Estuarine, Coastal and Shelf Science 125 (2013) 57–67

Contents lists available at SciVerse ScienceDirect

Estuarine, Coastal and Shelf Science

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

Coastal ‘Big Data’ and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric

David F. Millie[a,b,*], Gary R. Weckman[c], William A. Young II[d], James E. Ivey[e], David P. Fries[f], Ehsan Ardjmand[c], Gary L. Fahnenstiel[b,g]

[a] Palm Island Enviro-Informatics LLC, Sarasota, FL 34232, USA
[b] Michigan Technological University, Michigan Tech Research Institute, Ann Arbor, MI 48105, USA
[c] Ohio University, Russ College of Engineering and Technology, Department of Industrial and Systems Engineering, Athens, OH 45701, USA
[d] Ohio University, College of Business, Management Systems Department, Athens, OH 45701, USA
[e] Florida Fish & Wildlife Conservation Commission, Fish & Wildlife Research Institute, St. Petersburg, FL 33701, USA
[f] College of Marine Science, University of South Florida, St. Petersburg, FL 33701, USA
[g] Michigan Technological University, Great Lakes Research Center, Houghton, MI 49931, USA

ARTICLE INFO

Article history:
Received 20 December 2012
Accepted 2 April 2013
Available online 11 April 2013

Keywords:
artificial neural networks
decision trees
empirical models
environmental informatics
in situ chlorophyll
Sarasota Bay

ABSTRACT

Coastal monitoring has become reliant upon automated sensors for data acquisition. Such a technical commitment comes with a cost; particularly, the generation of large, high-dimensional data streams (‘Big Data’) that personnel must search through to identify data structures. Nature-inspired computation, inclusive of artificial neural networks (ANNs), affords the unearthing of complex, recurring patterns within sizable data volumes. In 2009, select meteorological and hydrological data were acquired via autonomous instruments in Sarasota Bay, Florida (USA). ANNs estimated continuous chlorophyll (CHL) *a* concentrations from abiotic predictors, with correlations between measured:modeled concentrations >0.90 and model efficiencies ranging from 0.80 to 0.90. Salinity and water temperature were the principal influences for modeled CHL within the Bay; concentrations steadily increased at temperatures >28° C and were greatest at salinities <36 (maximizing at ca. 35.3). Categorical ANNs modeled CHL classes of 6.1 and 11 μg CHL L^{-1} (representative of local and state-imposed constraint thresholds, respectively), with an accuracy of ca. 83% and class precision ranging from 0.79 to 0.91. The occurrence likelihood of concentrations > 6.1 μg CHL L^{-1} maximized at a salinity of ca. 36.3 and a temperature of ca. 29.5 °C. A 10th-order Chebyshev bivariate polynomial equation was fit (adj. $r^2 = 0.99$, $p < 0.001$) to a three-dimensional response surface portraying modeled CHL concentrations, conditional to the temperature–salinity interaction. The TREPAN algorithm queried a continuous ANN to extract a decision tree for delineation of CHL classes; turbidity, temperature, and salinity (and to lesser degrees, wind speed, wind/current direction, irradiance, and urea-nitrogen) were key variables for quantitative rules in tree formalisms. Taken together, computations enabled knowledge provision for and quantifiable representations of the non-linear relationships between environmental variables and CHL *a*.

> *… a radically new kind of “knowledge infrastructure” is materializing. A new era of ‘Big Data’ is emerging …*
>
> D. Bollier (2010)

1. Introduction

Coastal scientists traditionally have relied upon monitoring programs using invasive sampling at discrete locales/periods to derive conceptualizations of how local systems respond to anthropogenic stressors and natural disturbances. Limitations associated with such programs (e.g. the labor and time involved in

Abbreviations: ANN, artificial neural network; ATEMP, ambient temperature; BP, barometric pressure; CURDIR, current direction; CURSPD, current speed; DO, dissolved oxygen; HL, hidden layer; MLP, multi-layer perceptron; NIC, nature-inspired computation; NID, neural interpretive diagram; PAR, photosynthetic active radiation; PCA, principal component analysis; PE, processing element; pH, water acidity, basicity; PRECIP, precipitation; RH, relative humidity; SAL, salinity; SVM, support vector machines; TEMP, water temperature; TREPAN, TREes Parroting Network; TURB, turbidity; UR-N, urea-nitrogen; WNDDIR, wind direction; WNDSPD, wind speed.

[*] Corresponding author. Palm Island Enviro-Informatics LLC, 4645 Stone Ridge Trail, Sarasota, FL 34232, USA.

E-mail addresses: pi_enviro@comcast.net, dfmillie@mtu.edu, dfmillie@loyno.edu (D.F. Millie).

0272-7714/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecss.2013.04.001

sample acquisition over large spatial scales, the training, skill sets, and costs required for tedious in-laboratory sample preparation and analysis, the escalating expenses for personnel and equipment, etc.) result in data sets rendering ineffective spatial resolution, having discontinuous, often missed coverage and in effect, portraying ‘yesterday’s news.’ Coincident with the developmental explosion in computer technologies, scientists have committed to reliance upon automated sensors in coastal observatories, with their potential for continuous data acquisition, expansive spatial coverage, and data transfer in real/near-real time (e.g. Cole et al., 2003; Zappala and Azzaro, 2004; Paerl et al., 2005; Fries et al., 2008; Jannasch et al., 2008; Reed et al., 2010). However, such a technical commitment comes with a cost, particularly the generation of copious amounts of high-dimensional data streams (or ‘Big Data’; Bollier, 2010) that personnel must search through to identify complex, recurring data patterns.

The analysis of coastal ‘Big Data’ will be computationally-intensive, with a goal of describing system behavior from local observations of output (response) variables under conditions of interacting input (predictor) variables. Although complimentary, data mining and modeling have distinct utilities within ecological testing and interpretation. Data mining (the searching for unexpected patterns or relationships) facilitates reductions in database dimensionality via delineation of hetero-/homogenious observations and redundant predictors, while optimizing for lessened computational tasks within a minimized sample state space (Devlin, 1997; Müller and Lemke, 1999). Empirical modeling affords impartial prediction from historical observations (Elder and Pregibon, 1996), potentially enabling provision of user-friendly knowledge. Equally essential to knowledge derivation is the awareness for model reliability and the inherent uncertainties for the modeled-predictor relationship.

Nature-inspired computation (NIC) as an intelligent agent affords the unearthing of novel patterns and/or correlations within sizable volumes of data, thereby fueling scientific hypothesis formation and discovery (de Castro, 2007; Poole and Mackworth, 2010). Such an approach offers coastal managers the means with which to deconvolve and model the complexities of multifaceted systems and theoretically, an informational basis upon which to create rulings regarding water resources (see Abrahart et al., 2008). Artificial neural networks (ANNs) are biologically-motivated NIC technologies (de Castro, 2007) that are popular for the non-linear modeling of aquatic databases having high-dimensional space and displaying sizeable variance (e.g. Millie et al., 2006a,b; Suryanarayana et al., 2008; Jørgensen et al., 2009). However, ANNs act like ‘black boxes’ in that predictor–response relationships are encoded incomprehensibly as weight and bias values within the network’s complex topology (Fig. 1). Because they require little, if any expert knowledge for their application and exhibit a holistic lack of declarative information (c.f. Olden et al., 2004; Weckman et al., 2009), many scientists consider ANNs to have little, or no pragmatic ecological relevance.

The chlorophyll (CHL) *a* concentration of a water column is a metric for phytoplankton biomass and used within indices quantifying estuarine response to water-quality impairment (Borja et al., 2012). Together with hydrological and meteorological data, CHL concentrations routinely are acquired via automated instruments within monitoring programs; as such, they afford a pertinent example for conveying the synthesis, modeling, and interpretation of coastal ‘Big Data’. Here, ANNs were used to model CHL *a*

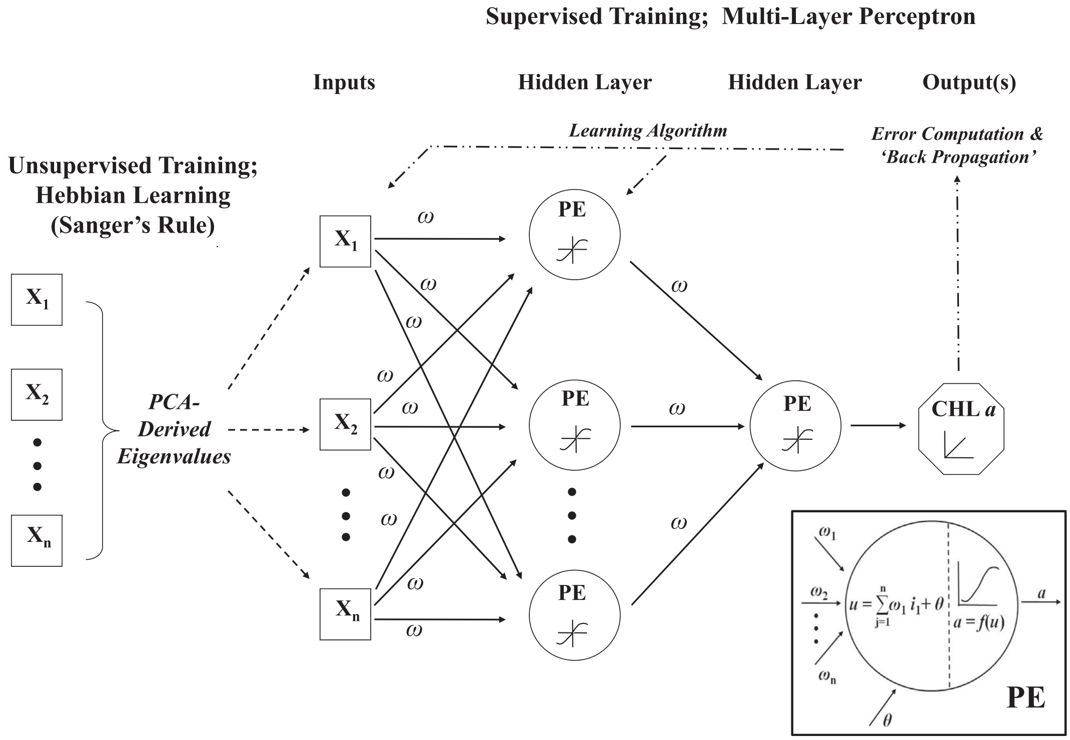

Fig. 1. Schematic of an artificial neural network. Supervised training: a multi-layer perceptron depicting interaction/influences among input variables ($X_{1\ldots n}$), hidden layers with processing elements ($\mathrm{PE}_{1\ldots j}$), synaptic weights (ω), and modeled CHL concentration/class (output). Unsupervised training: optimal linear features were extracted (as eigenvectors) from data, after which the eigenvalues were used as input for supervised training. Inset: Schematic depicting formulations of a hidden layer PE. For each presentation of the data set, inputs (i), whether from $X_{1\ldots n}$, or a net function from a PE, were multiplied by a synaptic weight ($\omega_{1\ldots n}$), the products summed, and combined with a bias value (θ) to produce (u) that then was transformed - via the illustrated sigmoid activation function, $f(u) = 1/(1 + e^{-u})$, to produce the output (a). Note: the PE in the output layer for continuous data utilized a linear function, $f(u) = au$.

concentrations from select environmental predictors for a lagoonal estuary. Specifically, we: (1) formulated network models for CHL *a*, as both continuous and categorical responses; (2) identified inconsistencies of the modeled CHL-predictor relationships; (3) via pedagogical perspectives, evaluated information potentials for networks, and (4) discussed the relevancy of NIC for coastal data applications.

2. Materials and methods

2.1. Sample site and data collection

Sarasota Bay (southwest Florida, USA) is a productive, lagoonal estuary of the Gulf of Mexico. Land use within the watershed of ca. 1200 km^2 primarily is residential and commercial, with only a small amount of agricultural activity. The exchange of Bay waters with Gulf waters is limited (restricted to four narrow 'passes') and the Bay has no major tributary. Almost all freshwater inputs arise from rainfall and associated storm-runoff. Within the Bay proper, hydrological conditions typically result in salinities (SALs) slightly greater (up to 39, practical salinity scale) than those of Gulf waters (ca. 36).

Data streams were acquired hourly from May to October, 2009 (resulting in ca. 3850 data vectors) via a moored instrument platform (27° 21′ 32″ N, 82° 36′ 14″ W). The platform housed a sensor array recording hourly data (units; abbreviation) for water temperature (°C; TEMP), turbidity (nephalometer turbidity units; TURB), acidity/basicity (−log [H+]; pH), SAL, dissolved oxygen (mg L^{-1}; DO), urea-nitrogen (μM N; UR-N) and CHL fluorescence (relative units) from sub-surface waters (ca. 1 m depth). A platform-mounted Acoustic Doppler Current Profiler recorded current velocity (m s^{-1}; CURSPD) and direction (compass degrees; CURDIR). A meteorological station provided data for ambient temperature (°C; ATEMP), wind speed (m s^{-1}; WNDSPD), wind direction (compass degrees; WNDDIR), precipitation (mm h^{-1}; PRECIP), barometric pressure (Hg; BP), relative humidity (%; RH), and photosynthetic active radiation (μE m^{-2} s^{-1}; PAR). Concentrations of CHL *a* (μg L^{-1}) were derived from fluorescence values (Holm-Hansen et al., 2000).

Because environmental variables at a coastal site reflect contemporary and proceeding conditions resulting from short-term meteorological events, water mass movement, nutrient transport, etc., a 'lagged' time series (3-hr delay) was added to typify the Bay's tidal cycles and/or short-term residence, effectively increasing two-fold the number of predictors. In addition to continuous data, CHL concentrations were assigned to classes (≤6.1 μg L^{-1}, >6.1 to ≤11 μg L^{-1}, >11 μg L^{-1}) reflecting management constraint thresholds of 6.1 μg L^{-1} (imposed by the Sarasota Bay Estuary Program, 2010) and 11 μg L^{-1} (imposed by the Florida Impaired Waters Rule, 62-303, FAC) for the Bay proper and all State of Florida impaired waters, respectively.

2.2. Predictor characterization & uncertainty

Diel means of the hydrological/meteorological variables were calculated from hourly data (except for daily PRECIP and PAR, in which hourly data were summed) to reduce the size of the database. Pair-wise scatter plots portrayed the distribution of and/or relationships among variables. The uncertainties of hydrological/meteorological predictors were assessed via a coefficient of variation (=standard deviation • mean^{-1}), with a variable having a greater coefficient being more dispersed than a variable with a lesser coefficient (Håkanson, 2000). A principal component analysis (PCA, *PRIMER* v6.1 software; Primer-E Ltd., Plymouth, UK), utilizing Euclidean distances, ordered sampling dates with respect to weekly mean values of environmental variables.

2.3. Network formulation, training, & testing

Concentrations of CHL *a* were modeled as classification and continuous problems using ANNs incorporating supervised learning or a hybrid integration of supervised and unsupervised learning (Fig. 1). For supervised learning (in which values for predictors are known), multi-layer perceptrons (MLPs) utilizing a back-propagation learning algorithm were constructed (*NeuroSolutions* v6.0 software; NeuroDimension, Inc., Gainesville, Florida USA). Hybrid networks implemented unsupervised learning (in which inputs were preprocessed to extract sets of uncorrelated linear features) prior to supervised learning. Support vector machines (SVMs), kernel-based algorithms using planar surfaces to separate classes, also were evaluated as an unsupervised feature (Dibike et al., 2001).

Networks were originated with intent to select the best architectures (after Millie et al., 2012); topologies were optimized for the number of processing elements (PEs) within hidden layers (HLs), and the types of transfer functions (sigmoid/hyperbolic tangent for continuous data, hyperbolic tangent/softmax for classification data) and learning rules (conjugate gradient, Levenberg–Marquardt, momentum, step). Data were assigned randomly to subsets for network training, cross-validation, and testing (60, 15, and 25% of data, respectively). Network training and testing followed that presented in Millie et al. (2012). During training, learning/momentum rates and step sizes were allowed to vary, thereby accelerating learning and ensuring convergence to a global minimum. Presentation of cross-validation data concurrent with training data provided for an unbiased estimation of prediction.

2.4. Network performance

In the modeling of continuous CHL concentrations, a correlation coefficient (ρ) measured the agreement between modeled:measured values. Performance metrics (MSE, normalized root MSE, reliability index, average error, average absolute error, modeling efficiency) afforded comparison between/among modeled outcomes. The MSE, normalized root MSE, and average absolute error quantified differences between modeled:measured values, with exact agreement producing a value of zero. The reliability index signified the mean correspondence factor between modeled:measured values, with exact agreement resulting in a value of one. Modeling efficiency assessed prediction relative to the mean measured value; values of one and zero signified exact similarity and that the measured mean value was no better than the modeled value, respectively (Stow et al., 2003).

Receiver operating characteristics (ROCs) afforded metrics for comparison of networks having categorical outcomes. By design, classification models produced tertiary outcomes, resulting in three-way confusion matrices. Because ROCs are performance evaluators for binary classification, tertiary classes were decomposed into dualistic outcomes via a 'one versus all' strategy (i.e. cases ordered as belonging to one group or not; Rifkin and Klautau, 2004). ROC metrics included accuracy in ordering of cases, the precision in ordering of distinct classes, the true positive/negative rates in classification, and the false positive/negative rates in classification for a confusion matrix (Brown and Davis, 2006). For tertiary classes within a network's test data subset, the true positive rate was plotted as a function of the false positive rate.

2.5. Network depiction & prediction uncertainty

To allow reproduction of computations, network architecture, weights, biases, transfer-threshold functions were incorporated into a data spreadsheet (after Millie et al., 2012). The magnitude/

direction of synaptic weights among nodes of trained networks were depicted via neural interpretive diagrams (NIDs; Olden, 2000). The uncertainties associated with predictor influences were characterized by sensitivity analyses and connected weights analyses. Sensitivity about-the-mean analysis (Saltelli et al., 2004) assessed variation in CHL *a* attributable to deviations of each predictor while other predictors remained fixed at their respective means; ± one and two standard deviations from the mean discerned the most influencing predictor(s) during common and disturbance variation, respectively (Jeong et al., 2003). A connected weights analysis (Olden and Jackson, 2002) determined the relative influence for predictors upon modeled outcomes. The variable pairs deemed to have greatest impacts upon CHL prediction were varied across their respective minimum–maximum ranges, with multidimensional output response surfaces generated via the reproduced network computations (after Millie et al., 2012). Planar equations then were fit to the modeled surfaces (*TableCurve 3D*, v4.0 software; Systat, Inc., Chicago, IL USA).

The algorithm, TREes Parroting Networks (TREPAN; see Young et al., 2011), extracted logical 'rules' (as a decision tree) for CHL classes from continuous networks. Briefly, TREPAN treated rule-extraction as an inductive learning problem with the MLP queried as an 'oracle' to induce decision outcomes representative of network prediction. Decision rules for class output were expressed in statements using single attributes and Boolean conditions (i.e. specified by an integer threshold, '*m*', for a set of '*n*' possible circumstances).

3. Results

The variables, PRECIP and UR-N, (and to a slighter degree, TURB) displayed the greatest variation among predictors (Fig. 2A). The lesser and consistent coefficients of variation for TEMP, ATEMP, SAL, BP, RH and DO, identified these variables to have the least dispersion across their data ranges. Intra-annual variability, with respect to sampling date, was evident within the PCA (Fig. 2B). The initial two components included descriptors indicative of hydrological/meteorological forcing (TEMP, SAL, CURSPD, RH) and proxy variables for algal production (DO, pH) in explaining ca. 55% of the variance within the data set. However, ca. 15% of the variance could not be accounted for by the PCA; after the initial two components, the cumulative variances explained by successively adding the third, fourth, and fifth component axes were 68.2%, 77.8%, and 85.2%, respectively.

Concentrations of CHL *a* ranged from 0.1 to 47.9 μg L^{-1} (mean ± SE, 10.52 ± 0.13; $n = 3376$), with a coefficient of variation of 0.72. Mean weekly concentrations varied in concert with both hydrological and meteorological regimes (compare Fig. 2B and C); the least concentrations occurred within the initial and final weeks of the study (June/July and October/November, respectively), with the greatest concentrations in mid/late August and early September. Maxima in CHL *a* concentrations tracked PRECIP (Fig. 3); following PRECIP exceeding ca. 30 mm hr^{-1}, SAL decreased and preceded an increase in CHL. Concentrations of CHL within this database previously were identified to display a non-normal probability distribution, heteroscedasticity, and conditions of nonlinearity (Millie et al., 2012).

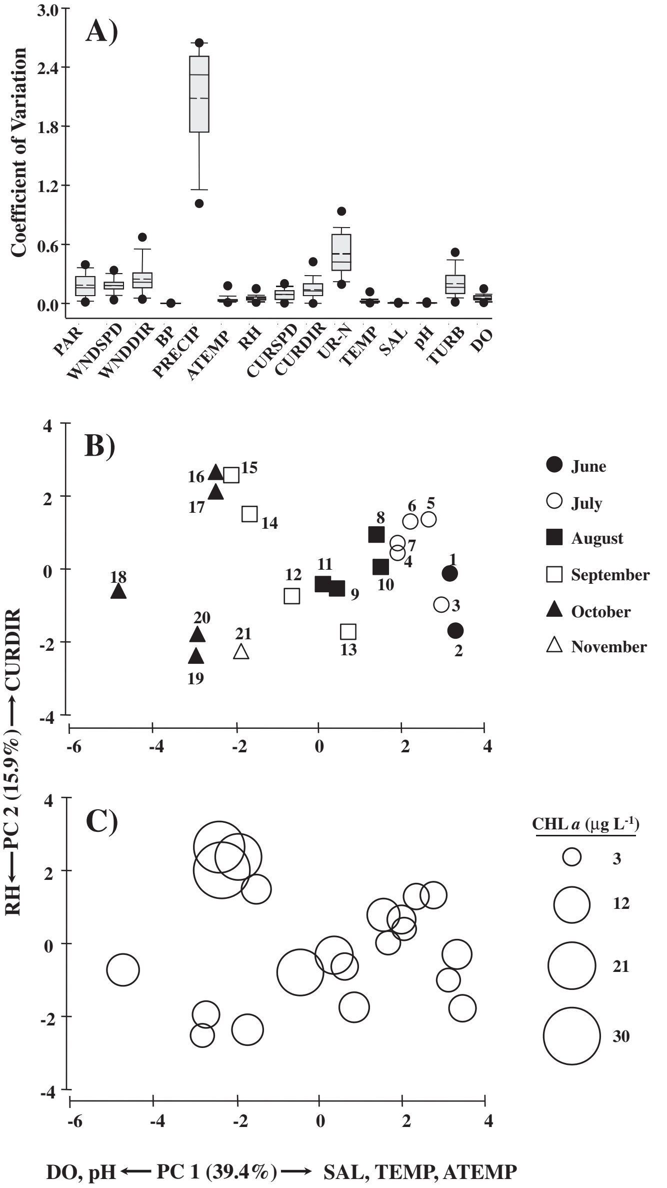

Fig. 2. A) Box and whisker plots for coefficients of variation, derived from weekly means of hydrological/meteorological variables. For each box, the dashed line indicates the mean value, boundaries closest/farthest to the solid (median) line signify the 25th/75th data percentiles, the whiskers (error bars) denote the 10th/90th data percentiles, and the dots represent the 5th/95th percentiles of data outliers. B and C) A principal component ordination for (weekly mean values of) environmental variables. Numbers in parentheses along axes represent the percent variability explained by the corresponding component. B) Sampling dates as a function of month. Symbols are numbered as the weekly interval for the study period. C) Mean chlorophyll concentrations superimposed as symbols of differing sizes—the larger the symbol, the greater the relative value, on ordination. See Methods for abbreviations.

3.1. Network modeling

MLPs, utilizing distinct supervised/unsupervised learning topologies, distinct transfer functions and learning algorithms, were surveyed as models for CHL *a* concentrations. For supervised learning of continuous data, an optimal topology (of six/two PEs within HL one/two, one output) was developed with training and cross-validated data sets ($n = 2026$ and 504, respectively), prior to its application to test data ($n = 844$). Hybrid networks utilized a PCA prior to an optimized MLP (with 10 PEs in one HL). Nearly all supervised networks performed well in terms of estimating continuous CHL *a* concentrations from predictors for the test data application (typically $\rho > 0.90$ for measured:modeled concentrations, with model efficiencies typically ranging from 0.80 to 0.90).

Prediction of CHL *a* via supervised ANNs involved complex interactions among predictors and produced an optimal model consisting of 360 synaptic influences throughout the network's

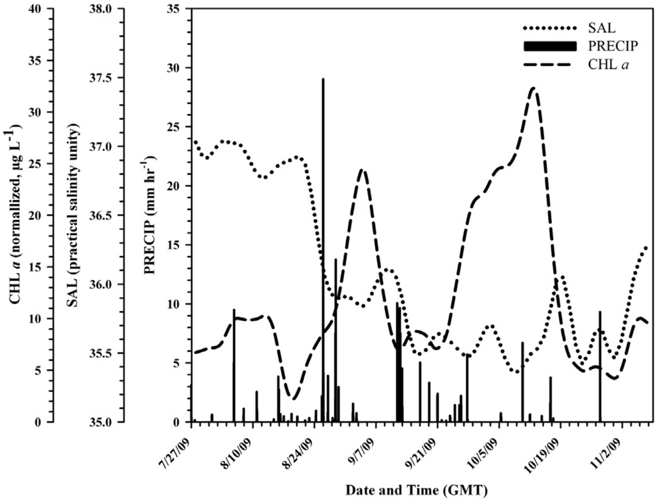

Fig. 3. Precipitation (PRECIP) and salinity (SAL) values and chlorophyll (CHL) *a* concentrations as a function of sampling date. A locally-weighted scatterplot smoothing algorithm was applied to the data to reveal holistic trends.

topology (Fig. 4). This network, incorporating sigmoid transfer functions and a Levenberg–Marquardt learning algorithm, provided a superior estimate of prediction (Fig. 5A), with the best performance metrics of the evaluated models (see Supplemental Table 1). Networks comprised of hyperbolic tangent transfer functions and either a Levenberg–Marquardt or momentum learning rule, provided slightly lesser performance than the aforementioned network. Collectively, those networks incorporating unsupervised and supervised learning underperformed networks incorporating supervised learning ($\rho \approx 0.87$ for measured:modeled concentrations with greater values for associated error/performance metrics and lesser modeling efficiencies).

Contemporary/time-lagged SAL and TEMP were identified via connected weights analyses (Fig. 5B) to have consistent, principal significant impacts upon modeled concentrations, with pH and TURB additionally having influences. SAL and TEMP$_\text{L}$ were confirmed via sensitivity analysis to have the greatest predictive influence upon CHL *a* (Fig. 5C). Concentrations of CHL *a* displayed converse, non-linear associations to these predictors; concentrations were greatest at SALs less than ca. 36 (maximizing at ca. 35.3, before declining) and steadily increased as TEMP increased beyond 28° C (Fig. 5D). The varying of these predictors across their respective minimum–maximum data ranges (while holding values of other predictors to their sample means) within network computations provided for a three-dimensional response surface (Fig. 6A), that depicted the interacting, curvilinear influences of SAL and TEMP upon modeled concentrations. A response surface generated via a 10th-order Chebyshev (x, *ln*y bivariate) polynomial equation was equivalent to the modeled surface (Fig. 6B, see Supplemental Equation).

CHL concentrations were partitioned into groups of $\leq$6.1 µg L^{-1} ($n = 983$, 29.12% of total data), >6.1 to $\leq$11 µg L^{-1} ($n = 1456$, 43.13%), and >11 µg L^{-1} ($n = 937$, 27.75%). The optimal training network identified for classification MLPs (of 10/eight PEs within HLs one/two and incorporating hyperbolic tangent transfer functions) produced an enormously complicated topology consisting of 7200 potential synaptic influences upon the modeled response (NID not shown). MLPs incorporating other transfer functions failed to successfully classify data. Classification networks ordered data well, with an accuracy of ca. 83% and class precision values ranging between 0.75 and 0.91 (see Supplemental Table 2). Across classes, the greater true negative rate (mean ca. 0.91) than that of true positive rate (mean ca. 0.84) signified that networks ordered an exemplar better in terms of the classes to which 'it did not belong' rather than that of the class to which 'it did belong'. The network incorporating the SVM did not improve upon classification performances of the MLPs.

Based upon plots arising from the decomposition of tertiary confusion matrices into dualistic outcomes, networks classified data more correctly in the extreme classes than the intermediate class (successively, >11 µg L^{-1}, $\leq$6.1 µg L^{-1}, >6.1 to $\leq$11 µg L^{-1}). The network incorporating hyperbolic tangent transfer functions and a momentum learning rule provided the best prediction across all

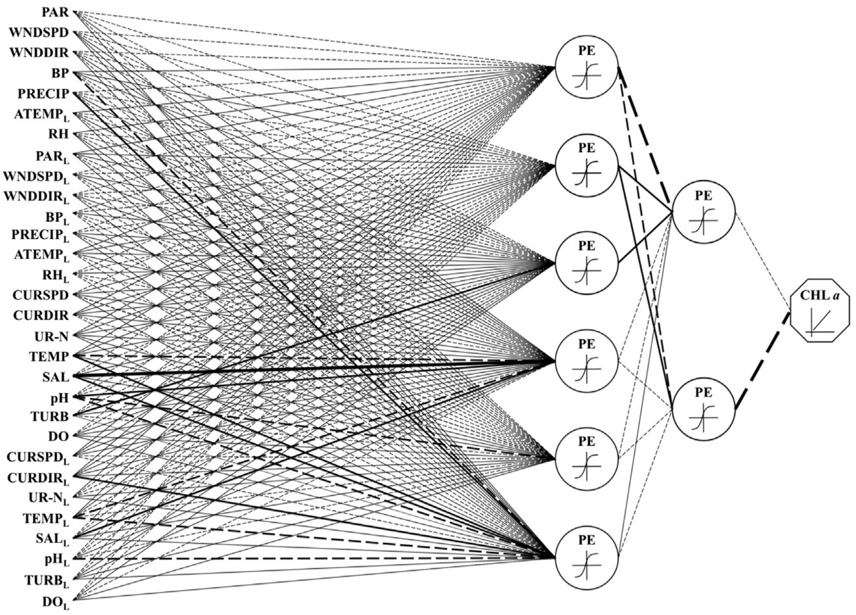

Fig. 4. A network interpretation diagram for the optimal continuous network (see Fig. 1). Dashed and solid lines depict negative (inhibitory) or positive (excitatory) effects, respectively, upon modeled chlorophyll (CHL) *a* concentrations by synaptic weights. Line thickness portrays the relative magnitude of the weight (i.e. greater values indicate more significance in prediction). Contrasting inhibitory/excitatory weights entering the same processing element (PE) denotes interaction. See Methods for variable abbreviations.

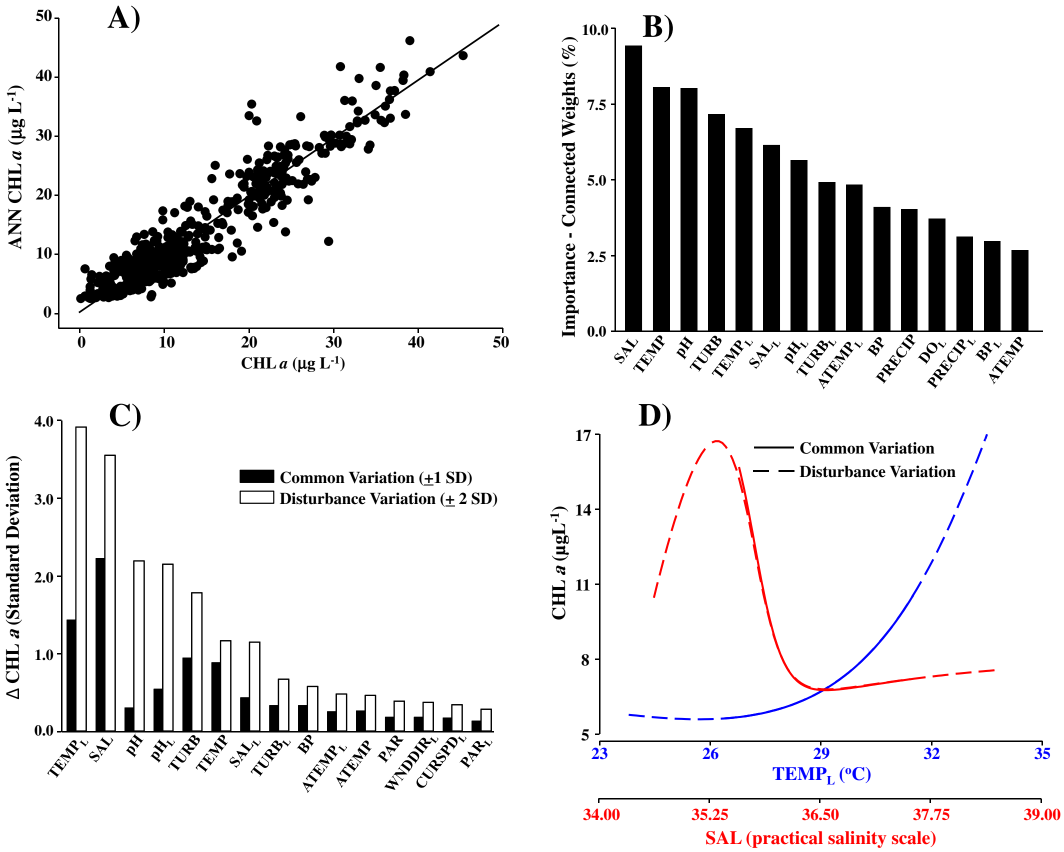

Fig. 5. A) Modeled chlorophyll (CHL) a concentrations as a function of measured concentrations for the optimal continuous network. The solid line represents a 1:1 relationship. B) The relative share of prediction associated with model inputs, as determined via the connected weights analysis upon training data. C and D) Results of sensitivity analyses performed on training data across common and disturbance variations of predictors. B and C) The initial 50% of variables having the greatest influence in sensitivity/connected weights analyses are depicted. See Methods for variable abbreviations. D) Modeled concentrations overlaid as functions of 'time-lagged' temperature ($TEMP_L$) and salinity (SAL).

CHL classes (Fig. 7A, see Supplemental Table 2). Connected weights analysis identified (contemporary and time-lagged) TURB and SAL and TEMP to have consistent, significant predictive influences upon CHL classification (Fig. 7B). Sensitivity analyses identified SAL and TEMP to be the most important variables for the ordering of data into the $\leq$6.1 µg L^{-1} and >6.1 to $\leq$11 µg L^{-1} classes (Fig. 8A).

The occurrence probabilities for concentrations of $\leq$6.1 µg L^{-1} and >6.1 to $\leq$11 µg L^{-1} were maximal and minimal at TEMPs of ca. 29.5 °C and 33.5 °C, respectively (Fig. 8B). Conversely, the occurrence probability for concentrations of >6.1 to $\leq$11 µg L^{-1} maximized at a SAL of ca. 36.5, whereas the probability for concentrations of $\leq$6.1 µg L^{-1} maximized at a SAL of ca. 35 (Fig. 8C). Concentrations of >11 µg L^{-1} had the greatest likelihood of occurrence within a water-column of SALs less than ca. 35.3 and greater than ca. 31.5 °C. Yet, the maximal probabilities of occurrence (ca. 0.18) for CHL concentrations of >11 µg L^{-1} were appreciably less than the greatest probabilities (ranging from 0.80 to 0.95) for concentrations $\leq$11 µg L^{-1}.

The TREPAN algorithm induced tree-based representations for CHL classifications across high-dimensional input space (i.e. 30 predictors). Although overall accuracy of classification for training data was noteworthy, the decision tree using single attributes underperformed a tree using M-of-N formalism (see Supplemental Table 2), undoubtedly due to lessened rule possibilities for derivation of 'leaves'. Logical rules applied to test data holistically ordered CHL classes at lesser modeling efficiencies than those of MLPs and the SVM, primarily due to lesser precisions in classifying concentrations $\leq$11 µg L^{-1} (see Supplemental Table 2, Fig. 6A). The primary variables identified for logical rules in the tree using M-of-N formalism (i.e. contemporary/time-lagged TURB, TEMP, SAL; Table 1) also were identified by the sensitivity and connected weights analyses as the variables having the greatest predictive influences. The tree relied (to a lesser degree) upon the variables WNDSPD, WNDDIR, CURDIR, PAR, and UR-N.

4. Discussion

An obstacle to the reliance upon instrumental technologies as means to guide coastal stewardship, is the resulting 'Big Data' that coastal scientists/managers must cope with. Accordingly, the discovery of relevant information arising via empirical modeling of such seemingly endless, sometimes extraneous data quantities is central to our comprehension of ecosystem alteration. Superficially, such a task appears ill-posed and dauntingly 'Herculean' in effort (after Müller and Lemke, 1999; Suryanarayana et al., 2008). Yet, continuous and categorical ANNs effectively reproduced the complex, non-linear interactions across high-dimensional input space and portrayed the intra-annual variance and magnitude of CHL *a* for Sarasota Bay. Sensitivity/connected weights analyses and multi-dimensional visualization deconvolved knowledge from the complexities of variable interactions and predictive uncertainties. The TREPAN algorithm induced symbolic translation of predictive outcomes, affording defined 'rule sets' for the delineation of CHL classes. Taken together, these analyses provided a comprehensive representation of the non-linear relationships between environmental variable(s) and CHL *a*.

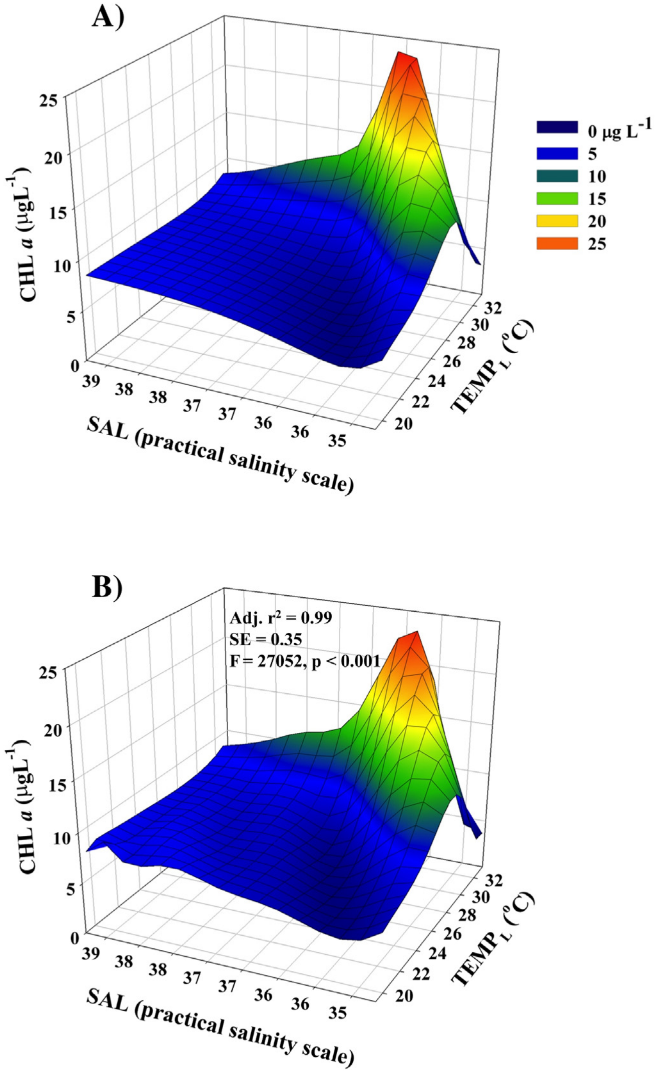

Fig. 6. A) Artificial neural network-derived response surface for chlorophyll (CHL) as a function of time-lagged temperature ($TEMP_L$) and salinity (SAL), varied across their data ranges. B) Response surface generated via a 10th-order Chebyshev bivariate polynomial equation for CHL as a function of $TEMP_L$ and SAL (see Results). Statistical information (adj. r^2, adjusted coefficient of determination; SE, standard error of fit; F, F-statistic for $n = 256$) denotes the goodness of fit arising from regression analysis in applying the polynomial equation to the ANN-derived response plane.

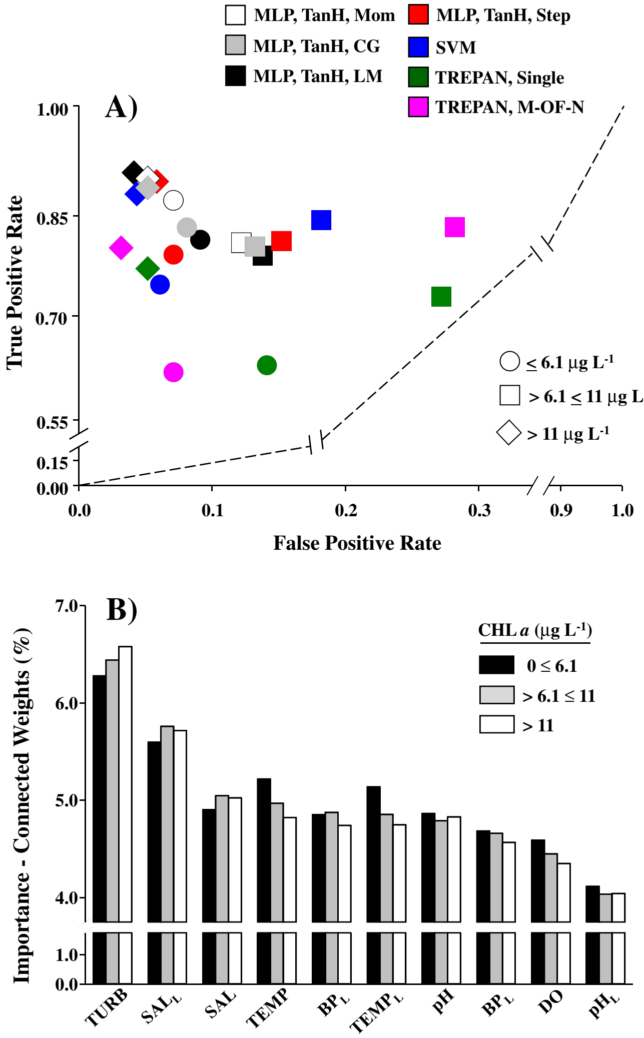

Fig. 7. A) True positive rate (TPR) as a function of false positive rate (FPR) for ordering of chlorophyll (CHL) classes within test data, as derived via categorical multi-layer perceptrons (MLPs), support vector machines (SVMs), and TREPAN algorithms. Note: 100% perfect and imperfect classification results in a datum at 0% FPR/100% TPR and 100% FPR/0% TPR, respectively. A datum lying upon the dashed line (0% FPR/0% TPR to 100% TPR/100% FPR) signifies ordering that is no better than random. B) The relative share of prediction associated with model inputs, as determined from connected weights analysis upon training data for the optimal categorical network (see text). The initial 50% of variables having the greatest influence are depicted. See Methods for variable abbreviations.

4.1. Predictor influences

SAL and TEMP are considered 'ecological master factors' within the estuarine environment (McKenney, 1987). Both variables were highly conservative within Sarasota Bay and provided the principal predictive influences upon CHL *a*. Influences were most prominent for concentrations <11 μg L^{-1} and appeared to be driven by seasonal meteorological conditions within tropical-temperate south Florida. The exchange of Bay waters with the GOMx is limited and the Bay has no major tributary. Episodic pulses of freshwater arising from runoff following heightened PRECIP during the 'wet' season (May to October) lessens the SAL of the Bay's metahaline waters, possibly exerting short-term osmotic or related cellular stress upon phytoplankton (Guillard, 1962; c.f.; Roubeix and Lancelot, 2008). Differences in SAL between runoff and Bay water masses may transform localized phytoplankton assemblages in their entirety, altering biomass distributions and compositional structures (Mikhail, 2008). Predictive uncertainty of CHL also can arise through controls of TEMP upon cell growth of an existing assemblage (Eppley, 1972) and/or responses for new populations arising via selection to changing TEMPs (i.e. a species' thermal niche, Thomas et al., 2012). Reflecting assemblage responses, in part, to altered SALs and TEMPs, concentrations of CHL within Florida's estuaries often are greatest during the wet summer months (e.g. Doering and Chamberlain, 1998; Millie et al., 2004). From plots of hourly data (Fig. 3), a time lag of several days appeared to occur between PRECIP, SAL minima, and increases in CHL. As such, the time lag utilized in network models would not have been sufficient to fully capture the temporal sequence of runoff-induced alterations in SAL and phytoplankton response (as CHL *a*).

Concentrations of 6.1 and 11 μg CHL *a* L^{-1} have been targeted as thresholds from which to infer eutrophication of the Bay and State of Florida-impaired waters, respectively. However, Brenden et al. (2008) noted that threshold identification is difficult during instances of dynamic, interacting environmental conditions (such as

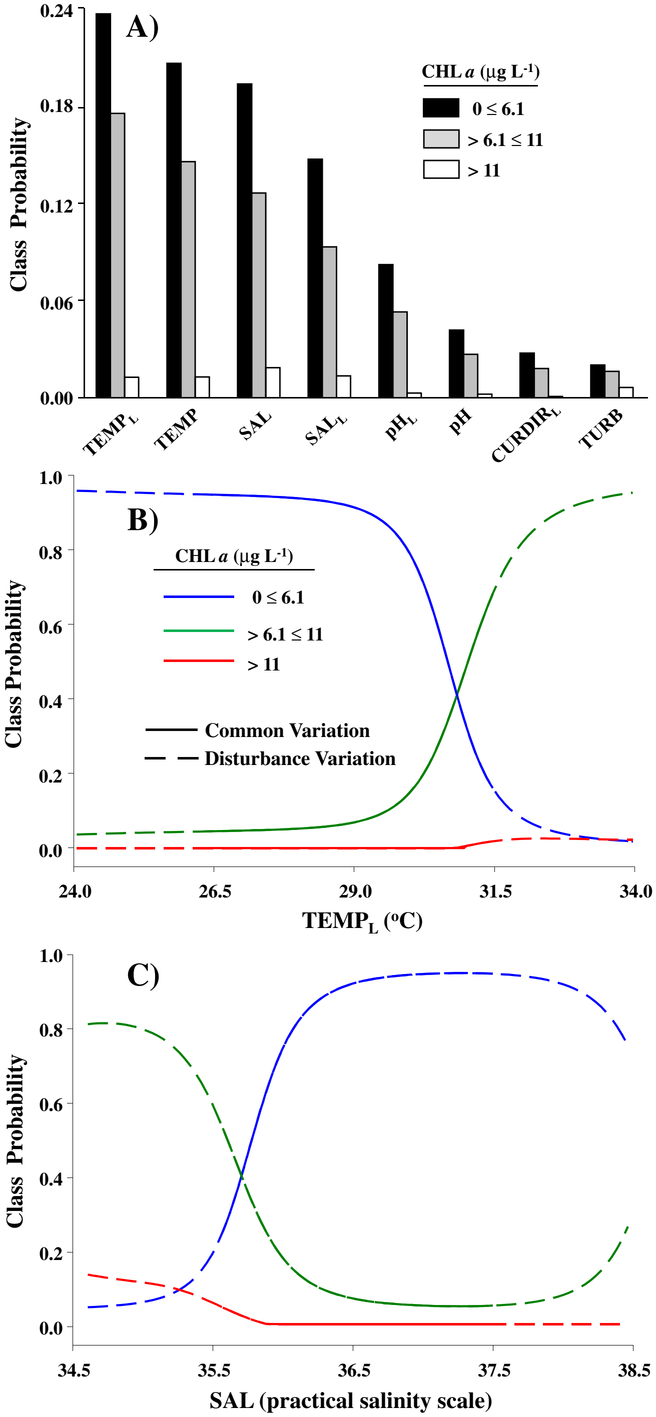

Fig. 8. A) Results of a sensitivity analyses performed on training data for the optimal, categorical network across common and disturbance variations of predictors. The initial 50% of variables having the greatest influence are depicted. See Methods for variable abbreviations. Note: because categorical networks utilized entropy error functions, modeled values represent the probability (≥0 to ≤1) for the classification state. B & C) Class probabilities are plotted as a function of C) 'time-lagged' temperature ($TEMP_L$) and D) salinity (SAL).

those in estuaries) and advocated the use of robust, quantitative approaches that included visualization to circumvent difficulties in data interpretation. Sensitivity analyses identified the occurrence likelihood of CHL concentrations > 6.1 µg L^{-1} in the Bay to occur independently at TEMPs greater than ca. 31 °C and SALs less than ca. 36. The output response surface portrayed non-linear uncertainties attributable to TEMP and SAL interaction throughout the range of modeled CHL concentrations, affording a multi-dimensional visualization of the environmental 'window of opportunity' for maximal biomass, conditional to these factors. Notably, the 10th-order Chebyshev polynomial equation (see Supplemental equation) provided an algorithmic function for the response surface. Modeled surfaces and algorithmic expressions for relationships between any interacting predictor pairs and CHL similarly can be derived (see Millie et al., 2012). Such pedagogical knowledge potentials for phytoplankton-environmental interactions are akin to niche modeling where species distributions in response to environmental, geographic and/or historic influences are visualized and quantified across environmental gradients.

Predictive uncertainties for CHL *a* concentrations also were identified via sensitivity/connected weights analyses and tree-conditional rules to be largely influenced by variables acting as proxy measurements of phytoplankton biomass and production (TURB and pH, respectively) and variables corresponding to descriptors of physical forcing (WNDSPD, WNDDIR) upon near-surface algal accumulations (see Paerl, 1988; Paerl et al., 1998; Millie et al., 2011). Identification of such influential variables (from candidate hydrological/meteorological predictors) for local algal accumulations would be anticipated within empirical models derived from a system, such as Sarasota Bay, that lacks the hydrological influences of a major tributary. Alternatively, the greater coefficients of variation for PRECIP and UR-N (than those for other potential predictors) signified these variables to be conditionally dependent upon and/or representative of episodic environmental influences, thereby highlighting their unreliability, and suitability as predictors for CHL.

4.2. Interpretation of network validity

The usefulness of ecological models is defined through the legitimacy of the variables available for incorporation. The validity of the generated ANNs was based upon the assumption that pertinent information for Sarasota Bay was contained within, and reproduced via the predictor-response variables for the observational database. The relationships between environmental predictors and CHL *a* were assessed over data ranges in their entirety, reducing the likelihood that data outliers might bias prediction and/or subsequent interpretation of outcomes. Sensitivity analysis conveyed quantitative uncertainties for SAL and TEMP upon modeled CHL concentrations. However, identification of single variable effects (while discounting the interacting complexities of other predictors) is somewhat disingenuous, particularly when multiple predictors control model output and an interactive non-linearity in response:predictor relationships exists (Sun et al., 2012; see Fig. 4). Supplemental analyses (connected weights, multi-dimensional visualization, TREPAN) facilitated in maintaining variable relationships within networks and identified complexities that typified the 'real world' of Sarasota Bay where dependent, linear associations among single variables do not exist.

ANNs require large amounts of data to ensure reproduction of variable interactions; a data luxury that coastal monitoring programs based upon invasive sampling characteristically do not provide, but programs utilizing instrumental detection would. Site-specific models (such as those presented here) achieve greater predictive accuracy than universal models and are appropriate for assessing phytoplankton dynamics within the spatially- and temporally-explicit scales required to evaluate system processes affecting such alterations (e.g. nutrient loading, wind mixing, frontal eddies, density gradients, tidal influences; Millie et al., 1995). Yet, despite modeling efficiencies of ca. 90% and the

Table 1

Decision tree (M-of-N formalism) induced via the TREPAN algorithm for CHL classes within the training data. Numbers in parenthesis following class designation denote instances of correct/total classifications for that tree node (bold type). See text for predictor abbreviations and units. Predictive accuracy and class precision for the training data follow rules (see Supplemental Table 2 for performance metrics of logical rule applications to the test data).

| 3 of {TEMP > 26.05, SAL$_L$ ≤ 36.04, TURB > 1.80}:

|| 2 of {ATEMP$_L$ > 27.40, CURSPD ≤ 0.05, TURB$_L$ > 1.85}: **> 11 µg L^{-1} (396/402)**

|| NOT 2 of {ATEMP$_L$ > 27.40, CURSPD ≤ 0.05, TURB$_L$ > 1.85}: **> 6.1 - ≤ 11 µg L^{-1} (35/58)**

| NOT 3 of {TEMP > 26.05, SAL$_L$ ≤ 36.04, TURB > 1.80}:

|| 5 of {UR ≤ 4.49, pH$_L$ ≤ 7.91, TURB$_L$ > 0.85 SAL ≤ 37.20, ATEMP$_L$ > 28.56, TURB$_L$ > 1.90, SA$_L$ ≤ 36.99}:

||| 3 of {WNDSPD > 7.45, WNDSPD$_L$ ≤ 5.90, RH ≤ 61.83, WNDSPD$_L$ > 6.05, TEMP$_L$ > 31.47, UR$_L$ ≤ 0.56, WNDSPD ≤ 2.85}:

||||2 of {WNDDIR > 166.80, DO ≤ 6.57, DO$_L$ > 7.16, BP$_L$ ≤ 101.47}: **> 11 µg L^{-1} (26/26)**

||||NOT 2 of { WNDDIR > 166.80, DO ≤ 6.57, DO$_L$ > 7.16, BP$_L$ ≤ 101.47}:

||||| 4 of {RH$_L$ > 66.47, pH ≤ 8.17, BP ≤ 101.61, ATEMP ≤ 26.26, ATEMP > 28.24, CURDIR$_L$ ≤ 148.29, CURDIR$_L$ > 177.14, pH ≤ 8.12}: **> 6.1 - ≤ 11 µg L^{-1} (26/27)**

||||| NOT 4 of { RH$_L$ > 66.47, pH ≤ 8.17, BP ≤ 101.61, ATEMP ≤ 26.26, ATEMP > 28.24, CURDIR$_L$ ≤ 148.29, CURDIR$_L$ > 177.14, pH ≤ 8.12}: **> 11 µg L^{-1} (29/40)**

||| NOT 3 of { WNDSPD > 7.45, WNDSPD$_L$ ≤ 5.90, RH ≤ 61.83, WNDSPD$_L$ > 6.05, TEMP$_L$ > 31.47, UR$_L$ ≤ 0.56, WNDSPD ≤ 2.85}: **> 6.1 - ≤ 11 µg L^{-1} (196/234)**

|| NOT 5 of { UR ≤ 4.49, pH$_L$ ≤ 7.91, TURB$_L$ > 0.85 SAL ≤ 37.20, ATEMP$_L$ > 28.56, TURB$_L$ > 1.90, SA$_L$ ≤ 36.99}:

||| SAL$_L$ ≤ 38.08:

||||6 of {BP ≤ 101.74, WNDDIR$_L$ ≤ 296.45, WNDDIR$_L$ > 305.90, PAR > 1207.30, PAR$_L$ ≤ 277.50, BP > 101.34, WNDDIR ≤ 100.25, RH$_L$ > 70.98, WNDDIR > 258.80, TEMP$_L$ ≤ 25.83}:

||||| TURB ≤ 1.95: **≤ 6.1 µg L^{-1} (205/234)**

||||| TURB > 1.95: **> 6.1 - ≤ 11 µg L^{-1} (12/16)**

||||NOT 6 of { BP ≤ 101.74, WNDDIR$_L$ ≤ 296.45, WNDDIR$_L$ > 305.90, PAR > 1207.30, PAR$_L$ ≤ 277.50, BP > 101.34, WNDDIR ≤ 100.25, RH$_L$ > 70.98, WNDDIR > 258.80, TEMP$_L$ ≤ 25.83}:

||||| TURB ≤ 1.21:

|||||| 2 of {UR$_L$ ≤ 7.23, pH > 7.98}: **≤ 6.1 µg L^{-1} (106/126)**

|||||| NOT 2 of {UR$_L$ ≤ 7.23, pH > 7.98}:

||||||| SAL$_L$ ≤ 37.05: **> 6.1 - ≤ 11 µg L^{-1}** (91/102)

||||||| SAL$_L$ ≤ 37.05:

|||||||| BP ≤ 101.88: **> 6.1 - ≤ 11 µg L^{-1} (36/51)**

|||||||| BP ≤ 101.88:: **≤ 6.1 µg L^{-1} (18/20)**

||||| TURB > 1.21:

|||||| 4 of {RH ≤ 69.81, CURDIR$_L$ > 290.89, pH ≤ 8.1, WNDSPD > 5.45, CURDIR > 141.62, ATEMP ≤ 25.66, pH ≤ 7.95}:

||||||| TURBL ≤ 2.75:

|||||||| pH$_L$ ≤ 8.15: **≤ 6.1 µg L^{-1} (63/80)**

|||||||| pH$_L$ > 8.15: **> 6.1 - ≤ 11 µg L^{-1} (12/17)**

||||||| TURB$_L$ > 2.75: **> 6.1 - ≤ 11 µg L^{-1} (10/10)**

|||||| NOT 4 of { RH ≤ 69.81, CURDIR$_L$ > 290.89, pH ≤ 8.1, WNDSPD > 5.45, CURDIR > 141.62, ATEMP ≤ 25.66, pH ≤ 7.95}:

||||||| TURB ≤ 1.65:

|||||||| 4 of {DO$_L$ > 6.09, DO ≤ 6.11, DO > 6.75, UR$_L$ > 0.31, WNDSPD ≤ 3.45, CURSPD ≤ 0.03}:

||||||||| pH$_L$ ≤ 8.01: **≤ 6.1 µg L^{-1} [7/7)**

||||||||| pH$_L$ > 8.01: **> 6.1 - ≤ 11 µg L^{-1} (29/57)**

|||||||| Not 4 of {DO$_L$ > 6.09, DO ≤ 6.11, DO > 6.75, UR$_L$ > 0.31, WNDSPD ≤ 3.45, CURSPD ≤ 0.03}:

||||||| TURB > 1.65: **> 6.1 - ≤ 11 µg L^{-1} (173/186)**

||| SAL$_L$ > 38.08: **> 6.1 - ≤ 11 µg L^{-1} (173/186)**

Accuracy: 0.75

Precision: ≤ 6.1 µg L^{-1}: 0.62; > 6.1 - ≤ 11 µg L^{-1}: 0.84; > 11 µg L^{-1}: 0.7

delineation of quantitative thresholds for SAL and TEMP upon CHL *a*, the derived networks provided only a limited understanding of the mechanisms underlying environmental forcing within Sarasota Bay. In actuality, the legitimacy of the networks is constrained; models afforded prediction of (and interpretation for) CHL *a* dynamics specific to the numerical limits and variable associations for the data 'at hand', that were presumed representative of and embodying the system's holistic variability. Consequently, the fidelity for such local models (and reliability outside of the data ranges from which they were generated) is conditional upon the amount of system variability omitted from the sample data (c.f. Peck et al., 2003) and any subjectivity introduced upon assigning data to training/testing subsets.

5. Conclusions and relevancy of NIC to coastal data

The ability to extract relationships from and form models for 'Big Data' is the basis of NIC. In applications to Sarasota Bay, NIC played a role in discovery science, leading to conceptualizations of CHL response to environmental forcing, (e.g. the delineation of SAL/TEMP thresholds, a quantitative expression for the SAL-TEMP interaction). Nevertheless, the inherent inability of ANNs to directly provide decipherable knowledge appears a significant obstacle to their application to coastal management. Network comprehensibility is important for user validation of model formulation and knowledge discovery, but mandatory for decision making in natural resource stewardship.

Overcoming the 'black box' nature of MLPs and despite the derivation of a polynomial equation for the visualized CHL response surface, the logical rules of decision trees appears most attractive to coastal managers, being easily interpretable through 'if-then' or Boolean conditional statements. Categorical tree algorithms (e.g. C4.5) only recently have been utilized for hydrological applications (e.g. Wei and Hsu, 2012; c.f. regression trees; Solomatine, 2006). However, the question remains, "What advantages might ANN-based decision trees, such as TREPAN, afford coastal management?" Simply, TREPAN integrates the superior non-linear modeling capabilities of ANNs with user-friendly interpretability, creating a 'best of both worlds' system for prediction accuracy and decision-making comprehensibility that is scalable to models having highly complex input space (c.f. Wei and Hsu, 2012). Moreover, a liability to classification ANNs is the requirement for balanced data for model training, validation and testing. If sample sizes for a classifier within an imbalanced data set are limited during the tree-induction process, TREPAN relies upon the network 'oracle' to induce artificial (yet realistic) sample instances.

Regional forecasting efforts in support of curtailing the eutrophication of coastal waters requires geographically-expansive monitoring capable of generating the high-resolution temporal data required to resolve local response(s) to natural and anthropogenic perturbations. However, to build comprehension of ecological structure from monitoring databases for dynamic coastal waters, one first must mine and model data in lesser, directed tasks (Suryanarayana et al., 2008), and in effect, "… build the big structure from estimates of local structure, rather than assuming a simple form for the big structure…" (McCune, 2006). To this end, NIC affords coastal informatics much more than local prediction. Data streams generated via instrumental platforms can be transmitted (in real/near-real time) to onshore computational facilities via wireless communication. This acquisition of contemporary information allows users to update existing models based upon historical data, while improving upon the utilization of pertinent coastal indicators and comprehension of response deviations to evolving perturbations. Models also could be programmed into data-logging processors onboard instrument platforms to provide 'smart' control of instrument operations and enable event-specific data acquisition.

Nonetheless, it is important to remember that environmental models serve as representations of our best, current understanding for a given system (i.e. not as statements of truth) and as such, should not be construed as a panacea for decision making (Starfield, 1997; Swannack et al., 2012). The usefulness of empirical models within coastal resource management is based upon our understanding that the data utilized in model development exemplify discontinuous 'snapshots' of the ecological continuum and are sufficient to inform, not dictate, decision-making processes (after Millie et al., 2006b; Swannack et al., 2012). Accordingly, the greatest impact that NIC may have upon coastal management is the provision of local knowledge in support of meso-/regional-scale modeling efforts. Definitions of site-specific biota-environmental relationships (e.g. CHL *a* as a polynomial function of SAL and TEMP) could be fed into stand-alone mechanistic models, ultimately providing for the projection of alterations in system-/event-specific structure across sizeable geographic and temporal scales. In supporting the analysis of real-world complexity, the synoptic modeling and knowledge-derivation potentials provided by NIC can assist and improve upon the heralded capabilities afforded through remote sensing and GIS in environmental monitoring, theory generation and decision support (after Boyd and Foody, 2011).

Acknowledgments

This work is contribution number 10 of the Great Lakes Research Center, Michigan Technological University. Funding was provided by the Florida Department of Environmental Protection, Florida Coastal Management Program (through grants from the National Oceanic and Atmospheric Administration, Office of Ocean and Coastal Resource Management under the Coastal Zone Management Act of 1972, as amended, Award #s NA10NOS4190178 and NA11NOS4190073). Reference to proprietary names are necessary to report factually on available data; however, the authors' institutions neither guarantee nor warrant the standard of a product and imply no approval of a product to the exclusion of others that may be suitable.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ecss.2013.04.001.

References

Abrahart, B., See, L., Solomatine, D.P. (Eds.), 2008. Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications. Springer-Verlag, Berlin-Heidelberg, p. 522.

Bollier, D., 2010. The Promise and Peril of Big Data. The Aspen Institute, Washington, District of Columbia USA, p. 56.

Borja, A., Basset, A., Bricker, S., Dauvin, J., Elliot, M., Harrison, T., Marques, J., Weisberg, S., West, R., 2012. Classifying ecological quality and integrity of estuaries. In: Wolanski, E., McLusky, D. (Eds.), Treatise on Estuarine and Coastal Science. Academic Press, Waltham, Massachusetts, pp. 125–162.

Boyd, D.S., Foody, G.M., 2011. An overview of recent remote sensing and GIS based research in ecological informatics. Ecological Informatics 6, 25–36.

Brenden, T.O., Wang, L., Zhenming, S., 2008. Quantitative identification of disturbance thresholds in support of aquatic resource management. Environmental Management 42, 821–832.

Brown, C.D., Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: a tutorial. Chemometrics and Intelligent Laboratory Systems 80, 24–38.

Cole, R., Weisberg, R., Donovan, J., Merz, C., Russel, R., Subramanian, V., Luther, M., 2003. The evolution of a coastal mooring system. Sea Technology 44, 24–31.

de Castro, L.N., 2007. Fundamentals of natural computing: an overview. Physics of Life Reviews 4, 1–36.

Devlin, B., 1997. Data Warehouse: From Architecture to Implementation. Addison Wesley, Boston, p. 448.

Dibike, Y.B., Velickov, S., Solomatine, D.P., Abbott, M.B., 2001. Model induction with support vector machines: introduction and applications. ASCE Journal of Computing in Civil Engineering 15, 208–216.

Doering, P.H., Chamberlain, R.H., 1998. Water quality in the Caloosahatchee estuary, San Carlos Bay and Pine Island Sound, Florida. In: Treat, S.F. (Ed.), Proceedings of the Charlotte Harbor Public Conference and Technical Symposium. Charlotte Harbor National Estuary Program, Tampa, Florida, pp. 229–240.

Elder IV, J.F., Pregibon, D., 1996. A statistical perspective on knowledge discovery in databases. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press-The MIT Press, Cambridge, Massachusetts, pp. 83–116.

Eppley, R.W., 1972. Temperature and phytoplankton growth in the sea. Fishery Bulletin 70, 1063–1085.

Fries, D.P., Bhanushali, P.H., Wilson, J.A., Broadbent, H.A., Sanderson, A.C., 2008. Broadband, low-cost, coastal sensor nets. Oceanography 20, 150–155.

Guillard, R.R.L., 1962. Salt and osmotic balance. In: Lewin, R.A. (Ed.), Physiology and Biochemistry of Algae: 929. Academic Press, New York-London, pp. 529–540.

Håkanson, L., 2000. The role of characteristic coefficients of variation in uncertainty and sensitivity analyses, with examples related to the structuring of lake eutrophication models. Ecological Modelling 131, 1–20.

Holm-Hansen, O., Amos, A.F., Hewes, C.D., 2000. Reliability of estimating chlorophyll *a* concentrations in Antarctic waters by measurement of in situ chlorophyll a fluorescence. Marine Ecology Progress Series 196, 103–110.

Jannasch, H.W., Coletti, L.J., Johnson, K.S., Fitzwater, S.E., Needoba, J.A., Plant, J.N., 2008. The land/ocean biogeochemical observatory: a robust networked mooring system for continuously monitoring complex biogeochemical cycles in estuaries. Limnology and Oceanography, Methods 6, 263–276.

Jeong, K.-S., Recknagel, F., Joo, G.-J., 2003. Prediction and elucidation of population dynamics of the blue-green algae Microcystis aeruginosa and the diatom Stephanodiscus hantzschii in the Nakdong River-Reservoir system (South Korea) by a recurrent artificial neural network. In: Recknagel, F. (Ed.), Ecological Informatics; Understanding Ecology by Biologically-inspired Computation. Springer-Verlag, Berlin, pp. 195–213.

Jørgensen, S.E., Chon, T.-S., Recknagel, F.A. (Eds.), 2009. Handbook of Ecological Modelling and Informatics. WIT Press, Southhampton, p. 431.

McCune, B., 2006. Nonparametric Multiplicative Regression for Habitat Modeling. MjM Software Design, Gleneden Beach, Oregon, p. 52.

McKenney Jr., C.L., 1987. Optimization of Environmental Factors During the Life Cycle of *Mysidopsis bahia*. Research Brief EPA/600/M-87–004. U. S. Environmental Protection Agency, Environmental Research Laboratory, Gulf Breeze, FL, p. 6.

Mikhail, S.K., 2008. Dynamics of estuarine phytoplankton assemblages in Mex Bay, Alexandria (Egypt): influence of salinity gradients. Egyptian Journal of Aquatic Biology and Fisheries 12, 231–251.

Millie, D.F., Vinyard, B.T., Baker, M.C., Tucker, C.S., 1995. Testing the temporal and spatial validity of site-specific models derived from airborne remote sensing of phytoplankton. Canadian Journal of Fisheries and Aquatic Sciences 52, 1094–1107.

Millie, D.F., Carrick, H.J., Doerong, P.H., Steidinger, K.A., 2004. Intra-annual variability of water quality and phytoplankton within the St. Lucie River Estuary (Florida, USA): a quantitative perspective. Estuarine Coastal and Shelf Science 61, 137–149.

Millie, D.F., Weckman, G.R., Paerl, H.W., Pinckney, J.L., Bendis, B.J., Pigg, R.J., Fahnenstiel, G.L., 2006a. Neural network modeling of estuarine indicators: hindcasting phytoplankton biomass and net ecosystem production in the Neuse (North Carolina) and Trout (Florida) Rivers. Ecological Indicators 6, 589–608.

Millie, D.F., Pigg, R., Tester, P.A., Dyble, J., Litaker, R.W., Carrick, H.J., Fahnenstiel, G.L., 2006b. Modeling phytoplankton abundance in Saginaw Bay, Lake Huron: using artificial neural networks to discern functional influence of environmental variables and relevance to a Great Lakes Observing System. Journal of Phycology 42, 336–349.

Millie, D.F., Fahnenstiel, G.L., Weckman, G.R., Klarer, D.M., Dyble Bressie, J., Vanderploeg, H.A., Fishman, D., 2011. An 'enviro-informatic' assessment of Saginaw Bay (Lake Huron USA) phytoplankton: characterization and modeling of *Microcystis* (Cyanophyta). Journal of Phycology 47, 714–730.

Millie, D.F., Weckman, G.R., Young, W.A., Ivey, J.E., Fahnenstiel, G.L., 2012. Modeling algal abundance with artificial neural networks: demonstration of a heuristic, 'Grey-Box' technique to deconvolve and quantify environmental influences. Environmental Modelling and Software 38, 27–39.

Müller, J.-A., Lemke, F., 1999. Self-organizing Data Mining, An Intelligent Approach To Extract Knowledge From Data, first ed. Berlin-Dresden, p. 225. Available: http://www.knowledgeminer.com/pdf/mining.pdf.

Olden, J.D., 2000. An artificial neural network approach for studying phytoplankton succession. Hydrobiologia 436, 131–143.

Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": understanding variable contributions in artificial neural net- works. Ecological Modelling 154, 135–150.

Olden, J.D., Joy, A.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178, 389–397.

Paerl, H.W., 1988. Nuisance phytoplankton blooms in coastal, estuarine, and inland waters. Limnology & Oceanography 33, 823–847.

Paerl, H.W., Dyble, J., Pinckney, J.L., Valdes, L.M., Millie, D.F., Moisander, P.H., Morris, J.T., Bendis, B., Piehler, M.F., 2005. Using microalgal indicators to assess human and climatically-induced ecological change in estuaries. In: Bortone, S.A. (Ed.), Estuarine Indicators. CRC Press, Boca Raton, Florida, pp. 145–174.

Paerl, H.W., Pinckney, J.L., Fear, J.M., Peierls, B.L., 1998. Eco-system responses to internal and watershed organic matter loading: consequences for hypoxia in the eutrophying Neuse River Estuary, North Carolina, USA. Marine Ecology Progress Series 166, 17–25.

Peck, M.S., Leffler, A.J., Flint, S.D., Ryel, R.J., 2003. How much variance is explained by ecologists? Additional perspectives. Oecologia 137, 161–170.

Poole, D.L., Mackworth, A.K., 2010. Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press, New York, p. 682.

Reed, R.E., Burkholder, J.M., Allen, E.H., 2010. Current online monitoring technology for surveillance of algal blooms, potential toxicity, and physicalechemical structure in rivers, reservoirs, and lakes. In: American Water Works Association, Manual M57, Algae. American Water Works Association, Denver, Colorado, pp. 1–24.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. Journal of Machine Learning Research 5, 101–141.

Roubeix, V., Lancelot, C., 2008. Effect of salinity on growth, cell size and silicification of an euryhaline freshwater diatom: *Cyclotella meneghiniana* Kütz. Transitional Waters Bulletin 1, 31–38.

Saltelli, A., Tarantola, S., Compolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice: a Guide to Assessing Scientific Models. John Wiley & Sons, Ltd, West, Sussex, England, p. 232.

Sarasota Bay Estuary Program, 2010. Nutrient Criteria for Sarasota Bay, p. 142. Technical report. Available at: http://www.sarasotabay.org/documents/SBEP-NNC-Final-Report.pdf (accessed 01.12.12.).

Solomatine, D.P., July 2006. Optimal modularization of learning models in forecasting environmental variables. In: Voinov, A., Jakeman, A., Rizzoli, A. (Eds.), Proceedings of the iEMSs 3rd Biennial Meeting: Summit on Environmental Modelling and Software. International Environmental Modelling and Software Society Burlington, Vermont, USA. CD ROM. http://www.iemss.org/iemss2006/sessions/all.html.

Starfield, A.M., 1997. A pragmatic approach to modeling for wildlife management. Journal of Wildlife Management 61, 261–270.

Stow, C.A., Roessler, C., Borsuk, M.E., Bowen, J.D., Reckhow, K.H., 2003. Comparison of estuarine water quality models for total maximum daily load development in Neuse River Estuary. Journal of Water Resources Planning and Management 129, 307–314.

Sun, X.Y., Newham, L.T.H., Croke, B.F.W., Norton, J.P., 2012. Three complementary methods for sensitivity analysis of a water quality model. Environmental Modelling and Software 37, 19–29.

Suryanarayana, I., Braibanti, A., Rao, R.S., Ramam, V.A., Sudarsan, D., Rao, G.N., 2008. Neural networks in fisheries research. Fisheries Research 92, 115–139.

Swannack, T.M., Fischenich, J.C., Tazik, D.J., 2012. Ecological Modeling Guide for Ecosystem Restoration and Management. Report # ERDC/EL TR-12–18. U.S. Army Engineer Research and Development Center, Vicksburg, Mississippi, USA, p. 60.

Thomas, M.K., Kremer, C.T., Klausmeir, C.A., Litchman, E., 2012. A global pattern of thermal adaptation in marine phytoplankton (published online 25 October 2012). Science. http://dx.doi.org/10.1126/science.1224836.

Weckman, G.R., Millie, D.F., Ganduri, C., Rangwala, M., Young, W., Fahnenstiel, G.L., 2009. Knowledge extraction from the neural 'black box' in ecological monitoring. Journal of Industrial Systems Engineering 3, 38–55.

Wei, C.-C., Hsu, H.-H., 2012. Neural-based decision trees classification techniques: a case study in water resources management. In: Qian, Z., Cao, L., Su, W., Wang, T., Yang, H. (Eds.), Recent Advances in Computer Science and Information Engineering. Lecture Notes in Electrical Engineering, vol. 124. Springer, Berlin-Heidelberg, pp. 377–382.

Young II, W.A., Weckman, G.R., Rangwala, M.H., Whiting II, H.S., Paschold, H.W., Snow, A.H., Mourning, C.L., 2011. An investigation of TREPAN utilizing a continuous oracle model. International Journal of Data Analysis Techniques and Strategies 3, 325–352.

Zappala, G., Azzaro, F., 2004. A new generation of coastal monitoring platforms. Chemistry and Ecology 20, 387–398.