A framework for coupling explanation and prediction in hydroecological modelling[☆]

Ben W.J Surridge [a,*], Simone Bizzi [b], Andrea Castelletti [c]

[a] Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster LA1 4YQ, UK
[b] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Via E. Fermi 2749, I-21027 Ispra, VA, Italy
[c] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy

Received 14 June 2013 Received in revised form
12 February 2014
Accepted 13 February 2014 Available online 18 March 2014

1. Introduction

Causal explanation and empirical prediction are distinct challenges facing environmental modelling. Explanatory modelling draws on hypotheses from *a priori* theory to specify causal relationships between predictor and response variables. These hypotheses are subsequently tested using statistical models, often involving the application of regression models to empirical data (Shmueli, 2010). In contrast, predictive modelling seeks optimal performance of a model, often defined in terms of accuracy of prediction for new or future observations of a response variable. This paper is concerned with coupling explanatory and predictive modelling of ecological data. The statistical challenges of ecological datasets, including complexity, non-linearity and multicollinearity, have generated particular interest in the use of machine learning methods to build predictive ecological models (e.g. Feio et al., 2013; Crisci et al., 2012; Goethals et al., 2007; Mouton et al., 2009; Shan et al., 2006). The predictive capability of models derived from machine learning depends on pattern recognition within an empirical dataset, rather than *a priori* representation of underlying ecological processes within a model structure. Underpinned by an inductive approach to modelling, machine learning methods draw on causal explanation primarily through the use of theory to provide process-based interpretations of statistical patterns identified within a dataset.

Explanation and prediction rarely coincide in ecological modelling (De'ath, 2007). Explanatory approaches yield models that are not usually evaluated for their predictive performance, or for which high explanatory power is erroneously conflated with

Abbreviations: SEM, structural equation modelling; ASPT, average score per taxon; LIFE, lotic-invertebrate index for flow evaluation; GQA, general quality assessment; RHS, river habitat survey; ET, extremely randomised trees; IIS, iterative input variable selection.

[☆] Thematic Issue on Novel Approaches to Challenges in Aquatic Ecosystem Modelling.

[*] Corresponding author. Tel.: +44 (0)1524 594516.

E-mail addresses: b.surridge@lancaster.ac.uk (B.W.J Surridge), simone.bizzi@jrc.ec.europa.eu (S. Bizzi), andrea.castelletti@polimi.it (A. Castelletti).

high predictive power (Shmueli, 2010). The utility of process-based, predictive models derived from explanatory approaches can be limited by complexity and by over-parameterisation, leading to equifinality with respect to model structure (e.g. Beven, 2006; Jakeman and Hornberger, 1993). Models with reasonable predictive power can be developed through machine learning, despite the statistical challenges of ecological datasets (e.g. Kocev et al., 2010; Tirelli et al., 2009). However, the structure of models derived from techniques including artificial neural networks and support vector machines often lacks ecological or physical interpretability (Geurts et al., 2009; Young, 2013). This limits the potential to use the outcomes of predictive modelling to inform the development of new theory regarding the causal organisation of ecological systems.

Despite philosophical and technical distinctions between explanatory and predictive modelling (*cf.* Shmueli, 2010), frameworks that support integration between these apparent extremes could enhance feedback among causal theory, empirical data and prediction. This paper proposes and evaluates a new framework in this context, based on novel explanatory and predictive modelling components applied to exemplar datasets from the field of freshwater ecology. In ecological terms, our research considered how spatial variation in benthic macroinvertebrate community indices from riverine ecosystems could be both explained and predicted.

Benthic macroinvertebrates are perhaps the most widely used bioindicator group in lotic ecosystems (e.g. Kokes et al., 2006; Wright, 2000), due to their hypothesised sensitivity to both chemical water quality (e.g. Sandin and Hering, 2004; Vaughan and Ormerod, 2012) and hydromorphological conditions including stream discharge and water velocity (e.g. Dunbar et al., 2010; Jowett, 2003; Monk et al., 2007). Benthic macroinvertebrates are also key components of riverine food webs, with the potential to influence the status of organisms at lower trophic levels, such as phytobenthic algae on which some macroinvertebrate taxa feed (e.g. Biggs, 1996), and organisms at higher trophic levels such as fish that prey on macroinvertebrate taxa (e.g. Winklemann et al., 2011). Whilst models have been developed to explain spatial variation in benthic macroinvertebrate communities (e.g. Lücke and Johnson, 2009; Murphy and Davy-Bowker, 2005; Poff et al., 2010; Townsend et al., 2003; Vaughan and Ormerod, 2012), few predictive models exist for these communities, particularly along a gradient of environmental stress that includes human-disturbed ecosystems. To our knowledge, no previous research has sought to couple explanatory and predictive modelling of benthic macroinvertebrate communities within a single framework.

Increasing spatial and temporal alignment of monitoring networks in riverine ecosystems (Vaughan and Ormerod, 2010) offers a basis for developing models with greater explanatory or predictive power. However, the ecological datasets that emerge from these monitoring networks suffer from the curse of dimensionality as they grow in both size and complexity. Such datasets contain variables that are informative alongside variables that are either redundant or irrelevant (Tirelli and Pessani, 2011). Under these conditions, alternative combinations of input variables selected from a range of candidate inputs can generate predictive models that differ in accuracy and in the extent to which the resulting model structure is ecologically interpretable (D'heygere et al., 2003). By avoiding the interference associated with non-relevant or redundant information, input variable selection (IVS) more effectively exploits the data available for model calibration, ultimately providing accurate and interpretable models (Guyon and Elisseeff, 2003; Mouton et al., 2010). The use of domain or expert knowledge is a common approach to IVS in the context of machine learning (Maier et al., 2010). Expert knowledge, alongside the availability of empirical data, is often used in an ad-hoc way to select input variables for predictive modelling. Alternatively, expert knowledge can be used to specify different combinations of input variables from which predictive models are built, assessing the predictive performance of each model to identify the most appropriate input variables. However, IVS based on expert knowledge can be perceived as unreliable due to significant variation in understanding (and therefore the input variables selected) between individual cases and individual experts (D'heygere et al., 2006), a particular challenge given complex and heterogeneous ecosystems at community levels of organisation (Lawton, 1999). Consequently, mathematical approaches to IVS that are largely disconnected from underlying theory are often used, including model-free (e.g. correlation) and model-based (e.g. pruning) techniques (Maier et al., 2010).

However, recent advances in explanatory modelling of ecological communities now enable a consistent, statistical evaluation to be made of expert knowledge as the basis for IVS. Specifically, through the Structural Equation Modelling framework (SEM, e.g. Bollen, 1989; Grace, 2006) *a priori* theoretical models of ecological systems can be specified and assessed against empirical data. Alongside a robust approach to variable selection given the statistical challenges of ecological datasets, including non-relevance and redundancy (e.g. Bollen, 1989), SEM also provides a simultaneous evaluation of multiple direct and indirect relationships between variables within an ecological system. Therefore, by using SEM to assess the role of individual variables within a complex system, sets of variables can be selected and carried forward for predictive modelling. This is important because the most informative individual variables may not correspond to the most informative sets of variables, meaning that any approach to IVS should seek the best combination of input variables from candidate inputs if optimal predictive models are to be built (Tirelli and Pessani, 2011 and references therein).

Our research coupled SEM for causal explanation and IVS with the development of predictive models using the Iterative Input variable Selection (IIS) algorithm (Castelletti et al., 2012; Galelli and Castelletti, 2013a). The IIS algorithm is a recently developed input variable selection algorithm that combines a ranking-based preselection of the candidate inputs with a subsequent model-based evaluation and filtering of the more useful preselected inputs in terms of model predictive capability. Extremely Randomised Trees (ET), a novel machine learning approach for building tree-based predictive models (Geurts et al., 2006; Galelli and Castelletti, 2013b), is used to both rank the candidate inputs and incrementally build the final predictive model. Our research did not involve an inter-comparison of the numerous predictive model classes that are available (for examples of such inter-comparisons see: Crisci et al., 2012; Geurts et al., 2006; Hoang et al., 2010). Instead, the IIS algorithm (and thereby ET) was used as an exemplar model class because it offers particular advantages in terms of handling large, non-linear data sets composed of heterogeneous variables in a computationally efficient way (Galelli and Castelletti, 2013b). Whilst tree-based methods have previously been applied to ecological datasets (e.g. Chen and Mynett, 2004; Jung et al., 2010), and ET have been extended to hydrological and water quality applications (e.g. Castelletti et al., 2010; Fornarelli et al., 2013), to our knowledge we report one of the first applications of ET in the context of hydroecological data. Although SEM could be operated in predictive mode, for applications such as that reported here that incorporate latent variables, a Bayesian framework is required in order to estimate metrics for the latent variables (Arhonditsis et al., 2006). Whilst comparing predictive models built using SEM with those from other predictive model classes, such as ET, provides opportunities for future work, it was beyond the scope of our research. Instead, we confine the use of SEM to confirmatory mode for causal explanation. Therefore, we propose a novel modelling framework (SEM-IIS), based on an explicit combination of explanatory (SEM) and predictive (IIS) components. Our objective was to assess whether SEM-IIS can effectively couple the ecological interpretability of models developed using SEM with the predictive capability of models built using the IIS algorithm.

2. Material and methods

2.1. Datasets

A common empirical dataset provided the basis for SEM and IIS modelling. Data were drawn from monitoring of rivers and streams conducted by the Environment Agency (EA) as part of the periodic General Quality Assessment (GQA) and River Habitat Survey (RHS) exercises across England, UK. Under the biological GQA, benthic macroinvertebrates are sampled once every three years at each monitoring station following a standardised 3-min kick sampling procedure using a 900-μm mesh net, where all habitats within a site are sampled in proportion to their occurrence (see Murray-Bligh, 1999 for further sampling details). Two benthic macroinvertebrate community indices were calculated from biological GQA data and used as response variables in our models. Firstly, the Average Score Per Taxon (ASPT), representing the average Biological Monitoring Working Party (BMWP) score given to scoring taxa identified in a sample. The BMWP score is an expert-defined metric that indicates the sensitivity of taxa to organic pollution. The ASPT score for a benthic macroinvertebrate community is hypothesised to be positively correlated with the sensitivity of the community to organic pollution (Armitage et al., 1983; Clarke et al., 2003). Secondly, the Lotic-invertebrate Index for Flow Evaluation (LIFE), also representing the average score assigned to taxa within a sample. The LIFE index is an expert-defined metric hypothesised to reflect the sensitivity of benthic macroinvertebrates to prevailing flow conditions, with higher LIFE scores indicating a preference for rapid flow types and lower LIFE scores indicative of a preference for lentic flow conditions (Extence et al., 1999). Both ASPT and LIFE are widely-established indices and are frequently included as response variables in statistical analyses of the controls on benthic macroinvertebrate communities (e.g. Dunbar et al., 2010; Lücke and Johnson, 2009; Vaughan and Ormerod, 2012).

Twenty-one candidate input variables were drawn from the GQA chemistry and nutrients and from the RHS databases (Table 1). Candidate input variables hypothesised to control APST and LIFE were identified from the theoretical basis to each index (see Armitage et al. (1983) for the ASPT index and Extence et al. (1999) for the LIFE index), from previous modelling that has sought to explain spatial variation in these indices (e.g. Dunbar et al., 2010; Lücke and Johnson, 2009; Feld and Hering, 2007; Vaughan and Ormerod, 2012), and from the availability of variables within the national monitoring databases. General quality assessment chemistry and nutrient data from three years of monthly grab samples immediately preceding the date of each GQA biology sample were used to calculate 90th percentile BOD_5 and ammonia concentrations, 10th percentile dissolved oxygen concentration, and average nitrate and reactive phosphate concentrations. Individual sites were surveyed under the RHS once during the period 1995–2003, and the date of macroinvertebrate sampling from the GQA biology network was selected to be closest to the RHS survey date at each individual monitoring station (difference in years: $\bar{x} = 2$ years, $\sigma = 3.7$ years). The number of riffles, pools, unvegetated and vegetated point bars, and the HMS were extracted directly from the RHS database. Indices describing the calibre of bed and bank sediments, the structure of bank and in-channel vegetation and stream power were calculated from RHS data following Emery et al. (2004). Map-derived variables (slope, altitude, distance to source and height of source) were defined on the basis of the RHS monitoring station location. Geographical coordinates were derived from the location of GQA biology monitoring stations. Sites were only included within the database if the location of GQA chemistry, nutrients, biology and RHS monitoring stations coincided within a stream length of 500 m, yielding a final database of 267 sites (see Fig. 1) that span a wide range of upland–lowland environments (altitude of sites: $\bar{x} = 54$ m, range 0–201 m; distance of site from river source: $\bar{x} = 18$ km, range 0.3–175 km). This database also spans a gradient from sites that have been minimally to those that have been heavily impacted by anthropogenic pressure (for example, minimum–maximum range for average dissolved oxygen concentration = 56.5–116.4% saturation, for average reactive phosphate concentration = 0.004–4.573 mg PO_4 L^{-1}, and for habitat modification score = 0–8210). Data covered the period 1995–2003 during which rates of biological change in benthic macroinvertebrate communities in the UK were modest (Vaughan and Ormerod, 2012), meaning that the causal system that was hypothesised to explain spatial variation in ASPT and LIFE (see Section 2.2.) was assumed to remain constant.

2.2. Explanatory framework: structural equation modelling

The SEM framework was used to evaluate causal hypotheses regarding the system controlling spatial variation in ASPT and LIFE, and thereby to select input variables from Table 1 using expert knowledge that were carried forward for subsequent predictive modelling using the IIS algorithm. The SEM framework has roots in the analysis of quantitative data from the fields of biology, economics, psychology and sociology (Grace, 2006; Hägglund, 2001). Statistically, SEM is a multivariate technique that incorporates both path and factor analysis (for detailed treatments of the statistical and broader basis to SEM see Bollen (1989) and Grace (2006)). Interest

Table 1
Summary of the 21 candidate input variables used for both SEM and IIS modelling. RHS = River Habitat Survey database and GQA = General Quality Assessment database.

Variable	Description	Abbreviation
Biochemical oxygen demand	90th percentile five-day biochemical oxygen demand (mg L^{-1}) from 3 years of monthly sampling before the biological sample	BOD_5
Nitrate concentration	Average nitrate concentration (mg NO_3 L^{-1}) from 3 years of monthly sampling before the biological sample	NO_3
Ammonia concentration	90th percentile total ammonia concentration (mg N L^{-1}) from 3 years of monthly sampling before the biological sample	NH_4
Orthophosphate concentration	Average reactive phosphate concentration (mg PO_4 L^{-1}) from 3 years of monthly sampling before the biological sample	PO_4
Dissolved oxygen concentration	10th percentile dissolved oxygen concentration (% saturation) from 3 years of monthly sampling before the biological sample	O_2
Number of riffles	Presence (in numbers) of each stream habitat feature within the 500-m RHS reach	Riffles
Number of pools		Pools
Number of unvegetated point bars		Unveg bars
Number of vegetated point bars		Veg bars
Habitat modification score	Dimensionless index related to artificial and engineering features present within the 500-m RHS reach, see Raven et al. (1998)	HMS
Bank material calibre index	Index describing average sediment size of bank material within the 500-m RHS reach (approximate phi units, after Emery et al., 2004)	BMCI
Bed sediment calibre index	Index describing average sediment size of channel substrate within the 500-m RHS reach (approximate phi units, after Emery et al., 2004)	BSCI
Bank vegetation structure	Dimensionless index representing average bank vegetation structure (bare, uniform, simple or complex) within the 500-m RHS reach, after Emery et al. (2004).	BVS
In-channel vegetation structure	Dimensionless index representing extent of flow resistance due to in-channel vegetation structure within the 500-m RHS reach, after Emery et al. (2004).	CVSI
Valley slope	Map derived data from the RHS database (m km^{-1})	Slope
Total stream power index	Cross sectional area (at bankfull)* slope (−)	TSPI
Altitude	Map derived data from the RHS database (m)	Altitude
Distance to source	Map derived data from the RHS database (m)	DtS
Height of source	Map derived data from the RHS database (m)	HoS
Eastings	Geographical coordinates based on location of GQA biology monitoring site	East
Northings	Geographical coordinates based on location of GQA biology monitoring site	North

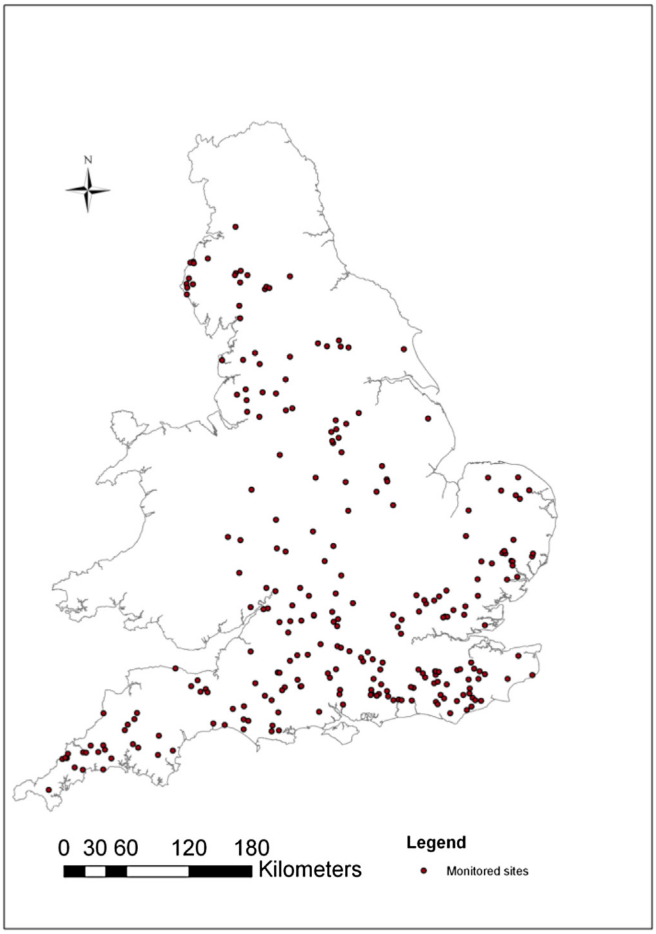

Fig. 1. Geographical location of the 267 monitoring sites across England, UK from which empirical data were drawn for explanatory and predictive modelling.

in using SEM to analyse ecological systems has emerged recently (e.g. Arhonditsis et al., 2007; Bernot et al., 2010; Mulholland et al., 2009; Reckhow et al., 2005), partly because of the potential to explicitly couple theoretical knowledge and empirical data analysis within a single modelling framework. This coupling relies on specifying constructs that are of general, theoretical importance within ecological systems as latent variables, and on these constructs being causally related to observed variables (Grace et al., 2010).

Latent and observed variables are connected in SEM through a network of direct and indirect paths, representing a hypothesis regarding the causal organisation of an ecological system. In the terminology of SEM, exogenous variables refer to independent variables and endogenous variables refer to dependent variables. The hypothesised model is specified as a structural equation (SE) model, yielding an expected covariance matrix which is evaluated against the covariance matrix from observed data. This comparison provides information about the overall fit of a SE model to observed data, usually on the basis of maximum likelihood estimation and chi-square tests. In contrast to statistical tests in which rejection of a null hypothesis is sought, SEM is theory-orientated and seeks acceptance of the *a priori* model. Given a *p*-value for the chi-square statistic ≥ 0.05, no significant difference between the observed and the model-implied covariance matrices is assumed at the 95% confidence level, and the model is deemed to be adequately supported by the observed data. In addition to an evaluation of overall model fit, information regarding the significance and strength of individual paths within a SE model is also provided. Standardised path coefficients are reported in this paper, calculated as the product of the unstandardized coefficients and the ratio of the standard deviations of the variables on either end of a path. Standardised path coefficients indicate the strength of the relationship between variables in the models, and are reported here because a common dataset (and therefore a common set of standard deviations for variables) underpinned the evaluation of SE models for both LIFE and ASPT.

A multivariate normal data set is required by SEM (Bollen, 1989). The variables selected for analysis in this research were tested for both univariate and multivariate normality. Logarithmic or square root transformations were applied to data where necessary. The existence of influential observations and outliers was also examined before and after transformations were applied. Linear relationships between predictor and response variables are assumed as part of SEM. Under conditions of nonlinearity, linearising transformations can be applied to data, for example, using a bivariate curvilinear regression (Grace and Pugesek, 1997; Weiher et al., 2004). Unimodal relationships, which are common in ecological data (ter Braak and Verdonschot, 1995), and quadratic relationships were also investigated through regression analysis applied to each relationship between a biotic index and a predictor variable. No evidence was found that nonlinear models improved model fit significantly compared with linear models. This may be due to the use of scoring indices for the benthic macroinvertebrate communities, which mask nonlinear responses of taxa to individual variables.

Structural equation models were specified for the ASPT and LIFE indices (Fig. 2). These models draw on earlier research reported in Bizzi et al. (2013) and are underpinned by the theoretical basis to each index, by the hypothesised controls on benthic macroinvertebrate communities and by the availability of candidate input variables. Alternative SE models, for example based on alternative theories regarding controls on benthic macroinvertebrate communities or on causal networks that differ in complexity, were not evaluated here. However, such comparisons could be undertaken based on the chi-square test statistic for alternative SE models, and by using additional evaluation metrics to account for differences in complexity between individual SE models, such as the Akaike information criterion (Akaike, 1974). The concentration of dissolved oxygen within the water column strongly influences the taxonomic composition of benthic macroinvertebrate communities (e.g. Friberg et al., 2010; Golubkov et al., 1992). Although observed dissolved oxygen concentration data are available from the GQA chemistry network, monthly spot sampling as part of this network will not capture all biologically-relevant dissolved oxygen concentrations, such as potential minima at night due to respiration. Therefore, the latent variable Effective Oxygen was specified in the SE model to represent the theoretical control exerted by dissolved oxygen concentration on benthic macroinvertebrate communities. The measured concentration of dissolved oxygen was used as a single effect indicator for this latent. Effective Oxygen is a second-order latent within the SE model, with the endogenous latent variables Organic Matter Load and Channel Hydromorphology acting as causal indicators. Paths are hypothesised between Channel Hydromorphology and Effective Oxygen, reflecting re-aeration during turbulent transfer of water through morphological features such as riffles, and between Organic Matter Load and Effective Oxygen based on the consumption of dissolved oxygen during aerobic respiration of biodegradable organic matter by microbial communities.

The latent variable Organic Matter Load is specified with the effect indicator BOD_5 representing the biodegradable fraction of this load. The causal indicator PO_4 is formatively related to Organic Matter Load, representing nutrient-related controls on primary production and the autochthonous production of organic matter within streams and rivers (Biggs, 2000; Kelly et al., 2008). Channel Hydromorphology is a theoretical construct that reflects interactions between channel hydraulics, sediment flux and channel morphology. This latent represents the ecological effects of these physical processes, and specifically the hypothesised influence of mesoscale physical habitat features and hydraulic conditions on benthic macroinvertebrate communities (e.g. Buffagni et al., 2004; Newson and Newson, 2000; Wiens, 2002). Over the longer timescales relevant to analysis of reach-scale spatial variation in fluvial systems, rather than shorter-term temporal trends, the interplay between channel hydraulics, sediment flux and channel morphology ultimately determines reach-scale characteristics such as bed sediment size and slope (Frissell et al., 1986). Therefore, the observed variables BSCI and slope are specified as effect indicators for the latent variable Channel Hydromorphology within the SE models. The latent variable Channel Hydromorphology represents a gradient between high slope-coarse sediment conditions (low values of the latent variable) and low slope-fine sediment conditions (high values of the latent variable).

The path between Vegetation Structure and Channel Hydromorphology represents the theoretical influence of in-channel vegetation on channel hydraulic conditions and sediment transport, for example through changes in flow resistance, flow velocity and sediment transport associated with the growth and senescence of aquatic vegetation (e.g. Asaeda et al., 2010; Gurnell et al., 2006). The in-channel vegetation structure index was specified as an effect indicator for the latent Vegetation Structure. This index reflects the extent to which in-channel vegetation increases resistance to flow (Emery et al., 2004), thereby reducing flow velocity and sediment transport capacity. A direct path was specified between HMS and LIFE or ASPT to represent the effect of hydromorphological disturbance on benthic macroinvertebrate communities (e.g. Feld and Hering, 2007). A further path was specified between NH_4 and ASPT or LIFE to represent the toxic effects of elevated unionised ammonia concentrations on macroinvertebrate taxa (e.g. Williams et al., 1986), based on the hypothesis that total ammonia and unionised ammonia concentrations are positively correlated (e.g. Berenzen et al., 2001). Whilst the ASPT and LIFE indices were originally conceived to reflect the sensitivity of benthic macroinvertebrate communities to distinct conditions within streams and rivers, Fig. 2 represents a hypothesis in which spatial variation in ASPT and LIFE can be explained by the same theoretical model. The hypothesis that a common causal system controls spatial variation in ASPT and LIFE is supported by the strong positive bivariate correlation between these indices ($\rho = 0.78$).

2.3. *Predictive modelling: iterative input variable selection algorithm*

In contrast to SEM, the IIS algorithm focuses solely on prediction and operates in black-box fashion without the use of *a-priori* domain knowledge to select input

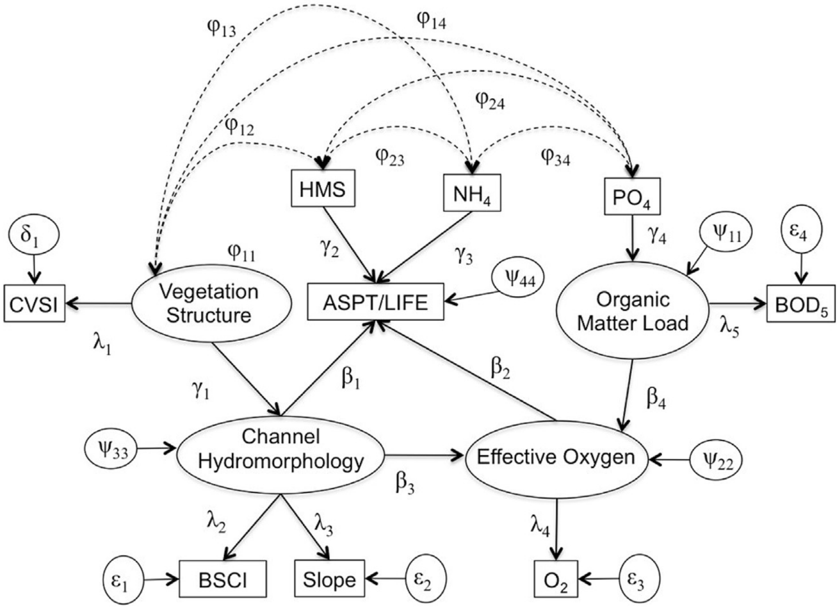

Fig. 2. Structural equation model specified for the ASPT and LIFE indices. Measured variables are represented by rectangles, latent variables by ellipses. For variable abbreviations see Table 1. Directed arrows between measured and latent variables represent hypothesised direction of causation. The term δ_i represents measurement error for the indicators of the exogenous latent variables, ε_i represents measurement error for the indicators of the endogenous latent variables, λ_i represents path coefficients between the latent variables and their indicators, γ_i represents path coefficients for exogenous (latent and measured) variables, β_i represents path coefficients for endogenous latent variables, φ_i represents variance of the exogenous latent variables, ψ_i represents variance of the structural (latent) errors.

variables or to specify the model structure. The IIS algorithm is a novel input variable selection method (Galelli and Castelletti, 2013a) that combines a ranking-based measure of input variable significance (Wehenkel, 1998) and a stepwise forward selection model-based approach (see Galelli and Castelletti, 2013b). Although we did not seek to formally compare the range of automatic input variable selection algorithms that have been proposed in the literature (e.g. Maier et al., 2010 and references therein), the IIS algorithm has been shown to be efficient in terms of high significance and low redundancy of the selected input variable set across a range of environmental applications (Fornarelli et al., 2013). The IIS algorithm was initially applied to the full candidate input variable dataset detailed in Table 1 (hereafter termed IIS). Separately, the IIS algorithm was also applied to an input variable dataset containing only those observed variables included in the SE models for ASPT or LIFE (hereafter termed SEM-IIS). Through these separate applications of the IIS algorithm, we sought to compare the effect of using explanatory models versus black box mathematical approaches for IVS on the performance of the final predictive models.

The IIS algorithm ranks each input variable by estimating its contribution, in terms of variance reduction, to the building of an underlying model of the output. The first p variables in the ranking are subsequently evaluated against the output by identifying p single-input-single-output models, with the best performing variable being added to the set of selected variables. The parameter p regulates the sensitivity of the algorithm in the presence of multiple redundant but informative inputs. It has been empirically demonstrated (Galelli and Castelletti, 2013a) that $p = 5$ ensures that informative inputs are not ignored in problems where the number of candidate input variables is of the same order as in Table 1. At the first iteration, the ranking algorithm is run on a data set composed of all the candidate input variables and the corresponding output values. For subsequent iterations, the original output values are replaced by the residuals of the underlying model identified in the previous step, ensuring that redundant inputs are not selected. The set of selected input variables is built incrementally by reiterating this procedure until the accuracy of the model built upon the selected set does not significantly improve with the addition of a further variable. Accuracy during this model-building phase using the IIS algorithm is assessed by calculating the average coefficient of determination (r^2) across the k validation folds of a k-fold cross-validation approach ($k = 5$ in our research). The two building blocks of the IIS algorithm are the ranking algorithm and the underlying model-building algorithm. Extremely Randomized Trees (Geurts et al., 2006) is used for both ranking and model-building purposes, because the particular structure and building algorithm of ET can be exploited to infer the relative importance of the different input variables (Fonteneau et al., 2008).

Extremely randomized trees is a recently developed tree-based method with strengths in terms of accurate and computationally-efficient modelling of strongly non-linear functions (e.g. Galelli and Castelletti, 2013b). The ET algorithm builds an ensemble of regression trees using randomization and averaging. A regression tree is a hierarchical cascade of rules able to predict numerical values of the output (Breiman et al., 1984). The tree-building process involves partitioning the input space into mutually exclusive regions according to a splitting criterion, progressively narrowing the size of the regions. The splitting criterion is defined by the identification of the best input to split a node and the corresponding splitting value. Eventually, when the number of instances in a region becomes smaller than a specific user-defined value (stopping condition), the partitioning of that region stops and a leaf is created. Whenever a new input is introduced into the tree, a specific path is followed according to the splitting rules adopted in the tree-building procedure, and the predicted output is then obtained by averaging of the values stored in the leaf. Extremely randomised trees differ from traditional regression trees in that the input and the cut-point to split a node (splitting value) are randomly selected. The effect of the randomization is then compensated for by creating an ensemble of M trees with the outcome of the ensemble being the average of the estimates produced by the M trees. Nodes are split using the following rule: K alternative cut-directions (inputs) are randomly selected and, for each one, a random cut-point is chosen; a score is then associated to each cut-direction and the cut-direction that maximises the score is adopted to split the node. The score in this case is proportional to the explained variance reduction of the output following the adopted splitting criterion, and can be re-formulated as r^2 (Geurts et al., 2006). When the number of instances within the node is smaller than a user-defined number ($N_{\min}$), the algorithm stops partitioning a node and a leaf is created.

The combined use of randomization and ensemble averaging provides more effective variance reduction than other randomized methods, while minimizing the bias of the final estimate (Geurts et al., 2006). Another important advantage of ET is that the values of the three hyper-parameters K, M, and $\mathrm{N}_{\min}$ does not require any optimal tuning, but can be simply fixed on the basis of empirical evaluations (Galelli and Castelletti, 2013b; Pianosi et al., 2013). In our simulations, we used K = the number of inputs, M = 200 and $N_{\min} = 5$. However, varying M between 100 and 500 and $N_{\min}$ between 2 and 15 did not alter model performance substantially, demonstrating robustness to values of these parameters within commonly adopted ranges (Galelli and Castelletti, 2013b). Moreover, the particular structure of ET can be exploited to rank the importance of the input variables in explaining the selected output behaviour (Wehenkel, 1998). Each input variable can be assigned a relevance score by estimating the variance reduction it can be associated with in the construction of the M different trees composing the ensemble (for further details see Galelli and Castelletti, 2013b).

2.4. *Evaluation of the final predictive models*

Evaluation of the final IIS and SEM-IIS models was performed by k-fold ($k = 5$ for our application) cross validation (Allen, 1974). The training data set was randomly split into k mutually exclusive subsets of equivalent size, and the IIS algorithm was run k times for both IIS and SEM-IIS models. Each time the underlying model was validated on one of the k folds and calibrated using the remaining $k - 1$ folds.

Metrics describing model prediction accuracy were calculated using observed and predicted data values from the k validation folds. The k-fold cross validation estimates the ability of a model to capture the behaviour of unseen or future observation data from the same underlying process, and, as such, it minimizes the risk of overfitting the data (Wan Jaafar et al., 2011 and references therein).

Our evaluation of the predictive performance of the IIS and SEM-IIS models draws on frameworks advocated by Bennett et al. (2013) and Willmott (1982). The aim of the evaluation was to assess whether IIS versus SEM-IIS approaches differed in terms of the predictive performance of the resulting models, rather than to identify a threshold of acceptable model performance against which both modelling approaches were assessed quasi-independently. We computed multiple performance metrics (see Bennett et al., 2013 for definitions of each metric), including summary statistics (mean, standard deviation) that enabled comparison of the entire modelled and observed datasets, alongside correlation measures (r^2 and coefficient of agreement, D) and difference measures (root mean squared error, RMSE, and mean absolute error, MAE) that enabled comparison of observed against predicted data values. Finally, we used the Bartlett test to evaluate differences in the predictive performance of IIS versus SEM-IIS models for each biotic index, based on the distributions of residuals from the final predictive models.

3. Results

3.1. Structural equation models

Fig. 3 reports final SE models for the ASPT and LIFE indices. The path between NH_4 and the response variable was not significant in either model ($p > 0.05$) and was removed without reducing the overall fit between the SE model and observed data. The path from Channel Hydromorphology to ASPT was also not significant ($p = 0.144$) and was removed, although this path retained a significant effect on the LIFE index and was included in the final SE model for this index.

The concentration of dissolved oxygen exerted the largest effect on both ASPT and LIFE indices, with compound path coefficients of 0.60 for APST and 0.55 for LIFE (compound path coefficients are given by the product of individual path coefficients linking variables, in this case $O_2 \leftarrow$ Effective Oxygen $\times$ Effective Oxygen $\rightarrow$ ASPT/LIFE). Path coefficients between the latent variable Organic Matter Load and both BOD_5 and PO_4 were positive, resulting in negative effects of these observed variables on ASPT and LIFE due to the negative path coefficient between Organic Matter Load and Effective Oxygen. The latent variable Channel Hydromorphology exerted an indirect effect on ASPT, mediated by the latent variable Effective Oxygen. Compound path coefficients indicated a negative effect of BSCI on ASPT (-0.30) and a positive effect of slope on ASPT (0.21), mediated by latent variables Channel Hydromorphology and Effective Oxygen. Similar indirect effects were exerted on LIFE by BSCI (-0.23) and slope (0.16). In addition, an indirect path (BSCI/slope ← Channel Hydromorphology → LIFE) connected the observed variables BSCI and slope to the LIFE index, with compound paths of 0.13 for slope and -0.18 for BSCI. In-channel vegetation structure index exerted a negative effect on ASPT, mediated by the latent variables Vegetation Structure, Channel Hydromorphology and Effective Oxygen (compound path coefficient $= -0.17$). A similar compound path connected CVSI with LIFE (path coefficient $= -0.13$), and an additional indirect path existed between CVSI and LIFE mediated by Vegetation Structure and Channel Hydromorphology (compound path coefficient $= -0.10$). The HMS variable exerted a direct negative effect on both ASPT (-0.16) and LIFE (-0.11). Moderate partial correlations existed between Vegetation Structure, HMS and PO_4 and between HMS and PO_4.

The final SE models for ASPT and for LIFE provided a strong fit to the observed data with $p \geq 0.05$ for models of both indices ($p = 0.128$ for ASPT and $p = 0.281$ for LIFE). Values of r^2 for ASPT (0.64) and LIFE (0.70) demonstrate that the SE models explained a substantial proportion of the among site variation in both indices within this dataset. This application of SEM represents a two-stage approach to IVS. Firstly, specification of a theoretical model on the basis of domain knowledge and available data (Fig. 2), and secondly statistical evaluation of the theoretical SE model against observed data (Fig. 3). Only measured variables included at both stages of this process were carried forward as input variables for subsequent predictive modelling as part of the SEM-IIS framework (see *Variables selected* in Fig. 3).

3.2. Predictive models derived from the IIS algorithm

The IIS algorithm was initially run using the full candidate input variable dataset detailed in Table 1 (termed IIS models) and, separately, using only those observed variables carried forward from the SE models reported in Fig. 3 (termed SEM-IIS models). Table 2 reports the variables included within the final IIS and SEM-IIS models for ASPT and LIFE in their order of selection, alongside r^2 determined during the model-building phase. Because the IIS algorithm is recursive, as each variable is incorporated within a model, the increase in cumulative r^2 and the individual contribution of each variable to the final model r^2 are reported. New variables are added to a predictive model during the model-building phase until inclusion of a further variable produces a decrease in r^2.

The IIS algorithm selected a reduced set of the 21 original candidate variables for inclusion in IIS models for the ASPT index (four variables selected) and the LIFE index (five variables selected). Within both models, O_2 was the most informative variable, contributing 62% and 78% to the ability of the final model to predict among site variation in ASPT and LIFE respectively. For LIFE, the physical habitat variables HMS and BSCI were also significant, together contributing 9% to the final model r^2. The concentration of ammonia was a significant predictor for both ASPT and LIFE, contributing 10% (ASPT) and 5% (LIFE) to the final model r^2. For both ASPT and LIFE, variables describing the geographical location of the monitoring site (East, North) or the height of source for the river were also selected by the IIS algorithm for inclusion within the final IIS models.

The SEM-IIS models for ASPT and LIFE were generated using the IIS algorithm, although starting from seven input variables selected from the original 21 candidate inputs through SEM (see *Variables selected* in Fig. 3). The final SEM-IIS models included six of the seven candidate input variables from the SE models, with PO_4 excluded from the ASPT model and CVSI excluded from the LIFE model by the IIS algorithm. The O_2 variable was again the most informative predictor for both ASPT and LIFE. However, there was no further correspondence between variables included within the SEM-IIS model compared to the IIS model for ASPT. For LIFE, only BSCI and HMS were selected for inclusion in both the SEM-IIS and IIS models, although in a reversed order of selection.

Metrics describing the performance of the final IIS and SEM-IIS models for both ASPT and LIFE are reported in Table 3. The mean value of both ASPT and LIFE predicted by either the IIS or SEM-IIS model was similar to that of the observed dataset ($\bar{y}_o{:}\bar{y}_p \approx 1$). However, both IIS and SEM-IIS models under-predicted the standard deviation observed in the ASPT and LIFE datasets ($\sigma_{o:}\ \sigma_p > 1$), due to over-prediction of lower values and under-prediction of higher values of both indices compared to observed data. In terms of r^2, the performance of the IIS and SEM-IIS models was directly comparable for the LIFE index. However, there was a modest reduction in r^2 for the SEM-IIS model compared to the IIS model for the ASPT index. Additional performance metrics (D, RMSE, MAE) indicated similar performance between IIS and SEM-IIS models for the LIFE index, alongside slightly reduced performance for the SEM-IIS model compared to the IIS model with respect to the APST index. The Bartlett test confirmed a significant difference between the residuals from IIS compared to SEM-IIS models for ASPT, but no

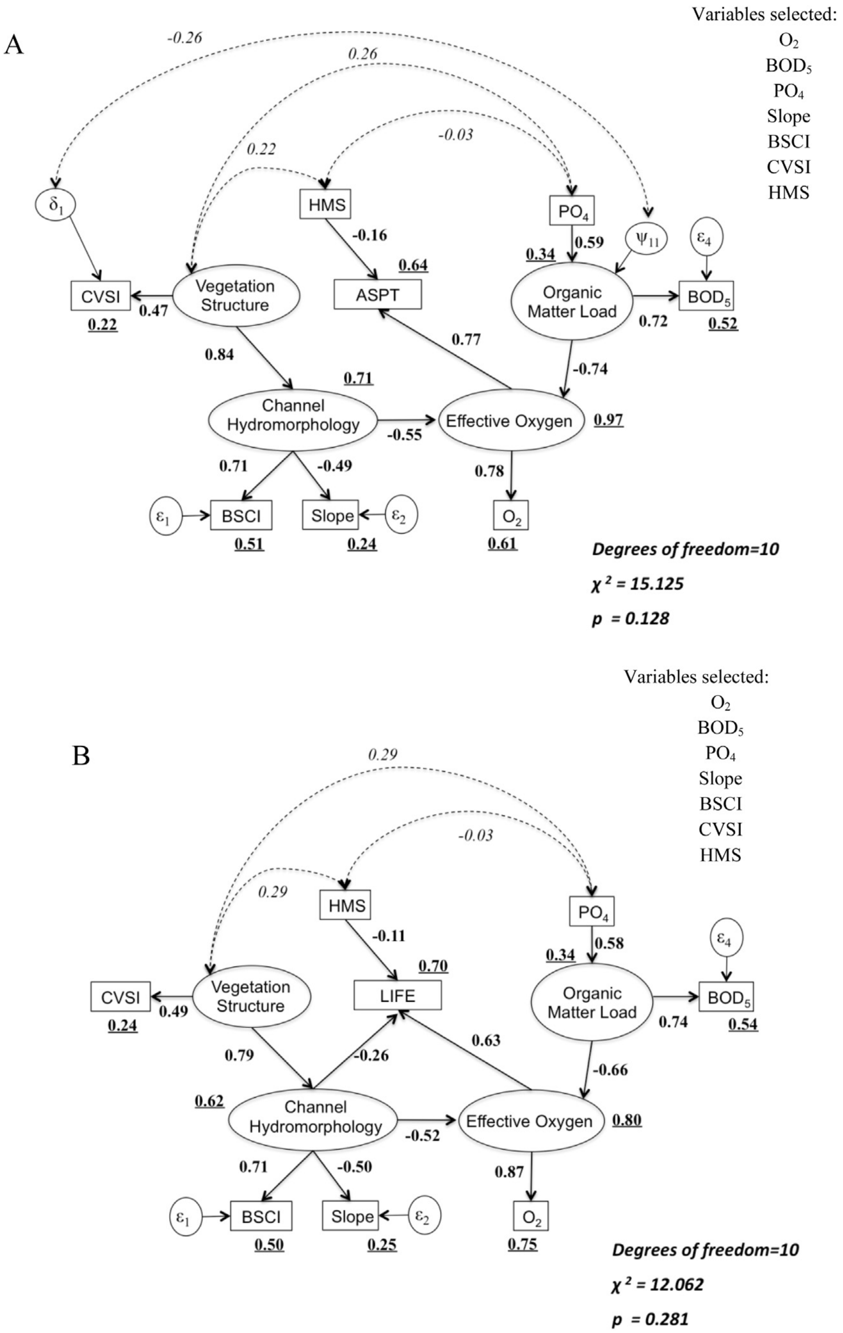

Fig. 3. Final structural equation models for the ASPT (A) and LIFE (B) index. Measured variables are represented by rectangles, latent variables by ellipses. For variable abbreviations see Table 1. The term δ_i represents measurement error for the indicators of the exogenous latent variables, ε_i represents measurement error for the indicators of endogenous latent variables, and ψ_i represents variance of the structural (latent) errors. Figures next to single-headed arrows are standardised path coefficients, next to double-headed dotted arrows are partial correlation coefficients and underlined are r^2 values. χ^2 corresponds to chi-square statistic and p-value to the probability level for rejecting the model based on the fit to observed data. Correlations between exogenous observed variables and the endogenous observed indicator variables for latents have been removed for simplicity (after Grace, 2006). Variable lists adjacent to each model identify measured variables carried forward from the SEM framework as input variables for subsequent SEM-IIS modelling.

significant difference between residuals from the two models for the LIFE index.

The predictive performance of the SEM-IIS models was explored further by sub-dividing the ASPT/LIFE variable space into lower quartile, upper quartile and interquartile ranges. Within each sub-area, the change in RMSE resulting from recursive inclusion of predictor variables within the SEM-IIS models was examined (Fig. 4). For predicted values of ASPT in the lower and upper quartiles, there was a monotonic improvement in RMSE with the addition of predictor variables (improvement of 15.4% in the lower quartile and 13.8% in the upper quartile for models containing all predictor variables compared to models containing only O_2). For values of ASPT in the interquartile range, the addition of predictor variables beyond O_2 resulted in a decrease in predictive performance, although all models exhibited relatively low RMSE consistent with $\bar{y}_o$:$\bar{y}_p$ ≈ 1. Similar patterns to those described above for ASPT were observed for predicted values of LIFE in the lower quartile, leading to a 19.7% improvement in RMSE compared to the

Table 2
Variables included in their order of selection, alongside evolution of r^2, during the predictive model-building phase using the IIS algorithm. For each individual variable, cumulative model r^2 and the percentage contribution of the variable to final r^2 are reported. IIS refers to predictive models built from the 21 original candidate input variables described in Table 1. SEM-IIS refers to predictive models built from the seven input variables selected following SEM (see *Variables selected*, Fig. 3). For variable abbreviations see Table 1. $N = 267$ based on 5-fold cross-validation.

Variable selected	Cumulative r^2	r^2 contribution [%] of individual variable	Variable selected	Cumulative r^2	r^2 contribution [%] of individual variable
ASPT					
IIS			*SEM-IIS*		
O_2	0.34	62	O_2	0.37	77
North	0.43	16	BOD_5	0.40	8
NH_4	0.48	10	BSCI	0.42	3
HoS	0.55	12	HMS	0.44	6
			Slope	0.46	3
			CVSI	0.47	3
LIFE					
IIS			*SEM-IIS*		
O_2	0.46	78	O_2	0.49	81
HMS	0.48	3	BSCI	0.53	7
BSCI	0.52	6	HMS	0.56	5
NH_4	0.54	5	Slope	0.57	2
East	0.60	9	BOD_5	0.58	1
			PO_4	0.60	3

model containing only O_2. For both the interquartile range and the upper quartile, the addition of BSCI, slope and HMS improved predictive performance for the LIFE model (8.9% decrease in RMSE for interquartile range, 11.0% decrease for upper quartile). However, inclusion of BOD_5 and PO_4 did not lead to further improvement in RMSE. Although RMSE for LIFE was low in the interquartile range, reflecting good predictive performance with respect to mean values of this index, RMSE remained relatively high in the upper quartile, consistent with results for the ASPT index.

4. Discussion

4.1. Explanatory modelling using the SEM framework

The SE models reported in Fig. 3 explained a substantial proportion of the among site variation in both ASPT and LIFE, consistent with strong performance of SE models specified for parallel datasets in our previous research (see Bizzi et al., 2013). This level of explanatory performance is comparable to that of ordination or regression models applied to similar benthic macroinvertebrate community indices (e.g. Feld and Hering, 2007; Lücke and Johnson, 2009). The SE models also meet or exceed the explanatory performance of ordination and regression techniques applied to spatial patterns in the taxonomic composition (e.g. Murphy and Davy-Bowker, 2005; Vaughan and Ormerod, 2012) or functional composition (e.g. Poff et al., 2010) of benthic macroinvertebrate communities. However, modelling spatial variation in indices compared to taxonomic descriptors of benthic macroinvertebrate communities can inflate model explanatory performance (e.g. Feld and Hering, 2007).

Direct and indirect gradient techniques have classically been applied in explanatory modelling of ecological datasets, yet ecological interpretation of the ordination axes resulting from these techniques remains a significant challenge (Graham, 2003; Vaughan and Ormerod, 2005). The SE models reported here match the explanatory power of these classical statistical techniques, whilst also providing an explicit and interpretable representation of the casual organisation of the underlying ecological system. Inferential errors can occur during the specification of SE models (Shipley, 2000), potentially leading to multiple models that differ in structure yet possess similar explanatory power. However, statistical evaluation of SE models will identify model structures that are fundamentally incompatible with the underlying organisation of ecological systems as expressed within available empirical data (e.g. Grace et al., 2010). Further, the consistency of the SE models reported in Fig. 3 with ecological theory (see below), alongside statistical support for the model structures, suggests that our SE models do provide a robust representation of the ecological processes underlying spatial patterns in ASPT and LIFE. Using the scoring indices ASPT and LIFE as response variables negates the challenges of non-linear taxonomic response to environmental gradients, including the unimodal relationships that are common in ecological data (e.g. ter Braak and Verdonschot, 1995). However, although not required for the ASPT or LIFE models reported in Section 3.1, variable transformation and the inclusion of polynomial regression structures enable SEM to address non-linear ecological relationships (e.g. Grace and Keeley, 2006; Wall and Amemiya, 2000). These extensions are likely to be required when extending SEM to explanatory modelling of taxonomic or functional trait descriptors of benthic macroinvertebrate communities.

Consistent with the theoretical basis to the SE models, the concentration of dissolved oxygen exerted a significant effect on both indices. The relationship between ASPT and dissolved oxygen concentration is underpinned by the sensitivity of BMWP scores to organic pollution and the oxygen demand associated with biodegradable organic matter (Armitage et al., 1983; Clarke et al., 2003). For the LIFE index, the importance of dissolved oxygen concentration is hypothesised to reflect the high oxygen demands of many taxa that display a habitat preference for fast flow conditions (Dunbar et al., 2010; Lorenz et al., 2004). Inverse relationships between PO_4 and ASPT/LIFE, and between BOD_5 and ASPT/LIFE, are consistent with nutrient-related controls on the autochthonous production of organic matter (e.g. Karl, 2000; Smith and Schindler, 2009), and the adverse effects of elevated biodegradable organic matter concentration on dissolved oxygen concentrations within streams and rivers, due to aerobic respiration by microbial communities. However, a positive bivariate correlation between BOD_5 and PO_4 ($\rho = 0.43$) indicates that a common allochthonous source for both pollutants, for example waste water effluent, cannot be excluded as an explanation for the relationship between these variables and ASPT or LIFE (e.g. Friberg et al., 2010). Previous application of SEM to analysis of spatial variation in the same benthic macroinvertebrate community indices (see Bizzi et al., 2013) supports a similar structure and size to the effects of the observed variables BOD_5, O_2, and PO_4 on the response variables

Table 3
Performance metrics for the final IIS and SEM-IIS models with respect to ASPT and LIFE indices. $\bar{y}_o$ = mean of the observed data, $\bar{y}_p$ = mean of the predicted data, σ_o = standard deviation of the observed data, σ_p = standard deviation of the predicted data, r^2 = coefficient of determination, D = coefficient of agreement, RMSE = root mean squared error; MAE = mean absolute error; p relates to the significance of the Bartlett test statistic, values <0.05 confirm a significant difference between residuals from IIS and SEM-IIS models at the 95% confidence level. $N = 267$ based on 5-fold cross-validation.

	ASPT		LIFE	
	IIS	SEM-IIS	IIS	SEM-IIS
$\bar{y}_o$:$\bar{y}_p$	1.00	1.00	1.00	1.00
σ_o: σ_p	1.36	1.47	1.29	1.30
r^2	0.59	0.50	0.61	0.61
D	0.90	0.87	0.90	0.90
RMSE	0.52	0.57	0.35	0.35
MAE	1.22	1.20	0.74	0.76
Bartlett's statistic	8.83		2.01	
p-value	0.01		0.35	

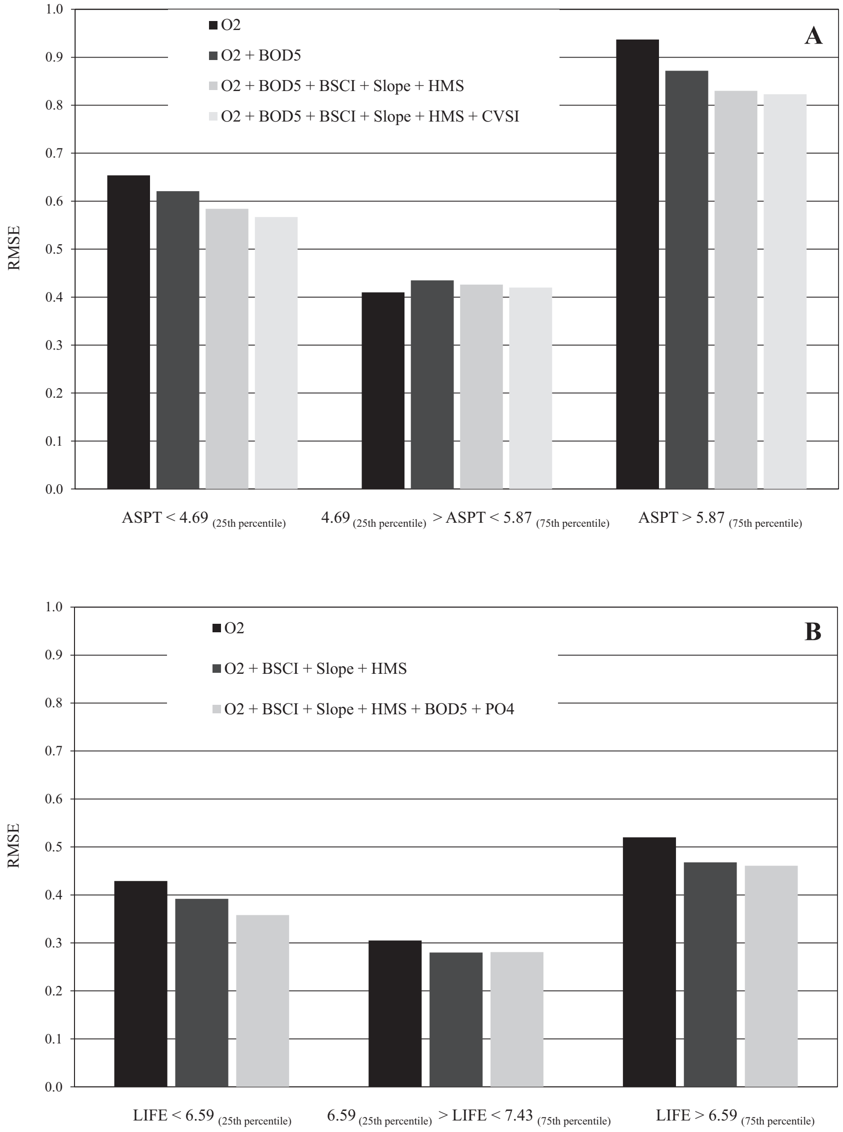

Fig. 4. Root mean square error (RMSE) for lower quartile, interquartile range and upper quartile values of ASPT (A) and LIFE (B) following addition of predictor variables to SEM-IIS models. For variable abbreviations see Table 1.

APST and LIFE, mediated by the latent variables Organic Matter Load and Effective Oxygen. A different, although related, dataset underpinned these earlier analyses, further validating these components of the SE models reported in Fig. 3.

Positive effects of slope and bed substrate size (BSCI is expressed in phi units which are inversely related to the physical size of sediment) on both ASPT and LIFE, mediated by Channel Hydromorphology and Effective Oxygen, are hypothesised to reflect reaeration of water during turbulent flow within steep-gradient channels and during flow through mesoscale bedform features, such as riffles, that are characterised by larger substrate size. The significant path from Channel Hydromorphology to LIFE suggests an additional relationship independent of that mediated by Effective Oxygen, likely reflecting the importance of mesoscale physical habitat structure as part of the hydromorphological template of lotic ecosystems (Maddock, 1999; Poff and Ward, 1990), and possibly the positive relationship between hydromorphological diversity and LIFE scores (e.g. Dunbar et al., 2010). Our SE models support a significant effect of in-channel vegetation structure on both ASPT and LIFE, mediated by the impact of in-channel vegetation growth and senescence on flow resistance, flow velocity and sediment transport (e.g. Asaeda et al., 2010; Gurnell et al., 2006). Negative effects of HMS on both LIFE and ASPT are consistent with the sensitivity of these indices to the direct, adverse effect of hydromorphological disturbance on benthic macroinvertebrate communities (e.g. Feld and Hering, 2007).

Variables defined at scales from micro-habitat to catchment can potentially contribute to explanation of spatial structure in benthic macroinvertebrate communities (Parsons et al., 2003). Across these spatial scales, explanatory variables are frequently correlated because of hierarchical process cascades that link rivers and their surrounding landscapes (e.g. Allan, 2004; Burcher et al., 2007; Frissell et al., 1986). Therefore, including variables describing spatial location in statistical models, whether geographical or longitudinal within a river network, can enhance explanation of spatial variation in benthic macroinvertebrate communities (e.g. Legendre and Legendre, 1998). However, spatial location often represents a surrogate for spatial structure in benthic macroinvertebrate communities that is driven by local responses to unmeasured physical habitat and hydrochemical conditions, biological interactions, or effects associated with historical legacy (Murphy and Davy-Bowker, 2005). Theoretically, these responses could be represented in explanatory models through incorporation of appropriate measured variables. The reliance on spatial variables often reflects constraints on understanding of how biota are causally related to meso-scale conditions and/or the lack of measured data through which causal relationships can be represented in explanatory models.

Spatial variables were not included in the SE models evaluated in our research because we hypothesised that spatial structure in ASPT and LIFE could be effectively explained by measured reach- or local-scale variables in the available dataset, consistent with the importance of meso-scale controls on benthic macroinvertebrate communities (e.g. Friberg et al., 2010). This hypothesis is supported by the concurrence between modelled and observed covariance matrices in our SE models, and by the substantial explanatory power of the SE models. However, this does not negate a role for spatial variables in explanatory modelling of benthic macroinvertebrate communities, particularly where taxonomic or functional response variables relate to biogeographical patterns in the evolution of taxa (Townsend et al., 2003) or to ecoregion-scale variation in unimpacted (Sandin and Johnson, 2000) or impacted (Sandin, 2003) communities. However, our application of SEM in the context of ASPT and LIFE emphasises the combination of explanatory power with an explicit representation of the causal organisation of an ecological system at meso-scale, rather than an attempt to build an explanatory model that maximises explanatory power at the expense of ecological interpretability.

4.2. *Predictive modelling of APST and LIFE using the iterative input variable selection algorithm*

Modelling of benthic macroinvertebrate communities has been predominantly explanatory to date, reflecting the importance within the discipline of ecology of understanding the variables and associated spatial or temporal scales that determine patterns in these communities (Begon et al., 1996). Apart from models designed to predict taxonomic composition at reference sites unimpacted by human pressures (e.g. Simpson and Norris, 2000; Wright et al., 2000), the development and evaluation of predictive models for benthic macroinvertebrate communities has been less common, particularly at sites along a gradient of environmental stress from minimally impacted sites to sites that are heavily impacted by physical and chemical disturbance.

Cross-fold validation of final IIS and SEM-IIS models for both ASPT and LIFE indicated that r^2 varied from 0.50 to 0.61 (Table 3). Similar predictive performance has been reported for tree-based models applied to other datasets from freshwater ecology, including in the prediction of lake diatom community composition (Kocev et al., 2010), planktonic chlorophyll-*a* concentration in reservoirs (Jung et al., 2010), and algal cell concentrations in coastal waters (Chen and Mynett, 2004). Compared to the reasonably high explanatory power of SE models, the weaker predictive performance of the final IIS and SEM-IIS models emphasises the need to avoid conflation of high explanatory and predictive power in the modelling of ecological data. Our IIS and SEM-IIS models demonstrated reasonable predictive performance for medium magnitude values of ASPT and LIFE ($\bar{y}_o$:$\bar{y}_p \approx 1$, Table 3), alongside relatively low RMSE in the interquartile range for SEM-IIS models of ASPT and LIFE (Fig. 4). However, performance as expressed through $\sigma_{o:}$ σ_p (Table 3), and through RMSE for SEM-IIS models at both higher and lower magnitude values of the biotic indices, indicated reduced predictive power. In particular, elevated values of RMSE were observed in the upper quartile of both ASPT and LIFE for SEM-IIS models, alongside little improvement in RMSE with inclusion of all predictor variables for LIFE in this quartile. These findings suggest that spatial patterns in ASPT and LIFE at sites that are less severely impacted by human disturbance (higher values of the biotic indices) are driven by ecological processes beyond those represented within our IIS and SEM-IIS models. Despite this, the monotonic improvement in RMSE for lower and upper quartile values of ASPT indicates that the SEM-IIS model was able to represent the effect of a range of chemical and physical habitat conditions on this benthic macroinvertebrate community index. The order of selection for input variables in the SEM-IIS model for ASPT was also consistent with a primary control exerted on this index by water quality parameters related to dissolved oxygen availability, followed by a secondary control exerted by physical habitat conditions. For the LIFE index, the order of variable selection in the SEM-IIS model was consistent with a primary control exerted by physical habitat conditions and a secondary control associated with water quality conditions. These observations are also consistent with the theoretical basis to both indices (Armitage et al., 1983; Extence et al., 1999).

Unmodelled deterministic behaviour within the IIS and SEM-IIS models may reflect the incomplete state of knowledge regarding controls on the ASPT and LIFE indices, particularly for SEM-IIS models in which expert knowledge was used as the basis for IVS. Alternatively, variables that are measured as part of additional monitoring networks but were not available within the candidate input dataset reported in Table 1 may have offered additional predictive power, for example by enabling the effect of controls on benthic macroinvertebrate communities associated with climate (e.g. Poff et al., 2010), discharge (Vaughan and Ormerod, 2012) or additional chemical parameters (e.g. Friberg et al., 2010) to be included within the predictive models. This may be particularly important for less heavily disturbed sites where benthic macroinvertebrate communities exhibit higher values of ASPT and LIFE, in which RMSE remained elevated despite inclusion of all informative variables in the SEM-IIS models (Fig. 4). Alternatively, the use of family-level indices such as ASPT or LIFE may have masked more sensitive, species-level responses to the environmental gradients that were included in the available empirical data (Vaughan and Ormerod, 2012). The use of species-level response variables may have produced more powerful predictive models, assuming species-level responses to the predictor variables in the available datasets were identified by the IIS algorithm.

4.3. *Comparison of predictive models built following expert-based and mathematically-based IVS*

Across a range of model evaluation criteria (Table 3), the predictive performance of IIS and SEM-IIS models was comparable with respect to the LIFE index. There was a modest reduction in predictive performance for the SEM-IIS model compared to the IIS model for ASPT, with a significant difference observed between the residuals from the two models for this biotic index. The SEM-IIS and IIS models were both based on a reduced set of the original candidate input variables, increasing the parsimony of the final predictive models. However, the IIS models included variables that were either excluded on a theoretical basis during the specification of the SE models (north, HoS for ASPT; east for LIFE), or during statistical evaluation of the SE models against empirical data (NH_4 for both ASPT and LIFE). Whilst the hierarchical organisation of riverine ecosystems provides some basis for ecological interpretation of variables such as north, east and HoS, other multicollinear datasets may lead to mathematical selection of input variables within modelling frameworks on the basis of spurious correlations without any link to ecological processes (Graham, 2003). Under these conditions, IVS based on domain knowledge through SEM provides a means of developing predictive models that remain ecologically-interpretable. However, in the specific application we report here, this comes at the expense of some reduction in model predictive performance for the ASPT index. This is a consequence of excluding variables (HoS and North) on a theoretical basis from the SE model that contributed to the predictive performance of the IIS model for ASPT. Although theoretically justified but statistically insignificant variables may be retained as part of explanatory modelling in some research fields (see for example Shmueli, 2010), we did not adopt this approach in our research and therefore NH_4 was removed from our SE models. The resulting fit of the SE models to empirical data, and the lack of a substantial decrease in predictive performance of SEM-IIS models compared to IIS models for the LIFE index, supports the removal of NH_4 from our models. However, retaining NH_4 within the SE model for ASPT may have improved the predictive performance of the resulting SEM-IIS model for this index, assuming that NH_4 was retained by the IIS algorithm in the final SEM-IIS model.

The net effect of the SEM-IIS framework is the potential to develop a predictive model of comparable parsimony and performance compared to a model generated through the IIS algorithm alone, at the same time as including only those input variables that are interpretable through meso-scale ecological processes. The SEM-IIS framework could help to address the lack of physical interpretability of some models derived from machine learning techniques (e.g. Kocev et al., 2010; Maier et al., 2010), without substantial sacrifices in model parsimony or predictive performance. Some previous research has explored both explanation and prediction of ecological data through tree-based modelling (e.g. Kocev et al., 2009). However, such research is based on changes in model parameters (e.g. pruning, use of ensembles of trees) within a modelling framework that is fundamentally predictive (regression trees), in an attempt to switch between a focus on either explanation or prediction. The SEM-IIS framework differs by coupling model components that are explicitly designed with either explanatory (SEM) or predictive (IIS) modes in mind, in order to model ecological systems. Our framework also incorporates components of both knowledge discovery (SEM) and operational use (IIS), offering the potential to implement knowledge-based scenario analyses related to the future condition of ecological systems. Understanding and predicting responses within riverine ecosystems to a range of drivers, including future climate change (e.g. Poff et al., 2010) or following stream restoration activities (e.g. Palmer et al., 2010), are pressing challenges in which modelling frameworks such as SEM-IIS could make significant contributions to both the development and application of knowledge.

5. Conclusions

Structural equation modelling represents a novel framework for explanatory modelling of ecological data, in which explicit links are made between causal theory and the analysis of empirical data. By incorporating constructs that are hypothesised to be of general, theoretical importance within ecological systems, SEM provides an appropriate modelling framework through which the existence of general ecological laws at community levels of organisation can be explored (Grace et al., 2010). The specific application of SEM reported here demonstrates that models with high explanatory power can be developed and evaluated against empirical data describing spatial patterns in benthic macroinvertebrate community indices within streams and rivers. Although based on robust theoretical justification of the causal paths within SE models, SEM remains correlative rather than causative. Significant paths within a SE model represent hypotheses regarding causal relationships within ecological systems that may be subject to further experimental validation (Iriondo et al., 2003), providing opportunities for feedback between explanatory modelling and empirical research.

The research reported here extends SEM from a purely explanatory modelling framework, using SE models as the basis for IVS through expert knowledge prior to predictive modelling. This represents an explicit coupling of explanatory and predictive modelling of ecological data, creating opportunities for feedback between causal explanation and empirical prediction. The predictive models developed here rely on the IIS algorithm, underpinned by ET, a robust model class for prediction given the complex, nonlinear and multicollinear nature of many ecological datasets. Through combining SEM with IIS, models can be built that maintain the predictive power and parsimony of those based on purely mathematical approaches to IVS, whilst at the same time addressing the lack of physical or ecological interpretability that can characterise predictive models built through machine learning alone. Therefore, the integrated SEM-IIS approach we propose offers the potential to develop environmental models that support both theory building and operational prediction within a single modelling framework.

Acknowledgements

The data for this research were released to the authors from the Environment Agency. We thank Prof. David Lerner for formative discussions regarding earlier versions of this paper. We acknowledge Dr Marc Naura and Mark Diamond and his team for useful discussions regarding the RHS database, and David Triggs from the Centre for Intelligent Environmental Systems at the University of Staffordshire for providing data on biological and chemical quality. The CatSci project (Early Stage Training in Catchment Science, funded by the European Commission, Marie Curie Actions Project No.: 21149) provided funding for SB to conduct research. We are grateful to the Associate Editor and three anonymous reviewers for their comments that helped to significantly improve the manuscript.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control AC 19, 716–723.

Allen, D., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125–127.

Allan, D.J., 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. Annu. Rev. Ecol. Evol. Syst. 35, 257–284.

Arhonditsis, G.B., Stow, C.A., Steinberg, L.J., Kenney, M.A., Lathrop, R.C., Mcbride, S.J., et al., 2006. Exploring ecological patterns with structural equation modeling and Bayesian analysis. Ecol. Model 192, 385–409.

Arhonditsis, G.B., Stow, C.A., Paerl, H.W., Valdes-Weaver, L.M., Steinberg, L.J., Reckhow, K.H., 2007. Delineation of the role of nutrient dynamics and hydrologic forcing on phytoplankton patterns along a freshwater-marine continuum. Ecol. Model 208, 230–246.

Armitage, P.D., Moss, D., Wright, J.F., Furse, M.T., 1983. The performance of a new water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Wat. Res. 17, 333–347.

Asaeda, T., Rajapakse, L., Kanoh, M., 2010. Fine sediment retention as affected by annual shoot collapse: sparganium erectum as an ecosystem engineer in a lowland stream. River Res. App. 26, 1153–1169.

Begon, M., Harper, J.L., Townsend, C.R., 1996. Ecology: Individuals, Population and Communities. Blackwell Science, Oxford.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Modell. Softw. 40, 1–20.

Berenzen, N., Schulz, R., Liess, M., 2001. Effects of chronic ammonium and nitrate contamination on the macroinvertebrate community in running water microcosms. Wat. Res. 35, 3478–3482.

Bernot, M.J., Sobota, D.J., Hall, R.O., Mulholland, J.P., Dodds, W.K., Webster, J.R., Tank, J.L., Ashkenas, L.R., Cooper, L.W., Dahm, C.N., Gregory, S.V., Grimm, N.B., Hamilton, S.K., Johnson, S.L., McDowell, W.H., Meyer, J.L., Peterson, B., Poole, G.C., Valett, H.M., Arango, C., Beaulieu, J.J., Burgin, A.J., Crenshaw, C., Helton, A.M., Johnson, L., Merriam, J., Niederlehner, B.R., O'Brien, J.M., Potter, J.D., Sheibley, R.W., Thomas, S.M., Wilson, K., 2010. Inter-regional comparison of land-use effects on stream metabolism. Freshw. Biol. 55, 1874–1890.

Beven, K.J., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36.

Biggs, B.J.F., 1996. Patterns in benthic algae of streams. In: Stevenson, R.J., Bothwell, M.L., Lowe, R.L. (Eds.), Algal Ecology: Freshwater Benthic Ecosystems. Academic Press, San Diego, California, pp. 31–56.

Biggs, B.J.F., 2000. Eutrophication of streams and rivers: dissolved nutrient-chlorophyll relationships for benthic algae. J. N. Am. Benthol. Soc. 19, 17–31.

Bizzi, S., Surridge, B.W.J., Lerner, D.N., 2013. Structural equation modelling: a novel statistical framework for exploring the spatial distribution of benthic macroinvertebrates in riverine ecosystems. River Res. App. 29, 743–759.

Bollen, K.A., 1989. Structural Equations with Latent Variables. John Wiley & Sons, New York.

Breiman, L., Friedman, J., Olsen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International.

Buffagni, A., Erba, S., Cazzola, M., Kemp, J.L., 2004. The AQEM multimetric system for the southern Italian Appennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. Hydrobiologia 516, 313–329.

Burcher, C.L., Valett, H.M., Benfield, E.F., 2007. The land-cover cascade: relationships coupling land and water. Ecology 88, 228–242.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2010. Tree-based reinforcement learning for optimal water reservoir operation. Water Resour. Res. 46, W09507. http://dx.doi.org/10.1029/2009WR008898.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2012. Data-driven dynamic emulation modelling for the optimal management of environmental systems. Environ. Modell. Softw. 34, 30–43.

Chen, Q., Mynett, A.E., 2004. Predicting *Phaeocystis globosa* bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression. Ecol. Model. 176, 277–290.

Clarke, R.T., Wright, J.F., Furse, M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecol. Model 160, 219–233.

Crisci, C., Ghattas, B., Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecol. Model. 240, 113–122.

De'ath, G.A., 2007. Boosted trees for ecological modelling and prediction. Ecology 88, 243–251.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol. Model. 160, 291–300.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecol. Model. 195, 20–29.

Dunbar, M.J., Pedersen, L.M., Cadman, D., Extence, C.A., Waddingham, J., Chadd, R.P., Larsen, S.E., 2010. River discharge and local-scale physical habitat influence macroinvertebrate LIFE scores. Freshw. Biol. 55, 226–242.

Emery, J.C., Gurnell, A.M., Clifford, N.J., Petts, G.E., 2004. Characteristics and controls of gravel-bed riffles: an analysis of data from the river habitat survey. Water Environ. J. 18, 210–216.

Extence, C.A., Blabu, D.M., Chadd, R.P., 1999. River flow indexing using British benthic macroinvertebrates: a framework for setting hydroecological objectives. River Res. App. 15, 543–574.

Feio, M.J., Viana-Ferreira, C., Costa, C., 2013. Combining multiple machine learning algorithms to predict taxa under reference conditions for streams bioassessment. River Res. App.. http://dx.doi.org/10.1002/rra.2707.

Feld, C.K., Hering, D., 2007. Community structure or function: effects of environmental stress on benthic macroinvertebrates at different spatial scales. Freshw. Biol. 52, 1380–1399.

Fonteneau, R., Wehenkel, L., Ernst, D., 2008. Variable selection for dynamic treatment regimes: a reinforcement learning approach. In: Proceedings of the 8th European Workshop on Reinforcement Learning, 2008.

Fornarelli, R., Galelli, S., Castelletti, A., Antenucci, J.P., Marti, C.L., 2013. An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by inter-basin water transfers. Water Resour. Res. 49, 3626–3641. http://dx.doi.org/10.1002/wrcr.20268.

Friberg, N., Skriver, J., Larsen, S.E., Pedersen, M.L., Buffagni, A., 2010. Stream macroinvertebrate occurrence along gradients in organic pollution and eutrophication. Freshw. Biol. 55, 1405–1419.

Frissell, C.A., Liss, W.L., Warren, C.E., Hurley, M.D., 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. Environ. Manage. 10, 199–214.

Galelli, S., Castelletti, A., 2013a. Tree-based iterative input variable selection for hydrological modelling. Water Resour. Res.. http://dx.doi.org/10.1002/wrcr.20339.

Galelli, S., Castelletti, A., 2013b. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrol. Earth Syst. Sci. 17, 1–6.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomised trees. Mach. Learn. 63, 3–42.

Geurts, P., Irrthum, A., Wehenkel, L., 2009. Supervised learning with decision tree-based methods in computational and systems biology. Mol. Biosyst. 5, 1593–1605.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat. Ecol. 41, 491–508.

Golubkov, S.M., Tiunova, T.M., Kocharina, S.L., 1992. Dependence of the respiration rate of aquatic insects upon the oxygen concentration in running and still water. Aquat. Insects 14, 137–144.

Grace, J.B., 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press.

Grace, J.B., Keeley, J.E., 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecol. Appl. 16, 503–514.

Grace, J.B., Pugesek, B.H., 1997. A structural equation model of plant species richness and its application to a coastal wetland. Am. Nat. 149, 436–460.

Grace, J.B., Anderson, M.T., Olff, H., Scheiner, S.M., 2010. On the specification of structural equation models for ecological systems. Ecol. Monog. 80, 67–87.

Graham, M.H., 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84, 2809–2815.

Gurnell, A.M., Van Oosterhout, M.P., De Vlieger, B., Goodson, J.M., 2006. Reach-scale interactions between aquatic plants and physical habitat: river Frome. Dorset. River Res. App. 22, 667–680.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hägglund, G., 2001. Milestones in the history of factor analysis. In: Cudeck, R., Du Toit, S.H.C., Sörbom, D. (Eds.), Structural Equation Modelling: Present and Future. Scientific Software International, Licolnwood Illinois, pp. 11–38.

Hoang, T.H., Lock, K., Mouton, A., Goethals, P.L.M., 2010. Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. Ecol. Inf. 5, 140–146.

Iriondo, J.M., Albert, M.J., Escudero, A., 2003. Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. Biol. Conserv. 113, 367–377.

Jakeman, A., Hornberger, G., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res. 29, 2637–2649.

Jowett, I., 2003. Hydraulic constraints on habitat suitability for benthic invertebrates in gravel-bed rivers. River Res. App. 19, 495–507.

Jung, N.C., Popescu, I., Kelderman, P., Solomatine, D.P., Price, R.K., 2010. Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. J. Hydroinf. 12, 262–274.

Karl, D.M., 2000. Phosphorus, the staff of life. Nature 406, 31–33.

Kelly, M., Juggins, S., Guthrie, S., Pritchard, R., Jamieson, S., Ripley, B., Hirst, H., Yallop, M., 2008. Assessment of ecological status in UK rivers using diatoms. Freshw. Biol. 53, 403–422.

Kocev, D., Džeroski, S., White, M.D., Newell, G.R., Griffioen, P., 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecol. Model 220, 1159–1168.

Kocev, D., Naumoski, A., Mitreski, K., Krstić, A., Džeroski, S., 2010. Learning habitat models for the diatom community in Lake Prespa. Ecol. Model 221, 330–337.

Kokes, J., Zahràdkova, S., Nemejcova, D., Hodovsky, J., Jarkovsky, J., Soldan, T., 2006. The PERLA system in the Czech Republic: a multivariate approach for assessing the ecological status of running waters. Hydrobiologia 566, 343–354.

Lawton, J.H., 1999. Are there general laws in ecology? Oikos 84, 177–192.

Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam.

Lorenz, A.W., Hering, D., Feld, C.K., Rolauffs, P., 2004. A new method for assessing the impact of hydromorphological degradation on the macroinvertebrate fauna of five German stream types. Hydrobiologia 516, 107–127.

Lücke, J.D., Johnson, R.K., 2009. Detection of ecological change in stream macroinvertebrate assemblages using single metric, multimetric or multivariate approaches. Ecol. Indic. 9, 659–669.

Maddock, I., 1999. The importance of physical habitat assessment for evaluating river health. Freshw. Biol. 41, 373–391.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Modell. Softw. 25, 891–909.

Monk, W.A., Wood, P.J., Hannah, D.M., Wilson, D.A., 2007. Selection of river flow indices for the assessment of hydroecological change. River Res. App. 23, 113–122.

Mouton, A.M., De Baets, B., Goethals, P.L.M., 2009. Knowledge-based versus data-driven fuzzy habitat suitability models for river management. Environ. Modell. Soft. 24, 982–993.

Mouton, A.M., Dedecker, A.P., Lek, S., Goethals, P.L.M., 2010. Selecting variables for habitat suitability of Asellus (Crustacea, Isopoda) by applying input variable contribution methods to artificial neural network models. Environ. Model. Assess. 15, 65–79.

Mulholland, J.P., Fellows, C.S., Tank, J.L., Grimm, N.B., Webster, J.R., Hamilton, S.K., Martí, E., Ashkenas, L., Bowden, W.B., Dodds, W.K., McDowell, W.H., Paul, M.J., Peterson, B.J., 2009. Inter-biome comparison of factors controlling stream metabolism. Freshw. Biol. 46, 1503–1517.

Murphy, J.F., Davy-Bowker, J., 2005. Spatial structure in lotic macroinvertebrate communities in England and Wales. Relationships with physical,chemical and anthropogenic stress variables. Hydrobiologia 534, 151–164.

Murray-Bligh, J., 1999. Procedures for Collecting and Analysing Macroinvertebrate Samples. Environment Agency, Bristol.

Newson, M.D., Newson, C.L., 2000. Geomoprhology, ecology and river channel habitat: mesoscale approaches to basin-scale challenges. Prog. Phys. Geog. 24, 195–217.

Palmer, M., Menninger, H.L., Bernhardt, E., 2010. River restoration, habitat heterogeneity and biodiversity: a failure of theory or practice? Freshw. Biol. 55, 205–222.

Parsons, M., Thoms, M.C., Norris, R.H., 2003. Scales of macroinvertebrate distribution in relation to the hierarchical organization of river systems. J. N. Am. Benthol. Soc. 22, 105–122.

Pianosi, F., Castelletti, A., Restelli, M., 2013. Tree-based fitted Q-iteration for multi-objective Markov decision processes in water resource management. J. Hydroinf. 15, 258–270.

Poff, N.L., Pyne, M.I., Bledzoe, B.P., Cuhaciyan, C.C., Carlisle, D.M., 2010. Developing linkages between species traits and multiscaled environmental variation to explore vulnerability of stream benthic communities to climate change. J. N. Am. Benthol. Soc. 29, 1441–1458.

Poff, N.L., Ward, J.V., 1990. Physical habitat template of lotic systems: recovery in the context of historical pattern of spatiotemporal heterogeneity. Environ. Manage. 14, 629–645.

Raven, P.J., Holmes, N.T.H., Dawson, F.H., Everard, M., 1998. Quality assessment using river habitat survey data. Aquat. Conserv. 8, 477–499.

Reckhow, K.H., Arhonditsis, G.B., Kenney, M.A., Hauser, L., Tribo, J., Wu, C., Elcock, K.J., Steinberg, L.J., Stow, C.A., McBride, S.J., 2005. A predictive approach to nutrient criteria. Environ. Sci. Technol. 39, 2913–2919.

Sandin, L., 2003. Benthic macroinvertebrates in Swedish streams: community structure, taxon richness, and environmental relations. Ecography 26, 269–282.

Sandin, L., Hering, D., 2004. Comparing macroinvertebrate indices to detect organic pollution across Europe: a contribution to the EC Water Framework Directive intercalibration. Hydrobiologia 516, 55–68.

Sandin, L., Johnson, R.K., 2000. Ecoregions and benthic macroinvertebrate assemblages of Swedish streams. J. N. Am. Benthol. Soc. 19, 462–474.

Shan, Y., Paull, D., McKay, R.I., 2006. Machine learning of poorly predictable ecological data. Ecol. Model 195, 129–138.

Shipley, B., 2000. Cause and Correlation in Biology. Cambridge University Press, Cambridge, UK.

Simpson, J.C., Norris, R.H., 2000. Biological assessment of river quality: development of AusRivAS models and outputs. In: Wright, J.F., Sutcliffe, J.W., Furse, M.T. (Eds.), Assessing the Biological Quality of Fresh Waters. Freshwater Biological Association, Ambleside, UK, pp. 125–142.

Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310.

Smith, V.H., Schindler, D.W., 2009. Eutrophication science: where do we go from here? Trends Ecol. Evol. 24, 201–207.

ter Braak, C.J.E., Verdonschot, P.M.E., 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat. Sci. 57, 255–289.

Tirelli, T., Pozzi, L., Pessani, D., 2009. Use of different approaches to model presence/absence of Salmo marmoratus in Piedmont (Northwestern Italy). Ecol. Inf. 4, 234–242.

Tirelli, T., Pessani, D., 2011. Importance of feature selection in decision-tree and artificial-neural-network ecological applications. *Alburnus alburnus* alborella: a practical example. Ecol. Inf. 6, 309–315.

Townsend, C.R., Doledec, S., Norris, R.H., Peacock, K., Arbuckle, C., 2003. The influence of scale and geography on relationships between stream community composition and landscape variables: description and prediction. Freshw. Biol. 48, 768–785.

Vaughan, I.P., Ormerod, S.J., 2005. The continuing challenges of testing species distribution and models. J. Appl. Ecol. 42, 720–730.

Vaughan, I.P., Ormerod, S.J., 2010. Linking ecological and hydromorphological data: approaches, challenges and future prospects for riverine science. Aquat. Conserv. 20, S125–S130.

Vaughan, I.P., Ormerod, S.J., 2012. Large-scale, long-term trends in British river macroinvertebrates. Glob. Change. Biol. 18, 2184–2194.

Wall, M.M., Amemiya, Y., 2000. Estimation for polynomial structural equation models. J. Am. Stat. Assoc. 95, 925–940.

Wan Jaafar, W.Z., Liu, J., Han, D., 2011. Input variable selection for median flood regionalization. Water Resour. Res. 47. http://dx.doi.org/10.1029/2011WR010436.

Wehenkel, L., 1998. Automatic Learning Techniques in Power Systems. Kluwer Academic Publishers, Dordrecht.

Weiher, E., Forbes, S., Schauwecker, T., Grace, J.B., 2004. Multivariate control of plant species richness and community biomass in blackland prairie. Oikos 106, 151–157.

Wiens, J.A., 2002. Riverine landscapes: taking landscape ecology into the water. Freshw. Biol. 47, 501–515.

Williams, K.A., Green, D.W.J., Pascoe, D., 1986. Studies on the acute toxicity of pollutants to freshwater macroinvertebrates. 3. Ammon. Arch. Hydrobiol. 106, 61–70.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. B. Am. Meteorol. Soc. 63, 1309–1313.

Winklemann, C., Hellmann, C., Worischka, S., Petzoldt, T., Benndorf, J., 2011. Fish predation affects the structure of a benthic community. Freshw. Biol. 56, 1030–1046.

Wright, J.F., 2000. An introduction to RIVPACS. In: Wright, J.F., Sutcliffe, J.W., Furse, M.T. (Eds.), Assessing the Biological Quality of Fresh Waters. Freshwater Biological Association, Ambleside, UK, pp. 125–142.

Wright, J.F., Sutcliffe, J.W., Furse, M.T., 2000. Assessing the Biological Quality of Fresh Waters. Freshwater Biological Association, Ambleside, UK.

Young, P.C., 2013. Hypothetico-inductive data-based mechanistic modeling of hydrological systems. Water Resour. Res. 49, 915–935.