WATER RESOURCES RESEARCH, VOL. 49, 7747–7762, doi:10.1002/2012WR013422, 2013

Coupling self-organizing maps with a Naïve Bayesian classifier: Stream classification studies using multiple assessment data

Nikolaos Fytilis[1] and Donna M. Rizzo[1]

Received 17 December 2012; revised 27 August 2013; accepted 31 October 2013; published 26 November 2013.

[1] Organizing or clustering data into natural groups is one of the most fundamental aspects of understanding and mining information. The recent explosion in sensor networks and data storage associated with hydrological monitoring has created a huge potential for automating data analysis and classification of large, high-dimensional data sets. In this work, we develop a new classification tool that couples a Naïve Bayesian classifier with a neural network clustering algorithm (i.e., Kohonen Self-Organizing Map (SOM)). The combined Bayesian-SOM algorithm reduces classification error by leveraging the Bayesian's ability to accommodate parameter uncertainty with the SOM's ability to reduce high-dimensional data to lower dimensions. The resulting algorithm is data-driven, nonparametric and is as computationally efficient as a Naïve Bayesian classifier due to its parallel architecture. We apply, evaluate and test the Bayesian-SOM network using two real-world hydrological data sets. The first uses genetic data to classify the state of disease in native fish populations in the upper Madison River, MT, USA. The second uses stream geomorphic and water quality data measured at ~2500 Vermont stream reaches to predict habitat conditions. The new classification tool has substantial benefits over traditional classification methods due to its ability to dynamically update prior information, assess the uncertainty/confidence of the posterior probability values, and visualize both the input data and resulting probabilistic clusters onto two-dimensional maps to better assess nonlinear mappings between the two.

Citation: Fytilis, N., and D. M. Rizzo (2013), Coupling self-organizing maps with a Naïve Bayesian classifier: Stream classification studies using multiple assessment data, *Water Resour. Res.*, *49*, 7747–7762, doi:10.1002/2012WR013422.

1. Introduction

[2] Over the past couple of decades, there has been an exponential explosion in the development of real-time sensor networks and other means for collecting and storing data in most areas of modern science [*Szalay and Gray*, 2006]. But to date, the analysis tools needed to sufficiently mine these data have not kept pace [*Gil et al*., 2007]. Recent emphasis on interdisciplinary research adds to the challenge because the data networks found, for example, in astronomy [*Lang and Hogg*, 2011], protein folding [*Khatib et al*., 2011], the Earth system [*Bickel et al*., 2007], and the hydrological sciences [*Wagner et al*., 2009] demand expertise across multiple disciplines for interpretation. These data-intensive issues result in the need for advanced statistical and computational tools capable of analyzing the complex, multivariate associations, and uncertainty inherent in these large data networks [*Emmott and Rison*, 2005].

[3] Pattern recognition techniques, such as clustering and classification, are important components of intelligent data preprocessing, data mining, and decision making systems [*Schalkoff*, 1992]. These tools are of particular interest in hydrological river research [*Dollar et al*., 2007; *Helsel and Hirsch*, 1992; *Wright*, 2000], where a number of statistical classification methods, both parametric and nonparametric, have been used to classify river regimes [*Harris et al*., 2000], explore the influence of streamflow on biological communities [*Monk et al*., 2006], and optimize the selection of input data to improve ecohydrological classification [*Snelder et al*., 2005] at multiple scales, including catchment [*Pegg and Pierce*, 2002], regional [*Nathan and Mcmahon*, 1990], national [*Poff*, 1996], and continental [*Puckridge et al*., 1998]. In addition, hydrologists often gain insights from well-gauged regions and classify or extrapolate to sparsely gauged regions using some limited number of stream characteristics. For example, *Kondolf* [1995] used geomorphologic characteristics to classify stream channel stability; *Alberto et al*. [2001] used select water quality parameters to identify variation at multiple temporal and spatial scales; *Rabeni et al*. [2002] used benthic invertebrates for stream habitat health classification; and *Besaw et al*. [2010] used local climate data to predict flow in small ungauged streams.

[4] Multiple correlated and cross-correlated data, missing data, binary data (i.e., presence or absence), and most importantly, the uncertainty inherent in these data pose significant limitations to existing classification and clustering

[1]Department of Civil and Environmental Engineering, College of Engineering and Mathematical Sciences, University of Vermont, Burlington, Vermont, USA.

Corresponding author: N. Fytilis, Department of Civil and Environmental Engineering, College of Engineering and Mathematical Sciences, University of Vermont, 211 Votey Hall, Burlington, VT 05405, USA. (nfitilis@uvm.edu)

0043-1397/13/10.1002/2012WR013422

algorithms and demand the development of new or hybrid clustering techniques [*Jain*, 2010]. A recent NSF-sponsored workshop, *Opportunities and Challenges in Uncertainty Quantification for Complex Interacting Systems* [*Ghanem*, 2009] recognized the need for new computationally efficient tools capable of improving the quantification of uncertainty in inferred models [*Roache*, 1997], network structure and model parameters [*Katz et al*., 2002].

[5] Bayesian methodology provides a fundamental approach to the problem of pattern classification [*Duda and Hart*, 1973], and offers the ability to quantify and reduce any kind of uncertainty given enough relevant new information (i.e., prior data) [*Malakoff*, 1999]. Combined with Monte Carlo Markov chain methods, Bayesian approaches have become popular for processing data and knowledge [*Steinschneider et al*., 2012] because of the relative computational ease with which they handle complex data sets [*Han et al*., 2012]. Bayesian methods have proven useful in hydrologic applications for parameter estimation and assessing uncertainty [*Smith and Marshall*, 2008]; and have been incorporated into stochastic simulation models [*Balakrishnan et al*., 2003; *Leube et al*., 2012; *Williams and Maxwell*, 2011] and optimization techniques [*Mariethoz et al*., 2010; *Reed and Kollat*, 2012] to reduce prediction uncertainty. In this work, we develop a new classification tool that couples the concept of a Naïve Bayesian classifier with an artificial neural network often used for nonparametric clustering and classification.

[6] Briefly, artificial neural networks (ANNs) are nonparametric statistical tools that specialize in nonlinear mappings given large amounts of data [*Haykin*, 1999; *Mitra et al*., 2002]. They have gained popularity in applications that require mining large numbers of multiple data types with both continuous and categorical responses [*Zhang et al*., 1998]. Along with other nonparametric statistical techniques, they are more suited than physics-based models [*Govindaraju and Artific*, 2000a, 2000b] when the objectives are classification or system characterization rather than an understanding of the physical system [*Kokkonen and Jakeman*, 2001]. Recently, ANNs have been shown to be more successful in many hydrology-related applications [*Maier and Dandy*, 2000; *Solomatine and Ostfeld*, 2008] than their traditional (parametric) statistical counterparts such as discriminant analysis [*Yoon et al*., 1993], regression techniques [*Paruelo and Tomasel*, 1997], principal component analysis [*Kramer*, 1991], and Bayesian analysis [*Cheng and Titterington*, 1994; *Richard and Lippmann*, 1991; *Wan*, 1990].

[7] Bayesian analysis has been incorporated into ANN algorithms used in hydrology for the purpose of improving the training procedure and overcoming the computational limitations associated with optimizing the ANN hidden weights [*Kingston et al*., 2005b]. Despite difficulties in coding these advanced Bayesian approaches into existing ANN learning algorithms (see *Titterington* [2004] for details), these predictive models provide a better means for computing uncertainty and model validation than the best deterministic ANN models while maintaining the high computational performance associated with traditional ANNs [*Kingston et al*., 2005a, 2008; *Zhang and Zhao*, 2012].

[8] In this work, we developed and applied a new framework that couples a Simple Bayesian analysis with a clustering ANN to advance the efficiency and statistical optimality of both techniques. Specifically, we use a Naïve Bayesian classifier in combination with an unsupervised ANN (the Kohonen Self-Organizing Map, SOM, Kohonen [1982, 1990]) to leverage prior information (or evidence) embedded in multiple data for the purpose of improving classification, while minimizing within class variance. To show proof-of-concept, we applied, evaluated, and tested the Bayesian-SOM network using two real-world data sets. The first uses genetic data and expert-assessed morphological data to predict the relative abundance of worm taxa related to the state of Whirling Disease in native fish populations in the upper Madison River, MT, USA. Specifically, we spatially estimate the relative abundance of stream sediment-dwelling worms. These worms are the definitive host of the parasite that causes Whirling Disease in fish that ingest these worms. The second application uses stream geomorphic and water quality data measured in ~2500 Vermont stream reaches (comprising 1371 stream miles) to assess habitat conditions. We compared the new classification tool with traditional classification techniques, a Simple Bayesian analysis, a traditional Naïve Bayesian classifier and Gaussian mixture models.

2. Methods

[9] To circumvent some of the classification challenges associated with large amounts of multiple data mentioned above while incorporating uncertainty and minimizing the classification error, we designed a new classification tool. This tool couples the concept of a Naïve Bayesian classifier with 1-D and 2-D Kohonen Self-Organizing Maps (SOMs). The Bayesian-SOM network described in this paper was coded in Matlab R2012a (MathWorks Inc.). This section briefly describes the Naïve Bayesian classifier, our choice of the SOM clustering algorithm, and a description of the coupled classification framework.

2.1. Naïve Bayesian Classifier

[10] A Bayesian classifier is a statistical classifier that leverages conditional probabilities using the degree of belief of an event (A) before and after accounting for new evidence (B). According to Bayes' theorem, the posterior probability, $p(\mathrm{A}|\mathrm{B})$, or the updated degree of belief in A having observed B, can be expressed as:

$$p(\mathrm{A}|\mathrm{B}) = \frac{p(\mathrm{A}) * p(\mathrm{B}|\mathrm{A})}{p(\mathrm{B})} \quad . \tag{1}$$

$p(\mathrm{A})$ is the prior probability of the event A. The prior distribution, supplied by the user, describes what is known about A before the collection of new observations. These prior data might be collected using different techniques, models, personal experience, or alternatively, might be an uninformative prior, indicating that we have only vague information about the variable of interest. The likelihood function, $p(\mathrm{B}|\mathrm{A})$, is the conditional probability of the observed data, B, given the initial belief in event A; and $p(\mathrm{B})$ is the probability of the observed data that serves as a normalization factor [*Good*, 1965].

[11] The Naïve Bayesian classifier assumes the observation of any feature, F_z, (i.e., particular parameter/data type)

on a given class, C_z, is independent of the presence of other features [*Lewis*, 1998]. This class conditional independence assumption greatly simplifies the computation. While this assumption is not met for many real-world applications, it is possible to preprocess the input variables to meet this class independence requirement. In this work, we use a principal component analysis. Despite its simplicity, the Naïve Bayesian classifier has been found to perform surprisingly well [*Friedman et al*., 1997]. To the best of our knowledge, the Naïve Bayesian classifier has not been used for stream classification purposes, despite its popularity in biological [*Wang et al*., 2007] and text classification related studies [*Androutsopoulos et al*., 2000]. To implement the classifier, one needs to (1) compute the individual conditional probabilities, $p(F_z|C_z)$, for each feature, z, (2) multiply each with the prior probability for each class, $p(C_z)$, and (3) classify the new information (features presented in a vector form) to the class with the highest product as:

$$p(C_z|F_1,\ldots,F_z)=p(C_z)\prod_{z=1}^{Z}p(F_z|C_z), \tag{2}$$

which is based on the maximum a posteriori decision rule. In this work, all the parameter values for the chosen model distribution (i.e., class priors and feature likelihood probability distributions) are computed using the Maximum-Likelihood Estimation (MLE) method utilizing the expectation-maximization technique because it provides a systematic method for simple statistical models and large data sets. The method has one disadvantage: the conditional probabilities may equal zero when a feature is not present in one of the classes, which cancels the effect of all other features during multiplication. To circumvent this problem, we use a Laplacian correction term [*Provost and Domingos*, 2003], which adds one (called a pseudocount) to all feature counts to guarantee that none of the conditional probabilities ($p(F_z|C_z)$) are set exactly to zero.

2.2. Kohonen Self-Organizing Map

[12] The Kohonen Self-Organizing Map (SOM) is a clustering algorithm developed in the 1980s by Teuvo Kohonen [*Kohonen*, 1982, 1990]. It is an unsupervised ANN that autonomously analyzes properties inherent in the input data. In general, unsupervised algorithms are used to extract relationships from data when the response variable or output classifications are unknown. This nonparametric clustering method uses what has become known as competitive learning to self-organize the input data into a topological map of the resulting clusters. The algorithm is data driven, and therefore does not require the development of site-specific, process-based models, or sets of if-then-else rules associated with expert systems to cluster the data. SOMs may be used to convert nonlinear, high-dimensional data to some user-defined lower dimension [*Fritzke*, 1994]. They are also capable of clustering large amounts of discrete and continuous data types, while relaxing many assumptions (e.g., normally distributed data) required by traditional statistical techniques [*Kohonen*, 1990; *Kohonen et al*., 1996].

[13] The SOM has been used in many data exploration studies [*Kaski*, 1997], water resources applications [*Kalteh et al*., 2008] and ecological studies to classify macroinvertebrate communities [*Chon et al*., 1996], cluster, and forecast flood events [*Moradkhani et al*., 2004], assess water quality [*Aguilera et al*., 2001], and classify streams [*Besaw et al*., 2009]. It outperforms many traditional clustering methods (e.g., hierarchical and K-means) on data with high dispersion, outliers and irrelevant variables [*Mangiameli et al*., 1996]. Figure 1 shows a schematic of the 1-D and 2-D SOM architectures used in tandem in this work.

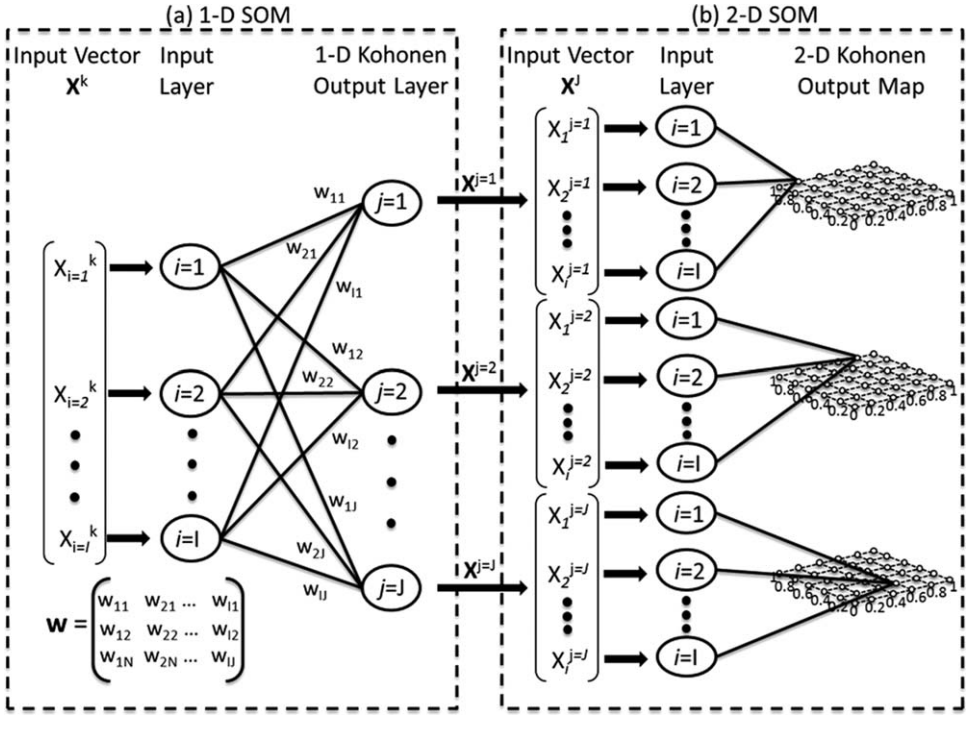

Figure 1. Schematic showing the (a) 1-D and (b) 2-D tandem SOM architecture.

2.2.1. Data Preprocessing

[14] Input data are presented to the SOM network as a matrix, **X**, which has rows corresponding to each sample, k, and columns, i, that correspond to the individual data types. All input data are normalized to values between 0 and 1 to avoid bias. The row vectors, $\mathbf{X}^k$ (called training vectors), associated with sample, k, were normalized independently across individual data types, i, such that:

[15] For i = 1:I (all data types)

$$\mathbf{X^k} = \frac{(x_i)^k\text{-min }(x_i)}{\max(x_i)\text{-min}(x_i)} \text{ for } k = 1, 2, \ldots, K. \tag{3}$$

[16] End for loop

[17] After normalization, each training vector, $\mathbf{X}^k$, is introduced to the 1-D or 2-D SOM input layers (Figure 1); each vector component corresponds to an input layer node. The SOM weights, w_{ij}, link the measured input nodes to the 1-D or 2-D Kohonen output maps and are initially set to random values between 0 and 1.

2.2.2. Self-Organization Phase

[18] During self-organization, each input vector passes to the Kohonen output map, where a distance metric is computed as:

$$z_j = ||\mathbf{X}^k - \mathbf{w}_j|| \quad \text{for } j = 1, 2, \ldots J. \tag{4}$$

[19] Here, $\mathbf{w}_j$ represents the weight vectors (rows of the Kohonen weight matrix, **w**) associated with each of the j Kohonen output nodes. In this work, we use the Euclidean distance as a measure of similarity between the input vector and each of the j Kohonen weight vectors. This distance metric, z_j, is passed through an activation function such that the Kohonen node with the minimum distance is declared the winner (or best matching unit). The "winning" Kohonen node, j, and all nodes in some neighborhood, N_c, of the winning node, are iteratively adjusted to be more similar to the input vector that caused it to become activated. The weights are updated as follows:

$$\begin{aligned} &\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t) * \left(\mathbf{w}_j(t) - \mathbf{X}^k\right) \text{ for } j \in N_c(t) \\ &\mathbf{w}_j(t+1) = \mathbf{w}_j(t) \text{ for } j \ni N_c(t), \end{aligned} \tag{5}$$

where α (a learning coefficient) decreases linearly during training from 0.9 to 0 and controls the magnitude of adjustment at each iteration t. An iteration is defined as a single pass through all of the input vectors. In this work, we use a total of 1000 iterations, and introduce the input vectors in random order to prevent the network from learning any one sequence of patterns. Initially, the neighborhood size, N_c, includes half of the nodes in the Kohonen output map and is decreased linearly to one over the total number of iterations. The reduction of N_c and α during this self-organization phase ensures that a global data structure is established in the early phases of self-organization and more local refinement is established in the latter stages.

2.2.3. Data Visualization Using the Unified-distance Matrix

[20] After self-organization is complete, the weights are fixed and the network clustering may be visualized in one- or two-dimensions using a postprocessing procedure known as the unified-distance matrix (U-matrix) formulated by *Ultsch and Siemon* [1989]; however, other techniques have been developed [*Kraaijveld et al*., 1995; *Manukyan et al*., 2012]. The U-matrix is computed by taking the average Euclidean distance between each of the fixed (i.e., self-organized) weight vectors and, in this work, the eight nearest neighbors. We plot the averaged Euclidean distances in gray scale on the 2-D SOM output maps to better visualize clustering. Any labeling (i.e., classification identifications, if known) of the data may also be superimposed onto the self-organized winning nodes. One additional benefit of the SOM is that individual vector components (i.e., original input data used for clustering) may be plotted on similar 2-D maps. The latter are referred to as "component planes" [*Kohonen*, 2001] and are useful in exploring the relationship of the input data parameters with resulting self-organized clusters.

2.3. Bayesian-SOM Classification Framework

[21] Our Bayesian-SOM classification framework (Figure 2) is executed in two phases: a self-organization phase where the likelihood data are self-organized and clustered using the tandem 1-D and 2-D SOMs and an operational phase where the fixed weights from the self-organization phase are used for classification/prediction. In this work, the 1-D output map has a toroidal structure; the 2-D map does not. The prior distribution parameters are provided by the user, as they would be using a Naïve Bayesian classifier. Whereas, the likelihood pdf parameters are generated during the self-organization phase, using the prior information to replace the parameter estimation method of a traditional Naïve Bayesian classifier. The 2-D grid sizes were selected using Kohonen's general rule of thumb (i.e., the optimal number of nodes is five times the square root of the number of observation data), but required some trial and error for both case studies. However, in our applications the clustering is not very sensitive to the grid size because the number observation data clustered to each parallel 2-D map varies and the 2-D SOMs are used simply to construct the likelihood pdf parameters.

2.3.1. Phase 1: Self-Organization (Training)

[22] The self-organization (training) phase enables estimation of the likelihood distribution parameters. It requires the available data be subdivided into two sets (*prior* and *likelihood*). The prior data for each of our known output classes forms the basis of the knowledge we wish to update and are not processed using a SOM. The distribution parameters for each class are simply estimated using the MLE method and all observation data in each class.

[23] The second subset of data is used to generate the parameters for the likelihood function distributions using the one-dimensional and two-dimensional SOMs. These data self-organize (i.e., cluster) onto the 1-D Kohonen output nodes (prespecified by the user to be at least equal to or greater than the number of expected classes). Specifically, the user creates an initial hypothesis as to the number of clusters needed. After self-organization, the weights are saved for use in the operational phase.

[24] The 2-D Kohonen maps, operating in parallel, follow the same iterative training procedure as the 1-D SOM maps. We again use the MLE method to estimate the

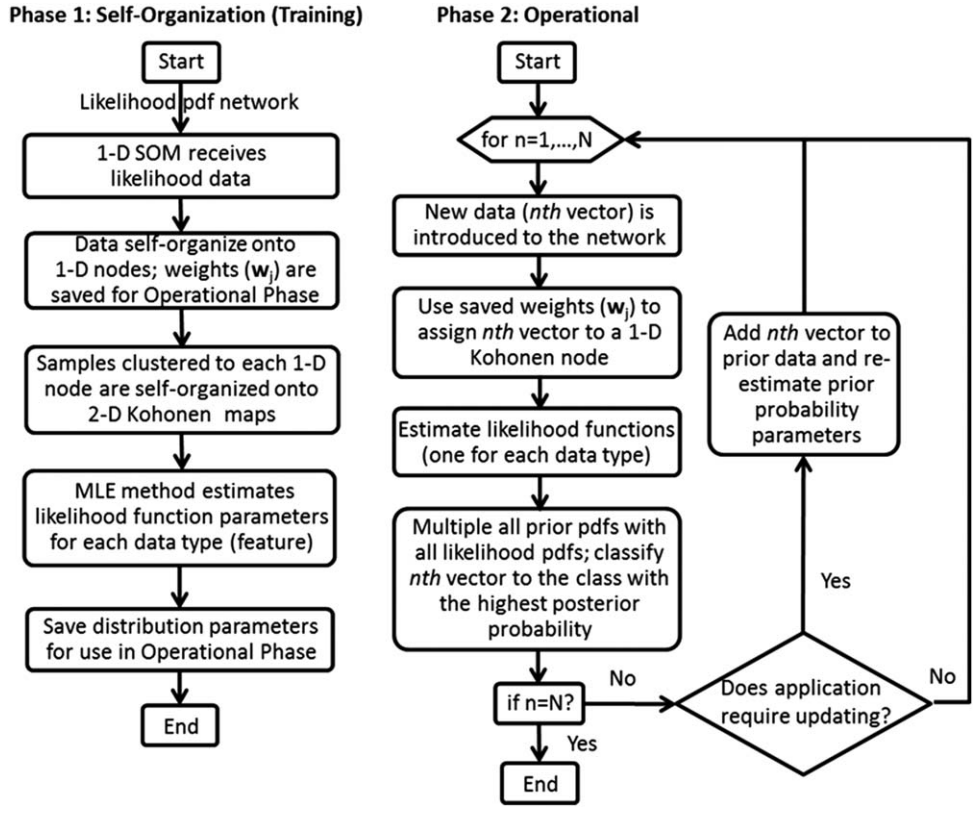

Figure 2. Classification framework flowchart: Training data are subdivided into two groups. The prior data are used to generate Maximum-Likelihood Estimates (MLE) of the class prior distribution parameters. Phase 1 outlines the training or self-organization phase and phase 2 uses the likelihood data to generate the MLE of the feature likelihood function distributions and then uses the fixed weights from the self-organization phase to classify new data. The user may opt to update the prior probability distributions dynamically.

distribution parameters for the likelihood pdf networks. Mapping the high-dimensional data clustered to each 1-D Kohonen node onto the 2-D Kohonen maps serves a dual purpose. First, it helps identify similarity (or subtle differences) when constructing the likelihood functions associated with each cluster. Second, these 2-D projections are useful for performing feature extraction (i.e., identifying the individual vector components (i.e., input data types) important for sample separation). The latter is accomplished by projecting the individual component planes (see section 2.2.3) onto the same 2-D map as the clustered data. We provide examples of these in sections 3 and 4.

2.3.2. Phase 2: Operational

[25] During the operational phase, *prediction* data (new measurement data the network has never seen before) are introduced to the network input layer. The saved weights (from each of the 1-D likelihood function SOMs (Figure 2, phase 1) are used to assign these new prediction vectors to one of the 1-D Kohonen output nodes (clusters). To accomplish this, the likelihood functions (the self-organized distribution parameters and each new prediction vector) are multiplied with each of the associated prior pdfs obtained from the self-organization phase to estimate the posterior pdfs. The new prediction vector is then assigned to the class with the highest posterior probability (maximum value). It is important to note that there are two ways of updating the prior probability density functions. One is to use the pdfs obtained during the self-organization phase; the other is to update the pdfs after each new prediction vector is presented to the network. The latter enables a dynamic analysis of the measurement data and is an important feature of the operational phase because it may be used when the initial prior information is noninformative or based on erroneous information. However, the former is arguably better in cases where the prior is based on more accurate (known) data. Finally, the operational phase (Figure 2, step 2) is computationally fast because it uses the fixed weights from the likelihood pdf network and the simple maximum *a posteriori* rule to determine the most probable class.

2.4. Tools to Evaluate the Bayesian-SOM Network Performance

[26] We compare the results of the Bayesian-SOM to traditional discriminant analysis, k-nearest neighbor using a Euclidean distance metric, a Simple Bayesian analysis and a traditional Naïve Bayesian classifier. We also used Gaussian mixture models as a surrogate for the 2-D SOM to estimate the likelihood distribution parameters of the resulting probabilistic models. Gaussian mixture models may use an unsupervised learning procedure similar to that of the SOM and provide a more fair comparison with the Bayesian-

SOM network because we use the same MLE technique using the expectation-maximization technique. The Gaussian mixture model is known to perform better when clusters overlap [*McLachlan and Peel*, 2000]. The discriminant, k-nearest and the Gaussian mixture models were performed using JMP PRO (version 9, SAS Institute Inc.). The Simple Bayesian analysis and Naïve Bayesian classifier were coded in Matlab R2012a (MathWorks Inc.).

3. Results

3.1. Case Study 1: Estimating Aquatic Worm Community Composition in the Madison River, MT

[27] Oligochaeta worms, specifically from the *Tubificidae* family are ingested by salmonid fish and are the intermediate host for a parasite, *Myxobolus cerebralis*, that causes Whirling Disease [*Wolf et al*., 1986]. Whirling Disease has caused a 90% reduction of rainbow trout populations in the area of Madison River, MT, USA [*Vincent*, 1996] and is responsible for the loss of millions of dollars in revenue from recreational fishing in the Intermountain West. One of the challenges in reducing the disease involves the spatial characterization and accurate identification of the oligochaete (worm) community [*Kerans and Zale*, 2002]. Three of the oligochaete taxa of particular interest vary widely in parasite susceptibility and constitute >90% of the oligochaete community. These taxa are *Tubifex tubifex, Rhyacodrilus spp*., and *Ilyodrilus templetoni*, hereafter referred to as Tt, Rhy, and Ily, respectively. Unfortunately, taxonomic identification of oligochaetes by experts depends largely on the morphological characteristics of sexually mature worms; and this life stage constitutes a relatively short period (∼2 weeks) of their life cycle [*Brinkhurst*, 1986]. To more easily distinguish between tubificid taxa of interest, we developed a multiplex probe-based qualitative real-time polymerase chain reaction assay (hereafter referred to as the PCR probe assay) that distinguishes the definitive host Tt from the nonhost Rhy. The latter approach is less labor-intensive and less costly than expert-based morphological identifications or DNA sequencing of worms. Details of the PCR probe assays may be found in *Fytilis et al*. [2013].

3.1.1. Study Area

[28] Our study area is a 39 km section of the river located between Quake Lake and Ennis Lake in Madison County, MT, USA. Six sites (Figure 3), previously showing variation in fish disease risk and tubificid composition [*Krueger et al*., 2006], were sampled until 250 worms had been collected. After field collection, we observed that Ily

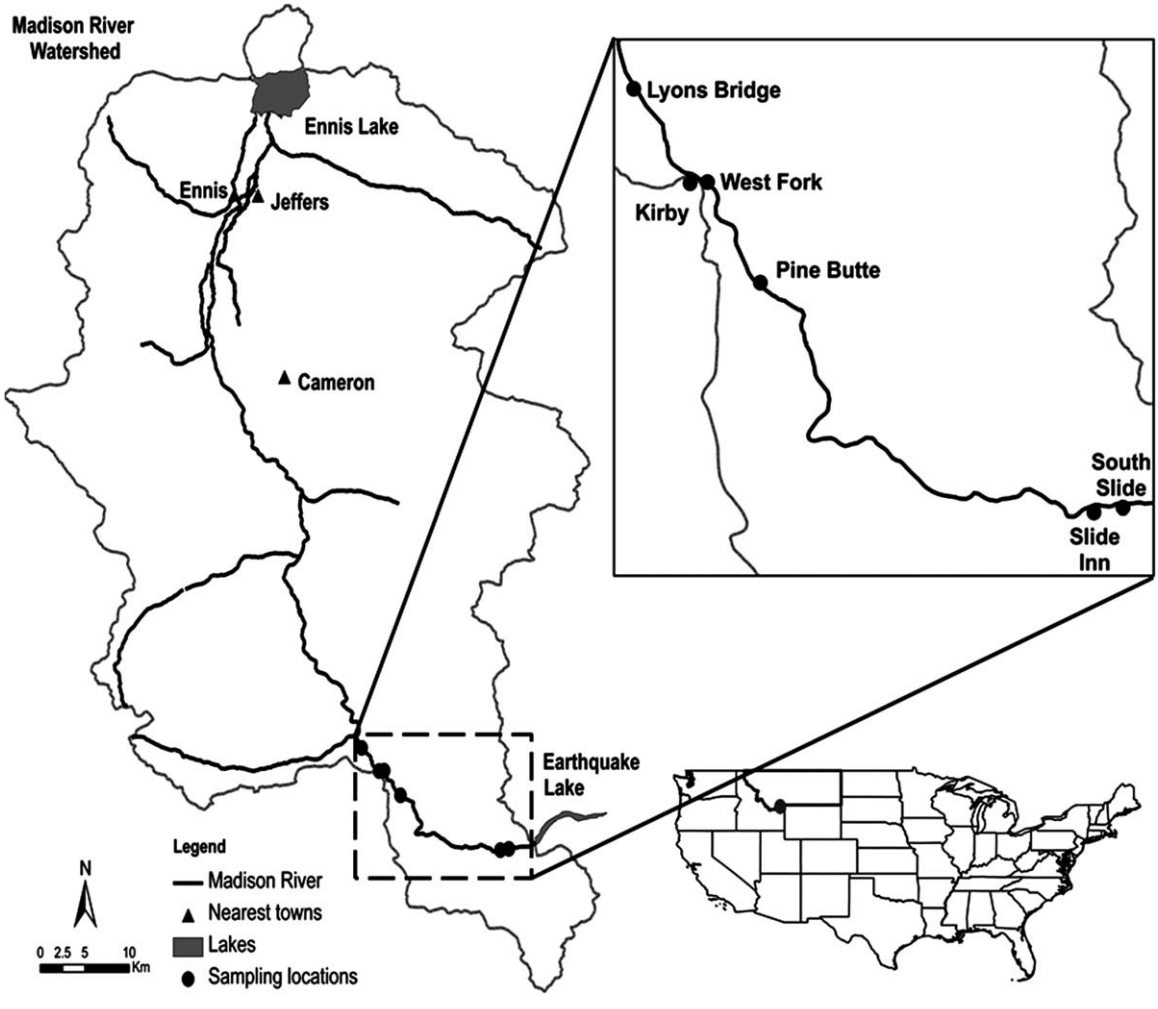

Figure 3. Madison River watershed showing our study reach and the six study site locations.

represented <2% of the total worm community composition and therefore did not include it in the PCR probe assay design. Eighty-eight worms (i.e., representative samples of the two taxa of interest, Tt and Rhy) from each site were selected randomly for the PCR probe assay. The data for Case Study 1 are comprised of the fluorescent signals from each of the two PCR probes (hereafter referred to as probe data) designed to detect Tt and Rhy taxa. Each probe value varies from 0 to 60 with units represented as derivatives of fluorescence with respect to temperature ($-dF/dT$). Since the units are the same for both probes, the minimum and maximum values used for normalization (equation (3)) were calculated over both sets of input data rather than each independently.

3.1.2. Overall Goals and Hypothesis (Case Study 1)

[29] Since it is not possible to identify immature worms morphologically, our goal is to accurately predict the relative abundance of the two worm taxa at individual stream sites using probe data from immature worms as our likelihood data and probe data from the less abundant sexually mature worms as our prior. All sexually mature worms were morphologically identified (labeled Rhy or Tt) by experts; and all morphological identifications were verified by DNA sequencing.

[30] Our hypothesis is: Can probe data (from immature worms), in combination with the prior morphological identification (Tt or Rhy) of sexually mature worms, improve site-specific relative abundance estimates of these two taxa?

[31] This hypothesis is predicated on the observation that the sexually mature worms, which only exist in this life stage for a limited time, comprise a relatively small percentage of the sampled worm community composition.

3.1.3. Bayesian-SOM Network Architecture

[32] The likelihood function distribution parameters are estimated using the 1-D SOM architecture with two input nodes (one corresponding to each of the normalized fluorescent signals obtained from the species-specific PCR probes) and three output nodes. The selection of three output nodes is based on the desire to cluster all data into one of three worm taxa categories (i.e., Tt, Rhy, and unknown). We assumed normal Gaussian distributions to represent each fluorescent signal. The classified data are projected onto a 2-D SOM to further identify the spread of each fluorescent signal. The 2-D maps for the likelihood pdf networks are 10×10 grids (Figure 4a). The grid size required some trial and error but is not sensitive (e.g., similar clustering occurs with only four nodes) or computational expensive (e.g., a 10-fold increase in map size only increases the run times from 11.9 to 15.8 s on a HP Pavilion Elite h8–1280t with Intel Core i7–3820 3.6 GHz and 16 GB RAM). The averaged Euclidean distances are plotted in gray scale to visualize clustering. One such U-matrix associated with Case Study 1 is provided in Figure 4b. In our example, labels T and R of Figure 4c are associated with our known prior information. The nodes that self-organize (cluster) worm samples of interest (i.e., Tt and Rhy) are expected to have a smaller spread than the samples that classify to the output node with ambiguous fluorescent signals (i.e., unknown). The likelihood functions represent confidence (or uncertainty) that the observed data are compatible with our initial hypothesis of the existence of two taxa.

[33] During the operational phase, we present new prediction data (probe data from site-specific worms of unknown taxa) to the network and multiply the self-organized likelihood functions with each of the associated prior pdfs to estimate the site-specific posterior probability of each new prediction vector being classified as Rhy or Tt. Once a new vector has been assigned to a class, we dynamically update the "winning" prior probability to make an "improved" estimate for the next unknown sample point.

3.1.4. Bayesian-SOM Network Performance

[34] We apply and compare the Bayesian-SOM network performance and computational efficiency against a Simple Bayesian and a traditional Naïve Bayesian classifier. Runtimes for the Bayesian-SOM and the Naïve Bayesian classifier were 12.59 and 11.84 s, respectively on a HP Pavilion Elite h8-1280t with Intel Core i7-3820 3.6 GHz and 16 GB RAM. Table 1 compares estimates of relative abundance for the two taxa at each of our six sites using each of the three methods. The column labeled "Ground Truth" represents our best estimate of relative abundance for the two taxa because it is based on DNA sequencing of all

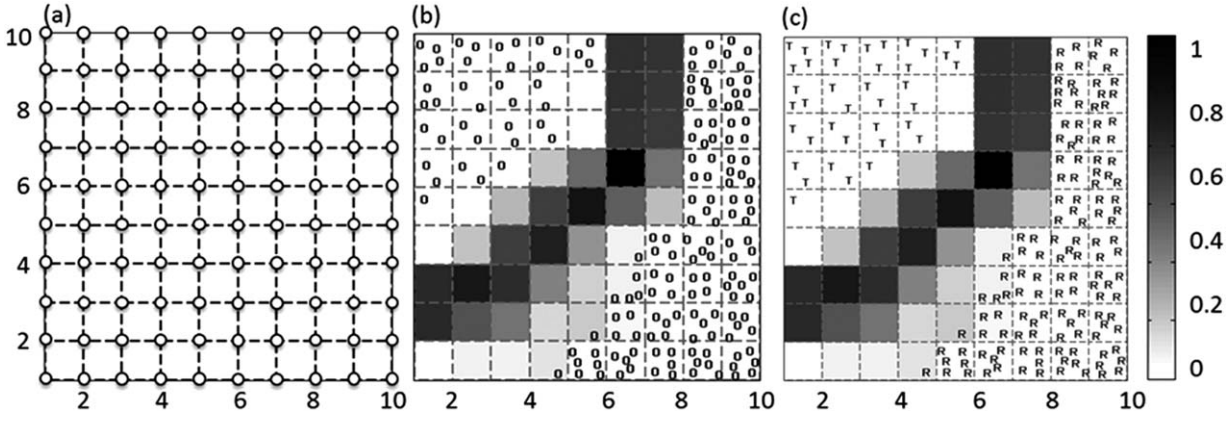

Figure 4. An example of the (a) 10×10, 2-D output map used in Case Study 1, (b) resulting (gray scale) U-matrix showing the separation distance between our clustered data, and (c) known labels superimposed on the associated self-organized samples.

Table 1. Relative Abundance Estimates for Each Worm Taxa Using the Bayesian-SOM Network (Column 6) Compared with the Morphological Taxonomic Identification of Sexually Mature Worms (Column 2), PCR-Probe Assay Results of Immature Worms (Column 3), Simple Bayesian Approach (Column 4), Naïve Bayesian Analysis (Column 5), and the Inferred "Ground Truth" Estimates (Column 7) at Each of Our Six Sites

Taxa	(*n*) Morph. ID (Mature Worms)		(*n*) PCR Assay (Immature Worms)		(*n*) Simple Bayesian		Naïve Bayesian	Bayesian-SOM Network	Ground Truth	Sites
Rhy	(77)	0.630	(90)	0.311	(167)	0.458	0.488	0.415	0.398	1. South Slide
Tt		0.370		0.689		0.542	0.512	0.585	0.602	
Rhy	(32)	1	(31)	1	(63)	1	0.984	0.936	1	2. Slide Inn
Tt		0		0		0	0.016	0.064	0	
Rhy	(33)	0.545	(135)	0.081	(168)	0.096	0.238	0.163	0.161	3. Pine Butte
Tt		0.455		0.919		0.904	0.762	0.837	0.839	
Rhy	(57)	0	(112)	0	(169)	0	0.006	0.027	0	4. West Fork
Tt		1		1		1	0.994	0.973	1	
Rhy	(83)	0.847	(78)	0.744	(161)	0.943	0.758	0.776	0.777	5. Kirby
Tt		0.153		0.256		0.057	0.242	0.224	0.223	
Rhy	(25)	0.059	(108)	0.028	(133)	0.003	0.023	0.029	0.039	6. Lyons Bridge
Tt		0.941		0.972		0.997	0.977	0.971	0.963	

available samples (mature and immature worms). The taxa relative abundances determined using expert morphological identification of sexually mature worms and the PCR probe-based estimates of the immature worms are presented in columns two and three, respectively. The relative abundances provided by the Simple Bayesian analysis (Table 1, column 4) were generated using a beta distribution as a prior distribution and sample size of 1000. The prior data comprised the morphological identifications of mature worms (column 2). The likelihood function distribution parameters were estimated using the PCR-probe data from immature worms (column 3) assuming a binomial distribution. The traditional Naïve Bayesian classifier estimates (column 5) use the same prior and likelihood information as in the Simple Bayesian method and the Bayesian-SOM network. The Bayesian-SOM estimates (Table 1, column 6) show sites 2 and 5 as having high Rhy relative abundance compared to Tt. At site 1, Rhy and Tt are more similar; and the remaining sites have higher relative abundance of Tt. In addition to generating estimates of relative abundance, we are able to produce high-density regions (HDRs). Figure 5 shows an example of the 90% posterior probability intervals generated for the relative abundance of Tt at site 1 and represents the probability of observing Tt at site 1 knowing that we collected 1000 samples. The dashed black line represents our best estimate of the relative abundance of Tt at that site using our Bayesian-SOM network and the solid line identifies the actual relative abundance of Tt for this site based on DNA sequencing (Ground Truth measurements).

3.2. Case Study 2: Estimating Stream-Reach Habitat Health Using Geomorphic and Water Quality Assessment Data in Vermont Streams

[35] Various divisions within the Vermont Agency of Natural Resources (VTANR) (e.g., Department of Environmental Conservation, River Management Program, Department of Fish and Wildlife and Vermont Geological Survey) launched a three-phase system to perform stream geomorphic and habitat assessments. The stream-reach Rapid Geomorphic Assessment (RGA) and the legacy Reach Habitat Assessment (RHA) field data used in the case study were collected at ~2500 stream segments or 1371 (6% of the total) stream miles over the period from 2003 to 2008. The RGA quantifies the degree to which the stream reach has departed from the reference or dynamic equilibrium condition via the assessment of four fluvial processes: degradation, aggradation, widening, and planform changes using various reach-scale metrics (e.g., width/depth ratio, sinuosity, bed substrate composition to name a few). These metrics are combined with expert opinion to assess an integer score between 0 (poor) and 20 (excellent) for each of the four processes. The four resulting scores are then summed to produce an overall reach-scale geomorphic score between 0 and 80 (higher scores indicate the stream reach is more likely to be in dynamic equilibrium).

[36] The reach habitat assessment (RHA) uses a combination of 10 parameters representative of the channel bed, bank and riparian vegetation to assess key processes

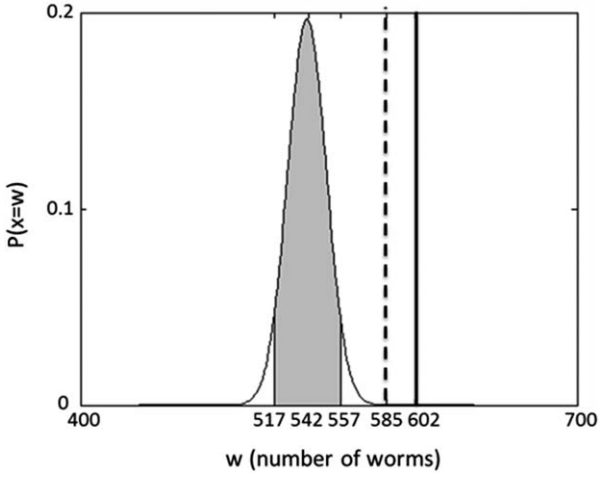

Figure 5. High-density region (90%) showing the site-specific credible interval where estimates of Tt relative abundance should lie knowing that 1000 samples were collected at site 1. The dashed line represents the relative abundance of Tt estimated using the Bayesian-SOM network; the solid black line identifies the relative abundance of Tt actually observed at site 1.

(physical, chemical, and biological) associated with the reach-scale aquatic habitat health. Each parameter has a score between 0 (poor) and 20 (excellent) that results, when summed, in a total RHA score no greater than 200. The final RGA and RHA stream-reach condition scores are further assigned to one of four categories—poor, fair, good, or excellent. The intersection of stream reaches with both types of RGA and RHA assessment data results in $n = 1363$ stream reaches (Figure 6). For details on RGA and RHA data as of 2008 see http://www.vtwaterquality.org/rivers/htm/rv_geoassesspro.htm.

[37] In 1990, the Vermont Department of Environmental Conservation established a long-term water quality and biological monitoring project that monitors ~70 water quality parameters, including 30 metals (both dissolved and total concentrations), chlorophyll a, conductivity, dissolved oxygen, *E. coli*, alkalinity, total hardness, total suspended solids, turbidity, visual total color, and pH. Water quality measurements were collected at anywhere from one to seven locations within individual stream reaches and therefore averaged over the reach scale. Temporally, water quality data were collected multiple times within a given season in a single year; thus, these data were seasonally averaged over each of the 5 years. The intersection of the stream reaches with RGA, RHA, and water quality data further reduced our data set from $n = 1363$ to $n = 235$ (Figure 7).

[38] In this work, we chose nine (i.e., *E. coli*, conductivity, total suspended solids, total phosphorus, total nitrogen, turbidity, dissolved chromium, dissolved iron, and dissolved arsenic) of the original 70 parameters for proof-of-concept. Selection was based on information provided in the 2011 Water Quality Integrated Assessment Report (http://www.anr.state.vt.us/dec/waterq/mapp/htm/mp_monitoring.htm), the availability of water quality data (i.e., we selected parameters that were monitored most consistently at the $n = 235$ reaches over the 5 year study period) using a buffer analysis in ArcGIS 10.0 (Esri Inc., Redlands, USA), and the amount of variance explained using a principal component analysis. The reduced set of nine water quality parameters provides a good representation of the three major causes of impairment (i.e., nutrients, metals, turbidity).

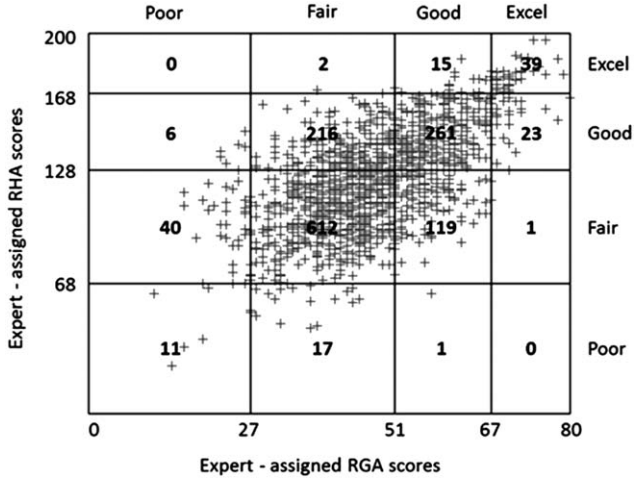

Figure 6. Scatter plot showing the number of Vermont stream reaches with both RGA and RHA data ($n = 1363$) over the period from 2003 to 2008. The vertical and horizontal lines mark divisions between VTANR assigned stream condition categories.

3.2.1. Overall Goal (Case Study 2)

[39] Here, we estimate the habitat health (RHA) at the stream-reach scale while constructing our likelihood pdf using both the RGA and the VT-DEC water quality data. This is desirable for environmental managers because these data are often less labor-intensive, more quantitative and easily obtained over a broader spatial coverage than biological integrity metrics that depend on a variety of aquatic species at specific life stages. Table 2 (matrix a) shows the raw expert-assigned assessment data (classified RGA against the RHA scores). A preliminary discriminant analysis (Table 2, matrix b), for all $n = 1363$ stream reaches accurately classifies 69.97% of the stream reaches using the four expert-assigned geomorphic RGA scores (i.e., degradation, aggradation, widening, and planform) as covariates. In an ideal world, RGA would align well with RHA and data might plot more along the matrix diagonal. In reality, this is not the case. This led us to hypothesize that using additional relevant information (i.e., water quality) might improve the overall stream-reach classification.

[40] Classification on the reduced data (i.e., $n = 235$ locations) with RGA, RHA, and water quality data was repeated using discriminant analysis and the same four RGA covariates (Table 3, matrix b). Interestingly, our ability to correctly classify stream habitat increases (76.17% from 69.97%) using this smaller $n = 235$ data set. The k-nearest neighbor classification analysis (Table 3, matrix c) produces the same classification error as the discriminant analysis.

3.2.2. Bayesian-SOM Network Architecture

[41] The input and output layers of the 1-D likelihood pdf network have four nodes each (Figure 8) corresponding to the four fluvial RGA processes and the expert-assigned RGA classes, respectively; whereas, the sequential 10×10 2-D SOM has nine input nodes, each corresponding to one of the nine VT-DEC water quality parameters. The prior information, $p(\mathrm{RHA})$, is generated using the original $n = 1363$ expert-assigned RHA data and describes the probability of stream-reach habitat health being classified as poor, fair, good, or excellent.

[42] The likelihood distribution parameters, on the other hand, were estimated using two probabilistic representations of the data. In the first approach, the likelihood pdf parameters, $p(\mathrm{WQ}|\mathrm{RHA})$, are constructed assuming a normal Gaussian distribution and the nine water quality parameters measured at 50% of the reduced ($n = 235$) stream reaches. The remaining reach data are held back for testing and validation. The posterior probability is calculated as

$$p(\mathrm{RHA}|\mathrm{WQ}) = p(\mathrm{RHA}) * p(\mathrm{WQ}|\mathrm{RHA}), \quad (6)$$

and represents the probability of stream-reach habitat being classified as poor, fair, good, or excellent given the nine select water quality measurements.

[43] In the second approach, the Gaussian distribution parameters for each likelihood function are estimated using

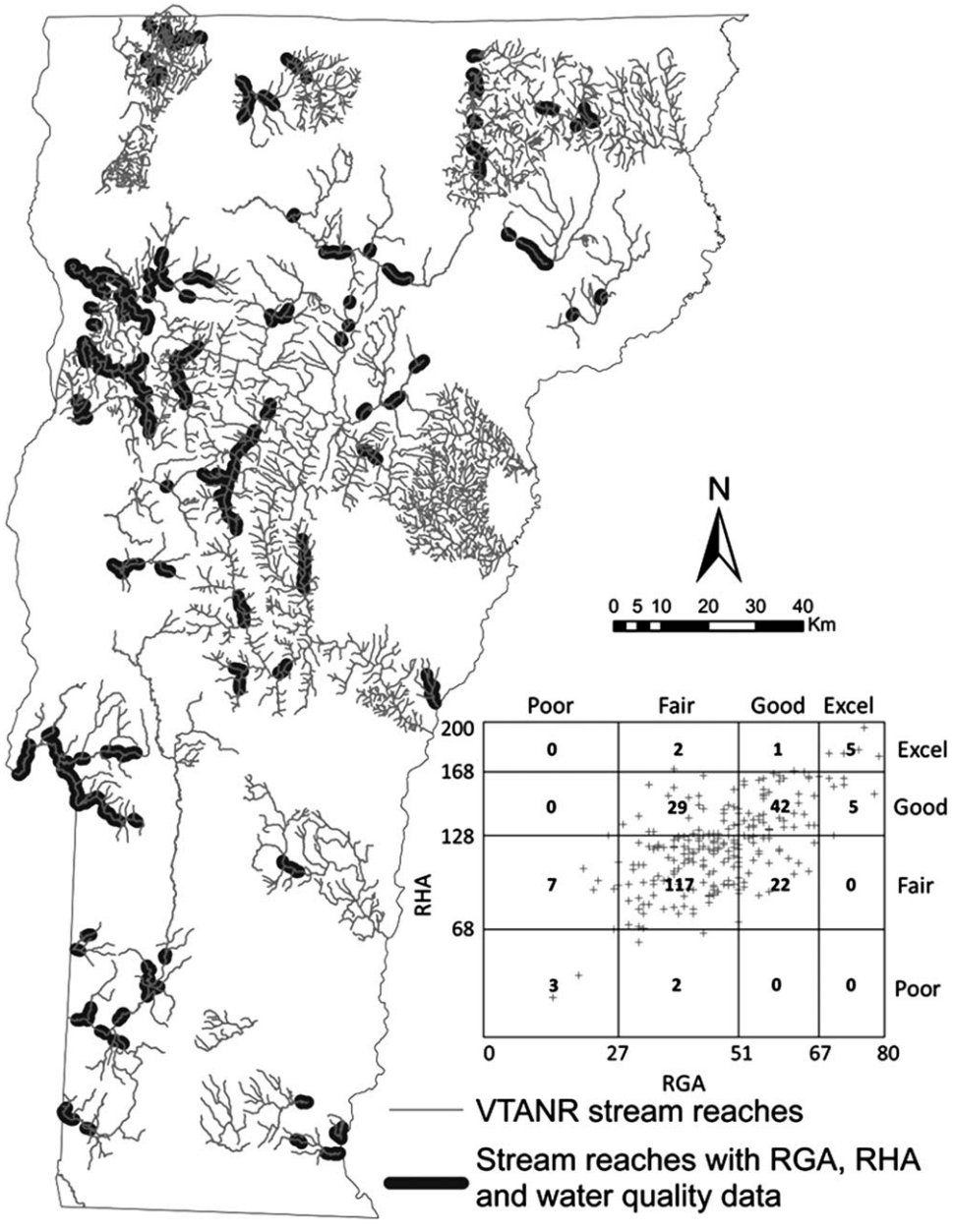

Figure 7. Vermont map showing the reduced ($n = 235$) number of stream reaches having RGA, RHA, and water quality data. The inset scatter plot shows the distribution of RGA and RHA scores for these 235 stream reaches. The vertical and horizontal lines mark divisions between VTANR assigned categorical stream conditions.

only the "worst case" water quality parameters instead of using all nine water quality parameters. "Worst case" refers to water quality parameter measurements that exceed the Vermont Water Quality Standards. At reaches where more than one water quality parameter exceeds the standards, all parameters that exceed the defined thresholds are

Table 2. (a) Expert-Assessed RGA and RHA Scores for ($n = 1363$) VT Stream Reaches and RHA Classified Using (b) Discriminant Analysis Using Four Fluvial RGA Processes as the Covariates

		(a) Reality ($n = 1363$)					(b) Discriminant Analysis (69.97%)				
RHA	Excellent	0	2	15	39	Excellent	0	0	8	22	
	Good	6	216	261	23	Good	1	112	277	37	
	Fair	40	612	119	1	Fair	27	684	241	1	
	Poor	11	17	1	0	Poor	1	0	0	0	
		Poor	Fair	Good	Excellent		Poor	Fair	Good	Excellent	
			RGA					RGA			

Table 3. (a) Observed ($n = 235$) Classified Stream Reaches and RHA Classified Using (b) Discriminant Analysis and (c) k-Nearest Neighbor Using Four Fluvial RGA Processes as the Covariates

	(a) Reality ($n = 235$) (Raw Classified Data)					(b) Discriminant Analysis Using RGA (76.17%)					(c) k-Nearest Neighbor Using RGA (76.17%)				
RHA	Excellent	0	2	1	5	Excellent	0	1	0	4	Excellent	0	0	0	2
	Good	0	29	42	5	Good	0	14	50	6	Good	0	13	45	9
	Fair	7	117	22	0	Fair	8	135	15	0	Fair	9	136	20	0
	Poor	3	2	0	0	Poor	2	0	0	0	Poor	1	0	0	0
		Poor	Fair RGA	Good	Excellent		Poor	Fair RGA	Good	Excellent		Poor	Fair RGA	Good	Excellent

used to estimate of the likelihood function parameters. As a result, the posterior probability for stream-reach habitat classified as poor, fair, good, or excellent given "worst case" water quality measurements is formulated:

$$p(\text{RHA}|\text{``worst''\ WQ}) = p(\text{RHA}) * p(\text{``worst''\ WQ}|\text{RHA}). \quad (7)$$

3.2.3. Bayesian-SOM Network Performance

[44] In this section, we compare the Bayesian-SOM network to a discriminant analysis and a nonparametric k-nearest neighbor classification analysis using the reduced $n = 235$ data. To ensure a fair comparison with discriminant analysis, we constructed priors by setting them proportional to the RHA class occurrence in the original $n = 1363$ RGA/RHA data set. Adding the nine water quality parameters as covariates to the discriminant analysis and the k-nearest neighbor (Table 4, matrixes a and b) only improves our classification error by ~1.7% compared to that of the traditional discriminant analysis (Table 3, matrix c).

[45] Next, we compare the results of the Bayesian-SOM network on the same reduced ($n = 235$) data set using two sets of likelihood data (i.e., nine water quality parameters versus "worst case" water quality parameters of Table 5). Using the nine water quality parameters to construct the likelihood shows only slight improvement (79.57% classification accuracy, Table 5, matrix a) compared to discriminant analysis (77.87%) and the k-nearest neighbor method (77.44%) using the same information. However, the Bayesian-SOM network, using only the "worst case" parameters, improves classification error by an additional 5% (Table 5, matrix c).

[46] Lastly, we replaced the 2-D SOM clustering analysis with a Gaussian mixture model using the nine water quality parameters and one that uses the "worst case" water quality parameters of the above example. In both cases, classifications using the mixture models were similar to, but less accurate than, using the 2-D SOM (Table 6). The results again show classifications using the "worst case" water quality parameters to be more accurate (80% versus 72.76%, Table 6).

4. Discussion

[47] We develop and apply a new classification framework that couples a Naïve Bayesian classifier with a

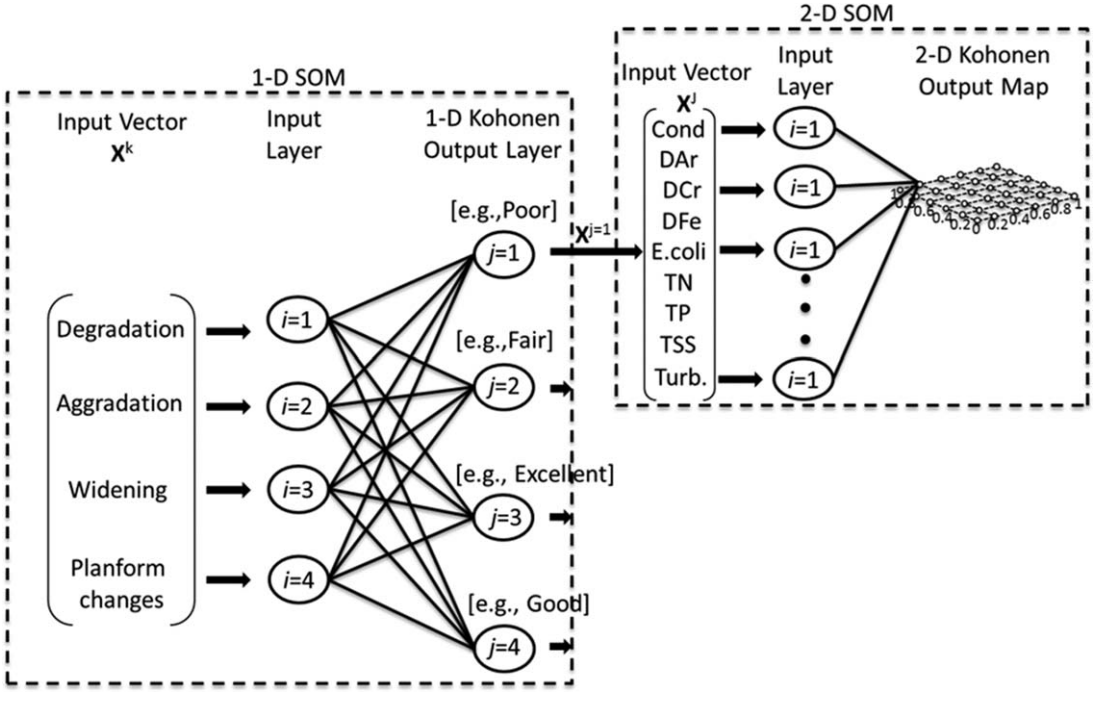

Figure 8. Network architecture of the likelihood pdf network used in Case Study 2. The input and output layers of the 1-D pdf network each have four nodes. The input vector of the 2-D SOM has nine nodes each corresponding to one of the nine select water quality parameters.

Table 4. RHA Classified Using (a) Discriminant Analysis and (b) k-Nearest Neighbor Using RGA and Water Quality Data for $n = 235$ VT Stream Reaches

	(a) Discriminant Analysis Using RGA and WQ (77.87%)					(b) k-Nearest Neighbor Using RGA and WQ (77.44%)				
RHA	Excellent	0	2	14	10	Excellent	0	0	0	1
	Good	0	30	40	0	Good	0	18	42	0
	Fair	7	109	11	0	Fair	10	131	32	0
	Poor	3	9	0	0	Poor	0	1	0	0
		Poor	Fair	Good	Excellent		Poor	Fair	Good	Excellent
			RGA					RGA		

nonparametric SOM clustering algorithm using two hydrological stream-monitoring studies to improve the clustering and feature extraction of large, nonlinear data networks. Identifying the number of clusters present in a given data set has been highlighted as one of the fundamental challenges associated with the clustering [*Jain and Dubes*, 1988]. It is still relevant today and the primary motivation for our using the 1-D classification and 2-D clustering SOMs in tandem. The 1-D Kohonen map serves to classify the likelihood data, where observations are assigned to one of some number of prespecified nodes. The 2-D Kohonen clustering algorithm works in tandem with the 1-D classifier to better identify the homogeneity of the input data classified to each of the 1-D Kohonen nodes. A commonly used technique (U-matrices of Figures 4 and 9) enables visualization of the sample spread associated within each class. We use the 2-D mappings of these high-dimensional data to estimate the probability distribution parameters. Even in the rare case of "perfect" classification (e.g., no misclassifications of Figure 4c), one can infer cluster similarity or at least check whether the data has a clustering tendency [*Smith and Jain*, 1984].

[48] One advantage of the network's operational phase (Figure 2) is the ability to dynamically update the prior based on recently "classified" information. In Case Study 1, the Bayesian-SOM network estimates the relative abundance of the two worm taxa slightly better than the Simple Bayesian and the Naïve Bayesian approaches (odds ratios of Table 1, columns 4–6) across all sites, except those with only one worm taxa (e.g., Naïve classifier slightly outperforms the Bayesian-SOM network at sites 2 and 4). We attribute this to the Bayesian-SOM's use of the Laplacian correction term; the latter ensures that sites without taxa have (pseudo) counts slightly greater than zero. It should be noted that in this case study, the prior data (morphologically identified sexually mature worms) comprise a relatively small number (n values of Table 1) of the sampled worm community and, therefore, are not a good representation of the site's overall worm composition. Since characterizing the community composition is critical to areas with high or low Whirling Disease risk, any method capable of leveraging the likelihood data associated with our more abundant worm population (i.e., immature worm population provided by the PCR probe data) will improve estimation, primarily because this information better represents the overall site-specific worm communities.

[49] The fact that we used a different statistical model (i.e., continuous Gaussian in the Bayesian-SOM rather than the discrete Beta binomial of the Simple Bayesian method) may explain the improved performance over the Simple Bayesian analysis. However, it should be noted that the Bayesian-SOM network performs better at sites where the observed relative abundances of Tt do not fall within the 90% credible interval calculated by the Naïve classifier (e.g., site 1 of Figure 5 where both the Bayesian-SOM network estimate and the real observation fall in the upper 2.5% of the relative abundance distribution). Overall, the Bayesian-SOM network suggests reliable estimation of the taxa relative abundance without the need for DNA sequencing.

[50] In Case Study 2, the Bayesian-SOM is used to classify stream-reach habitat health (RHA) using the VTANR reach-scale geomorphic assessments (RGA) and water quality data. In an ideal world, the correlation between RGA and RHA scores should be strongly positive. Figures 6 and 7 show positive correlation, but call attention to the need for additional information (e.g., water quality data) to improve habitat predictions. Both data sets are biased toward reaches classified as "fair" and "good" compared to the two extremes (poor and excellent). Preliminary discriminant analysis using the four components of RGA shows the classification accuracy of reach-scale RHA to be higher for the reduced $n = 235$ data set (76.17%; Table 3, matrix b) than the larger $n = 1363$ data set (69.97%; Table 2, matrix b) suggesting the reduced data set may be a better representation of the underlying RGA and RHA dynamics.

Table 5. RHA Classified Using the Bayesian-SOM Network and a Likelihood Pdf Generated Using (a) Nine Water Quality Parameters and (b) Only the "Worst Case" Water Quality Parameters for $n = 235$ VT Stream Reaches

	(a) Bayesian-SOM Network Using Nine WQ Parameters (79.14%)					(b) Bayesian-SOM Network Using "Worst Case" WQ Parameters (84.97%)				
RHA	Excellent	0	2	1	5	Excellent	0	1	0	3
	Good	3	17	56	0	Good	0	26	50	4
	Fair	2	122	22	0	Fair	7	107	32	0
	Poor	4	0	1	0	Poor	4	1	0	0
		Poor	Fair	Good	Excellent		Poor	Fair	Good	Excellent
			RGA					RGA		

Table 6. RHA classified Using Gaussian Mixture Models to Replace the 2-D SOM Clustering Using (a) Nine Water Quality Parameters and (b) "Worst Case" Water Quality Parameters for the $n = 235$ VT Streams

		(a) Gaussian Mixture Models Using Nine WQ Parameters (72.76%)					(b) Gaussian Mixture Models Using "Worst Case" WQ Parameters (80%)			
RHA	Excellent	0	1	1	4	Excellent	0	1	1	5
	Good	9	33	31	6	Good	1	33	30	5
	Fair	6	99	31	1	Fair	6	107	35	0
	Poor	3	8	2	0	Poor	3	5	2	0
		Poor	Fair RGA	Good	Excellent		Poor	Fair RGA	Good	Excellent

Adding nine water quality parameters to the discriminant analysis does not significantly increase RHA classification accuracy (76.17% versus 77.87%; Tables 3 matrix b and 4 matrix a, respectively); similar findings are observed when using the k-nearest neighbor classification method (Tables 3 matrix c and 4 matrix b). Given the substantial number of data provided by the water quality data set, we found this interesting, although not necessarily surprising, since the relationship between biological habitat health and water quality data are complex (i.e., fish can move, and macroinvertebrate data are often biased spatially to locations where one expects to find macroinvertebrates). As a result, we hypothesized that our nine water quality parameter data set, selected primarily on availability, might be at fault.

[51] To test the power of the Bayesian-SOM tandem network, we used the same four RGA covariates and two different sets of likelihood data—the nine water quality parameters versus "worst case" water quality parameters (i.e., those that exceed the state standards). The Bayesian-SOM showed only slight improvement over the better discriminant analysis approach (~79.14% versus 77.87%) using the nine water quality parameter likelihood data; whereas, using the "worst case" water quality data set improved classification error by an additional 6%. In addition, the worst-case likelihood data are particularly beneficial when classifying RHA scores in the stream-reach categories "fair" to "good." From a management perspective, this is useful because these two categories (1) comprise the majority of all assessed reaches, (2) have the weakest correlation between RGA and RHA, and (3) show stronger sensitivity to change compared to streams categorized in one of the two extremes. The latter approach not only provides more weight to water quality data of greatest concern at a local stream-reach scale, but also enables managers to identify spatial correlations between these reach-specific water quality data and reach-scale RHA for best management practice design (e.g., in areas where suspended sediments are the primary problem, the best

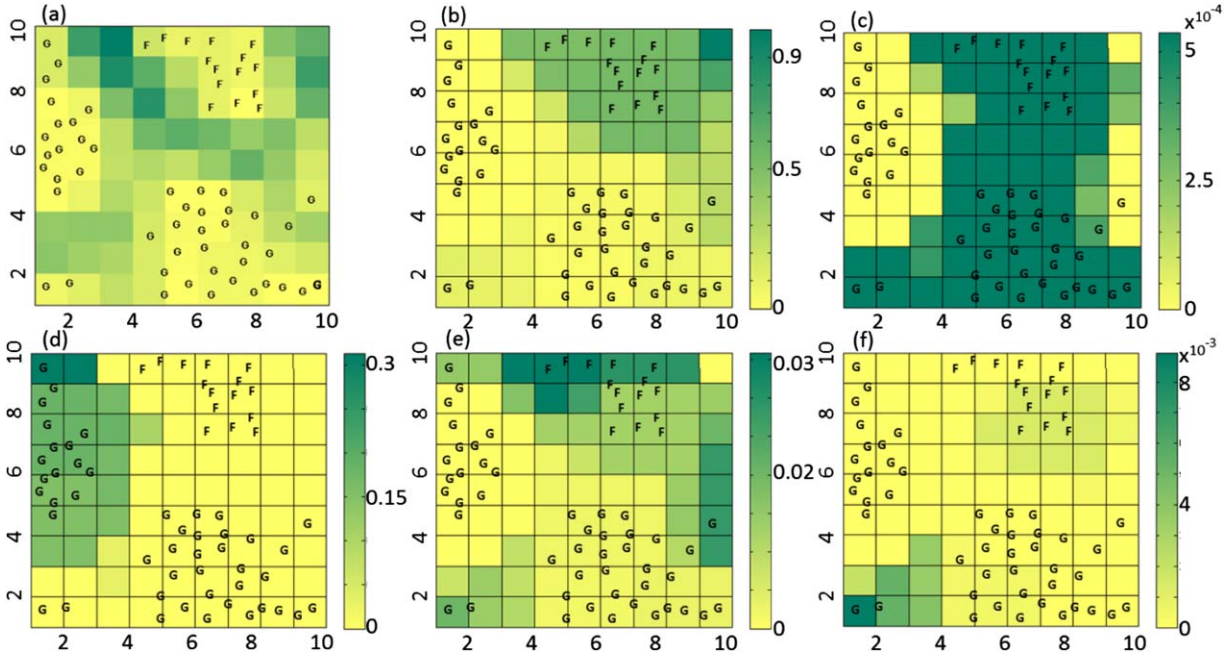

Figure 9. (a) U-matrix showing the three self-organized clustered to one of three 1-D SOM nodes. These three clusters correspond to stream reaches classified by experts as having fair and good stream habitat condition (labeled F and G, respectively). Five of the "worst case" water quality input data vectors are plotted in the same 2-D grid and represent (b) conductivity, (c) chromium/arsenic, (d) dissolved iron, (e) total phosphorus, and (f) total suspended solids. The latter are referred to as the component planes.

management practice would differ from areas where *E. coli* or phosphorus loadings are the major challenge). Since it has been noted that any clustering algorithm will improve the performance of a Naïve Bayesian classifier [*Chapelle et al*., 2006], we replaced the 2-D SOM clustering analysis with Gaussian mixture models (using the nine water quality parameters and one that uses the "worst case" water quality parameters, Table 6) to ensure that the improved classification was indeed a result of using the nonparametric SOM (as a choice of clustering method) in tandem with the Bayesian analysis.

[52] We also provide an example of the 2-D SOM "component planes" to further explore the relationship between the input data (e.g., in this case, water quality information) and the three self-organized stream habitat clusters (Figure 9a). The sample labels are superimposed corresponding to their classified RHA condition (e.g., expert-assigned categories good (G) or fair (F)). In this example, the component planes of the five "worst case" water quality parameters (Figures 9b–9f) for samples that clustered to a single 1-D node in the likelihood pdf network are plotted next to their associated U-matrix of Figure 9a. One advantage of plotting high-dimensional information onto 2-D maps is that a manager can readily see, for example, that the water quality parameter "conductivity" (Figure 9b) helps discriminate between RHA classified as fair and those classified as good. Whereas, chromium, arsenic and dissolved iron (Figures 9c and 9d) help discriminate between the two clusters classified as "good." Total phosphorus (Figure 9e) distinguishes fair from good; while total suspended solids (Figure 9f) do not appear to discriminate between any of the Figure 9a clusters, with perhaps the exception of the two samples classified as good (bottom left corner).

5. Conclusions

[53] Neither the SOM clustering algorithm nor Bayesian statistics are new; however, using the two in tandem leverage the prior information found in multiple types of data to minimize classification error and enhance the confidence in our classifications. The SOM clustering improves the performance of the Naïve Bayesian classifier; and we selected the latter because the class conditional independence assumption requires less training data and is simple to code. In addition, the method converges faster than discriminative models [*Jordan and Touretzky*, 2002]. In this work, we compare a variety of clustering methods and find the SOM provides two advantages over the more traditional clustering algorithms. First, combining the 1-D and 2-D SOMs to compute the likelihood pdf allows one to visualize the resulting 2-D likelihood data. This helps determine or optimize a sufficient number of nodes for the 1-D classification. Second, the component planes allow one to explore linkages between particular input features and the resulting classes or clusters. Visualizing relationships between high-dimensional inputs and the resulting clusters enables managers to hypothesize system processes that may explain the resulting groupings. For applications with a very large number of clusters, one can increase the size of the 2-D SOM map and assess the variability/similarity of the clusters. Next, the user can vary the credible interval to identify ambiguous samples, and if necessary, readjust the number of initial classes (1-D nodes) accordingly, and repeat the analysis. The network's ability to learn directly from the data and dynamically update both estimates and the associated uncertainty as the data collection evolves enables a more adaptive hydrological management approach. Bayesian uncertainty analysis and computational models in hydrology are clearly of vital importance and warrant future research that targets artificial neural network models coupled with Bayesian methods.

[54] **Acknowledgments.** This research was funded in part by NSF grant 1216193 as part of the joint NSF-NIH-USDA Ecology and Evolution of Infectious Diseases program, NSF grant 0842152 Division of Environmental Biology and Vermont Experimental Program to Stimulate Competitive Research (EPSCoR) with funds from the National Science Foundation grant EPS-1101317. We would like to thank Leslie Morrissey and Lori Stevens (from the University of Vermont) and Billie Kerans and graduate student Ryan Lamb (Ecology Department—Montana State University) for providing the morphological identifications for Case Study 1. Also, we would like to acknowledge Neil Kamman and Leslie Mathews from the Vermont Agency of Natural Resources, Monitoring, Assessment and Planning Program, and Gretchen Alexander and Mike Kline from the Vermont Agency of Natural Resources, Rivers Program at the Department of Environmental Conservation for providing data for Case Study 2 and fostering a productive collaboration. We sincerely thank the three anonymous reviewers, their comments and suggestions greatly enhanced the quality of this manuscript.

References

Aguilera, P. A., A. G. Frenich, J. A. Torres, H. Castro, J. L. M. Vidal, and M. Canton (2001), Application of the Kohonen neural network in coastal water management: Methodological development for the assessment and prediction of water quality, *Water Res.*, *35*(17), 4053–4062.

Alberto, W. D., D. M. Del Pilar, A. M. Valeria, P. S. Fabiana, H. A. Cecilia, and B. M. De Los Angeles (2001), Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River basin (Cordoba-Argentina), *Water Res.*, *35*(12), 2881–2894.

Androutsopoulos, I., J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos (2000), An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages, Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong editors, in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, edited by N. J. Belkin, P. Ingwersen, and M.-K. Leong, pp. 160–167, Association for Computing Machinery (ACM), Athens, Greece.

Balakrishnan, S., A. Roy, M. G. Ierapetritou, G. P. Flach, and P. G. Georgopoulos (2003), Uncertainty reduction and characterization for complex environmental fate and transport models: An empirical Bayesian framework incorporating the stochastic response surface method, *Water Resour. Res.*, *39*(12), 1350, doi:10.1029/2002WR001810.

Besaw, L. E., D. M. Rizzo, M. Kline, K. L. Underwood, J. J. Doris, L. A. Morrissey, and K. Pelletier (2009), Stream classification using hierarchical artificial neural networks: A fluvial hazard management tool, *J. Hydrol.*, *373*(1–2), 34–43.

Besaw, L. E., D. M. Rizzo, P. R. Bierman, and W. R. Hackett (2010), Advances in ungauged streamflow prediction using artificial neural networks, *J. Hydrol.*, *386*(1–4), 27–37.

Bickel, P., I. Johnstone, and B. Yu (2007), Session: Statistics in biological sciences and machine learning, paper presented at Workshop on Discovery in Complex or Massive Datasets: Common Statistical Themes, sponsored by the National Science Foundation's Division of Mathematical Sciences, Washington, D. C., 16–17 Oct.

Brinkhurst, R. O. (1986), *Guide to the Freshwater Aquatic Microdrile Oligochaetes of North America*, Dep. of Fish. and Oceans, Ottawa.

Chapelle, O., B. Schölkopf, and A. Zien (2006), *Semi-Supervised Learning*, MIT Press, Cambridge, Mass.

Cheng, B., and D. M. Titterington (1994), Neural networks—A review from a statistical perspective, *Stat. Sci.*, *9*(1), 2–30.

Chon, T. S., Y. S. Park, K. H. Moon, and E. Y. Cha (1996), Patternizing communities by using an artificial neural network, *Ecol. Modell.*, *90*(1), 69–78.

Dollar, E. S. J., C. S. James, K. H. Rogers, and M. C. Thoms (2007), A framework for interdisciplinary understanding of rivers as ecosystems, *Geomorphology*, *89*(1–2), 147–162.

Duda, R. O., and P. E. Hart (1973), *Pattern Classification and Scene Analysis*, vol. 17, 482 pp., John Wiley, New York.

Emmott, S., and S. Rison (2005), Session: Towards solving global challenges, paper presented at Workshop on Towards 2020 Science, sponsored by the Microsoft Research Organization, Venice, July.

Friedman, N., D. Geiger, and M. Goldszmidt (1997), Bayesian network classifiers, *Mach. Learning*, *29*(2–3), 131–163.

Fritzke, B. (1994), Growing Cell structures—A self-organizing network for unsupervised and supervised learning, *Neural Networks*, *7*(9), 1441–1460.

Fytilis, N., D. M. Rizzo, R. D. Lamb, B. L. Kerans, and L. Stevens (2013), Using real-time PCR and Bayesian analysis to distinguish susceptible tubificid taxa important in the transmission of *Myxobolus cerebralis*, the cause of salmonid whirling disease, *Int. J. Parasitol.*, *43*(6), 493–501.

Ghanem, R. (2009), Session: Uncertainty quantification, paper presented at Workshop on Opportunities and Challenges in Uncertainty Quantification for Complex Interacting Systems, Univ. of Southern Calif., Los Angeles, 12–14 Apr.

Gil, Y., E. Deelman, M. Ellisman, T. F. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers (2007), Examining the challenges of scientific workflows, *Computer*, *40*(12), 24–58.

Good, I. J. (1965), *The Estimation of Probabilities : An Essay on Modern Bayesian Methods*, vol. 12, 109 pp., MIT Press, Cambridge, Mass.

Govindaraju, R. S., and A. T. C. A. Artific (2000a), Artificial neural networks in hydrology. II: Hydrologic applications, *J. Hydrol. Eng.*, *5*(2), 124–137.

Govindaraju, R. S., and A. T. C. A. Artific (2000b), Artificial neural networks in hydrology. I: Preliminary concepts, *J. Hydrol. Eng.*, *5*(2), 115–123.

Han, J., M. Kamber, and J. Pei (2012), *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam.

Harris, N. M., A. M. Gurnell, D. M. Hannah, and G. E. Petts (2000), Classification of river regimes: A context for hydroecology, *Hydrol. Processes*, *14*(16–17), 2831–2848.

Haykin, S. S. (1999), *Neural Networks: A Comprehensive Foundation*, 2nd ed., vol. 21, 842 pp., Prentice Hall, Upper Saddle River, N. J.

Helsel, D. R., and R. M. Hirsch (1992), *Statistical Methods in Water Resources*, vol. 16, 522 pp., Elsevier, Amsterdam.

Jain, A. K. (2010), Data clustering: 50 years beyond K-means, *Pattern Recognition Lett.*, *31*(8), 651–666.

Jain, A. K., and R. C. Dubes (1988), *Algorithms for Clustering Data*, vol. 14, 320 pp., Prentice Hall, Englewood Cliffs, N. J.

Jordan, M. I., and D. S. Touretzky (2002), Advances in neural information processing systems, in Proceedings of the 2001 Conference, vol. 14, 841 pp., Kaufmann, San Mateo, Calif.

Kalteh, A. M., P. Hiorth, and R. Bemdtsson (2008), Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application, *Environ. Modell. Software*, *23*(7), 835–845.

Kaski, S. (1997), *Data Exploration Using Self-Organizing Maps*, *57* pp., Finn. Acad. of Technol., Espoo, Finland.

Katz, R. W., M. B. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Adv. Water Resour.*, *25*(8–12), 1287–1304.

Kerans, B. L., and A. V. Zale (2002), The ecology of *Myxobolus cerebralis*, paper presented at American Fisheries Society Symposium 29, Bethesada, Md.

Khatib, F., S. Cooper, M. D. Tyka, K. F. Xu, I. Makedon, Z. Popovic, D. Baker, and F. Players (2011), Algorithm discovery by protein folding game players, *Proc. Natl. Acad. Sci. U. S. A.*, *108*(47), 18,949–18,953.

Kingston, G. B., H. R. Maier, and M. F. Lambert (2005a), Calibration and validation of neural networks to ensure physically plausible hydrological modeling, *J. Hydrol.*, *314*(1–4), 158–176.

Kingston, G. B., M. F. Lambert, and H. R. Maier (2005b), Bayesian training of artificial neural networks used for water resources modeling, *Water Resour. Res.*, *41*, W12409, doi:10.1029/2005WR004152.

Kingston, G. B., H. R. Maier, and M. F. Lambert (2008), Bayesian model selection applied to artificial neural networks used for water resources modeling, *Water Resour. Res.*, *44*, W04419, doi:10.1029/2007WR006155.

Kohonen, T. (1982), Self-organized formation of topologically correct feature maps, *Biol. Cybern.*, *43*(1), 59–69.

Kohonen, T. (1990), The self-organizing map, *Proc. IEEE*, *78*(9), 1464–1480.

Kohonen, T. (2001), Self-organizing maps of massive databases, *Eng. Intell. Syst. Electr.*, *9*(4), 179–185.

Kohonen, T., E. Oja, O. Simula, A. Visa, and J. Kangas (1996), Engineering applications of the self-organizing map, *Proc. IEEE*, *84*(10), 1358–1384.

Kokkonen, T. S., and A. J. Jakeman (2001), A comparison of metric and conceptual approaches in rainfall-runoff modeling and its implications, *Water Resour. Res.*, *37*(9), 2345–2352.

Kondolf, G. M. (1995), Geomorphological stream channel classification in aquatic habitat restoration—Uses and limitations, *Aquat. Conserv.*, *5*(2), 127–141.

Kraaijveld, M. A., J. C. Mao, and A. K. Jain (1995), A nonlinear projection method based on kohonens topology preserving-maps, *IEEE Trans. Neural Networks*, *6*(3), 548–559.

Kramer, M. A. (1991), Nonlinear principal component analysis using autoassociative neural networks, *AIChE J.*, *37*(2), 233–243.

Krueger, R. C., B. L. Kerans, E. R. Vincent, and C. Rasmussen (2006), Risk of *Myxobolus cerebralis* infection to rainbow trout in the Madison River, Montana, USA, *Ecol. Appl.*, *16*(2), 770–783.

Lang, D., and D. Hogg (2011), Searching for comets on the World Wide Web: The orbit of 17PHolmes from the behavior of photographers, edited, *Astron. J.*, *144*(46), 11.

Leube, P. C., A. Geiges, and W. Nowak (2012), Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design, *Water Resour. Res.*, *48*, W02501, doi:10.1029/2010WR010137.

Lewis, D. (1998), Naive (Bayes) at Forty: *The Independence Assumption in Information Retrieval*, edited by C. Nédellec and C. Rouveirol, pp. 4–15, Springer, Berlin.

Maier, H. R., and G. C. Dandy (2000), Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications, *Environ. Modell. Software*, *15*(1), 101–124.

Malakoff, D. (1999), Bayes offers a 'new' way to make sense of numbers, *Science*, *286*(5444), 1460–1464.

Mangiameli, P., S. K. Chen, and D. West (1996), A comparison of SOM neural network and hierarchical clustering methods, *Eur. J. Oper. Res.*, *93*(2), 402–417.

Manukyan, N., M. J. Eppstein, and D. M. Rizzo (2012), Data-driven cluster reinforcement and visualization in sparcely-matched self-organinzing maps, *IEEE Trans. Neural Networks Learning Syst.*, *23*(5), 846–852.

Mariethoz, G., P. Renard, and J. Caers (2010), Bayesian inverse problem and optimization with iterative spatial resampling, *Water Resour. Res.*, *46*, W11530, doi:10.1029/2010WR009274.

McLachlan, G. J., and D. Peel (2000), *Finite Mixture Models*, vol. 22, 419 pp., John Wiley, New York.

Mitra, S., S. K. Pal, and P. Mitra (2002), Data mining in soft computing framework: A survey, *IEEE Trans. Neural Networks*, *13*(1), 3–14.

Monk, W. A., P. J. Wood, D. M. Hannah, D. A. Wilson, C. A. Extence, and R. P. Chadd (2006), Flow variability and macroinvertebrate community response within riverine systems, *River Res. Appl.*, *22*(5), 595–615.

Moradkhani, H., K. Hsu, H. V. Gupta, and S. Sorooshian (2004), Improved streamflow forecasting using self-organizing radial basis function artificial neural networks, *J. Hydrol.*, *295*(1–4), 246–262.

Nathan, R. J., and T. A. Mcmahon (1990), Identification of homogeneous regions for the purposes of regionalisation, *J. Hydrol.*, *121*(1–4), 217–238.

Paruelo, J. M., and F. Tomasel (1997), Prediction of functional characteristics of ecosystems: A comparison of artificial neural networks and regression models, *Ecol. Modell.*, *98*(2–3), 173–186.

Pegg, M. A., and C. L. Pierce (2002), Classification of reaches in the Missouri and lower Yellowstone rivers based on flow characteristics, *River Res. Appl.*, *18*(1), 31–42.

Poff, N. L. (1996), A hydrogeography of unregulated streams in the United States and an examination of scale-dependence in some hydrological descriptors, *Freshwater Biol.*, *36*(1), 71–91.

Provost, F., and P. Domingos (2003), Tree induction for probability-based ranking, *Mach. Learning*, *52*(3), 199–215.

Puckridge, J. T., F. Sheldon, K. F. Walker, and A. J. Boulton (1998), Flow variability and the ecology of large rivers, *Mar. Freshwater Res.*, *49*(1), 55–72.

Rabeni, C. F., K. E. Doisy, and D. L. Galat (2002), Testing the biological basis of a stream habitat classification using benthic invertebrates, *Ecol. Appl.*, *12*(3), 782–796.

Reed, P. M., and J. B. Kollat (2012), Save now, pay later? Multi-period many-objective groundwater monitoring design given systematic model errors and uncertainty, *Adv. Water Resour.*, *35*, 55–68.

Richard, M. D., and R. P. Lippmann (1991), Neural network classifiers estimate Bayesian a posteriori probabilities, *Neural Comput.*, *3*(4), 461–483.

Roache, P. J. (1997), Quantification of uncertainty in computational fluid dynamics, *Annu. Rev. Fluid Mech.*, *29*, 123–160.

Schalkoff, R. J. (1992), *Pattern Recognition: Statistical, Structural, and Neural Approaches*, vol. 19, 364 pp., John Wiley, New York.

Smith, S. P., and A. K. Jain (1984), Testing for uniformity in multidimensional data, *IEEE Trans. Pattern Anal.*, *6*(1), 73–81.

Smith, T. J., and L. A. Marshall (2008), Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques, *Water Resour. Res.*, *44*, W00B05, doi:10.1029/2007WR006705.

Snelder, T. H., B. J. F. Biggs, and R. A. Woods (2005), Improved eco-hydrological classification of rivers, *River Res. Appl.*, *21*(6), 609–628.

Solomatine, D. P., and A. Ostfeld (2008), Data-driven modelling: Some past experiences and new approaches, *J. Hydroinformatics*, *10*(1), 3–22.

Steinschneider, S., A. Polebitski, C. Brown, and B. H. Letcher (2012), Toward a statistical framework to quantify the uncertainties of hydrologic response under climate change, *Water Resour. Res.*, *48*, W11525, doi:10.1029/2011WR011318.

Szalay, A., and J. Gray (2006), Science in an exponential world, *Nature*, *440*(7083), 413–414.

Titterington, D. M. (2004), Bayesian methods for neural networks and related models, *Stat. Sci.*, *19*(1), 128–139.

Ultsch, A., and H. P. Siemon (1989), *Exploratory Data Analysis Using Kohonen Networks on Transputers*, 50 pp., Dekanat Informatik, Dortmund, Germany.

Vincent, E. R. (1996), Whirling disease and wild trout: The Montana experience, *Fisheries*, *21*(6), 32–33.

Wagner, W., N. E. C. Verhoest, R. Ludwig, and M. Tedesco (2009), Editorial 'Remote sensing in hydrological sciences', *Hydrol. Earth Syst. Sci.*, *13*(6), 813–817.

Wan, E. A. (1990), Neural network classification: A Bayesian interpretation, *IEEE Trans. Neural Networks*, *1*(4), 303–305.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole (2007), Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, *Appl. Environ. Microbiol.*, *73*(16), 5261–5267.

Williams, J. L., and R. M. Maxwell (2011), Propagating subsurface uncertainty to the atmosphere using fully coupled stochastic simulations, *J. Hydrometeorol.*, *12*(4), 690–701.

Wolf, K., M. E. Markiw, and J. K. Hiltunen (1986), Salmonid whirling disease—Tubifex-Tubifex (Muller) identified as the essential oligochete in the protozoan life-cycle, *J. Fish Diseases*, *9*(1), 83–85.

Wright, J. F. (2000), *Assessing the Biological Quality of Fresh Waters RIVPACS and Other Techniques*, vol. 24, 373 pp., Freshwater Biol. Assoc., Ambleside, U. K.

Yoon, Y. O., G. Swales, and T. M. Margavio (1993), A comparison of discriminant-analysis versus artificial neural networks, *J. Oper. Res. Soc.*, *44*(1), 51–60.

Zhang, G. Q., B. E. Patuwo, and M. Y. Hu (1998), Forecasting with artificial neural networks: The state of the art, *Int. J. Forecasting*, *14*(1), 35–62.

Zhang, X. S., and K. G. Zhao (2012), Bayesian neural networks for uncertainty analysis of hydrologic modeling: A comparison of two schemes, *Water Resour. Manage.*, *26*(8), 2365–2382.