Computers and Electronics in Agriculture 118 (2015) 350–360

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Predicting discharge using a low complexity machine learning model

Huma Zia [a,*], Nick Harris [a], Geoff Merrett [a], Mark Rivers [b]

[a] Electronics and Computer Science, University of Southampton, Southampton, United Kingdom
[b] Institute of Agriculture, University of Western Australia, Australia

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 13 August 2015
Accepted 12 September 2015
Available online 6 October 2015

Keywords:
Wireless sensor networks
Discharge prediction
Environmental modelling
Machine learning
M5 trees

ABSTRACT

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

1. Introduction

Over recent years, wireless sensor networks (WSNs), with their attractions of low cost and real time data availability, have received considerable attention in automating agricultural processes for economic benefits, e.g. in precision irrigation, pest control, and animal farming. However, a research gap still exists for mechanizing reutilization of resources (water and nutrients) amongst farms in order to additionally maximise environmental benefits. There is huge potential for leveraging existing networked agricultural activities into an integrated mechanism by sharing information about discharges (Zia et al., 2013). To illustrate this consider that the most commonly used irrigation method, surface irrigation, results in 40–60% of water losses in the form of runoff (Eisenhauer, 2011; Tindula et al., 2013). This runoff can transport up to 30–50% of applied nutrients to stream water and rivers (Liu et al., 2003) In the light of these figures, the motivation for this work is to develop a system that can potentially reduce water consumption and reduce outflows from farms, by predicting and monitoring discharge from local areas. This will enable the development of systems that can then proactively control irrigation strategies and also implement drainage reuse. This will also lead to improved water quality as it will allow nutrients to be kept in the place where they can be useful where previously they would have been discharged with no control into the local environment, eventually ending up in the streams and rivers. While drainage reuse has been advocated and adopted in farming (Adelman, 2000; Willardson et al., 1997; Harper, 2012), various resource constraints and farmer's concerns regarding real time availability of information on volumes, timings, and quality of discharges that will be delivered to the farms (Carr et al., 2011; Oster and Grattan, 2002), currently restricts wide adoption of this mechanism in agriculture.

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

To date, numerous physically-based hydrological models have been developed for the prediction of discharges, either measured as surface runoff, groundwater leaching or stream-flow. Although these models are popular in academic research and are very

* Corresponding author. Tel.: +44 (971) 55773242.
E-mail addresses: hz2g11@ecs.soton.ac.uk, huzach@hotmail.com (H. Zia), nrh@ecs.soton.ac.uk (N. Harris), gvm@ecs.soton.ac.uk (G. Merrett), mark.rivers@uwa.edu.au (M. Rivers).

http://dx.doi.org/10.1016/j.compag.2015.09.012
0168-1699/

useful in evaluating different scenarios, their dependence on acquiring numerous parameters, the need for calibrating models to individual areas, and the tremendous computational burden involved in running the models makes wide-spread application complicated and difficult (Basha et al., 2008; Galelli and Castelletti, 2013). In contrast, data-driven models have good prediction capability and require fewer parameters, which is consistent with the requirement for a reduction in the computational burden of decision making (Castelletti et al., 2010). Thus data-driven modelling, using machine learning algorithms, has been widely used in hydrological modelling (Wilby et al., 2003; Rasouli et al., 2012; Solomatine and Ostfeld, 2008) with artificial neural networks (ANN) being a popular choice (Dawson and Wilby, 1998; Minns and Hall, 1996; Wilby et al., 2003). Recently, decision tree modelling has been investigated (Galelli and Castelletti, 2013; Villa-Vialaneix et al., 2012; Fortin et al., 2014; Piñeros Garcet et al., 2006; Kuzmanovski, 2012) and an interesting example of this class are M5 model trees (Quinlan, 1992). The advantage of M5 model trees over ANNs are that they are faster to train and have guaranteed convergence (Solomatine and Dulal, 2003). However, there are two limitations in the existing work; either the existing models use simpler parameters but years of historical data with thousands of training samples to learn the heterogeneity of large areas (>1000 ha) (Galelli and Castelletti, 2013; Solomatine and Xue, 2004a), or they use more complex models with a significant number of parameters (Bhattacharya et al., 2005; Kuzmanovski, 2012). Additionally none of these approaches have been specifically targeted at sensor network applications, and the data used was obtained through traditional sampling methods in gauged catchments.

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

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

2. Experimental method

2.1. Specification of catchment data

The University of Cork carried out a study on the Dripsey catchment in the south of Ireland. The one-year study (2002) was aimed at understanding the underlying processes of nutrient loss from soil to water bodies within the catchment (Lewis, 2003) and thus fits the requirement for validating the *Q*-predictive model. This catchment consists of smaller nested sub-catchments. Fig. 1(a) shows the location of various data collection points in the stream network such as site 1, site 3 and site 4, which collect water drained from their associated sub-catchments. For the development of the *Q*-predictive model, data available for site 1 of the stream network is used. The sub-catchment which drains into this stream location is identified as 'catchment 1' (as shown in Fig. 1(a)) consisting of 17 ha of farmland. Precipitation (mm) and stream flow (mm) data, collected every 30 min for the year 2002 is used. The data is publically available for research and education purposes via the Environmental Protection Agency (EPA) website (Keily, 2003). The remainder of the data regarding field conditions is extracted from catchment descriptors available in the associated documentation (Lewis, 2003).

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

2.2. Modelling technique – M5 decision tree

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

As discussed in the introduction, decision tree modelling, specifically, is receiving increasing attention in the hydrological literature, in comparison to other learning models. Decision tree modelling is a method of approximating a target variable (output), with discrete values, from a given data set and represents the learned function in form of a decision tree (Mitchell, 1999), where each leaf contains the target values. Decision trees have been shown to perform well when compared to other model types (Galelli and Castelletti, 2013; Zhao and Zhang, 2008) but they do have one disadvantage. In decision trees, the predicted output is composed of discrete values and is reconstructed as a piecewise constant function. To ensure good prediction accuracy, the number