Validation of a Low Complexity Machine Learning Discharge Predictive Model

Huma Zia[1], Nick Harris[1], Geoff Merrett[1], Mark Rivers[2]

[1]*Electronics and Computer Science, University of Southampton, Southampton, United Kingdom*
[2]*Institute of Agriculture, University of Western Australia, Australia*
[1]{hz2g11, nrh, gvm} @ ecs.soton.ac.uk, [2]mark.rivers@uwa.edu.au

ABSTRACT

This paper reports on the validation of a simplified discharge prediction model that is suitable for implementation on a resourced constrained system such as a wireless sensor network, which will allow their operation to become more proactive rather than reactive. The data-driven model, utilising an M5 decision tree modelling technique, is validated using a 12-month training data set derived from published measured data. Daily runoff and drainage is predicted, and the results are compared with existing data-driven models developed in this domain. Results for the model give an R^2 of 0.82 and Root Relative Mean Square Error (RRMSE) of 35.9%. 80% of the residuals for the predicted test values fall within a $\pm$2 mm discharge depth/day error range. The main significance is that the proposed model gives comparable results with fewer samples and simpler parameters when compared to previous published research, which offers the potential for implementation in resource constrained monitoring and control systems.

Keyword: Wireless Sensor Networks, discharge prediction, environmental modelling, machine learning, M5 trees

1. Introduction

Over recent years, wireless sensor networks (WSNs), with their attractions of low cost and real time data availability, have received considerable attention in automating agricultural processes for economic benefits, e.g. in precision irrigation, pest control, and animal farming. However, a research gap still exists for mechanizing reutilization of resources (water and nutrients) among farms in order to additionally maximise environmental benefits. There is huge potential for leveraging existing networked agricultural activities into an integrated mechanism by sharing information about discharges (Zia et al., 2013). To illustrate this consider that the most commonly used irrigation method, surface irrigation, results in 40 to 60% of water losses in the form of runoff (Eisenhauer, 2011, Tindula et al., 2013). This runoff can transport up to 30-50% of applied nutrients to stream water and rivers (Liu et al., 2003) In the light of these figures, the motivation for this work is to develop a system that can potentially reduce water consumption and reduce outflows from farms, by predicting and

Corresponding author: Huma Zia, + (971) 55773242, hz2g11@ecs.soton.ac.uk

monitoring discharge from local areas. This will enable the development of systems that can then proactively control irrigation strategies and also implement drainage reuse. This will also lead to improved water quality as it will allow nutrients to be kept in the place where they can be useful where previously they would have been discharged with no control into the local environment, eventually ending up in the streams and rivers. While drainage reuse has been advocated and adopted in farming (Adelman, 2000, Willardson et al., 1997, Harper, 2012), various resource constraints and farmer's concerns regarding real time availability of information on volumes, timings, and quality of discharges that will be delivered to the farms (Carr et al., 2011, Oster and Grattan, 2002), currently restricts wide adoption of this mechanism in agriculture.

To address some of these issues, we have previously proposed a framework for water quality monitoring control and management (WQMCM) using collaborative WSNs in a catchment to investigate and enable such a mechanism (Zia et al., 2014a). The basic system architecture comprises various modules, one of which is a discharge prediction module (*Q*-predictive model). The validation of this model using field data from an instrumented catchment is the subject of this paper. Although previous work on the *Q*-predictive model has shown that it works well with simulated data (Zia et al., 2014b), this paper extends this by reporting on the validation of the model with field data from an instrumented catchment, and comparing its performance with other published models.

To date, numerous physically-based hydrological models have been developed for the prediction of discharges, either measured as surface runoff, groundwater leaching or stream-flow. Although these models are popular in academic research and are very useful in evaluating different scenarios, their dependence on acquiring numerous parameters, the need for calibrating models to individual areas, and the tremendous computational burden involved in running the models makes wide-spread application complicated and difficult (Basha et al., 2008, Galelli and Castelletti, 2013). In contrast, data-driven models have good prediction capability and require fewer parameters, which is consistent with the requirement for a reduction in the computational burden of decision making (Castelletti et al., 2010). Thus data-driven modelling, using machine learning algorithms, has been widely used in hydrological modelling (Wilby et al., 2003, Rasouli et al., 2012, Solomatine and Ostfeld, 2008) with artificial neural networks (ANN) being a popular choice (Dawson and Wilby, 1998, Minns and Hall, 1996, Wilby et al., 2003). Recently, decision tree modelling has been investigated (Galelli and Castelletti, 2013, Villa-Vialaneix et al., 2012, Fortin et al., 2014, Piñeros Garcet et al., 2006, Kuzmanovski, 2012) and an interesting example of this class are M5 model trees (Quinlan, 1992). The advantage of M5 model trees over ANNs are that they are faster to train and have guaranteed convergence

(Solomatine and Dulal, 2003). However, there are two limitations in the existing work; either the existing models use simpler parameters but years of historical data with thousands of training samples to learn the heterogeneity of large areas (>1000ha) (Galelli and Castelletti, 2013, Solomatine and Xue, 2004a), or they use more complex models with a significant number of parameters (Bhattacharya et al., 2005, Kuzmanovski, 2012). Additionally none of these approaches have been specifically targeted at sensor network applications, and the data used was obtained through traditional sampling methods in gauged catchments.

This highlights one of the main issues in that the historical data sets needed to develop these predictive models do not exist for every farm, and even for most catchments. In addition, the strengths of a WSN deployment (fine spatial and temporal measurements of dynamic parameters) requires a simplified underlying physical model, and a simple machine learning model based on fewer and, ideally, real-time field parameters acquired autonomously and shareable across neighbouring farms. Thus there is a requirement for a discharge predictive model, which takes into account field conditions (soil moisture, vegetation cover) of the farms and the drainage networks, and which could be generated with adequate performance using fewer training samples. Such a model, once implemented in the network, can adaptively learn and further improve its accuracy over the course of time.

In this paper, we recap the model simplification for the predictive model for completeness, as already proposed by Zia *et.al.* (Zia et al., 2013), which is based on (but not restricted to) the popular National Resource Conservation Method (NRCS curve number model). Furthermore, we explore the applicability of M5 decision trees, for discharge modelling based on the proposed parameters. A year-long dataset (200 event samples) consisting of daily values for precipitation, field conditions (soil moisture, vegetation cover) and discharges, obtained from a grassland catchment in Ireland is used for training and testing the model. Specifically, an assessment procedure with the following steps is used (i) evaluation of optimized input parameter combinations with optimal performance; (ii) random sampling of the observational dataset to ensure a robust evaluation of the model performance, and the use of 10-fold cross validation to avoid over fitting of the model; (iii) assessment of the model performance against selected criteria; (iv) uncertainty analysis on the model residuals; and (v) comparative assessment of the prediction accuracy against other similar research developed using M5 decision trees.

2. Experimental Method

2.1 Specification of Catchment Data

The University of Cork carried out a study on the Dripsey catchment in the south of Ireland. The one-year study (2002) was aimed at understanding the underlying processes of nutrient loss from soil to water bodies within the catchment (Lewis, 2003) and thus fits the requirement for validating the *Q*-predictive model. This catchment consists of smaller nested sub-catchments. Figure 1 (a) shows the location of various data collection points in the stream network such as site1, site3 and site4, which collect water drained from their associated sub-catchments. For the development of the *Q*-predictive model, data available for site1 of the stream network is used. The sub-catchment which drains into this stream location is identified as 'catchment 1' (as shown in Figure 1 (a)) consisting of 17 ha of farmland. Precipitation (mm) and stream flow (mm) data, collected every 30 minutes for the year 2002 is used. The data is publically available for research and education purposes via the Environmental Protection Agency (EPA) website (Keily, 2003). The remainder of the data regarding field conditions is extracted from catchment descriptors available in the associated documentation (Lewis, 2003).

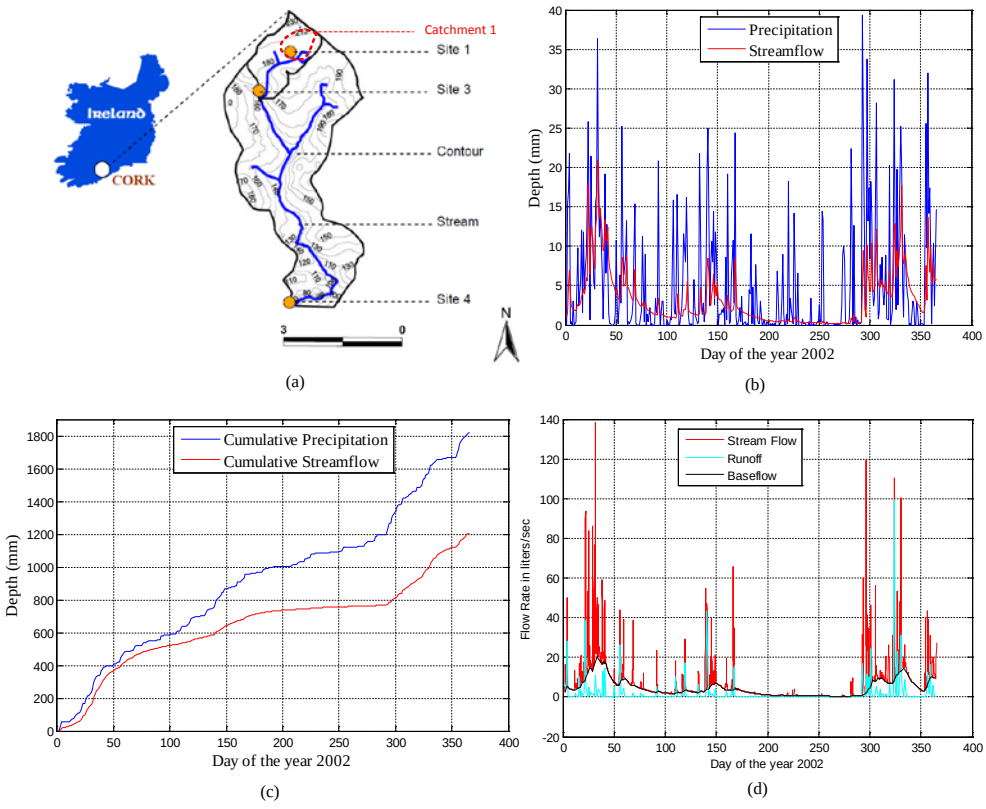


Figure 1: (a) Location and map of the Dripsey catchment (Khandokar, 2003); (b) precipitation and stream flow (mm) at site1; (c) cumulative precipitation and stream flow depth for site 1; (c) rate of stream flow, runoff, & base flow for site 1

For catchment 1, the cumulative rainfall for the year 2002 was 1812mm. The cumulative stream flow depth measured, at site1, was 1206mm of the rainfall (as shown in Figure 1 (c)). Stream flow here consists of water passing this point that originated as any surface runoff, sub-surface drainage or deeper groundwater contributions by catchment 1 (Khandokar, 2003). The monthly rainfall value ranges from less than 50 mm in the summer months to more than 250 mm in the winter months. The mean monthly temperature is 5^{o}C in the winter and 15^{o}C in the summer. The concentrations of total oxidised nitrate losses range from 0.5 to 6.5 mgl^{-1}. Land cover in the sub-catchment is dominated by agricultural grassland of high quality pasture and meadows. The growing season in Ireland is weather-dependant but generally summer-dominant, starting in early March and finishing in October. Grass is also cut as silage once or twice a year, typically at the end of May and at the end of July.

2.2 Modelling Technique – M5 decision Tree

With WSNs, it is now possible to obtain real time field data, which presents an opportunity for the development of simpler and more accurate data-driven models. These methods are based on the analysis of the data (of some simplified parameters) which characterises the system under study, thereby building models of physical processes. These models can complement or replace the knowledge-driven models describing behaviour of physical systems, and therefore can yield low computational complexity, making them well-suited for implementation on a resource constrained network.

As discussed in the introduction, decision tree modelling, specifically, is receiving increasing attention in the hydrological literature, in comparison to other learning models. Decision tree modelling is a method of approximating a target variable (output), with discrete values, from a given data set and represents the learned function in form of a decision tree (Mitchell, 1999), where each leaf contains the target values. Decision trees have been shown to perform well when compared to other model types (Galelli and Castelletti, 2013, Zhao and Zhang, 2008) but they do have one disadvantage. In decision trees, the predicted output is composed of discrete values and is reconstructed as a piecewise constant function. To ensure good prediction accuracy, the number of output classes (tree leaves) should be high; however, this increases the risk of over-fitting the observed data (Breiman, 1996). This can be resolved by replacing averaging in the tree leaves by fitting a linear regression function to the data and obtaining a continuous representation of the output (Galelli and Castelletti, 2013). This is known as M5 tree modelling, and was first introduced by Quinlan (Quinlan, 1992) and applied to hydrological modelling by Solomatine (Solomatine and Dulal, 2003, Solomatine and Xue, 2004b). The M5 tree is a piecewise linear model, so it takes an intermediate position between the linear models and truly nonlinear

models such as ANNs. Model trees have higher predictive accuracy and are able to make better predictions for values outside the training data range, when compared with regression trees (Kuzmanovski, 2012).

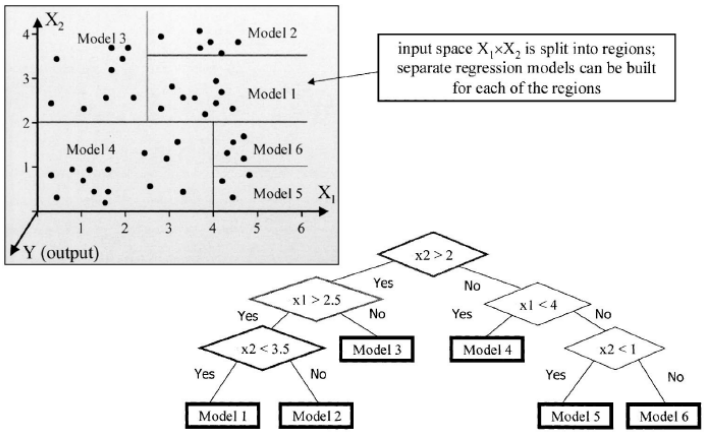


Figure 2: A generic M5 model tree, Models1-6 are linear regression models (Solomatine and Xue, 2004b)

The M5 model tree is a numerical prediction algorithm, and its splitting criterion is based on the standard deviation of the values in the subset of the training data that reaches a particular node. The construction of a model tree is similar to that of a decision tree. Figure 2 illustrates how the splitting of space is done in a generic M5 tree. Firstly, an initial tree is built and then is pruned (reduced) to overcome the over-fitting problem (that is a problem when a model fits the training data set very accurately, but fails on the test set). Finally, a smoothing process is employed to compensate for sharp discontinuities between adjacent linear models at the leaves of the pruned tree (this operation is not needed in building a decision tree) (Solomatine and Xue, 2004b). To implement the M5 model trees in this paper, MatLab toolbox M5PrimeLab (Jekabsons, 2010) is used.

2.3 Model Evaluation Criteria

The quantitative assessment of M5 tree modelling for prediction of *Q* based on the proposed model parameters is performed using a four-step procedure (as suggested in (Galelli and Castelletti, 2013)):

2.3.1 Random Sampling & Cross-Validation

To ensure a robust evaluation of the model performance, the data set was randomly partitioned into two groups: 75% of the observations were used for training the model while the remaining 25% are used for validation. When the available training data set is small, in order to overcome the problem of over-fitting (meaning the model fits the training data but not unseen test data) and reduce the sensitivity of the model to the selected training set, a cross-validation technique allows reliable model validation (Kohavi, 1995). Data is partitioned into subsets; one subset is used for training the model while the other is used for testing. Multiple rounds (folds)

of training and testing are performed using different partitions, and the validation results are averaged over the number of rounds. 10-fold is the most commonly used cross validation, which is used in this paper, where data is partitioned into 10 subsets.

2.3.2 Performance Measuring Parameters

To determine the performance (predictive accuracy) of the learned models, a multi-assessment criterion is used (Galelli and Castelletti, 2013). The performance parameters are:

- *Root mean square error (RMSE)*
- *Mean absolute error (MAE)*
- *Coefficient of determination (R^2)*
- *Root relative mean square error (RRMSE)*

RMSE and MAE can vary between 0 and infinity, with lower numbers indicating higher accuracy. R2 ranges between 1 and 0. It is equal to 1 if the predictions are perfect, i.e. a linear relationship exists between the predicted and measured values represented by a straight line. RRMSE is the ratio of the variance of the residuals to the variance of the target values themselves. Values of RRMSE can range between 0 and 1, where 0 means perfect forecasting. For comparing the performance of models developed using different datasets, R^2 and RRMSE are generally used. The definition of a good value for R^2 and RRMSE depends on the requirements of any specific application. For example, models developed for medical sciences generally need higher accuracy, whereas others might not. In this study, we compare these performance measures with those achieved by previous work in this area.

2.3.3 Uncertainty Analysis

As in any prediction there is a potential error which needs to be accounted for (Computing, 2004). The uncertainty of the predicted variable is investigated by the quantification of the residuals. Residuals represent the deviation of the predicted response from the observed or measured response obtained by subtracting the two. Since residuals are error, therefore, they are expected to be independently distributed. Ideally, the overall pattern of the residuals should be similar to the bell-shaped pattern observed when plotting a histogram of normally distributed values (Natrella, 2010). Through the analysis, we;

1. Find if residuals show any trend along the days of the year
2. Determine confidence intervals for residual errors

2.3.4 Comparative Assessment of *Q*-predictive model with other similar models

In order to evaluate the accuracy of the proposed *Q*-predictive model, we compare its performance with recent and relevant research efforts in the same area. Hydrologic models developed using M5 trees by *Solomatine et al.* (Solomatine and Xue, 2004b), *Kuzmanovski* (Kuzmanovski, 2012), *Corzo et al.* (Corzo et al., 2007) and *Galelli et al.* (Galelli and Castelletti, 2013), have been used for comparison. These works used various input parameters and thousands of training samples for developing the models. They were developed for daily or hourly predictions of discharges measured either as flow rates in a stream (for large catchments) or drainage volumes in field drainage collectors (for crop lands). For the first case, input parameters included precipitation, temperature and flow values for the previous several days. Whereas for the latter, parameters related to field condition were also included in the model. Most of these models did not use cross validation for performance evaluation. Furthermore, not all the performance metrics were used to show the performance accuracy of the models.

3. Development of *Q*-Predictive Model

3.1 Model Simplification

A simplified *Q*-predictive model was proposed by *Zia et.al.*(Zia et al., 2014b) based on the simplification of the most popular and simplest physically-based hydrological model: the NRCS method. This simplification is briefly explained here. For the prediction of discharges, the NRCS model uses rainfall, initial abstraction ratio (I_a) and curve number (*CN*) as input parameter where I_a is the volume of rainfall either retained in surface depressions or lost through evaporation or infiltration. *CN* is a coefficient reducing the total precipitation to runoff potential after surface absorption, and is computed considering the type of land use, land treatment, hydrological condition, hydrological soil group, and last 5-day rainfall (as a proxy for antecedent soil moisture condition). Although, the use of last 5-day rainfall to represent soil moisture conditions has been questioned (Vellidis et al., 2011) (Zia et al., 2014b), at the time the NRCS model was developed, proxy parameters and manual measurements were used to represent land conditions due to the absence of inexpensive sensing measures.

This simplification for the *Q*-predictive model is based on two steps: In the first stage, the transient parameters from the NRCS model parameters are selected because learning models are trained only on transient values. These parameters include rainfall value expected and land cover on the day of prediction as well as last 5-day rainfall value. In the second stage, the transient parameters are analysed for likely improvements made possible

by using WSNs. For example, methods such as field imaging and signal attenuation methods have been used to determine the plant biomass autonomously (Vellidis et al., 2011). This can be used to generate a crop stage (*CS*) parameter that can replace the land cover aspects of the NRCS model. Similarly, various applications have used sensors to monitor soil moisture conditions of the field for precision irrigation (Vellidis et al., 2008, Zia et al., 2013). Therefore, it has been proposed to use actual soil moisture values instead of the 5-day rainfall index. The model simplification is shown in Figure 3. The proposed model was developed by training it on simulated data and using M5 decision trees for learning. 10-fold cross validated results gave 90% accuracy when trained on 100 samples.

In order to validate the *Q*-predictive model further, it needs to be trained using real data acquired from a catchment. Although the increasing adoption of WSNs in agriculture means that it is now possible to extract real field conditions for some parameters, free and wide access to such long-term data required for model development is still not available. Therefore, for model validation, long-term soil moisture data was not available. Hence, we use the last-5-day rainfall value as a proxy. Nevertheless, this is the best available at present, and so offers a worst case performance baseline. When real, high frequency soil moisture readings become available to the model, performance should improve. The final mode includes precipitation (*P*), last-5-day rainfall (*L5PPT*) and crop stage (*CS*) data. This is still simpler than the NRCS model and also a comparable work (Kuzmanovski, 2012) for drainage discharge prediction which used M5 trees (This used 10 parameters). Later sections will demonstrate that the proposed model still gives comparable performance with the existing models developed for discharge prediction with the advantage of using far fewer training samples.

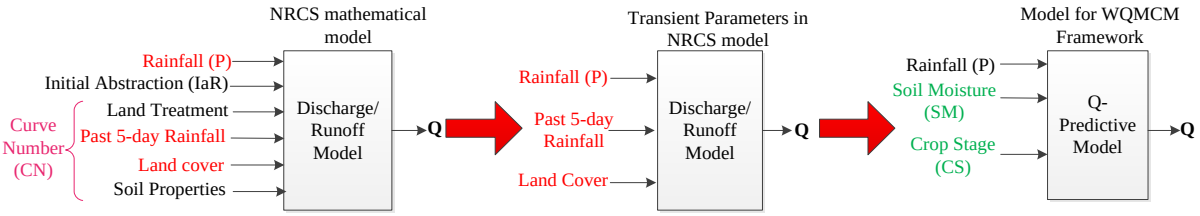


Figure 3: Simplification of model parameters for the *Q*-predictive model

3.2 Data Pre-processing & Sensitivity Analysis

The set of observations required for training the *Q*-predictive model was developed after implementing some pre-processing on the available dataset from the Dripsey catchment in Ireland (Keily). This is the only dataset that could be found with high temporal resolution data for an entire year. From the available dataset we used data related to 30 minute precipitation (mm) and stream flow ($lsec^{-1}$) for the year 2002. The remaining

parameters required for the *Q*-predictive model were either obtained using a proxy value or were extracted from the information available in the documentation for this study (Lewis, 2003).

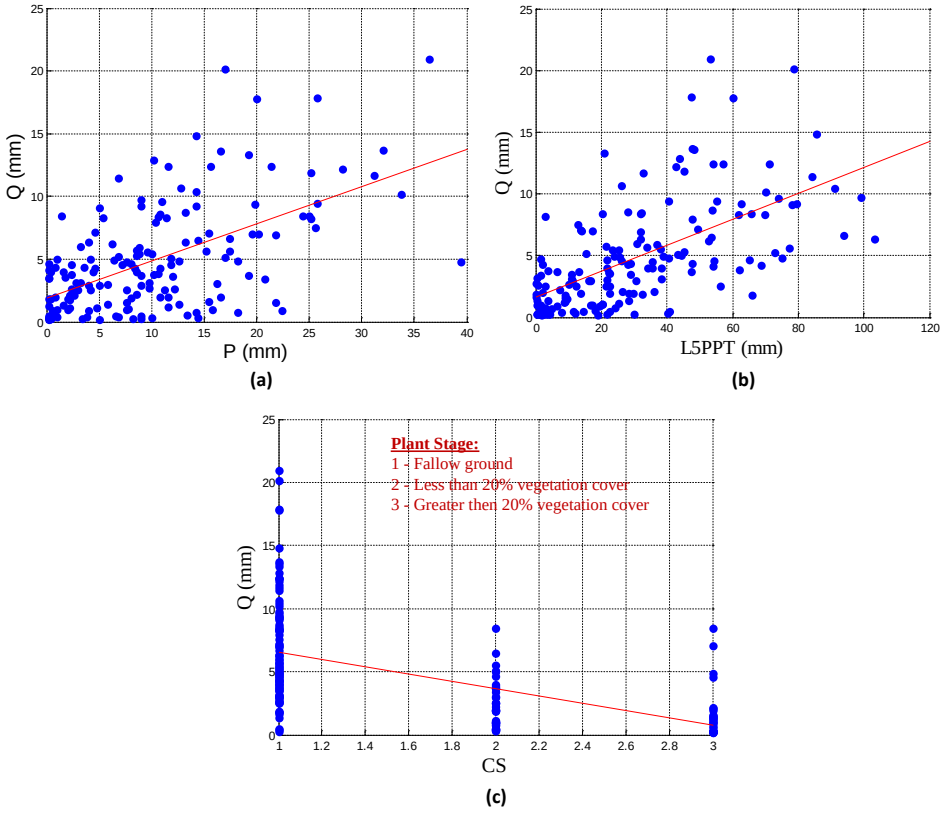


Figure 4: Sensitivity analysis of *Q* output with the input parameters, (a) *P*, (b) *L5PPT*, and (c) *CS*

Since the model is aimed at facilitating management decisions regarding drainage and nutrients reuse, we convert the 30-minute values of precipitation (*P*) into daily depths in mm. For each of the daily *P* values, the *L5PPT* value is computed by aggregating the depths of rainfall received in the last 5 days. For obtaining *CS* data, information regarding growing stages of grass in catchment 1 was assessed to obtain estimates for crop coverage throughout the year. According to crop coverage values, crop levels are assigned such that fallow land is referred to as stage 1, coverage less than 20% is termed as stage 2, and coverage greater than 20% is assigned stage 3.

For measures of output of the model, discharges from the catchment, runoff and stream flow rates were all available. We select stream flow as an output for our model for various reasons;

(i) runoff value is available as a single measurement per day, which does not provide complete information for the daily runoff depths,

(ii) due to the presence of high base flows, as shown in Figure 1 (c), stream flow better represents the discharges from the grass land

(iii) Pearson correlation coefficient (r) (which explains the strength and direction of the linear relationship between parameters) gives better values for stream flow as compared to runoff with the input parameters.

For stream flow, the values of 'r' for *P*, *L5PPT* and *CS* are 0.53, 0.32, and -0.38 respectively, while for runoff, the values of 'r' with *P*, *L5PPT* and *CS* are 0.52, 0.11, and -0.16 respectively. Positive 'r' values for *P* and *L5PPT* indicate that higher rainfall produces higher discharges. On the other hand, *CS* has a negative correlation, which indicates that higher vegetation cover absorbs more available water and also inhibits discharge flows.

Daily stream flow depth (mm), (*Q*) , is computed from 30 minute stream flow rates, as the model need daily depths for the decision making process. The values for 'r' show even stronger correlation of *Q* with the input parameters (0.57 for *P*, 0.61 for *L5PPT*, and -0.54 for *CS*). Figure 4 illustrate the sensitivity analysis of the input model parameters with *Q*. All the three plots indicate visible correlation with all input parameters.

4. Results and Discussion

In order to generate the *Q*-predictive model, training data of 200 event instances (75% of the total event dataset) are sampled randomly from the dataset which was generated as a result of the pre-processing discussed in the previous section. M5 decision trees were then used to generate the tree. The average value for *Q* in the training data is 3.80 mm, 25th percentile is 1.01, 75th percentile is 5.29, and 90th percentile is 9.22. The standard deviation of *Q* is 3.8.

Table 1: Performance measure for the *Q*-predictive model

	Performance Measure for Q-predictive Model			
	R^2	***MAE***	***RMSE***	***RRMSE (%)***
Without cross validation	0.87	0.81	1.34	35.9
10-fold cross validated	0.64	1.30	1.95	55.9

Performance parameters for the generated M5 decision tree indicate a high value for R^2 of 0.823 as listed in Table 1. In addition to this, the RMSE value for the model is 1.335. An accepted adequate value for RMSE in hydrological modelling is normally half of the standard deviation of the training data (Singh et al., 2005), which for this data is 1.9. The obtained RMSE of the trained model is well within this limit. Furthermore, the RRMSE

is 35.9%. Moreover, after 10-fold cross validation of the *Q*-predictive model, the results give R^2 as 0.63 and RRMSE as 55%. These results are comparable and in some cases even better than the performance of the existing models, which will be discussed in detail in the later section. However, when compared with our previous work on this model in Zia *et.al.*(Zia et al., 2014b), which was developed using simulated data (from NRCS simulator), R^2 was 0.99 and RRMSE was 7.5%. This is possibly because simulated data, though randomly sampled, is generated on the basis of an underlying mathematical model with obvious relationships between input and output, which is picked up by a learning model. Hence, models trained on simulated data, even if only few parameters of the actual mathematical model are selected, have higher predictive performance.

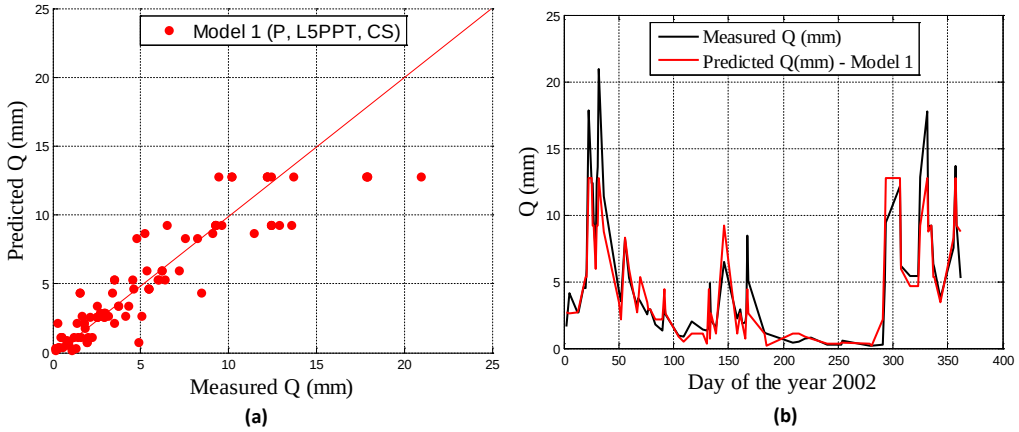


Figure 5: (a) Scatter plot of the predicted *Q* using the proposed model against measured *Q*, (b) Curve plot for predicted and measured *Q* plotted against days of the year

Furthermore, test data was used to predict *Q* values using the generated model. The predicted values are plotted against the known measured values and against days of the year in Figure 5. The model illustrates a good fit with R^2 of 0.868 as shown in Figure 5 (a). The predicted values for test data and measured values are also plotted against days of the year in Figure 5 (b) to represent the difference between the two curves. As is evident both curves almost overlap 50-60% of the time (covered more in section 4.2); however in the first 50 days of the year, the model seems to under-predict the stream values by about 25%. Also, it is evident in Figure 5 (a) that the prediction results flatten out for higher *Q* values. The reason is possibly because of the sparse training samples available above that threshold. Further work on verifying this with either more training samples or with classifying events with respect to their intensity (lower, medium and higher flows) would be able to explain these variations.

In the later section, we evaluate if subsets of the proposed model, i.e. any further simplification to the model parameters, impacts the model performance in any way. This would validate that using field condition

parameters in the proposed model is a better approach for prediction of outflows from smaller field plots, compared to relying only on the climatic conditions (Solomatine and Xue, 2004b, Corzo et al., 2007, Galelli and Castelletti, 2013).

4.1 Further Model Simplification – Viable or not?

In this section two models are developed using '*P* plus *L5PPT*' and only '*P*' as input parameters. These models are called 'model 2' and 'model 3' respectively. The full model previously discussed is referred to as 'model 1'. The performance parameters of the generated models are compared with the performance of the proposed model (as shown in Table 2). It is clear from the table that, model 1 show the best performance in comparison to the other models, although it is quite close to the performance of model 2. For instance, model 1 has 35.9% RRMSE value while model 2 has a 39.2% RRMSE value, although the R^2 values are almost similar. This shows that *L5PPT* (not included in model 2) is very significant for predicting outflows. Also, it is consistent with the correlation values of *L5PPT* with *Q* (0.609) as discussed in section 3.2. However, by using only *P* as an input in model 3, a very weak model is generated with 70% RRMSE value and 0.5 for R^2. This means that the predicted results using model 3 will have only 30% accuracy.

Table 2: Comparison of *Q*-predictive models developed using different combination of input parameters

Model No.	***Features for Q-predictive model***	***Performance Metrics***				***10-fold cross validated Performance Metrics***			
		R^2	***MAE***	***RMSE***	***RRMSE (%)***	R^2	***MAE***	***RMSE***	***RRMSE (%)***
1 (Proposed model)	*P*, L5PPT, CS	0.87	0.81	1.34	35.9	0.64	1.30	1.95	55.9
2	*P*, L5PPT	0.84	0.97	1.51	39.2	0.58	1.52	2.17	61.4
3	*P*	0.51	1.86	2.66	70.1	0.09	2.30	3.05	91.4

To illustrate the comparison among models further, test data is used to predict *Q* values for model 2 and model 3. The predicted values are plotted against the known measured values and against days of the year in Figure 6 and Figure 7. Model 2 show a good fit (as shown in Figure 6 (a) with R^2 of 0.844. The predicted values for test data and measured values are also plotted against days of the year in Figure 6 (b) to represent the difference between the two curves. Figure 6 (b) indicates that model 2 under-predicts during the initial 50 days

as well as last 50 days of the year during which highest discharges were observed (as shown in Figure 1 (b)). Furthermore, the predicted *Q* for test data using Model 3 is plotted in Figure 7 (a), which shows a poor fit, with an R^2 value of 0.505. This validates that reliance on only precipitation values, as an input for developing the *Q*-predictive model, will result in weak learning of the model hence poor prediction. Figure 7 (b) further illustrates a plot of the deviation of predicted *Q* compared to measured *Q*. The model seems to over-predict during summer and under-predict during winters leading to an unreliable system. To compare the prediction accuracy for these three models, we plot the results on a single graph as shown in Figure 8. This clearly illustrates the deviation of predicted results of these models from the measured results; model 1 has the closest similarity with the measured values as compared to other models. Hence, it is concluded that the proposed model parameters cannot be further simplified without degrading the performance.

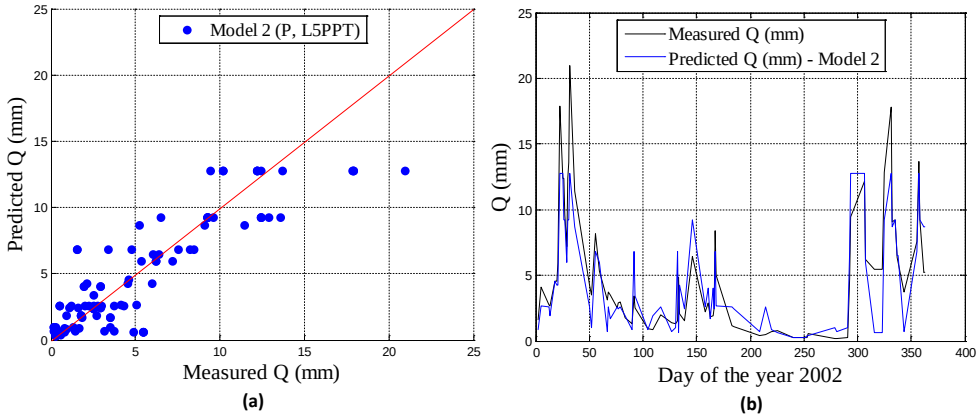


Figure 6: (a) Scatter plot of predicted *Q* using Model-2 against measured *Q*, (b) Curve plot for predicted and measured *Q* plotted against days of the year

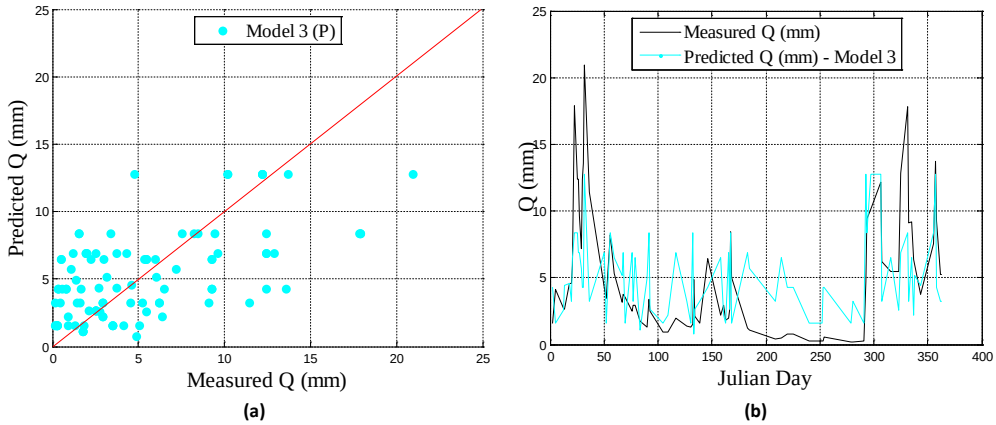


Figure 7: (a) Scatter plot of predicted *Q* using Model-3 against measured *Q*, (b) Curve plot for predicted and measured *Q* plotted against day of the year

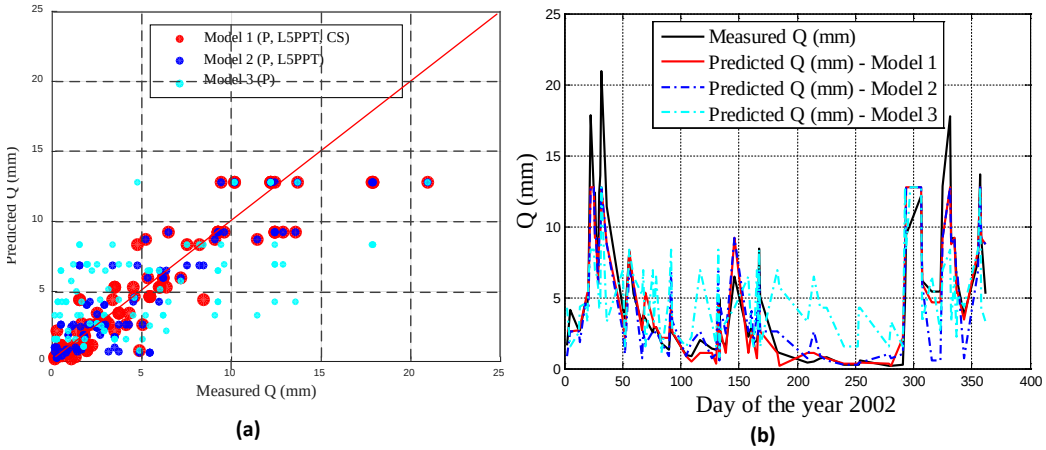


Figure 8: Comparison of predicted and measured *Q* for model 1, 2 & 3 in a (a) scatter plot and (b) curve plot

4.2 Uncertainty Analysis

To determine if any time dependency exists for the prediction error over summer and winter, residuals are plotted against day of the year. This provides us with a time-dependent confidence interval for the predicted values. Figure 9 (a) shows that 83% of the residuals for the predicted test values fall within a range of ±2 mm. A prediction error of this scale is not significant because this does not yield substantial outflows for this small catchment (17 ha). Therefore, incorrect estimates at this scale will not adversely impact decision making based on modelled results. Furthermore, as already pointed out in the earlier discussion of the results, there is a seasonal variation in the predicted data. This is linked to the performance of the model for predicting high *Q* events, and these high *Q* events tend to occur in winter, thus concentrating the uncertainty in this period. This is a feature of the current model structure that will be investigated in more detail in future work, but is currently limited by the availability of real data.

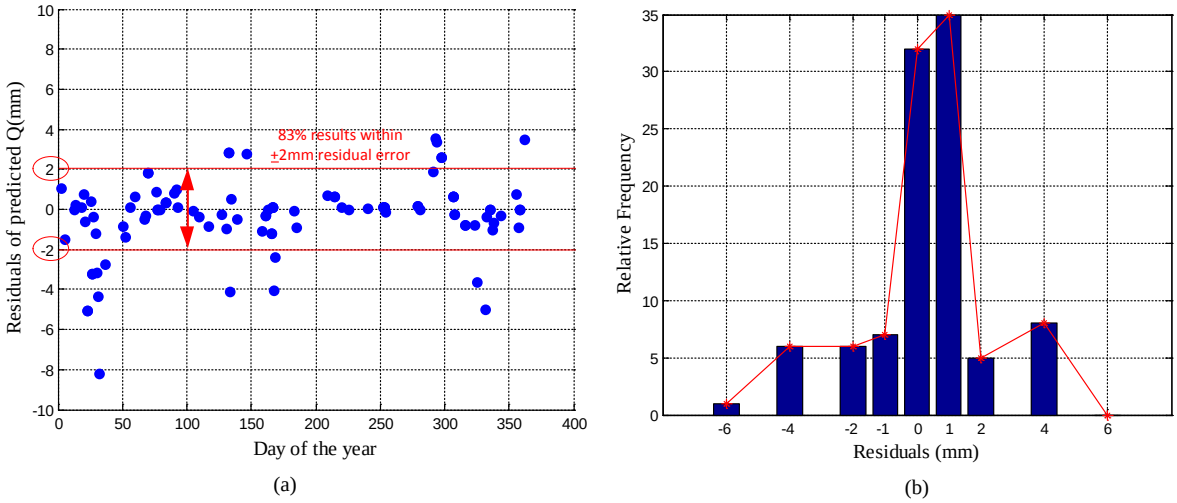
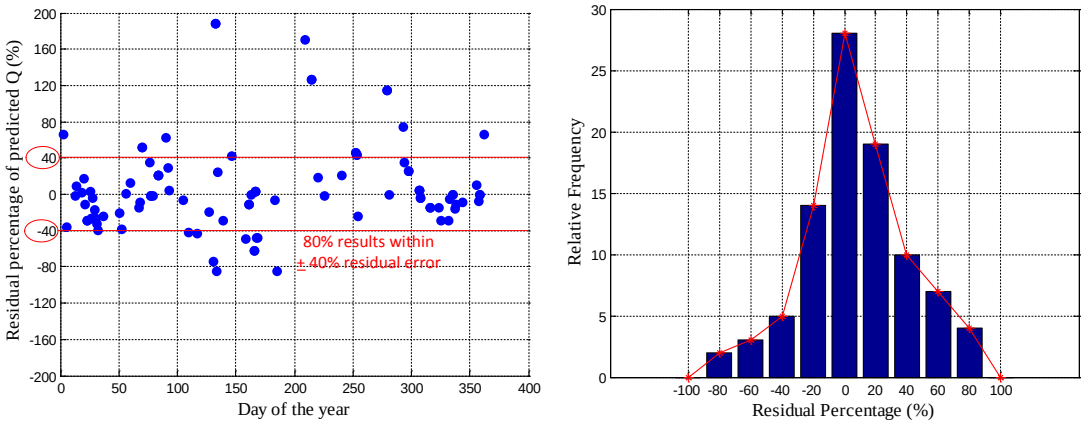


Figure 9: (a) Trend of residual error of predicted *Q*, (b) relative frequency of the residual error

A frequency plot for the residual error illustrates an approximately normal distribution of residuals produced by the model with highest frequency corresponding to 0 and 1 mm errors (Figure 9 (b)). To explain this further, Figure 10 (a) shows the residuals percentile of predicted *Q* against days of the year. This illustrates that 80% of the predicted values are within ±40% residual errors. A histogram for this shows a normal distribution having maximum number of values with 0% residual error (Figure 10 (b)).



(a) (b)

Figure 10: (a) Trend of percentile value of residual error versus, (b) relative frequency of per cent residual error

5. Comparative Assessment – *Q*-Predictive Model versus similar models

In order to evaluate if the proposed model has acceptable (or comparable) performance, we compare its results with the results of the most relevant existing models. For this, work by *Kuzmanovski* (Kuzmanovski, 2012), *Solomatine et al.* (Solomatine and Xue, 2004b), *Corzo et al.*(Corzo et al., 2007) and *Galelli et.al.* (Galelli and Castelletti, 2013) has been selected. All of these models use M5 decision trees. Although, all of these works were aimed at predicting discharges either in the form of drainage from a small field plot, or flow volumes and rates in a river or stream drained by large catchments, it is important to note that none of these works are entirely similar, in objective and methodology with the proposed *Q*-predictive model. Table 3 lists the experimental details and performance metrics for all the models including the *Q*-predictive model. The model listed first in the table is the *Q*-predictive model proposed in this paper.

Overall, it is clear that the *Q*-predictive model has a comparable performance compared to other models. More specifically, work done by *Kuzmanovski* (Kuzmanovski, 2012) is closer in objective to the *Q*-predictive model, hence provides better comparison. In this work, the aim was to measure drainage discharges from fields, ranging from 0.83 ha to 1ha, in order to control pollutant outflows. The model uses 10 parameters related to crop stage, day of the season, slope of the field, rainfall, temperature, runoff, drainage etc., and used 22 years of

daily data to train the model. Without cross validation, the performance measures for the models developed for various field plots was calculated; R^2 ranges between 0.75 and 0.89, and RMSE is between 45% and 65.9%. The performance (without cross validation) of the *Q*-predictive model is acceptable when compared with the best performing model by *Kuzmanovski*, although the former model is developed with only 3 input parameters. The reason for good performance of the *Q*-predictive model can be possibly attributed to the use of the most relevant parameters, especially the 5-day rainfall value. It is believed that results would further improve if actual soil moisture measurements are used. However, the models developed by *Kuzmanovski* use only 2 – day rainfall value, which may not accurately represent soil moisture conditions. The performance of the *Q*-predictive model would need to be validated with more data samples.

The models developed by *Solomatine et al. Corzo et al*. and *Galelli et al*., were all predicting discharges in a river and stream draining very large catchments, and hence used only 2 to 3 parameters related to climate and flow. This is because field conditions (soil moisture, vegetation cover) for such large heterogeneous catchments can vary tremendously; hence a single average value may not represent the field conditions for the whole catchment. Furthermore, these models are developed for hourly predictions; field conditions do not change in an hourly manner, however previous flow and rainfall intensity do. Hence, these models rely only on climatic and flow parameters for predictions and use thousands of samples so that the model could learn all possible samples over many years (decades in some cases). These models have slightly better performance when compared with the *Q*-predictive model for obvious difference in the training set size. For instance, the R^2 values of the model proposed by *Solomatine et al.* and *Corzo et al.* are 0.97 and 0.89 respectively as compared to the 0.86 for the *Q*-predictive model. Similarly, RRMSE of the model proposed by *Galelli et al.* is 48% as compared to 55% RRMSE of the *Q*-predictive model.

The proposed *Q*-predictive model is between the models developed by *Kuzmanovski* (Kuzmanovski, 2012) and the others (Solomatine and Xue, 2004b, Corzo et al., 2007, Galelli and Castelletti, 2013) in terms of the number of parameters used. The *Q*-predictive model has climatic conditions as used in the latter models. However it also uses field conditions but keeps the number of parameters simpler and fewer compared to *Kuzmanovski*.

Table 3: Performance comparison of *Q*-predictive models with the existing models developed using M5 trees

S. No.	Predictive	Output	No of Input	No of	Drainage	Cross	RRMSE	R2

	Models	**Value**	**variables**	**Training samples**	**area**	**Validation for training**	**(%)**	
1	Proposed *Q*-predictive Model	Daily Discharges from a farm	3 (rainfall, last 5-day rainfall, crop stage)	200 samples (2002)	17 ha	10- fold	55	0.63
						Not done	35	0.86
2	*Kuzmanovski* (Kuzmanovski, 2012)	Daily Drainage volume from farms	10 parameters	7000 samples (22 years data, 1987-2011)	0.83-1 ha field plots	Not done	45 - 65.9	0.75-0.89
3	*Solomatine et al*. (Solomatine and Xue, 2004b)	Flood discharge	3 parameters (11 sub values) Rainfall, Past rainfalls, Discharges	5000 samples (> 13 years data, 1976-1989)	106 ha	Not done	-	0.97
3	*Corzo et al.* (Corzo et al., 2007)	Hourly discharge in a stream	Previous runoff, effective rainfall	2000 (8 years data, 1988-1996)	37,000 ha	Not done	-	0.89
4	*Galelli et al.* (Galelli and Castelletti, 2013)	Hourly Stream flow	2 (rainfall & inflow value), 3 time lag sets	24120 (2.5 years data, 2009-2011)	10,000 ha	10-fold	48	-

Conclusions

In this paper, we have successfully validated a discharge (*Q*) predictive model for the proposed simplified parameters by employing M5 decision tree learning algorithms. The input parameters included daily precipitation, vegetation cover, and last-5-day rainfall value. The model is assessed for its prediction accuracy in comparison to major data-driven hydrological models. The *Q*-predictive model was evaluated by training it on a

year-long discharge dataset measured for a sub-catchment in Ireland. The significance of the proposed *Q*-predictive model is in the fact that it uses just a year-long training sample set which includes climatic as well as field parameters to develop accurate predictive model, and that it can adequately predict flows for smaller sub-catchments with simpler parameters and acceptable (comparable) prediction accuracy.

Results for the *Q*-predictive model show that:

i) Performance measures for the proposed *Q*-predictive model provide good performance; R^2 is 0.823, RMSE is 1.33, and RRMSE value is 35.9%. 10-fold cross-validated results yield R^2 of 0.63, RMSE of 1.95, and RRMSE of 55.7%.

ii) 83% of the residuals for the predicted test values fall within +2 mm range. It has been argued here that prediction errors of this scale are not significant, and a plot of the residuals percentile values illustrates that 80% of the predicted values are within ±40% residual errors.

iii) Investigation for further simplification of model parameters results in poor performance. 10-fold cross validated results for model developed using rainfall and last-5-day rainfall value as input parameters (and omitting the crop cover) result in RRMSE of 61.4%. For the model developed using only rainfall as the input parameter, RRMSE is 91.4%.

iv) In comparison to the existing models developed for discharge predictions using M5 decision trees, the proposed model presents great promise and comparable performance by only using a year-long dataset and simplifying the model parameters tremendously for small-scale land areas.

References

ADELMAN, D. D. 2000. Simulation of irrigation reuse system nitrate losses and potential corn yield reductions. *Environmental Science & Policy*, 3**,** 213-217.

BASHA, E. A., RAVELA, S. & RUS, D. Model-based monitoring for early warning flood detection. 6th ACM Conference onEmbedded network sensor systems 2008. ACM, 295-308.

BHATTACHARYA, B., PRICE, R. & SOLOMATINE, D. 2005. Data-driven modelling in the context of sediment transport. *Physics and Chemistry of the Earth, Parts A/B/C*, 30**,** 297-302.

BREIMAN, L. 1996. Bagging predictors. *Machine learning*, 24**,** 123-140.

CARR, G., POTTER, R. B. & NORTCLIFF, S. 2011. Water reuse for irrigation in Jordan: Perceptions of water quality among farmers. *Agricultural Water Management*, 98**,** 847-854.

CASTELLETTI, A., GALELLI, S., RESTELLI, M. & SONCINI-SESSA, R. 2010. Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, 46.

COMPUTING, W. N. 2004. Model-Independent Parameter Estimation. *User Manual:*.

CORZO, G., SIEK, M., SOLOMATINE, D. & PRICE, R. Modular data-driven hydrologic models with incorporated knowledge: neural networks and model trees. International congress IAHR, ASCE, Italy, 2007.

DAWSON, C. W. & WILBY, R. 1998. An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, 43**,** 47-66.

EISENHAUER, D. E. 2011. Irrigation Efficiency and Uniformity, and Crop Water Use Efficiency. The Board of Regents of the University of Nebraska - Lincoln Extension.

FORTIN, J., MORAIS, A., ANCTIL, F. & PARENT, L. 2014. Comparison of Machine Learning Regression Methods to Simulate NO 3 Flux in Soil Solution under Potato Crops. *Journal of Applied Mathematics*, 5**,** 832-841.

GALELLI, S. & CASTELLETTI, A. 2013. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrology & Earth System Sciences Discussions,* 10.

HARPER, H. H. 2012. Impacts of Reuse Irrigation on Nutrient Loadings and Transport in Urbanized Drainage Basins. *Florida Stormwater Association Annual Meeting.* Environmental Research & Design, Inc.

JEKABSONS, G. 2010. *M5PrimeLab: M5' regression tree and model tree toolbox for Matlab* [Online]. Institute of Applied Computer Systems Riga Technical University, Latvia. Available: http://www.cs.rtu.lv/jekabsons/Files/M5PrimeLab.pdf.

KEILY, G. 2003. *Phosphorus, Nitrogen and Suspended Sediment loss from Soil to Water from Agricultural Grassland* [Online]. Environmental Protection Agency (EPA). Available: http://erc.epa.ie/safer/resource?id=ad1f3acf-5035-102a-90c6-0593d266866d 2014].

KHANDOKAR, F. 2003. *Phosphorus in soils of pasture farms subject to chemical and slurry fertilization.* NUI, 2003 at Department of Civil and Environmental Engineering, UCC.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, 1995. 1137-1145.

KUZMANOVSKI, V. 2012. Integration of expert knowledge and predictive learning: Modelling water flows in agriculture - MS Thesis. Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

LEWIS, C. 2003. *Phosphorus, nitrogen and suspended sediment loss from soil to water from agricultural grassland.* Department of Civil and Environmental Engineering, University College Cork.

LIU, X., JU, X., ZHANG, F., PAN, J. & CHRISTIE, P. 2003. Nitrogen dynamics and budgets in a winter wheat-maize cropping system in the North China Plain. *Field Crops Research,* 83**,** 111-124.

MINNS, A. & HALL, M. 1996. Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal,* 41**,** 399-417.

MITCHELL, T. M. 1999. Machine learning and data mining. *Communications of the ACM,* 42**,** 30-36.

NATRELLA, M. 2010. *NIST/SEMATECH: e-Handbook of Statistical Methods,* Available online: http://www. itl. nist. gov/div898/handbook.

OSTER, J. & GRATTAN, S. 2002. Drainage water reuse. *Irrigation and Drainage Systems,* 16**,** 297-310.

PIÑEROS GARCET, J. D., ORDOÑEZ, A., ROOSEN, J. & VANCLOOSTER, M. 2006. Metamodelling: Theory, concepts and application to nitrate leaching modelling. *Ecological modelling,* 193**,** 629-644.

QUINLAN, J. R. Learning with continuous classes. 5th Australian joint Conference on Artificial Intelligence, 1992. Singapore, 343-348.

RASOULI, K., HSIEH, W. W. & CANNON, A. J. 2012. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology,* 414–415**,** 284-293.

SINGH, J., KNAPP, H. V., ARNOLD, J. & DEMISSIE, M. 2005. Hydrological modeling of the iroquois river watershed using HSPF and SWAT1. *JAWRA Journal of the American Water Resources Association,* 41**,** 343-360.

SOLOMATINE, D. & OSTFELD, A. 2008. Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics,* 10**,** 3-22.

SOLOMATINE, D. & XUE, Y. 2004a. M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *Journal of Hydrologic Engineering,* 9**,** 491-501.

SOLOMATINE, D. P. & DULAL, K. N. 2003. Model trees as an alternative to neural networks in rainfall—runoff modelling. *Hydrological Sciences Journal,* 48**,** 399-411.

SOLOMATINE, D. P. & XUE, Y. 2004b. M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering,* 9**,** 491-501.

TINDULA, G., ORANG, M. & SNYDER, R. 2013. Survey of Irrigation Methods in California in 2010. *Journal of Irrigation and Drainage Engineering,* 139**,** 233-238.

VELLIDIS, G., SAVELLE, H., RITCHIE, R., HARRIS, G., HILL, R. & HENRY, H. 2011. NDVI response of cotton to nitrogen application rates in Georgia. *Precision Agriculture***,** 359.

VELLIDIS, G., TUCKER, M., PERRY, C., KVIEN, C. & BEDNARZ, C. 2008. A real-time wireless smart sensor array for scheduling irrigation. *Computers and Electronics in Agriculture,* 61**,** 44-50.

VILLA-VIALANEIX, N., FOLLADOR, M., RATTO, M. & LEIP, A. 2012. A comparison of eight metamodeling techniques for the simulation of N2O fluxes and N leaching from corn crops. *Environmental Modelling & Software,* 34**,** 51-66.

WILBY, R., ABRAHART, R. & DAWSON, C. 2003. Detection of conceptual model rainfall—runoff processes inside an artificial neural network. *Hydrological Sciences Journal,* 48**,** 163-181.

WILLARDSON, L., BOELS, D. & SMEDEMA, L. 1997. Reuse of drainage water from irrigated areas. *Irrigation and Drainage Systems,* 11**,** 215-239.

ZHAO, Y. & ZHANG, Y. 2008. Comparison of decision tree methods for finding active objects. *Advances in Space Research,* 41**,** 1955-1959.

ZIA, H., HARRIS, N. R. & MERRETT, G. V. Empirical Modelling and Simulation for Discharge Dynamics Enabling Catchment-Scale Water Quality Management. The 26th European Modeling & Simulation Symposium, 2014a Bordeaux, France.

ZIA, H., HARRIS, N. R. & MERRETT, G. V. Water Quality Monitoring, Control and Management (WQMCM) Framework using Collaborative Wireless Sensor Networks. 11th International Conference on Hydroinformatics 2014b New York City, USA.

ZIA, H., HARRIS, N. R., MERRETT, G. V., RIVERS, M. & COLES, N. 2013. The impact of agricultural activities on water quality: A case for collaborative catchment-scale management using integrated wireless sensor networks. *Computers and Electronics in Agriculture,* 96**,** 126-138.