected [16, 17]. The models

were; Multivariate Adaptive Regression Splines (MARS), Artificial Neural Networks (ANN), Support Vector Machines (SVM) with radial basis, and Random Forest (RF).

Model Development Process: The flow chart in Fig. 2 showed the procedures to develop the final model for predicting canal outlet discharge:

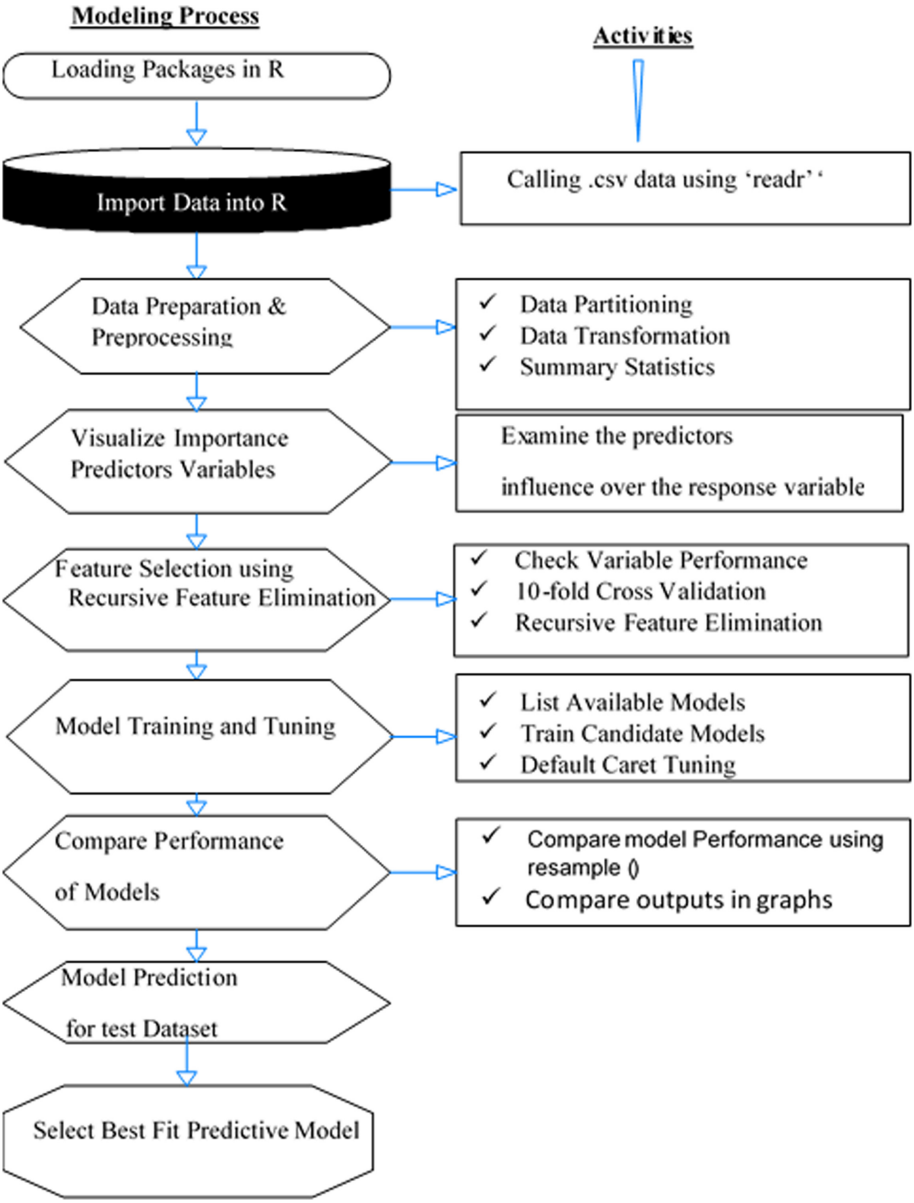

Fig. 2. Flow chart for model development

Model Evaluation Criteria: Evaluation of models refers to describing how well the trained model is performing to predict the targeted problem. The following regressive metrics were used in to evaluate the performance of models:

1. **Root Mean Square Error (RMSE)** is the standard deviation of the residuals (prediction errors). The residuals imply that how far the data points are from the regression line. The RMSE is calculated as,

$$RMSE = \sqrt{\frac{\sum \left(y_i - y_p\right)^2}{n-1}} \tag{3}$$

2. **Coefficient of determination ($\mathbf{R^2}$)** is interpreted as how much of the variability of the response variable is explained by the predictor variable. R^2 is an alternative measure of fit for the model over the RMSE, and is described as,

$$\mathrm{R}^2 = 1 - \frac{\sum \left(y_i - y_p\right)^2}{\sum \left(y_{i-\bar{y}}\right)^2} \tag{4}$$

3. **Mean Absolute Error (MAE)** is the average of the difference between the observed and predicted over the whole dataset, which is described as,

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |y_i \bar{y}| \tag{5}$$

Where, y_i is the i^{th} observed value, y_p is predicted value, and $\bar{y}$ is mean of observed values

3 Results and Discussion

3.1 Model Building and Variable Importance

The objective of this study was to develop an alternative approach to predict discharge in data scarce irrigation schemes. The observed discharge in ungated quaternary canals show significant flow rate variations both at spatial and temporal scales, compared with the planned flow rate per unit area at each quaternary canal. After several attempts of a randomized data splitting in boosting the best fit, a ratio of 70: 30% was used for model training and testing, respectively.

Six predictor variables, namely; water level per unit width (h), command area ratio (a), distance of outlets from tertiary canal head(l), Manning roughness coefficient of tertiary canals (n), rank of operated quaternary canal outlet along a tertiary canal (r), and modular outlet diameter were used to predict a response variable Q (outlet discharge).

Variable Importance: To examine how the predictors, influence the output discharge (Q), importance of variables was computed in R environment with caret for each model. The importance level of predictor variables was different for each model as shown at

Fig. 3. The canal outlet diameter (d) has the lowest importance level in developing all the models whereas, the area ratio(a) has the highest variable importance with ANN and MARS models.

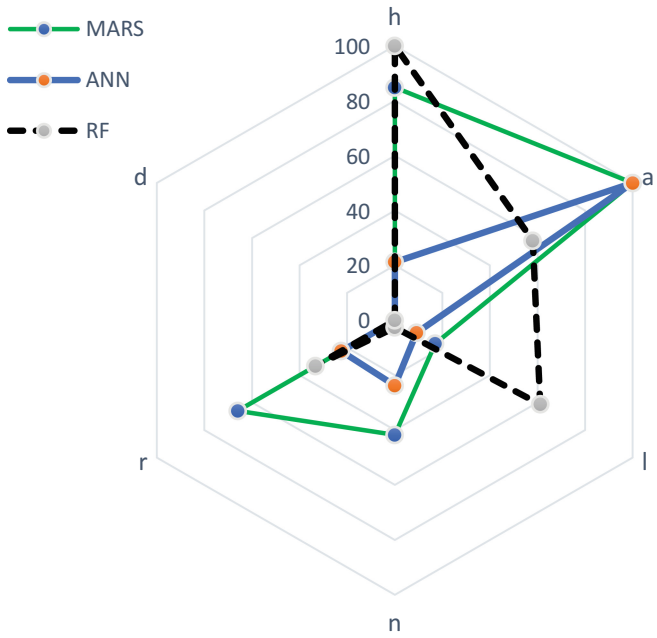

Fig. 3. Variable Importance for trained models

In Fig. 3, MARS is Multivariate Adaptive Regressive Spline, RF is Random Forest, and ANN is Artificial Neural Networks.

A recursive feature selection technique with 10-fold cross validation repeated three times, was used to analyze the resampling performance. The combination of top three model inputs variables h, a and l were selected as shown in Table 2.

Table 2. Recursive resampling performance over combination of inputs

Model inputs	RMSE	R^2	MAE
h	0.19	0.15	0.15
h, a	0.10	0.75	0.08
h, a, l	**0.08**	**0.85**	**0.06**
h, a, l, n	0.09	0.83	0.07
h, a, l, n, r	0.10 0.08	0.79 0.84	0.08 0.07

The recursive feature selection using recursive feature elimination (rfe) method, agrees with the variable importance selection of RF model. However, selection of variables with the rfe technique and eliminating the least selected variables is not advisable as different models have different importance levels. Thus, training and default tuning of models was performed using a Caret package in the R environment.

During model training using resamples, there were no significant performance differences between the models. RF slightly outperformed (RMSE = 0.075, R^2 = 0.87 and MAE = 0.056) followed by MARS (RMSE = 0.075, R^2 = 0.86 and MAE = 0.058) and ANN (RMSE = 0.078, R^2 = 0.85 and MAE = 0.06). All the metric values are with normalized datasets, including the values displayed in Fig. 4.

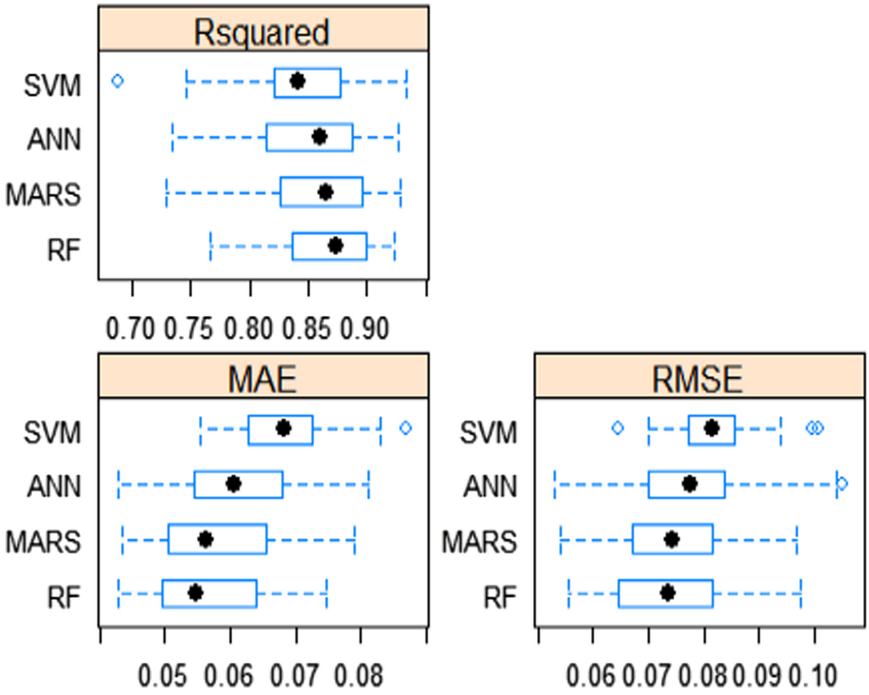

Fig. 4. Performance of models at training stage with normalized data

The learning speed of the models in the R environment for the training dataset was examined. MARS has taken the shortest duration (7.2 s) to train the model with 319 datasets and 6 predictors while, RF has taken the longest (60.1 s) in the normalized dataset (see Fig. 5).

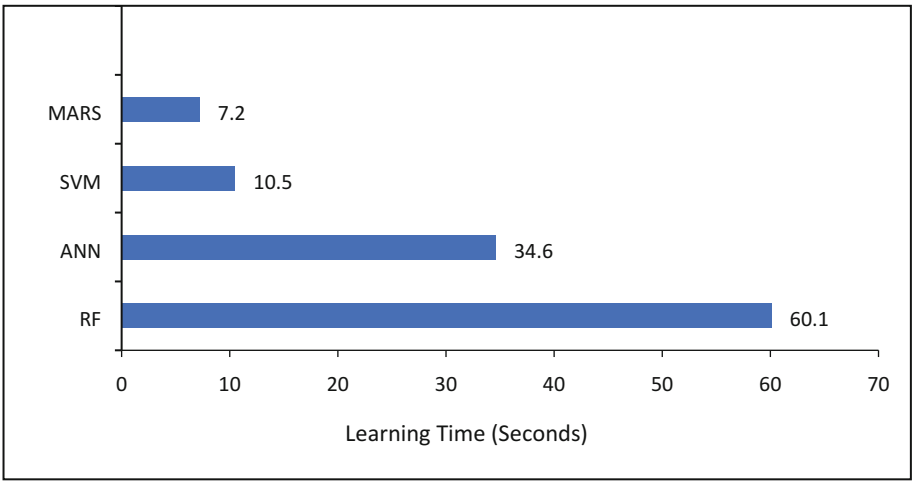

Fig. 5. Learning time of models

A randomized 30% test data was used to check the prediction performance of models with the normalized data and the outputs were rescaled to the original data structure. The results revealed that MARS was the best performing model followed by RF model (see Table 3).

Table 3. Performance of models with test dataset

	MARS	ANN	SVM	RF
R^2	**0.82**	0.76	0.76	0.78
RMSE	**3.89**	4.45	4.43	4.28
MAE	**2.91**	3.42	3.45	3.14

Model development process of MARS was relatively fast and interpretable compared to neuro and tree-based models. Furthermore, discharge prediction equation, which comprised of twelve basic functions and one intercept, was developed by this model. Three times repeated 10-fold cross validation of six variables were used to generate fifteen basic functions, out of which thirteen terms and five predictors were selected to develop the prediction equation. Figure 6 showed that the degree of prediction errors decreases as number of basic functions increase.

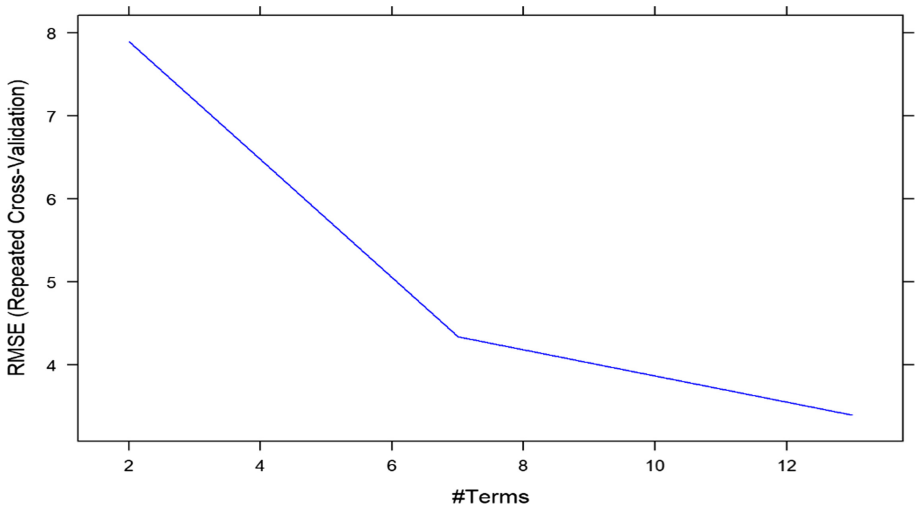

Fig. 6. Degree of errors with number of basic functions

The general Prediction equation developed by MARS was described as,

$$f(\mathrm{X}) = 26.5 + \sum_{i=0}^{12} \beta i \mathrm{BFi}(\mathrm{X}) \tag{6}$$

Where, BFi is the ith basic function, βi is the coefficient of basic functions and X is the predictor variable. Since all the basic functions generated by predictor variable (l), have zero multiplying coefficients, they were removed from Eq. 6. Using a combination of predictors, different equations were developed to predict canal outlet discharge, and the prediction performance is well described by Table 4.

Table 4. Discharge Prediction performance of MARS equations

Input variable	R^2	RMSE
a, h	0.38	5.40
h, n, r	0.60	6.34
a, h, r	0.71	5.66
a, h, r, n	0.80	4.27

The equation with all input variables, except canal outlet distance (l) has best prediction performance (RMSE = 4.27 m^3/s, R^2 = 0.8).

4 Conclusions

Water delivery performance of quaternary canal outlets were assessed in Koga irrigation scheme during the irrigation season of 2019, and significant water supply variations were found at spatial and temporal scales. Discharge prediction equation at

quaternary canal outlets was developed using six field parameters, using supervised machine learning algorithms in R environment. Four machine learning models were developed, and the optimal model both at the training stage (RMSE = 0.074 & R^2 = 0.86 with normalized data) and testing stage (RMSE = 3.89 m^3/s & R^2 = 0.85 with rescaled data) was Multivariate Adaptive Regressive Spline (MARS). The development process of MARS model was relatively interpretable compared to neuro and tree-based models. Furthermore, a prediction equation, which comprised of twelve basic functions and one intercept, from five predictors was developed by the Model. Since the distance parameter is eliminated due to its zero regression coefficients, the developed MARS equation was using four variables to predict discharge with prediction performance of RMSE = 4.27 m^3/s and R^2 = 0.80.

Acknowledgements. This study was made possible through the support of International Management Institute (IWMI) and Bahir Dar University Institute of Technology collaborative project entitled 'closing water productivity gaps' at Koga Irrigation Scheme, Ethiopia from 2017–2019. The sightings of this study are the full responsibilities of the Authors, and do not necessarily reflect the views of the above-mentioned organizations.

References

1. Agide, Z., et al.: Analysis of Water Delivery Performance of Smallholder Irrigation Schemes in Ethiopia: Diversity and Lessons Across Schemes, Typologies and Reaches (2016)
2. Zabihi, M., Pourghasemi, H.R., Pourtaghi, Z.S., Behzadfar, M.: GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. Environ. Earth Sci. **75**(8), 1–19 (2016). https://doi.org/10.1007/s12665-016-5424-9
3. Sanaee-Jahromi, S., Depeweg, H., Feyen, J.: Water delivery performance in the Doroodzan irrigation scheme Iran. Irrig. Drain. Syst. **14**(3), 207–222 (2000)
4. Adhakari, B.: Design of water distribution system: appropriateness of structured system in large irrigation projects in Nepal. Hydro Nepal J. Water Energy Environ. **19**, 25–30 (2016)
5. Tariq, J.A., Kakar, M.J.: Effect of variability of discharges on equity of water distribution among outlets. Sarhad J. Agric. **26**(1), 51–59 (2010)
6. Eriksson, S.: Water Quality in the Koga Irrigation Project, Ethiopia: A Snapshot of General Quality Parameters (2012)
7. Halefom, A., Sisay, E.: Performance assessment and diagnostic analysis of minor irrigation canal. Eng. Sci. Technol. Int. J. **7**, 10–17 (2017)
8. United States. Bureau of Reclamation: Water Measurement Manual. The Bureau (2001)
9. Chanson, H., Wang, H.: Unsteady discharge calibration of a large V-notch weir. Flow Meas. Instrum. **29**, 19–24 (2013)
10. Divya, K.S., Bhargavi, P., Jyothi, S.: Machine learning algorithms in big data analytics. Int. J. Comput. Sci. Eng. **6**(1), 64–70 (2018)
11. Weimer, M.: Machine Teaching–A Machine Learning Approach to Technology Enhanced Learning (Doctoral dissertation, Technische Universität) (2010)
12. Ciaburro, G., Venkateswaran, B.: Neural Networks with R: Smart Models Using CNN, RNN, Deep Learning, and Artificial Intelligence Principles. Packt Publishing Ltd. (2017)
13. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. (2012)
14. Ahmed, O.: Dataset Modification to Improve Machine Learning Algorithm Performance and Speed (Doctoral dissertation) (2014)

15. Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D.: Machine learning in agriculture: a review. Sensors **18**(8), 2674 (2018)
16. Gu, J., Yin, G., Huang, P., Guo, J., Chen, L.: An improved back propagation neural network prediction model for subsurface drip irrigation system. Comput. Electr. Eng. **60**, 58–65 (2017)
17. Parsaie, A., Haghiabi, A., Shamsi, Z.: Intelligent modeling of discharge coefficient of lateral intakes. AUT J. Civil Eng. **2**(1), 3–10 (2018)
18. Adnan, R.M., Liang, Z., Parmar, K.S., Soni, K., Kisi, O.: Modeling monthly streamflow in mountainous basin by MARS, GMDH-NN and DENFIS using hydroclimatic data. Neural Comput. Appl. **33**(7), 2853–2871 (2020). https://doi.org/10.1007/s00521-020-05164-3
19. Mehdizadeh, S., Behmanesh, J., Khalili, K.: Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. Comput. Electron. Agric. **139**, 103–114 (2017)
20. Friedman, J.H.: Multivariate adaptive regression splines. In: The Annals of Statistics, pp. 1–67 (1991)
21. Rezaie-Balf, M.: Multivariate adaptive regression splines model for prediction of local scour depth downstream of an apron under 2D horizontal jets. Iranian J. Sci. Technol. Trans. Civil Eng. **43**(1), 103–115 (2019)
22. Kuhn, M., Johnson, K.: Measuring performance in classification models. In: Kuhn, M., Johnson, K. (eds.) Applied predictive modeling, pp. 247–273. Springer New York, New York, NY (2013). https://doi.org/10.1007/978-1-4614-6849-3_11
23. Probst, P., Bischl, B., Boulesteix, A.L.: Tunability: Importance of hyperparameters of machine learning algorithms. arXiv preprint arXiv:1802.09596 (2018)
24. Kuhn, M.: A short introduction to the caret package. R Found. Stat. Comput. **1**, 1–10 (2015)
25. Grün, B., Leisch, F.: FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. **28**(4), 1–35 (2008). https://doi.org/10.18637/jss.v028.i04
26. Eguavoen, I., Tesfai, W.: Rebuilding Livelihoods After Dam-Induced Relocation in Koga, Blue Nile basin, Ethiopia (No. 83). ZEF Working Paper Series (2011)
27. Asres, S.B.: Evaluating and enhancing irrigation water management in the upper Blue Nile basin, Ethiopia: the case of Koga large scale irrigation scheme. Agric. Water Manag. **170**, 26–35 (2016)
28. Rochette, P., Desjardins, R.L., Pattey, E.: Spatial and temporal variability of soil respiration in agricultural fields. Can. J. Soil Sci. **71**(2), 189–196 (1991)