Contents lists available at ScienceDirect

Science of the Total Environment

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

Tree-based modeling of complex interactions of phosphorus loadings and environmental factors

S. Grunwald [a,*], S.H. Daroub [a,b], T.A. Lang [b], O.A. Diaz [c]

[a] Soil and Water Science Department, University of Florida, McCarty Hall 2169, Gainesville, Fl 32611, United States
[b] Everglades Research and Education Center, University of Florida, Belle Glade, Fl, 33430, United States
[c] South Florida Water Management District, West Palm Beach, Fl 33411, United States

ARTICLE INFO

Article history:
Received 14 December 2008
Received in revised form 22 February 2009
Accepted 23 February 2009
Available online 25 March 2009

Keywords:
Phosphorus
Phosphorus loads
Water quality
Regression trees
Committee trees
Everglades Agricultural Area

ABSTRACT

Phosphorus (P) enrichment has been observed in the historic oligotrophic Greater Everglades in Florida mainly due to P influx from upstream, agriculturally dominated, low relief drainage basins of the Everglades Agricultural Area (EAA). Our specific objectives were to: (1) investigate relationships between various environmental factors and P loads in 10 farm basins within the EAA, (2) identify those environmental factors that impart major effects on P loads using three different tree-based modeling approaches, and (3) evaluate predictive models to assess P loads. We assembled thirteen environmental variable sets for all 10 sub-basins characterizing water level management, cropping practices, soils, hydrology, and farm-specific properties. Drainage flow and P concentrations were measured at each sub-basin outlet from 1992–2002 and aggregated to derive monthly P loads. We used three different tree-based models including single regression trees (ST), committee trees in Bagging (CTb) and ARCing (CTa) modes and ten-fold cross-validation to test prediction performances. The monthly P loads (MPL) during the monitoring period showed a maximum of 2528 kg (mean: 103 kg) and maximum monthly unit area P loads (UAL) of 4.88 kg P ha^{-1} (mean: 0.16 kg P ha^{-1}). Our results suggest that hydrologic/water management properties are the major controlling variables to predict MPL and UAL in the EAA. Tree-based modeling was successful in identifying relationships between P loads and environmental predictor variables on 10 farms in the EAA indicated by high R^2 (>0.80) and low prediction errors. Committee trees in ARCing mode generated the best performing models to predict P loads and P loads per unit area. Tree-based models had the ability to analyze complex, non-linear relationships between P loads and multiple variables describing hydrologic/water management, cropping practices, soil and farm-specific properties within the EAA.

1. Introduction

The global phosphorus (P) cycle is simplified somewhat by the fact that there are no appreciable gaseous atmospheric P compounds, which are critical in the carbon and other biogeochemical cycles implicated in global climate change. The path of P from its release by chemical weathering to its transport and burial at sea is complex, because of the interaction of P with the biosphere and iron–manganese oxide particles in soils (Compton et al., 2000). It has been estimated that at global scale present-day P flux from rivers to the ocean is in the order of 17.7–30.4×10^{12} g yr^{-1}, which is about 2.5 times higher than in pre-human times (Compton et al., 2000). Land use shifts and management practices have caused P enrichment in naturally oligotrophic ecosystems. Large scale P enrichment has been documented in the Gulf of Mexico due to P flux from the Mississippi River Basin (2.9 million km^2) (Rabalais et al., 2002; Alexander et al., 2004) and in the Greater Everglades (~8250 km^2) due to P influx from upstream, agriculturally dominated, drainage basins (Everglades Agricultural Area – EAA) (Noe et al., 2001; Bruland et al., 2007; Grunwald et al., 2008). Phosphorus loads increased approximately 18 fold in the Chesapeake Bay and tributary estuaries since the pre-colonial period (Boesch et al., 2001). Changes in the relative proportions of nitrogen and P can exacerbate eutrophication, favor harmful algal blooms, aggravate oxygen depletion, alter food webs (Rabalais et al., 2002), and impact ecosystem functions, stability and resilience (Porter and Porter, 2001).

Alexander et al. (2004) pointed out that the estimated aquatic P removal rates declined with increasing stream size and rates of water flushing (i.e., areal hydraulic loads) using total phosphorus (TP) stream measurements from 336 watersheds in the U.S. This suggests that drainage basins, such as the ones in southern Florida, which are large in size with low relief, have longer residence times for P within the system. According to Heathwaite and Dils (2000) the magnitude and composition of the P load transported in surface and subsurface hydrological pathways depends on the discharge capacity of the flow route and the frequency with which pathways operate. They found that preferential flow pathways, particularly field drains, channels and

* Corresponding author. Tel.: +1 352 392 1951x204; fax: +1 352 392 3902.
E-mail address: sabgru@ufl.edu (S. Grunwald).

0048-9697/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2009.02.030

macropores, are important contributors to the overall P loads. Along these pathways most of the P is transported in the particulate fraction and associated with organic or colloidal P forms (Heathwaite and Dils, 2000). In south Florida's lowland drainage basins hydrologic manipulation has created an extensive system of such preferential pathways consisting of canals, ditches and water control structures. Diaz et al. (2006) reported that nutrient loading from the EAA and nearby urban communities as well as water flow rate and canal size have significantly influenced the amount of sediment and P pools stored in canals in Water Conservation Areas, Everglades, Florida. Phosphorus fractions associated with calcium and magnesium compounds and residual organic P were the dominant forms stored in the canal sediments. This suggests that more than 80% of the TP mass stored in surface sediments in these canals is fairly stable representing a long-term sink for P.

The effect of land use and management on TP loads has been well documented in studies of lowland river tributaries (Mander et al., 2000; Pieterse et al., 2003; Grunwald and Qi, 2006). These studies found positive correlations between agricultural coverage and intensity of land use and P loads exported from drainage basins. For example, most of the TP exported to the Gulf of Mexico from the Mississippi River with a TP mean of 0.22 kg ha^{-1} yr^{-1} (0.12–0.65 kg ha^{-1} yr^{-1}) originated in sub-basins where both agricultural and urban sources are large (Alexander et al., 2004). This has stimulated the implementation of best management practices (BMPs) and conservation measures to improve water quality (Sharpley et al., 2001; Daroub et al., in press). Despite these research findings our understanding of how different sub-basins with different land use characteristics and conservation measures affect P loads in lowland drainage basins is limited. Interactions between environmental landscape factors and P loads extending over multiple seasons are still poorly understood.

Commonly, multivariate regression models are used to characterize relationships between dependent and independent environmental attributes to explain their interactions or develop predictive models (Grunwald, 2006). Least square multivariate regressions are among the most commonly used analytical techniques in soil and water science applications (Trexler and Travis, 1993). However, classical regression methods are constrained due to assumptions about the statistical distribution of a response variable and the form of variance structure, which are often difficult to meet with environmental datasets (James et al., 2004). Tree-based modeling has been suggested to quantify relationships between soil, water and environmental landscape attributes, which are often complex, non-linear and show high-order interaction effects between environmental predictor attributes (DeAth and Fabricius, 2000). These methods include classification trees (ClT), regression trees (RT), and variants such as committee trees (CT) in Bagging and ARCing modes; the latter called boosted RT (bRT). Tree-based models are distribution free (non-

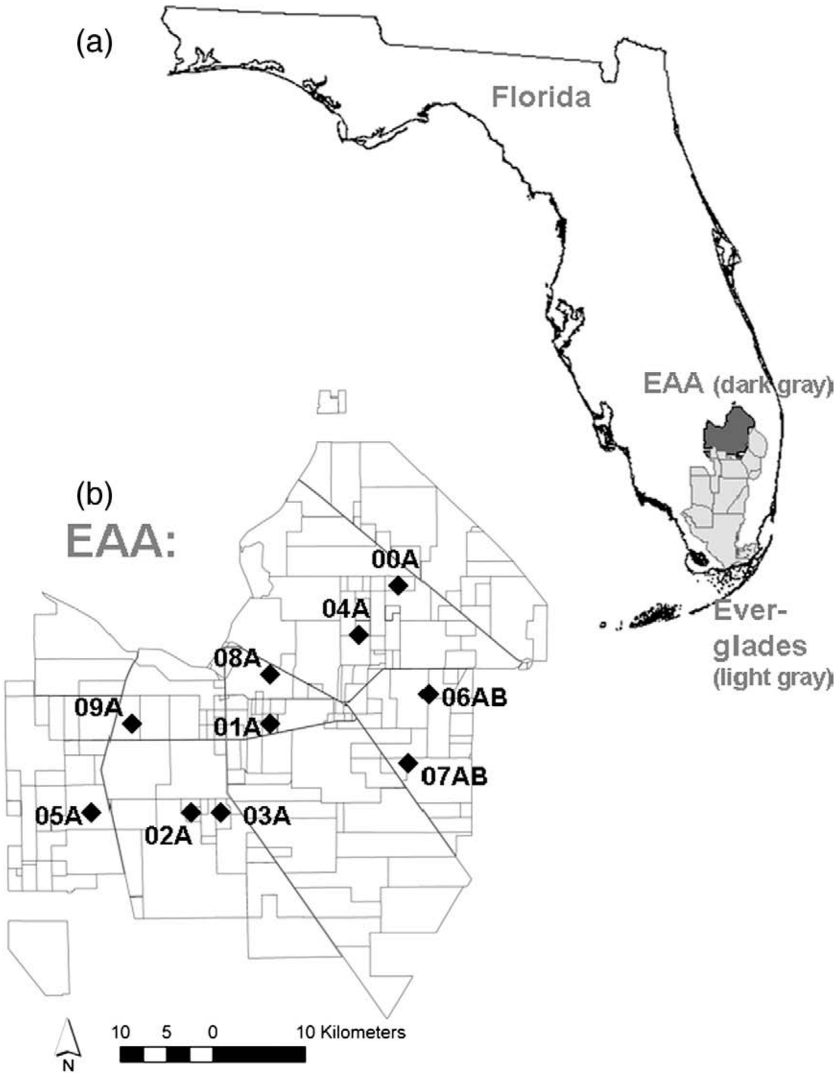

Fig. 1. (a) Everglades Agricultural Area (EAA) and Everglades located in the state of Florida; (b) Zoom-up of the EAA with ten farm basins.

parametric) and make no assumptions about regression variables or residuals (Breiman et al., 1984). The Classification and Regression Trees (CART) methodology is based on binary recursive partitioning; binary because parent nodes are always split into exactly two child nodes, and recursive because the process can be repeated by treating each child node as a parent node. Classification and Regression Trees use either categorical or continuous data types or both, which predict the data class (ClT) or the data values (RT). Optimized splitting rules are identified at each level of the tree. The goal of RT models is to partition the data into relatively homogenous (low deviation) terminal nodes, and the mean of the values in each node is the predicted value for that node. Tree-based modeling has been applied in various scientific disciplines to uncover hidden structures in complex datasets and to predict the characteristics of a chosen target variable by a set of meaningful predictor variables (Breiman et al., 1984). Trees have been used for ecological predictions and analysis (DeAth and Fabricius, 2000; Prasad et al., 2006), to quantify relationships between species and environmental landscape properties (DeAth, 2002; Moisen and Frescino, 2002), to delineate ecoregions (Hargrove and Hoffman, 2005), for environmental monitoring and epidemiology (Schröder, 2006), and in global change biological studies (Thuiller, 2003). In the soil science discipline RT, bRT and/or RF were used to predict soil organic carbon (Brown, 2007; Grimm et al., 2008; Vasques et al., 2008). Lilly et al. (2008) used RT to predict saturated hydraulic conductivity and Grinand et al. (2008) used CT to predict soil units based on environmental landscape properties. Tree-based modeling has been also used widely in the hydrology discipline in rainfall-runoff modeling (Solomatine and Dulal, 2003), to model water level–discharge relationships (Bhattacharya and Solomatine, 2005), to quantify relationships between snow water and environmental landscape properties (Mototch et al., 2005), to describe hydraulics (Pappenberger et al., 2006), to characterize water releases from various reservoirs in different time periods of the year (Reis et al., 2005), to investigate relationships between stream survey data and agricultural riparian buffers (Barker et al., 2006), to assess landscape conditions relative to water resources (Jones et al., 2000), and to predict P delivery to water bodies from agricultural land (Brazier et al., 2006). In this paper we explore single and committee RT to investigate complex interactions and effects of environmental basin attributes on P loads in various lowland agricultural-used drainage basins. Our specific objectives were to: (1) investigate relationships between various environmental factors and P loads, (2) identify those environmental factors that imparted major effects on P loads using three different tree-based modeling approaches, and (3) evaluate predictive models to assess P loads. The analyses were conducted in spatially-distributed sub-basins nested within an agricultural lowland drainage basin in south Florida.

2. Methods

2.1. Study area

The EAA basin (size 283,000 ha) (Fig. 1), located in south Florida, has been in agricultural use since 1948 and is divided hydrologically into four sub-basins, S5A, S6, S7, and S8 (compare Daroub et al., in press). It encompasses about 27% of the historic Everglades and is characterized by deep muck soils that have been undergoing subsidence at an annual average rate of 1.4 cm during the last 19 years due to organic matter oxidation (Shih et al., 1998). The Histosols (Suborder: Saprist) of the EAA have a soil organic matter content between 80 and 90% that is highly decomposed (Snyder, 1994). Soil series found in the EAA include Dania, Lauderhill, Pahokee, and Terra Ceia, which mainly differ in depth of the O horizon to the limestone bedrock (Rice et al., 2002a), and which decline from more than 1 m (S5A and S6 sub-basins) to less than 1 m (S7 and S8) (Table 1). The EAA is dissected by an extensive system of canals, ditches and water control structures to control seepage irrigation and drainage. Climate is sub-tropical with an average precipitation of 1270 mm yr^{-1}. The annual distribution of the rainfall is however uneven with 66% occurring during the months of June through October (Ali et al., 2000). Land area converted to sugarcane farms increased dramatically after the Cuban Revolution of 1959 causing nutrient enrichment in adjacent ecosystems (Lake Okeechobee and Greater Everglades) (Porter and Porter, 2001). Currently nearly 70% of the EAA is planted to sugarcane with lesser coverage of vegetables, sod, and rice (Rice et al., 2002b). U.S. Sugar Corp., the largest sugarcane producer in the nation, has sold about 74,800 ha of sugarcane land in the EAA, which will be converted into wetlands over the next 6 years aiming to accelerate recovery of the Everglades ecosystem (Stockstad, 2008). Growers in the EAA are required to implement a suite of BMPs and conduct monitoring of daily rainfall, drainage water volume and drainage water P concentrations. Growers choose BMPs from a list that has four main categories: (1) soil testing and application of P fertilizer according to a calibrated soil test, (2) controlled P fertilizer application methods, (3) water management practices, and (4) sediment source and transport controls. Details about BMP implementation on various farms in the EAA can be found in Daroub et al. (2004, in press). Topography across the EAA is nearly flat with minor elevation changes (in the cm range) leading to slow movement of water within fields and in ditches and canals.

Table 1
Characteristics of farm drainage basins in the Everglades Agricultural Area.

UF farm basin	Monitoring duration (months)	Sub-basin	Irrigation water structure/ canal[a]	Crops	Farm size (ha)	Average soil depth (m)	Rainfall detention (mm)	Soil series[b]
00A	118[c]	S5A	S352 WPB	Sugarcane	518	1.16	25.4	Pahokee
01A	90[d]	S6	S2 HB	Mixed[e]	518	0.61	12.7	Lauderhill
02A	118	S7	S2 NNR	Sugarcane	130	0.46	25.4	Dania
03A	118	S7	S2 NNR	Sugarcane	1865	0.43	12.7	Dania
04A	118	S6	S2 HB	Sugarcane	259	1.62	25.4	Terra Ceia
05A	90	S8	S3 Miami	Mixed	130	0.55	25.4	Lauderhill
06AB	118	S5A	S352 WPB	Mixed	710	0.88	12.7	Lauderhill
07AB	118	S6	S2 HB	Mixed	1012	0.98	25.4	Pahokee
08A	110[f]	S6	S2 HB	Sugarcane	106	0.73	25.4	Lauderhill
09A	118	S8	S3 Miami	Sugarcane	1243	0.98	25.4	Pahokee

[a] Canals being serviced by each irrigation structure: WPB West Palm Beach canal; HB = Hillsboro canal; NNR = North New River canal.
[b] Soil taxonomic descriptions: Dania – Euic, hyperthermic, shallow Lithic Haplosaprists; Lauderhill and Pahokee – Euic, hyperthermic Lithic Haplosaprists; Terra Ceia – Euic, hyperthermic Typic Haplosaprists.
[c] July 1992 to April 2002.
[d] July 1992 to December 1999.
[e] Mixed: Sugarcane, vegetables, sods, melons or rice.
[f] July 1992 to August 2001.

2.2. Environmental dataset

The following environmental variables for each farm drainage basin were used as predictor variables (Tables 1 and 2): (1) Farm size (ha) [farmsize], (2) Irrigation demand (cm) [irrdemand], (3) Irrigation P concentration (mg P L^{-1}) [irrpconc], (4) Monthly inside canal water level (meter a.m.s.l.) [monthlyinsidehead], (5) Monthly outside canal water level (meter a.m.s.l.) [monthlyoutsidehead], (6) Monthly rainfall (mm) [monthlyrain], (7) Percent sugarcane coverage (%) [percentcane], (8) Percent fallow and flood coverage (%) [percentfallflood], (9) Percent flood coverage (%) [percentflood], (10) Ratio between drainage to rainfall (mm mm^{-1}) [pumptorain], (11) Soil depth (m) [soildepth], (12) Soil type (Soil Series: Dania, Lauderhill, Pahokee, and Terra Ceia) [soils], (13) Location (farm basin codes: 00A to 09A) [location], and (14) Sub-basin (sub-basin codes: S5A, S6, S7, and S8) [sub-basin]. Water levels in farm canals and in surrounding conveyance canals can potentially indicate conditions conducive for seepage both into and off of a farm. Pressure transducers in canals were calibrated to measure canal water level as meters above mean sea level (a.m.s.l.). Canal water level was measured by a data logger on five minute intervals during drainage pumping and hourly otherwise, which were aggregated to derive monthly values for inside and outside canal water levels. Monthlyinsidehead was the average canal level adjusted to a.m.s.l. inside the farm at the farm exit drainage pump station. Outside canal level (monthlyoutsidehead) was the average canal level in the receiving canal outside the farm in close proximity to the main farm drainage pump station. Rainfall at each farm was measured in hourly intervals by tipping a bucket rain gauge located at the exit pump station(s) on each farm, which was aggregated to derive monthly rainfall totals. Irrigation water data were derived from the South Florida Water Management District-DBHYDRO database. Irrigation demand from pan evaporation was calculated as the difference of monthly evaporation totals from the Institute of Food and Agriculture Science (IFAS) – Belle Glade weather station minus farm monthly rainfall. Irrigation demand for sugarcane farms was estimated on a monthly basis from three components, farm monthly rainfall, monthly pan evaporation totals from IFAS Belle Glade weather station, and Penman sugarcane crop coefficients by month for evapotranspiration (compare Daroub et al., 2004). Average farm soil depth values were calculated from field measurements taken in 1995. Multiple soil depth measurements (six to ten) from within each sampled field were averaged to obtain a field soil depth average. 4 to 24 fields per farm were sampled and an overall farm soil depth average was determined from the field soil depth averages.

The descriptive statistics of environmental variables can be found in Table 2. Irrigation demand had a mean of 0.43 cm, median of 0.92 cm and a large range of 46.75 cm over the whole monitoring period. The measured P concentrations ranged from 0.02 to 0.39 mg P L^{-1} with mean and median of about 0.12 mg P L^{-1}. Monthly outside canal water levels showed a 1.3 times higher mean than monthly inside canal water levels with a modest range compared to other environmental variables. Monthly rainfall showed a mean of 122.5 mm and a median of 107.2 mm with a large range of 399.8 mm across the monitoring period. Percentages of sugarcane, fallow and flood coverage, and flood coverage showed large ranges among all farm basins. The ratio between pumping to rainfall showed a slightly skewed distribution with a mean of 0.60 mm mm^{-1}, median of 0.41 mm mm^{-1} and range of 8.54 mm mm^{-1}. The soil depth minimum was 0.43 and maximum was 1.62 m.

2.3. Drainage and water quality monitoring

The drainage systems in the EAA feature extensive networks of farm canals, ditches, and pump stations that are managed by both the South Florida Water Management District (SFWMD) and growers. The SFWMD manages the public canals and its own pump stations in the EAA while growers manage water levels on their farm drainage basins within the EAA basin. Normally, a main farm canal runs from the farm pump station to the far reaches of the farm, and sub-mains or farm laterals branch off the main canal at right angles, generally on 800 m spacing, on section and half section boundaries. Emanating at right angles from the farm laterals are equally spaced field ditches, which are parallel and subdivide the farm into rectangular areas with nominal dimensions of 200 by 800 m. These 16 ha blocks are considered the basin water management unit where sub-irrigation or open ditch drainage practices are accomplished by either raising or lowering field ditch water levels (Izuno and Bottcher, 1994). The seepage-based drainage control systems are managed to maintain the water level normally <1 m below the soil surface (Obreza et al., 1998).

We used a comprehensive water quality monitoring set which was focused on ten farm drainage basins covering a sampling period from 1992 to 1999 and 2002 (Daroub et al., 2004). The farm basins were selected to represent a typical range of farm sizes, soil types, crop rotations, water management, and geographical distribution across the EAA (Table 1). Six of the farm drainage basins (00A, 02A, 03A, 04A, 08A, and 09A) were dominated by sugarcane with more than 85% of the farm planted to sugarcane in most years. One exception was farm basin 08A which switched from being predominantly sugarcane culture from 1992–1999 to 55% and 29% of the area planted to vegetables in 2000 and 2001, respectively. The remaining four farms had mixed-cropping systems: farm 01A was strictly a vegetable monoculture; farm 05A was planted with sugarcane, sod, and melons; farm 06A/B was planted with sugarcane, vegetables, rice, sod and trees; and farm 07A/B grew sugarcane, vegetables, rice and sod. For mixed crop farms, vegetables are grown in the winter, and the fields in the summer are normally planted with flooded rice or kept flooded if left fallow.

Best management practices were similar for the 10 farms with minor differences for rainfall detention amount (Table 1). Each site's drainage flow was determined using a Campbell Scientific® CR-10 data logger that was programmed with the site's calibrated pump flow equations and was wired to upstream and downstream pressure

Table 2
Descriptive statistics of environmental variables used to predict monthly and unit area phosphorus loads on 10 farm basins in the Everglades Agricultural Area.

	Irrigation demand (cm)	Irrigation phosphorus conc. (mg P L^{-1})	Monthly inside canal water level (m a.m.s.l.)	Monthly outside canal water level (m a.m.s.l.)	Monthly rainfall (mm)	Percent sugarcane coverage (%)	Percent fallow and flood coverage (%)	Percent flood coverage (%)	Ratio between pumping to rainfall (mm mm^{-1})	Soil depth (m)
Mean	0.43	0.12	2.43	3.15	122.5	72.6	1.4	0.09	0.60	0.85
Std. error of mean	0.27	0.002	0.01	0.01	2.9	1.11	0.3	0.007	0.03	0.01
Median	0.92	0.11	2.44	3.19	107.2	86.7	0.0	1.0	0.41	0.88
Standard deviation	7.58	0.06	0.42	3.13	33.3	31.6	8.5	18.9	0.85	0.34
Skewness	−0.38	1.16	−0.03	−1.1	0.6	−1.1	8.2	2.9	5.4	0.2
Kurtosis	0.05	1.86	1.35	5.4	−0.4	−0.1	78.6	9.3	37.9	0.009
Range	46.75	0.37	2.85	4.92	399.8	100.0	100.0	100.0	8.54	1.19
Minimum	−26.37	0.02	1.06	0.63	0.0	0.0	0.0	0.0	0.00	0.43
Maximum	20.38	0.39	3.90	5.55	399.8	100.0	100.0	100.0	8.54	1.62

transducers and to drainage pump revs (pulses) per minute (RPM) sensors (shaft encoders). The data logger was also connected to a tipping bucket rain gauge (Texas Electronics® TE525-WSL) to measure rainfall and a water sampler (ISCO® 3700) that was triggered by the data logger to collect a water sample once the requisite amount of drainage flow had been pumped. The water samplers collected 100 ml volume for each sampling and composited the samples into a 4 L pre-acidified bottle. Sample bottles were collected after drainage pumping cessation or once the bottles became full. Water samples were transported, stored, and analyzed according to Everglades Research and Education Center (EREC) laboratory procedures (Chen, 2001). All drainage water samples were analyzed for TP. Samples were digested using the mercury oxide digestion with a block digester AS-4020 (Scientific Instruments Services, Inc., Ringoes, NC) (Method 365.4, USEPA, 1983). After digestion, solutions were analyzed using a Flow IV segmented flow analyzer (OI Analytical, College Station, TX) using the ascorbic acid method (Murphy and Riley, 1962). From the monitoring dataset daily loads were calculated from daily drainage flows and P concentrations that were then aggregated to monthly P loads [MPL] in kg and monthly unit area P loads [UAL] in kg P ha^{-1} ($n = 809$ records).

2.4. Analysis

We used CART methodology developed by Breiman et al. (1984) in three different modes: (i) Single RT (ST), (ii) CT with Bagging (CTb), and (iii) CT with ARCing (CTa) implemented in CART 5.0 software (Salford Systems, San Diego, CA). The environmental basin factors were used as inputs (independent variables) to predict target variables (MPL and UAL, respectively). Tree structure classifiers are constructed by repeated splitting of the set of observations (parent nodes) into two descendent subsets (child nodes). The splitting is continued as long as the child nodes become purer compared to the parent node. The entire process of tree construction revolves around three elements: (1) the selection of the splits, (2) the decision when to declare a node terminal or to continue splitting it, and (3) the assignment of each terminal node to a class. The aim of splitting is to increase child node purity when compared to parent node purity. We used the least squares splitting rule. Prime splitter and surrogate variables are included in trees, the latter ones to model high-order interaction effects among predictor variables.

Trees are grown until a maximum tree is reached depending on user specified restrictions or until further splitting is impossible. We set a minimum of three cases per terminal node for all trees. Fully grown trees are typically over-fitted to the training dataset and usually perform poorly when generalized to external data. Thus, we pruned trees to an optimal size to produce parsimonious trees in ST, CTb, and CTa modes. According to suggestions provided by Breiman et al. (1984) we selected the optimized trees using the minimum cost tree regardless of size assessed by the cross-validated relative error (CVRE). Ten-fold cross-validation was used to test prediction performances, which involves randomly dividing the data into partitions or folds. At each step, nine of these partitions are used to fit the model and the performance is assessed on the remaining partition held back as the test data. The procedure is repeated for each partition sequentially. The performance, averaged over all ten partitions held back generates the cross-validation performance assessment.

The selection of the splits is done by an exhaustive search on all possible variables, and for each variable, all possible splitting thresholds. The choice of the best splitter is based on a measure of accuracy. A least squares estimator is used and accuracy is assessed using the *resubstitution estimate* ($R(d)$) Eq. (1), which is the counterpart of the well-known mean squared error. The value that minimizes the resubstitution estimate at every tree node t is the average of the target variable ($\bar{y}_t$) for all cases falling into t. For every intermediate node t, a resubstitution estimate can be defined more specifically as $R(t)$ Eq. (2). The best split is then defined as the one which maximizes the decrease in $R(t)$ from the parent to the child nodes Eq. (3) (Breiman et al., 1984).

$$R(d) = \frac{1}{n}\sum_{i=1}^{n} [y_i - d(x_i)]^2 \tag{1}$$

$$R(t) = \frac{1}{n}\sum_{i}^{n} (y_i - \bar{y}_t)^2 \tag{2}$$

$$\Delta R(s,t) = R(t) - R(t_\mathrm{L}) - R(t_\mathrm{R}) \tag{3}$$

where: $R(d)$ = resubstitution estimate; n = number of observations; y_i = dependent variable at observation i; $d(x)_i$ = selected independent variables; x_i = threshold values for variable $d(x)_i$; $R(t)$ = minimum resubstitution estimate at node t, given that $d(x)_i = \bar{y}_t$; $\bar{y}_t$ = average of the target variable for all cases falling into t; $\Delta R(s,t)$ = decrease in $R(t)$ from the parent to the child nodes; s = split at node t; $R(t_\mathrm{L})$ = resubstitution estimate at the left child node; $R(t_\mathrm{R})$ = resubstitution estimate at the right child node.

At each node, after the best variable and threshold are selected, the dataset is split as following: For each sample in the training dataset, if the value of the selected variable exceeds the threshold, the sample is moved to the right (t_R); if not, the sample is moved to the left (t_L). Thus, at each node, the dataset is split into two new subsets. The partitioning continues at every node until the number of cases in the node reaches a predetermined minimum. The node then becomes a terminal node, and the average value for all cases falling into that node ($\bar{y}_t$) is assigned as the predicted value. In the process of minimizing $R(d)$ at each node, the overall tree resubstitution error ($R(T)$) is also minimized. Thus, a tree with minimum within-node error is found. This tree, however, is over-fitted to the training dataset and usually performs poorly when generalizing to external data. The tree can be pruned to an optimal tree size using either cross-validation or an independent validation dataset. The selection of the optimal tree is done by finding, among K different pruned trees, the one with minimum relative error ($REE(T_K)$) with respect to the validation data, according to Eqs. (4)–(6).

$$R(T) = \frac{1}{n}\sum_{j=1}^{\tilde{T}} R(t)_j \tag{4}$$

$$R(\bar{y}) = \frac{1}{n}\sum_{i}^{n} (y_i - \bar{y})^2 \tag{5}$$

$$REE(T_K) = R(T_K) \,/\, R(\bar{y}) \tag{6}$$

where: $R(T)$ = overall tree resubstitution estimate; n = number of observations; T = total number of terminal nodes of the tree; $R(t)$ = minimum resubstitution estimate at node j; $R(\bar{y})$ = average squared deviation of y around $\bar{y}$; y_i = dependent variable at observation i; $\bar{y}$ = average of y, for $i = 1$ to n; $REE(T_K)$ = relative mean squared error of a tree among K different pruned trees; $R(T_K)$ = overall resubstitution estimate of a tree among K different pruned trees.

Traditional CART modeling has been focused to model single trees fitting input to output variables (Steinberg and Colla, 1997). Committee trees assemble hundreds of single trees to predict a variable of interest (Prasad et al., 2006). In CART 5.0, bootstrap resampling is applied in a novel way, i.e., a separate analysis is conducted for each resample or replication generated and then the results are averaged. If the separate analyses differ considerably from each other (suggesting tree instability), averaging stabilizes the results, yielding much more accurate predictions. If the separate analyses are very similar to each other, the trees exhibit stability and the averaging will neither harm nor improve the predictions. Thus, the more unstable the trees, the greater the benefits of averaging (CART,

Table 3
Descriptive statistics of monthly phosphorus loads (kg) on each of the 10 farm basins in the Everglades Agricultural Area.

	Farm drainage basins										
	All 10 farm basins	00A	01A	02A	03A	04A	05A	06AB	07AB	08A	09A
Mean	103.1	94.3	357.1	7.3	71.3	18.3	15.8	222.2	159.3	3.3	50.4
Std. error of mean	8.5	18.5	66.9	1.6	9.1	2.6	2.7	34.4	20.8	0.4	5.2
Median	31.2	43.9	124.3	3.3	35.2	9.4	7.7	108.7	107.4	2.6	36.7
Standard deviation	240.4	164.0	543.4	11.5	90.9	22.2	23.4	347.5	196.0	3.2	53.7
Skewness	5.8	4.2	2.5	3.7	2.4	2.0	2.9	3.7	3.2	2.2	2.5
Kurtosis	42.5	20.6	6.3	17.4	7.1	5.1	10.6	16.8	15.5	6.7	8.9
Range	2527.9	1098.7	2526.0	70.1	512.8	114.2	133.0	2385.9	1351.0	17.6	336.1
Minimum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Maximum	2527.9	1098.8	2527.9	70.1	513.2	114.2	133.0	2386.3	1351.6	17.6	336.1

2002). Bagging and ARCing methods (Breiman, 1996; Freund and Schapire, 2000) were introduced to tree modeling to simulate multiple trees. These methods are called committee trees that are implemented in the form of bootstrap aggregation (Bagging) or ARCing (Boosting), a set of trees is generated by resampling with replacement from the original training data. The trees are then combined by averaging their outputs (RT mode). The key difference between Bagging and ARCing is the way each new resample is drawn. In Bagging, each new resample is drawn in an identical way (independent samples), while in ARCing the way a new sample is drawn for the next tree depends on the performance of the prior tree (CART, 2002). In ARCing the probability with which a case is selected for the next training set is not constant and is not equal for all cases in the original learn data set. Instead, the probability of selection increases with the frequency with which a case has been misclassified in previous trees. Cases that are difficult to classify receive an increasing probability of selection, while cases that are classified correctly receive declining weights from resample to resample. Multiple versions of the predictor are formed by making bootstrap replicates of the learning set and using these as new learning sets. The predicted value generated by the committee trees is an average over these multiple versions of predictors. We generated 250 trees in Bagging and ARCing modes, respectively.

3. Results

3.1. Monthly phosphorus loads and monthly unit area phosphorus loads

Tables 3 and 4 provide an overview of MPL and UAL distributions within farm basins over the whole sampling period. The mean MPL was highest in farm basin 01A with 357 kg, followed by farm basins 06AB (222 kg), 07AB (159 kg), 00A (94 kg) and 03A (71 kg), and 09A (50 kg). Phosphorus UAL showed the largest mean in farm basin 01A with 0.69 kg P ha^{-1}, followed by farm basins 06AB, 00A, 07AB and 05A. Since most of the distributions of MPL are skewed, as indicated by the skewness coefficients, the median is a more robust metric to highlight differences among farm basins. The MPL median for farm basin 01A with 124 kg was distinctly above all other medians (06AB>07AB>00A>09A>03A>04A>05A>02A>08A). The maximum MPL was found in 01A with 2527 kg followed by 06AB (2386 kg) and 07AB (1352 kg). Note that medians of UAL were more similar among farms with highest in farm basin 01A (0.24 kg P ha^{-1}) and lowest in 03A (0.02 kg P ha^{-1}). Maximum UAL were observed in farm basins 01A with 4.88 kg P ha^{-1} and declined in the sequence 06AB>00A>07AB>05A>02A>04A>03A~09A>08A.

3.2. Single tree regression models to predict monthly phosphorus loads

The variable importance to predict MPL in ST mode (29 terminal nodes) followed the ranking order: monthlyinsidehead (100)>irrdemand (94.6)>monthlyrain (90.0)>pumptorain (88.0)>percentfallflood (71.2)>location (44.8)>..... soils (2.7) (Table 5). This suggests that hydrologic properties were more important than the irrpconc and landscape properties (e.g. land use, soildepth, soils) to predict MPL. The geographic location expressed by the two variables location and sub-basin were also less important to explain the long-term MPL in the EAA.

Similar variable importance and surrogate behavior were found for the more parsimonious ST with 14 terminal nodes with a R^2 of 0.77. If only primary splitters are considered for the tree-building process of ST with 14 terminal nodes the following variables are ranked according to their scores: pumptorain (100)>irrdemand (80.4)> percentfallflood (63.2)>location (38.0)>monthlyrain (25.6)>farmsize (3.98). Thus, only hydrologic variables and the size of farms were important for splitting parent nodes into child nodes during the tree-building process. Less importance was placed on P concentrations and other landscape properties included in the model. If interactions, i.e. synergy between variables, were discounted during the tree-building process (14 terminal nodes) the variable importance shifted slightly with variable importance in the following order: pumptorain (100)> irrdemand (86.9)>percentfallflood (74.9)>monthlyrain (54.5)>location (43.2)........sub-basin (2.7). This indicates that monthlyinsidehead is dependent on interaction with other variables (possibly hydrologic variables) in the tree model to derive their high overall variable importance. To identify these high-order interactions is the strength of the tree-modeling approach.

To exemplify the tree results the ST build to predict MPL with 14 terminal nodes is shown in Fig. 2. Terminal node 1 showed the lowest mean with 12 kg (Standard deviation – STD: 19 kg) and terminal node 8

Table 4
Descriptive statistics of monthly phosphorus unit area loads (kg P ha^{-1}) on each of the 10 farm basins in the Everglades Agricultural Area.

	Farm drainage basins										
	All 10 farm basins	00A	01A	02A	03A	04A	05A	06AB	07AB	08A	09A
Mean	0.16	0.18	0.69	0.06	0.03	0.07	0.12	0.31	0.16	0.03	0.04
Std. error of mean	0.002	0.04	0.13	0.01	0.005	0.009	0.02	0.05	0.02	0.004	0.004
Median	0.05	0.08	0.24	0.03	0.02	0.04	0.06	0.15	0.11	0.02	0.03
Standard deviation	0.41	0.32	1.05	0.09	0.05	0.09	0.18	0.49	0.19	0.03	0.04
Skewness	6.6	4.2	2.5	3.7	2.4	2.0	3.0	3.7	3.3	2.2	2.5
Kurtosis	54.9	20.6	6.3	17.4	7.1	5.1	10.6	16.8	15.5	6.7	8.9
Range	4.88	2.12	4.88	0.54	0.28	0.44	1.03	3.36	1.34	0.17	0.27
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	4.88	2.12	4.88	0.54	0.28	0.44	1.03	3.36	1.34	0.17	0.27

Table 5
Summary 10-fold cross-validation results for single and committee tree models to predict monthly phosphorus loads on 10 farms in the Everglades Agricultural Area.

Tree models	Terminal nodes	Cross-validated relative error	Resubstitution relative error	R^2	Variable importance[a] (only the important ones are listed)
Single tree	29[b]	0.68	0.18	0.82	monthlyinsidehead (100)
	14[c]	0.70	0.22	0.77	**irrdemand** (94.6) **monthlyrain** (90.0) **pumptorain** (88.0) **percentfallflood** (71.2) **location** (44.8) **percentcane** (28.9) **monthlyoutsidehead** (27.9) **percentflood** (22.4) farmsize (8.9) sub-basin (7.9) irrpconc (7.3) soildepth (6.4) soils (2.7)
Committee tree (Bagging)	126[b]	0.44	0.11	0.88	monthlyoutsidehead (100)
	14[c]	0.51	0.18	0.82	monthlyinsidehead (97.9) irrdemand (74.6) pumptorain (62.6) percentcane (56.7) irrpconc (45.5) monthlyrain (44.8) location (34.3) percentfallflood (29.0) farmsize (21.9) sub-basin (15.8) percentflood (10.6) soildepth (8.4) soils (2.4)
Committee tree (ARCing)	106[b]	0.05	0.01	0.99	irrdemand (100)
	12[c]	0.21	0.16	0.84	pumptorain (84.4) monthlyrain (66.2) monthlyinsidehead (63.3) percentfallflood (58.4) percentflood (53.2) location (36.7) irrpconc (34.6) monthlyoutsidehead (34.5) percentcane (27.2) farmsize (22.8) sub-basin (22.6) soildepth (19.3) soils (5.6)

[a] Important primary splitters are in bold [marked only for single tree models].
[b] Model with smallest error ("best model").
[c] Parsimonious model with less terminal nodes and less complexity.

the largest mean with 1690 kg (STD: 597). This model shows the dominance of the primary split variables that were used multiple times in the tree to produce the tree branches. The right branches of the tree predicted low means of MPL with terminal nodes 9, 10, 11, and 12; whereas one of the center branches predicted high monthly P loads (nodes 7, 8 and 9). For example, to predict mean MPL of 1640 kg (terminal node 7) the following splitting rules have to be met in the following sequence:

percentfallflood <100 AND
location {"01A","06AB", OR "07AB"} AND
irrdemand ≤−7.62 AND
pumptorain ≤1.2 AND
monthlyrain >254 AND
pumptorain >0.9.

3.3. Committee tree models in bagging mode to predict monthly phosphorus loads

Bagging used repeated observations to run the bootstrap to generate 250 tree models to predict MPL for all 10 farms. It improved the prediction of MPL across the 10 farm basins with a 126 terminal node CT model emerging as the best one with a CVRE of 0.44, RRE of 0.11 and R^2 of 0.89 (Table 5). However, the complexity of the model structure was high with hundreds of nodes and branches in the tree. The CTb pruned to 14 terminal nodes showed comparable predictions with a CVRE of 0.51, RRE of 0.18 and R^2 of 0.81, but with much less complexity. This suggests that a parsimonious CT model with less nodes is likewise suited to predict MPL. The highly complex CT model with hundreds of terminal nodes improved predictions of MPL at the price of including additional variables that added less to the overall prediction accuracy. The relative importance of variables for 250 trees in Bagging mode yielded the following ranking list: monthlyoutsidehead (100)>monthlyinsidehead (97.9)>irrdemand (74.6)>pumptorain (62.6)>percentcane (56.7)>irrpconc (45.5)>monthlyrain (44.8)>location (34.3).....>soils (2.4) (Table 5). This confirmed results from the ST models illustrating that hydrologic variables have higher predictive power to infer on MPL when compared to irrpconc, location and soil properties. According to the CTb the farm and sub-basin locations showed less control to influence MPL. In summary, the CTb improved predictions of MPL across the 10 farms over ST. However, it was also shown that the ST model seemed robust to relate independent to the target variables. Committee tree models derived by Bagging may provide excellent fit to relate input to output variables; however, run the risk to over-parameterize if the tree is grown to full extent (126 terminal nodes). Pruned trees, such as the CTb with 14 terminal nodes performed nearly equally well when compared to the full tree but are likely more robust to predict MPL across all 10 farm basins.

3.4. Committee tree models in ARCing mode to predict monthly phosphorus loads

ARCing improved the prediction of MPL within 10 farm basins by reducing the sum of residuals square of 4,933,309 kg (initial tree) to 4,265,030 kg (committee tree). It used higher counts of repeated sampling from the total population when compared to Bagging. The predication capabilities in ARCing mode were striking and much improved when compared to ST and CTb. The best model with 106 terminal nodes showed a CVRE of 0.05, RRE of 0.005 and R^2 of 0.99 (Table 5). This indicates that the selected predictor variables were well suited to predict MPL but complex with hundreds of nodes, splits and tree branches. A more parsimonious 12 terminal node CT was identified with ARCing that had a CVRE of 0.21, RRE of 0.16 and R^2 of 0.84. This CTa was similar to the ones presented above derived from Bagging and ST modeling. The variable importance for the ARCing 106 node committee model showed the dominance of hydrologic properties that seemed to control MPL predications. However, the ranking of variable importance was slightly different when compared to the Bagging and ST models with: irrdemand (100)>pumptorain (84.4)>monthlyrain (66.2)>monthlyinsidehead (63.3)>percentfallflood (58.4)>percentflood (53.2)..... soils (5.6). The variable of importance for the location of farms and sub-basins as well as farmsize played a less important role to relate to MPL. In summary, ARCing performed best out of all tested tree modes to describe the relationships between input and target variable (MPL). The strength of the advanced boosting optimization method (ARCing) improved the regression modeling process to infer on MPL in the EAA.

3.5. Single tree models to predict monthly phosphorus unit area loads

The best ST regression model to predict monthly P UAL (kg P ha^{-1}) had 10 terminal nodes, a CVRE of 0.71, RRE of 0.31 and R^2 of 0.69. The

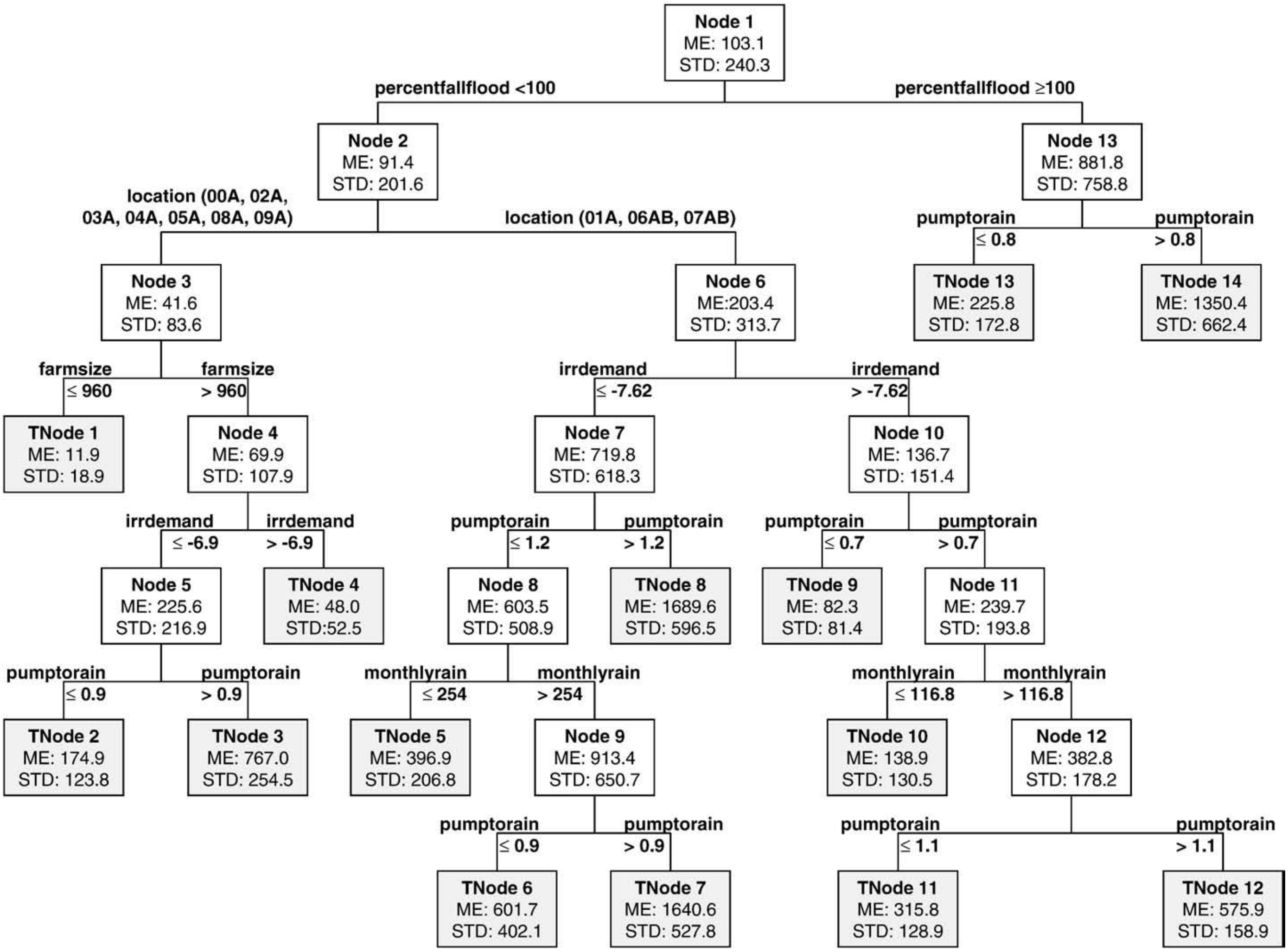

Fig. 2. Single tree regression model (14 terminal nodes) to predict monthly phosphorus loads in kilograms in farm basins in the Everglades Agricultural Area. The tree shows the splitting rules on top of each node. Terminal nodes (Tnodes) are shown in gray and other nodes in white. Each node shows the node number, mean (ME) and standard deviation (STD) grouped into a specific node.

predictions were not as good as for trees derived to predict MPL, but still moderately strong. The competing variables that improved the tree model were in the order: percentfallflood (100)>percentcane (87.6)> irrdemand (82.9)>location (69.1)>pumptorain (67.1)>...... soils (2.9). A large number of observations (549) were grouped into terminal node 6 with a mean of 0.06 kg P ha^{-1} (STD: 0.08), whereas much fewer observations were grouped into terminal nodes with larger UAL values (Fig. 3). For example, terminal node 1 (count: 5) had a mean UAL of 2.63 kg P ha^{-1} (STD: 1.32) and terminal node 10 (count: 7) a mean of 2.61 kg P ha^{-1} (STD: 1.27). This is in line with the skewness of the distribution of UAL data. Regression trees are well suited to make good predictions even with highly skewed input and output data.

The relative importance to predict UAL using primary splitters and surrogates in ST mode were in the order: percentfallflood (100)> percentcane (87.6)>irrdemand (82.9)>location (69.1)>pumptorain (67.1)>monthlyinsidehead (63.0)>monthlyrain (40.0)>monthlyoutsidehead (38.8) >...... soils (2.9) (Table 6). This ST model that predicted monthly P UAL ranked the variables percentfallflood and percentcane much higher than the ST model that predicted MPL. In contrast, monthlyinsidehead that had much more importance in the ST MPL prediction model showed much less importance to infer on UAL. The variables irrdemand, location, and pumptorain seemed to have equal importance in both ST models predicting MPL and UAL. As expected, farm size did not have much importance to predict UAL. The UAL model was simpler when compared to the MPL tree model as indicated by the number of primary split variables and only considered five variables including percentfallflood (100), irrdemand (80.0), percentcane (62.5), pumptorain (60.8), and location (31.0) as primary split variables of importance. The contribution of surrogates to improve predictions of UAL in ST mode seemed minor. This suggests that interaction between monthlyinsidehead and monthlyoutsidehead, among other variable interactions, contributed to the prediction success of UAL. A high monthly irrigation demand indicates low monthly rainfall and thus little need for drainage pumping. In contrast, a high negative irrigation demand is associated with high need for drainage pumping.

The tree topology of the UAL ST model is shown in Fig. 3. The left branch of the tree used the following splitting rules (conditions) to predict high UAL of mean 2.64 kg P ha^{-1} (terminal node 1):

percentfallflood ≤ 96 AND
irrdemand ≤ −7.26 AND
percentcane ≤ 29 AND
irrdemand ≤ −6.83.

The right branch of the tree used the following splitting rules to predict likewise high UAL of mean 2.61 kg P ha^{-1} (terminal node 10):

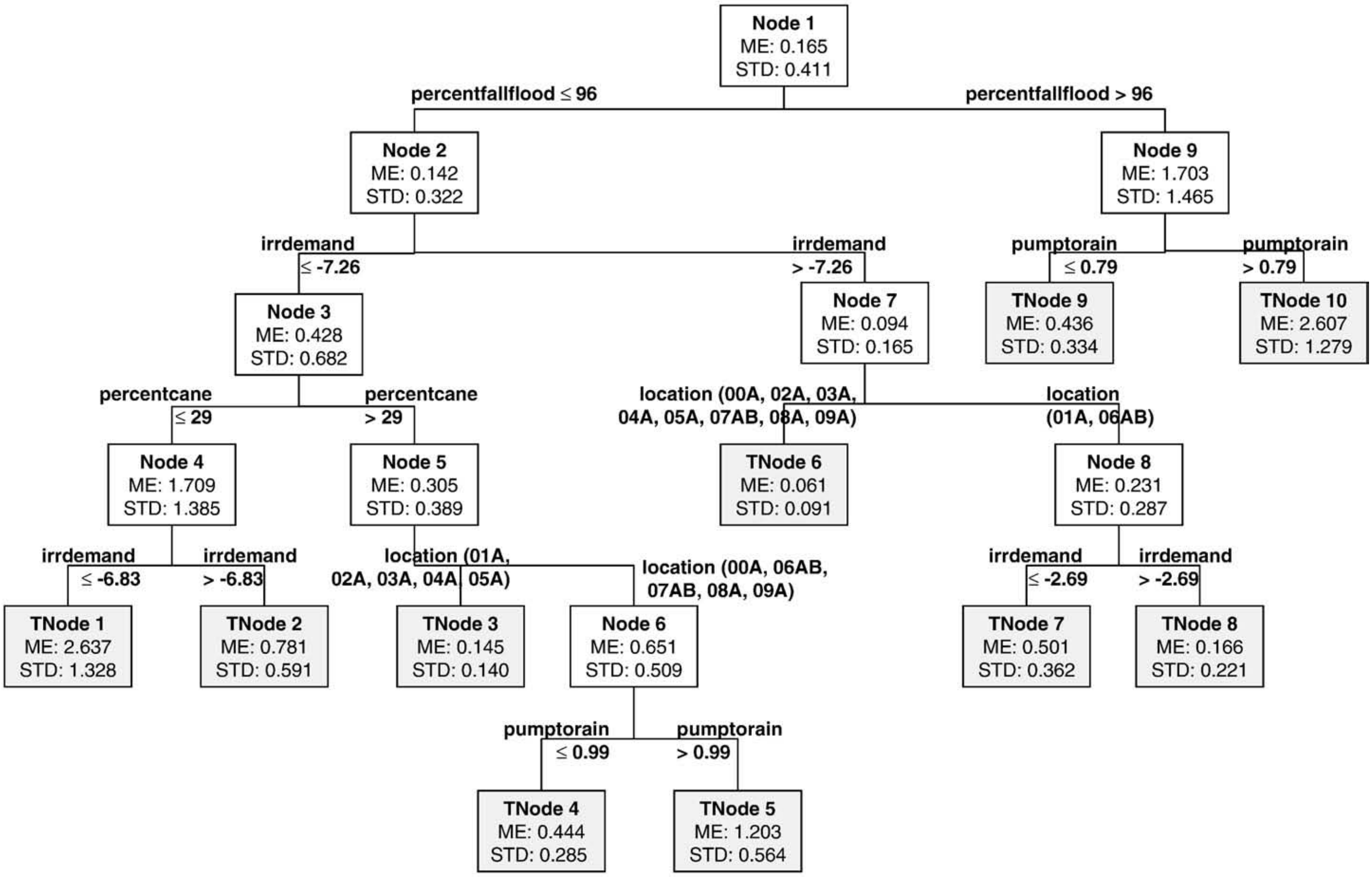

Fig. 3. Single tree regression model (10 terminal nodes) to predict monthly phosphorus unit area loads in kg ha^{-1} in farm basins in the Everglades Agricultural Area. The tree shows the splitting rules on top of each node. Terminal nodes (Tnodes) are shown in gray and other nodes in white. Each node shows the node number, mean (ME) and standard deviation (STD) grouped into a specific node.

percentflood >96 AND
pumptorain >0.79.

In contrast, very low UAL of mean 0.06 kg P ha^{-1} (terminal node 6) were generated using the following conditional rules:

percentfallflood ≤96 AND
irrdemand >−7.26 AND
location is {“00A”,“02A”,“03A”,“04A”,“05A”, “07AB”,“08A”, OR “09A”}.

3.6. Committee tree models in bagging mode to predict monthly phosphorus unit area loads

Bagging improved the predictions of UAL when compared to the ST regression model. However, the improvement was small as indicated by the R^2 that improved only 0.06 units. The best CTb model had 8 terminal nodes, a CVRE of 0.66, RRE of 0.25, and R^2 of 0.75 (Table 6). All CT generated with Bagging were parsimonious (i.e., had relatively few terminal nodes) and did not show the extreme overfitting with hundreds of terminal nodes that were generated with Bagging to predict MPL. Similar to the ST model to predict UAL the CTb confirmed the importance of the variable percentfallflood to predict UAL. The relative variable importance ranking for the UAL CTb was in the following order: percentfallflood (100)>irrdemand (85.6)>monthlyoutsidehead (79.8)>pumptorain (68.8)>monthlyrain (50.4)> monthlyinsidehead (42.7).......soils (0.0) (Table 6). The variable percentcane was less important in the Bagging model than the ST model.

3.7. Committee tree models in ARCing mode to predict monthly phosphorus unit area loads

The ARCing model improved much over the ST regression model shifting the R^2 from 0.69 to 0.95. The best ARCing model had 29 terminal nodes, a CVRE of 0.08, RRE of 0.04, and R^2 of 0.95. The CTa was able to accurately predict UAL with a parsimonious underlying model structure of 10 terminal nodes. The relative importance of variables was in the following order: pumptorain (100)>percentflood (72.5)>irrdemand (70.0)>percentfallflood (61.0)>irrpconc (57.5)> monthlyrain (54.5)>monthlyoutsidehead (54.5)>monthlyinsidehead (43.4)>......soils (2.7) (Table 6). In the CTa model the variable percentfallflood lost importance when compared to the Bagging model with scores dropping from 100 to 61. But pumptorain had more importance in the ARCing model with a score of 100 when compared to the Bagging model (68.8). It is interesting to note that irrpconc was much more important to predict UAL in the ARCing committee tree (relative importance score: 57.5) when compared to the Bagging tree (5.0). The variable percentflood with a relative importance score of 72.5 in the ARCing model showed much less importance in the Bagging model with a score of 21.1. Overall, variables that characterize geographic locations, such as sub-basin and location, had less importance in the ARCing model. Variables with comparable importance to predict UAL in ARCing and Bagging modes were pumptorain, irrdemand, percentfallflood and monthlyrain suggesting the importance of drainage / hydrology to predict UAL. The terminal nodes of the ARCing CT (trees not shown) showed a much more even distribution of observations grouped into different terminal nodes when compared to the Bagging committee tree. This may have also

Table 6
Summary 10-fold cross-validation results for single and committee tree models to predict monthly phosphorus unit area loads on 10 farms in the Everglades Agricultural Area.

Tree models	Terminal nodes	Cross-validated relative error	Resubstitution relative error	R^2	Variable importance[a] (only the most important ones are listed)
Single tree	10[b]	0.71	0.31	0.69	**percentfallflood** (100) **percentcane** (87.6) **irrdemand** (82.9) **location** (69.1) **pumptorain** (67.1) monthlyinsidehead (63.0) monthlyrain (39.9) monthlyoutsidehead (38.8) percentflood (34.8) irrpconc (15.9) sub-basin (8.7) farmsize (7.7) soildepth (7.5) soils (2.9)
Committee tree (Bagging)	8[b]	0.66	0.25	0.75	percentfallflood (100) irrdemand (85.6) monthlyoutsidehead (79.8) pumptorain (68.8) monthlyrain (50.4) monthlyinsidehead (42.7) percentflood (21.1) location (17.5) percentcane (14.2) irrpconc (5.0) sub-basin (4.3) soildepth (4.3) farmsize (1.6) soils (0.0)
Committee tree (ARCing)	29[b]	0.08	0.05	0.95	pumptorain (100) percentflood (72.5) irrdemand (70.0) percentfallflood (60.9) irrpconc (57.5) monthlyrain (54.5) monthlyoutsidehead (54.0) monthlyinsidehead (43.4) location (18.6) percentcane (14.2) soildepth (13.0) sub-basin (11.6) farmsize (8.8) soils (2.7)
	10[c]	0.26	0.22	0.78	

[a] Important primary splitters are in bold [marked only for single tree models].
[b] Model with smallest error ("best model").
[c] Parsimonious model with less terminal nodes and less complexity.

contributed to the superior performance of ARCing over Bagging to predict UAL.

4. Discussion

According to Walker (1999) P concentrations in the water column of the least impacted Everglades are typically <10 µg L^{-1}, which resemble the proposed goal set by the Everglades Forever Act of 10 µg L^{-1} TP for the Everglades Protection Area. The mean P concentrations in this study were elevated with 0.12 mg P L^{-1} (1992–2002) and comparable to those observed by Adorisio et al. (2006) with 0.13 mg P L^{-1} (water years 1994–2005) in the EAA basin. Water flow velocities are low in lowland drainage basins, but in the EAA, water flow velocities in farm drainage canals can be high depending on canal dimensions and canal water levels. This may lead to sediment re-suspension and transport. Substantial amounts of P can be stored in the sediment that is a magnitude of order higher when compared to P in the water column (Reddy et al., 1999). These P retention mechanisms include uptake and release by vegetation and micro-organisms, sorption and exchange reactions with soils and sediment, chemical precipitation in the water column, and sedimentation and entrainment. In this study the average MPL on sugarcane farms was 40.8 kg (3.3–94.3 kg), whereas MPL on mixed-used farm drainage basins was 4.5 times higher with 188.6 kg (15.8–357.1 kg). The same trend was confirmed by monthly UAL on sugarcane farms averaging 0.07 kg P ha^{-1} (0.18–0.03 kg P ha^{-1}), whereas mean UAL were much higher on mixed-use farms with 0.32 kg P ha^{-1} (0.12–0.69 kg P ha^{-1}). These results are not surprising given the higher P fertilizer requirement and more intensive water management that crops besides sugarcane generally need. On sugarcane farm basins UAL were controlled mainly by pumptorrain, irrdemand, monthlyrain and monthlyoutsidehead, whereas UAL on mixed-use farms were controlled by monthlyrain, pumptorain, percentcane, and irrpconc (ARCing mode). Rice et al. (2002a,b) observed average baseline UAL on sugarcane farms (1992) of 1.25 kg P ha^{-1} (0.51–2.47 kg P ha^{-1}) and 2.82 kg P ha^{-1} (0.51–5.75 kg P ha^{-1}) during a BMP monitoring period (1994–1998) in the EAA. On mixed-vegetable farm drainage basin in the EAA they found 5.60 kg P ha^{-1} (baseline) and 4.79 kg P ha^{-1} (BMP period), and on a sugarcane-rice farm drainage basin 1.29 kg P ha^{-1} (baseline) and 1.43 kg P ha^{-1} (BMP period), respectively. Daroub et al. (in press) observed a decreasing temporal trend (1992–2002) in flow-weighted monthly P concentrations and P loads on five out of ten farm drainage basins in the EAA with four of these basins being sugarcane monoculture. Changes in BMP implementation in the EAA since 1995 have resulted in average annual P load reductions of >50%, compared with baseline values (Daroub et al., 2004).

Factors that describe cropping practices (percentcane, percentfallflood, and percentflood) were less important to predict MPL than UAL. In CTb mode to predict MPL the importance of cropping practices declined with percentcane (56.7), percentfallflood (29.0), and pecentflood (10.6); and in CTa mode with percentfallflood (58.4), percentflood (53.2), and percentcane (27.2). Cropping practices were more pronounced to infer on UAL with percentfallflood (100), percentflood (21.1), percentcane (14.2) in CTb mode; and percentflood (72.5), percentfallflood (60.9), and percentcane (14.2) in CTa mode.

In our study rainfall ranked moderately high across all tree models with variable importance of 44.8 (CTb) and 66.2 (CTa) to control MPL; and 50.4 (CTb) and 54.5 (CTa) to predict UAL. This confirms the importance of rainfall relative to P loads emphasized earlier by Rice et al. (2002a,b). Farm-specific variables (farm size, location, and sub-basin) did not impart major inferences to MPL and UAL. This suggests that overall farm-specific differences are of less importance to control P loads in the EAA basin.

Our results derived from tree-based modeling suggest that hydrologic/water management properties are the major controlling variables to predict MPL and monthly P UAL in the EAA. The CTb model selected monthlyoutsidehead, monthlyinsidehead, irrdemand and pumptorain as most important variables; and the CTa model irrdemand, pumptorain, monthlyrain, and monthlyinsidehead to infer on MPL, respectively. Other sets of hydrologic/water management predictor variables were prominent to predict UAL with highest variable importance including percentfallflood, irrdemand, monthlyoutsidehead, and pumptorain (CTb model), and pumptorain, percentflood, irrdemand and percentfallflood (CTa model). Walker (1999), Stuck et al. (2001) and Diaz et al. (2006) emphasized that the magnitude of P loads is dependent on a combination of factors including nutrient loading, cross-sectional area (dimension) of canals and ditches that controls drainage volume, and pumping activities.

Izuno and Rice (1999) showed that particulate P (PP) accounted for 20% to 70% of the TP exported from EAA farms, subject to various cropping practices and rainfall characteristics, suggesting that the major mechanism for P is through sediment transport. They also showed that spikes in TP and PP loads coincided. In contrast, Stuck et al. (2001) indicated that farm canals in the EAA have a significant impact on TP loads discharged from agricultural farms, whereby the bulk of exported PP is sourced from biotic material growing in farm canals. They also showed that P adsorption–desorption was relatively unimportant for the high organic content particulates encountered in the EAA in field ditches.

The ratio between pumping to rainfall was identified as very important to predict MPL with high relative importance in ST (88.0), CTb (62.6), and CTa (84.4) models. Similarly, the relative importance of pumptorain was very high to predict UAL in ST (67.1), CTb (68.8), and CTa (100) modes. The relationship between pumping activities and P loads was investigated by Stuck et al. (2001) who found that farm pumping events produced specific total suspended solids (TSS) profiles in channels in the EAA with highest TSS concentrations coinciding with the beginning of pumping and then steadily declining followed by relatively low levels for the duration of the pumping event (Stuck et al., 2001). Interestingly, total TSS loads showed an inverse relationship with P loads across all farm pumping events during the monitoring period in the EAA (Stuck et al., 2001).

The pronounced structure of ditches and canals in this low relief landscape of the EAA provides preferential hydrologic pathways that seem to have a major impact on P load export from farm drainage basins as indicated by all tree models (ST, CTa and CTb). In contrast, field soil characteristics such as soil depth and soil type did not seem to impart major control on P loads, which may have been due to the relative homogeneous distribution of soils across the EAA that did not show major differences in soil characteristics. This can be confirmed by Stuck et al. (2001) who observed 9 to 20 times higher P content in ditch sediment samples when compared to field soil samples with 700 to 750 mg kg^{-1} suggesting the dominant role of ditches/canals to modulate drainage on P loads. They found that P content of the exported suspended solids and P content of macrophytes corresponded closely suggesting that exported solids are sourced from macrophytes rather than sediments (Stuck et al., 2001). An earlier study by Fiskell and Nicholson (1986) on 11 different fields under various crop management practices observed P content of the top 5 cm profile from 298 to 814 mg kg^{-1}, averaging 496 mg kg^{-1}. In the same study, the P content of field ditch sediments, and of the canal sediment samples not associated with macrophytes, was slightly higher ranging from 500 to 900 mg kg^{-1}.

5. Conclusions

Overall, tree-based modeling was successful in identifying relationships between P loads and environmental predictor variables on 10 farms in the EAA indicated by high R^2 and low prediction errors. Committee trees in ARCing mode generated the best performing models followed by CTb and ST to predict MPL as well as UAL. Tree-based models had the ability to analyze complex, non-linear relationships between P loads and multiple variables describing hydrologic/water management, cropping practices, soil and farm-specific properties within the study area. Hydrologic/water management variables showed the strongest relationships to MPL and UAL. A long-term BMP implementation program has reduced P loads from cropped fields to ditches, but it is not clear how much P is still retained within the EAA drainage basin. Given the importance of hydrologic preferential pathways (ditches/canals) in this lowland drainage basin, which imparts major control on P loads attention has to focus on mechanisms to either retain as much P within the system or reduce P export along these drainage pathways. The measured P loads in the EAA (1992–2002) do not resemble historic, oligotrophic conditions and may pose a future threat to the restoration of the Greater Everglades. However, agricultural use imposes some higher nutrient status when compared to areas not impacted by human management. Combining BMPs and storm water treatment areas has the potential to reduce P loads to 10 ppb which is the restoration goal of the Everglades. Expected land use shifts from sugarcane to more natural wetland ecosystems that will impact about 27% of the EAA in the near future may come at the risk of accelerating re-suspension and transport of P currently retained within the EAA. It will be critical to continue the monitoring of P concentrations and drainage from farm basins, which can be complemented by tree-based analysis of P data.

Acknowledgments

Funding for this project was provided by the Everglades Agricultural Area Environmental Protection District and the Florida Agriculture Experiment Station. Data provided by South Florida Water Management District is greatly appreciated.

References

Adorisio C, Bedregal C, Daroub S, McGinnes P, Miessau C, Pescatore D, et al. Chapter 3, Phosphorus controls for the basins tributary to the Everglades construction project. 2005 South Florida Environmental Report. West Palm Beach, FL: South Florida Water Management District; 2006.

Alexander RB, Smith RA, Schwarz GE. Estimates of diffuse phosphorus sources in surface waters of the United States using a spatially referenced watershed model. Water Sci Technol 2004;49(3):1-10.

Ali A, Abtew W, Van Horn S, Khanal N. Temporal and spatial characterization of rainfall over central and south Florida. J Am Water Resour Assoc 2000;36(4):833–48.

Barker LS, Felton GK, Russek-Cohen E. Use of Maryland biological stream survey data to determine effects of agricultural riparian buffers on measures of biological stream health. Environ Monit Assess 2006;117:1-19.

Bhattacharya B, Solomatine DP. Neural networks and M5 model trees in modeling water level-discharge relationship. Neurocomputing 2005;63:381–96.

Boesch DF, Brinsfield RB, Magnien RE. Chesapeake Bay eutrophication – scientific understanding, ecosystem restoration, and challenges for agriculture. J Environ Qual 2001;30:303–20.

Brazier R, Schärer M, Heathwaite L, Beven K, Scholefield P, Haygarth P, et al. A framework for predicting delivery of phosphorus from agricultural land using a decision-tree approach. Proceed. of Symposium on Sediment Dynamics and the Hydromorphology of Fluvial Systems, Dundee, UK July 2006, IAHS Publ., 306. ; 2006. p. 514–23.

Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. CA: Wadsworth International Group; 1984. 358 pp.

Brown DJ. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 2007;140:444–53.

Bruland GL, Osborne TZ, Reddy KR, Grunwald S, Newman S, DeBusk WF. Recent changes in soil total phosphorus in the Everglades: water conservation area 3. Environ Monit Assess 2007;129:379–95.

CART. Classification and Regression Trees – User's Guide. An implementation of the original CART methodology by Salford Systems; 2002. 302 pp.

Chen M. Everglades Research and Education Center quality manual. Everglades Res. and Educ. Center. Inst. Food and Agric. Sci.Belle Glade, FL: Univ. of Florida; 2001.

Compton J, Mallinson D, Glenn CR, Filippelli G, Föllmi K, Shields G, et al. Variations in the global phosphorus cycle. Marine anthigenesis: from global to microbial, vol. 66. SEPM Spec. Publ.; 2000. p. 21–33.

Daroub, S.H., Lang, T.A., Diaz, O.A., Chen, M., Stuck, J.D. Annual report phase XII: implementation and verification of BMPs for reducing P loading in the EAA and Everglades Agricultural Area BMPs for reducing particulate phosphorus transport. Submitted to the Everglades Agricultural Area Environmental Protection District and The Florida Department of Environmental Protection. Univ. of Florida, EREC, Belle Glade, FL; 2004.

Daroub, S.H., Lang, T.A., Diaz, O.A., Grunwald, S. Long-term water quality trends after implementing best management practices in south Florida; J Environ Qual in press.

DeAth G. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 2002;83(4):1105–17.

DeAth G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 2000;81(11):3178–92.

Diaz OA, Daroub SH, Stuck JD, Clark MW, Lang TA, Reddy KR. Sediment inventory and phosphorus fractions for Water Conservation Area canals in the Everglades. Soil Sci Soc Am J 2006;70:863–71.

Fiskell JGA, Nicholson IK. Organic phosphorus content of Pahokee muck and spodosols in Florida. Soil Crop Sci Soc Fla Proc 1986;45:6-11.

Freund Y, Schapire RE. Additive logistic regression: a statistical view of boosting – discussion. Ann Stat 2000;28(2):391–3.

Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using Random Forest analysis. Geoderma 2008;146:102–13.

Grinand C, Arrouays D, Laroche B, Martin MP. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 2008;143:180–90.

Grunwald S, editor. Environmental soil–landscape modeling – geographic information technologies and pedometrics. New York: CRC Press; 2006. 488 pp.

Grunwald S, Qi C. GIS-based water quality modeling in the Sandusky Watershed. J Am Water Resour Assoc 2006;42(4):957–73.

Grunwald S, Osborne TZ, Reddy KR. Temporal trajectories of phosphorus and pedo-patterns mapped in Water Conservation Area 2, Everglades, Florida, USA. Geoderma 2008;146:1-13.

Hargrove WW, Hoffman FM. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. Environ Manag 2005;34(1):39–60.

Heathwaite AL, Dils RM. Characterising phosphorus loss in surface and subsurface hydrological pathways. Sci Total Environ 2000;251/252:523–38.

Izuno FT, Bottcher AB. The history of water management in South Florida. In: Bottcher AB, Izuno FT, editors. Everglades Agricultural Area (EAA): water, soil, crop, and environmental management. Gainesville, FL: University Press of Florida; 1994. p. 13–26.

Izuno FT, Rice RW, editors. Implementation and verification of BMPs for reducing P loading in the EAA. Final project report submitted to the Florida Department of Environmental Protection and the Everglades Agricultural Area Environmental Protection District. Tallahassee, FL; 1999.

James MR, Turner MG, Smithwick EAH, Dent CL, Stanley EH. Spatial extrapolation: the science of predicting ecological patterns and processes. Bioscience 2004;54 (4):310–20.

Jones KB, Heggen DT, Wade TG, Neale AC, Ebert DW, Nash MS, et al. Assessing landscape condition relative to water resources in the western United States: a strategic approach. Environ Monit Assess 2000;64:227–45.

Lilly A, Nemes A, Rawls WJ, Pachepsky YA. Probabilistic approach to the identification of input variables to estimate hydraulic conductivity. Soil Sci Soc Am J 2008;72:16–24.

Mander Ü, Kull A, Kuusemets V, Tamm T. Nutrient runoff dynamics in a rural catchment: influence of land-use changes, climatic fluctuations and ecotechnological measures. Ecol Eng 2000;14:405–17.

Moisen GG, Frescino TS. Comparing five modeling techniques for predicting forest characteristics. Ecol Model 2002;157:209–25.

Mototch NP, Colee MT, Bales RC, Dozier J. Estimating the spatial distribution of snow water equivalent in an alpine basin using binary regression tree models: the impact of digital elevation data and independent variable selection. Hydrol Process 2005;19:1459–79.

Murphy J, Riley JP. A modified single solution method for the determination of phosphate in natural waters. Anal Chim Acta 1962;27:31–6.

Noe GB, Childers DL, Jones RD. Phosphorus biogeochemistry and the impact of phosphorus enrichment: why is the Everglades so unique? Ecosystems 2001;4:603–24.

Obreza TA, Anderson DL, Pitts DJ. Water and nitrogen management of sugarcane grown on sandy, high-water-table soil. Soil Sci Soc Am J 1998;62:992–9.

Pappenberger F, Iorgulescu I, Beven KJ. Sensitivity analysis based on regional splits and regression trees (SARS-RT). Environ Model Softw 2006;21:976–90.

Pieterse NM, Bleuten W, Jørgensen SE. Contribution of point sources and diffuse sources to nitrogen and phosphorus loads in lowland river tributaries. J Hydol 2003;271:213–25.

Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: bagging and random forest for ecological prediction. Ecosystems 2006;9:181–99.

Porter, JW, Porter, KG, editors. The everglades, Florida Bay and coral reefs of the Florida Keys. New York: CRC Press; 2001. 1000 pp.

Rabalais NN, Turner RE, Scavia D. Beyond science into policy: Gulf of Mexico hypoxia and the Mississippi River. BioScience 2002;52(2):129–42.

Reddy KR, Kadlec RH, Flaig E, Gale PM. Phosphorus retention in streams and wetlands: a review. Crit Rev Environ Sci Technol 1999;29(1):83-146.

Reis LFR, Walters GA, Savic D, Chaudhry FH. Multi-reservoir operation planning using hybrid genetic algorithm and linear programming (GA-LP): an alternative stochastic approach. Water Resour Manage 2005;19:831–48.

Rice RW, Gilbert RA, Daroub SH. Application of the soil taxonomy key to the organic soils of the Everglades Agricultural Area. EDIS document # SS-AGR-246. Florida Coop. Ext. Service. Inst. of Food and Agric. Sci.Gainesville, FL: Univ. of Florida; 2002a.

Rice RW, Izuno FT, Garcia RM. Phosphorus load reductions under best management practices for sugarcane cropping systems in the Everglades Agricultural Area. Agric Water Manag 2002b;56(1):17–39.

Schröder W. GIS, geostatistics, metadata banking, and tree-based models for data analysis and mapping in environmental monitoring and epidemiology. Int J Med Microbiol 2006;296(S1):23–36.

Sharpley AN, McDowell RW, Kleinman PJA. Phosphorus loss from land to water: integrating agricultural and environmental management. Plant Soil 2001;237 (2):287–307.

Shih SF, Glaz B, Barnes RE. Subsidence of organic soils in the Everglades Agricultural Area during the past 19 years. Soil Crop Sci Soc Fl Proc 1998;57:20–9.

Snyder GH. Soils of the Everglades Agricultural Area. In: Boettcher AB, Izuno FT, editors. Everglades Agricultural Area: water, soil, crop, and environmental management. Gainesville, FL: University Press of Florida; 1994. p. 27–41. chapter 3.

Solomatine DP, Dulal KN. Model trees as an alternative to neural networks in rainfall-runoff modeling. Hydrol Sci J 2003;48(3):399–412.

Steinberg D, Colla P. CART: tree-structured non-parametric data analysis. San Diego, CA: Salford Systems; 1997. 233 pp.

Stockstad E. Big land purchase triggers review of plans to restore Everglades. Science 2008;321(5885):22.

Stuck JD, Izuno FT, Campbell KL, Bottcher AB, Rice RW. Farm-level studies of particulate phosphorus in the Everglades Agricultural Area. Trans ASAE 2001;44:1105–16.

Thuiller W. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Glob Chang Biol 2003;9(10):1353–62.

Trexler JC, Travis J. Nontraditional regression analyses. Ecology 1993;76(6):1629–37.

U.S. Environmental Protection Agency (USEPA). Methods for chemical analysis of water and wastes. 600/4-79-020, p. 365.1-1 and 365.4-1. USEPA Environmental Monitoring and Support Laboratory. Cincinnati, OH, 1983.

Vasques GM, Grunwald S, Sickman JO. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 2008;146:14–25.

Walker WE. Long-term water quality trends in the Everglades. In: Reddy KR, O'Connor GA, Schelske CL, editors. Phosphorus biogeochemistry in subtropical ecosystems. Boca Raton, FL: Lewis Publ.; 1999. p. 447–66.