11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

# WATER QUALITY MONITORING, CONTROL AND MANAGEMENT (WQMCM) FRAMEWORK USING COLLABORATIVE WIRELESS SENSOR NETWORKS

HUMA ZIA, NICK R. HARRIS, GEOFF V. MERRETT
*Electronics and Computer Science, University Of Southampton, Southampton, United Kingdom*

## ABSTRACT

Improving water quality is of global concern, with agricultural practices being the major contributors to reduced water quality. The reuse of nutrient-rich drainage water can be a valuable strategy to gain economic-environmental benefits. However, currently the tools and techniques to allow this do not exist. Therefore, we have proposed a framework, WQMCM, which utilises increasingly common local farm-scale networks across a catchment, adding provision for collaborative information sharing. Using this framework, individual sub-networks can learn their environment and predict the impact of catchment events on their locality, allowing dynamic decision making for local irrigation strategies. Since resource constraints of network nodes (e.g. power consumption, computing power etc.) require a simplified predictive model for discharges, therefore low-dimensional model parameters are derived from the existing National Resource Conservation Method (NRCS), utilising real-time field values. Evaluation of the predictive models, developed using M5 decision trees, demonstrates accuracy of 84-94% compared with the traditional NRCS curve number model. The discharge volume and response time model was tested to perform with 6% relative root mean square error (RRMSE), even for a small training set of around 100 samples; however the discharge response time model required a minimum of 300 training samples to show reasonable performance with 16% RRMSE

## INTRODUCTION

Water quality degradation in a catchment is mainly attributed to outdated agricultural practices. Excessive or poorly timed application of irrigation water and fertilizer result in nutrient fluxes into the water system with main issues being due to phosphorous (P) and nitrogen (N) losses (EPA [1]). In addition, the inherent inefficiency of nutrient uptake by crops (up to 50% for N and 10% uptake for P) renders nutrient outflows inevitable. This implies that adopting a reutilization mechanism of drainage and nutrients within the farm system can prove to be a valuable strategy to manage these outflows before they end up in rivers (Harper [2]). However, it is challenging to make valid predictions about these outflows (what and when to expect).

Over recent years, wireless sensor networks (WSNs) have received considerable attention in precision agriculture, due to their low cost and real time data availability. It is believed that, despite their limitations, there is huge potential for leveraging existing networked agricultural activities into an integrated mechanism by sharing information about discharges and predicting their impact (Zia *et al.* [3]). However, there is no framework to investigate and implement such a mechanism. The authors have proposed a framework, WQMCM, which utilizes collaboration among networks in a catchment to investigate and enable such a mechanism (Zia *et al.* [4]). The basic architecture comprises modules to enable individual networks to learn its environment by correlating neighbor's events with events within their own zone, predict their impact in terms of discharges and runoffs and then adapt the local monitoring and management strategy. This paper focuses on the development and evaluation of the discharge prediction model.

For the prediction of discharges, various physical and mathematical hydrological models have been developed. Although popular in research, their dependence on acquiring numerous input parameters, the need for calibration, and the tremendous computational burden involved in running the models makes wide-spread application complicated and difficult for sensor networks (Basha *et al.* [5]). Furthermore, for implementing the WQMCM framework, constraints are associated with the practicality of information sharing among neighbors and the transmission costs linked with sharing high-dimensional input parameters for the predictive models. Therefore, a computing model running on WSNs for the WQMCM framework, requires a simplified underlying physical model, based on fewer and, ideally, real-time field parameters acquired autonomously. In this respect, data-driven techniques based on machine learning, are becoming increasingly popular in hydrological modelling (Solomatine *et al.* [6]), and feature low computational complexity. In this paper, we use a popular NRCS curve number model as a basis for deriving and evaluating simplified model parameters. We use an M5 decision tree machine learning algorithm to generate the predictive models based on proposed parameters. The effect of different feature sets and training sizes, on the prediction performance of the models, are evaluated and discussed.

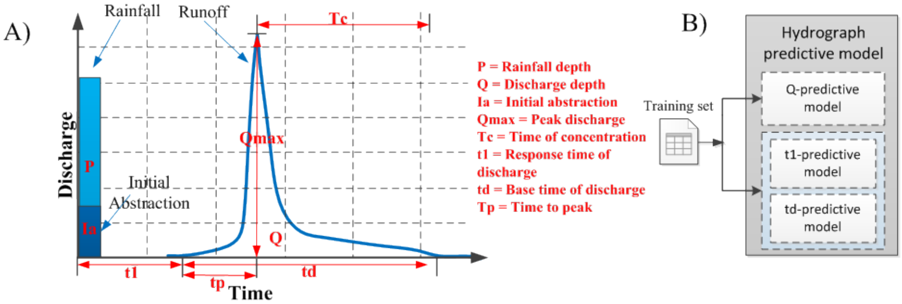

Figure 1: A) A discharge hydrograph, B) Predictive model for hydrograph dynamics

## MODEL SIMPLIFICATION FOR A DISCHARGE PREDICTIVE MODEL

A runoff or drainage discharge is represented using a hydrograph as shown in Figure 1A. For the WQMCM framework, the parameters of interest for discharge dynamics are ‘*Q*’, ‘*t1*‚ and ‘$t_d$’. These parameters provide information about the depth and timing of the expected discharges. Individual learning models are developed to obtain ‘*Q*’, ‘*t1*‚ and ‘$t_d$’, as shown in Figure 1B, for which we first derive the model parameters in the following.

### Mathematical Model for ‘*Q*’

One of the most popular methods to estimate the volume of surface runoff for a given rainfall event, is the NRCS Curve Number method (Hawkins *et al.* [7]). Using this method, *Q* is computed as follows;

$$Q = \frac{\left[P - 0.2\left(\frac{1000}{CN} - 10\right)\right]^2}{P + 0.8\left(\frac{1000}{CN} - 10\right)} \quad (1)$$

Where, *P* is the rainfall depth and *CN* is a coefficient reducing the total precipitation to runoff potential after surface absorption (with values in the range 0-100). The higher the *CN* coefficient, the higher is the runoff potential. It is computed considering the type of land use, land treatment, hydrological condition, hydrological soil group, and antecedent soil moisture condition (AMC). The volume of rainfall either retained in surface depressions or lost through evaporation or infiltration, termed as the initial abstraction (Ia), is assumed to be 20% of the potential soil moisture retention (Hawkins *et al.* [7]).

**Mathematical Model for '$t_1$' and '$t_d$'**

As evident from Figure 1A, $t_d$ is expressed as;

$$t_d = T_c + t_p \quad (2)$$

Where, $T_c$ is time for runoff to travel from the furthest distance in the watershed to the location where *Q* is to be determined, and $t_p$ is the time to peak discharge. Typically there are three distinct runoff patterns in a watershed such as sheet flow, shallow concentrated flow, and channel flow. Numerical equations based on the underlying physical model are described below.

$$T_\mathrm{c} = \frac{0.007(nL)^{0.8}}{(P_2)^{0.5}(S)^{0.4}} + \frac{L}{3600V} + \frac{L}{3600}\left(\frac{n}{1.49(R)^{\frac{2}{3}}(s)^{0.5}}\right) \quad (3)$$

Where, *L* is length (ft.) of flow pattern, n represents land cover, $P_2$ is 2-year return period 24 hour precipitation (in.) for a region, *R* is hydraulic radius (ft.), s is average ground slope (ft.-vertical/ft.-horizontal), $T_t$ is travel time (hr.), and *V* is average velocity (ft./s) of water.

As per the author's best knowledge, there is no direct mathematical equation to express *tp* in the NRCS method. The other parameter required is $t_1$, and once again there is no mathematical expression for this. However, both are extracted from hydrograph plots drawn using the convolution of incremental runoff depth and unit hydrograph flow rate for a specific region. The unit hydrograph is a hypothetical unit response of a watershed (in terms of runoff volume and timing) to a unit input of rainfall. It is specific to a particular watershed, rainfall distribution (*RD*), and rainfall duration ($P_d$) such as 1-hour, 6-hour, or 24-hour (Shaw *et al.* [8]).

**Limitations in Mathematical Model:**

The NRCS method, although simpler than the other models, still presents a challenge of acquiring a variety of permanent and transient parameters for every field under observation to determine discharge dynamics (Eq. (1) and Eq. (3)). Under the WQMCM framework paradigm, sharing these parameters among networks is not practical as it incurs high transmission costs resulting in low battery life of the deployed sensors. Moreover, at the time the NRCS method was developed, due to the absence of remote and inexpensive sensing measures, proxy parameters, average values or manual observations were used to represent land conditions. An example is AMC, which is used to determine CN. This is represented by using the amount of rainfall received in the five days preceding the storm event, which is a subjective judgment, instead of a physical reality (Fennessey *et al.* [9]). In addition, type and extent of land cover, slope and land treatment etc., is determined by manual observation of the field, which limits autonomous monitoring and renders result prone to error. Furthermore, determining $t_1$ and $t_d$ is computationally intensive. This implies that low-dimensional model parameters are required which should take into account real time field conditions in an autonomous manner.

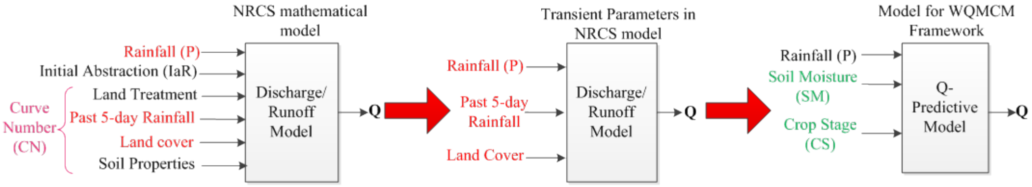

Figure 2: Model simplification for a *Q*-predictive model

### Model Simplification for *Q*, $t_1$ and $t_d$

During the last decade the area of empirical modelling received an important boost due to developments in the area of WSNs and machine learning. It is anticipated that learning models yield low computational complexity. Here, the authors derive a simplified model based on the NRCS model. This simplification is based on two steps; firstly the transient parameters from the NRCS model parameters are selected for each of the predictive models for *Q*, $t_1$ and $t_d$. This is because learning models are trained only on transient values. After this, the transient parameters are analyzed for likely improvements made possible by using WSNs.

For *Q*, model simplification is as shown in Figure 2. The transient parameters in the NRCS model are rainfall depth, past 5-day rainfall and land cover. With increasing adoption of WSNs in agriculture, it is more practical to use this technology to extract real field conditions for prediction. For example, methods such as field imaging and signal attenuation methods have been used to determine the plant biomass autonomously (Vellidis *et al.* [10]). This can be interpreted into the crop stage. Similarly, various applications have used sensors to monitor soil moisture conditions of the field for precision irrigation (Zia *et al.*, Vellidis *et al.* [3, 11]). Therefore, it is proposed to use actual soil moisture values instead of the 5-day rainfall index.

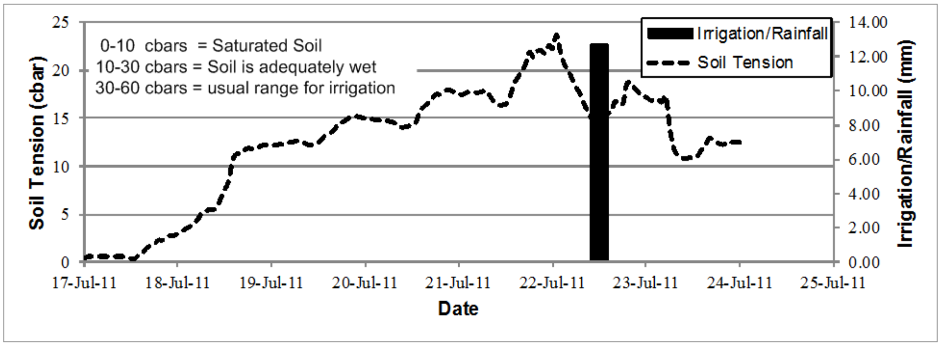

Figure 3: Soil moisture conditions in response to irrigation events in a field

In order to validate the limitation of 5-day rainfall index to represent AMC, we have analyzed season long data observed in a precision irrigation application, supplied by the University of Georgia (Vellidis *et al.* [11] ). The analysis shows that in many cases the soil moisture condition was measured as moderate, although the field did not receive any rainfall or irrigation in the last 5 days. Figure 3 plots a week long data of measured soil tension (represents soil moisture). Using the 5-day rainfall index, on 22nd July, dry soil conditions would be estimated, due to the fact that there was no rain in the preceding 5 days. However, the actual soil condition is measured as adequately saturated by the sensors. This leads to incorrect determination of drainage after a rainfall or irrigation. Therefore, rainfall, soil moisture and crop stage are proposed as the simplified model parameters for the prediction of *Q* (Figure 2).

As already discussed, for $t_1$ and $t_d$, the mathematical model and convolution method requires various parameters and historical data (Figure 4). Firstly the transient parameters are selected which include rainfall duration ($P_d$), rainfall ($P$), surface roughness ($n$) and 2-year average rainfall ($P_2$). This is further corroborated by analyzing an extensive set of simulated data (using NRCS based simulator, (Davis [12])) for which a routine in Matlab was written to extract $t_1$ and $t_d$. The data indicated strong correlation of the selected transient parameters with $t_1$ and $t_d$. This is because higher surface roughness inhibits flow rate and increases travel times. It is proposed in this thesis that crop stage may well represent the field roughness. Furthermore, instead of relying on historical data for estimating $P_2$ and *RD* for every region, it is proposed to use actual soil moisture conditions. Simulation results can be used to evaluate the effect of this substitution on prediction accuracy of $t_1$ and $t_d$.

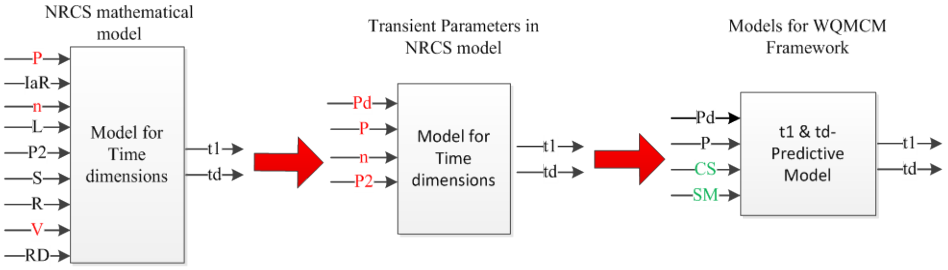
NRCS mathematical model
P
IaR
n
L
P2
S
R
V
RD
Model for Time dimensions
t1
td
Transient Parameters in NRCS model
Pd
P
n
P2
Model for Time dimensions
t1
td
Models for WQMCM Framework
Pd
P
CS
SM
t1 & td-Predictive Model
t1
td

Figure 4: Model simplification for a $t_1$ and $t_d$-predictive models

## SIMULATION AND RESULTS FOR HYDROGRAPH PREDICTVE MODELS

Using machine learning algorithms, the models are trained on the historical data describing the phenomenon in question. Historical data includes known samples that are combinations of inputs and corresponding outputs. The learned model is then used to predict the outputs from the new input values. Here historical data is generated using a simulator, developed in Matlab, for a combination of various event depths and field conditions, which is based on the NRCS method (Davis [12]). The obtained data set is then modified to substitute *CN* with the proposed simplified model parameters of *CS* and *SM*.

The prediction accuracy of the learned models is tested using RMSE (Root Mean Square Error), 10-fold cross validation (CVRMSE), Relative RMSE (RRSME) and R squared value (R2). A good value for RMSE and CVRMSE is stated as half of the standard deviation value for the output data (Singh *et al.* [13]). This comes out as 1.3 for *Q* and $t_1$, and 3.2 for $t_d$. Values of R2 and RRMSE can range between 0 and 1, where 1 means perfect forecasting. The value of RRMSE is represented as a percentage. The predicted models developed using different model parameters and training set sizes, are evaluated with test data to compare their performance with the NRCS model. For performance evaluation of these models, we use Matlab's M5 decision tree toolbox (Jekabsons *et al.* [14]) and the Java-based Weka machine learning simulator (Hall *et al.* [15]).

### *Q*-predictive model

For developing the model, a simple regression tree algorithm was first used to compare the performance with another, more advanced, algorithm - M5 decision trees [14]. For a small training set of 65 samples, the M5 model gives better performance (RMSE=0.317, R2=0.984, RRMSE=5.98%) when compared to the regression tree model (RMSE=0.915, R2=0.506, RRMSE=10%). Although the value of RMSE indicates that the regression tree model is also acceptable. However, the value of R2 indicates an extensive deviation of its predicted results; hence, the M5-tree algorithm was selected for further evaluation.

Table 1: Performance of the predictive models based on various training sizes using M5 trees

| | *Q*-Predictive Model (P, CS, SM) | | | $t_1$-Predictive Model ($P_d$, P, CS, SM) | | | $t_d$-Predictive Model ($P_d$, P, CS, SM) | | |
|---|---|---|---|---|---|---|---|---|---|
| Training set size | 250 | 125 | 65 | 450 | 300 | 100 | 450 | 300 | 100 |
| RMSE | 0.159 | 0.234 | 0.317 | 0.239 | 0.318 | 0.825 | 0.2755 | 0.299 | 0.598 |
| R2 | 0.998 | 0.997 | 0.984 | 0.985 | 0.976 | 0.835 | 0.997 | 0.977 | 0.991 |
| CVRMSE | 0.216 | 0.278 | 0.465 | 0.2935 | 0.381 | 1.042 | 0.3856 | 0.426 | 0.713 |
| RRMSE | 5.7% | 7.5% | 5.98% | 16.1% | 16.8% | 27% | 5% | 6% | 8.2% |

Figure 5A illustrates plots of predicted data for M5 tree models, calculated using the various models (see figure legend), against the output of NRCS mathematical model. The proposed model (*P*, *CS*, *SM*) shows an excellent match (R2=0.984, RRMSE=5.98%), however the predicted results of model developed using only *P* gives 30% RRMSE. Similarly, the performance of *Q*-predictive models generated using different training set sizes, based on the proposed parameters, is shown in Table 1. Figure 5 B) plots the result of test data for these models. Even a small training set of 65 samples retains the performance of the model.

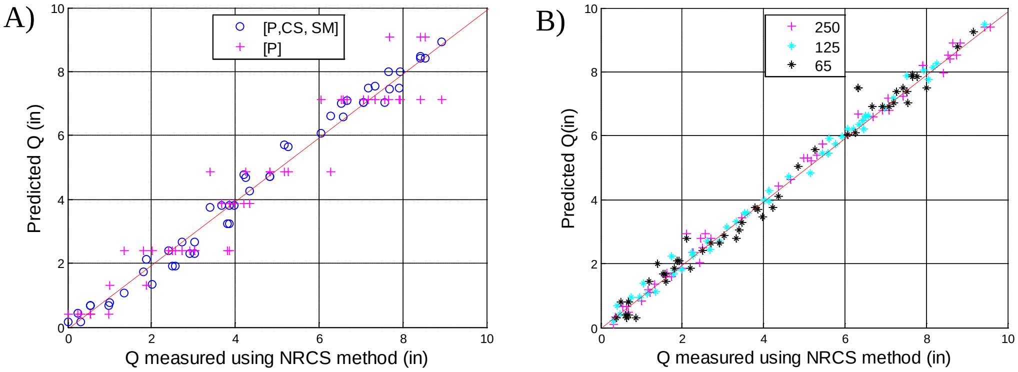

Figure 5: Plot of data predicted using *Q*-predictive models, generated using various A) model parameters and B) training sizes, against data measured using NRCS method

### $t_1$-Predictive and $t_d$-Predictive model

For $t_1$, initially 450 training instances, based on the same model parameters as in *Q* model (*P*, *CS*, *SM*), were used to generate the model for the sake of comparison using M5 decision tree. However, the model performance was very poor with RMSE of 1.433, which is higher than the acceptable value of 1.3, and RRMSE as 65%,. This shows that the same model parameters cannot be used for $t_1$. The new proposed parameters ($P_d$, *P, CS, and SM*) were then used to generate the model. This substantially improved the model performance (RMSE= 0.239, RRMSE=16.1%). This is further illustrated by plotting the result of predicted data using the above two models against the NRCS model in Figure 6A. Furthermore, to evaluate the impact of training set size on prediction performance, models are generated, based on the proposed parameters (*$P_d$, P, CS & SM*). As shown in Table 1, model developed using a training set of 300 shows reasonable performance (RMSE=0.318, RRMSE=16.8%). Although with 100 samples, the RMSE seems adequate (0.825), however the RRMSE increases to 27%. Therefore, although a training set of 65 samples gave good performance (RRMSE=5.98%) for *Q* model, similar results do not hold for $t_1$, and more training samples are required.

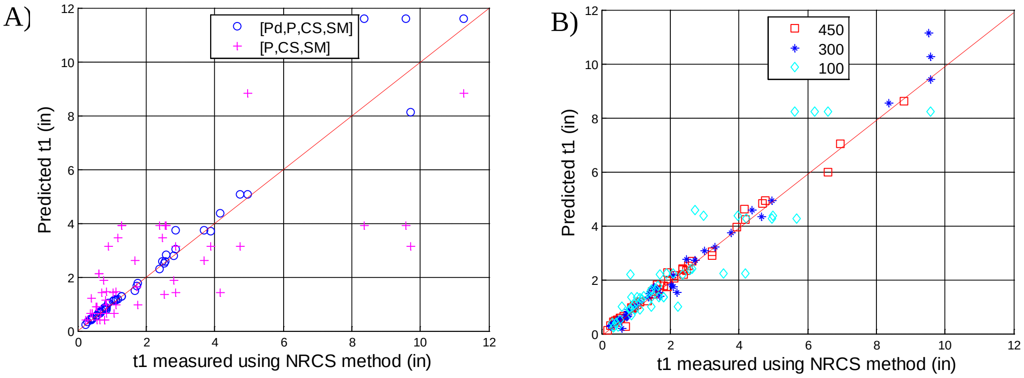

Figure 6: Plot of data predicted using $t_1$-predictive models, generated using various A) model parameters and B) training sizes, against data measured using NRCS method

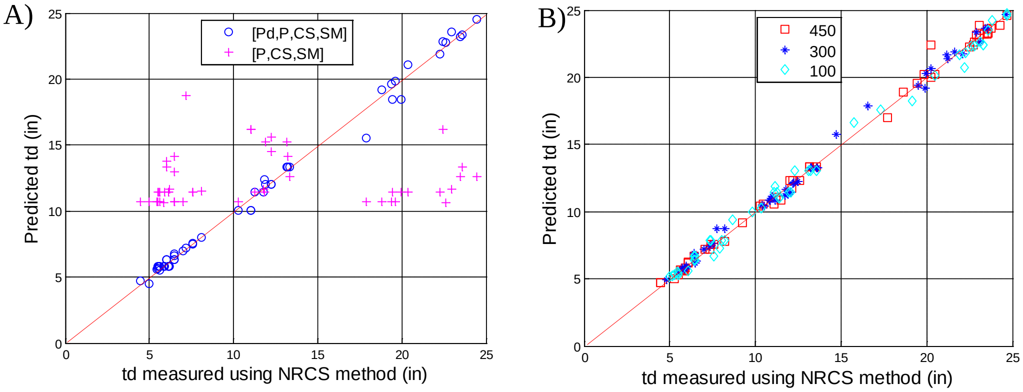

Figure 7: Plot of data predicted using $t_d$-predictive models, generated using various A) model parameters and B) training sizes, against data measured using NRCS method

Similarly, for $t_d$, the plot shown in Figure 7 A) demonstrates that the results of the model developed using parameter set (*P, CS, SM*) and 300 samples fits poorly to a 1:1 ratio (R2=0.107, RRMSE=98%), as compared to the results of the model generated using *$P_d$, P, CS, & SM* (R2=0.997, RRMSE =6%). As compared to the $t_1$-model, the $t_d$- model shows higher correlation of output with $P_d$. Similarly a comparison of prediction performance of M5 models developed using different training sets, based on *$P_d$, P, CS &SM*, is shown in Table 1. Unlike $t_1$, even a small training set of 100 shows good correlation (R2=0.990, RRMSE =8%) as shown in Figure 7 B).

## CONCLUSIONS

This paper has proposed that individual farm-scale networks can be integrated into a collaborative framework to support catchment-scale water quality monitoring and management to learn and predict the impact of catchment events. This enables reutilization and timely control of nutrient outflows within the farm system. Since a computing model on a sensor network, for the implementation of the collaborative WQMCM framework, requires a simplified underlying physical model therefore, low-dimensional model parameters are derived from the existing NRCS method for the prediction of discharge dynamics. An M5 decision tree algorithm is used to develop predictive models for discharge volume ($Q$) and response timing ($t_1$ and $t_d$), based on the proposed model parameters. Evaluation of these models has demonstrated high accuracy for $Q$ and $t_d$ (94%), even for a small training set of under 100 samples. However, for $t_1$, 300 samples are required to provide adequate performance (84%).

## ACKNOWLEDGEMENTS

We would like to thank Prof. George Vellidis and Mike Tucker at the University of Georgia for providing us with season long soil moisture data from a field.

## REFERENCES

[1] EPA, "National Water Quality Inventory: Report to Congress ; 2004 Reporting Cycle," US Environmental Protection Agency EPA 841-R-08-001, 2009.

[2] H. H. Harper, "Impacts of Reuse Irrigation on Nutrient Loadings and Transport in Urbanized Drainage Basins," Florida Stormwater Association, Environmental Research & Design, Inc.2012.

[3] H. Zia, N. R. Harris, G. V. Merrett, M. Rivers, and N. Coles, "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,* vol. 96, pp. 126-138, 2013.

[4] H. Zia, N. Harris, and G. Merrett, "Collaborative Catchment-ScaleWater Quality Management using IntegratedWireless Sensor Networks," presented at the EGU General Assembly, Vienna, Austria, 2013.

[5] E. A. Basha, S. Ravela, and D. Rus, "Model-based monitoring for early warning flood detection," in *Proc. of Embedded network sensor systems Conf.*, 2008, pp. 295-308.

[6] D. P. Solomatine and M. B. Siek, "Modular learning models in forecasting natural phenomena," *Neural Networks,* vol. 19, pp. 215-224, 2006.

[7] R. H. Hawkins, A. T. Hjelmfelt Jr, and A. W. Zevenbergen, "Runoff probability, storm depth, and curve numbers," *Jour. of Irrig. and Drainage Engg.*, vol. 111, 1985.

[8] E. M. Shaw, K. J. Beven, N. A. Chappell, and R. Lamb, *Hydrology in practice*: Taylor & Francis US, 2010.

[9] L. Fennessey and R. Hawkins, "The NRCS Curve Number, a New Look at an Old Tool," in *Proc. of Pennsylvania Stormwater Management Symp., Villanova Uni.*, 2001.

[10] G. Vellidis, H. Savelle, R. Ritchie, G. Harris, R. Hill, and H. Henry, "NDVI response of cotton to nitrogen application rates in Georgia," *Precision Agriculture*, p. 359, 2011.

[11] G. Vellidis, M. Tucker, C. Perry, C. Kvien, and C. Bednarz, "A real-time wireless smart sensor array for scheduling irrigation," *Computers and Electronics in Agriculture,* vol. 61, pp. 44-50, 2008.

[12] T. Davis, " SCS Unit Hydrograph Convolution : Hydrograph Generation and Analysis Tool," ed: Matlab.

[13] J. Singh, H. V. Knapp, J. Arnold, and M. Demissie, "Hydrological modeling of the iroquois river watershed using HSPF and SWAT1," *Journal of the American Water Resources Asso.,* vol. 41, pp. 343-360, 2005.

[14] G. Jekabsons. (2010). *M5PrimeLab: M5' regression tree and model tree toolbox for Matlab*. Available: http://www.cs.rtu.lv/jekabsons/Files/M5PrimeLab.pdf

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10-18, 2009.