Rusjan, S., Mikoš, M. 2008. Assessment of hydrological and seasonal controls over the nitrate flushing from a forested watershed using a data mining technique. Hydrology and Earth System Sciences 12, 2: 645-656.

**Hydrology and
Earth System
Sciences**

Assessment of hydrological and seasonal controls over the nitrate flushing from a forested watershed using a data mining technique

S. Rusjan and M. Mikoš

University of Ljubljana, Faculty of Civil and Geodetic Engineering, Chair of Hydrology and Hydraulic Engineering, Jamova 2, 1000 Ljubljana, Slovenia

Abstract. A data mining, regression tree algorithm M5 was used to review the role of mutual hydrological and seasonal settings which control the streamwater nitrate flushing during hydrological events within a forested watershed in the southwestern part of Slovenia, characterized by distinctive flushing, almost torrential hydrological regime. The basis for the research was an extensive dataset of continuous, high frequency measurements of seasonal meteorological conditions, watershed hydrological responses and streamwater nitrate concentrations. The dataset contained 16 recorded hydrographs occurring in different seasonal and hydrological conditions. Based on predefined regression tree pruning criteria, a comprehensible regression tree model was obtained in the sense of the domain knowledge, which was able to adequately describe most of the streamwater nitrate concentration variations (RMSE=1.02 mg/l-N; r=0.91). The attributes which were found to be the most descriptive in the sense of streamwater nitrate concentrations were the antecedent precipitation index (API) and air temperatures in the preceding periods. The model was most successful in describing streamwater concentrations in the range 1–4 mg/l-N, covering large proportion of the dataset. The model performance was little worse in the periods of high streamwater nitrate concentration peaks during the summer hydrographs (up to 7 mg/l-N) but poor during the autumn hydrograph (up to 14 mg/l-N) related to highly variable hydrological conditions, which would require a less robust regression tree model based on the extended dataset.

1 Introduction

In recent years, the export of nitrogen from forested watersheds has become an important research area and a public policy issue since nitrogen leaching can strip nutrients from forest soils, acidify streams and cause eutrophication (Vitousek et al., 1997; Fenn et al., 1998; Lovett et al., 2002; Wade et al., 2004; Fitzhugh et al., 2003). The variability in nitrogen loss from forested watersheds is high and has been ascribed to many causes, including differences in atmospheric nitrogen inputs (Stoddard, 1994; Aber et al., 2003), pedology (Gundersen et al., 1998), forest history (Goodale et al., 2000) and hydrology (Hornberger et al., 1994; Creed et al., 1996).

The hydrologically induced mobilization of nitrate as the most mobile form of nitrogen from the undisturbed, forested watersheds has received considerable attention in recent hydrological and biogeochemical studies (Creed et al., 1996; McHale et al., 2002; Beachtold et al., 2003; Weiler and McDonnell, 2006). Nitrate concentrations in the streamwater draining forested watershed provide the fundamental information about biogeochemical processing of nitrogen in the forest ecosystem (Burns, 1998; Goodale et al., 2002). At seasonal boundaries, accumulation of labile dissolved inorganic nitrogen in excess of physical and biological retention capacity tends to occur (Likens and Boremann, 1995; Cirmo and McDonnell, 1997; Lovett et al., 2002; Vanderbilt et al., 2003). Both autumn and spring streamwater nitrate pulses are usually observed, with autumn increases in the nitrate concentration associated with a greater amount of precipitation and diminished biological assimilation, whereas spring pulses are reported mainly from watersheds with snowmelt driven hydrology

Correspondence to: S. Rusjan
(srusjan@fgg.uni-lj.si)

(Arheimer et al., 1996; McHale et al., 2000; Langusch and Matzner, 2002; Sickman et al., 2003; Clark et al., 2004).

The understanding of how hydrological conditions trigger flushing of labile nutrients on a watershed scale is still rather poor, especially when we move from the timescale of seasonal variability towards the timescale of a single hydrological event. The main differences in the explanation of the hydrologically driven export of nitrate found in the literature are not necessarily contradictory as they can be ascribed to discrepancies in basic hydrological and climatological conditions, topography, forest soil characteristics and biogeochemical behaviour of forest ecosystems (Cirmo and McDonnell, 1997; Andersson and Lepisto, 1998; Aber et al., 2002; Worall et al., 2003; Stieglitz et al., 2003; Binkley et al., 2004).

The inability to obtain an insight into the interactions between the hydrological and biogeochemical states, which control the nitrate flushing mechanisms, lies in the complexity of event-scale hydro-biogeochemical observations. Studies of the hydrologically induced nitrate export behaviour from forests are observed mainly at low time frequencies, which do not allow tracing the behaviour of nitrate export during a particular hydrological event. In order to understand how hydrological flowpaths affect stream chemistry and, conversely, to use stream chemistry to decipher hydrological processes, we need chemical measurements at time scales that correspond to the hydrological dynamics of an observed hydrologic unit (Kirchner et al., 2004).

Several models have been used to describe the nutrient runoff and transport, for example ANSWERS (Beasley et al., 1980); AGNPS (Young et al., 1995); SWAT (Arnold and Allen, 1996); HBV-N (Pettersson et al., 2001); TOPCAT-NP (Quinn et al., 2007). These models are mainly applied to estimate the budgets of nutrient exports as a consequence of land use patterns and management. A conceptual explanation of mechanisms responsible for the observed peaks in solute concentration during changeable hydrological conditions was proposed by Hornberger et al. (1994) and is commonly referred to as the "Flushing Hypothesis". The hypothesis was later modelled by Creed et al. (1996) using TOPMODEL as the hydrological simulator. Temporally, the calculations in the models listed above are usually performed at daily or even longer time steps; therefore, their application for the modelling of hydrologically, event-induced nitrate mobilization is limited.

Our study presents an investigation of interacting seasonal and hydrological conditions which strongly influence the export of nitrate from a forested watershed in the SW part of Slovenia. The continuous high-frequency measurements of streamwater nitrate concentration in the periods of hydrological events in different seasons provided the ability to study the nitrate export behaviour predefined by seasonal meteorological settings and conditioned by the hydrological events observed. The results of the measurements show that seasonal hydrological and biogeochemical conditions play an important role in controlling the size of the forest soil nitrate pool which is available for further mobilization through hydrological mechanisms.

Continuous tracing of nitrate concentration in streamwater requires substantial effort and resources. Based on our continuous observations of nitrate export during 16 recorded floodwaves, a considerable amount of data was acquired. This paper aims at presenting the application of data mining (DM) and knowledge discovery from database (KDD) tool, namely regression or model trees, for gaining additional knowledge about the observed behaviour of seasonally and hydrologically influenced mobilization of nitrate during hydrological events and applying this knowledge to better understand the streamwater nitrate concentration behaviour.

2 Study area

The Padež stream watershed is situated in the southwestern part of Slovenia and comprises 42.1 km^2 (Fig. 1). The Padež stream is a tributary of the Reka river, one of the widest known sinking streams of the Classic Karst area in Slovenia; the Padež watershed reaches deeply into the hilly area of Brkini in the south (altitude up to 815 m a.s.l.). The studied area consists of Eocene flysch (mainly marl and sandstone layers) underlain by deep cretaceous carbonate bedrocks which also surround the wider area of the Brkini flysch pool. Spatially, the hydrogeological characteristics of the Padež watershed are uniform, characterized by the low permeability of erodible flysch layers and, consequently, a well developed, dense and highly incised stream channel network with a drainage density of 1.94 km/km^2. The lowest parts of the main valleys (the Padež and Suhorka stream valleys) are covered by up to 4-m thick alluvial deposits. The hydraulic conductivity of flysch is low (in the range 10^{-6} m/s to 10^{-5} m/s), the hillslopes are steep (average slope derived from the digital elevation model amounts to 33%), the average slope of the Padež stream channel being almost 3%. In 2006 the mean discharge of the Padež stream amounted 0.672 m^3/s; the long-term mean annual discharge of the Padež stream is 1.1 m^3/s. The hydrological response of the watershed is fast, which is reflected in the flushing, almost torrential regime of the Padež stream and short times to hydrograph peaks which can, in conditions of combined preceding wetting of the watershed and high rainfall intensities, vary between 2 and 3 h. For most of the year, stream water is present only in the Padež stream and its major tributary, the Suhorka stream, other smaller streams in the watershed being intermittent.

The Brkini hilly area is a climatic transitional area between the mediterranean and continental climate with a mean annual temperature of 9.6°C. The mean annual precipitation is approximately 1440 mm (Rusjan et al., 2006). The prevailing movement of the wet air masses is in the southwest-northeast direction. The majority of the precipitation falls during the October–March period with periodical snowfall

on the highest parts of the Brkini hills, which does not have substantial influence on the watershed hydrology. In 2006, the total annual precipitation amounted 1055 mm, more than 300 mm of rainfall was recorded during the intensive rainfall events of August 2006. The mean daily temperature in 2006 was 10.8°C.

Spatially, soils in the study area are uniform. According to the WRB 2006 soil classification they are classified as Haplic Cambisol (Humic, Hyperdystric, Endoskeletic); the hydraulic conductivity of clayish and silt soils is low (around 10^{-5} m/s).

The Padež stream watershed is minimally disturbed by human activity; it has already been used for drinking water supply and, as such, it is also foreseen as an additional source of drinking water for the water-deficient area of the Slovenian coastal region. According to the CORINE 2000 land cover data, 82% of the watershed is covered by forest (79% by broad-leaved forest), and 18% of the watershed comprises complex cultivation patterns (meadows with significant areas of natural vegetation), which discontinuously appear on top of the hills and are all in the state of successive afforestation. Few small settlements can be found on top of the hills and they are separated from the stream network by wide areas of forest. Steep hillslopes and narrow lower parts of the valleys are almost completely covered by deciduous forest; therefore the possible effect of anthropogenic disturbances on the stream chemistry was excluded. The main tree species that can be found in the Padež watershed are Sessile oak (*Quercus petraea*), Black alder (*Alnus glutinosa*), Beech (*Fagus sylvatica* L.), and Hornbeam (*Carpinus betulus*) (Slovenian Forest Service, 2000[1]).

3 Methods

3.1 Monitoring system

The monitoring system at the Padež watershed is shown in Fig. 1. Precipitation data were obtained from tipping bucket rain gauges located within the Padež watershed; the meteorological data were gathered from the automatic meteorological station positioned in the middle of the watershed (Fig. 1). Water level was recorded continuously with a 5-min time step on four locations using a 1-D Doppler instrument with an integrated logger. Flow was gauged on stream sections equipped with limnigraphs using two instruments. During low flow conditions, a salt-dilution flowmeter was used, whereas during middle to high flows, a 2-D/3-D handheld Doppler velocimeter was used. The resulting water-level records were converted to volumetric discharges by empirical ratings that were validated by gauging at different flow levels.

---

[1]Slovenian Forest Service: Silvicultural plans for the Kozina and Ilirska Bistrica local forest units (in Slovenian), unpublished documentation, 2000
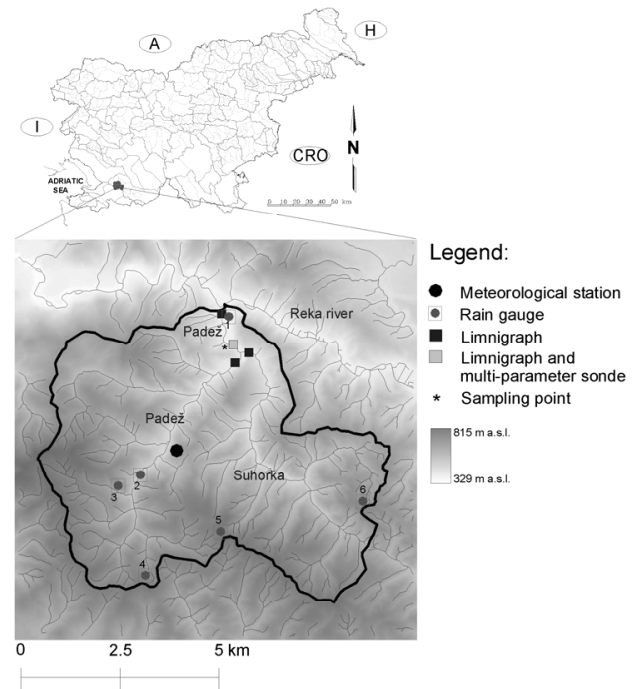

Fig. 1. The Padež stream watershed and the monitoring system.

Stream chemistry was measured continuously on a 30-min time step using a water quality multi-parameter data-sonde. The multi-parameter sonde is designed for on-site and flow-through applications and measures water chemistry parameters simultaneously (Brilly et al., 2006). The multiple parameters include: nitrate, temperature, electrical conductivity, depth, dissolved oxygen, Total Dissolved Solids (TDS), Oxidation Reduction Potential (ORP) and pH. Additionally, grab water samples were taken occasionally from January to November 2006 at the site where the multi-parameter sonde was installed for laboratory analysis in order to control the multi-parameter sonde readings. The samples were collected and preserved according to SIST EN ISO 5667-6 and SIST EN ISO 5667-3 standards, respectively. Nitrate was measured according to SIST EN ISO 10304-1 standard using the ion chromatograph. The differences in the nitrate concentration readings by the multi-parameter sonde and the laboratory analysis were not statistically significant (N=14; confidence level 0.95; p=0.12).

3.2 Regression trees

Regression trees as a subgroup of decision trees are a representation for piece-wise constant or piece-wise linear functions. Like classic regression equations, they predict the value of a dependent variable (called class) from the values of a set of independent variables (called attributes) (Džeroski, 2001; Witten and Frank, 2005). Regression trees are an especially attractive type of models for three main reasons.

Table 1. Attributes selected for the construction of regression trees.

Attribute	Description
NO_3	Nitrate concentration in streamwater [mg/l-N]
API_3	Antecedent precipitation index determined for 3 preceding days prior the day of the hydrograph peak occurrence [mm].
API_5	Antecedent precipitation index determined for 5 preceding days prior the day of the hydrograph peak occurrence [mm].
API_7	Antecedent precipitation index determined for 7 preceding days prior the day of the hydrograph peak occurrence [mm].
API_{14}	Antecedent precipitation index determined for 14 preceding days prior the day of the hydrograph peak occurrence [mm].
$Psum_{3h}$	Sum of rainfall during last 3 preceding h prior the occurrence of hydrograph peak [mm].
$Psum_{6h}$	Sum of rainfall during last 6 preceding h prior the occurrence of hydrograph peak [mm].
$Psum_{12h}$	Sum of rainfall during last 12 preceding h prior the occurrence of hydrograph peak [mm].
$Psum_{24h}$	Sum of rainfall during last 24 preceding h prior the occurrence of hydrograph peak [mm].
$Psum_{48h}$	Sum of rainfall during last 48 preceding h prior the occurrence of hydrograph peak [mm].
$avgT_1$	Average hourly air temperature during 1 preceding day prior the occurrence of the hydrograph peak [°C].
$avgT_3$	Average hourly air temperature during 3 preceding days prior the occurrence of the hydrograph peak [°C].
$avgT_7$	Average hourly air temperature during 7 preceding days prior the occurrence of the hydrograph peak [°C].
$avgT_{14}$	Average hourly air temperature during 14 preceding days prior the occurrence of the hydrograph peak [°C].
TW	Temperature of streamwater [°C].
Q	Discharge [m^3/s]
EW	Event water contribution to total discharge [-].

Firstly, they have an intuitive representation, the resulting model is easy to understand and assimilated by humans (Breiman et al., 1984). Secondly, the regression trees are nonparametric models, minimal intervention being required from the user, and thus they are very suited for data mining (DM) and knowledge discovery from database (KDD) (Džeroski, 2001; Atanasova and Kompare, 2002). Lastly, the accuracy of decision trees is comparable or superior to other models (Witten and Frank, 2005). Compared to neural networks, a more commonly used machine learning method in hydrological studies, regression trees can give a structural insight into the hydrological processes being modeled (Štravs and Brilly, 2007).

The regression tree construction proceeds recursively, starting with the entire set of training examples. At each step, the most discriminating attribute is selected automatically as the root of the (sub)tree and the current training set is split into subsets according to the values of the selected attribute. The split variable selection made by the algorithm is one of the main components of classification tree construction. The quality of the split selection criterion in tree nodes has a major impact on the quality (generalization, interpretability and accuracy) of the resulting tree. For regression trees, the automatically selected split is the one that maximizes the homogeneity of the two resulting groups with respect to the response variable (Prasad et al., 2006). In order to avoid the overfitting of the regression trees to the training data, many techniques, called pruning, have been proposed in the literature (Witten and Frank, 2005).

In our study of the nitrate flushing, a re-implementation of the well known regression tree induction algorithm M5 (Quinlan, 1992) within the software package WEKA (Wang and Witten, 1997; Witten and Frank, 2005) was used. Each leaf of the generated regression tree contains a linear regression equation which is used to model the dependant class inside the subset of instances classified to the particular leaf. The main intervention of the user in the application of the algorithm is the definition of the pruning factor – the threshold for the minimum number of instances which are classified into a particular regression tree leaf. The prediction accuracy of the constructed models was evaluated performing 10-fold cross-validation (Kohavi, 1995). In the 10-fold cross-validation, the dataset is randomly split into 10 disjoint subsets of approximately the same size, and 10 experiments are performed. In each of these, 1 of the 10 subsets is withheld, the prediction method trained on the union of the remaining 9, then tested on the unseen examples from the withheld subset. The reported accuracies are the averages of the 10 experiments.

The attribute selection considered in the data mining applications should be based on the domain knowledge of the modeled processes (Hall et al., 2002; Zaffron, 2005). The attributes which were considered in the dataset used for the construction of regression tree models of the streamwater nitrate concentration are listed in Table 1. The dataset consisted of 1257 records of attributes, which were temporally adjusted on an hourly time step containing 16 hydrographs, and every combination of attributes obtained at a certain time step represents an instance used in the data mining process.

Within a particular region or forest stand, mineralization and nitrification rates vary considerably in response to two key factors: temperature and moisture (Arheimer et al., 1996; Andersson and Lepisto, 1998; Welsch et al., 2001; Bernhardt et al., 2002; Vanderbilt et al., 2003). In temperate soils, mineralization displays a seasonal pattern; rates are greatest in summer or fall when soil temperatures are highest
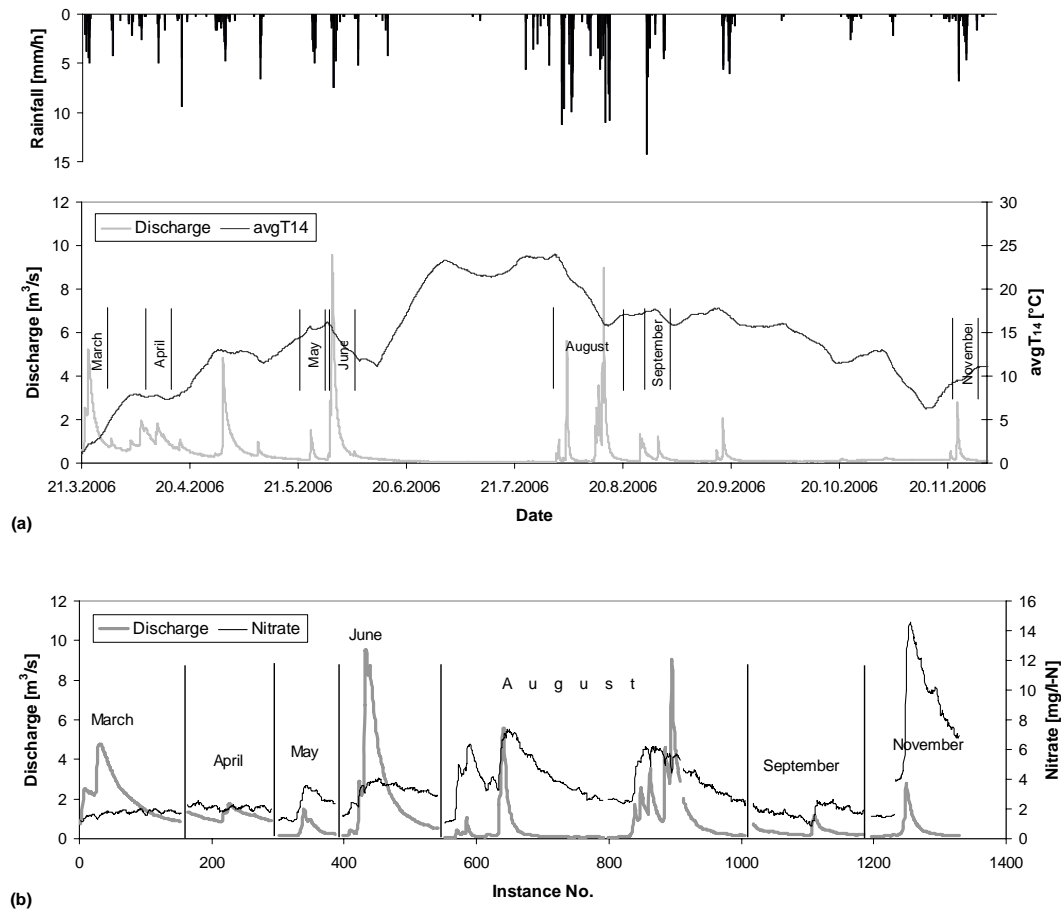
Fig. 2. (a) Recorded hydrological events inside the hydrological and seasonal temperature settings in period March–December 2006. **(b)** Hydrological events and associated streamwater nitrate concentrations included into the data mining process.

(Strader et al., 1989; Reich et al., 1997; Knoepp and Swank, 2002). Furthermore, Clark et al. (2004) used mean monthly air temperature to describe the mean monthly streamwater nitrate concentrations in forested watersheds, however, without a detailed consideration of the possible effect of changed hydrological conditions. The high temporal resolution of the detailed biogeochemical settings controlling the soil nitrogen transformations is almost impossible to obtain, however, we believe that the specific, spatially uniform hydrogeological and pedological settings at the Padež watershed enabled us to link the observed streamwater nitrate concentrations to the main driving factors of the nitrate formation considered by the given attribute selection. With particular attribute selections we therefore tried to capture the possible hydrological and seasonal biogeochemical characteristics which most likely play an important role in regulating streamwater nitrate responses during the observed hydrological events. In order to describe the preceding watershed wetness state we used the antecedent precipitation index (API_x) for a selected period of x preceding days using the method of Linsley et al. (1982), whereas the characteristics of the hydrological

events are captured through the sums of rainfall (Psum) for 3, 6, 12, 24 and 48 preceding hours and the observed discharges of the Padež stream. Additionally, the characteristics of the rainfall runoff formation are considered within the data about the proportion of event water in the total discharge (EW). The proportion of event water in the total discharge was obtained performing a two-component hydrograph separation using the electrical conductivity as a natural tracer. We measured the electrical conductivity of rainfall in a bulk sample and continuously in the streamwater. Seasonal biogeochemical implications in the sense of temperatures are considered through the data about average hourly air temperatures (avgT) for the periods of 1, 3, 7 and 14 preceding days and the streamwater temperatures (TW).

4 Results and discussion

Figure 2a shows the hydrological and seasonal air temperature settings in the period March–December 2006; Fig. 2b represents the recorded hydrological events and associated streamwater nitrate pulses which were included into the data
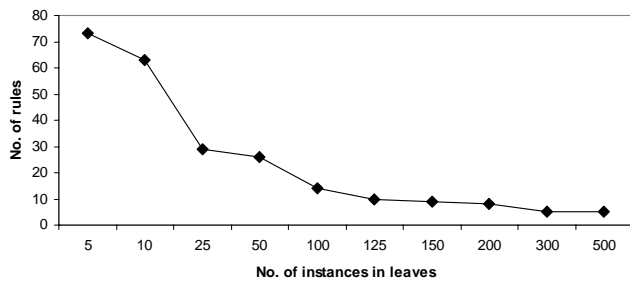
Fig. 3. Size of the generated regression trees performing 10-fold cross-validation method.

Fig. 4. Prediction accuracies of the generated regression trees performing 10-fold cross validation method.

mining process. The streamwater nitrate concentrations during the baseflow conditions were generally in the range of 1–1.5 mg/l-N. The measurements of the nitrate concentration in bulk rainfall samples during different episodes showed small concentrations of nitrate in rainfall (i.e. below 0.2 mg/l-N). Thus the wet deposition was not considered as an important source of stream water nitrate throughout the hydrological events. During the first two recorded hydrographs in early spring (March and April hydrographs), the streamwater nitrate concentrations showed no responsiveness to the changed hydrological conditions. The concentrations remained around 2 mg/l-N. Continuous measurements of the nitrate concentration during the early spring hydrographs and other dry periods disclosed a diurnal cycle of nitrate concentration oscillations with the appearance of maximum concentrations early in the morning and minimum concentrations late in the afternoon. The diurnal oscillations in the nitrate concentration are seasonally dependant and could be associated to the diurnal activity of aquatic photoautotrophs and surrounding, especially riparian forest vegetation, similarly to those described by Burns (1998). However, the diurnal cycle of the nitrate streamwater concentrations is not considered in detail in the paper as the main stress is on the nitrate streamwater concentrations during the rainfall events. The hydrological events in late spring and summer (observed hydrographs in May, June and especially August) expressed the strong influence of the changed hydrological state on the streamwater nitrate concentration increases of up to 7 mg/l-N. The greatest increase in the streamwater nitrate concentration was observed during the hydrological event in November, when streamwater nitrate concentrations exceeded 14 mg/l-N. The extreme rise of the streamwater nitrate concentration during the November hydrograph could be associated to the extremely dry (the total amount of precipitation in October and first half of November was only 19 mm) and warm (average daily temperatures in October around 15°C) preceding autumn period (Fig. 2a) coupled by the seasonally reduced N uptake by vegetation and additional input of nitrate through the litter decomposition. The possible additional source of nitrate to the stream during the observed rainfall events could be related to the presence of
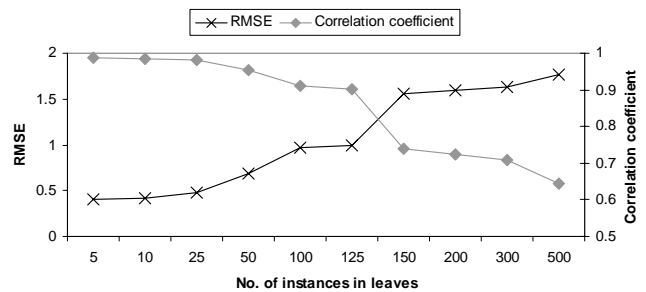
black alder in the riparian areas which is known because of the symbiotic relationship with a nitrogen-fixing bacterium *Frankia alni* (Zitzer et al., 1989; Cote and Camire, 2005).

4.1 Regression tree model

The size of the generated regression trees which depends on the predefined number of instances that reach a tree leaf as a pruning factor is shown in Fig. 3. The resulting performances of the trees in predicting the streamwater nitrate concentration expressed through RMSE and correlation coefficient r were obtained through a 10-fold cross validation and are given in Fig. 4. The regression trees with a small number of instances in the leaves are extremely big; in the case of only 5 instances in a leaf, the generated regression tree has 73 rules – linear regression equations in leaves. The performance of the trees with a large number of rules is suspiciously high (in the case of the tree with 73 rules: RMSE=0.40 mg/l-N and r=0.99).

However, a tree of such size is practically incomprehensive and very likely overfitted to the training data. In order to avoid the problem of overfitting and improve the comprehensiveness of the resulting regression trees, we have opted for a more drastic pruning of the regression tree by increasing the number of the instances in the leaves. Satisfactory prediction accuracies have been obtained generating regression trees with 100 and 125 instances in leaves which have 14 and 10 rules, respectively. If we further increased the number of instances in the leaves, the performance of the resulting regression trees decreased substantially (Fig. 4). The performance measures for the two selected regression trees are: in the case of a regression tree with 100 instances in leaves RMSE=0.97 mg/l-N and r=0.92; in the case of 125 instance regression tree RMSE=1.02 mg/l-N and r=0.91. The decrease in the performance of the two regression trees is relatively small whereas the size of the regression tree with 125 instances in leaves is additionally decreased (10 rules) vs. regression tree with 100 instances in the leaves (14 rules). We have therefore opted for the regression tree with 125 instances in the leaves (regression tree RT125). Figure 5 represents the structure of the resulting regression

tree. In Table 2, a list of linear regression equations which are included into the leaves of the resulting regression tree RT125 by performing 10-fold cross-validation of the dataset is given.

The splitting attribute selected in the root node of the resulting regression tree by the algorithm automatically is the antecedent precipitation index for the period of 5 preceding days API_5. According to the splitting value of the API_5=17.9 mm the two resulting branches interpret the states of high hydrological wetness of the watershed (values above 17.9 mm), whereas the branch defined by values of API_5 below 17.9 mm is used to describe the hydrologically less moist periods. On the second split level, $avgT_3$ (value 10.1°C) and $avgT_{14}$ (value 11.6°C) were selected by the algorithm to describe the seasonal character of the hydrological events. The result of the splitting on the first two split levels is four branches of the regression tree (Fig. 5). The split of the dataset into four branches according to the conditions imposed on the first two split levels of the regression tree is shown in Fig. 6. From the temporal point of view, branch 1 covers the data obtained during the hydrologically less moist, early spring period (March and April) and the short period before the occurrence of the November hydrograph. Branch 2 includes the data obtained during the rising limbs of the first hydrographs in the sequences of the late spring and summer hydrographs in May, June, August and September, whereas branch 4 covers the rest of the data obtained during the late spring and summer periods. Branch 3 comprehends the wettest periods during the November hydrograph and the March hydrograph.

On the lower, third split level (Fig. 5), we can find $Psum_{12h}$ and $Psum_{24h}$, which characterize the properties of the hydrological events in more detail. Attribute $avgT_1$ can be also found on the third split level which according to the particular split value 9.9°C draws a distinction between the data acquired during the seasonally and biogeochemically very contrasting periods, the March and November hydrographs.

On the fourth and fifth split levels $avgT_{14}$, $avgT_7$ and API_{14} are selected by the algorithm for further splitting. Interestingly, the resulting regression tree model does not include the data about discharge Q and streamwater temperature TW as a split attribute nor as an attribute included into the linear models in the leaves. However, the discharge is considered indirectly through the EW attribute, which appears in the linear models Nos. 8 to 10 (Table 2).

The measured streamwater nitrate concentrations vs. streamwater nitrate concentrations predicted by the regression tree model are shown in Fig. 7. The regression tree model successfully predicts low to medium nitrate concentration (1 mg/l-N to 4 mg/l-N). The accuracy of the model prediction decreases with an increase in the streamwater nitrate concentration. The regression tree model RT125 with 10 leaves is not adaptable enough to be able to more accurately predict high streamwater nitrate concentrations (above
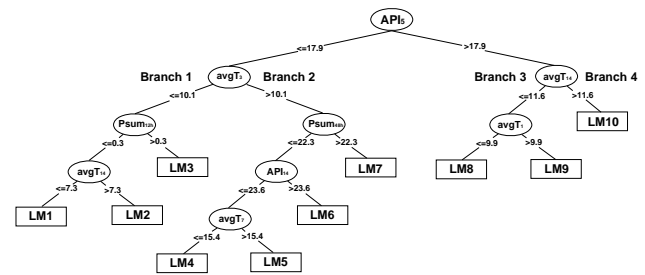

Fig. 5. Structural representation of the regression tree RT125.

5 mg/l-N concentrations) which occur only during short periods of hydrograph peaks (Fig. 2b). Furthermore, the model does not predict nitrate concentrations above 9 mg/l-N, while the measured concentrations during the November hydrograph peak discharges rose to 14 mg/l-N.

4.2 Interpretation of the model results in the light of domain knowledge

The structural transparency of the regression tree models offers additional opportunities to interpret not only the model results in the sense of performance but also model structure in the light of the domain knowledge of the modelled process. As the primary splitting attribute in the generated regression tree model, the antecedent precipitation index calculated for the period of 5 preceding days API_5 was chosen by the algorithm M5. The selection of the API as a primary split attribute was very likely imposed by its definition. The values of the API are defined on the daily time step, whereas other attributes were included into the dataset on the hourly time step. However, the time step definition of the API does not prevail over the further construction of the regression tree since on the second split level other attributes, namely average hourly air temperature during 3 ($avgT_3$) and 14 ($avgT_{14}$) preceding days, are selected automatically as splitting attributes.

The values of the antecedent precipitation indexes are defined empirically based on the selection of the recession constant and they help to simulate the drying of the watershed depending on the characteristics of the watershed. From the domain knowledge point of view, the exact values of the API do not offer the exact information about the hydrological nitrate mobilization process traced through continuous streamwater nitrate concentration measurements, however, they provide an insight into the temporal changes of the hydrological state of the Padež watershed responsible for nitrate flushing.

The values of the $avgT_3$=10.1°C and $avgT_{14}$=11.6°C are from the biogeochemical point of view much more interesting. Temperature together with soil moisture conditions significantly affects the mineralization and nitrification rates in forest soils (Arheimer et al., 1996; Andersson and Lepisto,

Table 2. Generated linear regression equations included into the leaves of the regression tree model RT125.

Linear regression equations
LM1: $NO_3 = + 0.0008 \times Psum_{24h} + 0.0006 \times Psum_{48h} + 0.0017 \times API_5 + 0.0017 \times API_{14} - 0.0008 \times avgT_3 + 0.0095 \times avgT_7 + 0.0017 \times avgT_{14} + 1.7664$
LM2: $NO_3 = + 0.0008 \times Psum_{24h} + 0.0006 \times Psum_{48h} + 0.0017 \times API_5 + 0.0017 \times API_{14} - 0.0008 \times avg_{T3} + 0.0095 \times avgT_7 + 0.0008 \times avgT_{14} + 1.9596$
LM3: $NO_3 = -0.0014 \times Psum_{12h} + 0.0008 \times Psum_{24h} + 0.0006 \times Psum_{48h} + 0.0017 \times API_5 + 0.0017 \times API_{14} - 0.0008 * avgT_3 + 0.0095 \times avgT_7 - 0.0029 \times avgT_{14} + 1.6483$
LM4: $NO_3 = 0.0016 \times Psum_{12h} + 0.0008 \times Psum_{24h} + 0.0025 \times Psum_{48h} + 0.0017 \times API_5 + 0.0071 \times API_{14} - 0.0006 \times avgT_3 - 0.008 \times avgT_7 - 0.0073 \times avgT_{14} + 2.1905$
LM5: $NO_3 = 0.0016 \times Psum_{12h} + 0.0008 \times Psum_{24h} + 0.0025 \times Psum_{48h} + 0.0017 \times API_5 + 0.0071 \times API_{14} - 0.0006 \times avgT_3 - 0.0068 \times avgT_7 - 0.0073 \times avgT_{14} + 1.6814$
LM6: $NO_3 = 0.0016 \times Psum_{12h} + 0.0008 \times Psum_{24h} + 0.0025 \times Psum_{48h} + 0.0017 \times API_5 + 0.0082 \times API_{14} - 0.0006 \times avgT_3 + 0.0082 \times avgT_7 - 0.0073 \times avgT_{14} + 2.3035$
LM7: $NO_3 = 0.0016 \times Psum_{12h} + 0.0008 \times Psum_{24h} + 0.0072 \times Psum_{48h} + 0.0017 \times API_5 + 0.0068 \times API_{14} - 0.0006 \times avgT_3 + 0.014 \times avgT_7 - 0.0073 \times avgT_{14} + 2.889$
LM8: $NO_3 = 0.2417 \times EW + 0.0004 \times Psum_{12h} + 0.0067 \times Psum_{24h} + 0.0181 \times API_5 - 0.0129 \times API_{14} + 0.2743 \times avgT_1 + 0.0013 \times avgT_3 + 0.0076 \times avgT_7 - 0.0277 \times avgT_{14} + 0.4859$
LM9: $NO_3 = 0.2417 \times EW + 0.0004 \times Psum_{12h} + 0.0067 \times Psum_{24h} + 0.0181 \times API_5 - 0.0129 \times API_{14} + 0.168 \times avgT_1 + 0.0013 \times avgT_3 + 0.0076 \times avgT_7 - 0.0277 \times avgT_{14} + 6.7974$
LM10: $NO_3 = 4.4079 \times EW + 0.0004 \times Psum_{12h} + 0.0183 \times Psum_{24h} - 0.0203 \times API_5 + 0.0587 \times API_{14} + 0.007 \times avgT_1 + 0.0013 \times avgT_3 + 0.0076 \times avgT_7 + 0.1929 \times avgT_{14} - 2.5505$
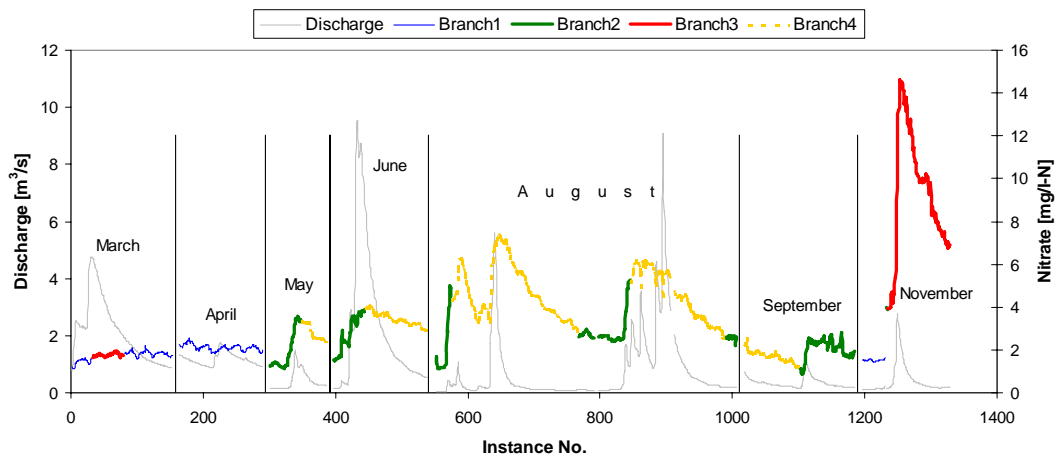
Fig. 6. Split of the dataset into four branches according to the conditions imposed on the first two split levels of the regression tree RT125.

1998; Knoepp and Swank, 2002). Whereas maximum mineralization is reported for the forest soil temperature range between 20 and 28°C (Nicolardot et al., 1994), the mineralization becomes strongly inhibited at temperatures below 10°C (Knoepp and Swank, 2002). Presuming that air temperatures indirectly represent the temperature of the forest soils, the values of $avgT_3 > 10.1$°C and $avgT_{14} > 11.6$°C, used for splitting in the regression tree model, define the periods of more effective mineralization, nitrification and nitrate accumulation in forest floors, which is in the periods of changed hydrological conditions available for hydrological mobilization. In Fig. 6 these conditions correspond to the periods

covered by branches Nos. 2 and 4. For further interpretation additional vegetation characteristics and detailed description of seasonal changes in soil characteristics would have to be considered. However, as in our case, such information is usually not available, also in the mechanistic approaches these characteristics are addressed differently. This is also beyond the mainly hydrological framework of the study.

Figure 8 shows the temporal performance of the regression tree model predicting streamwater nitrate concentrations against the measured streamwater nitrate concentrations during the observed hydrographs. The errors indicate the difference between the modelled and measured streamwater

nitrate concentrations. The regression tree model successfully predicts the streamwater nitrate concentrations during spring and summer hydrographs, when the error rarely exceeds 2 mg/l-N.

During the November hydrograph the concentrations of the nitrate were extremely high if compared to other observed hydrographs. Increased exports of the nitrate from forested watersheds are known to occur during seasonal transitions as a consequence of changed biochemical and hydrological conditions (Likens and Boremann, 1995). The high nitrate concentrations during the November hydrograph could be assigned to the extremely warm autumn period in 2006 (average daily air temperatures generally around 15°C) and absence of substantial rainfall in the early autumn period (September, October and first half of November), the biogeochemical and hydrological conditions being favourable for accumulation of nitrate in forest soils. Additionally, the vegetation nitrogen uptake is strongly reduced in autumn (Likens and Boremann, 1995; Beachtold et al., 2003). The streamwater nitrate concentrations are not satisfactorily described by the regression tree model during the November hydrograph as the error exceeds 4 mg/l-N. The changes in the nitrate concentrations occurred abruptly, therefore the regression tree model generated using the 125 instances in the leaf pruning criteria is not able to consider more accurately the extreme streamwater nitrate concentration changes during a single November hydrological event. In order to enable the regression tree to become more adaptive to the streamwater nitrate concentrations during the November hydrological and biogeochemical setting the pruning factor could be reduced – smaller number of instances in the leaves. This would result both, in the increased complexity of the regression tree model (Fig. 3) and slightly increased performance (Fig. 4), however such action usually leads to the increased risk of overfitting the algorithm to the extreme situations such as the one that occurred in November. Furthermore, the dataset included only one such extreme situation, therefore, to improve the regression tree algorithm performance, the dataset should be extended with more autumn and winter hydrological event observations, as it is the case in spring and summer rainfall events.

The data about the discharge Q and temperature of streamwater TW were not found to be important for the regression tree algorithm; the discharge is considered indirectly through the event water contribution to the total discharge (EW) in linear regression equations Nos. 8 to 10, which can be found under branches 3 and 4. These two branches cover the part of the dataset which can be characterized as hydrologically moist periods ($API_5 > 17.9$ mm) when the event water contribution to the total discharge can be substantial. It can be seen from the dataset (Fig. 2b) that beside some general positive relation between the discharge and streamwater nitrate concentration, similar hydrographs in the sense of the discharge peaks cause various nitrate concentration responses.
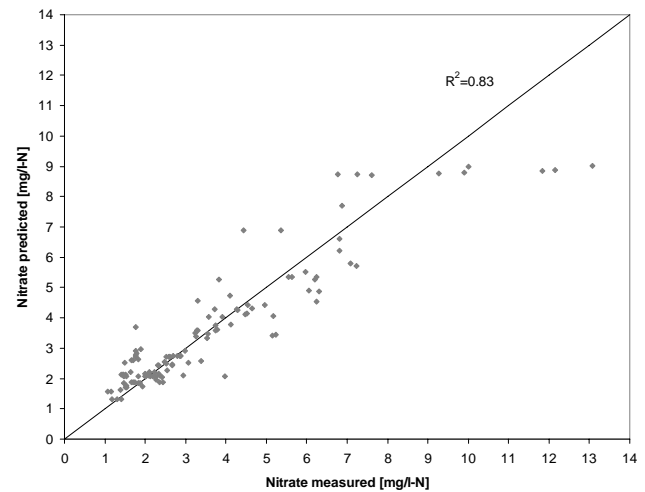

Fig. 7. Measured streamwater nitrate concentrations vs. predicted streamwater nitrate concentrations.

5 Conclusions

Regression trees proved to be a powerful and useful data mining tool in extracting additional knowledge from a given database, which helps to review and improve the existent domain knowledge about the mutual seasonal and hydrological controls of the streamwater nitrate pulses. Based on an extensive enlistment of attributes, which were expected to describe a general hydrological and seasonal biogeochemical framework of the forested watershed on the temporal scale of more than 50 days of hourly attribute collection, the regression tree generating algorithm successfully described complex streamwater nitrate concentration responses while enabling the conceptual explanation of the resulting regression tree structure. The regression tree model recognized the hydrological and seasonal patterns which lead the forested watershed from the states of being nitrate source limited (early spring hydrographs in March and April), in excess of hydrological mobilizing mechanisms, to the states when the availability of the hydrological mechanisms was exceeded by the size of the accumulated nitrate in the forested watershed (late spring, summer hydrographs and especially the autumn hydrograph). The regression tree model RT125 was obtained by automatic recognition of the associations between the attributes included into the dataset. The resulting model for the prediction of the streamwater nitrate concentration responses is derived from the data about the preceding hydrological and seasonal biogeochemical states of the forested watershed and is, by definition, empirical. However, the model structure can be explained by the domain knowledge of the modelled process, furthermore, the model structural transparency can be used to improve the conceptual understanding of the process from the event-based temporal point of view.

The possible limitations of the regression tree applications were also evident. Due to the extreme and fast streamwater
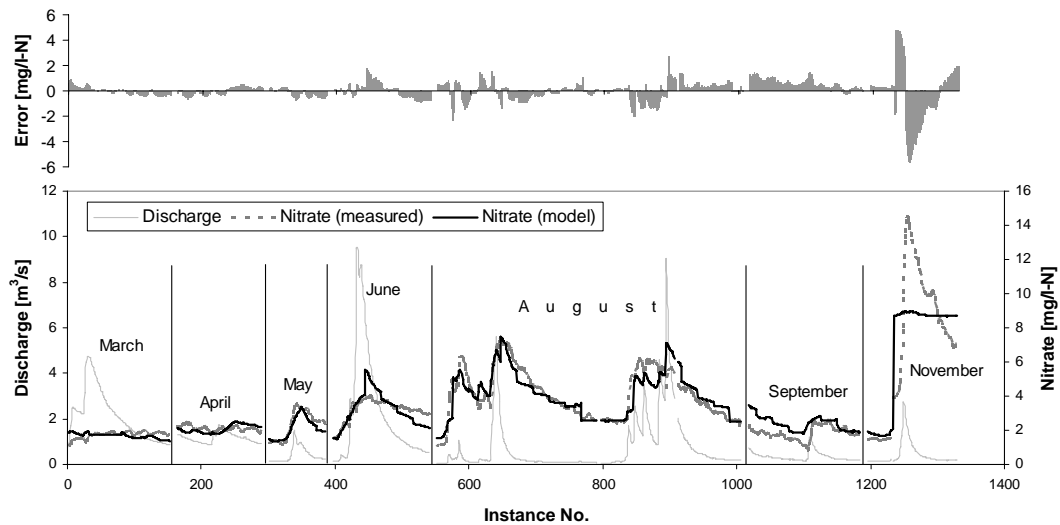
Fig. 8. Temporal performance of the regression tree model predictions.

nitrate response and the scarcity of the data in the autumn period (there was only one hydrograph in November) the resulting regression tree model was not able to adequately represent the streamwater nitrate concentrations during the November hydrograph with a given pruning threshold.

The question that remains to be addressed for future applications of the regression tree model is whether a given dataset really contains a range of "typical" hydrological and biogeochemical conditions. The model prediction would be improved by extending the dataset covering the sequences of rainfall events occurring in different hydrological and seasonal biogeochemical settings. These settings could be recognized by the regression tree models and used for the assessment of event-based nitrate flushing and the export of nitrate from a forested watershed on variable timescales. For future applications, it is worth considering combining the regression trees models able to describe the event-based streamwater nitrate pulses with conceptual and more process-based models, which are able to predict the nitrate export at lower temporal frequencies.

References

Aber, J. D., Ollinger, S. V., Driscoll, C. T., Likens, G. E., Holmes, R. T., Freuder, R. J., and Goodale, C. L.: Inorganic Nitrogen Losses from a Forested Ecosystem in Response to Physical, Chemical, Biotic, and Climatic Perturbations, Ecosystems, 5, 648–658, 2002.

Aber, J. D., Goodale, C. L., Ollinger, S. V., Smith, M., Magill, A. H., Martin, M. E., Hallett, R. A., and Stoddard, J. L.: Is Nitrogen Deposition Altering the Nitrogen Status of Northeastern Forests?, Bioscience, 53, 375–389, 2003.

Andersson, L. and Lepisto, H.: Links between runoff generation, climate and nitrate-N leaching from forested catchments, Water Air Soil Poll., 105, 227–237, 1998.

Arheimer, B., Andersson, L., and Lepisto, A.: Variation of nitrogen concentration in forest streams – influences of flow, seasonality and catchment characteristics, J. Hydrol., 179, 281–304, 1996.

Arnold, J. G., Allen, P. M.: Estimating hydrologic budgets for three Illinois watersheds, J. Hydrol., 176, 57–77, 1996.

Atanasova, N. and Kompare, B.: The use of decision trees in the modeling of a wastewater treatment plant, Acta Hydrotechnica, 20, 351–370, 2002.

Beachtold, J. C., Edwards, R. T., and Naiman, R. J.: Biotic versus hydrologic control over seasonal nitrate leaching in a floodplain forest, Biogeochemistry, 63, 53–72, 2003.

Beasley, D. B., Huggins, L. F., and Monke, E. J.: ANSWERS: A model for watershed planning, T. ASAE, 23, 938–944, 1980.

Bernhardt, E. S., Hall Jr., R. O., and Likens, G. E.: Whole-system Estimates of Nitrification and Nitrate Uptake in Streams of the Hubbard Brook Experimental Forest, Ecosystems, 5, 419–430, 2002.

Binkley, D., Ice, G. G., Kaye, J., and Williams, C. A.: Nitrogen and phosphorus concentrations in forest streams of the United States, J. Am. Water Resour. As., 40, 1277–1292, 2004.

Breiman, L., Freidman, J., Olshen, R., and Stone, C.: Classification and regression trees, Belmont (CA), Wadsworth, USA, 358 pp., 1984.

Brilly, M., Rusjan, S., and Vidmar, A.: Monitoring the impact of urbanisation on the Glinscica stream, Phys. Chem. Earth, 31, 1089–1096, 2006.

Burns, D. A.: Retention of NO3- in an upland stream environment: A mass balance approach, Biogeochemistry, 40, 73–96, 1998.

Cirmo, C. P. and McDonnell, J. J.: Linking the hydrologic and biogeochemical controls of nitrogen transport in near-stream zones of temperate-forested catchments: a review, J. Hydrol., 199, 88–120, 1997.

Clark, M. J., Cresser, M. S., Smart, R., Chapman. P. J., and Edwards, A. C.: The influence of catchment characteristics on the seasonality of carbon and nitrogen species concentrations in upland rivers of Northern Scotland, Biogeochemistry, 68, 1–19, 2004.

Cote, B. and Camire, C.: Nitrogen cycling in dense plantings of hybrid poplar and black alder, Plant Soil, 87, 195–208, 2005.

Creed, I. F., Band, L. E., Foster, N. W, Morisson, I. K., Nicolson, J. A., Semkin, R. S., and Jeffries, D. S.: Regulation of nitrate N release from temperate forests: A test of the N flushing hypothesis, Water Resour. Res., 32, 3337–3354, 1996.

Džeroski, S.: Applications of symbolic machine learning to ecological modeling, Ecol. Model., 146, 263–273, 2001.

Fenn, M. E., Poth, M. A., Aber, J. D., Baron, J. S., Bormann, B. T., Johnson, D. W., Lemly, A. D., McNulty, S. G., Ryan, D. F., and Stottlemyer, R.: Nitrogen excess in North American ecosystems: Predisposing factors, ecosystem responses, and management strategies, Ecol. Appl., 8, 706–733, 1998.

Fitzhugh, R. D., Lovett, G. M., and Venterea, R. T.: Biotic and abiotic immobilization of ammonium, nitrite, and nitrate in soils developed under different tree species in the Catskill Mountains, New York, USA, Glob. Change Biol., 9, 1591–1601, 2003.

Goodale, C. L., Aber, J. D., and McDowell, W. H.: The long-term effects of disturbance on organic and inorganic nitrogen export in the White Mountains, New Hampshire, Ecosystems, 3, 433–450, 2000.

Goodale, C. L., Lajhta, K., Nadelhoffer, K. J., Boyer, E. W., and Jaworski, N. A.: Forest nitrogen sinks in large eastern U. S. watersheds: estimates from forest inventory and an ecosystem model, Biogeochemistry, 57/58, 239–266, 2002.

Gundersen, P., Callesen, I., and de Vries, W.: Nitrate leaching in forest ecosystems is controlled by forest floor C/N ratio, Environ. Pollut., 102, 403–407, 1998.

Hall, M. J., Minns, A. W., and Ashrafuzzaman, A. K. M.: The application of data mining techniques for the regionalization of hydrological variables, Hydrol. Earth Syst. Sci., 6, 685–694, 2002, http://www.hydrol-earth-syst-sci.net/6/685/2002/.

Hornberger, G. M., Bencala, K. E., and McKnight, D. M.: Hydrological controls on dissolved organic carbon during snowmelt in the Snake River near Montezuma, Colorado, Biogeochemistry, 25, 147–165, 1994.

Kirchner, J. W., Feng, X., Neal, C., and Robson, A. J.: The fine structure of water-quality dynamics: the (high-frequency) wave of the future, Hydrol. Process., 18, 1353–1359, 2004.

Knoepp, J. D. and Swank, W. T.: Using soil temperature and moisture to predict forest soil nitrogen mineralization, Biol. Fert. Soils, 36, 177–182, 2002.

Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection, IJCAI-95, Morgan Kaufmann, San Mateo, 1137–1143, 1995.

Langusch, J. J. and Matzner, E.: N fluxes in two nitrogen saturated forested catchments in Germany: dynamics and modelling with INCA, Hydrol. Earth Syst. Sci., 6, 383–394, 2002, http://www.hydrol-earth-syst-sci.net/6/383/2002/.

Likens, G. E. and Bormann, F. H.: Biogeochemistry of a Forested Ecosystem, Second edition, Springer-Verlag, New York, 159 pp., 1995.

Linsley, R. K., Kohler, M. A., and Paulhus, L. H.: Hydrology for engineers, 3rd ed., McGraw Hill, New York, 512 pp., 1982.

Lovett, G. M., Weathers, K. C., and Arthur, M. A.: Control of nitrogen loss from forested watersheds by soil carbon: Nitrogen ratio and tree species composition, Ecosystems 5, 712–718, 2002.

McHale, M. R., Mitchell, M. J., McDonnell, J. J., and Cirmo, C. P.: Nitrogen solutes in an Adirondack forested watershed: Importance of dissolved organic nitrogen, Biogeochemistry, 48, 165–184, 2000.

McHale, M. R., McDonnell, J. J., Mitchell, M. J., and Cirmo, C. P.: A field-based study of soil water and groundwater nitrate release in an Adirondack forested watershed, Water Resour. Res., 38, 1031–1047, 2002.

Nicolardot, B., Fauvet, G., and Cheneby, D.: Carbon and nitrogen cycling through soil microbial biomass at various temperatures, Soil Biol. Biochem., 26, 253–261, 1994.

Pettersson, A., Arheimer, B., and Johansson, B.: Nitrogen concentrations simulated with HBV-N: new response function and calibration strategy, Nord. Hydrol., 32, 227–248, 2001.

Prasad, A. M., Iverson, L. R., and Liaw, A.: Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction, Ecosystems, 9, 181–199, 2006.

Quinlan, J. R.: Learning with continuous classes, in: Proceedings of the Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, 343–348, 1992.

Quinn, P. F., Hewett, C. J. M., and Dayawansa, N. D. K.: TOPCAT-NP: a minimum information requirement model for simulation of flow and nutrient transport from agricultural systems, Hydrol. Process., in press, 2007.

Reich, P. B., Grigal, D. F., Aber, J. D., and Gower, S. T.: Nitrogen mineralization and productivity in 50 hardwood and conifer stands on diverse soils, Ecology, 78, 335–347, 1997.

Rusjan, S., Brilly, M., Mikoš, M., Padežnik, M., and Vidmar, A.: Hydrologic and water chemistry characteristics of the Padež stream, in: Uncertainties in the monitoring-conceptualisation-modelling sequence of catchment research, edited by: Pfister, L., Matgen, P., van den Bos, R., and Hoffmann, L., Proceedings of the eleventh biennial conference of the euromediterranean network of experimental and representative basins (ERB), Luxemburg, CRP-Gabriel Lippmann, Belvaux, 19–22 September 2006, 212–215, 2006.

Sickman, J. O., Leydecker, A., Chang, C. C. Y., Kendall, C., Melack, J. M., Lucero, D. M., and Schimel, J.: Mechanisms underlying export of N from high-elevation catchments during seasonal transitions, Biogeochemistry, 64, 1–24, 2003.

Stieglitz, M., Shaman, J., McNamara, J., Engel, V., Shanley, J., and Kling, G. W.: An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport, Global Biogeochem. Cy., 17, DO1105, doi:10.1029/2003GB002041, 2003.

Stoddard, J. L.: Long term changes in watershed retention of nitrogen: Its causes and consequences, in: Environmental Chemistry of Lakes and Reservoirs, Am. Chem. Soc., Washington D. C., edited by: Baker, L.A., Adv. Chem. Ser., 237, 223–283, 1994.

Strader, R. H., Binkley, D., and Wells, C. G.: Nitrogen mineralization in high elevation forests of the Appalachians. I. Regional patterns in southern spruce-fir forests, Biogeochemistry, 7, 131–145, 1989.

Štravs, L. and Brilly, M.: Development of a low-flow forecasting model using M5 machine learning method, Hydrolog. Sci. J., 52, 466–478, 2007.

Vanderbilt, C. L., Lajhta, K., and Swanson, F. J.: Biogeochemistry of unpolluted forested watersheds in the Oregon Cascades: Temporal patterns of participation and stream nitrogene fluxes, Biogeochemistry, 62, 87–117, 2003.

Vitousek, P. M., Aber, J. D., Howarth, R. W., Likens, G. E., Matson, P. A., Schindler, D. W., Schlesinger, W. H., and Tilman, D. G.: Human alteration of the global nitrogen cycle: Sources and consequences, Ecol. Appl., 7, 737–750, 1997.

Wade, A. J., Neal, C., Butterfield, D., and Futter, M. N.: Assessing nitrogen dynamics in European ecosystems, integrating measurement and modeling: conclusion, Hydrol. Earth Syst. Sci., 8, 846–857, 2004,
http://www.hydrol-earth-syst-sci.net/8/846/2004/.

Wang, Y. and Witten, I. H.: Induction of model trees for predicting continuous classes, Proceedings of the poster papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague, 1997.

Weiler, M. and McDonnell, J. J.: Testing nutrient flushing hypotheses at the hillslope scale: A virtual experiment approach, J. Hydrol., 319, 339–356, 2006.

Welsch, D. L., Kroll, C. N., McDonnell, J. J., and Burns, D. A.: Topographic controls on the chemistry of subsurface stormflow, Hydrol. Process., 15, 1925–1938, 2001.

Witten, I. H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, 525 pp., 2005.

Worall, F., Swank, W. T., and Burt, T. P.: Changes in stream nitrate concentrations due to land management practices, ecological succession, and climate: Developing a systems approach to integrated catchment response, Water Resour. Res., 39, 1177–1191, 2003.

Young, R. A., Onstad, C. A., Bosch, D. D.: AGNPS: an agricultural nonpoint source model, in: Computer Models of Watershed Hydrology, edited by: Singh, V. P., WRP, Highlands Ranch, Colorado, USA, 1001–1020, 1995.

Zaffron, M.: Credible classification for environmental problems, Environ. Modell. Softw., 20, 1003–1012, 2005.

Zitzer, S. F., Dawson, J. O., and Gertner, G. Z.: Seasonal changes in nitrogen fixation activity of European black alder and Russian olive, Seventh Central Hardwood Forest Conference, Carbondale, Illinois, 134–140, 1989.