Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Low-Complexity Machine Learning Nitrate Loss Predictive Model – Towards Proactive Farm Management in a Networked Catchment

Huma Zia[1], Nick R. Harris[1], Geoff V. Merrett[1], Mark Rivers[2]
[1] *Electronics and Computer Science, University of Southampton, Southampton, United Kingdom*
[2] *ClearWater Research and Management Pty, Greenfields, WA, Australia.*

Corresponding author: Nick Harris (e-mail: nrh@soton.ac.uk).

ABSTRACT With the advent of Wireless Sensor Networks, the ability to predict nutrient-rich discharges, using on-node prediction models, offers huge potential for enabling real-time water reuse and management within an agriculturally-dominated catchment Existing discharge models use multiple parameters and large historical data which are difficult to extract and this, coupled with constraints on network nodes (battery life, computing power, sensor availability etc.) makes it necessary to develop low-dimensional models. This paper investigates a data-driven model for predicting daily total oxidized nitrate (TON) fluxes, and reduces the number of model parameters used to 5 – a reduction of at least 50%. Trained on only a 12-month training data set derived from published measured data, results for the model generated using an M5 decision tree, giving an R^2 of 0.92 and a relative root mean squared error (RRMSE) of 26%. 80% of the residuals for test data fall within +/-0.05 Kg $ha^{-1}day^{-1}$ error range, which is minimal, offering an improvement over results obtained by contemporary research.

INDEX TERMS Environmental modelling, nitrate loss prediction modelling, machine learning, M5 decision tree, wireless sensor networks, water quality management

I. INTRODUCTION

Fertilizers rich in phosphorus (P), potassium (K) and nitrogen (N) are added to soil to increase crop yields. However, agronomic nutrient recommendations are often far in excess of environmentally-appropriate levels [1]. Unlike P, which is less soluble than N (which in the form of nitrates and nitrites, is referred to as total oxidized nitrogen (*TON*)), N is more prone to be lost through leaching and drainage water [2-4]. In some cases, 30% - 50% of applied N is lost due to the coupled impact of over fertilization and irrigation runoff [5]. These substantial nutrient losses can have serious agronomic, economic and environmental implications [6, 7], as for example, nutrient-enrichment of waterways (eutrophication) can give rise to toxic and nuisance algal blooms which contaminate drinking water and harm aquatic life [8, 9]. As an example, in England, 70% of the landscape drains to waterways identified as “N-polluted” under the EC Nitrates Directive; and 60% of this N comes from agriculture [10].

Water quality control through drainage management has been widely advocated in the literature [11-13] and drainage reuse is the most suitable option to address the current water scarcity and quality problems [14, 15]. Drainage reuse can have many environmental as well as economic benefits [16]. For example, in Egypt, the amount of drainage water is 20%-30% of the applied water, which is reused downstream after mixing with fresh water to cater to increased demand on domestic water [17]. In another case, reapplication to the land of N-rich runoff waters provided more than the annual nutrient requirements for that land [18]. Some other research effort has 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 [19]. The results indicated a 33% reduction in fresh water usage for irrigation. Furthermore, it was determined that drainage water collected from the greenhouse contained 59% of the applied N. These studies, though small scale in some cases 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 intelligent and autonomous management mechanisms has not kept pace with the deteriorating water situation [20]. While drainage reuse has been advocated and adopted in farming [21-23], 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 [24, 25], currently restricts wide adoption of this mechanism in agriculture.

A. WSN FOR ENVIRONMENTAL MONITORING

Over recent years, wireless sensor networks (WSNs) with their attractions of low cost and continuous real time data availability, have received considerable attention in environmental monitoring. Since their emergence, WSNs have been successfully used for monitoring and managing farming activities [26, 27] and eco-hydrologic processes [28, 29]. These applications are discussed in detail in [30]. With WSNs receiving considerable attention over the last decade, there now exists huge potential for leveraging small-scale networked agricultural activities among farms into an integrated mechanism by sharing information about discharges across networks and farms and, thus, enabling better impact prediction and pre-emptive management. However, there was no framework to investigate and implement such a mechanism until the authors proposed one, Water Quality Monitoring Control and Management (WQMCM) which utilizes collaboration among networks in a catchment to investigate and enable such a mechanism [30]. The basic model architecture comprises modules to enable individual networks to learn their environment by correlating neighbours' events with events within their own zone (neighbour-linking model), predict their impact in terms of discharges (Q-predictive model) and nutrient losses (N-predictive model), and then adapt the local monitoring and management strategy (Classification and Decision model). Work on the initial two models have been completed [31, 32]. This work is a continuation and extension of the authors' earlier work on *Q*-predictive modelling [33] for water but applied in detail in this case to nitrogen, and is an extension of work previously presented by the authors [34]. Compared to the previous work, we extend the discussion to include more detail on the performance of the model, as well as an analysis of the model for simplified input parameters to investigate the trade-off between accuracy and complexity of the model. This work has also been included in a PhD Thesis by one of the authors [35]

B. RELATED WORK – DATA MODELS

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 [36-38]. 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.

Data-driven models in environmental sciences typically employ 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 [39]. Examples of this approach include models developed for N losses from grasslands in the UK [40, 41] and from agricultural properties in Australia [42]. 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 product, N lost through plant death and decomposition etc.. This approach generally works on annual time-steps, and is very suitable for policy making and management decisions regarding annual nutrient budgets but 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 [43]. The input parameters (11 variables) consisted of annual values related to N application, soil chemistry, and climatic conditions. 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, the study used 8000 training samples based on annual values, which were obtained from different sites, to achieve the optimal results. 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) [44]. 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 (R^2=0.78), however, the model was developed for gaseous N emissions and still relied on soil chemical data. Another study [45] 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. A further study applied data mining tools such as artificial neural networks, polynomial regression etc. to predict weekly nitrate concentrations at a gauging station on the Sangamon River near Decatur, Illinois to determine unsuitability of water (nitrate concentration >8.5 mgl^{-1}) [46]. Input parameters included weekly average values for nitrate concentration (mgl^{-1}), flow discharge (m^{3}s^{-1}), temperature, and total weekly precipitation obtained from dataset between 1994 and 1999. All models were reported to perform reasonably well but it was noted that prediction accuracy could be improved by using hydro-meteorological forecasts, spatially distributed model inputs or by separating surface and base flows. More recently, to design the operation of drip fertigation system for nutrient management, an understanding of nutrient leaching behaviour is required especially for shallow rooted crops like potatoes [47]. In this study, simulated nitrate leaching data obtained from a solute transport model (HYDRUS-2D) was used to train and validate an adaptive network based fuzzy inference system. 528 data points for input parameters such as emitter discharge rate (l h^{-1}) and fertilizer amount (Kg ha^{-1}) along with the output parameter of nitrate leaching (mg) were used. Correlation of coefficient between measured and obtained data from HYDRUS was obtained as 0.99. Furthermore in an another study, regression methods were used in a 2 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 with R^2 of 0.95, however it used percentage of clay and soil depth besides other input parameters, and was based on sparse yearly samples [48].

C. CONTRIBUTION

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 N losses within the farm system with the aim of enabling reutilization and alerts in real time, i.e. by using WSNs. Furthermore, 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. In addition, 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 data 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.

In this paper we extend the concept of abstraction used in [43] and extend on [34] to further simplify the model parameters with a view towards eventual deployment within a WSN. This would enable wide-scale field management applications using WSNs without the need for complicated geo-chemical data. Furthermore, we explore the applicability of an M5 decision tree algorithm for nitrate loss modelling based on the proposed parameters. M5 decision tree is a simpler algorithm compared with ANN and can give comparable performance with less computational time, making it more suitable for implementation within a network. However it has not yet been applied in nitrate modelling.

II. 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 information acquired autonomously and shared by the neighbouring farms.

In order to simplify the model parameters, we extended in [34]the concept of model abstraction done by *Villa-Vialaneix et al.* [43] where 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. Those authors used a data-driven model to simplify and calibrate a physically based model. It is important to mention, that our data-driven model includes new inputs and is not just a simplification of their model. Indeed, our model is not developed from *Villa-Vialaneix et al.* but it is useful to compare the two models. TABLE 1 lists input parameters for the two models (columns 2 and 3) under various input categories. Based on this abstracted list of parameters, which still contains soil chemistry and N sources data, we further abstract this, and add a few additional parameters to achieve the simplified parameters (3rd level abstraction, column 4). This 3rd level abstraction is explained below.

For the input category of climatic conditions, we select precipitation only. Temperature is not selected because the same information can be directly measured by soil moisture sensors. Temperature readings are used in most models 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 [49]. 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 (see [30]), the dependence on proxy parameters can be minimized. Furthermore, we have also elaborated previously on the impracticalities attached with the methods used for acquiring these other soil parameters (porosity, texture etc.). In our model, 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 and so are effectively constants. Therefore, in the category of soil properties, we propose using only soil moisture.

Moving on to the category of N input sources, we select N in fertilizers and in manure from the parameters listed in the 2nd level abstraction. This is because of the ease of 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 [49]. As well as these two parameters (N in fertilizers and manure), we propose to use two additional parameters – 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 [48, 50]. 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.

TABLE 1
ABSTRACTION OF INPUT PARAMETERS FOR THE TON-LOSS PREDICTIVE MODEL USING TRADITIONAL BIO-GEOCHEMICAL MODELS

Input variable category	**Parameters used in existing DNDC model *Li et al.* [49]**	**Parameters used by *Villa-Vialaneix et al.* [43]** ***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

In the management information category, we propose using none of the parameters suggested in the 2nd level abstraction done by *Villa-Vialaneix et al.* [43]. This is because the dataset was comprised of annual values for the parameters, therefore using annual averages for these variables would not have significantly contributed to the model development. Instead, we propose to use crop cover 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 [51]. 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 [52]. 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 [48, 50] 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

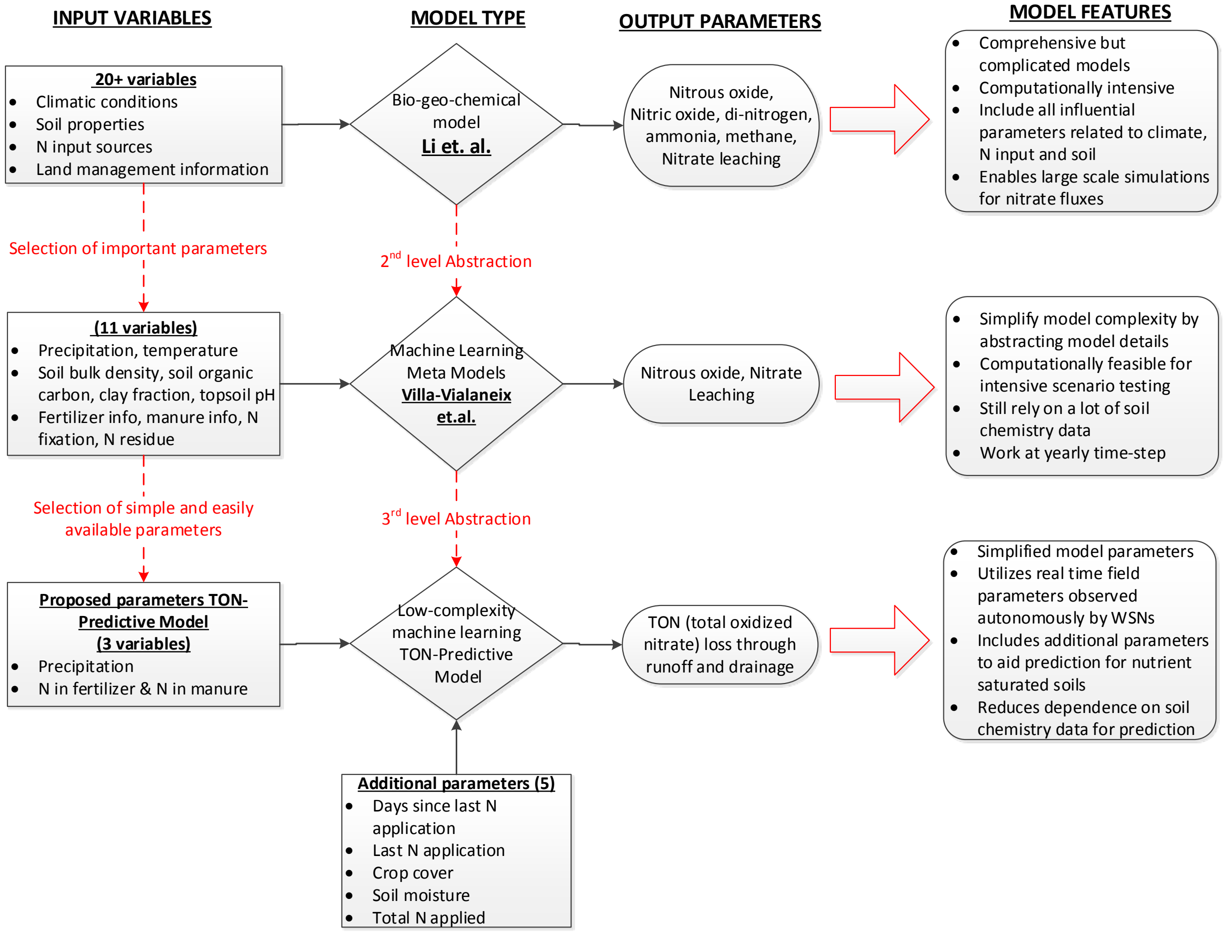

INPUT VARIABLES
MODEL TYPE
OUTPUT PARAMETERS
MODEL FEATURES
20+ variables
Climatic conditions
Soil properties
N input sources
Land management information
Bio-geo-chemical model
Li et. al.
Nitrous oxide, Nitric oxide, di-nitrogen, ammonia, methane, Nitrate leaching
Comprehensive but complicated models
Computationally intensive
Include all influential parameters related to climate, N input and soil
Enables large scale simulations for nitrate fluxes
Selection of important parameters
2nd level Abstraction
(11 variables)
Precipitation, temperature
Soil bulk density, soil organic carbon, clay fraction, topsoil pH
Fertilizer info, manure info, N fixation, N residue
Machine Learning Meta Models
Villa-Vialaneix et.al.
Nitrous oxide, Nitrate Leaching
Simplify model complexity by abstracting model details
Computationally feasible for intensive scenario testing
Still rely on a lot of soil chemistry data
Work at yearly time-step
Selection of simple and easily available parameters
3rd level Abstraction
Proposed parameters TON-Predictive Model
(3 variables)
Precipitation
N in fertilizer & N in manure
Low-complexity machine learning TON-Predictive Model
TON (total oxidized nitrate) loss through runoff and drainage
Simplified model parameters
Utilizes real time field parameters observed autonomously by WSNs
Includes additional parameters to aid prediction for nutrient saturated soils
Reduces dependence on soil chemistry data for prediction
Additional parameters (5)
Days since last N application
Last N application
Crop cover
Soil moisture
Total N applied

FIGURE 1: Model abstraction from high complexity bio-geo-chemical models to low complexity *TON*-loss predictive mode

[4]. 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, we propose using day of the year in the *TON*-loss predictive model to investigate its impact on prediction accuracy of the model.

Thus, the proposed parameters for the *TON*-loss predictive model TABLE 1(column 4,), 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 1.

III. EXPERIMENTAL METHOD

A. DESCRIPTION OF CATCHMENT DATA

A study was carried out by the University of Cork in the Dripsey catchment located in the south of Ireland [50]. The one-year study (2002) was aimed at understanding the underlying processes of nutrient losses from soil to water bodies. The Dripsey catchment is a sub-catchment of the broader catchment of the river Lee in Cork, Ireland. This catchment consists of smaller nested sub-catchments. FIGURE 2a 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*–loss 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 2(a) consisting of 17 ha of farmland. Precipitation (mm) and *TON* export concentration (mgl^{-1}) data, collected every 30 minutes for the year 2002 is used. The dataset is available for research and education purposes via the Environmental Protection Agency (EPA) website [53]. The remainder of the data regarding field conditions is extracted from catchment details available in the associated documentation [50].

The cumulative rainfall for the year 2002 was 1812 mm, whereas the cumulative stream flow depth was measured as 1206 mm of the rainfall at Site 1 (as shown in FIGURE 2(c). The monthly rainfall values range from less than 50 mm in the summer months to over 250 mm in the winter months.

The monthly temperature is 5°C in the winter and 15°C in the summer.

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 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 occur in winter.

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 starts in early March and ends in October. Grass is cut as silage (conserved feed) once or twice a year, typically at the end of May and at the end of July. 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 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.

B. 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. 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 we used data related to half hourly precipitation (mm) and *TON* losses (litres sec^{-1}) for the year 2002. 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 [50].

Since the *TON*-loss predictive model is aimed at facilitating

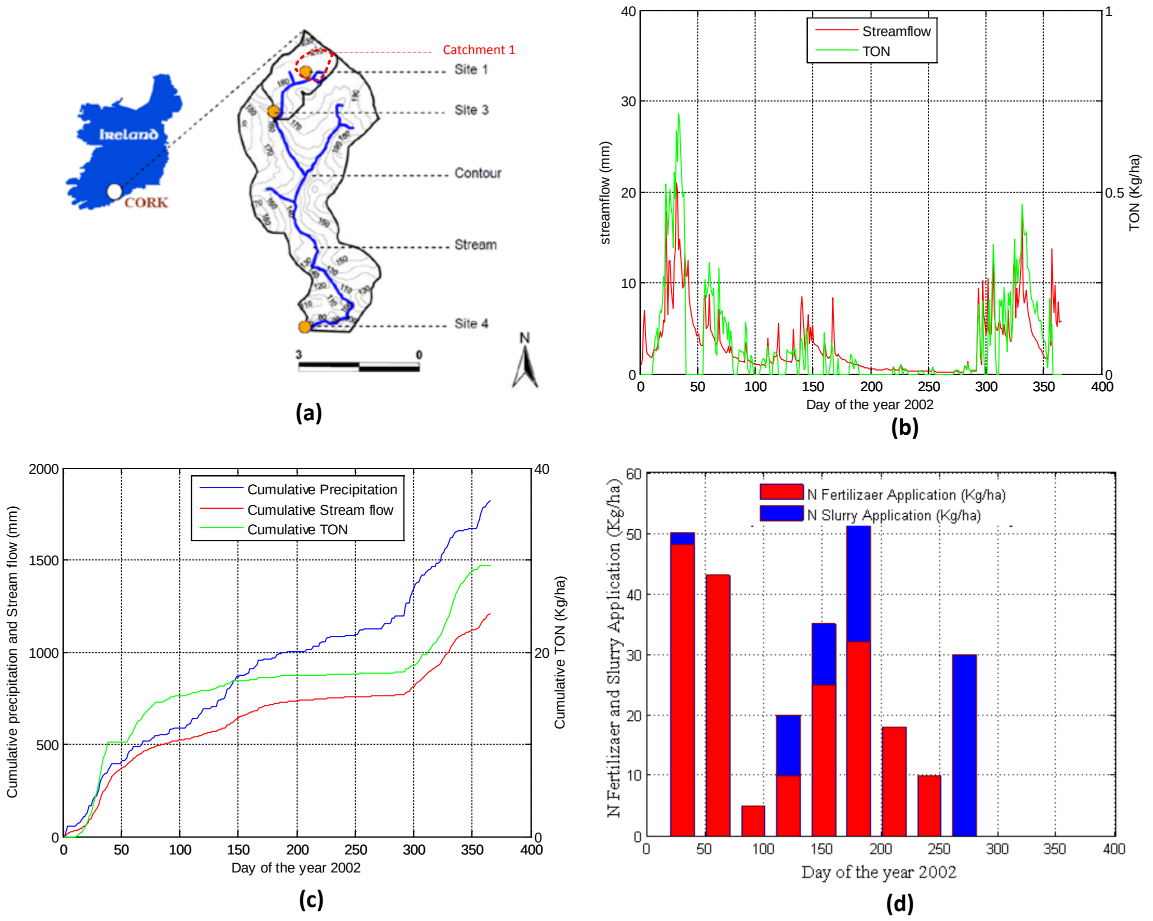


FIGURE 2:(a) Location and map of the Dripsey catchment (reproduced from [50]); (b) plot of observed stream flow and TON at site1 and; (c) cumulative precipitation, stream flow and TON losses observed at site1 for year 2002; (d) fertilizer and slurry applications

daily management decisions regarding nutrient loads (KgNha^{-1}), we convert the hourly values 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), we use an alternative method. Instead of using the mathematical method suggested in the DNDC model [49] (because of the complexity of the required parameters) we use a proxy parameter - last-five-day-rainfall. 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 [54]. 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-five-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 [55]. For example, in a fallow season, field conditions are considered dry, medium and wet respectively if rainfall depths are less than 13mm, between 13mm and 28mm, 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, 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 [50] 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}.

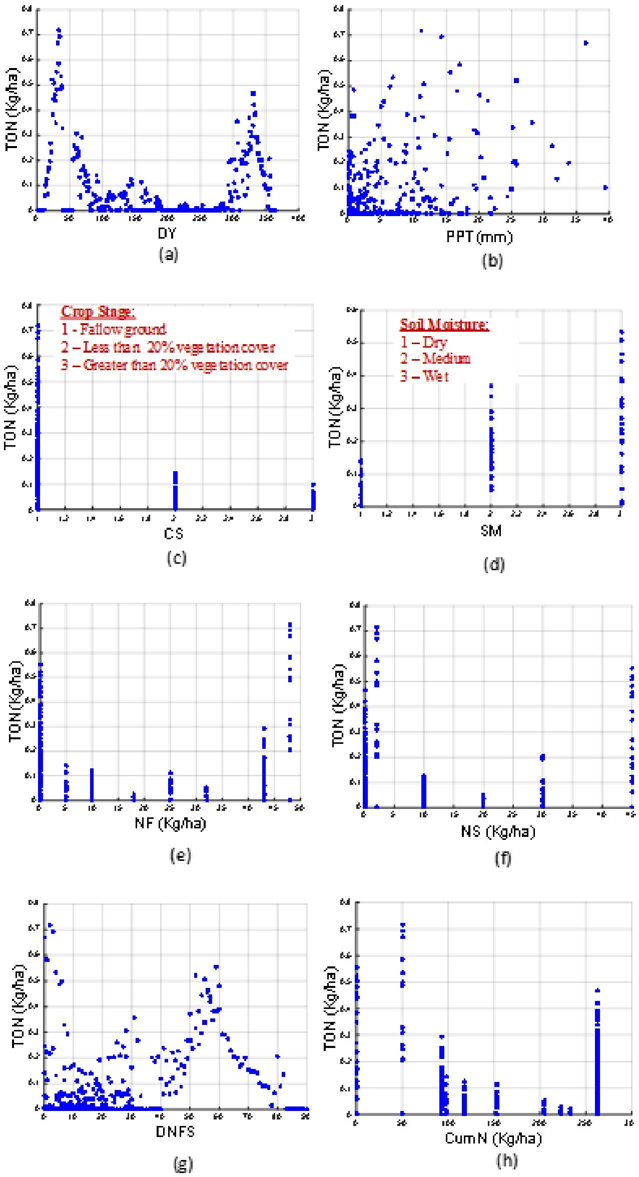


FIGURE 3: Sensitivity analysis of input parameters ((a) DY, (b) PPT, (c) CS, (d) SM, (e) NF, (f) NS, (g) DNFS, (h) CumN) for the TON-loss predictive model with the output parameter TON

TABLE 2
PEARSON CORRELATION COEFFICIENT FOR THE DATA FROM DRIPSEY CATCHMENT

Independent Parameters	**Acronym**	**Pearson (*r*) Correlation Coefficient with *TON* (Kg/ha)**
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 (Kgha^{-1})	*NF*	0.12
Last N Slurry Application (Kgha^{-1})	*NS*	0.05
Days since last Fertilizer/Slurry Application	*DNFS*	0.31
Cumulative N applied so far this year (Kgha^{-1})	*CumN*	-0.27

In the obtained dataset, we exclude instances with zero precipitation values although the impact of no rain remains in the model (using the 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 correlation coefficient (*r*) are given in TABLE 2, along with the acronyms for these variables which will be used from here on in this paper.

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 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 and later scrutinized in detail for exclusion of any parameter.

As discussed, *TON* losses show a definite trend with *DY* (FIGURE 3(a)), though negative with a 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, as shown in FIGURE 3 (a). With *PPT*, there is a positive correlation with *TON*, though not very strong, as one would expect, as shown in (b). The strongest relationship of *TON* losses is with *SM* (*r* = 0.71), as illustrated in **FIGURE 3** (d), which shows that the higher the soil moisture level, the greater the *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 3c), the higher the crop stage, the lesser are the *TON* losses.

It is interesting to note that *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 4 (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, further N applications to N-saturated soil are lost rapidly [4]. However, as proposed in the previous section, *DNFS* and *CumN* show a relatively better correlation with nitrate losses (0.30 and -0.27 respectively).

C. MODELLING TECHNIQUE – DECISION TREE MODELS

After being widely used in hydrologic modelling [56-58], data-driven modelling using machine learning has also started being adopted in nutrient loss modelling [43]. Among the learning algorithms used in hydrology and nutrient modelling, artificial neural networks (ANN) and decision trees have been widely used [37, 43, 45, 48]. 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 [59] developed in Matlab is used for generating the model.

D. MODEL EVALUATION CRITERIA

The quantitative assessment of M5 decision tree modelling for *TON* losses based on the proposed model parameters is performed using a six-step procedure (as suggested in [37]) and adopted by the authors for discharge prediction model [33]:

(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);
(vi) Uncertainty analysis on the model residual.

1. Comparative Assessment against other traditional data-driven approaches

Once an optimized *TON*-loss predictive M5 tree model is identified, based on the model parameters with the best prediction performance, it is compared against other machine learning algorithms used in the literature. For this, multiple linear regression (MLR), REPTree and multi-layer perceptron (MLP) are used. For ANN comparison, two different techniques, namely radial basis function network (RBFN) and multi-layer perceptron (MLP) are used. WEKA [60], a data mining package, is used for the comparative assessment of these models.

MLR attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. 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 [61]. A three-layer MLP is the most widely used type of back-propagation ANN used in hydrology. In a MLP, a network is made up of a number of interconnected nodes, arranged into three basic layers of input, output and hidden. 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 [62].

2. Uncertainty Analysis

As in any prediction, there is a potential error which needs to be accounted for. In this section, the uncertainty of the predicted variable is investigated by the quantification of the residuals. Residuals represent deviation of 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 [63].

Through the analysis, we try to find if residuals show any trend along the days of the year

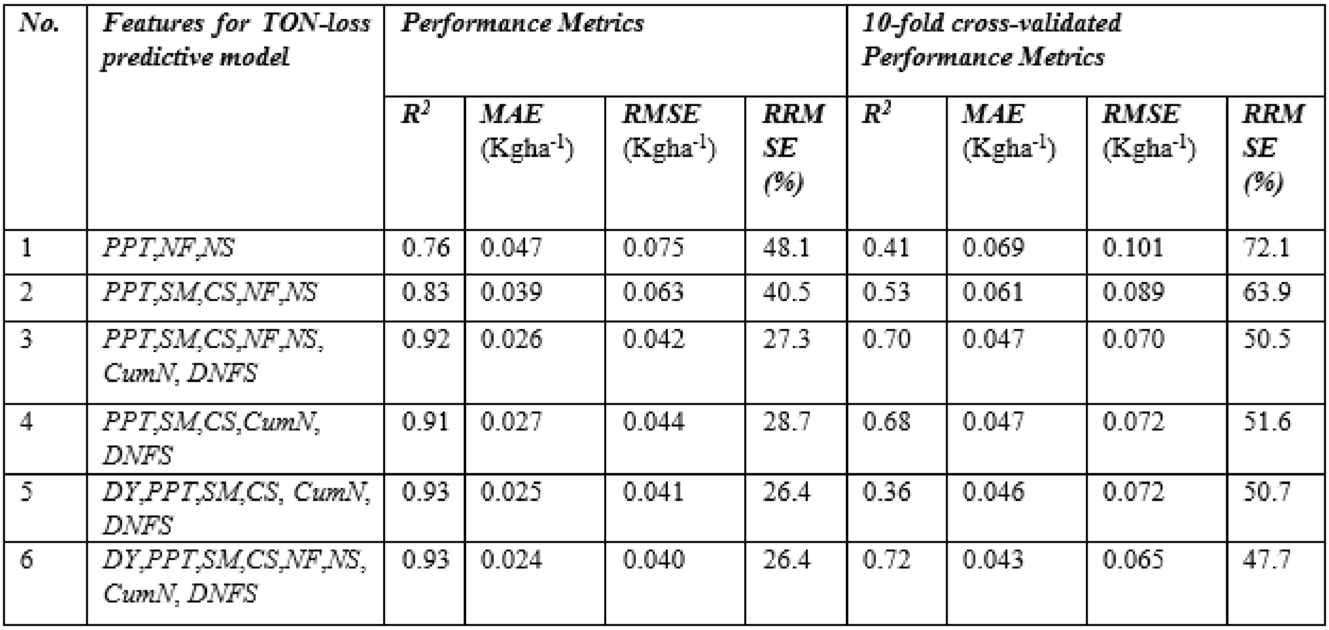
IV. RESULTS AND DISCUSSION FOR VALIDATION OF *TON*-LOSS PREDICTIVE MODEL

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 the M5 toolbox in Matlab. The generated 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 [64].

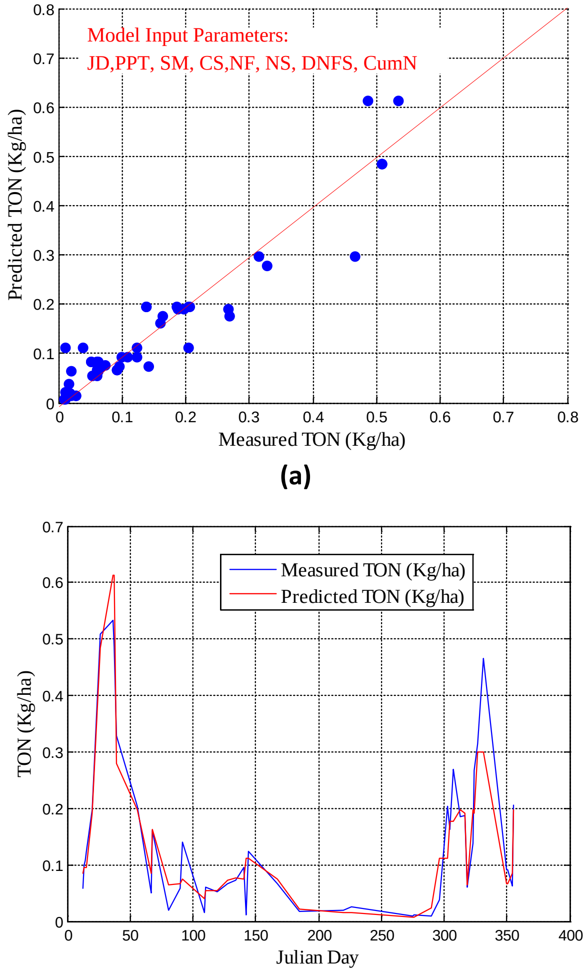
For the training dataset, this value is calculated as 0.078. Even in 10-fold cross-validated result for the model, the value for RMSE (0.065) falls well below this threshold.

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 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

TABLE 3
PERFORMANCE MEASURES FOR VARIOUS *TON*-LOSS PREDICTIVE MODELS DEVELOPED USING DIFFERENT INPUT PARAMETERS



No.	*Features for TON-loss predictive model*	*Performance Metrics*				*10-fold cross-validated Performance Metrics*			
		R^2	*MAE* ($Kgha^{-1}$)	*RMSE* ($Kgha^{-1}$)	*RRMSE (%)*	*R^2*	*MAE* ($Kgha^{-1}$)	*RMSE* ($Kgha^{-1}$)	*RRMSE (%)*
1	*PPT,NF,NS*	0.76	0.047	0.075	48.1	0.41	0.069	0.101	72.1
2	*PPT,SM,CS,NF,NS*	0.83	0.039	0.063	40.5	0.53	0.061	0.089	63.9
3	*PPT,SM,CS,NF,NS, CumN, DNFS*	0.92	0.026	0.042	27.3	0.70	0.047	0.070	50.5
4	*PPT,SM,CS,CumN, DNFS*	0.91	0.027	0.044	28.7	0.68	0.047	0.072	51.6
5	*DY,PPT,SM,CS, CumN, DNFS*	0.93	0.025	0.041	26.4	0.36	0.046	0.072	50.7
6	*DY,PPT,SM,CS,NF,NS, CumN, DNFS*	0.93	0.024	0.040	26.4	0.72	0.043	0.065	47.7



(a)

(b)

FIGURE 4: (a) Scatter plot of predicted TON, using the proposed TON-loss predictive model, against the measured TON; (b) plot of predicted and measured TON against day of the year

against day of the year (Figure 4b). It is apparent that both curves overlap significantly, however the model seems to under predict in the last 50 days.

A. 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, we compare its results with the meta-models developed by *Villa-Vialaneix et al.* [43]. 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 3.

In that study, the decision tree based meta-model resulted in 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 decision trees and simplified parameters gives 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, we have used a different learning algorithm, the M5 decision trees. 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.

B. FURTHER MODEL SIMPLIFICATION

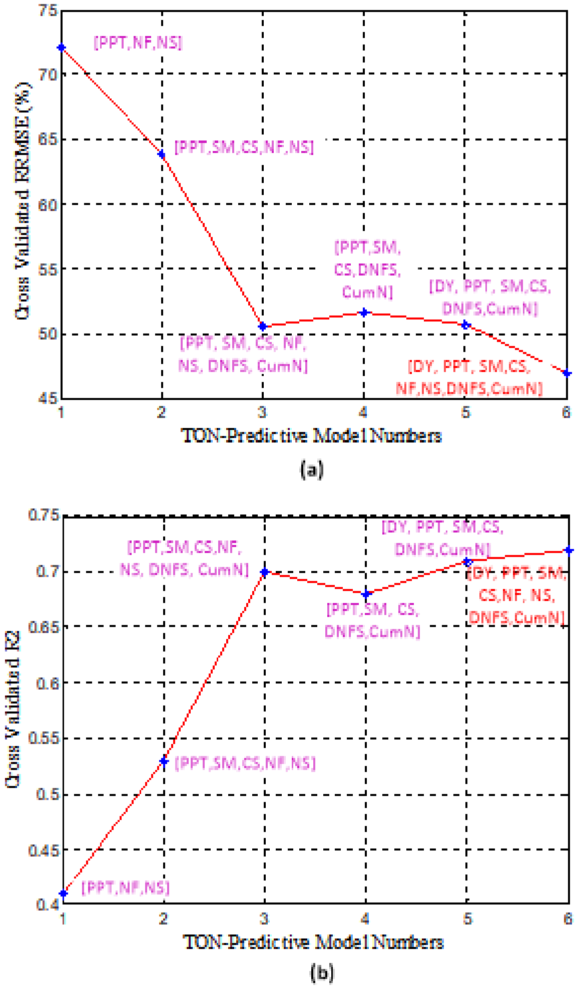
In this section we evaluate the impact of further model parameter simplification on the predictive performance of the model. This is because, some input parameters, e.g. NS and NF, do not have a strong correlation with the output parameters despite apparently being very important variables. Furthermore, 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. 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 we develop six models, using different combination of the proposed input parameters, as discussed below. The performance parameters of the generated models are listed in Table 4.

TABLE 4
PERFORMANCE COMPARISON OF THE *TON* PREDICTIVE MODEL WITH AN EXISTING WORK BASED ON META MODELS

	Learning Algorithm used	**Performance Metric (R^2)**
Meta models by *Villa-Vialaneix et al.* [43]	Decision tree	0.74
	MLP	0.82
TON-loss predictive model	M5 tree	0.92

Firstly, we start with developing a model (model no.1) comprised of the available input parameters which are common with the 2nd layer abstraction model we used in our simplification by *Villa-Vialaneix et al.* [43]. The cross validated performance indicate low performance with R^2 of 0.413, MAE of 0.069, RMSE of 0.101 which is beyond the acceptable value of RMSE (0.078), and RRMSE of 72.1%.

This validates inclusion of additional parameters, as we have proposed. We then include variables related to field conditions such as *SM* and *CS* to develop model no.2. The performance increases with relative error, RRMSE, reducing to 63.9%. The RMSE value (0.703) also falls under the acceptable value substantially with RRMSE value of 50.5% and R^2 of 0.703.



(a)

(b)

FIGURE 5: Comparison of (a) RRMSE and (b) R2 values, for the TON-loss predictive models developed using various input parameter combinations

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 of 0.78. This validates the inclusion of additional parameters related to N in the *TON*-loss predictive model.

We then go on to remove N input parameters, *NS* and *NF*, to develop model no.4 which gives 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 we include the *DY* to develop model no.5, for which the performance increases slightly by 1% relative error. Finally we include all the proposed input parameters (3rd level abstraction) for the *TON*-loss predictive model no.6, which gives 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 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).

C. COMPARATIVE ASSESSMENT OF *TON*-LOSS PREDICTIVE MODEL USING VARIOUS LEARNING ALGORITHMS

Using the optimal input parameter set, consisting of *DY*, *PPT*, SM, *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. In this case, all the models are developed using WEKA [60] for ease of use. TABLE 5 lists the performance metrics of the generated models. It is apparent that the M5 tree algorithm 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.

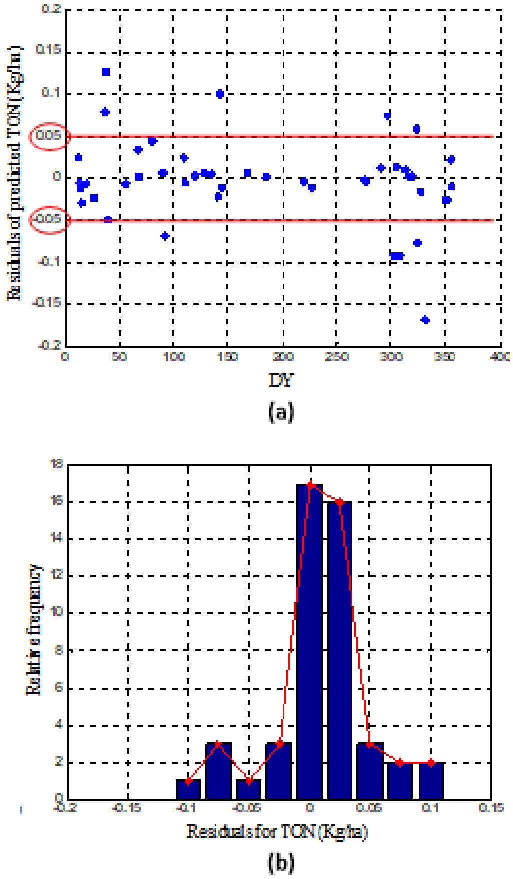
TABLE 5
PERFORMANCE COMPARISON OF THE TON-LOSS PREDICTIVE MODELS DEVELOPED USING VARIOUS MACHINE LEARNING ALGORITHMS

Model Type	10-fold cross-validated performance metric			
	R^2	MAE ($Kgha^{-1}$)	RMSE ($Kgha^{-1}$)	RRMSE (%)
M5 tree algorithm	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 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 ANN model developed using MLP, gives 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 various input parameter combinations 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 we conclude that the M5 tree solution is currently the best performing learning algorithm.

D. UNCERTAINTY ANALYSIS



(a)

(b)

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

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 (a) shows that 80% of the residuals for the predicted test values falls within +0.05 Kgha^{-1} error range. A 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, incorrect estimations of this scale do not impact the 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 (b)).

V. CONCLUSION

In this paper, we evaluate a nitrate loss model that uses a simplified set of parameters when compared with other models reported. In addition, we have investigated the change in performance as the parameter set is reduced further, allowing a relatively small amount of performance to be traded against model complexity, depending on which parameters are chosen. By using an M5 decision tree learning approach and parameters that avoid reliance on complex bio-geo-chemical parameters, a data driven model with low computing complexity is possible, and this could be implemented on node. Using a nitrate loss dataset measured for a sub-catchment in Ireland the model variations are assessed for their relative prediction accuracy in comparison to other popular data-driven methods in hydrological and nitrate loss modelling. 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 result in an R^2 of 0.72 and an RRMSE of 47.7%. 80% of the residuals for the predicted test values fall within +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 (by 2-3%) performance difference. Hence we conclude that for the considered catchment, the proposed parameters list is the most appropriate to accomplish optimal results. We may also conclude 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 has superior performance to 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 paper show great potential for enabling data driven real time nutrient control and management applications within collaborative networked farm systems, as these require simplified models due to sensor node resource constraints precluding implementing complex bio-geo-chemical models locally. This is due to both the non-availablilty of sensors to measure the required parameters directly, and also to then compute efficiently the required model.

V. ACKNOWLEDGEMENTS

We recognise the support of the EPSRC under grant number EP/P010164/1

REFERENCES

[1] D. K. Mueller, D. R. Helsel, and M. A. Kidd, *Nutrients in the nation's waters: too much of a good thing?*, Circular 1136 ed.: US Government Printing Office, Circular 1136, 1996.

[2] V. H. Smith, G. D. Tilman, and J. C. Nekola, "Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems," *Environmental pollution*, vol. 100, pp. 179-196, 1999.

[3] B. Vanlauwe, J. Wendt, J. Diels, G. Tian, F. Ishida, D. Keatinge, *et al.*, "Combined application of organic matter and fertilizer," 2001, pp. 247-279.

[4] L. A. McKergow, D. M. Weaver, I. P. Prosser, R. B. Grayson, and A. E. G. Reed, "Before and after riparian management: sediment and nutrient exports from a small agricultural catchment, Western Australia," *Journal of Hydrology*, vol. 270, pp. 253-272, 2003.

[5] X. Liu, X. Ju, F. Zhang, J. Pan, and P. Christie, "Nitrogen dynamics and budgets in a winter wheat-maize cropping system in the North China Plain," *Field Crops Research*, vol. 83, pp. 111-124, 2003.

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

[7] S. P. Syswerda, B. Basso, S. K. Hamilton, J. B. Tausig, and G. P. Robertson, "Long-term nitrate loss along an agricultural intensity gradient in the Upper Midwest USA," *Agriculture, Ecosystems & Environment*, vol. 149, pp. 10-19, 2012.

[8] V. H. Smith and D. W. Schindler, "Eutrophication science: where do we go from here?," *Trends in ecology & evolution*, vol. 24, pp. 201-207, 2009.

[9] W. Quan 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, pp. 291-299, 2002.

[10] D. f. Environment, R. Affairs, and N. England, *Protecting our water, soil and air: a code of good agricultural practice for farmers, growers and land managers*: DERECHO INTERNACIONAL, 2009.

[11] J. Rhoades, "Intercepting, isolating and reusing drainage waters for irrigation to conserve water and protect water quality," *Agricultural Water Management*, vol. 16, pp. 37-52, 1989.

[12] Q. X. Fang, R. W. Malone, L. Ma, D. B. Jaynes, K. R. Thorp, T. R. Green, *et al.*, "Modeling the effects of controlled drainage, N rate and weather on nitrate loss to subsurface drainage," *Agricultural Water Management*, vol. 103, pp. 150-161, 2012/01/01/ 2012.

[13] M. R. Williams, K. W. King, and N. R. Fausey, "Drainage water management effects on tile discharge and water quality," *Agricultural Water Management*, vol. 148, pp. 43-51, 2015/01/31/ 2015.

[14] K. K. Tanji and N. C. Kielen, *Agricultural drainage water management in arid and semi-arid areas*: Food and Agriculture Organization of the United Nations, 2002.

[15] A. A. Rashed and E. El-Sayed, "Simulating agricultural drainage water reuse using QUAL2K Model: case study of the Ismailia canal catchment area, Egypt," *Journal of Irrigation and Drainage Engineering*, vol. 140, p. 05014001, 2014.

[16] L. S. Pereira, T. Oweis, and A. Zairi, "Irrigation management under water scarcity," *Agricultural Water Management*, vol. 57, pp. 175-206, 2002.

[17] H. I. Abdel-Shafy and M. S. Mansour, "Overview on water reuse in Egypt: present and future," *Sustainable Sanitation Practice*, vol. 14, pp. 17-25, 2013.

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

[19] H. S. Grewal, 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, pp. 841-846, 3// 2011.

[20] W. G. M. Bastiaanssen, R. G. Allen, P. Droogers, G. D'Urso, and P. Steduto, "Twenty-five years modeling irrigated and drained soils: State of the art," *Agricultural Water Management*, vol. 92, pp. 111-125, 9/16/ 2007.

[21] D. D. Adelman, "Simulation of irrigation reuse system nitrate losses and potential corn yield reductions," *Environmental Science & Policy*, vol. 3, pp. 213-217, 2000.

[22] L. Willardson, D. Boels, and L. Smedema, "Reuse of drainage water from irrigated areas," *Irrigation and Drainage Systems*, vol. 11, pp. 215-239, 1997.

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

[24] G. Carr, R. B. Potter, and S. Nortcliff, "Water reuse for irrigation in Jordan: Perceptions of water quality among farmers," *Agricultural Water Management*, vol. 98, pp. 847-854, 2011.

[25] J. Oster and S. Grattan, "Drainage water reuse," *Irrigation and Drainage Systems*, vol. 16, pp. 297-310, 2002.

[26] N. Wang, N. Zhang, and M. Wang, "Wireless sensors in agriculture and food industry—Recent development and future perspective," *Computers and Electronics in Agriculture*, vol. 50, pp. 1-14, 2006.

[27] L. Ruiz-Garcia, L. Lunadei, P. Barreiro, and I. Robla, "A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends," *Sensors*, vol. 9, pp. 4728-4750, 2009.

[28] F. Regan, A. Lawlor, B. O. Flynn, J. Torres, R. Martinez-Catala, C. O'Mathuna, *et al.*, "A demonstration of wireless sensing for long term monitoring of water quality," 2009, pp. 819-825.

[29] P. W. Rundel, E. A. Graham, M. F. Allen, J. C. Fisher, and T. C. Harmon, "Environmental sensor networks in ecological research," *New Phytologist*, vol. 182, pp. 589-607, 2009.

[30] H. Zia, N. R. Harris, G. V. Merrett, M. Rivers, and N. Coles, "The impact of agricultural activities on water quality: A case for collaborative catchment-scale management using integrated wireless sensor networks," *Computers and Electronics in Agriculture*, vol. 96, pp. 126-138, 2013.

[31] H. Zia, N. R. Harris, and G. V. Merrett, "Water Quality Monitoring, Control and Management (WQMCM) Framework using Collaborative Wireless Sensor Networks," presented at the 11th International Conference on Hydroinformatics HIC2014 New York City, USA, 2014.

[32] H. Zia, N. R. Harris, and G. V. Merrett, "Empirical Modelling and Simulation for Discharge Dynamics Enabling Catchment-Scale Water Quality Management," presented at the The 26th European Modeling & SImulation Symposium, Bordeaux, France, 2014.

[33] H. Zia, N. Harris, G. Merrett, and M. Rivers, "Predicting discharge using a low complexity machine learning model," *Computers and Electronics in Agriculture*, vol. 118, pp. 350-360, 2015.

[34] H. Zia, N. Harris, G. Merrett, and M. Rivers, "Data-driven low-complexity nitrate loss model utilizing sensor information—Towards collaborative farm management with wireless sensor networks," in *Sensors Applications Symposium (SAS), 2015 IEEE*, 2015, pp. 1-6.

[35] H. Zia, "Enabling proactive agricultural drainage reuse for improved water quality through collaborative networks and low-complexity data-driven modelling," PhD, Electronics and Computer Science, University of Southampton 2015.

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

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

[38] Q. Chen, Y. Morales-Chaves, H. Li, and A. E. Mynett, "Hydroinformatics techniques in eco-environmental modelling and management," *Journal of Hydroinformatics*, vol. 8, p. 297, 2006.

[39] X. Ding, Z. Shen, Q. Hong, Z. Yang, X. Wu, and R. Liu, "Development and test of the Export Coefficient Model in the Upper Reach of the Yangtze River," *Journal of Hydrology*, vol. 383, pp. 233-244, 2010.

[40] H. Rodda, D. Scholefield, B. Webb, and D. Walling, "Management model for predicting nitrate leaching from grassland catchments in the United Kingdom: 1. Model development," *Hydrological Sciences Journal*, vol. 40, pp. 433-451, 1995.

[41] P. J. Johnes, "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, pp. 323-349, 1996.

[42] S. D. Neville, D. Weaver, K. Lavell, M. Clarke, R. Summers, H. Ramsey, *et al.*, "Nutrient balance case studies of agricultural activities in south west Western Australia," in *7th International River Symposium, Brisbane, Australia, Aug 31st–3rd Sept*, 2004.

[43] N. Villa-Vialaneix, M. Follador, M. Ratto, and A. Leip, "A comparison of eight metamodeling techniques for the simulation of N2O fluxes and N leaching from corn crops," *Environmental Modelling & Software*, vol. 34, pp. 51-66, 2012.

[44] F. Oehler, J. C. Rutherford, and G. Coco, "The use of machine learning algorithms to design a generalized simplified denitrification model," *Biogeosciences*, vol. 7, pp. 3311-3332, 2010.

[45] J. D. Piñeros Garcet, A. Ordoñez, J. Roosen, and M. Vanclooster, "Metamodelling: Theory, concepts and application to nitrate leaching modelling," *Ecological modelling*, vol. 193, pp. 629-644, 2006.

[46] M. Markus, M. Hejazi, P. Bajcsy, O. Giustolisi, and D. Savic, "Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois," *Journal of Hydroinformatics*, vol. 12, pp. 251-261, 2010.

[47] S. Hosein, A. Majid, H. M. Ali, N.-p. Hossein, A. Fariborz, and S. Farid, "Nitrate leaching from a potato field using adaptive network-based fuzzy inference system," *Journal of hydroinformatics*, 2012.

[48] J. Fortin, A. Morais, F. Anctil, and L. Parent, "Comparison of Machine Learning Regression Methods to Simulate NO 3 Flux in Soil Solution under Potato Crops," *Applied Mathematics*, vol. 2014, 2014.

[49] C. Li, N. Farahbakhshazad, D. B. Jaynes, D. L. Dinnes, W. Salas, and D. McLaughlin, "Modeling nitrate leaching with a biogeochemical model modified based on observations in a row-crop field in Iowa," *Ecological modelling*, vol. 196, pp. 116-130, 2006.

[50] C. Lewis, "Phosphorus, nitrogen and suspended sediment loss from soil to water from agricultural grassland," NUI, 2003 at Department of Civil and Environmental Engineering, UCC., 2003.

[51] R. P. Udawatta, P. P. Motavalli, H. E. Garrett, and J. J. Krstansky, "Nitrogen losses in runoff from three adjacent agricultural watersheds with claypan soils," *Agriculture, Ecosystems & Environment*, vol. 117, pp. 39-48, 2006.

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

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

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

[55] D. Allan, D. Erickson, and J. Fay, "The influence of catchment land use on stream integrity across multiple spatial scales," *Freshwater Biology*, vol. 37, pp. 149-161, 1997.

[56] R. Wilby, R. Abrahart, and C. Dawson, "Detection of conceptual model rainfall—runoff processes inside an artificial neural network," *Hydrological Sciences Journal*, vol. 48, pp. 163-181, 2003.

[57] K. Rasouli, 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, pp. 284-293, 2012.

[58] D. Solomatine and A. Ostfeld, "Data-driven modelling: some past experiences and new approaches," *Journal of hydroinformatics*, vol. 10, pp. 3-22, 2008.

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

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

[61] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Advances in Space Research*, vol. 41, pp. 1955-1959, 2008.

[62] D. Solomatine 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, pp. 491-501, 2004.

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

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