UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

School of Electronics and Computer Science

Enabling Proactive Agricultural Drainage Reuse for Improved Water Quality through Collaborative Networks and Low-Complexity Data-Driven Modelling

by

Huma Zia

Thesis for the degree of Doctor of Philosophy

June 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

ENABLING PROACTIVE AGRICULTURAL DRAINAGE REUSE FOR IMPROVED WATER QUALITY THROUGH COLLABORATIVE NETWORKS AND LOW-COMPLEXITY DATA-DRIVEN MODELLING

by Huma Zia

With increasing prevalence of Wireless Sensor Networks (WSNs) in agriculture and hydrology, there exists an opportunity for providing a technologically viable solution for the conservation of already scarce fresh water resources. In this thesis, a novel framework is proposed for enabling a proactive management of agricultural drainage and nutrient losses at farm scale where complex models are replaced by in-situ sensing, communication and low complexity predictive models suited to an autonomous operation. This is achieved through the development of the proposed Water Quality Management using Collaborative Monitoring (WQMCM) framework that combines local farm-scale WSNs through an information sharing mechanism.

Under the proposed WQMCM framework, various functional modules are developed to demonstrate the overall mechanism: (1) neighbour learning and linking, (2) low-complexity predictive models for drainage dynamics, (3) low-complexity predictive model for nitrate losses, and (4) decision support model for drainage and nitrate reusability. The predictive models for drainage dynamics and nitrate losses are developed by abstracting model complexity from the traditional models (National Resource Conservation Method (NRCS) and De-Nitrification-DeComposition (DNDC) model respectively). Machine learning algorithms such as M5 decision tree, multiple linear regression, artificial neural networks, C4.5, and Naïve Bayes are used in this thesis. For the predictive models, validation is performed using 12-month long event dataset from a sub-catchment in Ireland.

Overall, the following contributions are achieved: (1) framework architecture and implementation for WQMCM for a networked catchment, (2) model development for low-complexity drainage discharge dynamics and nitrate losses by reducing number of model parameters to less than 50%, (3) validation of the predictive models for drainage and nitrate losses using M5 tree algorithm and measured catchment data. Additionally modelling results are compared with existing models and further tested with using other learning algorithms, and (4) development of a decision support model, based on Naïve Bayes algorithm, for suggesting reusability of drainage and nitrate losses.

Table of Contents

List of tables

List of figures

Declaration of Authorship

I, Huma Zia, declare that this thesis entitled 'Enabling Proactive Agricultural Drainage Reuse for Improved Water Quality through Collaborative Networks and Low-Complexity Data-Driven Modelling' and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as listed in section 1.3 of this thesis.

Signed: H. Zia

Date: 30 Apr 2015

Acknowledgements

I would like to acknowledge the contributions of all those who supported me throughout this amazing and unforgettable journey for my doctorate degree. First and foremost my sincere gratitude to my supervisors Nick Harris, Geoff Merrett and Prof. Neil White for their supervision, guidance and support. I'm incredibly grateful to Nick Harris for putting in confidence in my abilities, showing the enthusiasm and openness to adopt new ideas, and challenging me to achieve more than I ever thought was possible. Furthermore, I would also like to thank Prof. Mark Rivers, from the University of Western Australia, for his continuous support all through my research.

I wish to thank the Commonwealth Scholarship Commission (CSC) for granting me the chance to undertake my doctoral studies in the UK. I am also grateful to the World University Network (WUN) for providing travel support in order for me to collaborate with researchers in Australia. Furthermore, I am thankful for the research facilities and support provided by the School of Electronics and Computer Sciences (ECS) throughout my studies at Southampton.

In addition, a special thanks to all my friends in the ESS and EEE group, especially Teng Jiang and Sei Ping Lau, for motivating and inspiring me, and making this an interesting place to work. Outside of work, I was especially lucky to have wonderful friends like Zunaira, Tayyaba, Wasifa and Sana, who were always there for discussions, stress relief and providing a better perspective.

My family particularly my parents who dealt with quite a lot and supported me through it all. My thanks to my siblings for all their support and love. Finally, I would like to express my special gratitude to my husband Moiz for being extremely loyal to the cause of my studies and for accommodating the hectic routine, constant travelling, and tension that occurred during this time. This thesis is dedicated to my parents and my husband, without whom none of this would have been possible.

Chapter 1: Introduction

"Water and air, the two essential fluids on which all life depends, have become global garbage cans"

Jacques-Yves Cousteau

Less than 1% of the earth's total fresh water resources are available to humans and that which is available is distributed inequitably. Because of this, billions of people around the globe live in water-stressed basins [1]. This is exacerbated by poor water management and out-dated agricultural activities, which result in further depletion [2] and degradation of already scarce water resources. Furthermore, due to the increasing food demand to cater to the growing population, water withdrawals for human consumption and food production have increased [3]. As a result, it is feared that, by 2025, half of Earth's population will be facing water-based vulnerability [4]. Figure 1-1 illustrates the role of various activities in a catchment contributing towards water quality degradation. Surface and ground water contamination in a catchment is mainly attributed to outdated farm management practices [5, 6].

In traditional agricultural practices, excessive or poorly timed application of irrigation and fertilizers result in nutrient fluxes into the water mainly containing phosphorous (P) and nitrogen (N). Poor water use efficiency from surface irrigation, still the most commonly used irrigation method, results in 40 to 60% of water losses in the form of runoff [7, 8]. In England, 70% of the landscape drains to waterways identified as "N-polluted" under the EC Nitrates Directive; 60% of this N comes from agriculture [9, 10]. This is majorly because agronomic nutrient recommendations are often far in excess of environmentally-appropriate levels [11]. N becomes a concern to water quality when soil N is converted into nitrate (NO_3^-). This form of N is not held tightly by soil particles, therefore residual nitrates in the soil after the growing season will be lost to leaching and runoff into ground and surface water [12, 13]. In some cases, 30%-50% of applied N is lost due to the coupled impact of over fertilization and irrigation runoff [14]. These substantial nutrient losses can have serious agronomic, economic and environmental implications [15], e.g. nutrient-enrichment of waterways (eutrophication), can give rise to toxic and nuisance algal blooms which contaminate drinking water and harm aquatic life [16, 17].

To control the increasing environmental impact of the agriculture sector, various legislative approaches have been formulated; of particular note are the 'Clean Water Act' [18, 19], the 'Water Framework Directive' [20] and the 'Nitrates Directive' [21]. These approaches include adherence to the implementation of best management practices (BMPs) related to soils, nutrients, manure, and water management. Furthermore, prior soil testing for N and P is done to develop fertilizer application budgets accordingly. However, their effectiveness needs to be assessed from farm scale to catchment scale using continuous long term data. While BMPs have been successful, it is important to note that the inherent inefficiency of nutrient uptake by crops and of commonly used surface irrigation method renders excess nutrient inputs (and their likely loss) inevitable.

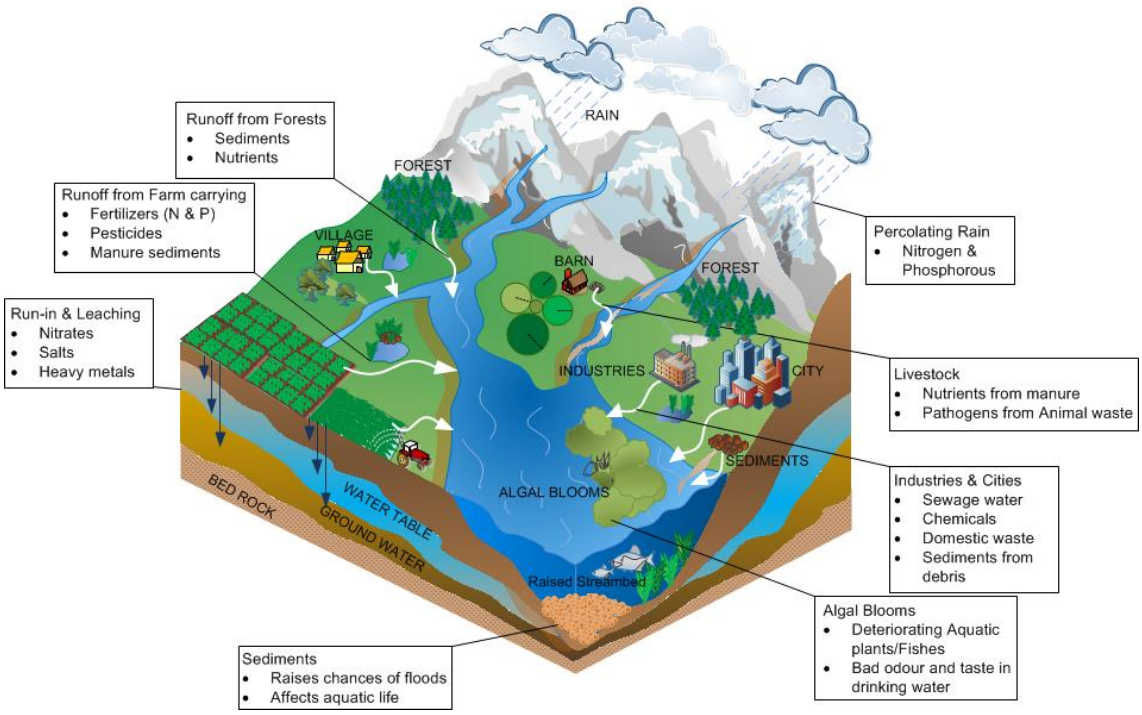


Figure 1-1: Role of catchment-scale activities in water quality degradation

Water quality control through drainage management has been widely advocated in the literature [22]. Basic options for this include drainage water reuse, disposal and treatment. It has been advocated in the literature that drainage reuse is the most suitable option to address the current water scarcity and quality problems [3]. Drainage reuse can have many environmental as well as economic benefits [23] (see appendix A for a case study on economic value analysis of drainage losses). Unlike phosphorous, which is less soluble than nitrogen, N (in the form of nitrates and nitrites), is more prone to be lost through leaching and drainage water [24]. The sum of nitrates and nitrites is also referred to as total oxidized nitrogen (*TON*), while the levels of nitrites are usually minimal. *TON* does not volatilize and is therefore, likely to remain in water until used by

the plants. For example, in some intensive farming areas, farmers have begun to test their groundwater for nitrate concentrations and therefore change their nutrient budgets accordingly [25]. In another case, reapplication to land of N-rich runoff waters provided more than the annual nutrient requirements for that land [26]. Some other research efforts have been aimed at local drainage reuse for hydroponic systems maintained in greenhouses to increase water and nutrient use efficiency and to reduce the environmental impact [27]. The results indicated 33% reduction in fresh water usage for irrigation. Furthermore, it was determined that drainage water collected from the greenhouse contained 59% of applied N. These studies, though small scale and based on local drainage reuse, are very encouraging. Despite tremendous technological advancements and the great promise of the benefits of drainage water reuse, implementation of an intelligent and autonomous management mechanism has not kept pace with the deteriorating water situation [28]. Various resource constraints and farmer's concerns regarding real time availability of information on volumes, timings, and quality of discharges that can be delivered to the farms [29, 30], restricts wide adoption of this mechanism in agriculture.

The demand for better and effective integration of technology and decision making is ubiquitous in environmental management [31]. Over recent years, WSNs, with their attractions of low cost and continuous real time data availability, have received considerable attention in environmental monitoring. WSNs provided a much needed alternative to the previously used methods which either lacked fine-grained spatio-temporal data, such as grab sampling and data loggers [32, 33], or were very expensive, like satellite remote sensing [34]. Furthermore, these methods monitoring at weekly and monthly time step have the potential to misclassify water quality status by missing pollution events [35]. WSNs are networks of small embedded computers referred to as sensor nodes that monitor environment and transmit data wirelessly amongst themselves. Information collected from all the nodes is aggregated at the data sink and transmitted to and analysed at an off-field control node where a decisions might be taken. Since their emergence, WSNs have been successfully used for monitoring and managing farming activities [36, 37] and eco-hydrologic processes [38, 39]. Sensor networks have monitored microclimate of farms to prevent moulds in crops [40, 41], provided precision irrigation in farms to better manage water [42, 43], and monitored animal health in barns [44, 45]. Similarly in hydrology, sensor networks have been employed to monitor water quality in rivers [38, 46], lakes [47], dams [48] and ground water [49, 50] for applications seeking to investigate suitability of drinking water, health threats, degradation of aquatic life and water. Water quality parameters, such as pH, dissolved oxygen,

turbidity and salinity, have been monitored in these applications. These applications are discussed in detail in the literature review in chapter 2.

Despite wide adoption of WSNs in agriculture, none of the applications focus on monitoring and controlling the impact of individual farms and fields for their contributions to degraded water quality. Only a couple of small-scale applications [51, 52] have focused on monitoring local drainage water for nutrients outflow. Furthermore, most of the water quality monitoring applications are motivated by only reporting the quality status of water system bodies. Some applications have also been intended to monitor large-scale impact of weather on nutrient outflows in large catchments [53-55] encompassing water systems, mountains and agricultural lands. In these projects, data observations were made using a single network and were transmitted to a central depository using GSM or internet. Using a central processing unit for high resolution heterogeneous data from multiple networks leads to processing and filtering overhead, and delays decisions. In addition it requires availability of internet or GSM which renders this solution unsuitable in a remote area of crop lands. Moreover, these are still prototype deployments based on limited monitoring points across the catchment aimed at developing the technology and algorithms for large-scale remote monitoring and data processing. To conclude, all the above discussed applications in water quality domain renders WSNs operation reactive rather than proactive. To the author's best knowledge, none of the applications focused on drainage monitoring at farm scale with the aim of enabling reuse or any other control management. Furthermore, there is no framework to investigate a distributed mechanism, comprising autonomous networks, for dynamic monitoring and control without using a central control unit.

To preserve water resources and quality, a proactive, intelligent and integrated *monitoring and management mechanism* is required. For enabling such a mechanism, it is believed that there is huge potential for leveraging individual networked farms and streams into an integrated mechanism. This is more natural to this specific catchment-scale problem as it is not feasible to have a single monitoring system covering the entire catchment as individual stakeholder's monitoring requirements vary as they undertake different activities. Allowing event information to be transmitted across multiple networks as they are detected will allow prediction of when the repercussions of that event might be seen downstream. This facilitates other stakeholder networks in the vicinity to adjust their monitoring and management strategy. A de-centralized approach comprising autonomous networks presents a flexible methodology where independent networks, in addition to local monitoring objectives, seek to opportunistically utilize neighbouring events. This approach fundamentally differentiates the undertaken research from the existing work.

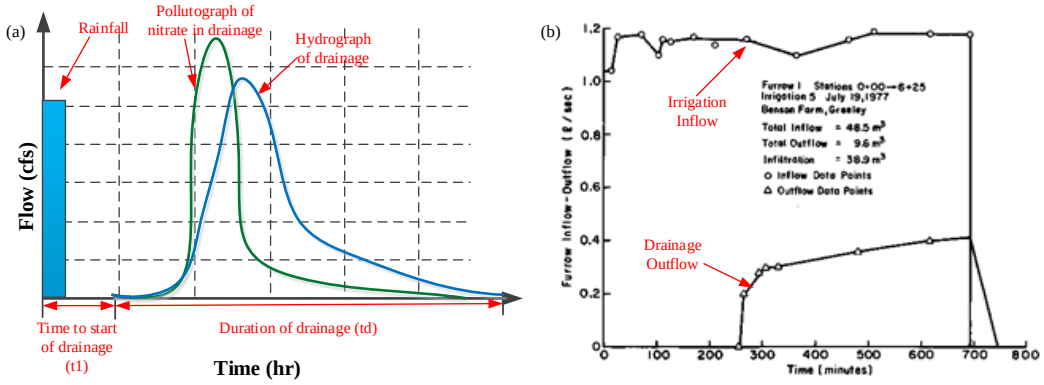


Figure 1-2: (a) Plot of a typical discharge hydrograph and pollutograph for a rainfall event, (b) plot of drainage outflow (reproduced from [56]) for an irrigation inflow

As it is already established that drainage reuse is very suitable for addressing the existing water crisis by reducing the overall drainage and nutrient losses, however availability of prior information on volumes and timings of drainage is crucial to implementing proactive drainage management. The predictive information about availability of drainage is important for several reasons. Firstly, because these drainage events can be intense and short-lived, most of the nutrient fluxes flush away at the beginning and so normally are missed (completely or partially) at usual sampling rates [35]. Figure 1-2 illustrates a typical plot of a discharge hydrograph and pollutograph with respect to time, in response to a rainfall and irrigation event. Secondly, the networks receiving the drainage (farms, drainage bay, streams) would need to adjust the sampling rate of their sensors to capture the actual drainage, nutrients, or pollutant (if any) contributions by farms. It can be argued to have the network sensors sample continuously however, this strategy substantially reduces network lifetime (see appendix B for a related case study). Thirdly, to manage the drainage outflows, predictive information is required to plan ahead about reuse, or disposal of the drainage.

To date, numerous physically-based and mathematical models have been developed for the prediction of hydrological discharges and nutrient losses. Although these models are quite 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 [57, 58]. Furthermore, constraints on network nodes with respect to battery life, computing power, and availability of sensors, necessitates development of low-dimensional and simplified models for deployment within the networks. In comparison to

traditional modelling approaches, 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 [59]. Such models, once implemented in the network, can adaptively learn and further improve their accuracy over the course of time. Thus data-driven modelling, using machine learning algorithms, has been widely used in hydrological modelling [60-62] with artificial neural networks (ANN) being a popular choice [60, 63, 64]. Recently, decision tree modelling has been investigated [58, 65-68] and an interesting example of this class is the M5 tree algorithm [69]. M5 tree algorithm has been used in hydrology for modelling drainage discharge [68], rainfall-runoff [70-72], streamflow [58], flood forecasting [73] and sediment transport in discharge [74]. It is important to mention that none of these applications except for the work by *Kuzmanovski* [68] were based on farm scale drainage discharges. *Kuzmanovski* [68] 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. Although, machine learning has been widely used in hydrological modelling; it has only recently been adopted in the modelling of N losses [66, 67]. In this regard, a modelling framework was developed to calculate annual nitrous oxide flux and nitrate leaching by abstracting the complexity of a complex bio-geo-chemical model [65]. Various algorithms, such as multi-layer perceptron, random forests and support vector machine were used for comparison. Similarly, other work further simplified the input parameters and training size requirement and used neural networks to simulate total N emissions from the de-nitrification process (N_2O) [75].

There are two major limitations in the existing work for both of nitrate and drainage modelling; either the existing models use simpler parameters but years of historical data with thousands of training samples to learn the heterogeneity of large areas (above 1000 hectare (ha)), or they use more complex models with a significant number of parameters that may require manual sampling and expensive sensing equipment (e.g. bio-geo-chemical parameters). 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. Therefore, there is a need for developing low-complexity predictive models for drainage and nitrate losses based on fewer and, ideally, real-time field parameters acquired autonomously and shareable across neighbouring farms.

1.1 Research Aims

In this thesis, the research mainly investigates a mechanism for enabling proactive reuse of agricultural drainage, intended for water quality control, through an intelligent mechanism of collaborative networks. Such a mechanism fundamentally requires a technological framework to enable communication among farms, low complexity models to allow drainage forecasts based on the communicated information, and a management tool for advising on drainage reuse.

The themes of this research, and hence this thesis, are:

- **A Technological Framework:** This research investigates the design of a framework for facilitating real time utilization or disposal of agricultural drainage among farms using collaboration among prevalent farm networks. The basic system architecture comprises modules for environmental learning, prediction of the impact of neighbouring events in terms of drainage and nutrients losses, and a local decision support mechanism. The overall functionality of the framework is explored in terms of stages of learning, training and testing. A network learning model is required to identify flow links of a network with neighbouring networks.

- **Low-Complexity Predictive Models:** In order to enable on-node prediction for drainage, simplified models, describing the underlying physical phenomena, based on fewer and real time field parameters are required. A key theme of this research is developing low-complexity models for drainage depth/volume, drainage response time (duration and start time), and N losses, by abstracting detail from traditional hydrological and nutrient loss models. The developed models should be validated through measured field data. Further exploration is undertaken for performance evaluation of models developed using varying training set sizes, different machine learning algorithms, and comparison with other relevant studies.

- **Decision Support Mechanism:** This research considers the functional blocks required for implementing the management decision mechanism for reusing drainage and nutrients in a farm. The key aspects are an economic evaluation of the drainage, a decision support model for classifying reusability of drainage and nutrients, and adaptive sampling of sensors in the farm.

1.2 Research Contributions

Overall, this thesis contributes to the goal of an intelligent, proactive and autonomous water quality control sensor network. The major contributions of this thesis are:

1. **Water Quality Management using Collaborative Monitoring (WQMCM) Framework:** Design and development of a novel proactive framework for integrated and autonomous water quality control and management by using collaboration among sensor networks. This framework couples real-time water quality monitoring and collaborative networks in a catchment through information sharing, thereby facilitating opportunistic management and policy making. The basic system architecture comprises modules for neighbour linking model, predictive models for Q (drainage depth/volume), t_1 (drainage start time), t_d (drainage depth), *TON* (Total Oxidized Nitrogen), and decision support model for drainage reuse.

2. **Low-Complexity Q, t_1, and t_d Predictive Models:** Development and simulation of a low complexity models by seeking simplification in model parameters of the traditional mathematical National Resource Conservation Method (NRCS) curve number model. The number of parameters for modelling Q, t_1, and t_d are reduced to less than 50% of the parameters used by the NRCS model. Evaluation of the predictive models, developed using simulated data and M5 decision tree algorithm, demonstrates accuracy of 84-94% for training samples under 300, compared with the traditional NRCS model.

3. **Low-Complexity *TON*-loss Predictive Model:** Development of a low complexity *TON*-loss predictive model by abstracting details from an existing complex bio-geo-chemical model (by *Villa-Vialaneix et.al.* [65]) to select an optimized input parameter combination with optimal performance. This reduces the number of model parameters used to 5 – a reduction of at least 50%.

4. **Validation of Q and *TON* Predictive Models:** Using M5 tree algorithm and a yearlong event dataset from a real catchment in Ireland, the models have been validated with promising results. For the Q-predictive model, the training dataset consisting of daily values for precipitation, field conditions and discharges, is used for training and testing the model. Results for the Q-predictive model provide R^2 of 0.82 and RRMSE of 35.9%. With 10-fold cross-validation, R^2 is 0.63 and RRMSE is 55%. 80% of the residuals for the predicted test values fall within ± 2 mm discharge depth/day error range. These results

are comparable with the state-of-the-art in this domain which utilizes thousands of training samples. Furthermore, for the *TON*-loss predictive model, training data consists of daily values for N input, climatic conditions and N losses. The results provide R^2 as 0.92 and RRMSE as 26%. 80% of the residuals for test data fall within $\pm$0.05 $Kgha^{-1}day^{-1}$ error range, 10-fold cross-validation results indicate R^2 as 0.72 and RRMSE as 47.7%. These results are better than the existing research efforts.

5. **Decision Support Model**: While not a primary contribution of this research, a simple illustration of a decision support model for drainage and nitrate reuse has been presented. Using Naïve Bayes algorithm, the developed decision model for drainage reuse classifies the learned ruleset with 90.47% accuracy, whereas for *TON* reuse the model correctly classifies 94.44% of the samples. These results are important because they demonstrate that simple ruleset based on the proposed model parameters has the potential to generate a useful model which can support decision making.

1.3 Published Papers

The research undertaken in this thesis has contributed in part or full to the following publications:

1. 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" in Journal of Computers and Electronics in Agriculture, (96), 126-138, 2013.

2. H. Zia, N. R. Harris, G. V. Merrett, "Collaborative catchment-scale water quality management using integrated wireless sensor networks (poster)", presented at European Geosciences Union General Assembly 2013, Austria, Republic of, AT, 07 - 12 Apr. 2013.

3. P. Hall, N. Coles, J. Camkin, A. Cranny, N. R. Harris, and H. Zia, "Water quality management: the case for real time monitoring and reporting", presented at 16th International River symposium, Brisbane, Australia, 23 - 26 Sep 2013.

4. N. Coles, J. Camkin, N. R. Harris, A. Cranny, P. Hall and H. Zia, " Water, boundaries and borders, the great intangibles in water quality management: can new technologies enable more effective compliance?", presented at Transboundary water management

across borders and interfaces: present and future challenges (TWAM2013), Aveiro, PT, 16 - 20 Mar 2013.

5. H. Zia, N. R. Harris, G. V. Merrett, "Water quality monitoring, control and management (WQMCM) framework using collaborative wireless sensor networks", presented at 11th International Conference on Hydroinformatics (HIC) 2014, New York City, US, 17 - 21 Aug 2014.

6. H. Zia, N. R. Harris, G. V. Merrett, "A Low Complexity Data Driven Model of Environmental Discharge Dynamics for Sensor Network Applications", presented at Eurosensors XXVIII, 8-10 Sep 2014.

7. H. Zia, N. R. Harris, G. V. Merrett, "Empirical modelling and simulation for discharge dynamics enabling catchment-scale water quality management", presented at 26th European Modelling & Simulation Symposium, 10 - 12 Sep 2014.

8. H. Zia, N. R. Harris, G. V. Merrett, and M. Rivers, "Data-driven Low-Complexity Nitrate Loss Model utilizing Sensor Information – Towards Collaborative Farm Management with Wireless Sensor Networks", in 2015 IEEE Sensors Applications Symposium (SAS 2015), Zadar, Croatia, 13-15 Apr 2015.

9. M. Rivers, N. Coles, H. Zia, N. R. Harris, R. Yates, "How could sensor networks help with agricultural water management issues? Optimizing irrigation scheduling through networked soil-moisture sensors", in 2015 IEEE Sensors Applications Symposium (SAS 2015), Zadar, Croatia, 13-15 Apr 2015.

10. H Zia, N. R. Harris, G. V. Merrett , "Validation of a Low Complexity Machine Learning Discharge Predictive Model – Enabling Proactive Water Management in a Catchment with Wireless Sensor Networks", in Journal of Computers and Electronics, 2014, (under review).

11. H Zia, N. R. Harris, G. V. Merrett , "A Low-Complexity Machine Learning Nitrate loss Predictive Model – Towards Proactive Farm Management with Collaborative Wireless Sensor Networks", in Journal of Hydroinformatics, 2015, (under review).

1.4 Thesis Structure

The structure of the remainder of the thesis is as follows:

Chapter 2 reviews the applications of WSNs in agriculture and hydrology to establish and critique the state-of-the-art. It further elaborates on the existing data driven modelling approaches. Chapter 3 discusses WQMCM architecture, a novel framework proposed in this thesis, along with development of the neighbour linking model. Chapter 4 presents the modelling and simulation of low-complexity *Q*-predictive, t_1-predictive and t_d-predictive models. Chapter 5 investigates the validation of the *Q*-predictive model. Chapter 6 explores the modelling and validation of the *TON*-loss predictive model. Chapter 7 presents the decision support modelling which interprets predicted *Q* and *TON* values so that a water reuse decision can be taken. Chapter 8 concludes the thesis, and presents an outline for future work.

Appendix A presents a case study on economic valuation and resulting benefit of reusing the average drainage losses in Australia for grass land and rice crop.

Appendix B presents a case study illustrating the impact of sampling rate on network life and information quality.

Appendix C incudes a table representing the NRCS curve numbers for an agricultural land.

Chapter 2: Literature Review

This chapter provides an overview of the state-of-the-art for all aspects of this thesis. Section 2.1 presents a discussion about water quality control through drainage management, specifically drainage reuse. Section 2.2 explores the work related to the applications of WSNs in hydrology and agriculture with the aim of highlighting strengths and limitations in existing work for monitoring and managing water quality. Sections 2.3 and 2.4 discusses related work for discharge and TON loss prediction with respect to traditional and data-driven modelling respectively. In the end, section 2.5 presents the machine learning algorithms that have been commonly used for hydrological and TON loss modelling, and hence will be adopted in this thesis for model development.

2.1 Water Quality Control through Drainage Management

As already discussed in the introduction chapter, there is an increasing strain on existing fresh water resources globally to cater to rising food demands by producing higher crop yields. Higher levels of nutrients in soil which cannot be utilized productively are either stored in the soil or are lost through leaching and runoff into ground and surface water. As the soil becomes saturated with nutrients, any excess input in the form of fertilizers or manure is lost completely [13, 76]. Owing to the coupled impact of traditional farming practices and inherent inefficiency of nutrient uptake by crops (up to 60%), there is an inevitable release of drainage water (35-60% of surface irrigation water) rich in nutrients [7]. For example, in a watershed (190,000 ha) in Western Australia, facing high nutrient fluxes, phosphorous losses are measured to be 140 tonnes per annum (tpa) which is twice its target. It is expected to rise to 1300 tpa in the next 100 years if current practices continue [78]. In another study on the North China Plain, the recovery of fertilizer N by the crop at the conventional N fertilizer rate (300 $KgNha^{-1}$) was approximately 25%, while 30%-50% of the applied N was lost [14].

In order to avoid exacerbating the water crisis and to prevent food shortages, an advantageous strategy is the conservation and reuse of agricultural drainage (and the dissolved nutrients) before it ends up in fresh water system [3]. Reusing drainage water and nutrients emanating from one farm in another farm, before they enter the water system, can have huge environmental as well as economic benefits. In particular, reuse reduces the amount of fresh water extracted from the environment, thus lowering its diversion from sensitive ecosystems. In regions where irrigation water supplies are limited, drainage water can be used to supplement them [22]. In addition, agricultural drainage reuse can benefit the farmers (or any stakeholders) by saving cost on not using fresh irrigation water and fertilizer inputs.

The only concern about drainage water reuse is whether or not the water is safe for reuse, i.e., does not contain high concentration of salts and pesticides. Highly saline water cannot be used for salt-sensitive crops. However it can successfully be used for salt-tolerant crops, trees, fodder, natural wetland and even for salt-sensitive crops at later growth stages [3, 79]. Conjunctive use of saline water with fresh water increases the suitability of drainage water. With regards to pesticides, in areas where strong environmental safeguards exist for pesticide usage, there is little risk associated with the reuse of surface runoff or tail water drainage [79]. Hence, drainage water can safely be used if appropriate considerations are taken into account.

Drainage water and dissolved nutrients have been globally utilized for crops and greenhouses. In some intensive farming areas, farmers have begun to test their groundwater for nitrate concentrations and therefore change their nutrient budgets accordingly [25]. In another case, reapplication of N-rich runoff waters provided more than the annual nutrient requirements for that land [26]. In one study, reuse of saline water for salt tolerant forages, has been investigated under varying salinity level treatments (between 15 and 25 dS/m). For this experiment, sand tanks were used in a greenhouse. Almost all forages showed promise with regards to biomass production, whereas wheatgrass, Bermuda-grass, and paspalum performed particularly well [80].

Some work in local drainage reuse is reported for hydroponic systems maintained in a greenhouse (in which plants are grown in water instead of soil). In one application, high quality tomato was grown with drainage reuse [81]. During seedling stage, fresh nutrients were supplied with irrigation of which 20-30% overflowed as drainage. At the final stage of ripening, the preserved drainage was reused with no wastage being drained out from the greenhouse. In a similar work based on greenhouses in Australia, drainage reuse was used for growing cucumber and tomato [27]. The study was aimed at investigating the use of drainage water of the greenhouse to increase water and nutrient use efficiency and reduce the environmental impact. Flow meters were installed to gauge the volumes of water applied to the crops. Water samples were taken five times a day for inflows and outflows, and were analysed for pH, salinity and concentrations of nutrients. The results indicated 33% reduction in fresh water usage for irrigation. Furthermore, it was determined that drainage water collected from the greenhouse contained 59% of applied N and 25% of applied P. These studies, though small scale and based on local drainage reuse, are very encouraging.

Existing work though promising, is based on spatially and temporally-limited manual sampling of soils and waters and on hypothetical guesses as to the processes involved in the N cycling. Furthermore, 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 [29, 30], restricts wide adoption of this mechanism in agriculture. Despite tremendous promise of the benefits of drainage water reuse and technological advancements, implementation of an intelligent and autonomous management mechanism has not kept pace with the deteriorating water situation. Some of the reasons as outlined in a detailed study [28] are; (i) insufficient awareness of available technology, (ii) unavailability of soil, weather and crop data, (iii) inappropriate model selection which inadequately capture the system details, and (iv) gap between decision makers and scientists. The integration of useful and relevant scientific information is necessary and critical to enabling informed decision-making for drainage reuse or disposal [31]. Recent adoption of WSNs in agriculture and hydrology presents huge promise for improving water management, and the next section discusses the applications of WSNs for water quality monitoring and agriculture and identifies huge opportunities available with real time, dense and remote data availability.

2.2 Introduction to WSNs in Hydrology and Agriculture

Advancements in micro-electro-mechanical systems (MEMS), low-power and low-cost microcontrollers, and radio modules have enabled WSNs for environmental monitoring [82]. The use of WSNs overcomes the limitations of previous expensive, and bulky monitoring equipment with low spatio-temporal resolution to a certain extent. WSNs are networks of small embedded computers referred to as sensor nodes or 'motes', spatially distributed to cooperatively monitor an environment and transmit data wirelessly. The relatively low cost of a WSN allows in principle, the deployment of a dense population of nodes that can adequately represent the variability present in the environment.

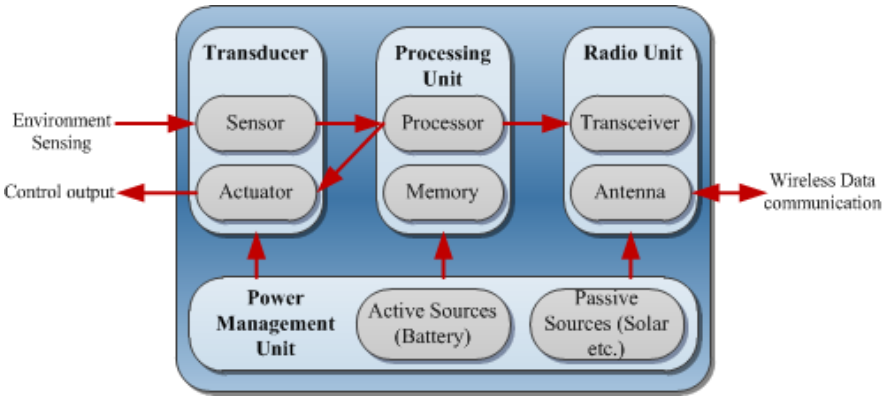


Figure 2-1: Sensor node architecture

WSNs have been used for monitoring and managing farming activities [36, 37] and eco-hydrologic processes [38, 39]. Water quality parameters such as pH, dissolved oxygen (DO), turbidity, salinity and nitrates have been measured in the reviewed applications using WSNs.

A basic sensor node, as shown in Figure 2-1, contains units for sensing and actuating, data processing, RF communication and power supply management [83]. A node can have many different types of sensors, for example measuring humidity, temperature, turbidity and nitrate but it should be noted that the quality of the measured data is dependent on the quality of the sensor. The acquired analogue samples are converted into digital signals using an on-board analogue-to-digital converter. The microprocessor may collect samples over time and aggregate them before transmitting them through the local radio transceiver to another node or a data sink (also called a gateway). Information collected from all the nodes is aggregated at the data sink and is generally transmitted to and analysed at an off-field control node where decisions might be taken. In some cases on-field control nodes are also employed. This could involve sending some actuation signal back to the network, for instance turning on sprinklers to irrigate the field. Furthermore, alternative networking approaches exist where information is processed and used entirely within the network, without the need for a central control node.

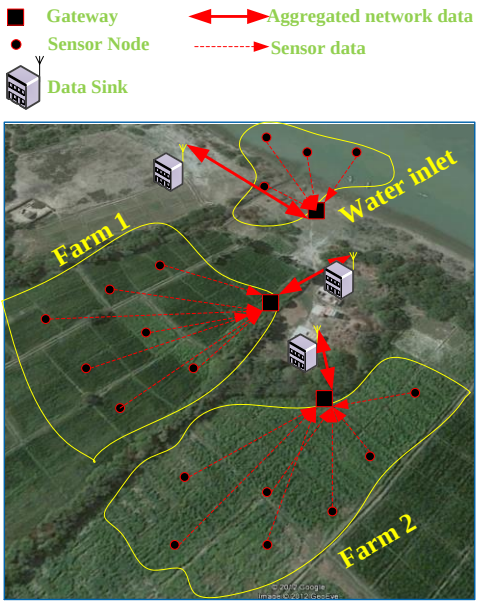


Figure 2-2: Data flow through WSNs deployed in a catchment

Figure 2-2 illustrates an example of autonomous networks deployed in individual farms and water inlet with off-field control nodes. In the existing applications, coverage areas varies from small individual fields to a few hectares of land.

2.2.1 Application of WSNs for Monitoring Water Fluxes and Quality in Different Environmental Compartments

WSNs have been employed for water quality monitoring of surface water (rivers, lakes, and dams), ground water, soil-water and catchments. These water bodies receive contaminants as a result of agricultural activities in the catchment (as discussed in chapter 1 and shown in Figure 1-1). Preliminary results have been fundamental in advancing the knowledge of contaminant trends, through spatially constrained observations. These applications are discussed below with respect to the area they monitor in a catchment.

2.2.1.1 Surface Water Monitoring

A test deployment project, 'Lakenet', was undertaken for water quality monitoring in a eutrophic lake by University of Notre Dame [47]. The sensor pod system, consisting of MICA2 and MDA300 sensor board measuring temperature, DO, and pH was deployed for a short term (10 days). The primary objective of this deployment was to test the design of both the pods and the network under field conditions and to examine its ability to collect data continuously and reliably. This project was defined as successful on the basis of collected data which showed diurnal fluctuations.

To demonstrate the need for availability of high temporal rate data, as a requirement of the Water Framework Directive [20], a WSN project (DEPLOY) was deployed at five selected sites along the length of the river Lee in Ireland. These sites extended from the Inniscarra reservoir to the Trivoli Dorks in Cork city. It achieved continuous sampling of parameters including conductivity, pH, chlorophyll, DO and temperature. The system consisted of the Tyndall programmable system-on-chip sensor interface and data telemetry system using IDS data pod, and an optimized RF section to provide up to 4km transmission. Results showed fluctuations in measurements due to activities such as runoff, rainfall or release of water from dams in the river. This demonstrated the potential for catchment behavioural analysis using statistical processing and interpretation. However, sensor fouling and maintenance issues were raised [38, 46].

Similarly, another deployment involved performance evaluation of a prototype network of wireless nodes (Sunspot) on a dam in Malawi, Africa to counter severe health threats caused by degrading water quality. Water parameters such as acidity (pH), DO and turbidity were measured by integrating a sensor board with a sensor node. These measurements were then compared with numeric standards to decide on the suitability of water for drinking [48].

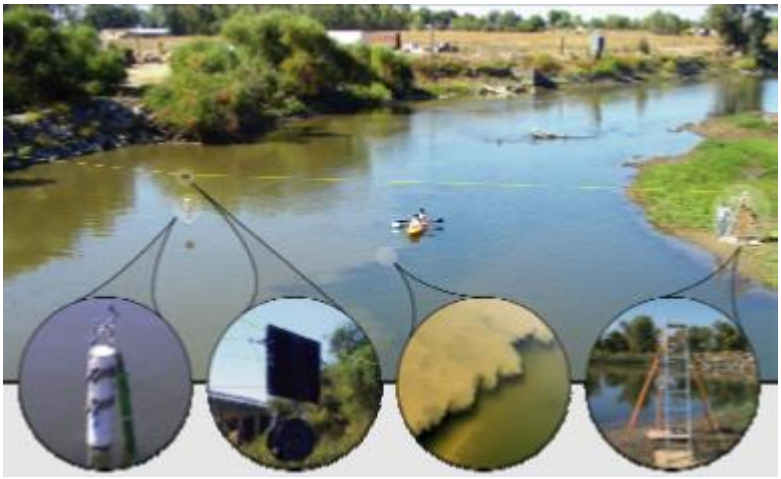


Figure 2-3: Scanning of flow and water quality using NIMS RD system within the San Joaquin (reproduced from [84])

To enable contaminant transport assessment, and urban stream quality and algal dynamics' characterization, considerable research has been undertaken by CENS (Centre for Embedded Networked Sensing), UC Los Angeles on developing the required technology [84, 85]. This started in 2007, by developing a test-bed, using automatic irrigation control, for salinity, nitrate and soil moisture control in a 12 ha land area [86]. Later progress on developing nitrate sensors for groundwater, soils and aquatic systems was made in 2009 [87]. In 2010-2011, the project continued to focus on high resolution river observations using rapidly deployable networked infomechanical system (NIMS RD) (as shown in Figure 2-3). Sensing is focused on temperature, DO and nitrate changes in the river [84]. For monitoring impact of poor water quality on aquatic life, parameters such as pH, DO, turbidity, salinity and electrical conductivity have been measured in various applications [88, 89].

2.2.1.2 Soil-Water Monitoring

Monitoring drainage water of farms enables assessment of its quality. In this application [51], wireless lysimeter sensing stations were used 90 cm below the soil surface to monitor water flux leached below root zone under an irrigated cropping system. Continually monitored soil-water and weather data, sensed by a distributed network of sensors, was wirelessly transmitted to a base station using Bluetooth, which was connected to a web server for broadcasting the data via the internet. This provided remote online access to drainage water flux and field conditions for the whole growing season. This was targeted at measuring volume of drainage water based on irrigation frequency and rainfall events for different crops. However, this work could be easily extended to also monitor nutrient or salt parameters leached out from the soil. Another important

project demonstrated the potential of WSNs for measuring soil water content, which was shown to be helpful in controlling hydrologic fluxes [52].

2.2.1.3 Groundwater Monitoring

Groundwater contamination from human impacts is less obvious than surface water and cleaning it is more costly and technically demanding [90]. Once polluted it may transport nutrients and chemicals to coastal waters [91, 92]. WSNs have been used to measure rate, quantity and direction of water contamination flow in two different scenarios, one being to understand prevalence of arsenic in Bangladeshi ground water, and the other to monitor nitrate propagation through soils and groundwater in parts of the USA.

In order to monitor saltwater intrusion into coastal aquifers, WSN has been deployed for monitoring water quality in a coastal area in Australia. Preliminary data showed potential for supporting decisions related to the placement of extraction bores for irrigation to avoid groundwater degradation. However, it was not possible to set up a well-connected radio link in this particular humid environment [50].

2.2.1.4 Catchment Monitoring

Some work has been reported for water quality monitoring at catchment scale. Such catchments include mountains, forests, lakes, rivers, and agricultural lands. These are prototype deployments aimed at developing the technology and algorithms for large-scale remote monitoring and data processing for large volumes of heterogonous data for water quality. These are fundamental research efforts in this domain but improvements can be made.

To find causes and interactions of physical/chemical variability on rivers and estuaries, work was started by US geological survey on creating a wireless catchment in Santa Margarita ecological reserve in Southern California. It was intended for high density monitoring of the catchment encompassing mountains, canyons and river channels. The system was primarily focused on developing and testing a network of radios and routers for wireless data transmission and its availability on the internet. Various issues were faced due to the rugged topography and the difficulty of obtaining line-of sight communication. Despite these challenges, this provided an effective way of remote monitoring and was used further for monitoring meteorological properties related to pollutant loadings [55]. Similarly, for evaluating variations of regional weather, at mountain-to-sea scale, on water quality/quantity, and flooding in a catchment in Hawaii, a

prototype system to provide 3D geospatial data visualization utilizing WSN was proposed. The system consisted of an array of nodes consisting of water quality sensors (temperature, pH, conductivity, DO, and turbidity), weather stations, sensor data loggers (InteleCell), and video monitors. Preliminary results showed improvement over traditional sampling methods [53].

The 'SoilWeather' network was used in Finland for providing high temporal resolution data on weather conditions, soil moisture and water quality [54]. The work was reported to have been used in developing a leaching model at sub catchment scale, including an improved hydrological model of the river, for monitoring water quality and nutrient retention in rivers and wetland. The network was deployed on a catchment covered by forests, river, lakes, and agricultural lands owned by private owners, who were the main users of the data. This is an important work highlighting data quality and processing issues with high volumes of heterogeneous data, received from catchment-scale observation using a single network, varying in content, accuracy, and spatial-temporal resolution. Additionally, maintenance of sensors due to such issues as bio-fouling, freezing, and rain gauge problems, which resulted in erroneous and missing data were viewed as major challenges in obtaining high data quality.

2.2.2 Application of WSNs in Monitoring Agricultural Processes

Some of the key research areas in precision agriculture which have received considerable attention include microclimate monitoring for pest control and precision irrigation [36, 37]. Precision agriculture is defined as a farm management system which utilizes farm observation and monitoring in order to respond to the field requirements [93]. Instead of taking decisions based on hypothetical average conditions, the real time information from these systems provides a solid basis for farmers to adjust strategies at any time [37]. There is a potential of leveraging existing networked agricultural activities for their extension into an integrated water quality management mechanism. In this regard, these agricultural practices, monitored and controlled using WSN, are discussed below.

2.2.2.1 Microclimate Monitoring for Pest Control

Microclimate monitoring enables the measurement and understanding of many parameters such as soil condition, local temperature and humidity, nutrient level in the soil and in water used for irrigation. Such information enables decisions, such as when to irrigate or fertilize, or risk analysis, such as the likelihood of pest infestation. To determine and prevent the onset of a fungal disease, phytophtora, in a potato field, a network of sensor nodes was deployed to sense humidity

and temperature. Sensed data was sent periodically to a gateway at the edge of the field from where it was forwarded to a server using Wi-Fi. The accumulated field data for a day was then fed into a decision support system server where it was combined with weather forecast to determine the treatment policy each morning, rather than spraying pesticides based on hypothetical assumptions.

This experiment, though based on specialized hardware, still presented many challenges for radio coverage owing to crop canopy effect, routing protocol, software bugs and in-field deployment [94]. It has had tremendous impact on improving deployment and practicality issues in WSN applications [50, 95]. Furthermore, to prevent frost, pest and mould damage in vineyards, which are prone to quick damage if overstressed, WSNs have been used to monitor ambient temperature, soil moisture and soil temperature [40, 41]. This information enables vineyard owners to monitor the plants remotely and manage priorities accordingly [96]. These pest control applications could be extended, to include chemical sensing, in order to assess chemical levels in the crops and soil after pesticides and herbicides applications. Information obtained through such a mechanism could then be used to inform on and thus potentially control chemical fluxes into water bodies.

2.2.2.2 Animal Farming

In order to maintain and improve animal health within animal feeding operations (AFO) and to prevent its negative impact on the ecosystem, information related to the local environment is very important. Moreover, by having a mechanism to monitor and control animal herds in pasture lands, pathogen and organic additions to the environment may be controlled and assessed. The deployment of WSNs in many AFOs has enabled monitoring of parameters such as temperature, humidity, noise and ammonia content in the air which has allowed the management and maintenance of healthy environments in these facilities [44, 45].

For managing cattle herds in pasture land, an algorithm called ‘moving virtual fence’ was developed. Each animal in the herd was given a smart collar consisting of a GPS (for determining animal’s location), a PDA, a WLAN card and a sound amplifier. As the animal approached a perimeter, it is presented with an acoustic stimulus to drive it away from crossing over a designated virtual boundary. By dynamically shifting fence lines, the usage of feedlots is improved. In addition, it reduced the overhead of installing and moving physical fences [97]. Other work has investigated online monitoring of cattle presence and pasture time using WSNs to register the time animals spend in specific areas of the field [98]. This information can be used

to indicate grass quality and quantity and can also determine the right time to provide access to a new grass strip.

These applications can be clearly extended to monitor pathogen levels of animal waste in a barn and their fluxes to water bodies and therefore manage it by raising alerts. Also, by managing animal herds with a virtual fence, the possible situation of them contaminating water bodies or other fields can be avoided or reduced.

2.2.2.3 Precision Irrigation

In a US geological survey done in 2005, it was found that irrigation withdrawals were 62% of the total fresh water withdrawals for the nation [99]. Therefore an efficient way of utilizing water resources by people in general and agriculture sector in particular is crucial to avoiding a global water crisis [100].

By monitoring soil and environmental parameters (such as soil moisture, soil temperature, ambient temperature and humidity), an informed decision can be made by farmers to control the timing and quantity of irrigation. Unlike expensive, sophisticated, high maintenance and high energy sensor-based technologies like Adcon Telemetry and Automata, and labour intensive data collection from logging stations, WSNs provide real-time and less expensive sensing mechanisms for improving irrigation automation [43].

Another example made a comparison between traditional and sensor scheduling systems in a cotton field (2.3 ha) for the whole growing season [43]. The East of the field, divided into four different irrigation zones based on soil type and historic yield map, was equipped with 9 sensor nodes and centre pivot variable rate irrigation (VRI) system. On the western side, 3 sensor nodes were deployed to monitor soil water tension, and here irrigation was triggered based on traditional assessment of crops. Each sensor node was equipped with three moisture and temperature sensors (tensiometers) buried at 0.2, 0.4 and 0.6 m depth under the soil. Results showed water tension at more than double the trigger points in western zone despite irrigation, whereas the eastern side never surpassed the trigger points. However, it is stated that further work needs to be done on the VRI controller for varying quantity of irrigation relative to the needs of individual areas, or for different crops in a single field. To address such a need, site specific VRI was employed in a 3.6 ha plot by using six in-field sensor stations. The irrigation machine was controlled by a programmable logic controller that would update the geo-referenced location of sprinklers from a differential GPS, and wirelessly communicate with a computer at the base station. This enabled variable rate irrigation according to the needs of individual crops in a large field. [101].

As discussed above, a significant amount of successful work has been published using automatic and variable rate irrigation with some limitations regarding reliability. However, what is lacking is the inclusion of real time information about events elsewhere in the farm to alter irrigation decisions. For example, information about possible weather changes, like rain, from other networks in a catchment could be used to change time and quantity of irrigation by allowing forward planning and prediction. Also, information regarding irrigation in one farm can be used elsewhere in the catchment to enable reutilization of discharges in another farm.

2.2.3 Discussion on research gaps in the existing applications of WSNs

The literature review for the application of WSN for water quality monitoring and agricultural activities presents many possibilities and potential, yet identifies numerous limitations. Following are the major gaps commonly observed in the reviewed literature.

2.2.3.1 Absence of catchment scale integrated monitoring and management

WSNs have been successful in facilitating the understanding of the trends of nutrients and contaminants fluxes owing to non-point source activities in a catchment. However, this understanding is limited as it can currently only use small-scale, localized and specific measurements at rivers, lakes, dams and groundwater reservoirs [38, 46-50, 102, 103]. These water resources receive contaminants as a result of activities in the catchment, thus providing insufficient information for quantifying explicit responsible activities. Most of the applications were small scale prototype deployments and were mainly aimed at developing and testing the technology for remote, continuous and real-time data availability. Fewer applications are intended for large-scale observations [53-55], but these are prototype deployments based on limited monitoring points across the catchment. Furthermore, these applications rely on GSM or internet to enable availability of data, originating from various networks, on a central repository [53, 54]. Such a reliance renders this solution unsuitable in a remote area of crop lands. In addition, using a central processing unit for high resolution heterogeneous data from multiple networks leads to processing and filtering overhead, and delays decisions. This establishes a need for a distributed large-scale and multi-level monitoring mechanism.

WSNs have been used in agricultural activities but not specifically for water quality monitoring. For example a number of applications have been developed for precision irrigation for automatic and controlled irrigation, reducing the chances of water wastage (runoff which is a carrier of

nutrients and contaminants) [42]. However, runoff coming from elsewhere (e.g. another farm) cannot be avoided. This runoff can be carrying nutrients. In that case, the current precision irrigation system could be extended to use information about runoff to modify the irrigation and fertilizer application of its field to save water and fertilizer usage.

Moreover, as already discussed in the introduction chapter, that reuse of reclaimed drainage water reduces environmental impact on the river network. However, no technical mechanism exists to facilitate management and reuse of reclaimed water in an autonomous and dynamic manner among farms. These applications have yet to explore the implementation and impact of this technology for management and control decisions, in order to minimize and prevent individual stakeholder's contaminant contributions, in an autonomous and dynamic manner. When this happens, the loop would be closed and better management policies by farmers would come into place. If methods existed, then networks for small-scale, individually targeted stakeholders in a catchment could be included in an 'integrated' and 'collaborative' monitoring mechanism to share information about runoff, contaminants or nutrient fluxes.

2.2.3.2 Absence of suitable sensors

Table 2-1 lists various sensor modalities available for use in WSNs. It is evident through the surveyed literature that existing work does not monitor all of the contaminants required for water quality monitoring, for example phosphorous, nitrates, and ammonia. Although simple and low-rate sensors, such as those typically used for monitoring temperature, humidity and wind, require little energy, this is not the case for other sensor modalities [39]. For example, the high cost of some sensors (e.g. ammonium) is a major barrier for dense deployment of WSNs [49]. Furthermore, suitable in-situ phosphate and nitrate sensors for terrestrial deployment are not available. The closest suitable sensors for nitrate monitoring are based on ultraviolet absorption and flow cell analysers which are very expensive for dense deployment [104]. Both electromechanical [105] and optical sensing modes [106] relevant to nitrogen and phosphorous species are still an active research area [39]. Furthermore, research on the development of miniature, reliable and inexpensive water quality solid-state sensors using metal oxide sensing electrodes is underway by CSIRO, CMSE and SSN TCP laboratories [107].

Table 2-1: Major sensor modalities with specifications on cost, reliability and power requirements (reproduced from [39])

Sensor category	Example	Comments
Physical	Temperature (e.g. thermocouple, thermistor, IR sensor)	Inexpensive to intermediate cost, reliable, low power requirements
	Relative humidity	Intermediate, reliable, low power
	Leaf wetness	Inexpensive, reliable, low power
	Soil moisture	Inexpensive to moderate, issues with calibration and measurement units, low power; many choices
	PFD, total irradiance	Intermediate, reliable with calibration issues, low power
	Wind speed and direction	
	Cup anemometer	Inexpensive to intermediate, reliable, fails at low wind speed, low power
	Hot wire anemometer	Intermediate, less reliable, higher power
	2-D/3-D sonic anemometer	Intermediate to expensive, very reliable, moderate power
Chemical	Atmospheric carbon dioxide	Expensive, reliable, moderate power, requires careful calibration
	Soil carbon dioxide	Intermediate, reliable, low power, calibration?
	Soil carbon dioxide efflux	Expensive, reliable, moderate power, requires careful calibration
	Nitrate sensor	Expensive, under development for reliable terrestrial deployments
	Phosphorus sensor	Not available for terrestrial deployments
Biological	Digital imagers	Moderately expensive, reliable, moderate power; high bandwidth, software requirements
	Minirhizotron camera	Expensive, variable power requirements
	Sap flow sensors	Commercial probes moderate, control system needed; calibration issues
	Acoustic sensors	Moderate, reliable, moderate power, high bandwidth; software needs

Moreover, in the absence of suitable inexpensive and miniature in-situ phosphate and nitrate sensors for terrestrial deployment, for which development is underway [107], sophisticated analytical instrumentation such as chemi-bio sensors present inherent limitations because of the number of samples that they can take [108]. In order to maximize the deployment lifetime, these power hungry sensors should adaptively sample at higher resolution only when discharges occur. These discharges may be intense but are short-lived with high nutrient flux flushing away in the beginning. This has the potential to result in the misclassification of water quality status [35] which requires continued reporting of water quality status. In order to capture discharge and pollutant events in real time, prior information of such an event would be required to adjust sampling at higher rates. However, there is no mechanism for such event information to be available to monitoring networks.

2.3 Modelling Hydrological Discharges

In the development of the hydrological predictive models, a range of more traditionally used physical, and lumped conceptual models along with more recently adopted empirical based methods are used [109]. Below an overview of these models, with respect to their strengths and limitations, is presented.

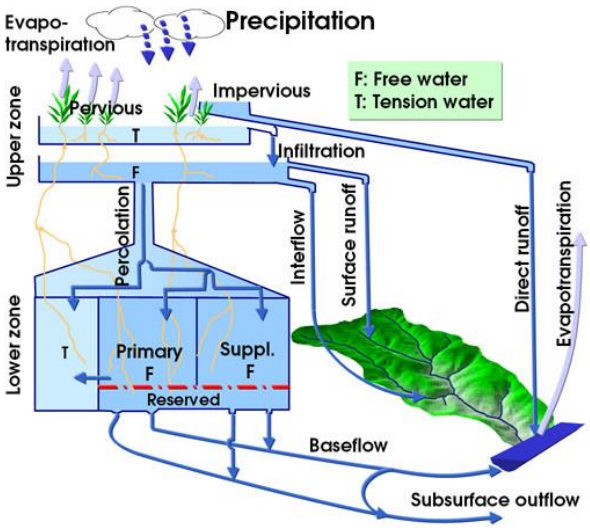


Figure 2-4: Physically based hydrology model - SAC-SMA (reproduced from [110])

2.3.1 Traditional Modelling

Hydrological phenomena are extremely complex, highly nonlinear and exhibit a high degree of spatial and temporal variability [70]. There are various traditional hydrological models, based on physical and conceptual basis, to predict rainfall-runoff, irrigation-drainage, and pollutants discharge dynamics [110-114]. The general frameworks that supports some well-known models are discussed below. Traditional models infer the discharge dynamics of pollutants (surface and ground) into the river, and subsequent stream flow using mathematical implementations (through partial differential equations) of the processes governing transport through the soil and the river channels [115]. In other words, most hydrological models are represented by a sequence of 'stores' and 'fluxes' between stores. An example is SAC-SMA (Sacramento Soil Moisture Accounting model), which models surface runoff by considering three different flow zones: a zone describing the direct runoff from rain falling on impervious soils, a zone describing water flowing into the river after exceeding the soil moisture capacity of pervious soils, and a zone describing runoff occurring after soil moisture capacity is exceeded above water impervious regions [116]. Figure 2-4 illustrates these zones. Furthermore, various hydrology models have been developed under EPA, which include watershed models like Storm Water Management Model (SWMM), and water quality models like Water quality Analysis Simulation Program (WASP) [114, 117].

In order to predict runoff and drainage, a model requires knowledge about various permanent and transient parameters including meteorological conditions, rainfall/irrigation depth, river/channel

network parameters and the soil's time dependent response to the event. This response is dependent on land cover, topography, soil composition; hydrological properties of the soil and antecedent moisture condition of the soil (see Figure 2-5 as an example for flood discharge).

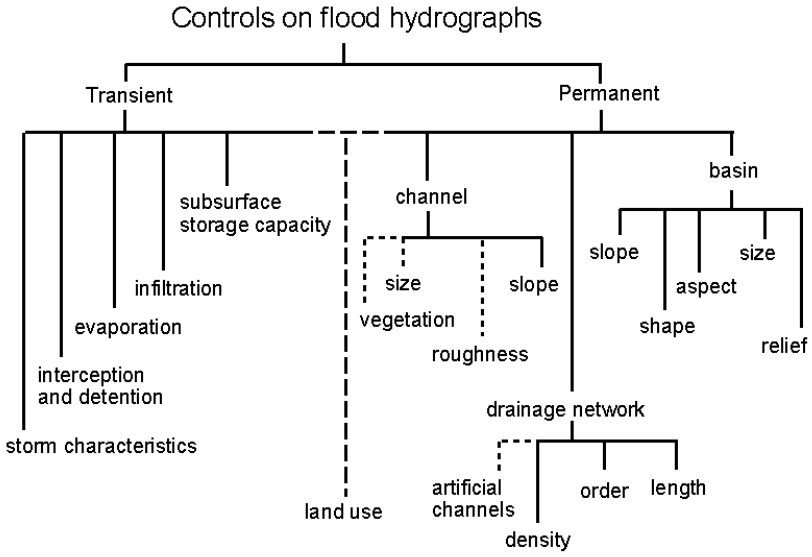


Figure 2-5: Parameters affecting discharge hydrograph in a river (reproduced from [118])

With regards to drainage modelling, a popular model is DRAINMOD which is a process-based and field-scale model to describe the hydrology of poorly drained soils based on water-balance calculations [119]. In comparison to other drainage models, such as WAVE and SWAP (Soil-Water-Atmosphere-Plant), DRAINMOD has the advantage of giving more accurate estimate of discharges [120]. The DRAINMOD model has now been extended to watershed scale by including routing methods for surface runoff from fields to predict flow rates in drainage channels and streams. The model simulates output such as infiltration, subsurface drainage, surface runoff, evapotranspiration, vertical and lateral seepage, water table depth etc. [121]. DRAINMOD is a comprehensive, hence a complex model requiring a plethora of variables related to weather, soil properties, seepage details, and drainage system specifications.

One of the most popular and simpler hydrological models is the Soil Conservation Service Curve Number (SCS-CN) method. It has been recently renamed as the Natural Resource Conservation Service (NRCS) Curve Number method [122]. This method is used to estimate the volume of direct surface runoff and, its response and travel times for a given rainfall event. This model, although designed to estimate direct surface runoff from a given rainfall event, provides huge potential and basis for simplistic hydrological modelling. Below the mathematical models for the

volume of direct surface runoff, its response and travel times, is presented along with a discussion on the strengths and limitations of these models.

2.3.1.1 NRCS Curve Number Method

SCS runoff curve number method is used for predicting direct runoff or infiltration from excess rainfall after initial abstraction (see Figure 2-6 for detailed parameters list). This method has been discussed at a great length in the recent hydrologic literature [123, 124]. The scale over which the SCS curve number method has been applied ranges widely from 20 acres to 150 acres of drainage area [125]. Runoff curve number is based on the area's hydrologic soil group, land use, hydrological condition and antecedent moisture condition.

Using this method, direct surface runoff (in) is calculated as follows ([126, 127]);

$$Q = P - (I_a + F) \quad (1)$$

Where, P (in.) is the rainfall depth, I_a (in.) is initial abstraction and F (in.) is the retention after runoff begins. Q (in.) is direct runoff which consists of channel runoff, surface runoff and subsurface runoff. I_a is the sum of all rainfall losses incurred by either retention in surface depression, or through evaporation or infiltration. For the development of the SCS runoff equation, Q and P are measured from the field, whereas a generalized relationship is developed for I_a and F based on observed data. I_a and F are a function of S, which represents hypothetical limit of storage. S is defined as potential maximum retention after runoff begins.

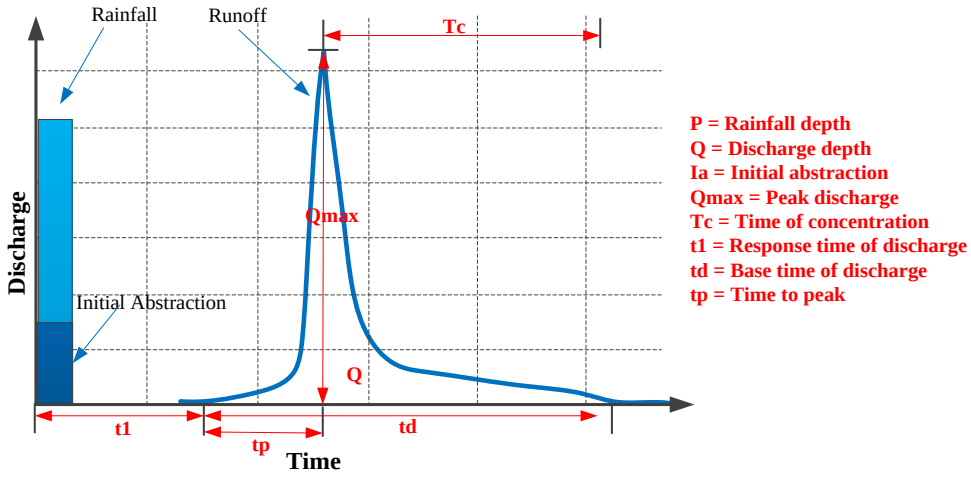


Figure 2-6: A discharge hydrograph

To develop the SCS runoff equation from the water balance equation (Eq. (1)), two assumptions are considered. Firstly, the ratio of the amount of water retained (F) to the maximum potential

retention (*S*) is equivalent to the ratio of the amount of runoff (*Q*) to the maximum rainfall available for runoff (*P* - I_a). This is expressed as;

$$\frac{F}{S} = \frac{Q}{P - I_a} \tag{2}$$

Substituting the value of *F* from Eq. (2) in Eq. (1), Eq. (1) becomes;

$$Q = \frac{(P - I_a)^2}{(P - I_a + S)} \tag{3}$$

The second assumption is that I_a is 20% of the maximum potential retention of soil. This is based on the mean observed data for various sites. It is expressed as;

$$I_a = 0.2\,S \tag{4}$$

Substituting I_a in Eq. (4) results in;

$$Q = \frac{(P - 0.2\,S)^2}{(P - 0.8\,S)} \tag{5}$$

Later, *S* was converted into another variable *CN*. As compared to S which had a range of values from 0 to infinity, range for *CN* is limited by 0 and 100. The practical range of *CN* is from 40 to 98. 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) [128]. Land use type include row crops, pastures, legumes, meadow and woodlands etc. Different land treatments comprise straight row, contoured and terraced practices. Soils are divided into hydrological groups A, B, C and D according to their profile and infiltration capacity. Soil group A has high infiltration meaning low runoff, whereas soil group B has moderate infiltration and moderate runoff. Soil group C has moderate to high runoff due to low infiltration, whereas soil group D has very high runoff due to very low infiltration. A table listing NRCS curve number assignment, with respect to the dependent field variables, is attached at appendix C. The soil moisture condition in the drainage basin, before a runoff occurs, is another important factor influencing the final value for *CN*. In the Curve Number Method, AMC is classified into three groups:

- AMC I: The soils in the drainage basin are practically dry (i.e. the soil moisture content is at wilting point).

- AMC 11: Average condition.
- AMC 111: The soils in the drainage basins are practically saturated from antecedent rainfalls (i.e. the soil moisture content is at field capacity).

These groupings are determined on the basis of 5-day antecedent rainfall depth value (i.e. the accumulated total rainfall preceding the runoff under consideration). Table 2-2 illustrates the range of 5-day rainfall values for the three AMC groups.

Table 2-2: Seasonal rainfall limits for AMC classes (reproduced from [119])

Antecedent Moisture Condition Class	5-Day Antecedent Rainfall (in)		
	Dormant Season	Growing Season	Average
I	< 0.5	< 1.4	< 0.9
II	0.5 – 1.1	1.4 – 2.1	0.9 – 1.5
III	> 1.1	> 2.1	> 1.5

The relationship between *CN* and *S* is given as;

$$S = \frac{1000}{CN} - 10 \tag{6}$$

Substituting the value of *S* in Eq. (5);

$$Q = \frac{\left[\left(P - 0.2\left(\frac{1000}{CN} - 10\right)^2\right)\right]}{P + 0.8\left(\frac{1000}{CN} - 10\right)} \tag{7}$$

For runoff volume, following relationship is used;

$$Runoff\ Volume = Q(ft) * Area\ of\ drainage(acres) \tag{8}$$

For computing t_d, it is evident from Figure 2-6 that t_d can be expressed as;

$$t_d = T_c + t_p \tag{9}$$

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.

$$T_c = \frac{0.007\ (nL)^{0.8}}{(P_2)^{0.5}\ (S)^{0.4}} + \frac{L}{3600V} + \frac{L}{3600}\left(\frac{n}{1.49\ (R)^{\frac{2}{3}}\ (s)^{0.5}}\right) \tag{10}$$

Where, L is length (ft.) of flow pattern, n represents surface roughness, 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./sec) of water.

Hence, Eq. (9) becomes;

$$t_d = t_p + \frac{0.007\ (nL)^{0.8}}{(P_2)^{0.5}\ (S)^{0.4}} + \frac{L}{3600V} + \frac{L}{3600}\left(\frac{n}{1.49\ (R)^{\frac{2}{3}}\ (s)^{0.5}}\right) \tag{11}$$

As per the author's best understanding, there is no direct mathematical equation to express *tp* in the NRCS method. The other parameter required is t_l, 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.* [129]).

2.3.1.2 Limitations of the NRCS Curve Number Method

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. (3) and Eq. (11)). 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 [130]. In addition, the 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_l and t_d is computationally intensive. Additionally, for I_a a mean value of 20% was considered which may work for many regions,

however in many studies I_a was observed as less as 5% to 15% [131-133]. Therefore, a constant value for I_a is not applicable for every region.

2.3.2 Data-Driven Modelling

During the last decade, the area of empirical modelling received an important boost due to developments in the area of WSNs and machine learning. With contemporary WSNs, it is now possible to extract real time field data, which presents a huge opportunity for the development of simpler models. These models can complement or replace knowledge-driven models describing behaviour of physical systems, and therefore can yield models of low computational complexity, making them well-suited for implementation on a network. Using machine learning algorithms, the models are trained on the historical data describing the phenomenon in question.

Data-driven modelling, using machine learning algorithms, has been widely used in hydrological modelling [60-62]. Examples of the most popular supervised learning methods used for hydrological predictive systems are: statistical methods, artificial neural networks (ANN), and decision model trees. An example of a statistical method is multiple linear regression (MLR). The algorithmic detail of these methods will be discussed in section 2.5 of this chapter.

MLR is a linear model and has been used for flood forecasting, in which a model is learnt based on data for parameters such as precipitation, air temperature and river flow [57]. However, the training dataset is collected from historical data of 7 years. Due to the linear modality of MLR, it is reported that there are other algorithms that perform better and are more robust when compared to MLR [61]. One such algorithm is ANN.

ANNs have been a popular choice for rainfall-runoff modelling and stream-flow forecasting [60, 61, 63, 64]. It is stated that attractiveness of ANN for flood forecasting is mainly because; ANNs can represent a non-linear function, ANNs can identify relationships between input parameters, ANNs perform in a robust manner even in the presence of faulty or missing data [63]. An ANN in most cases is used as a black-box representation of hydrological processes, with input parameters of rainfall and inflows [134]. In one study, while investigating river forecasting in two flood-prone gauged UK catchments, it was possible to construct a robust model for a small data set which performed with comparable results with respect to the traditional forecasting systems [63]. Furthermore, another work used weekly values for precipitation and evaporation rates for a river basin in southern England. The land cover of the river basin is mainly agricultural land. The weekly values were converted to daily mean rates and were used to simulate daily stream flows.

Various input combinations of rainfall and evaporation values spanning over several days were used to test the results, with a seven input model giving the best performance [60].

Due to the complex structure of ANNs, decision tree modelling has been investigated recently and adopted in hydrological forecasting [58, 65-68] and an interesting example of this class are M5 model trees [69]. The advantage of M5 model trees over ANNs are that they are faster to train [70]. M5 model trees have been used in hydrology for modelling drainage discharge [68], rainfall-runoff [70-72], streamflow [58], flood forecasting [73] and sediment transport in discharge [74]. It is important to mention that none of these applications except for the work by *Kuzmanovski* [68] were based on farm scale drainage discharges. In this work, the aim was to measure drainage discharges from fields, ranging from 0.83 ha to 1 ha, 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. The model performed with 45% to 66.9% predictive accuracy. The models developed by *Solomatine et al.* [73], *Corzo et al.*[71], and *Galelli et al.* [58] were all used to predict 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), which can influence flow rates vastly, for such large heterogeneous catchments can vary tremendously; hence a single average value may not represent the field conditions for the whole catchment. 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 most cases).

2.3.3 Discussion

Traditional modelling approaches, popular in academic research, are very suitable for large scale catchment applications of flood and stream monitoring. However, the need for calibrating spatially-distributed models to individual basins, reliance on numerous parameters, and the tremendous computational burden involved in running them makes their application complicated and, in resource-constrained underdeveloped areas, nearly impossible [135]. The NRCS curve number model although simpler than the other discussed methods has its own limitations of using proxy parameters.

With increased monitoring and real data availability, empirical modelling based on machine learning algorithms is prevalent in hydrological modelling. 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 [59]. However, these modelling projects are limited to flood forecasting and stream flow monitoring, in a gauged catchment, by using decades of historical data for training. As per the author's best knowledge, only the research work by *Kuzmanovski* [68] is similar in part to the objective of this research, i.e., modelling discharges from a farm. However, the model uses 20 years of historical data for field and climatic parameters. This highlights one of the main issues in that the historical data sets needed to develop predictive models do not exist for every farm, and for most catchments.

2.4 Modelling Nitrogen Losses

Fertilizers rich in phosphorous, potassium and nitrogen are added to soil to increase crop yields. However, agronomic nutrient recommendations are often far in excess of environmentally-safe levels [11]. Nitrogen (N) becomes a concern to water quality when soil N is converted into nitrate (NO_3^-). This form of nitrogen is not held tightly by soil particles, therefore residual nitrates in the soil after the growing season will be lost to leaching and runoff into ground and surface water [12, 13]. The rate of leaching depends on soil drainage, rainfall, amount of N present in the soil and crop uptake. Well-drained soils, low crop yield, high N inputs (especially outside of the growing season) and high rainfall are all conditions that increase the potential for nitrate leaching and runoff. Losses realised in this way result in dual concerns – decreased soil fertility and degraded water quality.

Two main approaches for nitrate loss modelling have been reported in the literature so far: the bottom-up approach, in which the physical mechanisms that contribute to nitrate losses in the N cycle are conceptualized, using mathematical equations, into *process based models*; and the top-down approach, in which mapping from the space of predictor space to that of the response variable is inferred totally and directly from observational data to a more general class of *data-driven models* [58, 136]. Depending on the objective of the modelling exercise, one approach can be more suitable than the other.

2.4.1 Traditional Modelling

The complexity of process-based models is the key to improving our understanding of the various processes involved in N losses, and is very advantageous in evaluating different scenarios and clarifying the basic principles of N transport within the soil-plant-water continuum. Some of those models are more hydrology-oriented with less details about N-biogeochemical processes, such as MIKESHE and MODFLOW [137, 138], while others have focused more on the NO_3^- leaching

processes but with little focus on hydrological aspects, such as CENTURY and SOILN [139, 140].

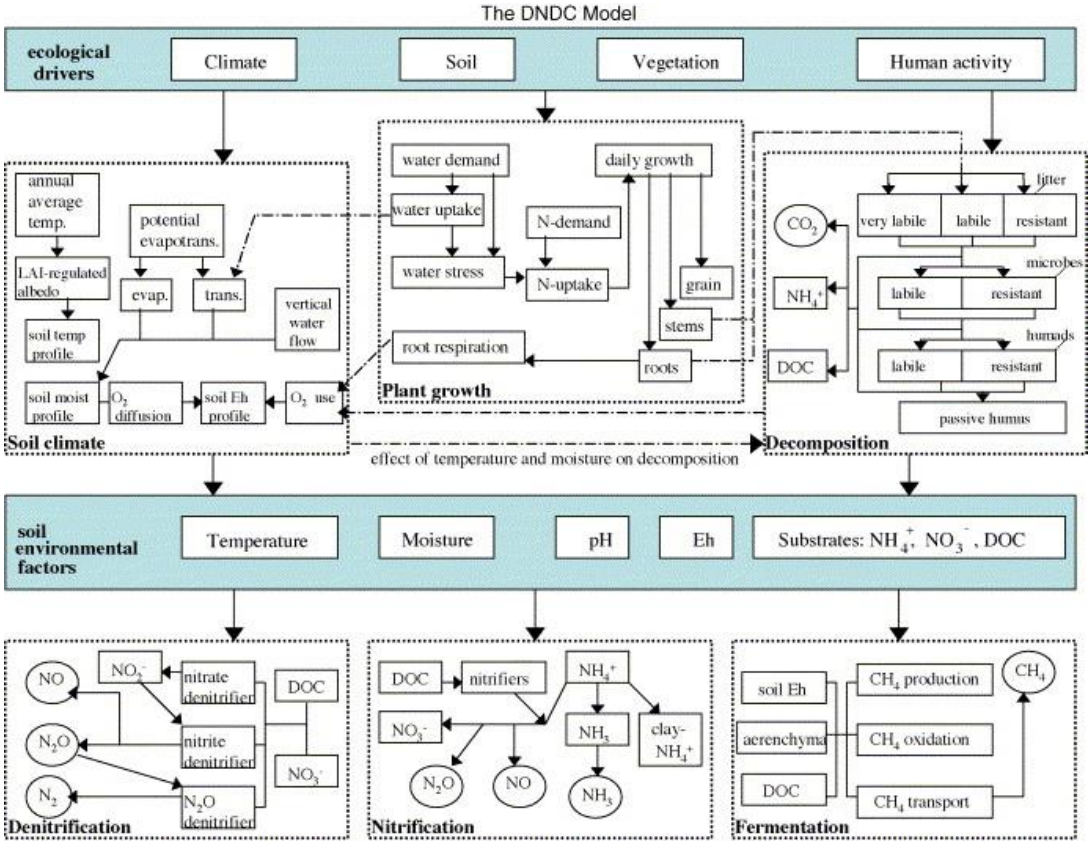


Figure 2-7: The structure of DNDC Model (reproduced from [141])

In comparison to above models, a modified De-Nitrification-DeComposition (DNDC) model possesses an N leaching module in addition to a complete set of N transformation processes parameterized by mathematical equations [142-144]. DNDC model consists of two major components driven by ecological drivers, and soil environmental factors as shown in Figure 2-7. The first component consists of soil climate, crop growth and decomposition sub-modules, which computes the soil environmental variables (soil temperature, moisture, pH etc.) based on the ecological drivers (climate, soil, vegetation, and human activity). The second component consists of nitrification, denitrification and fermentation sub modules, which predicts microbial activities and trace gas products based on the soil environmental variables [141]. DNDC is a comprehensive model, however its dependence on acquiring numerous parameters (above 20) renders its application quite complex. Furthermore, the miscalibration and over-parameterization results in low predictive capability of the model [58].

Another model used to simulate nitrate losses in artificially drained soils is DRAINMOD-N [145], an extension of the hydrological model DRAINMOD discussed earlier in section 2.3.1. DRAINMOD-N was developed to simulate a simplified version of the nitrogen cycle, with main emphasis on nitrate. The influencing parameters in the model include rainfall, fertilizers, plant N uptake, denitrification, surface runoff losses and sub-surface drainage losses [146]. It is based on water balance calculations of DRAINMOD to determine average soil-water fluxes. This model has been used in applications with promising results, such as in one study, the simulated results showed an average monthly deviation of 3 $Kgha^{-1}N$ [147].

2.4.2 Data-Driven Modelling

In contrast to traditional models, data-driven models have high prediction capability and require fewer parameters, which combine well with the computational burden of decision making [59]. Data-driven models in nitrate modelling uses two approaches: The export coefficient approach and the machine learning approach. The export coefficient approach is used to calculate the N loads delivered annually to a water body by summing the individual loads exported from each nutrient source within a catchment [148]. Examples of this approach include models developed for N losses from grasslands in the UK [149, 150] and from agricultural properties in Australia [151]. These models include sub-models with empirical equations based on annual values for all possible sources of N inputs and outputs, e.g. N input through fertilizers and the atmosphere, plant and animal N-uptake, N lost through dung/urine, N present in dairy products, N lost through plant death and decomposition. This approach generally works on annual time-steps, and is very suitable for policy making and management decisions regarding annual nutrient budgets. However, it is not appropriate for daily field scale management and control decisions.

The second approach of the data-driven models, i.e. machine learning, has been widely used in hydrological modelling; however, it has only recently been adopted in the modelling of N losses. In this regard, a modelling framework was developed to calculate annual nitrous oxide flux and nitrate leaching by abstracting the complexity of the DNDC model [65]. The input parameters (11 variables) consisted of annual values related to N application, soil chemistry, and climatic conditions were selected from the original DNDC dataset as shown in Figure 2-8. Various algorithms, such as multi-layer perceptron, random forests and support vector machine were used for comparison. Although this research effort reduced the number of parameters by a half, it still required 8000 training samples based on annual values, which were obtained from different sites, to achieve the optimal results.

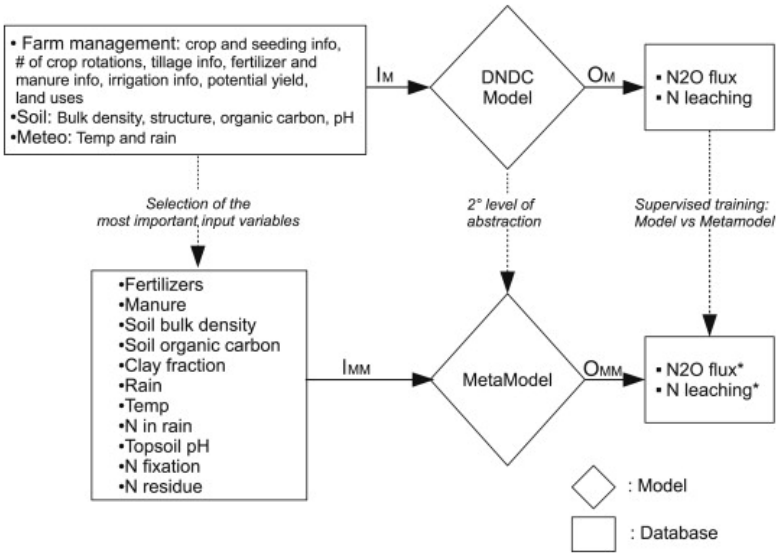


Figure 2-8: Selection of parameters abstracted from the DNDC dataset for the metamodel (reproduced from [65])

Similarly, another work further simplified the input parameters and training size requirement and used neural networks to simulate total N emissions from the de-nitrification process (N_2O) [75]. The training set consisted of only 536 records based on input variables such as water filled pore space, nitrate concentration, soil denitrifying potential, organic matter, soil pH, bulk density and soil depth. The model gave optimal performance, however, the model was developed for gaseous N emissions and still relied on soil chemical data. Another study [67] used a machine learning algorithm to build a meta-model (an abstracted model of a complex model) for a deterministic N leaching model called WAVE. Recently, regression methods were used in a two-year study to simulate seasonal nitrate concentration dynamics in soil water extracted from 36 suction lysimeters in potato plots about seven to eight times each year. The model achieved maximum performance, however it used percentage of clay and soil depth besides other input parameters, and was based on sparse yearly samples [66].

2.4.3 Discussion

Similar to traditional hydrological modelling, nitrate models using physical and mathematical basis, though comprehensive and useful for simulating nitrate losses, rely on numerous complex parameters. The reliance of models on acquiring chemical and geologic data, which either requires grab sampling and laboratory analysis or very expensive equipment, limits wide-scale adoption of this technology for high resolution output. Therefore, data-driven modelling has recently been

adopted for modelling complex N-cycling processes with the aim of simplifying modelling relatively fewer parameters. One such recent and promising example of low complexity modelling is done by *Villa-Vialaneix et al.* [65], which predicts annual nitrous oxide flux and nitrate leaching by abstracting the complexity of the DNDC model. The model reduces the input parameters to half as compared to the original DNDC model (11 variables), consisting of annual values related to N application, soil chemistry, and climatic conditions. However, the study used 8000 training samples, obtained from different sites, to get the optimal results.

Based on the literature review and to the best of the authors' knowledge, it is apparent that existing modelling has not been intended for predicting daily *TON* losses within the farm system with the aim of enabling reutilization and alerts in real time, i.e. by using WSNs. Furthermore, these data-driven models rely on thousands of training samples, containing biological and chemical data, acquired using traditional grab sampling methods. To enable an intelligent and real time mechanism for managing nutrients, influencing parameters should be simpler, and such which could be acquired autonomously. This is because constraints on WSNs with regards to battery life, and computing power requires simplified model for on-node deployment.

2.5 Machine Learning Algorithms for Environmental Modelling

In machine learning algorithms, the task of learning is often also characterised as supervised or unsupervised learning. Algorithms that require a set of data points with known outputs are referred to as supervised learning models. These are in contrast to unsupervised algorithms where the target outputs are not known. Typically, supervised learning models use the methods of computational intelligence to build (numerical prediction) models linking a system's input and output variables [71, 152]. This phase is called as the training phase. Such learning models are trained on 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 in the prediction phase [109]. Furthermore, if the outputs are classes, then the learning task is defined as a classification, whereas if the outputs are real values then it is referred to as a prediction [68, 153]. In the next sections, the most important supervised learning techniques are explored with respect to prediction and classification.

2.5.1 Machine Learning for Predictive Models

Below, the three most popular machine learning algorithms adopted in hydrological and nutrient predictive modelling (section 2.3.2 and 2.4.2), are discussed and will be used in this thesis – ANN, M5 tree algorithm, and MLR.

2.5.1.1 Artificial Neural Networks

The concept of ANNs is inspired by the biological neural networks of the human brain. Mathematically, an ANN is a complex nonlinear function with many parameters that are adjusted in such a way that the ANN output becomes similar to the measured output on a known data set. More specifically, a network is made up of a number of interconnected nodes (neurons) arranged in three layers – input, hidden and output [63], as shown in Figure 2-9.

The input layer simply forwards the inputs to the network without performing any computation. The links represent weighted connections between the nodes. A processing element multiplies the input by a set of weights and transforms the result into an output value. By changing the weights, ANNs work towards producing an output which is closer to the measured value [154]. This is also defined as feedforward networks or multiple layer perceptron (MLP). The number of input and output variables depend on the complexity of the problem. Fewer nodes may not be able to interpret a new data set and too many nodes result in longer computational time for training.

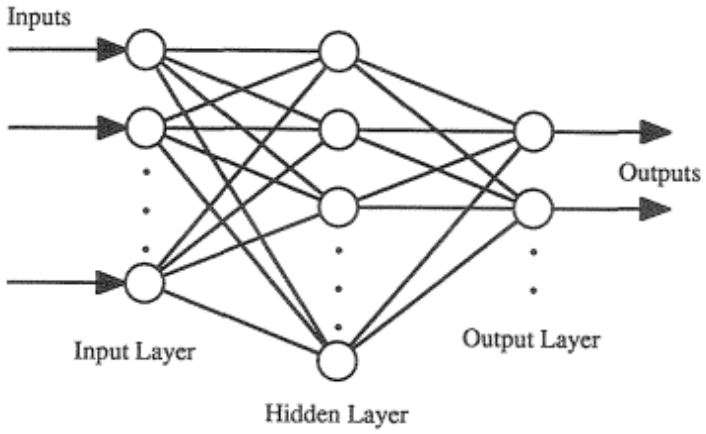


Figure 2-9: A basic diagram of artificial neural network topology (reproduced from [63])

Despite many advantages of ANNs, such as being able to represent a non-linear function and robust performance in presence of faulty or missing data, there are few disadvantages. One of the drawbacks of ANNs is that for a decision maker it is very difficult to analyse the structure of the

resulting ANN and to relate it to the outputs [70]. Furthermore, there are approaches to numerical prediction that use piece-wise linear approximations, which are much easier to interpret and can be trained in less computational time than ANNs, such as M5 decision model trees.

2.5.1.2 M5 Decision Model Tree

Due to the complex structure of ANNs, a rather simpler approach, decision model trees [69], has been demonstrated as an alternative [70]. 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 [152], where each leaf contains the target values. If the target value is a class, then it is termed as a classification tree, whereas if the target value is a real value then it is termed as a regression tree.

Decision trees have been shown to perform well when compared to other model types [58, 155] 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 [156]. The problem of over-fitting occurs when a model fits the random error or noise in the training data set instead of the underlying relationship. This problem transpires in an excessively complex model which has too many parameters in relevance to the number of training observations [157]. Over-fit models generally have poor predictive performance. This problem is solved by M5 tree algorithm.

The problem of over-fitting in decision trees 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 [58]. This is known as M5 tree modelling, and was first introduced by Quinlan [69] and applied to hydrological modelling by Solomatine [70, 73]. 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 [68].

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 2-10 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. 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) [73]. The M5 tree algorithm available through the MatLab toolbox M5PrimeLab [158] is used in this thesis.

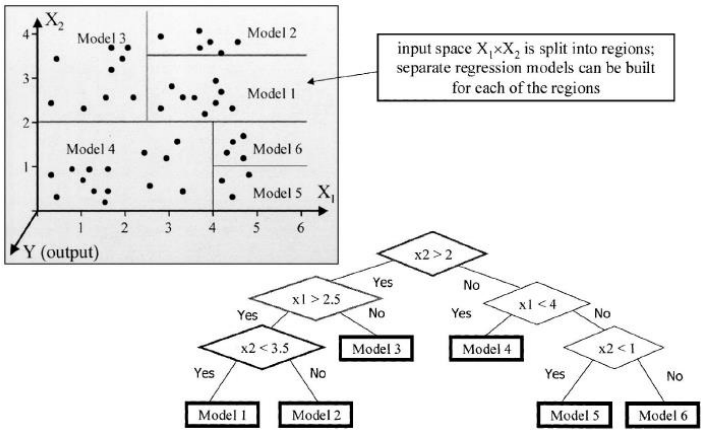


Figure 2-10: A generic M5 model tree, Models1-6 are linear regression models (reproduced from [73])

2.5.1.3 Multiple Linear Regression

MLR analysis determines the relationship of a single dependent variable given the set of independent input variables [159]. It postulates a linear function and then finds the parameter values that maximizes the fit to the training data [69]. Every value of the input variable x is associated with a value of the dependent variable y. the regression line for p input variables x_1, x_2,……., x_p is defined as [160, 161]:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p \quad (12)$$

Here, β_0 is a constant and β_i,……,β_p are coefficients relating input variables with the output variables. This line explains the variability of mean response μ_y with respect to the independent variables. The observed values for y vary about their means and are assumed to have the same standard deviation. The fitted values estimate the parameters β_0, β_i,……,β_p of the regression line.

2.5.2 Machine Learning for Classification Models

Classification is a basic task in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of attributes [162]. A large number of techniques have been developed for classification based on artificial intelligence and statistics [163, 164]. Artificial intelligence based methods include

decision tree algorithms such as C4.5, and perceptron-based algorithms such as neural networks. On the other hand, Statistical techniques encompass Bayesian networks and instance-based methods (k-nearest neighbour, k-NN). These algorithms are reviewed in detail in [164]. Selecting a specific classification algorithm is a critical step and depends on the requirements of an application and the particular dataset. Generally, the simplest approach is to use prediction accuracy to evaluate the various classifiers and select the one with the best performance. Nevertheless, it is stated in the literature [164] that neural networks tend to perform better with continuous features as compared to decision trees which perform better with discrete or categorical features. Moreover, Naïve Bayes require smaller dataset and takes less computational time for training as compared to neural networks. Furthermore, Naïve Bayes require little storage space during training and classification as compared to k-NN. Decision tree algorithms and Naïve Bayes are the easier to interpret, whereas neural networks and k-NN have poor interpretability. Based on the above analysis, decision trees (C4.5) and Naïve Bayes are selected for developing the decision support model in this thesis. These algorithms are discussed below in detail.

2.5.2.1 Naïve Bayes

Naive Bayesian is a statistical classifier, and one of the most effective one in terms of its high predictive performance in comparison to the state-of-the-art, C4.5 [162]. Naïve Bayes are very simple Bayesian networks which are composed of direct acyclic graphs with only one parent and several children. The classification process consists of two steps – training and prediction. In the training phase, using the training data, the classifier learns the conditional probability of each attribute A_i for a given class C_j. In the prediction step, classification is performed for any unseen test data by applying Bayes rule. The method computes the posterior probability of that sample (A_1, A_2 .....A_n) belonging to each class (C_1,C_2, ..., C_m) and then classifies the test data based on the class with the highest posterior probability [162, 165]. Using Bayes' law, the classification is expressed as;

$$p(C_j|A_1, \dots \dots . A_n) = \frac{p(C_j)p(A_1, \dots \dots . A_n|C_j)}{p(A_1, \dots \dots . A_n)} \quad (13)$$

The naive Bayes classifier is designed for use when predictors are independent of one another within each class, but it appears to work well in practice even when the independence assumption is not valid. The assumption of class independence allows the Naive Bayes classifier to better estimate the parameters required for accurate classification while using less training data than many other classifiers [166].

2.5.2.2 C4.5 (J48) – Decision Tree Algorithm

The C4.5 algorithm is based on decision trees. As mentioned earlier, decision tree modelling is a method of approximating a target variable (output), with discrete values, from a given data set. It represents the learned function in the form of a decision tree, where each leaf contains the target value. If the target value is a class, then it is termed as a classification tree. The C4.5 algorithm constructs the decision tree based on a 'divide and conquer' strategy [167]. A decision tree algorithm creates a tree model by using values of only one attribute at a time. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is based on the information gain (entropy) of the attribute. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then repeats this on the smaller sub-lists.

2.6 Summary

This chapter has established the state-of-the-art in the available monitoring and modelling themes of this research. Application principles, strengths, and limitations have been discussed, which form the basis for this thesis.

Reusing drainage water and nutrients emanating from one farm in another farm, before they enter the water system, can have huge environmental as well as economic benefits. Predicted information on volumes, timings, and quality of discharges that will be delivered to the farms becomes vital for adoption of this mechanism in agriculture. Despite tremendous promise of the benefits of drainage water reuse and technological advancements, implementation of an intelligent and autonomous management mechanism has not kept pace with deteriorating water situation. With contemporary WSNs, it is now possible to extract real time field data with spatio-temporal detail, hence it has been adopted widely in environmental monitoring in the last decade. Existing work in the area of real-time and remote monitoring of agriculture and hydro systems, using WSNs, is either aimed towards enabling high crop yields or to report on water quality status respectively. So far, no technical mechanism exists to facilitate management and reuse of reclaimed water in an autonomous and dynamic manner among farms.

For the prediction of hydrological discharges and nitrates, various modelling approaches have been adopted. With increasing monitoring and real data availability, empirical modelling based on machine learning algorithms is prevalent in hydrological and nitrate loss modelling. In

hydrological modelling, these projects are limited to flood forecasting and stream flow monitoring, in a gauged catchment, by using decades of historical data for modelling. As per the author's best knowledge, only the research work by *Kuzmanovski* [68] is similar in part to the objective to this research, i.e., modelling discharges from a farm. However, the model uses 20 years of historical data for field and climatic parameters. This highlights one of the main issues in that the historical data sets needed to develop predictive models do not exist for every farm, and for most catchments. For nitrate loss modelling, the models rely on enormous biological, chemical and geological data acquired using traditional grab sampling methods requiring expensive equipment. For both modelling approaches, it is apparent that existing modelling has not been intended for predicting losses within the farm system with the aim of enabling reutilization and alerts in real time, i.e. by using WSNs.

Data-driven modelling, using machine learning algorithms, has been widely used in hydrological modelling. Examples of the most popular supervised learning methods used for hydrological predictive systems are: statistical methods (e.g. MLR), artificial neural networks (ANN), and decision model trees (e.g. M5 tree). Recently, decision model trees have been investigated and adopted in hydrological modelling due to higher prediction accuracy and easier interpretation as compared to the other methods. Model trees are an extension of regression trees, and include first order linear models at the leaf nodes, compared to zero-order models in regression trees. Model trees are able to make predictions for values outside the training data range, which is not the case with regression trees.

Chapter 3: WQMCM Framework Architecture

In chapter 2, the concept and the importance of proactive and autonomous monitoring, through using the predicted prevalence of WSNs in agriculture, was presented for addressing water quality control through drainage reuse, and the need for low-complexity modelling was highlighted. This chapter presents the architecture of Water Quality Management using Collaborative Monitoring (WQMCM), which uses existing networked farms and water systems and low-complexity predictive models to enable real time drainage water management. Section 3.1 provides functional overview of the WQMCM framework with the design of a modified drainage network. Section 3.2 provides architectural detail on various model components; the neighbour linking model, predictive models, and the decision support model for drainage reuse, while section 3.3 discusses the development of neighbour linking model. Later chapters present the development of the remaining model components.

3.1 Functional Overview

WQMCM is an integrated control and management strategy which requires that individually targeted monitoring units or local networks, representing different stakeholders in a catchment e.g. a farm, should be able to share information with each other about runoff, drainage or nutrient fluxes. These events may be intense but are short-lived and so information sharing becomes important as they may be very fast, and so may normally be missed with the usual sampling rate. Allowing event information to be transmitted across multiple networks as they are detected will allow prediction of when the repercussions of that event might be seen downstream, allowing other stakeholder networks in the vicinity to adjust their monitoring and management strategy. This will include taking decisions about reusing or disposing the drainages, or increasing their sample rate to catch transient events. As emphasized in the literature, drainage reuse strategy reduces the overall stress of nutrient losses to the water system and provides economic benefit as well by reducing fertilizer usage. The proposed framework enables stakeholders to manage and benefit from this reutilization by sharing information about their availability and presence. Such a de-centralized approach comprising autonomous networks presents a flexible methodology where independent networks, in addition to local monitoring objectives, seek to opportunistically utilize neighbouring events.

To demonstrate the mechanism of the proposed framework with respect to agricultural drainage reuse, a modified drainage network is designed. Figure 3-1 illustrates an example irrigation and drainage system in which various farms and drainage regions are linked with each other through water flow paths. The figure shows an additional bay, drainage reuse bay, linked with individual

farm's irrigation and drainage bay to implement the drainage reuse mechanism. Each farm would have the option to either use drainage from another farm or fresh irrigation water for irrigation. As mentioned earlier, the WQMCM framework aims to combine local individual networks into an integrated mechanism, therefore, it is assumed that these farms are monitored by individual networks with local application objectives. These objectives are to facilitate farming decisions with regards to e.g. irrigation or pesticides scheduling by monitoring microclimate (soil moisture, crop cover, and soil temperature) of the field. For implementing the framework, an additional network on the water system, the drainage reuse bay in this case, is required to monitor drainage and nutrient contributions by each farm. As shown in Figure 3-1, individual sensors in the drainage network are deployed at the outlet of each farm to monitor its drainage outflow. Other nodes monitor base flow in the drainage bay. This network will be either deployed by an official governing body working towards maintaining water quality or by local farmers for a collaborative cause.

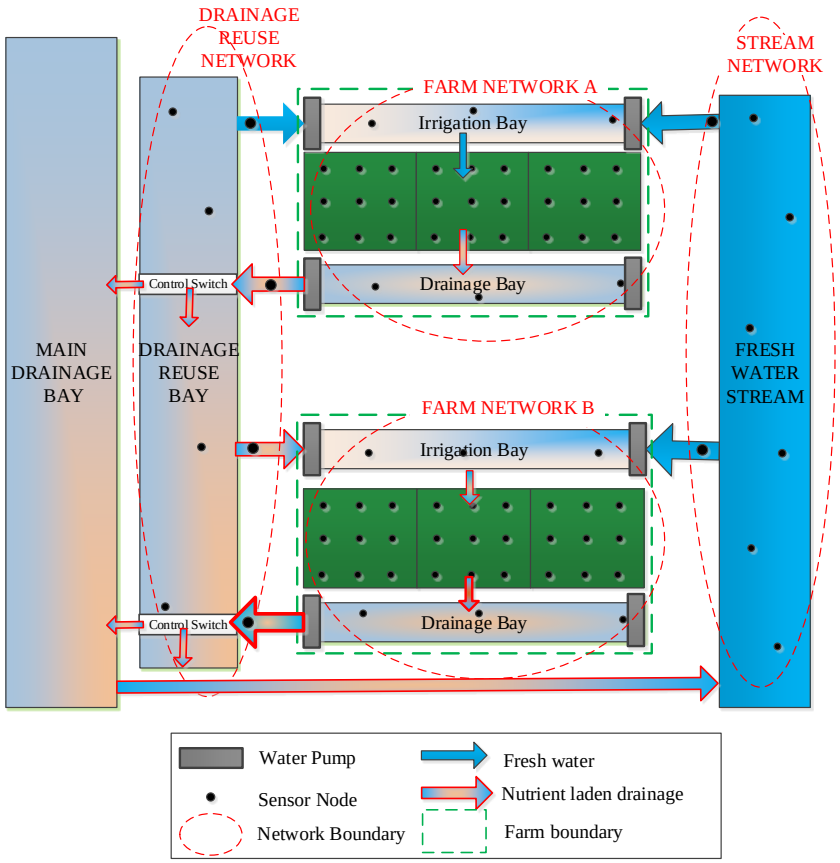


Figure 3-1: A modified drainage network design to implement drainage reuse for the WQMCM framework

These networks, under the proposed WQMCM framework, share information about the start of a daily event with each other, e.g. an irrigation event in a farm or high pollutant drainage discharge from drainage bay. When event information is received from a farm network (e.g. farm A), the drainage network node associated with that farm uses on-node predictive models to forecast the values for expected drainage and nutrient dynamics as a result of that event. The forecasting of drainage dynamics is undertaken by the drainage network for the following reasons. Firstly, because drainage network links all the farms and the stream networks, hence it's ideal to have the drainage network disseminate the predicted information about drainage to all the other farms and stream network for reuse, treatment or disposal. Secondly, the main drainage bay could be distantly located from the drainage ditch of a farm, hence volumes of actual drainage outflows received by a drainage bay from a farm may change owing to evapotranspiration and absorption during its transport. Additionally, running predictive models is a computational overhead which should naturally be taken up by the network responsible for decision making.

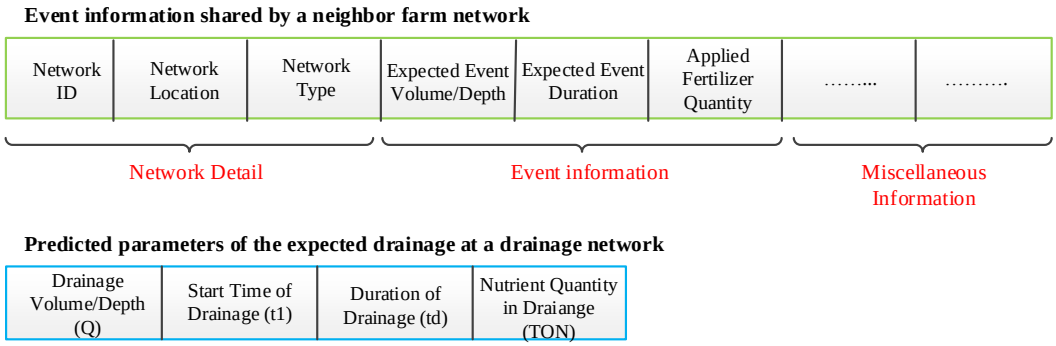


Figure 3-2: Parameters related to upcoming event shared by neighbour networks and the predicted variables for the resulting drainage event

Figure 3-2 illustrates the format of information shared by a farm and the parameters predicted by a drainage network. The shared event information packet from a farm includes network and event details. To identify a network, information such as network id, type and location is included. Network type is related to whether it is a farm, drainage or a stream network, which helps filter out received messages. For instance, a farm network may only want to receive information from drainage or stream network, or a drainage network may only be interested in information coming from farms for obvious reasons. Network location filters out geographically dislocated networks or the ones located down-stream, which are unlikely to impact upstream networks. Further to that, event detail in the information packet includes event depth/volume, event duration, fertilizer quantity applied. Any additional event information will be governed by the requirements of a

predictive model, which is discussed in the next chapter. As far as the predicted parameters for expected drainage are concerned, as discussed in chapter 1, the relevant information necessary to implement a proactive monitoring and management system is drainage depth/volume (Q), fertilizer loads in the drainage (TON), start time (t_1) and duration (t_d) of the drainage.

Predicted values of drainage and nutrient dynamics by the sensor node are transmitted to the gateway of drainage network, from where it is relayed to the neighbouring farm and stream networks. The farm networks (e.g. farm B) uses the predicted information and local decision support model to decide whether to reuse the drainage or not, and transmit a reuse acknowledgement to the drainage network. In the former case in which network B intends to reuse the drainage, the drainage water, once available from farm A, is allowed to drain into the drainage reuse bay (through a control pump) instead of the main drainage bay. From the reuse bay, the drainage is then pumped into the irrigation bay of farm B. In case none of the networks send reuse acknowledgements, the drainage would be drained into the main drainage bay. The stream network can then decide, based on the predicted values for nutrients, whether to divert the flow in case of high nutrient outflows or to otherwise allow it to enter the stream.

3.2 Architectural Detail

The fundamental part of the WQMCM framework is that individual networks learn their environment to predict the impact of events elsewhere in the catchment on their own zone of influence. The predicted drainage information can be beneficial for adjusting management strategy in a farm or in a stream network accordingly, by adopting drainage reuse, disposal or treatment. For managing agricultural reuse, the overall architectural detail comprises various modules encompassing drainage, farm and stream networks, as illustrated in Figure 3-3. For enabling forecasting of drainage dynamics expected as a result of an event in a neighbouring farm, two key modules are developed in the drainage network; neighbour linking model and, drainage and pollutant dynamics module. The neighbour linking model uses neighbour event information and sensed drainage data to link the impactful neighbours. The predictive module further comprises individual models to predict Q, t_1, t_d, and TON, which forms the primary contribution of this thesis. The predicted drainage information is used by decision support models in farm and stream network to enable decision making about its reuse or disposal. Furthermore, this information is further used to adjust sampling rate of the sensors, to capture the approaching drainage flow, at the predicted response time.

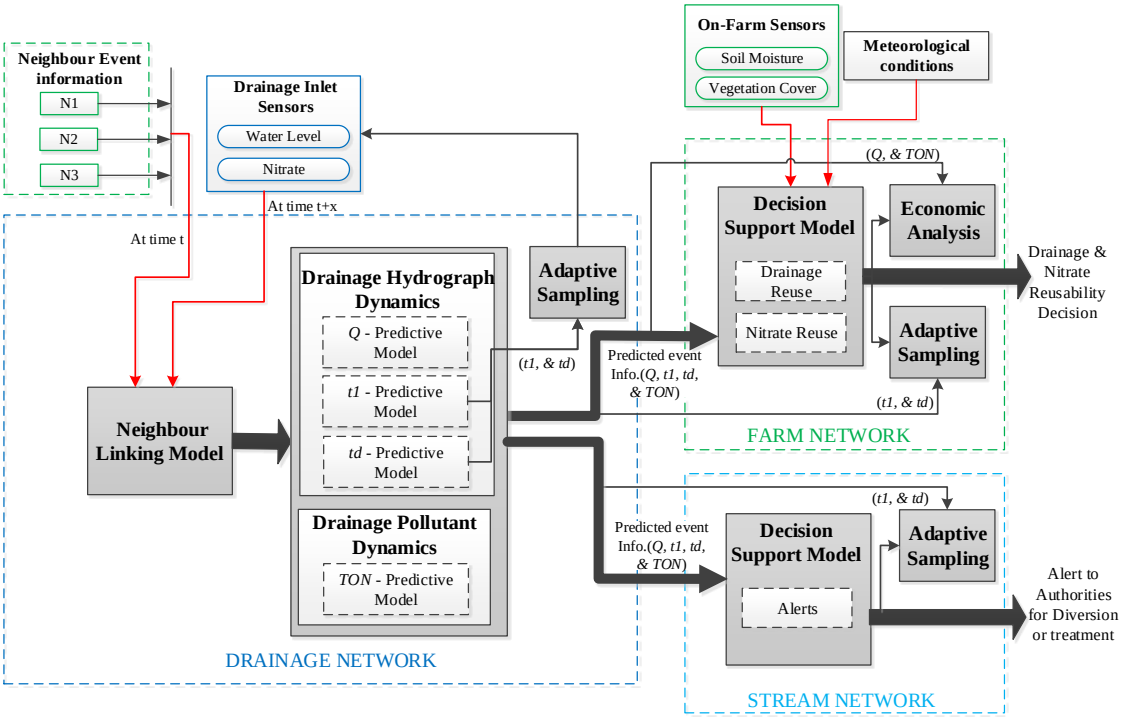


Figure 3-3: Block diagram of the WQMCM framework architecture

It has been emphasized in this thesis that due to inevitable drainage and nutrient losses despite adopting BMPs, it is important to enable mechanism for their reutilization. Therefore, a simplified decision support model is developed just as an example to illustrate the utilization of predicted information for enabling reuse mechanism. Below the modules of drainage network and farm network blocks are briefly introduced. Figure 3-4 and Figure 3-6 illustrates the functional flow of these modules for both the blocks.

3.2.1 Neighbour Linking Model

The main purpose of this module in the drainage network is to identify the farm networks which drain into this drainage network. These links are identified over a period of time using a learning process. Firstly, dislocated networks (e.g. located at lower altitudes of the catchment) are filtered out using network location in the shared information packet by the neighbours. Secondly, for the filtered neighbours, training dataset is acquired over a period of time, as respective event information is received from these individual neighbours. Training data consists of; (i) the event information packets that were received that particular day, e.g. at time *t*, and (ii) the sensed values of the received drainage and nutrients at the drainage bay at time *t+x*. Here, *x*, refers to the time

it takes for drainage to appear in the drainage bay from the time event information is received, and *t* is ranged between 00 hrs to (*x-1*) hrs.

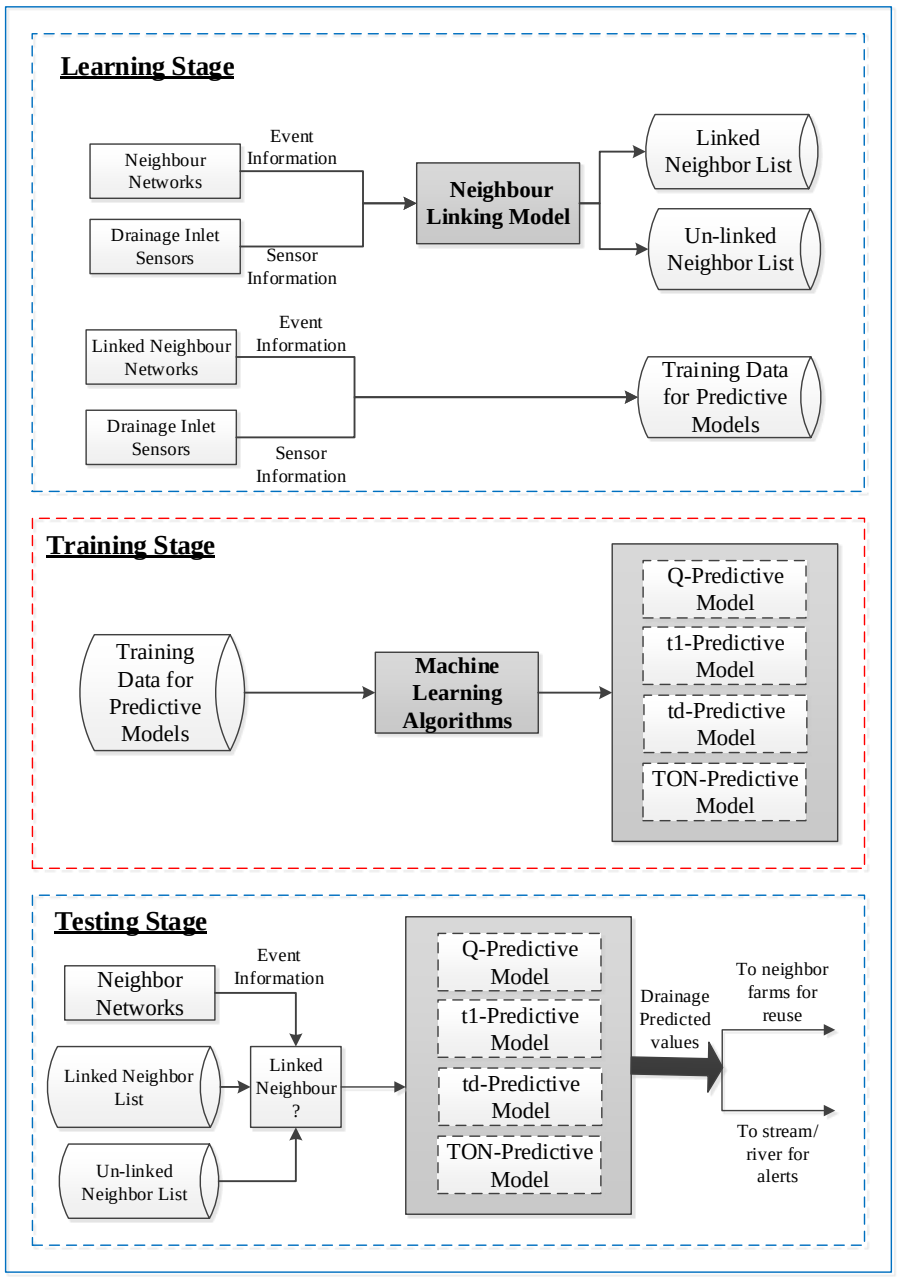


Figure 3-4: Functional stages of the drainage network modules, under the WQMCM framework, such as 'Learning', 'Training', and 'Testing'

For each set of the acquired training data for a particular neighbour, a linear regression model is used to identify the relationship between the sensed drainage and the received event details. This process divides the neighbours, which sent information over the learning phase, into two lists; linked neighbours and un-linked neighbours. Later, for the linked neighbours, sufficient training

data are acquired to provide for the development of the predictive models. Figure 3-4 illustrates the mechanism of the neighbour linking model in the learning stage section. The development of neighbour linking model will be discussed in detail in the later section 3.3.

3.2.2 Drainage Hydrograph and Nutrient Dynamics Predictive Models

Once training dataset is acquired for the linked neighbours, the next step is to develop the models for predicting the drainage hydrograph and nutrient dynamics. Constraints on network nodes (battery life, computing power, availability of sensors etc.) 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. Ideally, the model should be based on minimal training samples so that the model can be implementable soon after the deployment of the network. Such models are local in the sense of being valid for a given site (farm in this case). Once developed at the gateway, these models are deployed on the relevant node associated with the particular farm. This facilitates distributed computing where individual nodes of the drainage network, deployed at the outlets of farms, run the learned predictive models for forecasting drainage from those farms. These models can then generate expected drainage hydrograph and nutrients dynamics, which are transmitted to the gateway for further action regarding transmission to neighbouring networks.

These models intrinsically self-calibrate because the evolving record of the observations allow them to adapt to the latest condition. This creates portability from one season to the next, and from one climate regime to the next. With new data regarding a farm, the models are calibrated at the gateway and re-deployed at the relevant node. However, it is important that a model must maintain a balance between the complexity of the model and the predictive accuracy of the model.

Based on the discussion in the literature review of sections 2.3 and 2.4, existing state-of-the-art predictive models are used as a basis to derive low-complexity models for Q, t_1, t_d and TON. A machine learning algorithm, M5 tree, is then used to train the individual models as shown in the training stage of Figure 3-4. The development and validation of these four predictive models is the primary contribution of this thesis and will be discussed later in the thesis in chapters 4, 5, and 6. Once the models are trained with acceptable prediction performance, the drainage network progresses to the testing stage. In the testing stage, neighbour event information is firstly interpreted using developed neighbour linking lists and then used to predict drainage dynamics using the predictive models, in case of a linked neighbour as illustrated in the testing stage of

Figure 3-4. As mentioned earlier, the model accuracy can be continuously improved by learning the evolving instances in the testing stage.

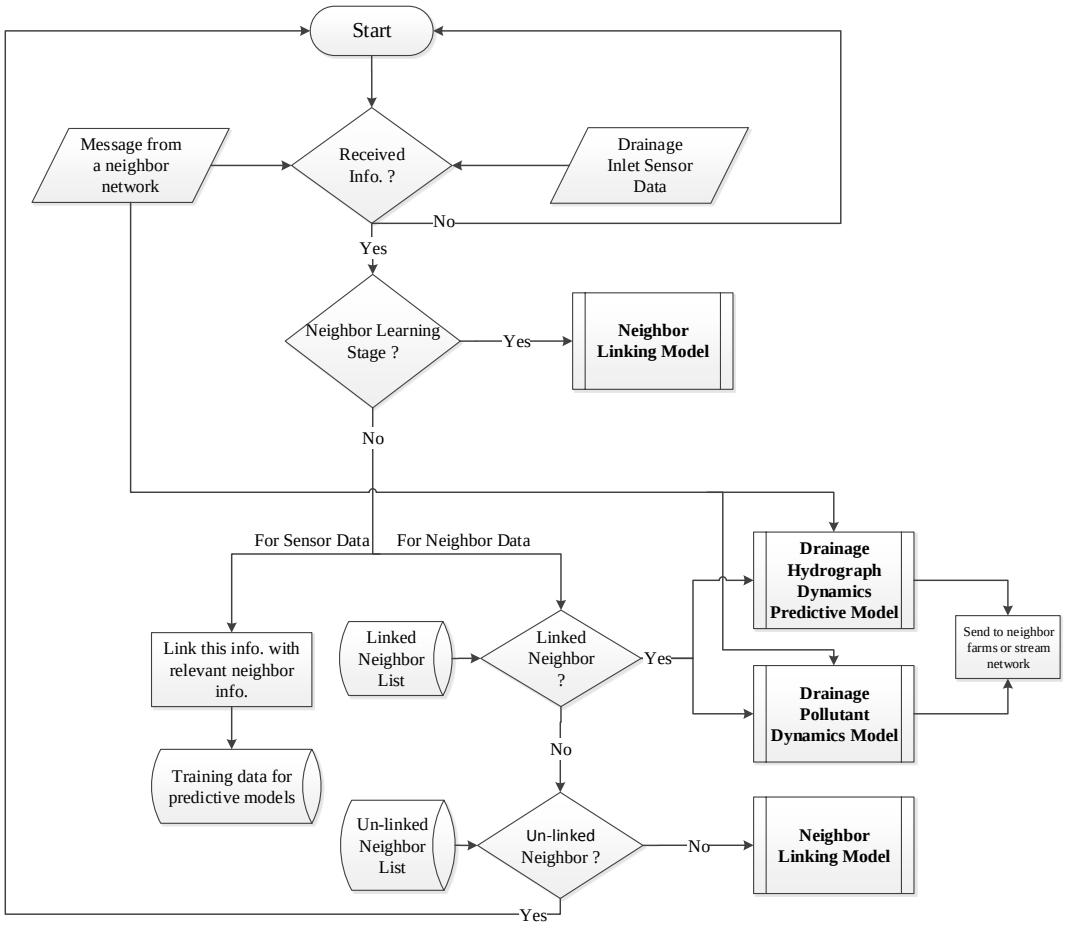


Figure 3-5: Basic algorithm running at a sensor node of a drainage network under WQMCM framework

The algorithmic flow of these stages for a drainage network is illustrated using a flow diagram in Figure 3-5. When information is received at the data sink of the drainage network by either a drainage network sensor or a neighbour farm, firstly it is checked whether the network is in the learning stage or not. In such a case, the information is passed on to the neighbour linking model. If the model is in testing stage, then in case the received data packet is from a neighbour, it is checked if the neighbour ID is in the linked neighbour list. If it is an already linked neighbour, then the relevant trained predictive models for that particular neighbour are used to predict the event values. Otherwise, it is determined if the neighbour is an un-linked neighbour, in which case the event packet information is disregarded. In case the received data packet is from the drainage sensor, then the data within the packet are linked with the relevant neighbour information and saved for improving the models later.

3.2.3 Decision Support Model (for drainage reuse in a farm)

For the decision support model, the challenge lies in designing a model which takes into account local field conditions, predicted dynamics, and expert knowledge. Unlike the predictive models in the drainage network, this model essentially runs on the gateway of the farm network. The model complexity can substantially vary depending upon the requirements set by the farmer. For example, the farmer may want the model to advice on the possible repercussions of drainage reuse on crop. Furthermore, in case the available drainage is not enough or high in N, the model may also advise on mixing drainage and fresh water for irrigation to fulfil its requirements or to disregard the excess nutrients in the drainage which the farm may not want to reuse. These complexities are highly scenario dependent and require sufficient expert knowledge and data to address. In this thesis, a simplified decision support model is developed as an example to demonstrate the utilization of predicted information for enabling the reuse mechanism. It is not a primary contribution of this thesis, but adds useful context.

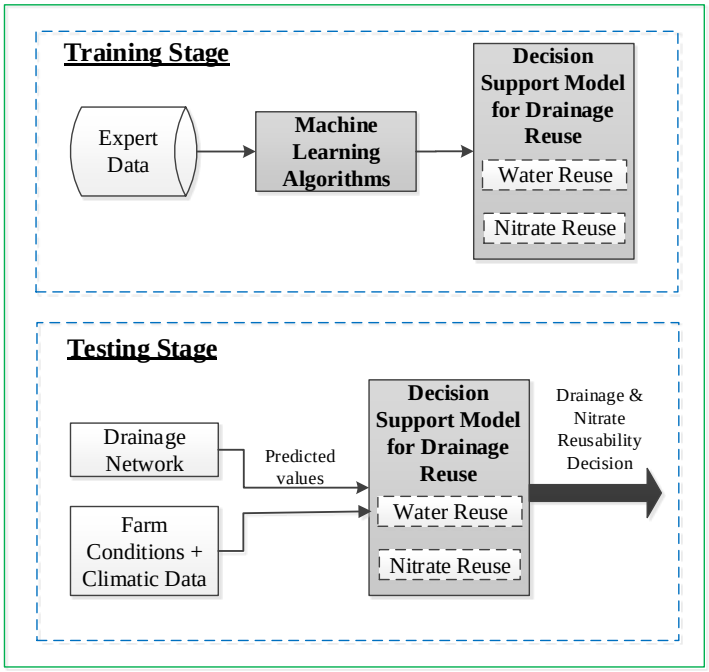


Figure 3-6: Functional stages of a decision support model at a farm network, which intends to reuse drainage water, under the WQMCM framework

Figure 3-6 illustrates the functional stages of the development and use of the decision model for drainage reuse in a farm. In the training stage, expert knowledge and machine learning algorithms are used to implement simplified models for drainage and nutrient reuse. Once the model is

trained, predicted drainage information received from a drainage network, and local field conditions and climatic data are used to classify the usability of the drainage water and nutrients (as shown in the testing stage of Figure 3-6). The development of the decision model is discussed in chapter 7.

3.3 Development of Neighbour Linking Model

To demonstrate the development of this model, consider an example layout of a few farms and their respective drainage bays. As illustrated in Figure 3-7, there are two separate drainage reuse bays represented with networks, D1 and D2. Similarly the farms are monitored using networks F1, F2, F3, F4, and F5. It is to be noted that Figure 3-7 is a simplified layout illustrating only the relevant units of the drainage network, as compared to the one presented in Figure 3-1. As already introduced in the earlier section 3.2.1, the neighbour linking model consists of two stages; pre-linking stage with geographical filtering, and the linking stage with linear regression modelling.

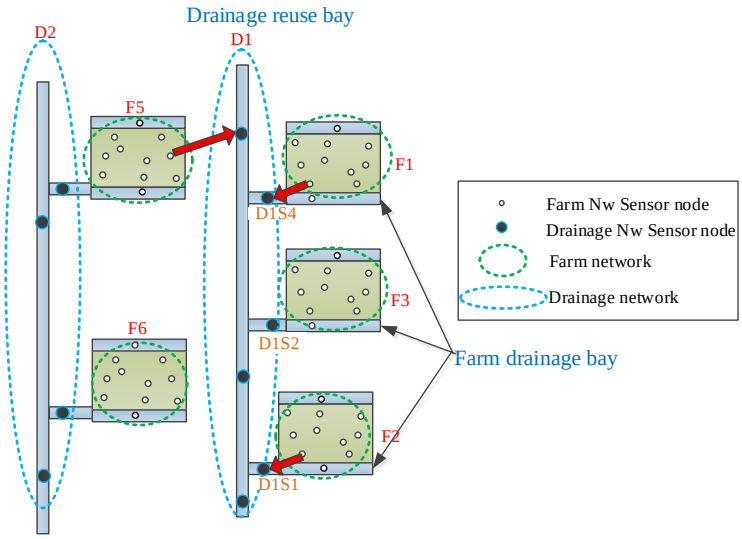


Figure 3-7: Example scenario to demonstrate neighbour linking modelling

The network topology for the drainage network is designed in such a way to identify and measure individual events in case drainage outflows emanate from more than one farm at the same time. Therefore, some sensors are placed right before where the drainage water joins the drainage reuse bay, such as D1S4, D1S2, and D1S1 as shown in Figure 3-7. Hence, these sensors are able to capture the actual contribution of each farm in the drainage and nutrient loads.

This incurs a neighbour linking design in which each neighbour farm is not just linked with the drainage network it drain into, but to the specific sensor node of the drainage network. For

instance, F1 will be linked with D1S4, F3 with D1S2, and F2 with D1S1. In the testing stage of the framework, when event information is received from farm F1, the D1S1 sensor selects and run the relevant predictive models for F1. The predicted values are transmitted back to the data sink of the drainage network, which then shares this information with other neighbour farms. Data aggregation and transmission in such a scenario is another detailed topic, but is not the focus of this thesis. Furthermore, D1S1 activates adaptive sampling with respect to the predicted timing of the expected drainage. This would prolong network life by not having the network sense constantly and only activating the relevant sensor for sensing instead of the whole network.

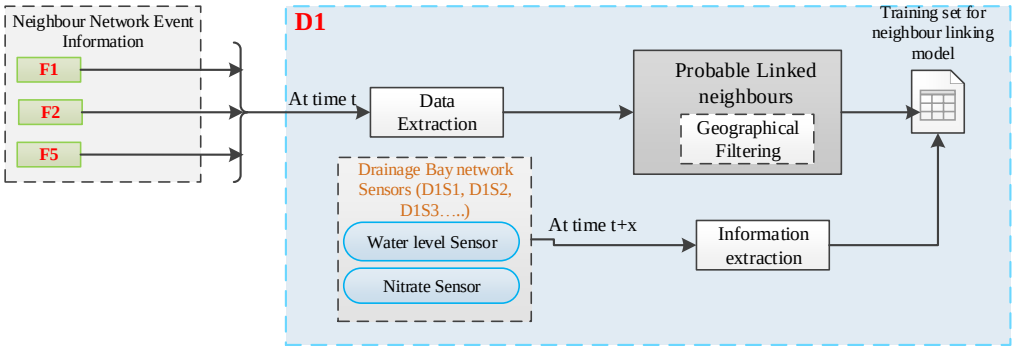


Figure 3-8: Stage 1 - Pre-linking stage of neighbour linking modelling

Moving on to the development of the neighbour linking model, firstly the pre-linking stage is presented. In this stage, geographical filtering based on the network location of the farm networks is used to filter out probable linked neighbours. Figure 3-8 illustrates this stage. Over a period of time, training samples are accumulated for the probable linked neighbours. For example, as illustrated in Figure 3-7, on a particular day, event information is received by D1 from F1, F2, and F5. After geographical filtering, it is concluded that none of the networks are dislocated. Hence the drainage network activates all its sensors to constantly sample and awaits data. After a while, sensor D1S4, and D1S1 starts to respond with data. Once the entire event is sensed by both the sensors, training instances are created for drainage data sensed by D1S4 and D1S2 against the event information received from the neighbours F1, F2, and F5. Over a period of time, when enough samples of data are accumulated for each of the sensors of the drainage bay, linear regression is used, in stage 2, to identify the link with a particular farm, as illustrated in Figure 3-9. Whether the number of samples are enough to identify the links depends on the outcome of the regression model. If a clear relationship is identified between a particular sensor and a farm, the learning can be stopped, and linked neighbour list be updated.

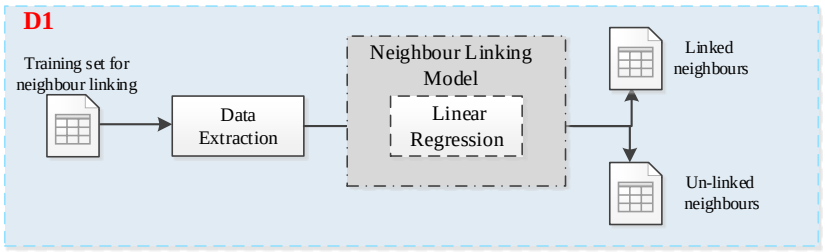


Figure 3-9: Stage 2 - Learning in the neighbour linking model

To demonstrate the regression process, simulated drainage dataset is generated (e.g. for sensor D1S1), from a simulator based on NRCS curve number method [168], by creating various event details (event intensity, crop cover, soil moisture conditions) for three farms (e.g. F1, F2, and F5). It is important to mention that even in a highly unlike scenario in which various farms have similar event intensities and farm area, drainage volumes can be different based on different vegetation cover and soil moisture conditions. However, there can be instances when different farms generate similar drainage, hence the need for sufficient training samples to overlook such discrepancy. Both of the above situations were taken into account for generating the training samples.

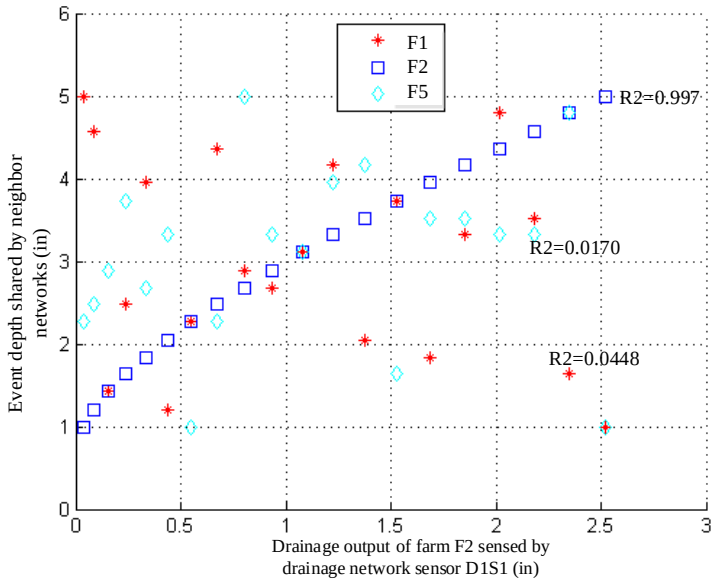


Figure 3-10: Plot of regression modelling applied on drainage outcome sensed by D1S1 and the event depths shared by neighbour farms F1, F2, and F5 over a period of time

Figure 3-10 illustrates the outcome of regression modelling applied on the simulated data set containing event depths for the three farms and drainage depths received by the drainage network. The plots demonstrate that in this case there is a clear linear relationship between event depths shared by the neighbour farm F2 and the drainage depths measured by the drainage network

sensor D1S1 during the same time window. This is demonstrated by a high value of 0.99 for the correlation of determination, representing very high correlation between the events and generated drainage from farm F2. Whereas, very weak correlation of 0.017 and 0.044 is represented between the drainage sensed by D1S1 and event depths shared by farms F1 and F5 respectively.

3.4 Summary

This chapter presented the WQMCM framework for a collaborative control and management of agricultural drainage water for addressing the issue of prevalent water crisis. The framework leverages individual networked farms and streams into an integrated water management mechanism. Such a monitoring system should enable each farm to share information about its drainage flow with neighbour networks, e.g. with a drainage bay network, which can then process the information for timely treatment, disposal or reuse of the drainage.

To implement the drainage management, the architecture of the WQMCM framework comprises various modules. Modules for a drainage bay network include neighbour linking model, and predictive models for drainage and pollutant dynamics. Whereas, for a farm network, a decision support model is used to ascertain the reusability of the predicted drainage event. The overall functionality of the framework is explored in terms of stages of learning, training and testing. In the learning stage, neighbour linking model is used to determine the correlation of events in various farm networks with the events received in the drainage bay by the drainage network. The model results in identifying linked and un-linked farm networks by using a combination of geographical filtering and linear regression methods. For the linked networks, training dataset is acquired to provide for the development of the predictive models for drainage dynamics and nitrate losses. When the drainage network has learned the environment and the predictive models for individual farms, it is brought into a testing stage. In this stage, neighbour event information is firstly interpreted using developed neighbour linking lists and then, in case of a linked neighbour, used to predict drainage dynamics. These predicted values are transmitted by a drainage network to other farms and stream networks so that they can take a decision for the reuse, disposal or conservation of the expected drainage.

Chapter 4: Modelling and Simulation of Low-Complexity Discharge-Dynamics (Q, t_1, and t_d) Predictive Modules

Chapter 3 presented the architecture of WQMCM framework, and the development of the neighbour linking module. This chapter investigates the modelling and simulation of low-complexity Q-Predictive, t_1-predictive and t_d-predictive models. Section 4.1 presents the development of models for Q, t_1, and t_d by simplifying the model complexity in terms of influencing parameters. Section 4.2 outlines the experimental setup for the simulation of the models with respect to the simulated data that will be used, and various criteria that will be adopted for model evaluation. Using the proposed low-complexity model parameters and simulated data, section 4.3 investigates the training and testing for Q, t_1, and t_d predictive models.

4.1 Model Development of a Low-Complexity Discharge-Dynamics Predictive Module

As established in chapter 2 (section 2.3 and 2.4), that resource constraints associated with network nodes (e.g. power consumption, computing power etc.), and unavailability of in-situ, inexpensive, and low-power bio-geo-chemical sensors, require a simplified predictive model for in-network deployment. Detailed review of various hydrological methods both for runoff and drainage suggests that the NRCS curve number model is the simplest approach to predict discharges from a land, hence has been popularly adopted for decades. To address the limitations pointed out, in chapter 2 (section 2.3.1.2), with regards to the use of proxy parameters in the NRCS method, it is proposed that modifications in the model parameters for Q, t_1, and t_d are sought.

For the prediction of discharges (Q), as expressed in Eq. (7) in chapter 2, the NRCS model uses rainfall (P), and curve number (CN) as the input parameters. A constant value for I_a (initial abstraction) is assumed as was shown in Eq. (4) of chapter 2. Recalling from chapter 2, Eq. (7) was expressed as;

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

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, AMC). Although, the use of last 5-day rainfall to represent soil moisture conditions has been questioned [169], at the time the NRCS model was developed, proxy parameters and manual measurements were used to represent land conditions due to the absence of direct sensing measures.

In order to validate the limitation of last-5-day-rainfall index to represent AMC, a season long data observed in a precision irrigation application is analysed. The data was supplied by the University of Georgia (*Vellidis et al.* [43]). Figure 4-1 plots a week-long dataset of measured soil tension (which represents soil moisture). 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. This is contrary to the traditional last-5-day-rainfall indexing method, according to which AMC is interpreted as dry in case 5-day antecedent rainfall or irrigation depth is less than 36mm (chapter 2,Table 2-2). For example, on the 22nd of July in the Figure 4-1, dry soil conditions would be estimated as per the last-5-day-rainfall index, due to the fact that there was no rain in the preceding 5 days. However, the actual soil condition is measured as adequately wet by the sensors (soil tension = 24 cbars). In the absence of moisture sensors, when the last-5-day-rainfall index is used to determine the drainage/runoff for a rainfall event, a lower value is calculated which would obviously be based on the wrong assumption of prior dry land conditions. This assumption consequently leads to inaccurate decisions about irrigation, resulting in inevitable drainage losses. Therefore, rainfall depth along with soil moisture and crop stage are equally important.

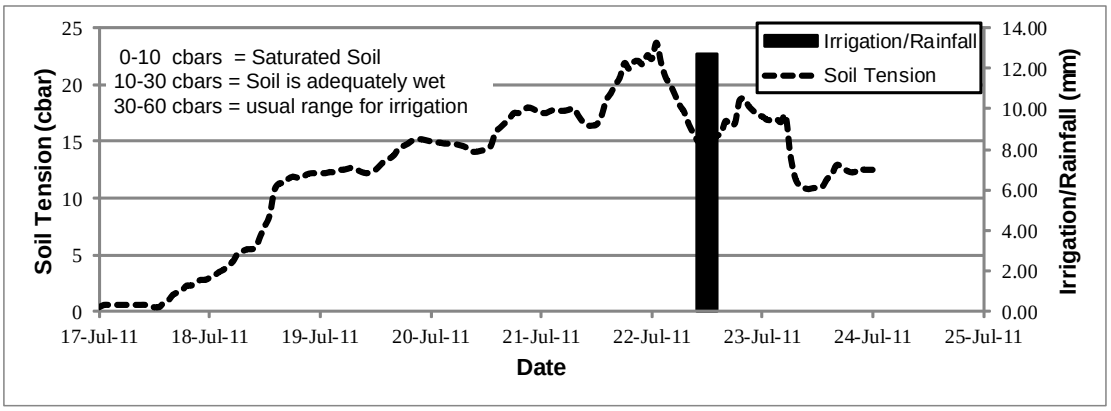


Figure 4-1: Measured soil water tension values in response to irrigation events in a networked field

The proposed simplification for the Q-predictive model is based on two steps (as shown in Figure 4-2). 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 expected rainfall value for that day 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 [169]. This can be used to generate a crop stage (CS) parameter, without manual observation, 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 [43]. Therefore, it has been proposed to use actual soil moisture values instead of the last-5-day-rainfall index.

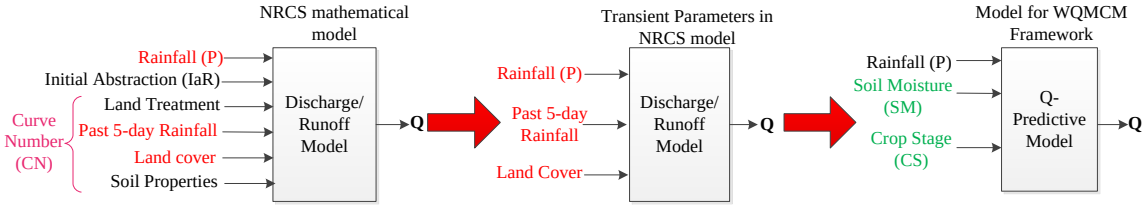


Figure 4-2: Model simplification for a Q-predictive model

As already discussed in chapter 2, for t_1 and t_d, the mathematical model and convolution method requires various parameters and historical data. Recalling from Eq.(11) in chapter 2, t_d is expressed as;

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

To simplify the complexity and dependence of this model, a low-complexity model is abstracted from the available model detail as shown in Figure 4-3. 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 analysing an extensive set of simulated data (using an NRCS based simulator [168]) for which a routine in Matlab is written by the author to extract t_1 and t_d. The acquired dataset indicates 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. 'n' in the NRCS is based on the kind and density of land cover for various land types including cultivated land, smooth surface, grass land and woods. For a cultivated land, the density of crop cover is used to determine the 'n' coefficient. Therefore, crop stage is used to represent the field roughness in the model parameters, as cultivated farms are under consideration in this

thesis. 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.

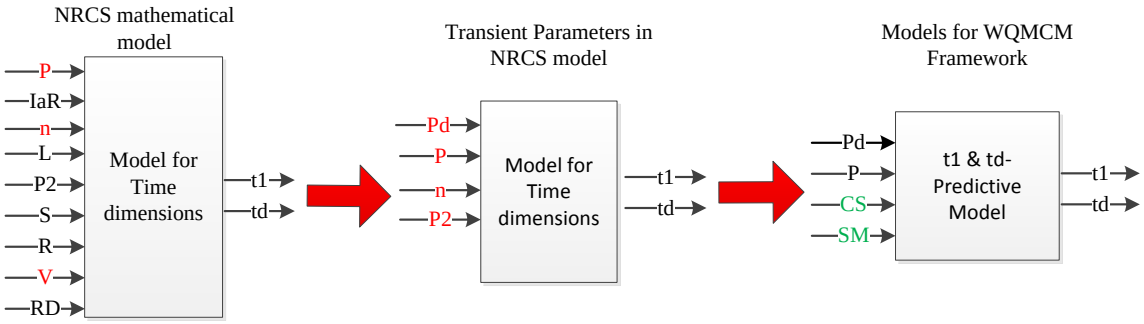


Figure 4-3: Model simplification for a t_1 and t_d predictive models

4.2 Experimental Method

4.2.1 Simulated Data

For training and testing the model, historical dataset is generated using a simulator based on the NRCS method [168], which was developed in Matlab. This is because at the time of this experiment, despite many efforts, sufficient measured field data could not be found for training the model.

A combination of various values for P, CN and P_d is considered to generate two sets of data – one for Q-predictive model and the other for t_1 and t_d predictive model. For values of CN, field condition values ranged from fallow land to more than 20% coverage, whereas AMC values ranged from dry, medium to wet land. Other parameters for determining CN include land treatment and hydrological group as discussed earlier, for which straight row practice and group C (moderate runoff) is considered. These parameters, although normally remain constant for a particular farm, and will not be considered for the learning model, are required by the simulator to generate the discharge values. The obtained data set is then modified to substitute CN with the proposed simplified model parameters of CS and SM. Values for CS and SM are assigned according to the field conditions that were considered for CN. For P_d, three event durations, 6 hours, 12 hours and 24 hours were considered as per general practice for national weather forecasts [170]. These input dataset values resulted in 360 and 512 training samples for developing the Q-predictive, and t_1 and t_d predictive models respectively.

4.2.2 Modelling Technique

As established in chapter 2, that decision model trees have higher prediction accuracy, easier model interpretation and ability to predict outside the training range in comparison to other adopted method in hydrological modelling. An example of decision model trees is M5 model tree algorithm which has been adopted in hydrology for modelling drainage discharge, rainfall runoff, streamflow, and flood forecasting. Therefore, the M5 tree algorithm is used for generating the Q, t_1 and t_d predictive models in this chapter. For comparative analysis, MLR algorithm is also used to relate its prediction performance with the M5 tree model. For developing these models, M5 tree toolbox [158] and MLR algorithm from Matlab are employed.

4.2.3 Model Evaluation Criteria

The quantitative assessment of the generated models is performed using a three-step procedure (as suggested in [58]) consisting of (1) random sampling and cross-validation, (2) performance measuring parameters, and (3) comparative assessment. Comparative assessment is undertaken with respect to other learning algorithms. It is important to mention that similar criteria will be used in the later chapters 5 and 6 as well. Below we discuss the evaluation criteria.

4.2.3.1 Random Sampling & Cross-Validation

To ensure a robust evaluation of the model performance in this chapter, the data set was randomly partitioned into two groups: training set and testing set. 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 [171]. Data are 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 research, where dataset is partitioned into 10 subsets.

4.2.3.2 Performance Measuring Parameters

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

- Root Mean Square Error (RMSE)

RMSE represents the error associated with the model, by giving an average measure of the error in predicting the dependent variable. The difference between predicted and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means that the RMSE is most useful when large errors are particularly undesirable. This is a valuable index as it indicates error in the same units of the constituents of interest, which aids in analysis of the results.

- Mean Absolute Error (MAE)

The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. The MAE is the mean over the verification sample of the absolute values of the differences between forecast and the corresponding observation. Unlike RMSE, all the individual differences are weighted equally in the average. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. Both the MAE and RMSE can range from 0 to ∞.

- Coefficient of determination (R^2)

R^2 represents the variability that can be explained by the model in terms of goodness of fit. R^2 ranges between 1 and 0. It is equal to 1 if the predictions are perfect, i.e. a linear relationship exist between the predicted and measured values represented by a straight line. It thus provides a way to quantify the accuracy of the model to predict the dependent variable.

- Root Relative Mean Square Error (RRMSE)

The 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. The value is normally multiplied by 100 to show a percentage of relative error.

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, these performance measures are compared with those achieved by previous work in this area.

4.2.3.3 Comparative Assessment

In order to compare the prediction performance of the Q, t_1 and t_d predictive models developed using M5 tree algorithm with another modelling technique, MLR is used here. The predicted results are plotted for each of the Q, t_1 and t_d models developed using the M5 tree algorithm and the MLR model.

4.3 Model Training & Testing for Q, t_1 and t_d Predictive Modules

4.3.1 Results and Discussion for Q-Predictive Model

For training the model, initially 100 training instances, are sampled from the simulated dataset generated earlier based on the proposed parameters (P, CS, and SM) and the output parameter (Q). M5 tree algorithm is used on the training set along with 10-fold cross-validation settings to generate the Q-predictive model.

The prediction performance for the generated Q-predictive model with 10-fold cross-validation results in RMSE value of 0.23 and RRMSE of 10.2%. To interpret the result of RMSE, a standard criteria is considered. An accepted adequate value for RMSE in hydrological modelling is normally equivalent to or less than half of the standard deviation (SD) of the training data [172]. For the conducted experiment, the SD for Q is equivalent to 2.6, so the acceptable value of RMSE for this model is equivalent to or less than 1.3. Hence, it is stated that the performance of the developed Q-predictive model is acceptable as its RMSE value, at 0.23, is even better than the acceptable value. To understand the value of RRMSE value, a good acceptable value which depends highly on an application, the normal standard with anything less than 30% as good (meaning 70% accuracy) is used. Hence, RRMSE value of 10.2% for the generated model certainly can be interpreted as an acceptable result.

Now, test data are drawn from the remaining training data to predict the values of Q for unseen input data. The predicted Q values are plotted against the measured values of the NRCS method as shown in Figure 4-4. The plot of predicted output for Q-predictive model shows a very good fit (R^2=0.98) as compared to the plot of measured output using the NRCS model. These results are very promising for validating the low-complexity model parameters that were proposed in this thesis. The proposed model will further be validated using measured field data in the next chapter.

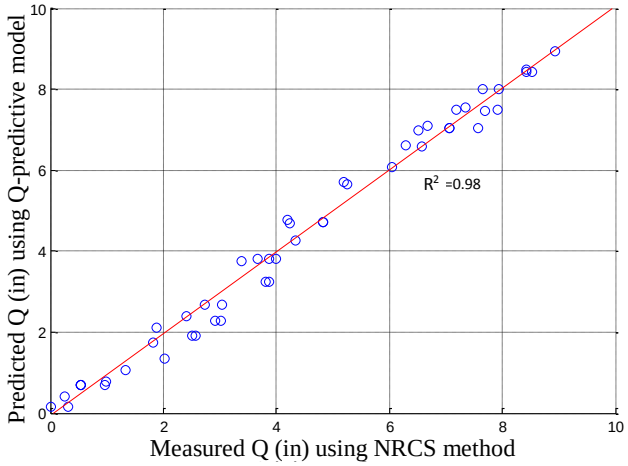


Figure 4-4: Plot of test data for Q-predictive model, developed using M5 tree algorithm, against measured values from NRCS method

The proposed Q-predictive model is further investigated for three things; (1) the impact of further simplification of model parameters on the performance, (2) performance comparison with the model developed using MLR algorithm, and (3) the impact of varying training set size on model performance.

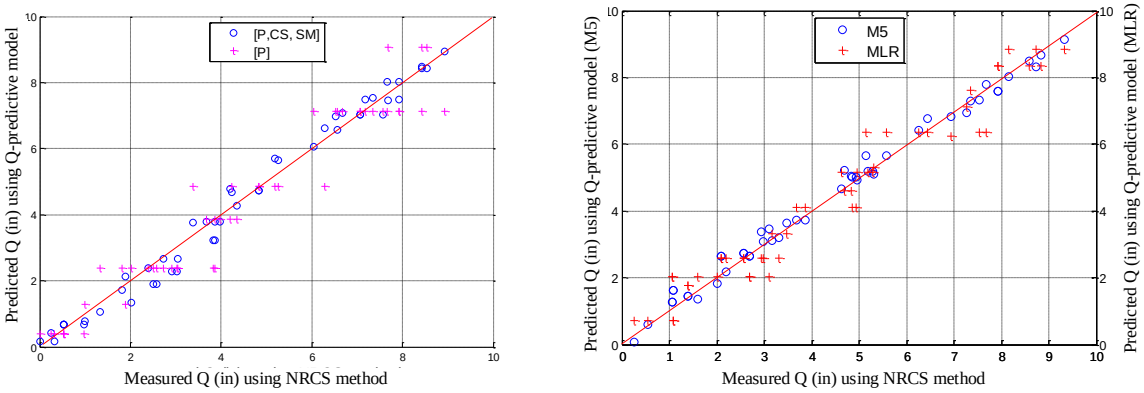


Figure 4-5: (a) Plot of test data for a modified and original Q-predictive model, (b) Plot of test data for Q-predictive models developed using M5 and MLR algorithm for the proposed parameters.

The first investigation is aimed at evaluating if the proposed model parameters are an optimum set or further simplification can also provide comparable results. To explore this, a model is generated using training set consisting of only P instances. Test dataset is used to plot the results of this model against the results of the original Q-predictive model as shown in Figure 4-5 (a). The figure illustrates that the plot for the modified model, developed using only P as input

parameter, shows very poor fit. The modified model resulted in RRMSE value of 30% as compared to the value of 10.2% for the original *Q*-predictive model. The reason for poor performance of the modified model is certainly because other parameters such as *CS* and *SM* have a substantial impact on the volume of runoff or drainage generated from a field for a given rainfall. This validates that using *CS* and *SM*, along with *P*, are fundamental to get better prediction performance for the *Q*-predictive model.

Table 4-1: Results for the *Q*- predictive model developed using M5 tree and MLR model

	***Q*-Predictive Model**							
Algorithm	**10-fold cross-validation**				**Testing**			
	R^2	**MAE**	**RMSE**	**RRMSE**	**R^2**	**MAE**	**RMSE**	**RRMSE**
M5 tree	0.99	0.19	0.23	10.2%	0.98	0.16	0.20	8%
MLR	0.98	0.24	0.30	13.5%	0.99	0.20	0.26	10.2%

The second investigation is related to the use of another learning algorithm – MLR. Table 4-1 presents the performance comparison of the *Q*-predictive model developed using MLR and M5 decision tree algorithm for the training and testing phase. For a small training set of 100 samples, both the models provide almost comparable performance. The model developed using MLR resulted in RMSE of 0.30 and RRMSE of 13.5%. These values are very close to the performance values for the *Q*-predictive model generated using M5 tree algorithm, which resulted in RMSE of 0.23 and RRMSE as 10.2%. Figure 4-5 (b) illustrates the plot of the predicted output for test data using the two models. While trying to analyse this further, the number of training samples is reduced to 50 for developing the models which showed a significant performance difference. M5 tree model results in RMSE of 0.317 and R^2 of 0.984, whereas for MLR model, RMSE increases to 0.915 and R^2 drops down to 0.506, which shows a poor fit of the MLR model for smaller training sample size. This analysis validates the precedence of M5 tree algorithm over MLR, specifically in scenarios with limited available training samples.

For the third assessment, various training set sizes (65, 125 and 250 samples) are drawn randomly from the simulated dataset to develop the *Q*-predictive model. M5 tree algorithm is used here. The performance of the generated *Q*-predictive models is then compared with each other as shown in Table 4-2. As evident in the table, the performance does not vary much and even a small training set of 65 samples retains the optimal performance. The value of RMSE does increase on reducing the number of training samples, however it is still within acceptable range. For

illustration, the predicted Q for test instances of the three models is plotted against the measured data from NRCS method in Figure 4-6, which does not show much deviation.

Table 4-2: Performance of the Q-predictive models based on various training sizes

Q-Predictive Model			
Performance Metric	**Training set size**		
	250	***125***	***65***
RMSE	0.159	0.234	0.317
$\boldsymbol{R^2}$	0.99	0.99	0.98
RRMSE	5.7%	7.5%	5.98%

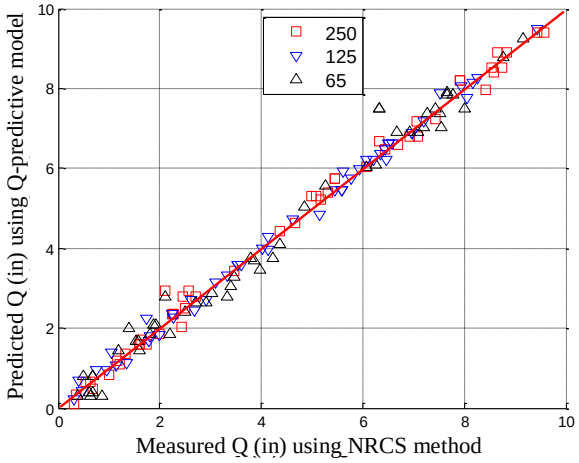


Figure 4-6: Plot of test data for the Q-predictive models developed using different training sizes

4.3.2 Results and Discussion for t_1-Predictive Model

For the model development and evaluation of t_1, a similar approach to the one used for the Q–predictive model, is taken. Initially 100 training instances, drawn from the training set generated for t_1 modelling, are used to develop the t_1-predictive model. The training set contains instances for P, CN, P_d and t_1.

The prediction performance of the 10-fold cross-validated generated model results in RMSE of 0.61, and RRMSE of 30%. As established previously, an acceptable value of RMSE for a model is equal to or less than half of the standard deviation of the values of t_1 in the training data. The acceptable value of RMSE, hence, is calculated as 1.3. Although the RMSE of the generated t_1-predictive model is quite better than the acceptable value, the value of RRMSE is still high. Later in this section the model is developed using more training samples to evaluate if the performance

improves. On the generated t_1-predictive model, test dataset is used to predict values for t_1. These predicted values are plotted against measured values of NRCS method in Figure 4-7.

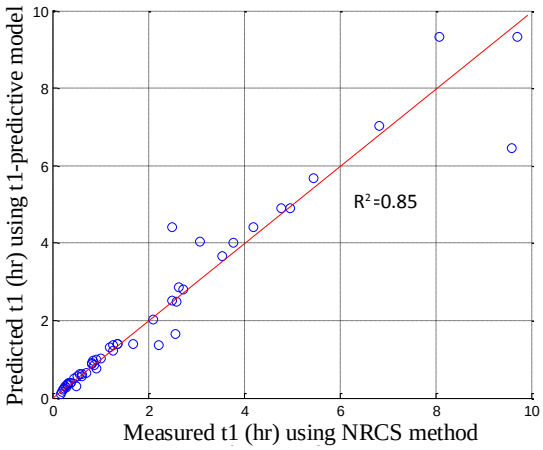


Figure 4-7: Plot of test data for t_1-predictive model with measured values from NRCS method

Similar to the Q-predictive model, the developed t_1-predictive model is further investigated for; (1) the impact of further simplification of model parameters on performance; (2) performance comparison of model developed using MLR algorithm, and (3) the impact of varying training set size on model performance.

For further manipulation of model parameters, the t_1-predictive model is developed using the same model parameters as the ones used for the Q-predictive model, i.e. P, CS, and SM. However, the model performance of the developed model is very poor with RMSE of 1.47, which is greater than the acceptable value of 1.3, and RRMSE as 74%. This validates that the same model parameters cannot be used for predicting t_1. This is further illustrated by plotting the result of test data predicted using the above two models against the measured output of the NRCS model in Figure 4-8 (a). It is evident from the plot that the modified model (developed using P, CS & SM) deviates enormously from 1:1 ratio. R^2 for the modified model comes as 0.667, whereas for the original t_1-predictive model developed using the proposed parameters, it comes as 0.85.

For the next assessment, another learning algorithm, MLR, is used to develop the t_1-predictive model based on the proposed model parameters. The performance of this model is compared against the performance of the t_1-predictive model developed using M5 decision algorithm. The performance evaluation parameters for cross-validation and testing dataset is given in Table 4-3. As evident from the table, there is a substantial difference in predictive performance between the two models. The model developed using M5 tree algorithm results in RRMSE of 30%, whereas

the other model developed using MLR has RRMSE of 50%. This is unlike the modelling results of the Q-predictive models developed using M5 tree and MLR, which showed comparable performance. RMSE values for both the t_1-predictive models are, though, within acceptable range.

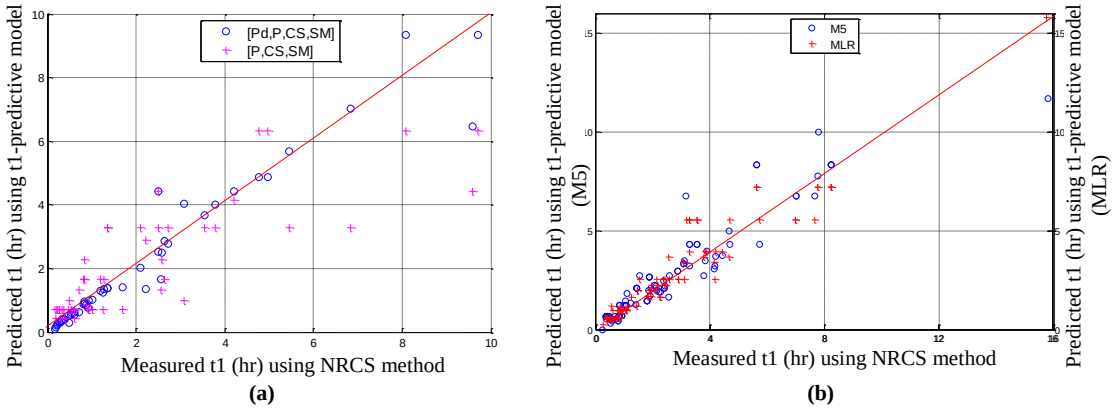


Figure 4-8: (a) Plot of test data for t_1-predictive models developed using M5 tree algorithm for different model parameters, (b) Plot of test data for t_1-predictive models developed using M5 and MLR algorithms for the proposed parameters.

Table 4-3: Results for the t_1 -predictive models developed using M5 tree and MLR models

t_1-Predictive Model								
Algorithm	**10 fold cross-validation**				**Testing**			
	R^2	**MAE**	**RMSE**	**RRMSE**	**R^2**	**MAE**	**RMSE**	**RRMSE**
M5 tree	0.91	0.45	0.61	30%	0.85	0.46	0.56	38.1%
MLR	0.74	0.73	1.00	50%	0.70	0.90	0.95	60%

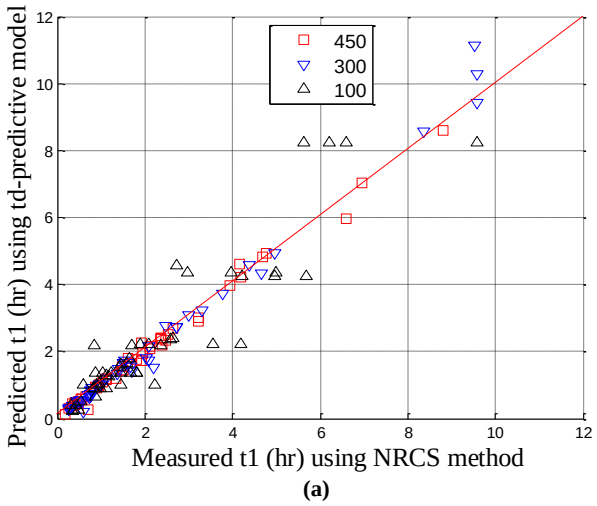
Moreover, test dataset is used for both the models to plot the results against the measured values of NRCS method. As illustrated in Figure 4-8 (b), the plot of the MLR based model shows poor fit, specifically for higher values of t1. R^2 for MLR based model is 0.74 as compared to 0.91 for the M5 based model, as listed in Table 4-3. The low performance of MLR model is attributed to its model architecture. In the MLR based model, the predicted output is simply the mean of the output values associated to the inputs falling in a specific leaf. Whereas, M5 decision tree based model show better performance because they have a linear function model in each leaf. This validates the use of M5 decision model instead of a simpler MLR model.

In the end, to evaluate the impact of training set size on prediction performance, models are generated, based on various training set sizes and M5 tree algorithm. The results of this comparison is listed in Table 4-4. As shown in Table 4-4, the model developed using a training set of 300 samples shows reasonable performance (RMSE=0.318, RRMSE=16.8%). Although

with 100 samples, the RMSE seems adequate (0.825), however RRMSE increases to 27%. Therefore, although a training set of 65 samples results in good performance for the Q-predictive model (RRMSE=5.98%), similar results do not hold for t_l, and more training samples are required. The values of R^2 for the t_l-predictive models developed using different training samples do not vary much. To illustrate this comparison, results are plotted for the predicted values of t_l, for the test data, generated using the three models in Figure 4-9. This figure corroborates that test results for model developed using 100 training samples shows relatively poor fit.

Table 4-4: Performance of the t_l-predictive models based on various training sizes

t_l-Predictive Model			
Performance Metric	**Training set size**		
	450	***300***	***100***
RMSE	0.239	0.318	0.825
R^2	0.985	0.976	0.835
RRMSE	16.1%	16.8%	27%



(a)

Figure 4-9: Plot of test data for t_l- predictive model, generated using various training sizes

4.3.3 Results and Discussion for t_d-Predictive Model

For the model development and evaluation of t_d, initially 100 training instances, are drawn from the training set generated for t_d modelling. The training set contained instances for P, CN, P_d and t_d. The prediction performance of the 10-fold cross-validated t_d-predictive model, developed using M5 tree algorithm, results in RMSE of 0.59, and RRMSE of 8.2%. The acceptable value of RMSE

for t_d-predictive model, is any value equivalent to or less than 1.3. The value of RMSE is well within the acceptable value and RRMSE indicates very good model performance. This is further illustrated by the plot (shown in Figure 4-10) of test data predicted using the t_d-predictive model against the measured values of t_d with NRCS method. The plot shows excellent fit.

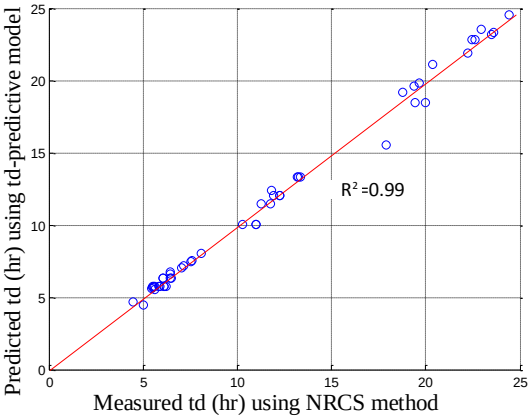


Figure 4-10: Plot of test data for t_d-predictive model, developed using M5 tree algorithm

To demonstrate the impact of further model simplification on the performance of t_d-predictive models, a model is generated using model parameters of *P, CS* and, *SM*. 100 training samples are used to train the model based on M5 tree algorithm. Test dataset is used on the developed model to predict values of t_d. These values are plotted along with the values predicted using the original t_d model (based on P_d, *P*, *CS* and, *SM*) against the measured values of NRCS method in Figure 4-11 (a). The predicted values from the modified model fits poorly to a 1:1 ratio. The RRMSE value of the modified model is 98%, whereas R^2 is 0.10. This demonstrates that the duration of runoff discharge is highly dependent on the duration of the event (rainfall/irrigation).

Table 4-5: Results for the t_d -predictive models developed using M5 tree and MLR models

t_d-Predictive Model								
Algorithm	**10 fold cross-validation**				**Testing**			
	R^2	***MAE***	***RMSE***	***RRMSE***	***R^2***	***MAE***	***RMSE***	***RRMSE***
M5 tree	0.99	0.45	0.62	9.6%	0.99	0.45	0.56	7.4%
MLR	0.96	0.75	1.01	15%	0.96	0.82	1.05	13.9%

Now, a comparison of the two learning algorithms (M5 tree and MLR) is done for the development of the t_d -predictive model. Table 4-5 presents the performance comparison of the t_d -predictive model developed using MLR and M5 decision model algorithms for the training and testing phase. The results indicate that the performance of the M5 based model is better than the

one developed using MLR. RRMSE of the former model is 9.6% and RMSE is 0.62, whereas for the latter model, these values are 15% and 1.01 respectively. However, the models show adequate fit on the plot as illustrated in Figure 4-11 (b).

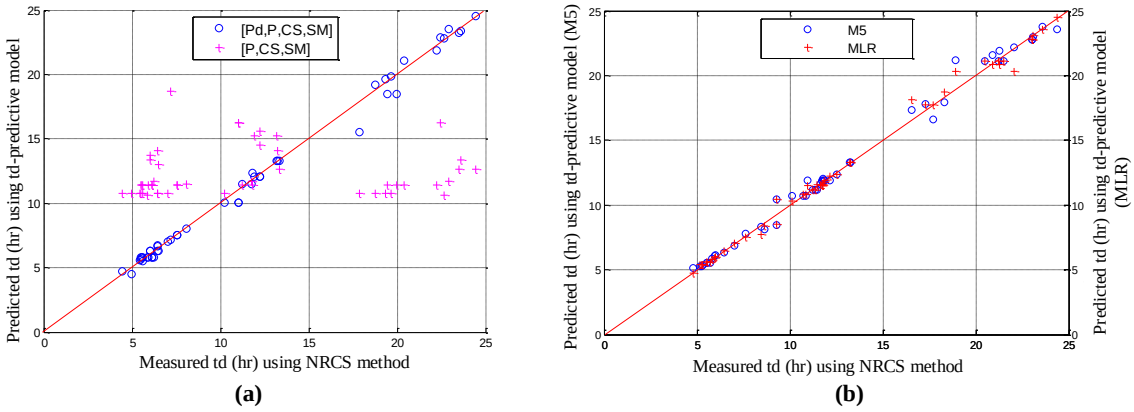
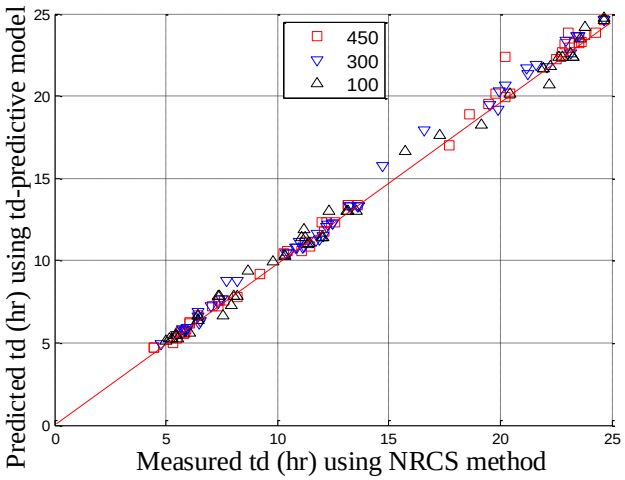


Figure 4-11: (a) Plot of test data for t_d -predictive models developed using M5 tree algorithm for different model parameters, (b) Plot of test data for t_d -predictive models developed using M5 tree and MLR algorithm for the proposed parameters.

Furthermore, to evaluate the impact of training set size on prediction performance, models are generated using different training set sizes and M5 tree algorithm. As shown in Table 4-6, the model developed using a training set of 300 shows reasonable performance (RMSE=0.299, RRMSE=6%). Therefore, although a training set of 65 samples gave good performance (RRMSE =5.98%) for the Q-predictive model, similar results do not hold for t_1, and more training samples were required. However, unlike t_1, for t_d even a small training set of 100 shows good correlation (R^2=0.990, RRMSE =8%) as shown in Figure 4-12).

Table 4-6: Performance of the t_d predictive models based on various training sizes

t_d-Predictive Model			
Performance Metric	***Training set size***		
	450	***300***	***100***
RMSE	0.275	0.299	0.598
R^2	0.997	0.977	0.991
RRMSE	5%	6%	8.2%



(b)

Figure 4-12: Plot of test data for t_d-predictive model, generated using various training sizes

4.3.4 Summary

Since resource constraints of network nodes (e.g. power consumption, computing power etc.), and unavailability of in-situ, inexpensive, and low-power bio-geo-chemical sensors, require a simplified predictive model for in-network deployment. Therefore low-complexity model parameters are derived from the existing NRCS model, utilising real-time and transient field values. In this chapter, low complexity models for Q, t_1, and t_d are developed by reducing the number of parameters to less than 50% in comparison to the NRCS model. Evaluation of the predictive models, developed using simulated training data and M5 tree algorithm, demonstrates accuracy of 84-94% compared with the traditional NRCS curve number model. The Q and t_d model was tested to perform with 94% and 92% accuracy respectively, even for a small training set of around 100 samples; however the t_1 predictive model required a minimum of 300 training samples to show reasonable performance with about 84% accuracy. 10-fold cross-validation of these models demonstrates RRMSE of 10.2%, 30% and 9.6% for Q, t_1 and t_d respectively. Furthermore, performance of these models is validated using MLR model, while it is observed that models developed using M5 tree algorithm demonstrated better performance.

Chapter 5: Validation of the *Q*-Predictive Module

In the previous chapter, the modelling and simulation of Q, $t_{1,}$ and t_d predictive models was presented. This chapter explores the validation of the discharge dynamics models using a year-long dataset from a catchment in Ireland. Due to the unavailability of relevant data required for $t_{1,}$ and t_d models, the chapter only presents validation for the *Q*-predictive model. In this chapter, section 5.1 discusses the experimental setup undertaken for the validation of the model, with respect to detail on the acquired dataset, model evaluation criteria, and pre-processing of the dataset along with sensitivity analysis. Section 5.2 demonstrates the results and discussion for training and testing of the *Q*-predictive model, as well as its performance comparison with state-of-the-art modelling approaches in this domain.

5.1 Experimental Method

5.1.1 Description 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 responsible for nutrient losses from soil to water bodies within the catchment [173], and thus fits the requirement for validating the *Q*-predictive model. This catchment consists of smaller nested sub-catchments. Figure 5-1 (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 5-1 (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 datasets are publically available for research and education purposes via the Environmental Protection Agency (EPA) website [174]. The remainder of the data regarding field conditions is extracted from catchment descriptors available in the associated documentation [173].

For catchment 1, the cumulative rainfall for the year 2002 was 1812 mm. The cumulative stream flow depth measured, at site1, was 1206 mm of the rainfall (as shown in Figure 5-1 (c)). Stream flow here consists of water passing this point that originated as any surface runoff (above ground), sub-surface drainage (below surface) or deeper groundwater contributions by catchment 1 [175]. 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 TON 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, starts in early March and ends 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.

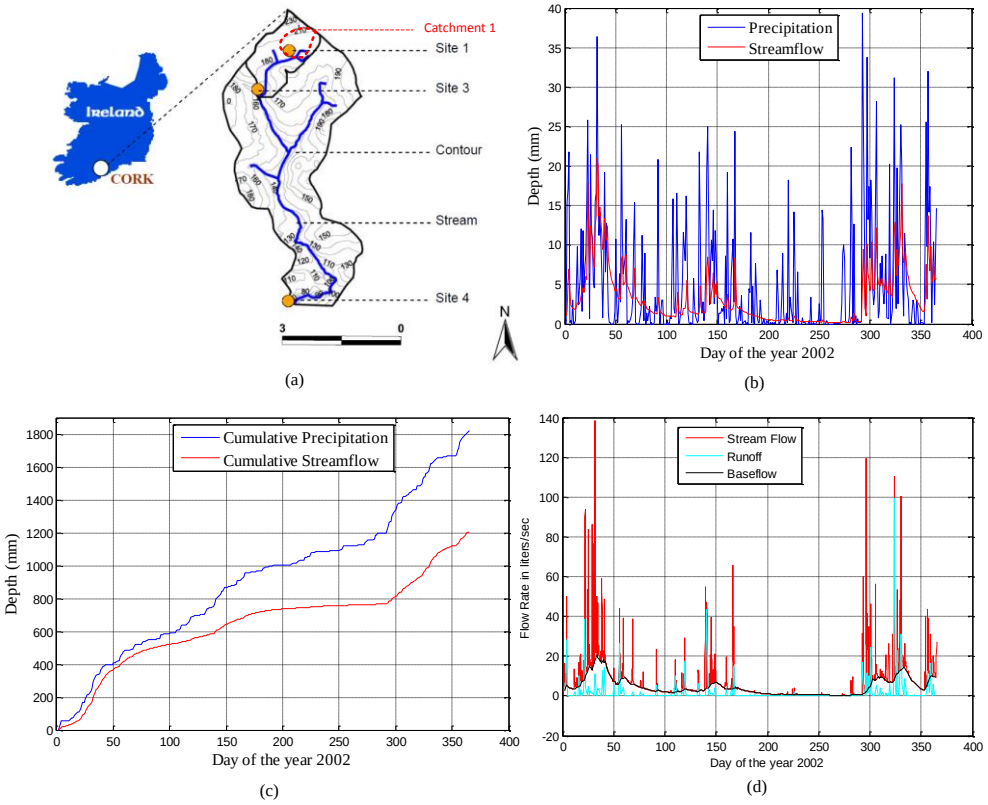


Figure 5-1: (a) Location and map of the Dripsey catchment (reproduced from [175]); (b) precipitation and stream flow (mm) at site1; (c) cumulative precipitation and stream flow depth for site 1; (c) rate of stream flow, runoff, & base flow for site 1

5.1.2 Data Pre-processing & Sensitivity Analysis

In order to validate the *Q*-predictive model, it needs to be trained using real measured 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, the last-5-day-rainfall value is used 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 [68] for drainage discharge prediction which uses M5 trees and 10 input parameters. Later sections will demonstrate that the proposed model still perform comparably well with the existing models developed for discharge prediction with the advantage of using far fewer training samples.

The set of observations required for training the *Q*-predictive model was obtained after some pre-processing of the available dataset from the Dripsey catchment [174]. This is the only dataset that could be found with high temporal resolution event data for an entire year. From the available dataset, instances related to 30 minute precipitation (mm) and stream flow ($lsec^{-1}$) are used for the year 2002. Pre-processing of data is required as the available data is not in the form, e.g. in terms of resolution and units, which can be used directly. Since the model is aimed at facilitating management decisions regarding drainage and nutrients reuse, the 30-minute values of precipitation (*P*) is converted 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.

The remaining parameters required for the *Q*-predictive model such as *SM* and *CS* were respectively obtained using a proxy value and from the information available in the documentation for this study [173]. 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 measuring output of the model, discharges from the catchment, runoff and stream flow rates were all available. Stream flow is selected as an output for our model for various reasons;

1. runoff value is available as a single measurement per day, which does not provide complete information for the daily runoff depths,
2. due to the presence of high base flows, as shown in Figure 5-1 (c), stream flow better represents the discharges from the grass land
3. Pearson correlation coefficient (r) (which explains the strength and direction of the linear relationship between parameters) provides better values for stream flow as compared to runoff with the input parameters as explained below.

To support the above statement 3, the value of 'r' for stream flow with *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 needs 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 5-2 illustrates the sensitivity analysis of the input model parameters with *Q*.

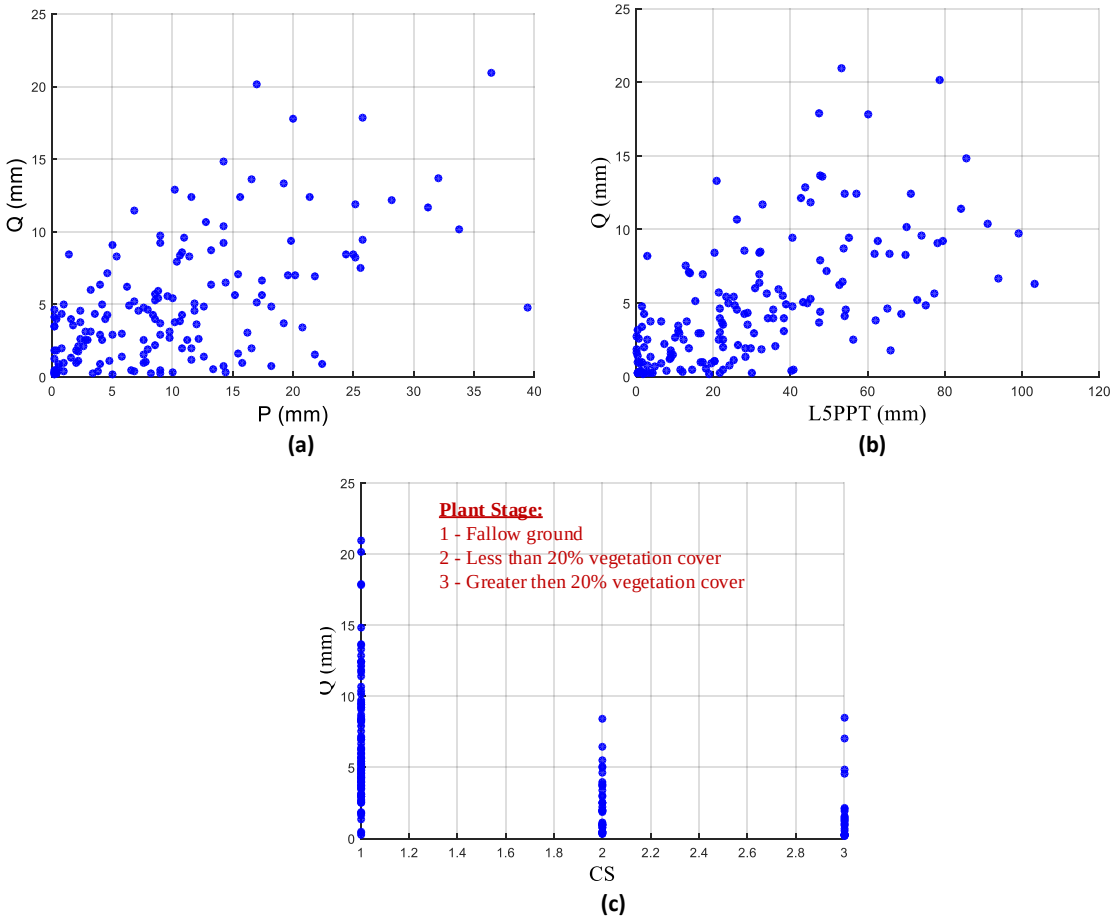


Figure 5-2: Sensitivity analysis of the output parameter, *Q*, with the input parameters, (a) *P*, (b) *L5PPT*, and (c) *CS*

5.1.3 Model Evaluation Criteria

An assessment procedure for model evaluation of the *Q*-predictive model includes;

(i) Evaluation of optimized input parameter combinations with optimal performance;
(ii) Random sampling and 10-fold cross-validation to avoid over-fitting of the model;
(iii) Performance measuring parameters (RMSE, R^2, MAE, and RRMSE);
(iv) Comparative assessment of the prediction accuracy against other similar research studies developed using M5 tree algorithm, and;
(v) Uncertainty analysis on the model residuals

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. In chapter 4, for comparative assessment of the *Q*-predictive model developed using M5 algorithm, another algorithm was used - MLR model. Since the results in chapter 4 established that M5 tree model delivers better prediction accuracy compared to MLR, therefore, for the validation of the *Q*-predictive model, the M5 tree algorithm is used. For the comparative assessment of the validated model, the state-of-the-art research studies, also developed using M5 tree algorithm, are used. Similar to the previous chapter, the M5 tree toolbox [158] developed in Matlab is used for generating the model. Additionally, another criterion is used to analyse the prediction error in the validated model – uncertainty analysis. Below, the detail on the undertaken comparative assessment and the uncertainty analysis is presented.

5.1.3.1 Comparative Assessment of *Q*-Predictive Model with Similar Models

In order to evaluate the accuracy of the proposed *Q*-predictive model, its performance is compared with recent and relevant research efforts in the same area. Hydrologic models, developed using M5 trees and real experimental data, by *Solomatine et al.* [73], *Kuzmanovski* [68], *Corzo et al.* [71] and *Galelli et al.* [58], have been used for comparison. These works used various input parameters and thousands of training samples for developing the models. They were either 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.

5.1.3.2 Uncertainty Analysis

As in any prediction, there is a potential error which needs to be accounted for [176]. 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 [177]. Through the analysis;

1. it is investigated if residuals show any temporal trend over the year, and ;
2. confidence intervals for residual errors are determined

5.2 Results and Discussion for *Q*-Predictive Module

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 tree algorithm was then used to generate the tree. The average value for *Q* in the training dataset is 3.8 mm, 25th percentile is 1.01 mm, 75th percentile is 5.29 mm, and 90th percentile is 9.22 mm. The standard deviation of *Q* is 3.8.

Table 5-1: 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	1.33	35.9
10-fold cross-validation	0.64	1.3	1.95	55.9

Performance parameters for the generated M5 tree indicate a high value for R^2 of 0.823 as listed in Table 5-1. 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 [172], which for this dataset is 1.9. The obtained RMSE of the trained model is well within this limit. Furthermore, the RRMSE is 35.9%. Moreover, 10-fold cross-validation of the *Q*-predictive model results in R^2 of 0.63 and RRMSE of 55.9%. 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 chapter 4, 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 fewer parameters of the actual mathematical model are selected, have higher predictive performance.

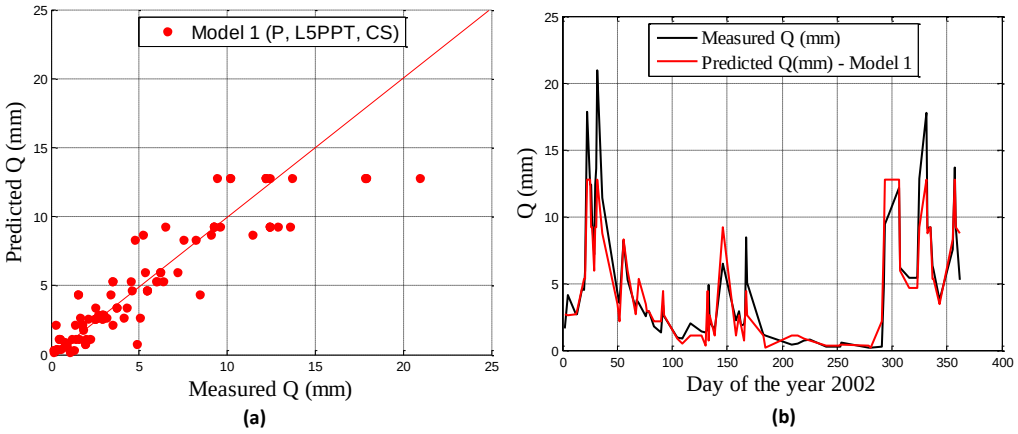


Figure 5-3: (a) Scatter plot of the predicted *Q* using the proposed model against measured *Q*, (b) Curve plot for predicted and measured *Q* plotted against days of the year

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 5-3. The model illustrates a good fit with R^2 of 0.868 as shown in Figure 5-3 (a). The predicted values for test data and measured values are also plotted against days of the year in Figure 5-3 (b) to represent the difference between the two curves. As is evident both curves almost overlap 50-60% of the time; however in the first 50 days of the year, the model seems to under-predict the stream values by about 25%. Further analysis on determining confidence intervals provides insight into the potential of these results. This is covered in section 5.2.2. Also, it is evident in Figure 5-3 (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 in the future would need to verify this.

5.2.1 Further Model Simplification – viable or not?

In this section, it is evaluated if subsets of the proposed model, i.e. any further simplification to the model parameters, impacts the model performance in any way. This will 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 [58, 71, 73]. For

this, two models are developed using '*P* and *L5PPT*' and only '*P*' as input parameters. These models are called 'model-2' and 'model-3' respectively. The proposed 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 5-2). It is clear from the table that, model-1 shows 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 the same. 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 5.1.2. However, by using only *P* as an input in model-3, a very weak model is generated with 70% RRMSE value and R^2 of 0.5. This means that the predicted results using model-3 will have only 30% accuracy.

Table 5-2: Comparison of *Q*-predictive models developed using different input parameters

Model No.	**Input Parameters**	**Performance Metrics**				**10-fold Cross-Validated Performance Metrics**			
		R^2	***MAE***	***RMSE***	***RRMSE (%)***	***R^2***	***MAE***	***RMSE***	***RRMSE (%)***
1	***P, L5PPT, CS***	0.86	0.80	1.33	35.9	0.63	1.29	1.95	55.9
2	***P, L5PPT***	0.84	0.97	1.51	39.2	0.58	1.51	2.17	61.4
3	***P***	0.50	1.85	2.65	70.1	0.08	2.26	3.05	91.4

To illustrate the comparison among models further, test dataset 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 5-4 and Figure 5-5. Model-2 shows a good fit (as shown in Figure 5-4 (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 5-4 (b) to represent the difference between the two curves. Figure 5-4 (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 was shown in Figure 5-1 (b)). Furthermore, the predicted *Q* for test data using model-3 is plotted in Figure 5-5 (a), which shows a poor fit, with an R^2 value of 0.5. 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 5-5 (b) further illustrates a plot of the deviation of predicted *Q* compared to the measured *Q*. The model seems to over-predict during summer and under-predict during winters leading to an unreliable system.

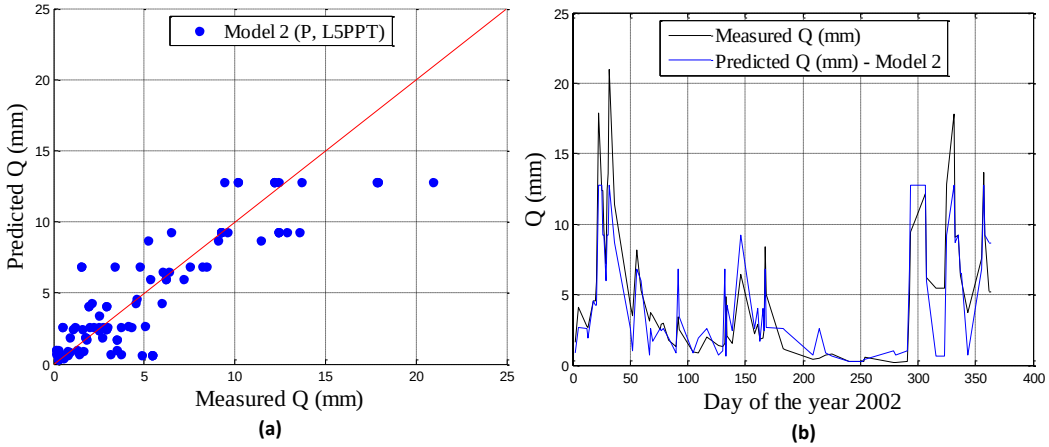


Figure 5-4: (a) Scatter plot of predicted *Q* using model-2 against measured *Q*, (b) Curve plot for predicted and measured *Q* plotted against days of the year

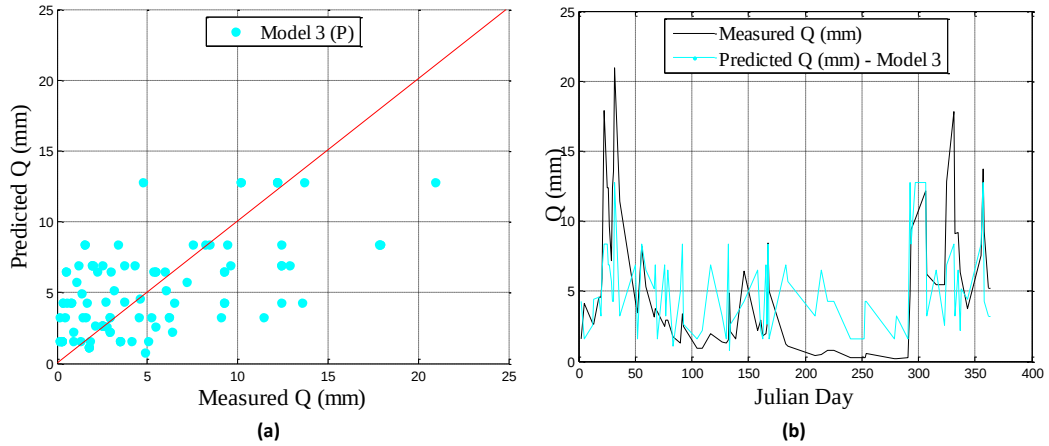


Figure 5-5: (a) Scatter plot of predicted *Q* using model-3 against measured *Q*, (b) Curve plot for predicted and measured *Q* plotted against day of the year

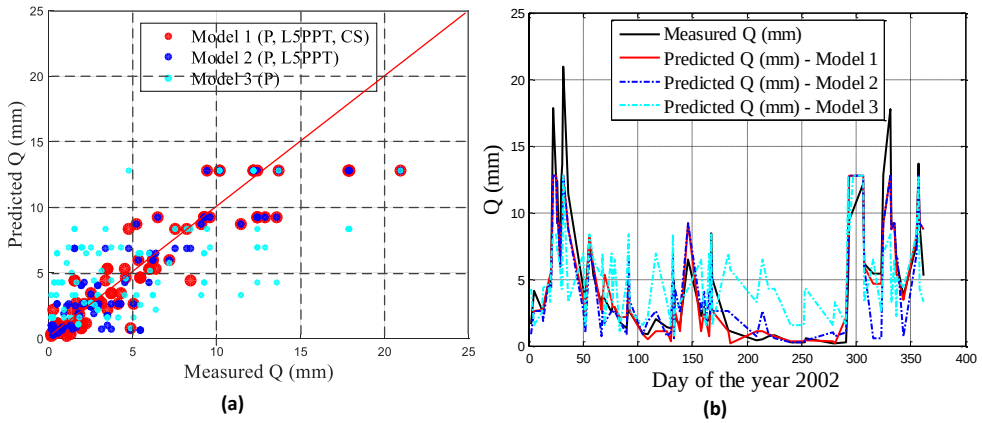


Figure 5-6: Comparison of predicted and measured *Q* for model 1, 2 & 3 in a (a) scatter plot and (b) curve plot

To compare the prediction accuracy for these three models, the results are plotted on a single graph as shown in Figure 5-6. 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 the other models. Hence, it is concluded that the proposed model parameters cannot be further simplified without degrading the performance.

5.2.2 Uncertainty Analysis for *Q*-Predictive Model

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 5-7 (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.

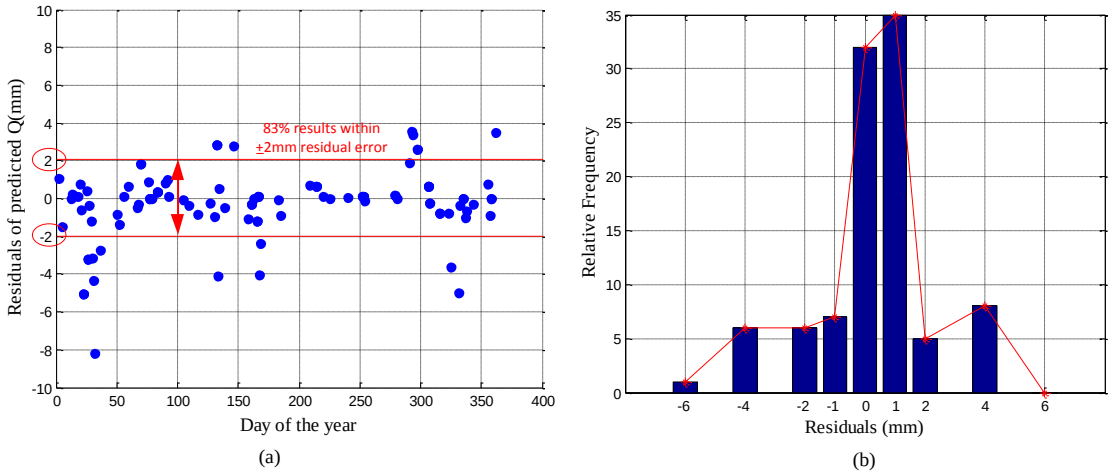


Figure 5-7: (a) Trend of residual error of predicted *Q*, (b) relative frequency of the residual error

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 5-7 (b)). To explain this further, Figure 5-8 (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 error.

A histogram for this shows an approximately normal distribution having maximum number of values with 0% residual error (Figure 5-8 (b)).

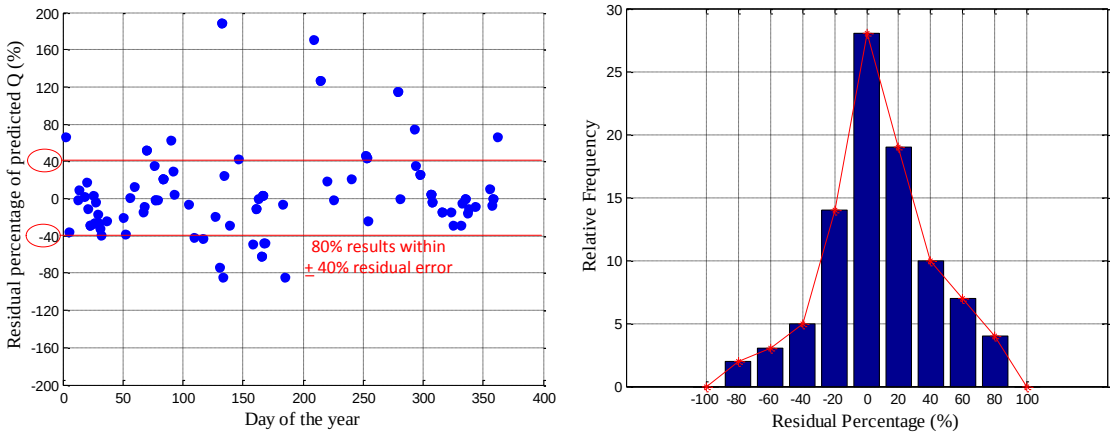


Figure 5-8: (a) Trend of percentile value of residual error versus, (b) relative frequency of percent residual error

5.2.3 Comparative Assessment of *Q*-Predictive Model with Similar Models

In order to evaluate if the proposed model has acceptable (or comparable) performance, the results are compared with the results of the most relevant existing models. Although, it is difficult to establish a reliable comparison in a data-driven paradigm given different data sets or dissimilar application objectives, and assessment criteria adopted. However, an effort is made to list all the available experimental detail of the comparative studies for an honest analysis.

For this, works by *Kuzmanovski* [68], *Solomatine et al.* [73], *Corzo et al.*[71], and *Galelli et al.* [58] have been selected. All of these models are based on the M5 tree algorithm. 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 5-3 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 research.

Overall, it is clear that the *Q*-predictive model has a comparable performance compared to other models. More specifically, work done by *Kuzmanovski* [68] 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 1 ha, in order to control pollutant outflows. The model uses 10 input 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.

Table 5-3: Performance comparison of *Q*-predictive models with the existing models

S. No.	Predictive Models	Output Value	No. of Input variables	No. of Training samples	Drainage area	Cross-Validation for training	RRMSE (%)	R^2
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	55.9	0.63
						Not done	35	0.86
2	*Kuzmanovski* [68]	Daily drainage from farms	10	7000 samples (22 years data)	0.83 - 1 ha field plots	Not done	45 - 65.9	0.75-0.89
3	*Solomatine et al*. [73]	Flood discharge	3 (11 sub values), Rainfall, Past rainfalls, flow	5000 samples (> 13 years data,	106 ha	Not done	-	0.97
3	*Corzo et al.* [71]	Hourly discharge in a stream	Previous runoff, effective rainfall	2000 (8 years data)	37,000 ha	Not done	-	0.89
4	*Galelli et al.* [58]	Hourly Stream flow	2 (rainfall & inflow value), 3 time lag sets	24120 (2.5 years data)	10,000 ha	10-fold	48	-

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 *Kuzmanovsk*i used only 2–day

rainfall value, which may not accurately represent soil moisture conditions. The performance of the *Q*-predictive model would need to be further 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 input 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, the RRMSE of the model proposed by *Galelli et al*. is 48% compared to 55.9% RRMSE of the *Q*-predictive model.

The proposed *Q*-predictive model is ranked between the models developed by *Kuzmanovski* [68] and the others [58, 71, 73] in terms of the number of parameters used. The *Q*-predictive model utilizes climatic conditions as an input parameter similar to the latter models. However the model also uses field conditions, crop cover, and soil moisture, unlike the model proposed by *Kuzmanovski* [68]. Furthermore, the *Q*-predictive model keeps the number of parameters much simpler and fewer compared to *Kuzmanovski* [68].

5.3 Summary

In this chapter, a discharge (*Q*) predictive model for the proposed simplified parameters is successfully validated by employing M5 tree algorithm. 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:

- Performance measures for the proposed *Q*-predictive model provide good performance; R^2 is 0.868, 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 55.9%.
- 83% of the residuals for the predicted test values fall within $\pm$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 $\pm$40% residual errors.
- 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%.
- In comparison to the existing models developed for discharge predictions using M5 tree algorithm, 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.

Chapter 6: Modelling and Validation of *TON*-loss Predictive Module

This chapter presents the modelling and validation of low-complexity *TON*-loss Predictive models. Furthermore, the applicability of the M5 tree algorithm is explored for nitrate loss modelling based on the proposed parameters. Section 6.1 presents the development of model for *TON* by seeking simplification in the complexity of model parameters. Section 6.2 outlines the experimental method undertaken for the model validation, with respect to the adopted datasets, and the evaluation criteria. Section 6.3 demonstrates the results and discussion for training and testing of the *TON*-loss predictive model, as well as its performance comparison with the state-of-the-art modelling approaches in this domain.

6.1 Model Development of a Low-Complexity *TON*-loss Predictive Module

As discussed in the introduction, the adoption of WSNs for nutrient management in general and the implementation of WQMCM framework specifically on a farm, requires simplified predictive models based on fewer, and ideally, real-time field parameters acquired autonomously and shared by the neighbouring farms.

In order to simplify the model parameters, the model abstraction done by *Villa-Vialaneix et al.* [65] is further extended and simplified. In that work, the input parameters consisting of 11 variables were themselves drawn from the dataset of the DNDC model based on a preliminary sensitivity analysis and expert evaluation. Table 6-1 lists input parameters for the two models (columns 2 and 3) under various input categories. The abstracted list of parameters, which still contains soil chemistry and N sources data, is further abstracted, with the addition of a few other parameters to get the simplified parameters (3rd level abstraction, column 4). This 3rd level abstraction is explained below.

For the input category of climatic conditions, only precipitation is selected. Temperature is not selected because temperature readings are taken into account in a model to imply the rate of evapotranspiration that would have occurred, which when combined with other soil properties (such as soil field capacity, soil porosity, soil texture) indicate the soil moisture conditions and the eventual discharge flux from the soil [142]. To put it simply, the higher the temperature, the higher is the rate of evapotranspiration and the drier is the land, resulting in low discharge fluxes.

Since with WSNs, it is now possible to measure soil moisture directly with small and cheap sensors (as discussed in chapter 2), the dependence on proxy parameters can be minimized.

Table 6-1: Model parameter abstraction for the *TON*-loss predictive model

Input variable category	**Parameters used in DNDC model *Li et al.* [142]**	**Parameters used by *Villa-Vialaneix et al.* [65] *(2nd level Abstraction)***	**Proposed parameters for *TON*-loss predictive model *(3rd level Abstraction***
Climatic conditions	Precipitation	Precipitation	Precipitation
	Temperature	Temperature	
Soil Properties	Soil type		
	pH	pH	
	Redox		
	Carbon content	Carbon content	
	Bulk density	Bulk density	
	Clay content	Clay content	
	Temperature		
	denitrifying potential		
	Field capacity		Soil moisture
N input sources	Profile Mass		
	N in fertilizer	N in fertilizer	N in fertilizer
	N in manure	N in manure	N in manure
	N from precipitation	N from precipitation	
	N in plant residue	N in plant residue	
	N from fixation	N from fixation	
	N from mineralization		
			Total N applied so far
			Days since Last N application
Management Information	Crop cover		Crop cover
	Tillage		
	Crop rotation		
Additional Parameters			Day of the year

Additionally, the impracticalities attached with the methods used for acquiring these soil parameters have also been elaborated previously. Moreover, these soil parameters can be ignored since the geographical and time extent of the model in this research study is limited, as these parameters do not change over the scales considered. Therefore, in the category of soil properties, it is proposed to only use soil moisture.

Moving on to the category of N input sources, N in fertilizers and in manure are selected from the parameters listed in the 2nd level abstraction. This is because of the ease of the availability of this information unlike other parameters listed in this category (N from precipitation, plant residue and fixation) which require laboratory analysis of soil samples and mathematical modelling [142]. Besides these two parameters (N in fertilizers and manure), it is proposed that two additional parameters be used – days since last N application and cumulative N applied so far that year.

The reason for this is attributed to the fact that nitrate applications and subsequent nitrate fluxes do not always have a linear relationship. This means that high monthly exports of nitrate do not always coincide with large monthly inputs of nitrogen fertiliser [66, 173]. Therefore, additional information related to N application is needed to develop a better relationship between N inputs and N fluxes. This will be corroborated in the later sections by the sensitivity analysis of the considered dataset and the model evaluation.

In the management information category, it was proposed to use none of the suggested parameters in the 2nd level abstraction done by *Villa-Vialaneix et al.* [65]. This is because the dataset was comprised of annual values for the parameters, therefore using annual averages for these variables would not have possibly contributed much to the model development. However, crop cover is selected in 3rd level abstraction for two reasons. Firstly, our model is looking at daily nitrate fluxes at the field scale in which vegetation cover can play an important role. Crop cover hinders outflows as well as impacting nutrient losses as nutrients are absorbed more in the initial stages of a crop [178]. Secondly, the availability of methods, using WSNs, enables autonomous monitoring of crop cover. For example, methods such as field imaging and signal attenuation methods have been used to determine the plant biomass autonomously [169]. This leads to autonomous interpretation of the crop stage.

An additional parameter, which was not used in either of the two previous models, is day of the year. Daily nitrate fluxes tend to present a seasonal pattern [66, 173] for small scale land areas irrespective of the times of N application. The reason is attributed to the fact that in a nutrient saturated soil, any nutrients applied to the soil are simply lost through drainage [13]. Furthermore,

during crop growing stages, nutrient losses tend to be lower owing to absorption of nutrients by the crops. The conclusion from this is that there would be a pattern of nutrient losses for a particular land-area throughout the year. Therefore, it is proposed to use day of the year in the *TON*-loss predictive model to investigate its impact on prediction accuracy of the model.

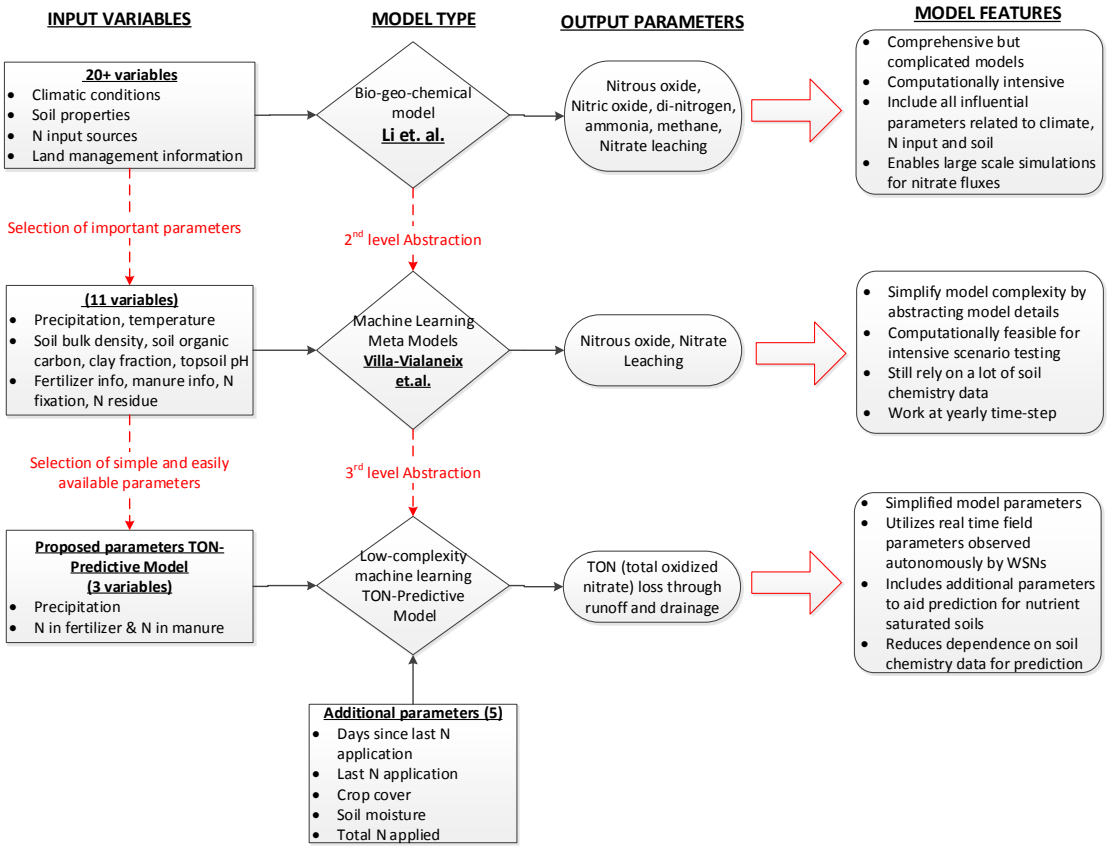


Figure 6-1: Model abstraction from high complexity bio-geo-chemical models to low complexity *TON*-loss predictive model

Thus, the proposed parameters for the *TON*-loss predictive model (column 4, Table 6-1), abstracted from the two complex models, are precipitation, soil moisture, N in fertilizer, N in manure, total N applied so far, days since last N application, crop cover, and day of the year. All of these parameters are easily available. The three levels of abstraction from high to low complexity model parameters, along with model inputs and corresponding output parameters are shown in Figure 6-1.

6.2 Experimental Method

6.2.1 Description of Catchment Data

Once again, data collected at the Dripsey catchment was used [173]. Figure 6-2 (a) show 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 *TON*–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 6-2 (a)) consisting of 17 ha of farmland.

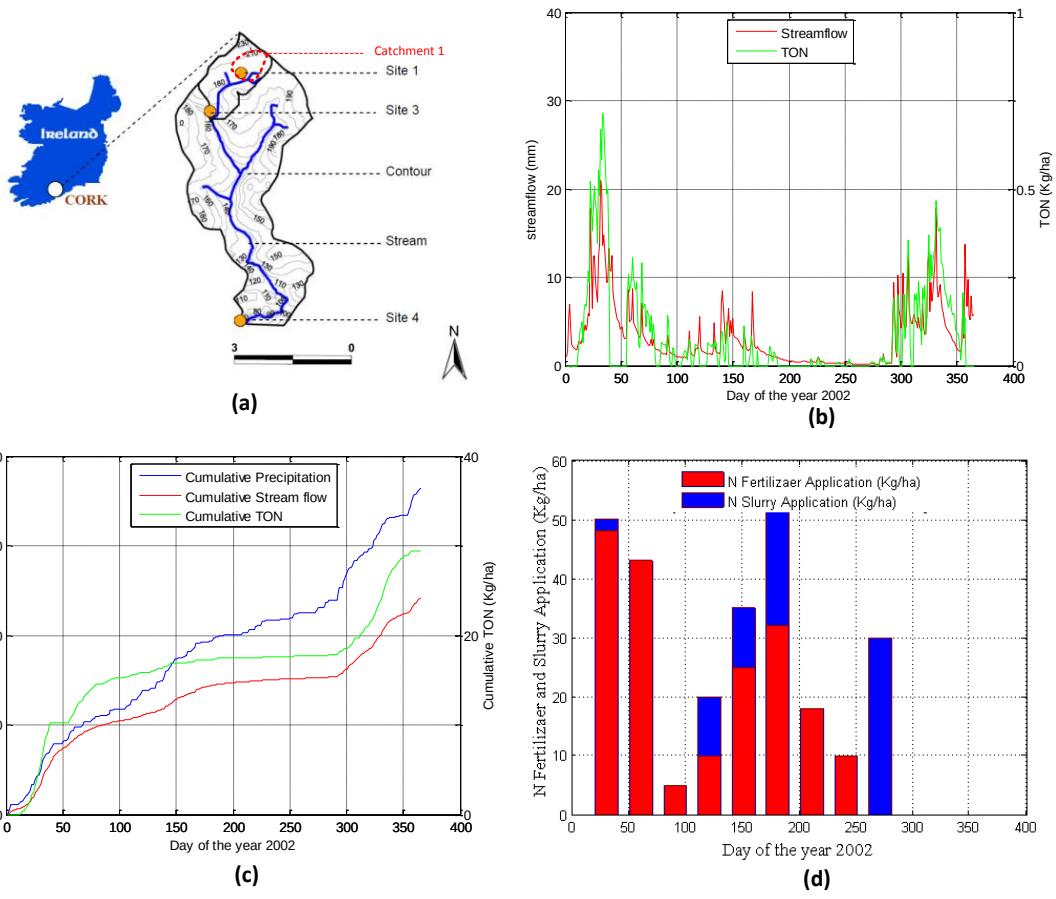


Figure 6-2: (a) Location and map of the Dripsey catchment [175]; (b) plot of observed stream flow and *TON* at site1 and; (c) cumulative precipitation, stream flow and *TON* losses observed at site 1 for year 2002; (d) fertilizer and slurry applications at site1 in the year 2002

Precipitation (mm) and *TON* export concentration (mgl^{-1}) data, collected every 30 minutes for the year 2002 are used. The lower bound of the estimated annual export of *TON* for catchment 1 was at 29 $KgNha^{-1}$, whereas the upper bound was 69 $KgNha^{-1}$. This was the highest nitrate export loads among all the sub-catchments measured in Dripsey. It can be clearly seen from Figure 6-2 (b) that the export of *TON* observed at site1 is strongly related to stream flow emanating from catchment 1. The lowest concentrations of *TON* losses occurred in early to mid-summer. The highest concentrations occurred in winter.

Chemical fertiliser is applied on all farms. The fertiliser applications generally begin in February and continue at about four to six week intervals until September (as shown in Figure 6-2 (d)). Manure in slurries from livestock shelters is applied irregularly with the amount and timing dependent on what had accrued from winter housing of the stock. In catchment 1, there was a significant application of slurry in December 2001 and in October 2002.

6.2.2 Data Pre-processing & Sensitivity Analysis

The set of observations required for training the *TON*-loss predictive model was created after some pre-processing of the available dataset from the Dripsey catchment in Ireland. The datasets are available for research and educational purposes at the Environmental Protection Agency (EPA) website [174]. Despite many efforts (and perhaps surprisingly) this is the only dataset (of this type) that could be found with high temporal resolution data of *TON* losses for an entire year. From the available dataset, half hourly precipitation (mm) and *TON* losses (litres sec^{-1}) for the year 2002 is used. The remaining parameters required for the *TON* predictive model were either obtained using a proxy value or were extracted from the information available in the documentation for this study [173].

Since the *TON*-loss predictive model is aimed at facilitating daily management decisions regarding nutrient loads ($Kg\ Nha^{-1}$), the hourly values are converted into daily loads. For soil moisture data, which are not available in the dataset (as sensor technology, still new at the time, was not adopted in this study), an alternative method is used. Instead of using the mathematical method suggested in the DNDC model [142] (because of the complexity of the required parameters) a proxy parameter is used - the last-5-day-rainfall value. This proxy value has been widely used in hydrological models, such as in the NRCS curve number model, to represent soil moisture conditions, although there are questions about its accuracy [130]. Nevertheless, this is the best available at present, and so offers a worst case performance baseline. When real soil moisture readings become available to the model, performance should improve. Therefore, for

each of the daily precipitation values, last-5-day-rainfall is computed. Using this value, moisture levels are determined according to the thresholds provided for growing and dormant seasons in the NRCS curve model [179]. For example, in a fallow season, field conditions are considered dry, medium and wet respectively if rainfall depths are less than 13 mm, between 13 mm and 28 mm, and greater than 28 mm. Respective thresholds for rainfall depths are set for a growing season, which are higher than the ones for dormant season.

For crop cover data, similar to the previous chapter, information regarding the growing stages of grass in catchment 1 was used to estimate crop cover throughout the year. According to crop cover values, crop levels are assigned such that fallow land is referred to as stage 1; coverage less than 20% is defined as stage 2 and; coverage greater than 20% is assigned stage 3. Similarly, information regarding N fertilizer and slurry inputs to catchment 1 were extracted from a thesis based on the same project [173] and added to the dataset. Based on the N application rates and timings, cumulative-N-application for each day and days-since-last-N-application are computed. The final dataset contains all the proposed attributes for the *TON*-loss predictive model. The mean daily *TON* for the dataset is 0.099 $Kgha^{-1}$, 25th percentile is 0.040 $Kgha^{-1}$, 75th percentile is 0.22 $Kgha^{-1}$, and 90th percentile is 0.716 $Kgha^{-1}$. The standard deviation is calculated as 0.156 $Kgha^{-1}$.

Table 6-2: Pearson correlation coefficient for the data from Dripsey catchment

Independent Parameters	**Acronym**	**Pearson (*r*) Correlation Coefficient with *TON* ($Kgha^{-1}$)**
Day of the year	*DY*	-0.20
Precipitation (mm)	*PPT*	0.32
Soil Moisture	*SM*	0.71
Crop Stage	*CS*	-0.52
Last N Fertilizer Application (Kg/ha)	*NF*	0.12
Last N Slurry Application (Kg/ha)	*NS*	0.05
Days since last Fertilizer/Slurry Application	*DNFS*	0.31
Cumulative N applied so far this year (Kg/ha)	*CumN*	-0.27

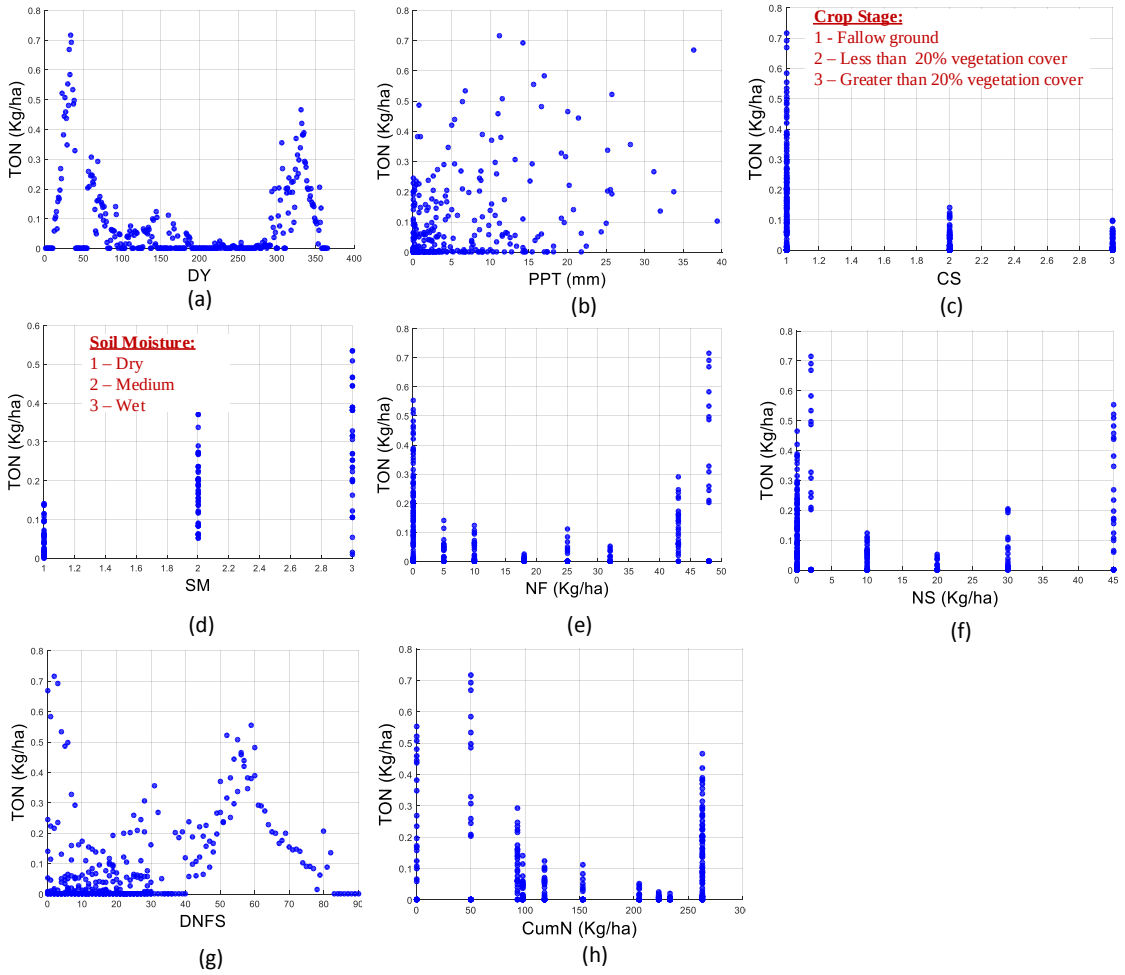


Figure 6-3: Sensitivity analysis of input parameters ((a) *DY*, (b) *PPT*, (c) *CS*, (d) *SM*, (e) *NF*, (f) *NS*, (g) *DNFS*, (h) *CumN*) for the *TON*-loss predictive model with the output parameter *TON*

In the obtained dataset, instances with zero precipitation values are excluded although the impact of no rain remains in the model (using the last-5-day-rainfall value). This exclusion reduces the sample set to 200 instances. A sensitivity analysis was done on the obtained dataset to analyse the correlation of various independent variables with *TON* losses. The values for Pearson coefficient of correlation (*r*) are given in Table 6-2, along with the acronyms for these variables which will be used from here on in this thesis. Furthermore, the sensitivity plots have also been drawn for all the independent variables in the dataset against daily *TON* loss parameter as shown in Figure 6-3. It is important to note in the sensitivity plots that a non-linear relationship is illustrated for most of the input parameters with *TON*, e.g. for *DY*, *NS*, *NF*, *DNFS* and *CumN*, which cannot be explained with a linear regression line. Hence, all these parameters would be used initially for

model development (in section 6.3) and later scrutinized in detail for exclusion of any parameter (in section 6.3.2).

As discussed, *TON* losses show a definite timely trend with *DY* (Figure 6-3 (a)), though negative with low correlation value of -0.20. This means that more *TON* losses are observed during the initial months of the year, possibly owing to high slurry application and rainfalls. With *PPT*, there is a positive correlation with *TON*, though not very strong, as one would expect, as shown in Figure 6-3 (b). The strongest relationship of *TON* losses is with *SM* (r = 0.71), as illustrated in Figure 6-3 (d), which shows that the higher the soil moisture level, the greater are *TON* losses. Conversely, *TON* has a strong negative correlation with *CS* (r = -0.52). As illustrated in the sensitivity plot of *CS* with *TON* in Figure 6-3 (c), the higher the crop stage, the lesser are the *TON* losses. It is interesting to note that the *NF* and *NS* have the weakest correlation value with the *TON* (0.12 and 0.05 respectively), as is also visible through the sensitivity plot in Figure 6-3 (e) and (f). This is possibly attributed to the fact that specific N inputs do not correlate temporally with day-specific N losses for a nutrient saturated soils. Since N-saturated soils have a large "N store" from previous N-application events therefore, further N applications to N-saturated soil are lost rapidly [13]. However, as proposed in the previous section, *DNFS* and *CumN* show a relatively better correlation with nitrate losses (0.30 and -0.27 respectively).

6.2.3 Modelling Technique

For nitrate model, as discussed in chapter 2 (section 2.4.2), various machine learning algorithms have been used including ANN, decision trees, and statistical methods. However, for *TON*-loss predictive model, the applicability of the M5 tree algorithm is explored for nitrate loss modelling based on the proposed parameters. This is because M5 tree algorithm is a simpler algorithm compared with ANN and provides comparable performance with less computational time, however it has not yet been applied in nitrate modelling. The M5 tree toolbox [158] developed in Matlab is used for generating the model.

6.2.4 Performance Evaluation Criteria

Similar to chapter 5, a six-step assessment procedure is adopted including;

(i) Selection of an optimized input parameter combination with optimal performance;

(ii) Random sampling of the observational dataset to ensure a robust evaluation of the model's performance, and the use of 10-fold cross-validation to avoid over-fitting of the model;
(iii) Multi-criteria assessment of the model performance (RMSE, R^2, MAE, and RRMSE);
(iv) Comparative assessment of predictive accuracy efficiency of M5 tree based *TON*-loss predictive model against state-of-the art modelling approach in this domain;
(v) Comparative assessment of predictive accuracy efficiency of M5 tree based *TON*-loss predictive model against other traditional data-driven approaches (ANNs, REPTree and MLR). And;
(vi) Uncertainty analysis on the model residual.

6.3 Results and Discussion for Validation of *TON*-loss Predictive Module

Using 75% of the pre-processed dataset as a training set consisting of parameters *DY, PPT, SM, CS, NF, NS, CumN, DNFS* and *TON*, the predictive model is generated by utilizing M5 toolbox in Matlab. The generated *TON*-loss predictive model shows good performance with R^2 equivalent to 0.927, MAE as 0.024, RMSE as 0.040 and RRMSE as 26.4%. The 10-fold cross-validated results indicate R^2 as 0.727, MAE as 0.043, RMSE as 0.065 and, RRMSE as 47.7%. As discussed earlier, RMSE values less than half of the standard deviation of the measured data may be considered adequate for model evaluation [180]. For the training dataset, this value is calculated as 0.078. Even in 10-fold cross-validated result for the model, the value for RMSE (0.065) falls well below this threshold.

For testing the model, test samples are drawn from the remaining 25% of the dataset. Test results of the predicted *TON* values are plotted against measured *TON* as shown in Figure 6-4 (a). The scatter plot shows a very good fit with R^2 equal to 0.91. To illustrate the difference between the predicted and measured *TON* curves, these values are plotted against day of the year. It is apparent that both curves overlap significantly, however the model seems to under predict in the last 50 days.

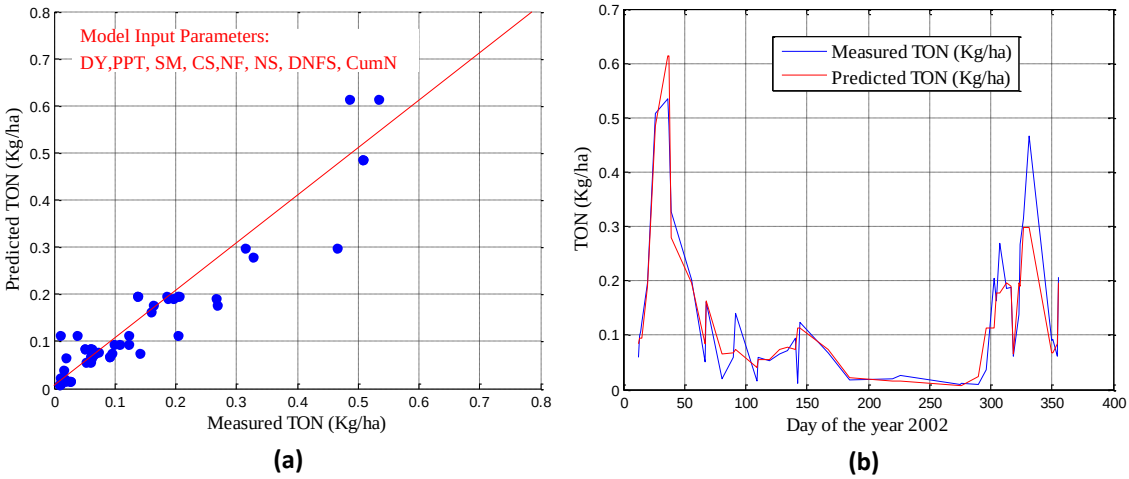


Figure 6-4: (a) Scatter plot of predicted *TON per day*, using *TON*-loss predictive model; (b) plot of predicted and measured *TON* against day of the year.

6.3.1 Comparative Assessment of *TON*-loss Predictive Model with Previous Models

In order to evaluate if the proposed model performs comparably (or better) than the other models, the results are compared with the meta-models developed by *Villa-Vialaneix et al.* [65]. For that research, various machine learning algorithms were used to develop the meta-models with different training set sizes. For performance measurement, only R^2 has been evaluated, whereas cross-validation was not done. The comparison is listed in Table 6-3. In that study, the decision tree based meta-model resulted with R^2 of 0.74 for a 200 training set sample size. The MLP based model gave R^2 of 0.82 for the same training set size. This indicates that our proposed model developed with M5 tree algorithm and simplified parameters results in better performance for daily nitrate losses with R^2 equivalent to 0.92. The reason may possibly be attributed to the fact that the models developed by *Villa-Vialaneix et al.* are for yearly estimates of nitrate losses which can overlook, by oversimplification, the complicated heterogeneous conditions through the year. Furthermore, a different learning algorithm, the M5 tree algorithm, is used. One of the benefits of WSN-based models such as this is that they can gather catchment-specific data for developing accurate models. Although our modelling results show that the proposed model simplification is promising, a more rigorous prospective study with datasets obtained from different catchments is required to validate this further.

Table 6-3: Performance comparison of the *TON*-loss predictive model with an existing work

Model	**Learning Algorithm used**	**Performance Metric (R^2)**
Meta models by *Villa-Vialaneix et al.* [65]	Decision tree	0.74
	MLP	0.82
***TON*-loss predictive model**	M5 tree	0.92

6.3.2 Further Model Simplification

In this section, the impact of further simplification of the model parameters on its predictive performance is evaluated. This is because, some input parameters, e.g. *NS* and *NF*, do not have a stronger correlation with the output parameters despite apparently being very important variables. Furthermore, a non-linear relationship between parameters has also been observed, e.g. with *DY*, *DNFS* and *CumN*, which cannot be explained with linear regression lines. Hence, model parameters are shuffled, in an informed manner, to develop the models and observe the impact on predictive performance.

Table 6-4: Comparison of *TON*-loss predictive models developed using different input parameters

Model No.	***Features for TON-loss predictive model***	***Performance Metrics***				***10-fold cross-validated Performance Metrics***			
		R^2	***MAE* (Kg /ha)**	***RMSE* (Kg/ha)**	***RRMSE (%)***	***R^2***	***MAE* (Kg /ha)**	***RMSE* (Kg /ha)**	***RRMSE (%)***
1	*PPT,NF,NS*	0.764	0.047	0.075	48.1	0.413	0.069	0.101	72.1
2	*PPT,SM,CS,NF ,NS*	0.832	0.039	0.063	40.5	0.532	0.061	0.089	63.9
3	*PPT,SM,CS,NF ,NS,CumN, DNFS*	0.923	0.026	0.042	27.3	0.703	0.047	0.070	50.5
4	*PPT,SM,CS, CumN, DNFS*	0.915	0.027	0.044	28.7	0.681	0.047	0.072	51.6
5	*DY,PPT,SM,CS ,CumN, DNFS*	0.927	0.025	0.041	26.4	0.364	0.046	0.072	50.7
6	*DY,PPT,SM,CS ,NF,NS, CumN, DNFS*	0.927	0.024	0.040	26.4	0.727	0.043	0.065	47.7

It is important to note that this particular evaluation is only valid for the selected catchment, as similar potential simplification might have a different impact on the performance parameters for a different catchment. For this, six models are developed using different combination of the proposed input parameters. The performance parameters of the generated models are listed in Table 6-4.

Firstly, it is started with developing a model (model no.1) comprised of the available input parameters which are common with the 2nd layer abstraction model that are used in our simplification by *Villa-Vialaneix et al.* [65]. The cross-validated performance indicates low performance with R^2 of 0.413, MAE of 0.069, RMSE of 0.101 which is beyond the acceptable value of RMSE of 0.078 or lesser, and RRMSE of 72.1%. This validates inclusion of additional parameters. Therefore the variables related to field conditions such as *SM* and *CS* are included to develop model no.2. The performance increases with relative error, RRMSE, reducing to 63.9%. After this, parameters related to additional information for N application, *CumN* and *DNFS* are included in the previous input list to develop model no.3. The performance improves substantially with RRMSE value of 50.5% and R^2 of 0.703. The RMSE value (0.703) also falls under the acceptable value of 0.78. This validates the inclusion of additional parameters related to N in the *TON*-loss predictive model.

Furthermore, the N input parameters, *NS* and *NF*, are removed to develop model no.4 which provides interesting results though consistent with low correlation values as discussed in the sensitivity analysis. The performance of model no.4, developed with input parameters of *PPT*, *SM*, *CS*, *CumN*, and *DNFS*, drops slightly by just 1% relative error. Then *DY* is included to develop model no.5, for which the performance increases slightly by 1% relative error. Finally all the proposed input parameters (3rd level abstraction) are included for the *TON*-loss predictive model no.6, which results in the best performance with R^2 of 0.727, and RRMSE of 47.7%. This exercise validates that the proposed model parameters are the optimal set for best prediction performance.

Figure 6-5 illustrates the performance curve of the *TON*-loss predictive model by plotting the R^2 and RRMSE values of the different models. This shows how the performance is improving by adding the proposed parameters one by one thus reaching the optimal performance with the whole set. However, depending on the specific requirements of any particular model application, model accuracy might be traded for model simplicity by choosing models trained on fewer input parameters (as in models 3, 4 and 5).

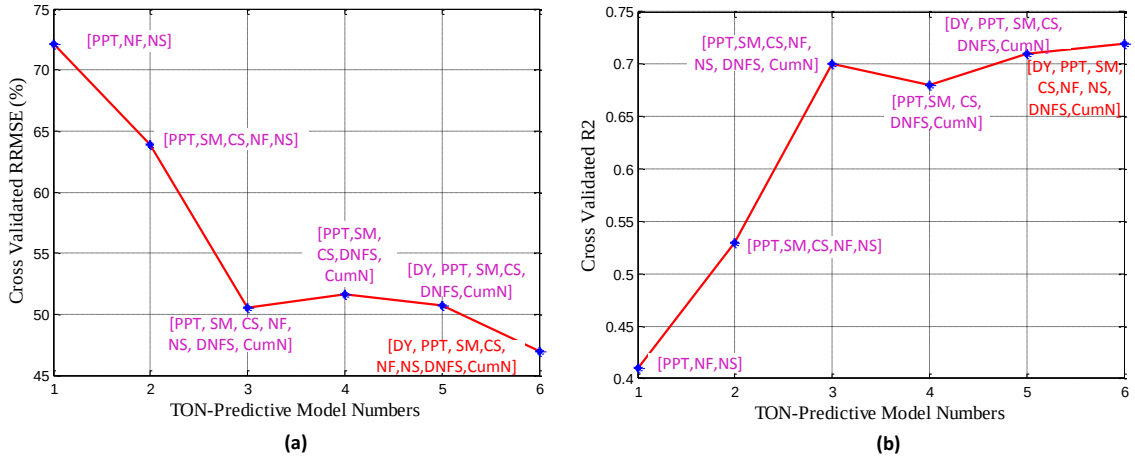


Figure 6-5: Comparison of (a) RRMSE and (b) R^2 values, for the *TON*-loss predictive models developed using various input parameter combinations

6.3.3 Comparative Assessment of *TON*-loss Predictive Model Using Various Learning Algorithms

Using the optimal input parameter set, consisting of *DY*, *PPT*, *CS*, *NF*, *NS*, *DNFS* and *CumN*, a *TON*-loss predictive model is developed based on various machine learning algorithms such as REPTree, MLR and MLP for comparison with M5 tree algorithm. REPTree is a fast decision tree learner which builds a regression/decision tree using information gain as the splitting criterion, and prunes it using reduced error pruning [155]. The models are developed using the Java-based data mining package, Weka [181] for ease of use. Table 6-4 lists the performance metrics of the generated models. It is apparent that the M5 tree based model has the best performance. This is explained by the model architecture of M5 tree, which has linear models in the pruned leaves allowing prediction for unseen events.

Table 6-5: Performance comparison of *TON* models using various learning algorithms

Algorithm used	**10-fold cross-validated performance metric**			
	R^2	**MAE ($Kgha^{-1}$)**	**RMSE ($Kgha^{-1}$)**	**RRMSE (%)**
M5 tree	0.88	0.053	0.068	47.4
REPTree	0.87	0.052	0.076	48.8
MLP - ANN	0.81	0.071	0.092	58.8
MLR	0.75	0.077	0.102	65.1

The model developed using REPTree, also a decision tree, is very close to the one developed with M5 tree with RRMSE of 48.8%. However, RMSE value of the REPTree (0.076) based *TON* model is just at the border of the acceptable value (0.078). The model developed using MLP, provides a good value for R^2, which is 0.81; however the RMSE value is unacceptable at 0.092. The RRMSE for this model is 58.8%. The *TON* model developed using MLR has the lowest performance with 65.1% RRMSE and an unacceptable RMSE value of 0.102. The reason for this low performance is attributed to the architecture of the MLR model. MLR attempts to fit a linear equation to observed data, which does not work well with non-linearly related input and output variables. Therefore it is concluded that the M5 tree algorithm is currently the best performing learning algorithm.

6.3.4 Uncertainty Analysis for *TON*-loss Predictive Model

Residuals are firstly plotted against day of the year to determine if any time dependency exists for the prediction error over summers and winters. This also provides a confidence interval for the predicted values.

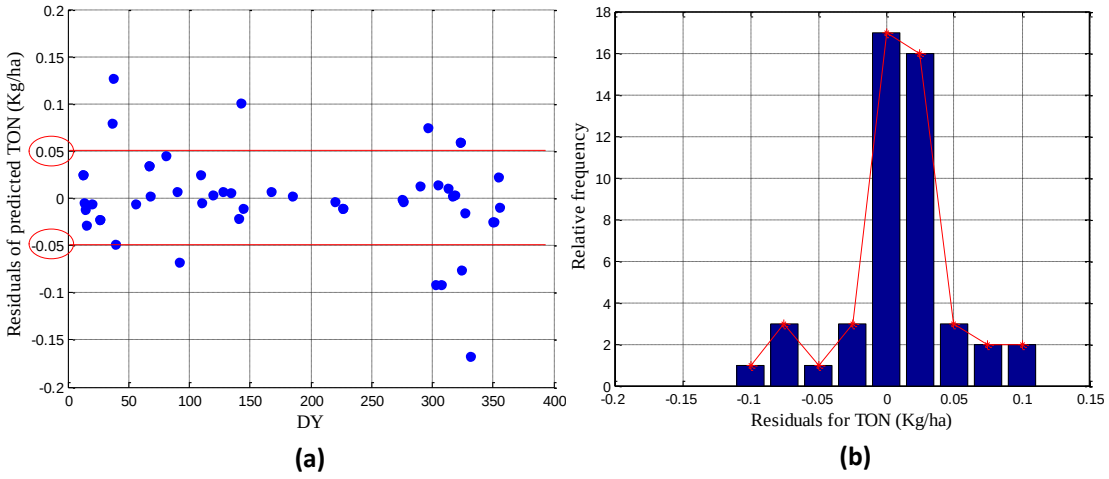


Figure 6-6: (a) Trend of residual error of predicted stream flow, (b) relative frequency of the residual error

Figure 6-6 (a) shows that 80% of the residuals for the predicted test values falls within $\pm$0.05 Kgha^{-1} error range. Prediction error of this scale does not seem significant for the reason that this does not yield substantial nitrate losses, which would amount to 0.85 Kg of nitrate for this 17 ha catchment. Furthermore, the minimum monthly application of N in this catchment was about 5 Kgha^{-1}, which means that about 85 Kg of nitrate was used on 17 ha of land. Hence, even for the

minimum application of N input, the daily error of 0.85 Kg is just 3% of the total application. Therefore, erroneous estimations of this scale do not impact decision making.

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 0.025 $Kgha^{-1}$ error (Figure 6-7 (b)). To explain this further a plot is made for the residuals percentile value with respect to day of the year, as shown in Figure 6-7 (a). This graph illustrates that 84% of the predicted values are between +20% and -40% residual errors. A histogram for this shows a normal distribution having maximum number of values with 0% residual error (Figure 6-7 (b)).

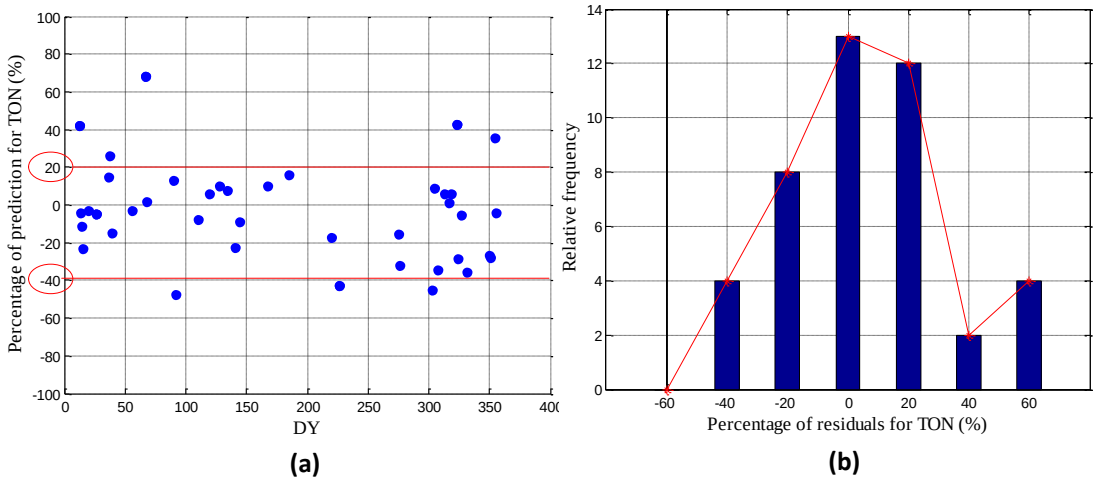


Figure 6-7: (a) Trend of percentile value of residual error versus, (b) relative frequency of residual error percentage

6.4 Summary

In this chapter, a nitrate loss (*TON*) predictive model is successfully evaluated for the proposed simplified parameters by employing an M5 tree learning approach. Simplified parameters related to climate, N input and field conditions were defined and used to avoid reliance on complex bio-geo-chemical parameters used in the existing nitrate loss predictive models. The model is assessed for its prediction accuracy in comparison to other popular data-driven methods in hydrological and nitrate loss modelling. The analysis was conducted on a nitrate loss dataset measured for a sub-catchment in Ireland. Results show that:

- The proposed *TON*-loss predictive model provides good performance on the real dataset in terms of the performance measures used. For the generated model, R^2 is 0.92, RMSE is 0.04, and RRMSE as 26%. 10-fold cross-validation results indicate R^2 as 0.72 and

RRMSE as 47.7%. 80% of the residuals for the predicted test values fall within $\pm$0.05 $Kgha^{-1}day^{-1}$ error range, which is quite minimal.

- The performance of the M5 tree based proposed model is better than the existing data-driven nitrate loss models generated for same training set size (150 samples). To achieve the same performance as the *TON*-loss predictive model, existing nitrate loss models require thousands of samples of complex parameters for training.
- Various simplifications in the model parameters are used to develop the models and hence evaluate their performance differences. The proposed parameter list results in the best performance, however, some subsets of this list resulted in very little (2% to 3%) performance difference. Hence it is concluded that for the considered catchment, the proposed parameters list is the most appropriate to attain optimal results. It may also be concluded that it may be appropriate to trade model complexity for model performance depending on the objectives of the specific modelling application.
- The *TON*-loss predictive model developed using M5 tree algorithm produces best results when compared with models developed using other commonly used learning algorithms such as MLR, MLP and REPTree. This is explained by the model architecture of M5 tree, which has linear models in the pruned leaves allowing prediction for unseen events.

The results presented in this chapter show great potential for enabling real time nutrient control and management applications within collaborative networked farm system, which already require simplified models due to resource constraints associated with (i) measuring complex parameters, and (ii) implementing complex bio-geo-chemical models on sensor nodes.

Chapter 7: Decision Support Model for Drainage Reuse

Chapter 3 presented the WQMCM framework, a water quality management framework for a networked catchment. Subsequently, chapter 4, 5 and 6 investigated development and validation of low complexity drainage dynamics (Q, t_l, & t_d) and nitrate (*TON*) predictive models. As highlighted in chapter 3, the predicted information regarding drainage and nutrients can be vital for implementing a proactive drainage management. The predicted drainage information can be beneficial for adjusting management strategy in a farm or in a stream network by adopting drainage reuse, disposal or treatment (section 3.2). This chapter demonstrates the use of predicted information for supporting decision process about drainage reuse in a farm, for which an example decision model is developed and discussed.

7.1 Functional Detail of a Farm Network

As already emphasized in chapter 2, reuse of agricultural drainage water within the farm system, before it ends up in streams and rivers, can have huge environmental and economic benefits. Currently there are various constraints associated with implementing a drainage mechanism due to the unavailability of predictive information about volumes and timings of drainage that will be delivered to the farms. These limitations are attributed to the inadequate available technological and management solution, which is further constrained by complex predictive models, and expensive sensing equipment. These constraints are addressed by the research contributions of this thesis (chapter 3, 4, 5, and 6). Now, to demonstrate how the predicted information about drainage can enable and support decision making regarding its reuse, some key functional blocks required in a farm network under the WQMCM framework are now proposed and discussed. These functional blocks include; economic evaluation of expected drainage and nutrients (*TON* in this case), adaptive sampling of the sensors in the farm, and a decision support model for classifying reusability of expected drainage and nutrients. Figure 7-1 illustrates these functional blocks.

7.1.1 Economic Evaluation:

The predicted volumes (*Q, & TON)* are used to evaluate the economic benefit of drainage reuse. For this, market value of fresh water and N fertilizer can be used as a standard for setting the price for the drainage water and *TON* losses (lower than the fresh water and N prices). A farmer may choose to reuse drainage for irrigation and fertigation, which will naturally be cost-effective due

to the lower market value of agricultural drainage and dissolved *TON* in comparison to fresh water and fertilizers. To evaluate the economic benefit of drainage reuse, a detailed case study for Australian price structure and land type is attached in appendix A. For the undertaken case-study, cost for drainage and *TON* is assumed to be 50% of the market prices for fresh water and N fertilizer. The minimum cost saving in case of re-usage of drainage and *TON* losses from 1 ha of a pasture land, lost at the lowest rate of 35% and 30% respectively, is $144.7. Whereas, the economic value of the total drainage and *TON* goes up to $2436.2, if the loss rate for drainage and *TON* are considered as maximum, at 60% and 50% respectively. The detailed case study is attached at appendix A.

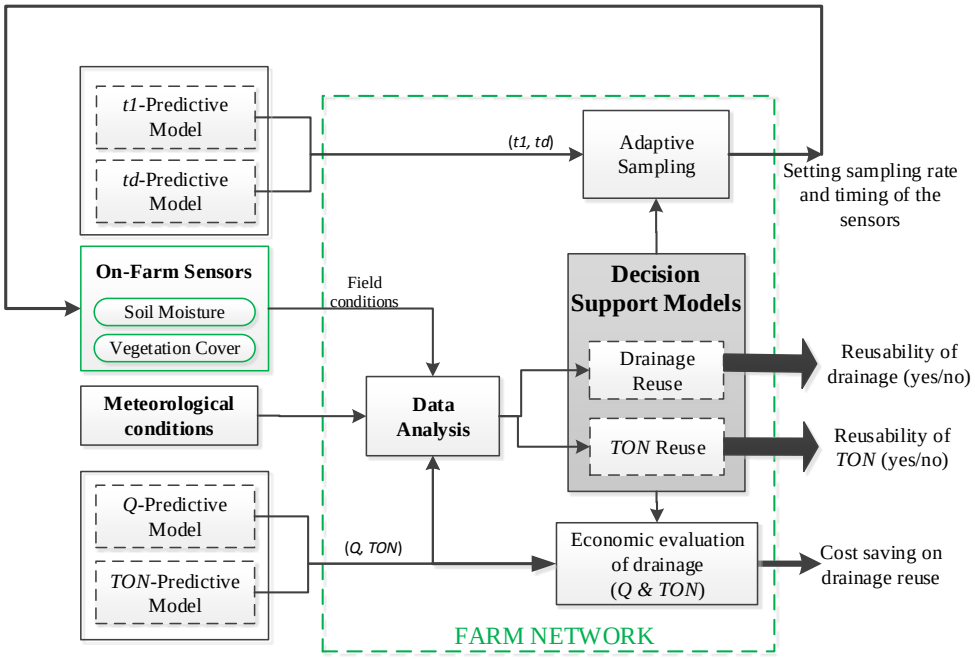


Figure 7-1: Block diagram of decision making in a farm network in response to drainage prediction from a drainage network

7.1.2 Adaptive Sampling

In the absence of low-power and miniature solid-state sensors, as discussed in chapter 2, currently power hungry chemi-bio sensors are used which present inherent limitations because of the number of samples that they can take [108]. In order to maximize the deployment lifetime, these sensors should adaptively sample at higher resolution only when an event occurs. As demonstrated in the case study attached at appendix B, with a single weekly or daily reading events can be missed or inaccurately captured. This has been highlighted in the literature [35]. On the other hand, continuously sensing every minute guarantees accurate event sensing but can

substantially reduce the battery lifetime for a Tyndall 90-FLT sensor board [127] (normally used in most of the WSN applications) to just 4.8 days. Sampling at an hourly rate increases the battery life to 288 days while maintaining 95% accurate measurements. This highlights the importance of adaptive sampling in maintaining high battery life while accurately sensing the events. Therefore, the predicted values for t_1, & t_d are used to enable adaptive sampling of sensors at the predicted times of the expected drainage (as shown in Figure 7-1), which helps prolong network life and measure events accurately.

7.1.3 Decision Support Model

Lastly, the main block of the farm network is the decision support model for advising farmers on the reusability of the drainage water and nutrients. Figure 7-1 illustrates the various inputs going into a decision support model, including sensed farm conditions, climatic data, and predicted *Q* and *TON* values. The development of an example decision model is discussed below.

7.2 Architecture of a Decision Support Model for Drainage Reuse

For a decision support model, the challenge lies in designing a model which takes into account relevant influencing parameters, predicted drainage values, and expert knowledge. This is shown as the dataset development stage in Figure 7-2. Expert knowledge is utilized to develop guidelines or rules for supporting decision about drainage and nutrients reuse, resulting in a dataset of rules for individual models. E.g., if rainfall is expected in next 2 days, and the predicted drainage level is low, then do not reuse the drainage water. The model complexity, in terms of influencing parameters, can substantially vary depending upon the requirements of the farmer or application. Influencing parameters can likely include local field conditions, climatic forecast, and predicted dynamics (*Q* and *TON*). The selection of influencing parameters, or model parameters, for each of *Q* and *TON* models are discussed later in the section. Additionally, the predicted *Q* and *TON* values are translated into well-defined levels, with respect to the required irrigation and fertilizer quantities. In this regard, simple percentage rules are used to interpret the values as ‘low’, ‘medium’ or ‘high’. These percentage rules are subjective and may vary according to a farmer’s requirements. Here, the percentages are selected keeping in view the maximum loss rates for drainage and *TON* which are 60% and 50% respectively. Assuming a 10% loss due to evaporation or absorption, and similar farm sizes and irrigation and fertilizer requirements, the received drainage losses (at 40% or more) are the maximum losses that can emanate from a 1 ha of a land.

For example, if the predicted value of the expected *Q* or *TON* from neighbouring farm is less than 10% of the required values for water (irrigation) or N (fertilization) on the farm respectively, then this predicted value is assigned a 'low' level. Similarly, other assignments are listed in Table 7-1. A 'medium' level is assigned if predicted values range between 10% and 40% of the required values, and a 'higher' value is assigned if the predicted values are greater than 40% of the required ones.

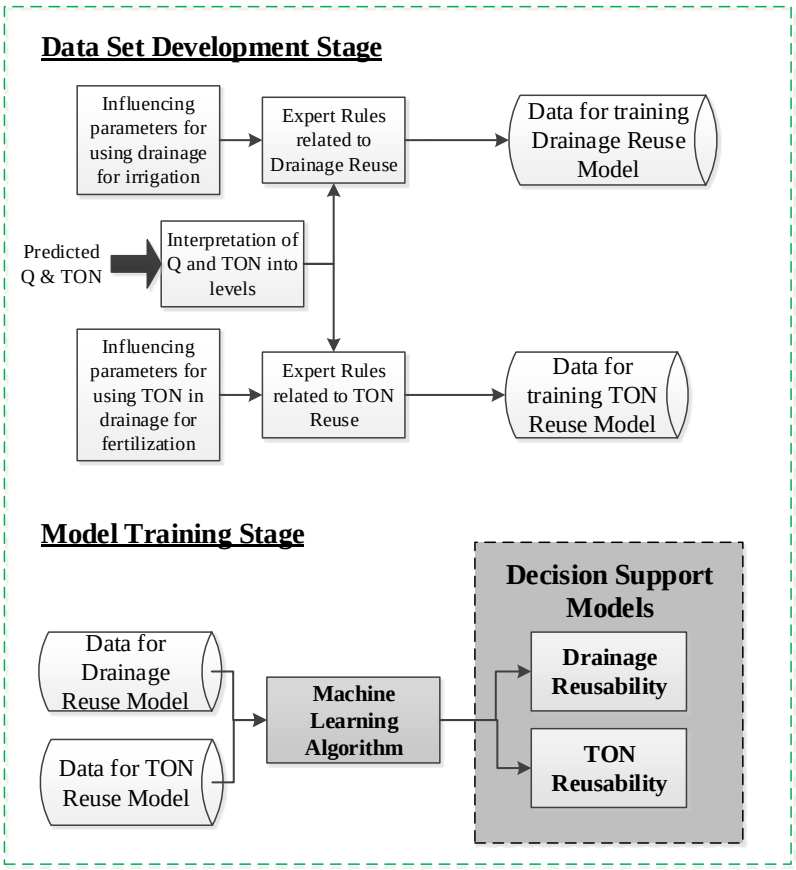


Figure 7-2: Data development and training stages for the development of decision support model

The dataset of resulting rules from the data development stage, is then used to train a machine learning algorithm to produce the decision support models for drainage water and *TON*. This is illustrated as model training stage in Figure 7-2. The trained decision models are then used to determine reusability of drainage water and *TON*. Figure 7-1 illustrates the functional overview of the decision support model, in which the trained models for drainage and *TON* reuse, classifies the reuse decision in affirmative or negative. Influencing parameters, such as meteorological forecast and sensed farm conditions, along with predicted *Q* and *TON* values are used as input

parameters for the decision models. In case of affirmative decision about reusability, a farm network may reuse the drainage water and N nutrients.

Table 7-1: Interpretation of quantified values for predicted *Q* and *TON* into classes

S.No.	Predicted Value	Level
1	If *Q*, or *TON* < 10% of required irrigation and fertilizer (N) quantity	Low
2	If *Q*, or *TON* >10% & <40% of required irrigation and fertilizer (N) quantity	Medium
3	If *Q*, or *TON* >40% of required irrigation and fertilizer (N) quantity	High

7.2.1 Model Parameters for the Decision Support Models

For model development, the fundamental part is the selection of appropriate input parameters. As mentioned in the previous section, key parameters required for decision making about drainage reuse can vary substantially depending upon the complexity of the application or specific requirements of a farm owner. Here, simple parameters and common knowledge is used to illustrate the concept.

For instance, before making a decision about the reuse of drainage water for irrigation, weekly forecast information about impending rainfall can be very useful. In case there is no imminent rainfall, the drainage water reuse is a likely beneficial option as drainage water is usually priced at a lower value compared to fresh irrigation water [182]. In case of a heavy rainfall forecast (exceeding irrigation requirements), a farmer would naturally prefer to rely on rainfall instead of either of fresh or drainage water for irrigation. However, in case of a light rainfall or when there are a few days in rainfall, a farmer may rely partly on drainage water to fulfil immediate irrigation requirements. This implies that information about the possibility of rainfall, its expected depth, and number of days to rainfall are all vital for the decision model. Hence, influencing parameters, for decision model of drainage reuse, consists of the following input parameters; predicted *Q* level, rainfall forecast (yes/no), rainfall depth forecast and, days to rainfall (as shown in Figure 7-3 (a)).

In case of factors which can likely influence a reusability decision about nutrients, simple model parameters are selected. It is reported in the literature [79], that during the early stage of seedling and germination, when crops are more susceptible to damage, farmers may avoid reusing drainage water. Although usually applicable to water reuse of marginal quality, it may still be included in the influencing parameters. For instance, farmers may choose to not use nutrients in drainage water on crops at an early crop stage, i.e. seedling and germination. Therefore, the model parameters for *TON* reuse model includes crop stage, predicted *Q* level, and predicted *TON* level, as shown in Figure 7-3 (b).

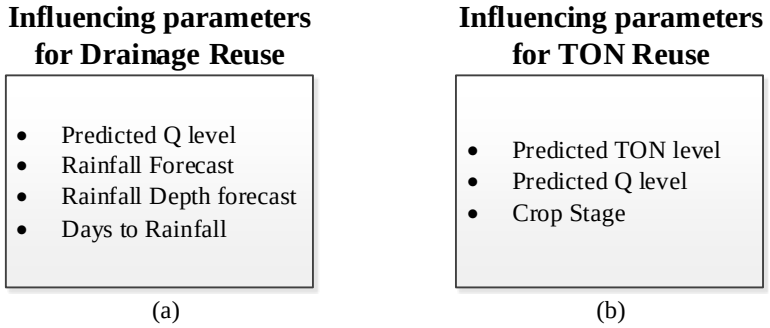


Figure 7-3: Influencing model parameters for; (a) drainage water reuse model, and (b) *TON* reuse model

7.3 Experimental Method

7.3.1 Data Set

Based on the proposed parameters for each decision model in the previous section, an example rule set is defined for classifying reusability decision about drainage and nutrient reuse. Based on the combination of values for various input parameters of each model, reusability is assigned a 'yes' or 'no'. A tree representation of the defined ruleset for *TON* reuse is illustrated in Figure 7-4. For example, an instance of a rule includes that in case of low *Q* level, a later crop stage, and medium *TON* level, *TON* reusability is in affirmative.

In an actual implementation, data for this model should be generated in consultation with a particular stakeholder's preferences for setting the reusability criteria. For example, one may choose strict criteria by wanting to reuse water only when higher levels of drainage and N are available. Hence, the rule set would be updated accordingly to train the model so that it prompts

with affirmative decision about reusability only when higher drainage fluxes are expected and negative otherwise.

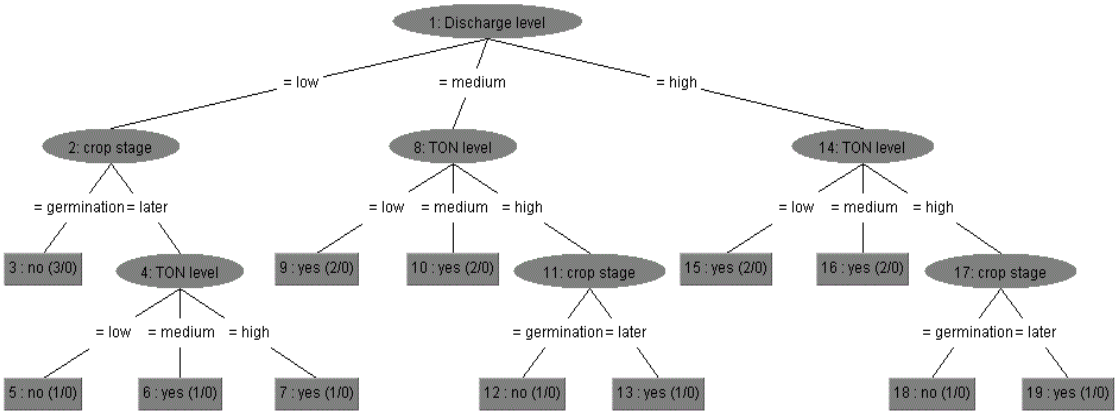


Figure 7-4: Defined ruleset for decision support model of *TON* reuse

7.3.2 Modelling Technique

For developing the decision support models, classification algorithms are used. Classification is a basic task in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of attributes [162]. Among various classification algorithms, two that are popular and simple to implement are the Naïve Bayes and C4.5 as discussed in section 2.5.2. Therefore, these algorithms are used to develop the decision classifier models for discharge and *TON* reuse. A Java-based data mining package, Weka [181], is used here for ease of use.

7.3.3 Performance Metric

The performance of the decision classifier model is measured using different parameters such as RMSE, percentage of correct and incorrect classification and, kappa coefficient value. For representing the correct and incorrect instances, a confusion matrix, also known as error matrix, is used [183]. As shown in Table 7-2, each row represents the instances in each class for the observed data, whereas each column represents instances in each class for the predicted data. In this table, ‘a’ and ‘d’ represents the number of instances that are correctly predicted or classified as class ‘Yes’ or ‘No’ respectively, while ‘b’ and ‘c’ indicates the number of instances that are incorrectly predicted as class ‘No’ and ‘Yes’ respectively. The percentage for correctly and incorrectly classified instances is calculated as;

$$Correct(\%) = \frac{(a+d)}{(a+b+c+d)} \tag{14}$$

$$Incorrect(\%) = \frac{(c+b)}{(a+b+c+d)} \tag{15}$$

Table 7-2: Confusion Matrix

REUSABILITY		
Observed Class	**Predicted Class**	
	Yes	**No**
Yes	a	b
No	c	d

Another performance metric used for classification models is Cohen's kappa coefficient (*k*) [184]. It is a measure of inter-rater agreement for classified instances. More specifically, it compares observed accuracy with an expected accuracy (random chance). It is considered to be a more robust measure than percent accuracy calculation, since it takes into account the agreement occurring by chance [185]. This performance metric is specifically useful for comparing multiple classifier models used for the same classification task.

The equation for kappa coefficient is;

$$k = \frac{p(o) - p(e)}{1 - p(e)} \tag{16}$$

Where, *p(o)* is the total agreement probability or simply put as the number of correctly classified instances in the confusion matrix. Furthermore *p(e)* is the hypothetical probability of chance agreement, which is explained as the accuracy that any random classifier would be expected to achieve based on the confusion matrix. This is more simply expressed as;

$$k = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy} \tag{17}$$

As mentioned above, *p(o)* or observed accuracy is calculated as;

$$p(o) = \frac{(a+d)}{(a+b+c+d)} \tag{18}$$

Whereas, *p(e)* or expected accuracy is computed as follows;

$$p(e) = \left(\frac{(a+b)(a+c)}{a+b+c+d} + \frac{(c+d)(b+d)}{a+b+c+d}\right) / (a+b+c+d) \tag{19}$$

In order to interpret *k* values, although there is not a standardized method, arbitrary guidelines have appeared in various studies. For example in some papers, *k* over 0.75 is rated as excellent, below 0.4 as poor and in-between values as fair to good [186, 187]. In simple terms, for comparing different classifiers, the better performing classifiers should then have the higher *k*.

7.4 Results and Discussion for the Decision Support Models

For the decision support models for drainage reuse, the rule-based dataset of the proposed influencing parameters, is used to train the C4.5 and Naïve Bayes classifiers. The performance results for the two classifiers are illustrated in Table 7-3. Using C4.5, the generated decision model correctly classified 80.95% of the data with error percentage of 19.04. The value of *k* has come out as 0.61 which is fair to good as per standards. The RMSE for this model is 0.38. On the other hand, using Naïve Bayes, the classifier model for drainage water correctly classified 90.47% samples with incorrect classifications of 9.52%. The value of *k* is 0.81 which is categorized as excellent, and RMSE is 0.29 which is also better than the previous model. Clearly, the performance of Naïve Bayes model indicates that it is a strong model with high prediction accuracy and is better when compared with the model performance of C4.5 for this application. Results are encouraging in the sense that even a simple rule set for model parameters of the decision model has successfully classified the training instances.

Table 7-3: Decision classifier model for reusability of drainage water

Classification Algorithm	**Performance**			
	RMSE	**Correct (%)**	**Incorrect (%)**	***k***
C4.5	0.38	80.95	19.04	0.61
Naïve Bayes	0.29	90.47	9.52	0.81

Similarly, for modeling *TON* reusability, the two machine learning algorithms are applied on the training data to generate the models for comparison. The model generated using C4.5 correctly classifies 83.33% of the samples with 16.6% incorrect classifications. The value of k is 0.57 which is just fair and RMSE is 0.35. In comparison, the model developed using Naïve Bayes correctly classifies 94.44% of the samples with 5.55% incorrect classifications. The value of *k* is 0.86 which is considered excellent, meaning the model can be described as a high performance predictive model. RMSE is 0.32 for this model. Similar to the decision model for drainage reusability, the performance of Naïve Bayes model is significantly better compared to the one generated using C4.5, as summarized in Table 7-4.

Table 7-4: Classifier model for reusability of *TON*

Classification Algorithm	**Performance**			
	RMSE	**Correct (%)**	**Incorrect (%)**	***k***
C4.5	0.35	83.33	16.66	0.57
Naïve Bayes	0.32	94.44	5.55	0.86

7.5 Summary

This chapter presents the functional blocks associated with implementing a management mechanism in a farm network with regard to using expected drainages from neighbouring farms or drainage bay for irrigation. These functional blocks include; economic evaluation of expected drainage and nutrients (*TON* in this case), a decision support model for classifying reusability of expected drainage and nutrients, and adaptive sampling of the farm.

This chapter has presented the development of a decision support model, which supports amendment of a farm's management strategy. The model development comprises three key elements; (1) exploration of the key model parameters that can influence a reusability decision; (2) outlining of rule set based on the selected model parameters and expert knowledge defining reusability of drainage and nutrients; and, (3) use of machine learning classification algorithm on the outlined ruleset, for which C4.5 and Naïve Bayes are utilized and compared.

The proposed model parameters for decision classifier model for drainage reuse consists of predicted *Q* level, rainfall forecast (yes/no), rainfall depth forecast and, days to rainfall. Whereas, the proposed model for decision classification for nutrient reuse includes stage of the crop along with predicted *Q* level and *TON* level as model parameters.

Using Naïve Bayes algorithm, the developed decision model for drainage reuse classifies the learned ruleset with 90.47% accuracy with 9.52% incorrect classifications. The kappa index for the model comes as 0.81. On the other hand, the developed decision model for *TON* reuse correctly classifies 94.44% of the samples with 5.55% inaccurate classifications. The kappa coefficient for the model is 086. Both the models have clearly high prediction accuracy. Furthermore, Naïve Bayes algorithm performed better than the C4.5 algorithm. These results are significant because they demonstrate that simple ruleset based on the proposed model parameters has the potential to generate a useful model which can support decision making. This chapter, though not a primary contribution of this thesis, is quite significant in outlining the usefulness of information sharing among networks in a farm, with regards to water quality control and drainage management with associated economic benefits.

Chapter 8: Conclusions and Future Work

8.1 Summary of Work

Globally, there is an increasing strain on existing, already scarce, fresh water resources for producing higher crop yields to cater to rising food demands. In order to avoid exacerbating the water crisis and to prevent food shortages, an advantageous strategy is the conservation and reuse of agricultural drainage (and the dissolved nutrients) before it ends up in the fresh water system. The availability of predictive drainage information is necessary and critical to enabling informed decision-making for drainage reuse or disposal. However, various limitations owing to available inadequate technological and management solutions, which is further constrained by complex predictive models, and expensive and power-hungry sensing equipment, are associated with enabling such a mechanism. This thesis has addressed the above limitations by providing a technological framework, which is based on collaborative infrastructure enabled using WSNs, and low-complexity predictive models. The proposed framework has been developed consisting of various modules, while the key contribution of this thesis is the development of low-complexity predictive modules which have been validated using measured catchment data. Below, a brief summary of aims and contributions of each chapter is outlined along with unaddressed points. Next section will present the future work based on these unaddressed areas.

Chapter 3 presented the WQMCM framework, consisting of individual networked farms, drainage bays and streams, to facilitate high resolution monitoring and to enable information sharing about drainage flow among neighbour networks. The shared information is processed for timely treatment, disposal or reuse of the drainage. The basic system architecture comprises modules to process shared information in order to enable individual networks to learn environment (neighbour linking model), predict the impact of neighbouring events in terms of drainage dynamics (Q, t_1, & t_d predictive models) and nutrient losses (*TON*-loss predictive model), and then adapt the local monitoring and management strategy. The overall functionality of the framework is explored in terms of stages of learning, training and testing. In the learning stage, the neighbour linking model is used to determine the correlation of events in various farm networks with the events received in the drainage bay by the drainage network. The model results in identifying linked and un-linked farm networks. The acquired training data, from the neighbour linking stage, is used to train the predictive models for drainage dynamics and nitrate losses. In the testing stage, neighbour event information is firstly interpreted using developed neighbour linking lists and then used to predict drainage dynamics in case of a linked neighbour. These

predicted values are transmitted by a drainage network to other farms and stream networks so that a decision is taken for their reuse, disposal or conservation.

Chapter 4 demonstrated development of low complexity models for Q, t_1, and t_d by reducing the number of parameters to less than 50% in comparison to the NRCS model. The input parameters for the Q-predictive model included daily precipitation, vegetation cover, and last-5-day-rainfall value, whereas the t_1 and t_d predictive models include an additional parameter of event duration. Evaluation of the predictive models, developed using M5 tree algorithm, demonstrates accuracy of 84-94% compared with the traditional NRCS curve number model. The Q and t_d model was tested to perform with around 94% and 92% accuracy respectively, even for a small training set of around 100 samples; however the t_1 predictive model required a minimum of 300 training samples to show reasonable performance with 84% accuracy. Furthermore, performance of these models is validated using MLR model, while it is observed that models developed using M5 tree algorithm demonstrated better performance.

Chapter 5 investigated validation of the Q-predictive model for the proposed simplified parameters by employing M5 tree algorithm. Because of the lack of relevant data for validating t_1 and t_d, these predictive models could not be validated, hence will need addressing in the future. The Q-predictive model was evaluated by training it on a year-long discharge dataset measured for a sub-catchment in Ireland. Evaluation of the Q-predictive model, developed using M5 tree algorithm, demonstrates RRMSE of 35.9%, with 10-fold cross-validated RRMSE of 55.7%. The model is assessed for its prediction accuracy in comparison to major data-driven hydrological models. 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. Due to the unavailability of soil moisture data, a proxy parameter, last-5-day-rainfall, is used as suggested in NRCS model. When real soil moisture readings become available to the model, performance should improve. This will be investigated in the future work.

Chapter 6 presented development and validation of a low-complexity *TON*-loss predictive model for the proposed model parameters. The proposed input parameters, consisting of 5 simple real time variables, are related to climate, N input and field conditions. These parameters were proposed with the aim to avoid reliance on complex bio-geo-chemical parameters used in the existing nitrate loss predictive models. Model validation is conducted on a nitrate loss dataset measured for a sub-catchment in Ireland, and by employing an M5 tree learning approach. The

proposed *TON*-loss predictive model provides good performance on the real dataset in terms of the performance measures used. For the generated model, R^2 is 0.92, RMSE is 0.04, and RRMSE as 26%. 10-fold cross-validation results indicate R^2 as 0.72 and RRMSE as 47.7%. The performance of the M5 tree based proposed model is better than the existing data-driven nitrate loss models generated for same training set size (150 samples). To get the same performance as the *TON*-loss predictive models, existing nitrate loss models require thousands of samples for training and complex parameters. The *TON*-loss predictive model developed using M5 tree algorithm is assessed for its prediction accuracy in comparison to other popular data-driven methods such as MLR, MLP and REPTree, and is found to perform better than these algorithms. The proposed *TON*-loss predictive model shows huge potential by using 5 simple model parameters and a year-long data set. Furthermore, the performance of the proposed model is better than the existing data-driven nitrate loss models generated for same training set size. To get the same performance as the *TON*-loss predictive model, existing nitrate loss models require thousands of samples for training and complex parameters. It is important to mention that out of the input parameters for training, rainfall, and *TON* losses were reported using sensors, whereas rest of the parameters were extrapolated from field information. If sensor networks are employed, actual soil moisture data (instead of last-5-day-rainfall), and autonomous readings for vegetation cover (crop stage) would be available. Additionally, if cheap in-situ sensors for nitrate existed (for which research is underway, section 2.2.3.2), measured soil nitrate content readings could be used in the model instead of a proxy parameter CumN (cumulative N applied). Nonetheless, the proposed *TON* model performs better than the existing models.

In the end, chapter 7 presents the functional blocks associated with implementing a management mechanism in a farm network with regards to the expected drainages from neighbouring farms or drainage bay. These functional blocks include; economic evaluation of expected drainage and nutrients (*TON* in this case), a decision support model for classifying reusability of expected drainage and nutrients, and adaptive sampling of the farm. This chapter, though not the primary contribution of this thesis, is quite significant in outlining the usefulness of information sharing among networks in a farm, with regards to water quality control and drainage management with associated economic benefits. Using Naïve Bayes algorithm, the developed decision model for drainage reuse classifies the learned ruleset with 90.47% accuracy, whereas for *TON* reuse the model correctly classifies 94.44% of the samples. These results are important because they demonstrate that simple ruleset based on the proposed model parameters has the potential to generate a useful model which can support decision making.

This thesis has highlighted the importance of an autonomous technological framework for enabling water quality management through agricultural drainage reuse among farms. Through a novel combination of real time field information made available by using WSNs, and decision tree learning algorithm, low-complexity predictive models have been investigated and proposed. The predictive models for drainage discharge and *TON* losses have been validated through year-long measured data from a real catchment in Ireland. The obtained results demonstrate that comparable performance in the case of *Q* and even better performance in the case of *TON* can be achieved by using the proposed low-complexity models for a much smaller training set. The results are promising for enabling autonomous and real time water and nutrient management by employing simple models due to the constraints associated with measuring complex parameters and implementing complex models. The undertaken research has resulted in international conference and journal publications as mentioned in section 1.3 [188-196].

8.2 Future Work

The research undertaken in this thesis has successfully addressed the research aims outlined in section 1.1. However, this is a comprehensive application area with complex underlying details which could not possibly be covered entirely in this thesis. Therefore, to take this research forward, numerous extensions are possible. This section highlights some of the areas for future work in terms of specific and general extensions of this research.

8.2.1 Specific Extensions

- **Model Validation Using Sensor Data**

  For *Q* and *TON* predictive models, data validation is currently done using the available dataset from an agricultural grassland consisting of rainfall, streamflow and nitrate loss values. Due to the constraints associated with either free access or availability to long term field data (soil moisture, and soil N content), measured using WSNs, proxy parameters (last-5-day-rainfall, Cumulative N applied) were used instead for the validation. In future, it is planned to validate the model by acquiring sensed field data for those parameters and evaluate model performance. Such a scenario is most likely going to improve the model performance e.g. as we argued in section 4.1 that last-5-day-rainfall index is not a true representative of antecedent soil moisture conditions. Similarly, t_l and t_d predictive models need to be validated by acquiring temporal values for individual

events. This will involve superimposed event detection and separation to identify individual events.

- **Decision Support Model**

  In this thesis, out of the three functional blocks mentioned in section 7.1 for the management mechanism, the implementation of the decision support model for reusability was considered with a simplified set of data. A generic illustration of the economic benefits that can likely be achieved by drainage reuse and the role of adaptive sampling for longer lifetime of sensors is represented in appendix A and B. For validating economic benefits that are actually possible through drainage and nutrient reuse and actual lifetime of sensors that can be achieved with adaptive event sampling, application specific experimental data is required.

  As mentioned in section 3.2, the predicted drainage information can be quite beneficial for adjustment of management strategy regarding drainage reuse, disposal or treatment. In this thesis, we considered the management strategy for drainage reuse in farms. In the similar manner, decision models can be developed for supporting management decisions, taken by receiving stream network, about disposal and treatment of drainage and nutrients in case it cannot be reused by farms. Furthermore, alert models can be formulated for scenarios where disposal and treatment mechanisms are not in place. Such a model can make use of additional information like pollutants and pesticides and alert the adjoining stream system or river system for taking appropriate action such as diverting the drainage flow from entering into sensitive ecosystems.

- **Network Deployment and Data Acquisition**

  The most vital part of the future works is the design and implementation of this framework using a networked catchment capable of operating autonomously for drainage reuse. This will involve network platform design, establishing direct interconnection between networks, and data acquisition for neighbour linking model. Once the neighbours are linked, training data is acquired for developing the predictive models, which are later deployed on the sensor nodes of a drainage network.

  For the network platform, a low cost, low power and reliable system is required. An example of such a system is a low cost wireless soil moisture sensing system developed by University of Georgia [197]. A sensor node contains watermark soil moisture sensors, temperature sensors, radio frequency transmitter and a circuit board. A 20 node system

consists of $3500. They plan to add sensors (pH, nutrients) to the node in the future. In addition various off-the-shelf sensor boards are commercially available (as reviewed in chapter 2). They differ from each other in terms of e.g. sensor modalities, transmitter range, robustness, and network life that they provide. From the available platform solutions, a sensor board can be selected for deployment initially and, if needed, could be redesigned to cater to application specific needs or add more sensors. For enabling opportunistic direct interconnection between co-located WSNs, another PhD thesis is underway, and so far the simulations have shown promising results [198, 199]. The next step is the implementation of the protocol on actual networks.

- **Low complexity model for estimating phosphorous and other nutrients losses**
  As mentioned in the introduction chapter, phosphorous and nitrate are the major contaminants responsible for non-point source water quality degradation. In this thesis, due to the availability of real catchment data for only *TON* losses and time constraints, nitrate loss modelling was investigated. To further this research, modelling phosphorous losses from agricultural land, based on real field conditions and fewer parameters, is quite significant for quantifying expected phosphorous losses in drainage.

8.2.2 General Extensions

- **Sensor Network Platform Design**
  Current applications of WSN for water quality monitoring are based on the use of off the shelf components which often fail to address diverse range of environmental and degradation problems that can coalesce at catchment scales. Existing devices are not made to address the specific requirements of a particular application. Various limitations have been solved in the existing literature by the use of off the shelf components [50] such as the coverage problem [38], signal attenuation [43], configuration and operational difficulties [200], and absence of a network model for a broad environment. Existing literature showed that the nodes lasted from a few days to only a few months, therefore effective design changes like wake up and power control need to be considered to prolong their lifetime to years.

  Furthermore, although separate sensor boards [48, 201] can be integrated with nodes to enable a multi sensing system with various sensor types, improvements can be made to take this further. The plug and play system constrains sensors to be compatible with that system, whereas a more reconfigurable interface may well extend the choice of sensors

at the expense of specifying to the node what it is. Therefore, there is a requirement for a reconfigurable sensor interface for a sensor node as this would allow a more rapid deployment for a specific application and thereby reducing the costs of the WSN.

- **In-situ Sensors for Nutrient Measurement**

  As discussed in section 2.2.3.1, simple and low-rate sensors, such as those typically used for monitoring temperature, humidity and wind are readily available at low-cost, this is not the case for other sensor modalities [39]. For example, the high cost of some sensors (e.g. ammonium) is a major barrier for dense deployment of WSNs [49]. Furthermore, suitable in-situ phosphate and nitrate sensors for terrestrial deployment are not available. The available sensors for nitrate monitoring are based on ultraviolet absorption and flow cell analysers which are very expensive for dense deployment [104]. Both electromechanical [105] and optical sensing modes [106] relevant to nitrogen and phosphorous species are still an active research area [39]. Furthermore, research on the development of miniature, reliable and inexpensive water quality solid-state sensors using metal oxide sensing electrodes is still underway by CSIRO, CMSE and SSN TCP laboratories [107].

Appendix A Case Study 1 – Economic benefit of drainage reuse

In order to evaluate costs and benefits of drainage reuse, a general approach can be taken into account based on the market value set for drainage water and nutrients. Market value of water and fertilizers can be used as a yardstick for setting the price for valuing the drainage water and nutrients found within [182] either by mutual settlement between farmers or by a governing body. In this analysis, the price rate for drainage water is set at 50% of the market cost for fresh water and N as suggested by hydrologists in the area. An analysis of the possible economic benefits that can be achieved in general due to drainage and fertilizer reuse is given below.

Irrigation costs are rising worldwide, due to a combination of increasing scarcity, rising energy prices, and increasing awareness of climate change and its potential implications on rainfall and the availability of surface water resources. Cost for irrigation water varies in various regions of different countries depending upon the availability and access to fresh water resources. A detailed report for water tariffs in Europe is available at [202]. In United Kingdom, irrigation is supplementary because of abundance of water resources owing to high rainfall, for example in Scotland and Northern Ireland, farmers can take water from adjacent rivers simply by applying for a paid license and some annual charges. Prices vary from € 0.008 m^{-3} to € 0.021 m^{-3} depending upon the location, water quality, and the season in which the abstraction is made [203]. Whereas in the United States, given the variety of water rights allocations and contractual arrangements for irrigation, cost of irrigation water ranges between $ 0.02 m^{-3} and $ 0.1 m^{-3} (€ 0.02 to 0.12 m^{-3}) [204].

To evaluate the economic benefits of reuse, the price structure of Australia is considered as an example. This is because of the prevalent water scarcity and pollution issues which provides a perfect example for adopting the reuse mechanism and framework. Australia is an arid continent with relatively little water. Ecosystems of the Great Barrier Reef, world heritage area, are widely considered to be at risk from agricultural runoff carrying nutrients (specifically nitrogen) and sediments [205]. Water is managed in Australia through a system of entitlements. Entitlements represent a permanent claim on some amount of water from a given source and are variously classified according to reliability. In 2008, entitlement prices ranged from $0.13 m^{-3} to $1.67 m^{-3} for various entitlement categories based on reliability of water availability [206]. The fertilizer price was reported as $1.24 Kg^{-1} of N in one of the recent studies conducted in Australia [207].

Water and fertilizer application rates are considered for two of the major agricultural sectors of Australia – dairy pastures and rice. It is estimated that 70% of the annual water usage of the agricultural industry is used to irrigate dairy pastures in Australia [42] with an average application rate of 4.4 mega liters per ha ($MLha^{-1}$) . On the other hand, rice is considered the thirstiest crop, with an application of 14.1 $MLha^{-1}$, more than three times the national average across all crops and pasture [208]. For dairy pastures, average nitrogen fertilizer use is 122 $KgNha^{-1}$, whereas the usage goes up to 374 $KgNha^{-1}$ [209]. On the other hand, for a rice crop, the fertilizer application rate is about 140 $KgNha^{-1}$ [99].

As already discussed, 35% to 60% of irrigation water is lost due to runoff in various methods adopted in surface irrigation [7]. Similarly, 30%-50% of the applied N is lost due to the coupled impact of over fertilization and irrigation runoff [14]. Based on the average nutrient and drainage loss rates and the discussed price structure, the economic value is calculated for the drainage water and N released from 1 ha of a land. The evaluated cost can also be termed as economic savings, which can be achieved by reusing those fluxes in another farm.

Table 8-1: Economic savings possible on irrigation water

Crop	**Irrigation Rate**	**% of Drainage Loss**	**Drainage Loss (m^3ha^{-1})**	**Cost of drainage water per ha ($) (at 50% of Fresh water cost)**	
				Fresh water cost= $1.67 m^{-3}	**Fresh water cost= $0.13** m^{-3}
Pasture	4.4 $MLha^{-1}$ ($4400m^3\ ha^{-1}$)	**35%**	1540	1285.9	100.1
		60%	2640	2204.4	171.6
Rice	14.1 $MLha^{-1}$ ($14100m^3\ ha^{-1}$)	**35%**	4935	4120.7	320.7
		60%	8460	7064.1	549.9

Table 8-1 presents the economic value of the expected drainage water, per ha of a land at given application, loss rates and price rates, for the two land types - pasture and rice crop. As mentioned, 4.4 $MLha^{-1}$ (4400 m^3ha^{-1}) of water is used for irrigation in a pasture land. This results in a minimum drainage loss of 1540 m^3ha^{-1} considering a 35% loss and a maximum of 2640 m^3ha^{-1} considering a 60% maximum loss. On the other hand, for a rice crop, 14.1 $MLha^{-1}$ (14100 m^3ha^{-1}) of water is applied for irrigation. At 35% loss rate, drainage loss is estimated at 4935 $m^3\ ha^{-1}$, whereas this goes up to 8460 m^3ha^{-1} of drainage loss at 60% loss rate. Using the standard cost of fresh irrigation water and N fertilizer in Australia (which ranges between $0.13 and $1.67 m^{-3} for water and $1.24 for N) as a standard, economic value of the drainage and N losses is calculated

at 50% of the standard price. In case of drainage coming from 1 ha of a pasture land, minimum economic value is $100 and a maximum value of $2204. When the drainage originating from 1 ha of a rice crop is reused completely, the cost savings can range between $320.7 and $7064. Hence, this implies that if this drainage is reused in another farm which requires same amount of water as the received drainage water, 50% savings are made. Total cost savings on the irrigation water for entire crop for the farm depends upon its requirements and availability of drainage water.

A similar analysis is done for estimating the economic value of available N nutrients in the drainage water for the two land types, as shown in Table 8-2. For a pasture land, minimum fertilizer loss for an average application of 120 $Kgha^{-1}$ is estimated at 36 $Kgha^{-1}$ for a 30% loss rate, whereas the maximum value goes up to 60 $Kgha^{-1}$ at 50% loss rate. For calculating the cost of N available in drainage, it is assumed that this should be rated at half the price of the new fertilizers. Economic value of N in drainage water, originating from 1 ha of a pasture land which had an average N application, is estimated to range between $44.64 and $74.4 at 30% and 50% loss rate respectively. Whereas, at a maximum application rate of 374 $KgNha^{-1}$ in a pasture land, the minimum loss is estimated at 112 $KgNha^{-1}$ and maximum value goes up to 187 $KgNha^{-1}$. The economic value of these fluxes range between $138.88 and $231.8, for 1 ha of a pasture land. For a rice crop, an application of 140 $Kgha^{-1}$ of N results in losses that range between 42 $Kgha^{-1}$ and 70 $Kgha^{-1}$ at 30% and 50% loss rate respectively. The economic value of N losses contained in drainage, leaving a rice crop, is estimated between $52 and $86 per ha.

Table 8-2: Economic savings possible on N fertilizers

Crop	**Fertilizer Application**	**% of Fertilizer Loss**	**Fertilizer loss (Kg ha^{-1})**	**Cost per Kg of N fertilizer in drainage ($) (at 50% of fresh N fertilizer cost $1.24/Kg)**
Pasture	120 $Kgha^{-1}$ (average application rate)	30	36	44.64
		50	60	74.4
	374 $Kgha^{-1}$ (maximum application)	30	112	138.88
		50	187	231.88
Rice	140 $Kgha^{-1}$	30	42	52.08
		50	70	86.8

Based on the results discussed above, for individual economic values under multiple loss percentages and application rates, the total economic value on drainage and N reuse for a worst and best case scenario with application rates is now presented, as illustrated in Table 8-3.

Table 8-3: Combined economic savings possible on irrigation water and N fertilizers

Crop Type	Case	Case Specification		drainage cost ($)	Fertilizer cost ($)	Total cost of drainage & N per ha ($)
		Drainage	N Fertilizer			
Pasture	Worst Case	35% loss rate, 4400 m^3ha^{-1} application, cost at 50% of \$0.13 m^{-3}	30% loss rate, 120 $KgNha^{-1}$ application, cost at 50% of \$1.24 m^{-3}	100.1	44.6	144.7
	Best Case	60% loss rate, 4400 m^3ha^{-1} application, cost at 50% of \$1.67 m^{-3}	50% loss rate, 374 Kg Nha^{-1} application, cost at 50% of \$1.24 m^{-3}	2204.4	231.8	2436.2
Rice	Worst Case	35% loss rate, 14100 m^3ha^{-1} application, cost at 50% of \$0.13 m^{-3}	30% loss rate, 140 $KgNha^{-1}$ application rate, cost at 50% of \$1.24 m^{-3}	320.7	52.1	372.8
	Best Case	60% loss rate, 14100 m^3ha^{-1} application, cost at 50% of \$1.67 m^{-3}	50% loss rate, 140 $KgNha^{-1}$ application rate, cost at 50% of \$1.24 m^{-3}	7064.1	86.8	7150.9

For a pasture land, firstly a minimum loss rate is considered for drainage and N at 30% and 35% respectively, and average application rates for water and N. Applying a cost rate of 50% of the market cost for fresh water and N, total cost savings for this case come at \$144.7 per ha of a land. For estimating cost savings for a best case scenario, maximum loss rates of 60% for drainage and 50% for N are considered. Furthermore, average available application rate for drainage is used, whereas, for N, a maximum application rate of 374 $KgNha^{-1}$ is used. This provides us with the maximum savings of \$2436.2 per ha, which can be achieved if total drainage and N losses are reused. These numbers would be much higher as farm sizes range from 5 ha to 50 ha whereas only per ha rates are being considered here. A similar analysis is done for the rice crop. For a worst case scenario with minimum loss rate for an average application of water and N on 1 ha

land, $372.8 can be saved if such discharges are reused. Furthermore, cost savings of $7150 are possible in a best case scenario with maximum loss rates for an average application of water and N in 1 ha land with rice crop.

A.1.1 Summary

Reusing drainage water and nutrients emanating from one farm in another farm, before they enter the water system, can have huge environmental as well as economic benefits. To evaluate the economic benefits of reuse, the price structure of Australia is considered as an example. This is because of established water scarcity and pollution issues which exist there and hence provide a perfect example which can benefit from adopting the reuse mechanism and framework. Water and fertilizer application rates are considered for two of the major agricultural sectors of Australia – dairy pastures and rice. Using the standard cost of fresh irrigation water and N fertilizer in Australia as a standard, economic value of the drainage and N losses is calculated at 50% of the standard price. The total economic value of drainage and N originating from 1 ha of a pasture land, at a minimum loss rate for drainage at 30% and N at 35%, and average application rates for irrigation and N fertilizer, is estimated as $144.7. Whereas, at maximum loss rates (60% for drainage and 50% for N), and maximum application rate for N, the economic value of drainage and N come as $2436.2 per ha of a pasture land. This implies that minimum $144.7 and maximum $2436.2 savings can be achieved if total drainage and N losses originating from 1 ha of a land are reused. A similar analysis is done for rice crop. For a worst case scenario with minimum loss rate for an average application of water and N on 1 ha land, $372.8 can be saved if such discharges are reused. Furthermore, cost savings of $7150 are possible in a best case scenario with maximum loss rates for an average application of water and N in 1 ha land with rice crop. These numbers would be much higher as farm sizes range from 5 ha to 50 ha.

Appendix B Case Study 2 – Longer network life and better information availability

As the previous section validates the economic benefits provided by a collaborative and integrated water quality monitoring management mechanism through drainage reuse. Next step is to demonstrate the importance of high resolution adaptive sampling as the event occurs, and how information sharing can facilitate adaptive sampling. As it has been mentioned in chapter 2, that in the absence of low-power and miniature solid-state sensors, currently used power hungry chemi-bio sensors present inherent limitations because of the number of samples that they can take [108]. In order to maximize the deployment lifetime, these sensors should adaptively sample at higher resolution only when an event occurs. Low resolution constant sampling (say once every hour) has the potential to miscalculate/misrepresent a short-lived intense event of less than a couple of hours. Figure B-1 illustrates a pollutant discharge event being captured using sensors at various sampling rates. Long sampling intervals (per 2 hours, or 6 hours) tend to miss peak discharge values and misrepresent the discharge value.

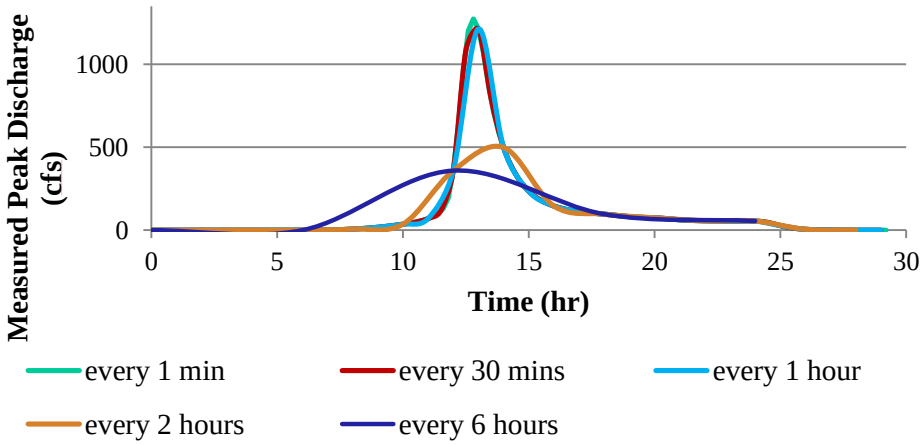


Figure B-1: Runoff discharge measured using various sampling rates

Similarly, high resolution continuous sampling depletes sensor board's battery in a few days. To demonstrate this, a case is considered where a sensor board, Tyndall 90-FLT [127], samples at various rates to capture an event in a day. Assuming that N is the number of measurements per day, the device wakes up for 50 seconds each time, and the 90-FLT meter consumes zero energy when it is off. The device can last for 34560/50N days [48]. By taking continuous readings every minute (1440 readings /day), the sensor board's battery only lasts 4.8 days. Battery life at various sampling rates is shown in Figure B-2 against the values of measured discharge precision.

Sampling every half an hour or an hour provides 95% accurate measurements, however the battery lasts for 144 to 288 days respectively. To aim for a longer lifetime, adaptive sampling only in case of an event occurrence is critical.

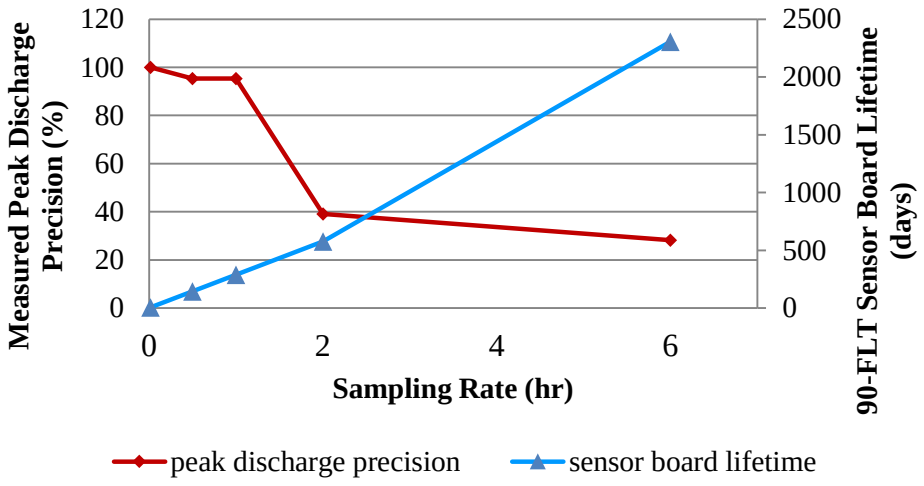


Figure B-2: Variation in sensed peak discharge values and sensor board lifetime with sampling rate

Appendix C *NRCS Curve Number Table*

Table 8-4: Curve numbers for agricultural land [128]

Land use or cover	Treatment or practice	Hydrological condition	Hydrological soil group			
			A	B	C	D
Fallow	Straight row	Poor	77	86	91	94
Row crops	Straight row	Poor	72	81	88	91
	Straight row	Good	67	78	85	89
	Contoured	Poor	70	79	81	88
	Contoured	Good	65	75	82	86
	Contoured/terraced	Poor	66	74	80	82
	Contoured/terraced	Good	62	71	78	81
Small grain	Straight row	Poor	65	76	84	88
	Straight row	Good	63	75	83	87
	Contoured	Poor	63	74	82	85
	Contoured	Good	61	73	81	84
	Contoured/terraced	Poor	61	72	79	82
	Contoured/terraced	Good	59	70	78	81
Close-seeded legumes or rotational meadow	Straight row	Poor	66	77	85	89
	Straight row	Good	58	72	81	85
	Contoured	Poor	64	75	83	85
	Contoured	Good	55	69	78	83
	Contoured/terraced	Poor	63	73	80	83
	Contoured/terraced	Good	51	67	76	80
Pasture range		Poor	68	79	86	89
		Fair	49	69	79	84
		Good	39	61	74	80
	Contoured	Poor	47	67	81	88
	Contoured	Fair	25	59	75	83
	Contoured	Good	6	35	70	79
Meadow (permanent)		Good	30	58	71	78
Woodlands (farm woodlots)		Poor	45	66	77	83
		Fair	36	60	73	79
		Good	25	55	70	77
Farmsteads			59	74	82	86
Roads, dirt			72	82	87	89
Roads, hard-surface			74	84	90	92

Bibliography

1. Bernstein, S., "Freshwater and human population: a global perspective," *Human Population and Freshwater - US cases and international perspective*, p. 0-9, 2002.

2. Rosegrant, M.W., X. Cai, and S.A. Cline, "World water and food to 2025: Dealing with scarcity." International Food Policy Research Institute. 348, 2002.

3. Tanji, K.K. and N.C. Kielen, "Agricultural drainage water management in arid and semi-arid areas." vol. 61, Food and Agriculture Organization of the United Nations, 2002.

4. Kulshreshtha, S.N., "A global outlook for water resources to the year 2025," *Water Resources Management*, vol. 12(3), p. 167-184, 1998.

5. Moss, B., "Water pollution by agriculture," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363(1491), p. 659-666, 2008.

6. EPA, "National Water Quality Inventory: Report to Congress ; 2004 Reporting Cycle," United States Environmental Protection Agency, (EPA 841-R-08-001), 2009.

7. Eisenhauer, D.E., "Irrigation Efficiency and Uniformity, and Crop Water Use Efficiency," The Board of Regents of the University of Nebraska - Lincoln Extension., (EC 732), 2011.

8. Tindula, G., M. Orang, and R. Snyder, "Survey of Irrigation Methods in California in 2010," *Journal of Irrigation and Drainage Engineering*, vol. 139(3), p. 233-238, 2013.

9. Defra, "Protecting our Water, Soil and Air - A Code of Good Agricultural Practice for Farmers, Growers and Land Managers." Department for Environment, Food & Rural Affairs. 579, 2011.

10. Defra. "Nitrate vulnerable zones in England," Available from: http://archive.defra.gov.uk/environment/quality/water/waterquality/diffuse/nitrate/nvz2008.htm, 2010.

11. Mueller, D.K., D.R. Helsel, and M.A. Kidd, "Nutrients in the nation's waters: too much of a good thing?," US Geological Survey, vol. 1136, 1996.

12. Vanlauwe, B., et al., "Combined application of organic matter and fertilizer." Sustaining Soil Fertility in South Africa. Soil Science Society of America Inc., 2001.

13. McKergow, L.A., et al., "Before and after riparian management: sediment and nutrient exports from a small agricultural catchment, Western Australia," *Journal of Hydrology*, vol. 270(3), p. 253-272, 2003.

14. Liu, X., et al., "Nitrogen dynamics and budgets in a winter wheat-maize cropping system in the North China Plain," *Field Crops Research*, vol. 83(2), p. 111-124, 2003.

15. Ball, P.R. and J.C. Ryden, "Nitrogen relationships in intensively managed temperate grasslands." Biological Processes and Soil Fertility, ed. J. Tinsley and J.F. Darbyshire. vol. 11: Springer Netherlands. 23-33, 1984.

16. Smith, V.H. and D.W. Schindler, "Eutrophication Science: Where Do We Go From Here?," *Trends in ecology & evolution*, vol. 24(4), p. 201-207, 2009.

17. Quan, W. and L. Yan, "Effects of Agricultural Non-point Source Pollution on Eutrophica tion of Water Body and Its Control Measure " *Acta Ecologica Sinica*, vol. 22(3), p. 291-299, 2002.

18. Parry, R., "Agricultural Phosphorus and Water Quality: A U.S. Environmental Protection Agency Perspective," *Environmental Quality*, vol. 27(2), p. 258-261, 1998.

19. UNEP, "Best Management Practices for Agricultural Non-Point Sources of Pollution," Caribbean Environment Programme, United Nations Environment Programme (UNEP), vol. 41, 1998.

20. Kaika, M., "The Water Framework Directive: a new directive for a changing social, political and economic European framework," *European Planning Studies*, vol. 11(3), p. 299-316, 2003.

21. Directive, C., "Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources," *Official Journal L*, vol. 375(31), p. 12, 1991.

22. Rhoades, J., "Intercepting, isolating and reusing drainage waters for irrigation to conserve water and protect water quality," *Agricultural Water Management*, vol. 16(1), p. 37-52, 1989.

23. Pereira, L.S., T. Oweis, and A. Zairi, "Irrigation management under water scarcity," *Agricultural Water Management*, vol. 57(3), p. 175-206, 2002.

24. Smith, V.H., G.D. Tilman, and J.C. Nekola, "Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems," *Environmental pollution*, vol. 100(1), p. 179-196, 1999.

25. Adelman, D.D., "Simulation of irrigation reuse system nitrate losses and potential corn yield reductions," *Environmental Science & Policy*, vol. 3(4), p. 213-217, 2000.

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

27. Grewal, H.S., B. Maheshwari, and S.E. Parks, "Water and nutrient use efficiency of a low-cost hydroponic greenhouse for a cucumber crop: An Australian case study," *Agricultural Water Management*, vol. 98(5), p. 841-846, 2011.

28. Bastiaanssen, W.G.M., et al., "Twenty-five years modeling irrigated and drained soils: State of the art," *Agricultural Water Management*, vol. 92(3), p. 111-125, 2007.

29. Carr, G., R.B. Potter, and S. Nortcliff, "Water reuse for irrigation in Jordan: Perceptions of water quality among farmers," *Agricultural Water Management*, vol. 98(5), p. 847-854, 2011.

30. Oster, J. and S. Grattan, "Drainage water reuse," *Irrigation and Drainage Systems*, vol. 16(4), p. 297-310, 2002.

31. Liu, Y., et al., "Linking science with environmental decision making: Experiences from an integrated modeling approach to supporting sustainable water resources management," *Environmental Modelling & Software*, vol. 23(7), p. 846-858, 2008.

32. Glasgow, H.B., et al., "Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies," *Journal of Experimental Marine Biology and Ecology*, vol. 300(1), p. 409-448, 2004.

33. Lemos, S.G., et al., "Soil Calcium and pH Monitoring Sensor System," *Journal of Agricultural and Food Chemistry*, vol. 55(12), p. 4658-4663, 2007.

34. Basha, E. and D. Rus, "Design of early warning flood detection systems for developing countries," in *Proc. of Information and Communication Technologies and Development Conference*, IEEE, p. 1-10, 2007.

35. Lloyd, C., et al., "Assessing the effects of sampling design on water quality status classification," *EGU General Assembly Conference Abstracts*, vol. 15, p. 13032, 2013.

36. Wang, N., N. Zhang, and M. Wang, "Wireless sensors in agriculture and food industry—Recent development and future perspective," *Computers and Electronics in Agriculture*, vol. 50(1), p. 1-14, 2006.

37. Ruiz-Garcia, L., et al., "A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends," *Sensors*, vol. 9(6), p. 4728-4750, 2009.

38. Regan, F., et al., "A demonstration of wireless sensing for long term monitoring of water quality," in *Proc. of 34th Local Computer Networks*, IEEE, p. 819-825, 2009.

39. Rundel, P.W., et al., "Environmental sensor networks in ecological research," *New Phytologist*, vol. 182(3), p. 589-607, 2009.

40. Morais, R., et al., "A ZigBee multi-powered wireless acquisition device for remote sensing applications in precision viticulture," *Computers and Electronics in Agriculture*, vol. 62(2), p. 94-106, 2008.

41. Beckwith, R., D. Teibel, and P. Bowen, "Report from the field: results from an agricultural wireless sensor network," in *Proc. of IEEE Local Computer Networks*, IEEE, p. 471-478, 2004.

42. McCulloch, J., et al., "Wireless sensor network deployment for water use efficiency in irrigation," in *Proc. of 3rd Workshop on Real-World Wireless Sensor Networks*, Glasgow, United kingdom, Association for Computing Machinery, p. 46-50, 2008.

43. Vellidis, G., et al., "A real-time wireless smart sensor array for scheduling irrigation," *Computers and Electronics in Agriculture*, vol. 61(1), p. 44-50, 2008.

44. Cugnasca, C., et al., "Ad Hoc Wireless Sensor Networks applied to animal welfare research," in *Proc. of 8th International Symposium on Livestock Environment* American Society of Agricultural and Biological Engineers (ASABE), 2008.

45. Darr, M. and L. Zhao, "A wireless data acquisition system for monitoring temperature variations in swine barns," in *Proc. of Agriculture and Biosystems Engineering Conference* vol. 288, 2008.

46. O'Flynn, B., et al., "Experiences and recommendations in deploying a real-time, water quality monitoring system," *Journal of Measurement Science and Technology*, vol. 21, p. 12, 2010.

47. Seders, L.A., et al., "Lakenet: an integrated sensor network for environmental sensing in lakes," *Environmental Engineering Science*, vol. 24(2), p. 183-191, 2007.

48. Zennaro, M., et al., "On the Design of a Water Quality Wireless Sensor Network (WQWSN): An Application to Water Quality Monitoring in Malawi," in *Proc. of International Conference on Parallel Processing Workshop*, IEEE, p. 330-336, 2009.

49. Ramanathan, N., et al., "Designing wireless sensor networks as a shared resource for sustainable development," in *Proc. of International Conference on Information and Communication Technologies and Development*, IEEE, p. 256-265, 2006.

50. Le Dinh, T., et al., "Design and deployment of a remote robust sensor network: Experiences from an outdoor water quality monitoring network," in *Proc. of 32nd Conference on Local Computer Networks*, IEEE, p. 799-806, 2007.

51. Kim, Y., J.D. Jabro, and R.G. Evans, "Wireless lysimeters for real-time online soil water monitoring," *Irrigation Science*, vol. 29(5), p. 423-430, 2011.

52. Bogena, H., et al., "Potential of wireless sensor networks for measuring soil water content variability," *Vadose Zone Journal*, vol. 9(4), p. 1002-1013, 2010.

53. Kido, M.H., et al., "Integration of Wireless Sensor Networks into Cyberinfrastructure for Monitoring Hawaiian "Mountain-to-Sea" Environments," *Environmental Management*, vol. 42(4), p. 658-666, 2008.

54. Kotamäki, N., et al., "Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in Southern Finland: Evaluation from a data user's perspective," *Sensors*, vol. 9(4), p. 2862-2883, 2009.

55. Cayan, D., et al., "The Wireless Watershed at the Santa Margarita Ecological Reserve," *Southwest Hydrology*, vol. 2(5), p. 18-19, 2003.

56. Walker, W.R., "Guidelines for designing and evaluating surface irrigation systems," FAO (Food and Agriculture Organization of the United Nations), 1989.

57. Basha, E.A., S. Ravela, and D. Rus, "Model-based monitoring for early warning flood detection," in *Proc. of 6th ACM Conference onEmbedded network sensor systems* ACM, p. 295-308, 2008.

58. Galelli, S. and A. Castelletti, "Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling," *Hydrology & Earth System Sciences Discussions*, vol. 10(2), 2013.

59. Castelletti, A., et al., "Tree-based reinforcement learning for optimal water reservoir operation," *Water Resources Research*, vol. 46(9), 2010.

60. Wilby, R., R. Abrahart, and C. Dawson, "Detection of conceptual model rainfall—runoff processes inside an artificial neural network," *Hydrological Sciences Journal*, vol. 48(2), p. 163-181, 2003.

61. Rasouli, K., W.W. Hsieh, and A.J. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," *Journal of Hydrology*, vol. 414–415(0), p. 284-293, 2012.

62. Solomatine, D. and A. Ostfeld, "Data-driven modelling: some past experiences and new approaches," *Journal of hydroinformatics*, vol. 10(1), p. 3-22, 2008.

63. Dawson, C.W. and R. Wilby, "An artificial neural network approach to rainfall-runoff modelling," *Hydrological Sciences Journal*, vol. 43(1), p. 47-66, 1998.

64. Minns, A. and M. Hall, "Artificial neural networks as rainfall-runoff models," *Hydrological Sciences Journal*, vol. 41(3), p. 399-417, 1996.

65. Villa-Vialaneix, N., et al., "A comparison of eight metamodeling techniques for the simulation of N2O fluxes and N leaching from corn crops," *Environmental Modelling & Software*, vol. 34(0), p. 51-66, 2012.

66. Fortin, J., et al., "Comparison of Machine Learning Regression Methods to Simulate NO 3 Flux in Soil Solution under Potato Crops," *Journal of Applied Mathematics*, vol. 5, p. 832-841, 2014.

67. Piñeros Garcet, J.D., et al., "Metamodelling: Theory, concepts and application to nitrate leaching modelling," *Ecological modelling*, vol. 193(3–4), p. 629-644, 2006.

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

69. Quinlan, J.R., "Learning with continuous classes," in *Proc. of 5th Australian joint Conference on Artificial Intelligence*, Singapore, vol. 92, p. 343-348, 1992.

70. Solomatine, D.P. and K.N. Dulal, "Model trees as an alternative to neural networks in rainfall—runoff modelling," *Hydrological Sciences Journal*, vol. 48(3), p. 399-411, 2003.

71. Corzo, G., et al., "Modular data-driven hydrologic models with incorporated knowledge: neural networks and model trees," in *Proc. of International congress IAHR, ASCE, Italy*, 2007.

72. Solomatine, D.P. and M.B. Siek, "Modular learning models in forecasting natural phenomena," *Neural Networks*, vol. 19(2), p. 215-224, 2006.

73. Solomatine, D.P. and Y. Xue, "M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China," *Journal of Hydrologic Engineering*, vol. 9(6), p. 491-501, 2004.

74. Bhattacharya, B., R. Price, and D. Solomatine, "Data-driven modelling in the context of sediment transport," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 30(4), p. 297-302, 2005.

75. Oehler, F., J.C. Rutherford, and G. Coco, "The use of machine learning algorithms to design a generalized simplified denitrification model," *Biogeosciences*, vol. 7(10), p. 3311-3332, 2010.

76. Carpenter, S.R., et al., "Nonpoint pollution of surface waters with phosphorus and nitrogen," *Ecological applications*, vol. 8(3), p. 559-568, 1998.

77. Rivers, M., K. Smettem, and P. Davies, "Estimating future scenarios for farm-watershed nutrient fluxes using dynamic simulation modelling–Can on-farm BMPs really do the job at the watershed scale?," 2010.

78. Rivers, M., et al., "Estimating future scenarios for farm–watershed nutrient fluxes using dynamic simulation modelling," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 36(9), p. 420-423, 2011.

79. Madramootoo, C.A., W.R. Johnston, and L.S. Willardson, "Management of agricultural drainage water quality." vol. 13, Food & Agriculture Organisation, 1997.

80. Grattan, S.R., et al., "Evaluation of salt-tolerant forages for sequential water reuse systems: I. Biomass production," *Agricultural Water Management*, vol. 70(2), p. 109-120, 2004.

81. Okano, K., Y. Sakamoto, and S. Watanabe, "Reuse of drainage water for the production of high quality fruits in single-truss tomato grown in a closed hydroponic system," in *Proc. of XXV International Horticultural Congress, Part 1: Culture Techniques with Special Emphasis on Environmental Implications*, p. 277-288, 1998.

82. Estrin, D., et al., "Instrumenting the world with wireless sensor networks," in *Proc. of Acoustics, Speech, and Signal Processing*, IEEE, vol. 4, p. 2033-2036 2001.

83. Jafer, E., et al., "Design of miniaturized wireless sensor mote and actuator for building monitoring and control," in *Proc. of 17th International Conference n Telecommunications*, IEEE, p. 887-892, 2010.

84. Harmon, T., et al., "An Overview of CENS Contaminant Transport Observation and Management Research," University of California, Los Angeles - Center for Embedded Network Sensing, 2009.

85. Gilbert, R.M., "A multi-method approach for the characterization of urban stream quality and algal dynamics - Thesis," University of California, Los Angeles. 2010

86. Park, Y. and T. Harmon, "Soil Moisture, Salinity, and Nitrate Control for Soil and Groundwater Protection in Support of Wireless Sensor Networks and Optimal Irrigation Strategy," in *Proc. of American GeoScience Union (AGU) Fall Meeting Abstracts*, vol. 1, p. 1514, 2007.

87. Ratâko, A., et al., "Micro- and Mini-nitrate Sensors for Monitoring of Soils, Groundwater and Aquatic Systems, UC Los Angeles: Center for Embedded Network Sensing, http://www.escholarship.org/uc/item/7td3g73w." 2009

88. He, D., et al., "A Water-Quality Dynamic Monitoring System Based on Web-Server-Embedded Technology for Aquaculture," *Computer and Computing Technologies in Agriculture IV*, p. 725-731, 2011.

89. Zhu, X., et al., "A remote wireless system for water quality online monitoring in intensive fish culture," *Computers and Electronics in Agriculture*, vol. 71, Supplement 1(0), p. S3-S9, 2010.

90. Zektser, I.S. and E. Lorne, "Groundwater resources of the world: and their use," *IHP Series on groundwater*, (6), 2004.

91. Gallagher, D.L., et al., "Ground water discharge of agricultural pesticides and nutrients to estuarine surface water," *Ground Water Monitoring & Remediation*, vol. 16(1), p. 118-129, 1996.

92. Valiela, I., et al., "Transport of groundwater-borne nutrients from watersheds and their effects on coastal waters," *Biodegradation*, vol. 10(3), p. 177-197, 1999.

93. Davis, G., W.W. Casady, and R.E. Massey, "Precision agriculture: An introduction," *Extension publications*, University of Missouri, 1998.

94. Langendoen, K., A. Baggio, and O. Visser, "Murphy loves potatoes: Experiences from a pilot sensor network deployment in precision agriculture," in *Proc. of 20th International Parallel & Distributed Processing Symposium*, IEEE, p. 155, 2006.

95. Liang, C.J.M. and A. Terzis, "Koala: Ultra-low power data retrieval in wireless sensor networks," in *Proc. of International Conference on Information Processing in Sensor Networks*, IEEE Computer Society, p. 421-432, 2008.

96. Burrell, J., T. Brooke, and R. Beckwith, "Vineyard computing: Sensor networks in agricultural production," *Pervasive Computing, IEEE*, vol. 3(1), p. 38-45, 2004.

97. Butler, Z., et al., ""Virtual fences for controlling cows"," in *Proc. of Robotics and Automation, 2004.* , IEEE, vol. 5, p. 4429-4436 Vol. 5, 2004.

98. Nadimi, E., et al., "ZigBee-based wireless sensor networks for monitoring animal presence and pasture time in a strip of new grass," *Computers and Electronics in Agriculture*, vol. 61(2), p. 79-87, 2008.

99. USGS. "Irrigation & Water Use," Available from: http://water.usgs.gov/edu/wuir.html, 2014.

100. Chartres, C.J. and S. Varma, "Out of water: from abundance to scarcity and how to solve the world's water problems." International Water Management Institute, 2011.

101. Kim, Y., R.G. Evans, and W.M. Iversen, "Remote sensing and control of an irrigation system using a distributed wireless sensor network," *Instrumentation and Measurement, IEEE Transactions on*, vol. 57(7), p. 1379-1387, 2008.

102. Papoutsa, C., et al., "Smart monitoring of water quality in Asprokremmos Dam in Paphos, Cyprus using satellite remote sensing and wireless sensor platform," *Remote Sensing*, p. 78310Q, 2010.

103. Harmon, T.C., et al., "High-resolution river hydraulic and water quality characterization using rapidly deployable networked infomechanical systems (nims rd)," *Environmental Engineering Science*, vol. 24(2), p. 151-159, 2007.

104. Robertson, C., G. Smith, and M. Tamburri, "Performance Demonstration Statement PMEL Map CO2/Battelle Seaology pCO2 Monitoring System," Alliance for Coastal technologies, [UMCES] CBL 10-092, 2009.

105. Bobacka, J., A. Ivaska, and A. Lewenstam, "Potentiometric ion sensors," *Chemical Reviews*, vol. 108(2), p. 329-351, 2008.

106. McDonagh, C., C.S. Burke, and B.D. MacCraith, "Optical chemical sensors," *Chemical Reviews*, vol. 108(2), p. 400-422, 2008.

107. Zhuiykov, S., "Solid-state sensors monitoring parameters of water quality for the next generation of wireless sensor networks," *Sensors and Actuators B: Chemical*, vol. 161(1), p. 1-20, 2012.

108. O'Connor, E., et al., "A neural network approach to smarter sensor networks for water quality monitoring," *Sensors*, vol. 12(4), p. 4605-4632, 2012.

109. IHE, "Data Driven Modelling," Institute for water education , UNESCO.

110. CHRS. "Hydrological-Predictions - On-going Activities," Available from: http://chrs.web.uci.edu/research/hydrologic_predictions/activities07.html, 2004.

111. Pechlivandis, I., et al., "Catchment scale hydrological modelling: a review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications.," *Global NEST Journal*, vol. Vol 13(No 3), p. pp 193-214, 2011.

112. Beven, K., et al., "Topmodel," *Computer models of watershed hydrology.*, p. 627-668, 1995.

113. Walker, W.R., "SIRMOD III Surface Irrigation Simulation, Evaluation and Design," *Guide and technical documentation. Dept of Biological and Irrigation Engineering, Utah St Univ, Logan, UT, USA*, 2003.

114. Wool, T.A., et al., "Water quality analysis simulation program (WASP)," *User's Manual, Version*, vol. 6, 2006.

115. Reed, S., et al., "Overall distributed model intercomparison project results," *Journal of Hydrology*, vol. 298(1), p. 27-60, 2004.

116. Georgakakos, K.P., "Analytical results for operational flash flood guidance," *Journal of Hydrology*, vol. 317(1), p. 81-103, 2006.

117. Huber, W.C., et al., "Storm water management model, version 4." US Environmental Protection Agency, Environmental Research Laboratory, 1988.

118. Hoehler, F.K., "Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity," *Journal of Clinical Epidemiology*, vol. 53(5), p. 499-503, 2000.

119. Skaggs, R., "Evaluation of Drainage- Water Table Control Systems Using a Water Management Model," in *Proc. of 3rd Drainage Symposium*, Chicago, p. 61-88, 1977.

120. El-Sadek, A., J. Feyen, and J. Berlamont, "Comparison of models for computing drainage discharge," *Journal of irrigation and drainage engineering*, vol. 127(6), p. 363-369, 2001.

121. Skaggs, R., M. Youssef, and G. Chescheir, "DRAINMOD: model use, calibration, and validation," *Transactions of the ASABE*, vol. 55(4), p. 1509-1522, 2012.

122. Hawkins, R.H., A.T. Hjelmfelt Jr, and A.W. Zevenbergen, "Runoff probability, storm depth, and curve numbers," *Journal of Irrigation and Drainage Engineering*, vol. 111(4), p. 330-340, 1985.

123. Hawkins, R., "Asymptotic Determination of Runoff Curve Numbers from Data," *Journal of Irrigation and Drainage Engineering*, vol. 119(2), p. 334-345, 1993.

124. Schneider, L. and R. McCuen, "Statistical Guidelines for Curve Number Generation," *Journal of Irrigation and Drainage Engineering*, vol. 131(3), p. 282-290, 2005.

125. Kowalik, T. and A. Walega, "Estimation of CN Parameter for Small Agricultural Watersheds Using Asymptotic Functions," *Water*, vol. 7(3), p. 939-955, 2015.

126. Hawkins, R.H., A.T. Hjelmfelt Jr, and A.W. Zevenbergen, "Runoff probability, storm depth, and curve numbers," *Journal of Irrigation and Drainage Engineering*, vol. 111(4), 1985.

127. USDA, N., "SCS Runoff Equation," *Engineering Hydrology Training Series*, National Employee Development Centre, Natural Resource Conservation Service, US Department of Agrcilture, 1999.

128. USDA, S., "Urban hydrology for small watersheds," *Technical release*, US Deptartent of Agriculture, Soil Conservation Service, Engineering Division, vol. 55, p. 2-6, 1986.

129. Shaw, E.M., et al., "Hydrology in practice." vol. 4, Taylor & Francis US, 2010.

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

131. Woodward, D.E., et al., "Runoff curve number method: examination of the initial abstraction ratio," in *Proc. of World Water and Environmental Resources Congress*, ASCE, 2004.

132. Shi, Z.-H., et al., "Research on the SCS-CN initial abstraction ratio using rainfall-runoff event analysis in the Three Gorges Area, China," *CATENA*, vol. 77(1), p. 1-7, 2009.

133. Jiang, R., "Investigation of Runoff Curve Number Initial Abstraction Ratio," in *Renewable Natural Resources*Univeristy of Arizona. 2001

134. Lees, M., "Data-based mechanistic modelling and forecasting of hydrological systems," *Journal of Hydroinformatics*, vol. 2, p. 15-34, 2000.

135. Basha, E.A., S. Ravela, and D. Rus, "Model-based monitoring for early warning flood detection," in *Proc. of Proceedings of the 6th ACM conference on Embedded network sensor systems*, ACM, p. 295-308, 2008.

136. Radcliffe, D. and M. Cabrera, "Modeling Phosphorus in the Environment." CRC Press, Boca Raton, USA, 2006.

137. Harbaugh, A.W., et al., "MODFLOW-2000, the US Geological Survey modular ground-water model: User guide to modularization concepts and the ground-water flow process." US Geological Survey Reston, VA, USA, 2000.

138. Refshaard, J., B. Storm, and V. Singh, "MIKE SHE," *Computer models of watershed hydrology.*, p. 809-846, 1995.

139. Liu, S., et al., "Simulation of nitrous oxide and nitric oxide emissions from tropical primary forests in the Costa Rican Atlantic Zone," *Environmental Modelling & Software*, vol. 15(8), p. 727-743, 2000.

140. Johnsson, H., et al., "Simulated nitrogen dynamics and losses in a layered agricultural soil," *Agriculture, Ecosystems & Environment*, vol. 18(4), p. 333-356, 1987.

141. Li, C., et al., "Modeling nitrate leaching with a biogeochemical model modified based on observations in a row-crop field in Iowa," *Ecological Modelling*, vol. 196(1–2), p. 116-130, 2006.

142. Li, C., et al., "Modeling nitrate leaching with a biogeochemical model modified based on observations in a row-crop field in Iowa," *Ecological modelling*, vol. 196(1), p. 116-130, 2006.

143. Li, C., S. Frolking, and T.A. Frolking, "A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity," *Journal of Geophysical Research: Atmospheres (1984–2012)*, vol. 97(D9), p. 9759-9776, 1992.

144. Zhang, Y., et al., "An integrated model of soil, hydrology, and vegetation for carbon dynamics in wetland ecosystems," *Global Biogeochemical Cycles*, vol. 16(4), p. 9-1-9-17, 2002.

145. Brevé, M., et al., "DRAINMOD-N, a nitrogen model for artificially drained soils," *Transactions of the ASAE*, vol. 40(4), p. 1067-1075, 1997.

146. Brevé, M.A., et al., "Using the DRAINMOD-N model to study effects of drainage system design and management on crop productivity, profitability and NO3–N losses in drainage water," *Agricultural Water Management*, vol. 35(3), p. 227-243, 1998.

147. Northcott, W., et al., "Application of DRAINMOD-N to fields with irregular drainage systems," *Transactions of the ASAE*, vol. 44(2), p. 241-249, 2001.

148. Ding, X., et al., "Development and test of the Export Coefficient Model in the Upper Reach of the Yangtze River," *Journal of Hydrology*, vol. 383(3–4), p. 233-244, 2010.

149. Rodda, H., et al., "Management model for predicting nitrate leaching from grassland catchments in the United Kingdom: 1. Model development," *Hydrological Sciences Journal*, vol. 40(4), p. 433-451, 1995.

150. Johnes, P.J., "Evaluation and management of the impact of land use change on the nitrogen and phosphorus load delivered to surface waters: the export coefficient modelling approach," *Journal of Hydrology*, vol. 183(3), p. 323-349, 1996.

151. Neville, S.D., et al., "Nutrient balance case studies of agricultural activities in south west Western Australia," in *Proc. of 7th International River Symposium, Brisbane, Australia, Aug 31st–3rd Sept*, 2004.

152. Mitchell, T.M., "Machine learning and data mining," *Communications of the ACM*, vol. 42(11), p. 30-36, 1999.

153. Balagatabi, Z.N., "Comparison of Decision Tree and Naïve Bayes Methods in Classification of Researcher's Cognitive Styles in Academic Environment," *Journal of Advances in Computer Research*, vol. 3(2), p. 23-34, 2012.

154. Solomatine, D. and Y. Xue, "M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China," *Journal of Hydrologic Engineering*, vol. 9(6), p. 491-501, 2004.

155. Zhao, Y. and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Advances in Space Research*, vol. 41(12), p. 1955-1959, 2008.

156. Breiman, L., "Bagging predictors," *Machine learning*, vol. 24(2), p. 123-140, 1996.

157. Dietterich, T., "Overfitting and undercomputing in machine learning," *ACM computing surveys (CSUR)*, vol. 27(3), p. 326-327, 1995.

158. Jekabsons, G. "M5PrimeLab: M5' regression tree and model tree toolbox for Matlab," Available from: http://www.cs.rtu.lv/jekabsons/Files/M5PrimeLab.pdf, 2010.

159. Aiken, L.S., S.G. West, and S.C. Pitts, "Multiple linear regression," *Handbook of psychology*, 2003.

160. Lacey, M. "Multiple Linear Regression," Available from: http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm, 1998.

161. Tranmer, M. and M. Elliot, "Multiple linear regression," *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 2008.

162. Friedman, N., D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Machine Learning*, vol. 29(2-3), p. 131-163, 1997.

163. Athanasiadis, I.N., K.D. Karatzas, and P.A. Mitkas, "Classification techniques for air quality forecasting," in *Proc. of Proceeding of the 5th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, Riva del Garda, Italy*, 2006.

164. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques." Emerging Artificial Intelligence Applications in Computer Engineering. IOS Publisher, 2007.

165. Karim, M. and R.M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *Journal of Software Engineering and Applications*, vol. 6, p. 196, 2013.

166. Karlık, B. and E. Öztoprak, "Personalized Cancer Treatment by Using Naïve Bayes Classifier," *International Journal of Machine Learning and Computing*, vol. 2(3), p. 339-344, 2012.

167. Quinlan, J.R., "C4. 5: programs for machine learning." vol. 2, 1993.

168. Davis, T. " SCS Unit Hydrograph Convolution : Hydrograph Generation and Analysis Tool," Available from: http://www.mathworks.co.uk/matlabcentral/fileexchange/10420-scs-unit-hydrograph-convolution/content/hydrograph/hydrographui.m.

169. Vellidis, G., et al., "NDVI response of cotton to nitrogen application rates in Georgia," *Precision Agriculture*, p. 359, 2011.

170. Davis, R. "Detecting Time Duration of Rainfall: A Controlling Factor of Flash Flood Intensity," Available from: http://www.erh.noaa.gov/pbz/ambr1.htm, 1996.

171. Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. of IJCAI*, vol. 14, p. 1137-1145, 1995.

172. Singh, J., et al., "Hydrological modeling of the iroquois river watershed using HSPF and SWAT1," *JAWRA Journal of the American Water Resources Association*, vol. 41(2), p. 343-360, 2005.

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

174. Keily, G. "Phosphorus, Nitrogen and Suspended Sediment loss from Soil to Water from Agricultural Grassland," Secure Archive For Environmental Research Data, Available from: http://erc.epa.ie/safer/resource?id=ad1f3acf-5035-102a-90c6-0593d266866d, 2003.

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

176. Computing, W.N., "Model-Independent Parameter Estimation," *User Manual:*, 2004.

177. Natrella, M., "NIST/SEMATECH e-handbook of statistical methods," 2010.

178. Udawatta, R.P., et al., "Nitrogen losses in runoff from three adjacent agricultural watersheds with claypan soils," *Agriculture, Ecosystems & Environment*, vol. 117(1), p. 39-48, 2006.

179. Allan, D., D. Erickson, and J. Fay, "The influence of catchment land use on stream integrity across multiple spatial scales," *Freshwater Biology*, vol. 37(1), p. 149-161, 1997.

180. Singh, J., et al., "Hydrological modeling of the iroquois river watershed using HSPF and SWAT1," *Journal of the American Water Resources Asso.*, vol. 41(2), p. 343-360, 2005.

181. Hall, M., et al., "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11(1), p. 10-18, 2009.

182. Intizar Hussain, et al., "Wastewater Use in Agriculture - Review of Impacts and Methodological Issues in Valuing Impacts," International Water Management Institute, (ISBN 92-9090-472-0).

183. Stehman, S.V., "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62(1), p. 77-89, 1997.

184. Carletta, J., "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22(2), p. 249-254, 1996.

185. Vieira, S.M., U. Kaymak, and J.M. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," in *Proc. of International Conference on Fuzzy Systems*, IEEE, p. 1-8, 2010.

186. Altman, D.G., "Practical statistics for medical research." Chapman & Hall 1990.

187. Lantz, C.A. and E. Nebenzahl, "Behavior and interpretation of the κ statistic: Resolution of the two paradoxes," *Journal of clinical epidemiology*, vol. 49(4), p. 431-434, 1996.

188. Zia, H., N. Harris, and G. Merrett, "Collaborative Catchment-Scale Water Quality Management using Integrated Wireless Sensor Networks," in *Proc. of EGU General Assembly Conference Abstracts*, vol. 15, p. 10248, 2013.

189. Zia, H., N. Harris, and G.V. Merrett, "A Low Complexity Data Driven Model of Environmental Discharge Dynamics for Sensor Network Applications," in *Proc. of Eurosensors XXVIII*, 2014.

190. Zia, H., N.R. Harris, and G.V. Merrett, "Water Quality Monitoring, Control and Management (WQMCM) Framework using Collaborative Wireless Sensor Networks," in *Proc. of 11th International Conference on Hydroinformatics* New York City, USA, 2014.

191. Zia, H., N.R. Harris, and G.V. Merrett, "Empirical Modelling and Simulation for Discharge Dynamics Enabling Catchment-Scale Water Quality Management," in *Proc. of The 26th European Modeling & Simulation Symposium*, Bordeaux, France, 2014.

192. Zia, H., et al., "Data-driven Low-Complexity Nitrate Loss Model utilizing Sensor Information – Towards Collaborative Farm Management with Wireless Sensor Networks," in *Proc. of IEEE Sensors Applications Symposium (SAS)*, Zadar, Croatia, 2015.

193. Zia, H., et al., "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(0), p. 126-138, 2013.

194. Coles, N., et al., "Water, boundaries and borders, the great intangibles in water quality management: can new technologies enable more effective compliance?," in *Proc. of Transboundary water management across borders and interfaces: present and future challenges (TWAM2013)*, Aveiro, PT, 2013.

195. Hall, P., et al., "Water quality management: the case for real time monitoring and reporting," in *Proc. of 16th International River symposium*, Brisbane, Australia, 2013.

196. Rivers, M., et al., "How could sensor networks help with agricultural water management issues? Optimizing irrigation scheduling through networked soil-moisture sensors," in *Proc. of IEEE Sensors Applications Symposium (SAS 2015)*, Zadar, Croatia, 2015.

197. Vellidis, G. "Vellidis Research Group at The Univerity of Georgia," Available from: http://www.nespal.org/vellidis/research-projects/smart-irrigation/uga-smart-sensor-array/, 2013.

198. Jiang, T., G.V. Merrett, and N.R. Harris, "Enabling opportunistic energy trading between overlapping energy harvesting wireless sensor networks," in *Proc. of 2nd International Workshop on Energy Neutral Sensing Systems*, ACM, p. 25-30, 2014.

199. Jiang, T., G.V. Merrett, and N.R. Harris, "Opportunistic energy trading between co-located energy-harvesting wireless sensor networks," in *Proc. of 1st International Workshop on Energy Neutral Sensing Systems*, ACM, p. 11, 2013.

200. Trubilowicz, J., K. Cai, and M. Weiler, "Viability of motes for hydrological measurement," *Water Resour. Res*, vol. 45, p. W00D22, 2009.

201. O'Flynn, B., et al., "SmartCoast: a wireless sensor network for water quality monitoring," in *Proc. of*, Ieee, p. 815-816, 2007.

202. Garrido, A. and J. Calatrava, "Agricultural water pricing: EU and Mexico," *Consultant report for the OECD.*, 2010.

203. Berbel, J., J. Calatrava, and A. Garrido, "1 3 Water Pricing and Irrigation: A Review of the European Experience," *Irrigation Water Pricing: The Gap Between Theory and Practice*, vol. 4, p. 295, 2007.

204. Wichelns, D., "Agricultural Water Pricing: United States," Organisation for Economic Co-operation and Development, 2010.

205. Bainbridge, Z.T., et al., "Identifying the land-based sources of suspended sediments, nutrients and pesticides discharged to the Great Barrier Reef from the Tully–Murray Basin, Queensland, Australia," *Marine and Freshwater Research*, vol. 60(11), p. 1081-1090, 2009.

206. Aylward, B., et al., "The Economic Value of Water for Agricultural, Domestic and Industrial Uses: A Global Compilation of Economic Studies and Market Prices," vol. 31, p. 1-29, 2010.

207. FarquharsonA, R.J., D. ChenB, and Y. LiC, "What is the impact on farmer nitrogen fertilizer use of incorporating the effects of nitrous oxide emissions?," in *Proc. of 19th World Congress of Soil Science: Soil solutions for a changing world, Brisbane, Australia, 1-6 August 2010.*, International Union of Soil Sciences (IUSS), c/o Institut für Bodenforschung, Universität für Bodenkultur, p. 274-277, 2010.

208. ABS, "Water Use on Australian Farms 2002–03: Pasture for grazing the biggest user of irrigation water," Australian Bureau of Statistics, (4618.0), 2005.

209. Martin Staines, et al., "Nitrogen for intensively grazed dairy pastures," Department of Agriculture and Food, Australia, (ISSN: 1833-7236), 2011.