Ecological Informatics 34 (2016) 13–21

Contents lists available at ScienceDirect

Ecological Informatics

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

Two-phase approach to improve stream health modeling

Mohammad Abouali, A. Pouyan Nejadhashemi *, Fariborz Daneshvar, Sean A. Woznicki

Biosystems and Agricultural Engineering, Michigan State University, USA

ARTICLE INFO

Article history:
Received 24 February 2016
Received in revised form 19 April 2016
Accepted 24 April 2016
Available online 27 April 2016

Keywords:
PLSR
ANFIS
Biological indicators
Error prediction
Stream health
Biotic indices

ABSTRACT

Direct measurement of biotic indices used in monitoring stream health is time consuming, costly, and usually limited to few sites and locations. This severely limits the spatial extent and the temporal interval of assessment; hence, continuous long-term monitoring of all reaches becomes impossible. Therefore, modeling approaches are commonly used as an alternative. However, modeling complex natural systems are not without challenges and the error in modeling these systems is usually high. This study focuses on modeling four biotic indices, including one fish and three macroinvertebrate indices, using 171 water quantity and 78 water quality variables. This study introduces a new two-phase approach in modeling biotic indices. In the first phase, an initial estimate of the biotic index along with an estimate of the error associated with those initial predictions is obtained. In the second phase, these initial estimates are combined to develop a new predictive model. Although different modeling methods can be used in each phase, to demonstrate the concept, in this study we tested Partial Least Square Regression (PLSR) and Adaptive Neuro-Fuzzy Inference System (ANFIS). The proposed approach was evaluated based on monitoring data form the Flint River watershed, located in Michigan, USA. The results demonstrate that the two-phase approach that uses PLSR (first phase) and ANFIS (second phase) is superior to common-single-phase approach (R^2 for the stream health predictive models increased on average from 0.5 in the first phase to over 0.9 in the second phase). Additionally, the two-phase approach eliminates the need for variable selection, a common pre-processing step, and provides satisfactory results despite the limited number of samples, which makes the approach more reliable, robust, and applicable. Although in this study the proposed two-phase approach is applied to biotic indices, the process can be extended to other natural and physical systems.

1. Introduction

A stream ecosystem is comprised of a community of biotic organisms that interact with abiotic components of their environment in a systematic way (Fisher and Likens, 1972; Angelier, 2003). Human activities, such as urbanization and agricultural practices, disrupt these interactions and have negative impacts on stream ecosystems (Vitousek et al., 1997; Chapagain and Hoekstra, 2008; Brander et al., 2010). In order to monitor and evaluate anthropogenic impacts on stream ecosystems, biotic indices were developed (Karr et al., 1986; Kerans and Karr, 1994; Karr, 1999; Maddock, 1999). Biotic indices based on species abundance, richness, and trophic composition are commonly used for assessments of biological integrity (Herman and Nejadhashemi, 2015). Studies show that fish and macroinvertebrate indices as measures of biological integrity provide a reliable assessment of stream health/degradation (Brazner et al., 2007; Flinders et al., 2008; Pelletier et al., 2012). Due to the limited mobility of macroinvertebrates and their low tolerance to pollutants, macroinvertebrate indices are sensitive to local degradation (Kerans and Karr, 1994; Compin and Céréghino, 2003). On the other hand, fish indices are ideal for regional and long-term impact assessment due to abundance and mobility of fish within a stream network (Karr, 1981; Herman and Nejadhashemi, 2015).

Due to the cost of monitoring, biotic indices are only measured at a few sites within a river basin and it is almost impossible to perform continuous long-term measurements for all reaches. Modeling approaches such as empirical and linear regression are commonly used to fill this data gap (Dodds et al., 2002; Van sickle et al., 2004; Wang et al., 2007; Maret et al., 2010; Einheuser et al., 2012; Merriam et al., 2015).

Abbreviations: ANFIS, Adaptive Network Fuzzy Inference System; DH17, average high flow duration; DH21, 25th percentile of high flow duration; DH22, flood interval; DL16, low flow pulse duration; EPT, abundance of three species (*Ephemeroptera*, *Plecoptera*, and *Trichoptera*); FH2, variability in high pulse count; FH4, high flood pulse count; FIBI, Family-level Index of Biological Integrity; FL1, low flood pulse count; gauss2mf, Gaussian combination membership function; gaussmf, Gaussian curve membership function; gbellmf, generalized bell-shaped member ship function; HBI, Hilsenhoff Biotic Index; IBI, Index of Biotic Integrity; MA27, variability of flow values for month of April; MA34, variability of flow values for month of November; MA37, variability across monthly flows; MA40, skewness in the monthly flows; MH9, mean maximum flows for September; ML20, base flow; RA7, change of flow; PLSR, Partial Least Square Regression; RA8, number of reversals; Sed12, mean sediment for month of January; TA3, seasonal predictability of flooding; TH2, variability in Julian date of annual maxima; TH3, seasonal predictability of non-flooding; TL1, Julian date of annual minimum; TN21, mean total nitrogen for month of October; TN29, variability of total nitrogen for month of June; TP19, mean total phosphorus for month of August.

* Corresponding author.
E-mail address: pouyan@msu.edu (A.P. Nejadhashemi).

http://dx.doi.org/10.1016/j.ecoinf.2016.04.009
1574-9541/

However, linear models usually perform poorly in modeling most ecosystem responses to stressors (Scheffer et al., 2001; Johnson and Host, 2010); Using complex nonlinear methods, such as artificial neural networks (Céréghino et al., 2003; Park et al., 2004; Lencioni et al., 2007; Mathon et al., 2013) or fuzzy logic (Einheuser et al., 2012, 2013a, 2013b; Woznicki et al., 2016) has become more favorable. Nevertheless, in these types of models, the number of input variables should be limited (up to six variables) in order to reduce the number of fuzzy rules and the required computational power and time (Chen and Mynett, 2003; Sanikhani and Kisi, 2012; Woznicki et al., 2015).

While in general it is desirable to limit the number of input variables due to the associated cost for data collection or modeling limitations, it is important to include all significant variables in order to create a robust model. Additionally, having too few measurements for the model calibration, as is usually the case for macroinvertebrate and fish monitoring sites, imposes more restrictions on the number of input variables. As the number of input variables to the model increases, more data points are needed for training phase of model development. One solution is to perform additional field measurements that are expensive, time consuming, and in many cases cannot be accommodated due to the budget restrictions. Therefore, this study aims to develop a predictive stream health model that can use all the relevant variables, even if the number of measured data is limited. The ultimate goal of this paper is to introduce a new modeling approach that can be used in establishing a predictive model that relates a set of input variables or predictors to an output variable.

2. Materials & methods

In this study, the modeling process for estimating stream health indices is illustrated in Fig. 1. The process started with setup, calibration, and validation of a hydrologic model in order to estimate the stream flow and water quality variables for all stream segments within the study area. Next, the hydrologic model outputs were further parameterized using the Hydrologic Index Tool (Kennen et al., 2009) to calculate ecologically relevant variables that include 171 variables related to water quantity and 78 variables related to water quality for every stream segment within the study area. In the next step, variable selection can be performed, if necessary, to identify the most relevant variables. The generated variables are finally used to develop stream health models for one fish and three macroinvertebrate indices. Stream health models were developed using a two-phase approach. In the first phase, biotic indices were estimated along with the associated errors. In the second phase, these initial estimates (biotic indices and associated errors) were used as inputs for development of final stream health models.

2.1. Study area

The study area is the Flint River watershed (Hydrologic Unit Code 04080203) located in southeast Michigan, USA. The Flint River watershed has more than 3800 reach segments, while the Flint River is one of the major tributaries of the Saginaw River Watershed (040802) that drains to Lake Huron (Fig. 2). The Saginaw River Watershed (including Flint) is identified as an area of concern by U.S. Environmental Protection Agency (EPA) due to polluted sediments and significant loss of recreational values (EPA, 2015).

The total drainage area for the Flint River watershed is 3445 km^2 and consists of 40% forest, 25% agriculture, 18% pasture, 16% urban, and 1% water. The Flint River is an important food resource and spawning habitat for fish and invertebrates (Flint River Watershed Coalition, 2007). More than 300,000 people live in the watershed and use the Flint River as a drinking water source (Flint River Watershed Coalition, 2007). In addition, the Michigan Department of Natural Resources identified the Flint River watershed as a priority site for water quality and toxic waste load control (McIlroy et al., 1986).

2.2. Data collection

2.2.1. Physiographical data

The physiographical data including topography, landuse, and soil characteristics were mainly used as inputs to the hydrologic model. Topographic data was obtained from the USGS National Elevation Dataset (NED) at a 30 m spatial resolution (NED, 2015). Land cover data was obtained from the US Department of Agriculture (USDA)—National Agricultural Statistics Services (NASS). This dataset

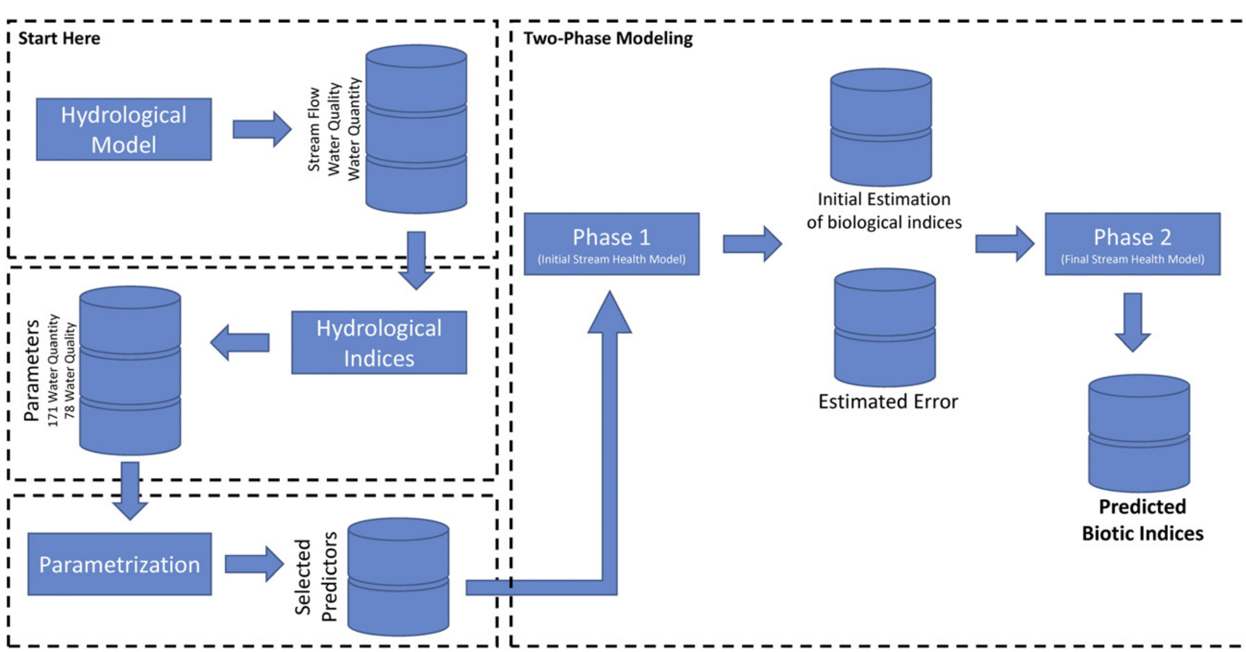

Fig. 1. Flow chart explaining the process how biotic indices are predicted using the proposed two-phase approach.