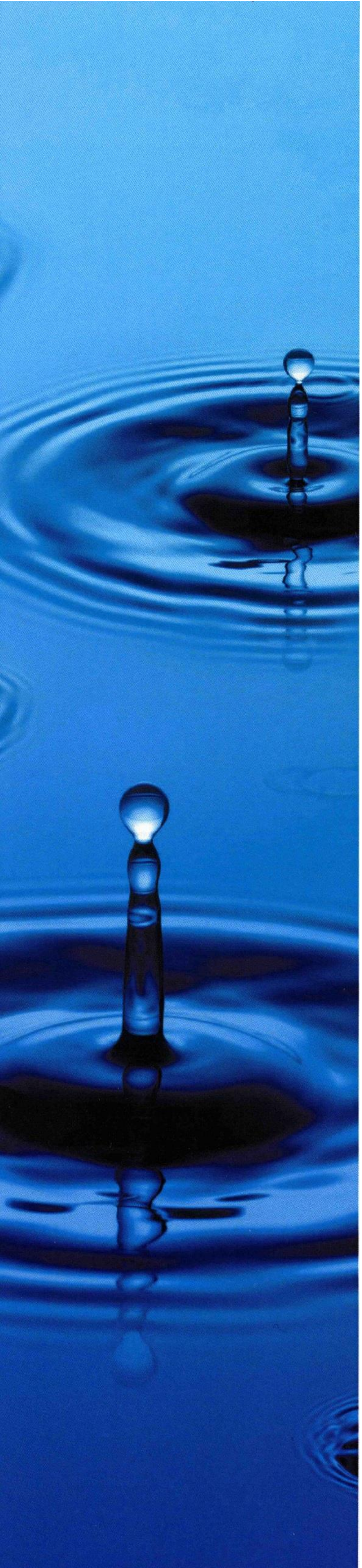

Freshwater Biological Association

GIS Re-calibration of the hydromorphology-independent RIVPACS predictive model (M37): new model M44

A Report to the
Scottish Environment Protection Agency

Final 24July2018

R.T. Clarke
J. Davy-Bowker

July 2018

Research Contractor

This document was produced by the Freshwater Biological Association[†]:
Using sub-contractors: Ralph T. Clarke and John Davy-Bowker

[†]The Freshwater Biological Association, River Laboratory, East Stoke, Wareham, Dorset,
BH20 6BB, United Kingdom.

Project Funders

This project was funded by the Scottish Environment Protection Agency (SEPA).

Disclaimer

Whilst this document is considered to represent the best available scientific information and expert opinion available at the stage of completion of the report, it does not necessarily represent the final or policy positions of the project funders or contractors.

Dissemination status

Unrestricted

Scottish Environment Protection Agency Project Manager

Scottish Environment Protection Agency's project manager for this contract was:
Ian Milne (SEPA)

FBA Project Manager

FBA's project manager for this contract was:
John Davy-Bowker

FBA Project Code

S/0025/R

The Freshwater Biological Association

The Freshwater Biological Association
The Ferry Landing
Far Sawrey, Ambleside
Cumbria, LA22 0LP, United Kingdom

The Freshwater Biological Association
River Laboratory
East Stoke, Wareham
Dorset, BH20 6BB, United Kingdom

Web site: www.fba.org.uk
Email: info@fba.org.uk

Registered Charity No. 214440
Company Limited by Guarantee No. 263162, England
UKPRN No. 10018314

EXECUTIVE SUMMARY

GIS Re-calibration of the hydromorphology-independent RIVPACS predictive model (Model M37): New model M44

Project funder: Scottish Environment Protection Agency (SEPA)

Background to research

The Regulatory Agencies in the UK (the Environment Agency; Scottish Environment Protection Agency; and the Northern Ireland Environment Agency) now use the River Invertebrate Classification Tool (RICT) to classify the ecological quality of rivers for Water Framework Directive compliance monitoring. RICT incorporates RIVPACS IV predictive models.

While RICT currently classifies waters for general degradation and organic pollution stress, producing assessments of status class and uncertainty, WFD compliance monitoring also requires the UK Agencies to assess the impacts of a wide range of pressures including hydromorphological and acidification stresses. Some of these pressures alter the predictor variables that current RIVPACS models use to derive predicted biotic indices.

A previous SNIFFER project WFD119 (Clarke *et al.,* 2011) developed and assessed a range of new RIVPACS models that do not use predictor variables that are affected by these stressors, but instead use alternative GIS based variables that are wholly independent of these pressures. The recommended best of these new models involved geological and physical features of the upstream catchment area of each river site and were Model 24 (hydromorphology independent), Model 35 (alkalinity independent) and Model 13 (hydromorphology and alkalinity independent) – see Clarke *et al.,* (2011) for further details.

Models 13, 24 and 35 all involved the variable PROPWET, which estimates the proportion of time upstream catchment soils are wet, based on CEH Flood Estimation Handbook data. Due to potential IPR and licensing issues with the estimation of the variable PROPWET for reference and other stream sites, the Environment Agency contracted Clarke and Davy-Bowker (2017a) to develop and assess the relative effectiveness of a new predictive model (Model 37) which involves the same predictive variables as Model 24 except for the exclusion of the variable PROPWET. Clarke and Davy-Bowker (2017a) found that new model (model M37) gives almost the same accuracy of predictions as the previous best model 24 found by Clarke *et al.,* (2011) in terms of predicting biotic fauna and the WHPT and LIFE biotic indices without using flow-dependent predictor variables.

Objectives of research

- Use values from the new CEH-derived RICT replacement variables database for the RICT reference sites to recalibrate the replacement variables **Model 37** (without the variable PROPWET
- Report on model performance
- Deliver the recalibrated model with files for incorporation into the RICT software, a test dataset, and updated RIVPACS database

Summary of findings

- Existing model M1 (the current RICT software default), model M2 (model M1 but excluding stream width, depth and substratum composition) and model M37 were re-fitted and compared to the equivalent new models M41, M42 and M43 for which the site's distance from source, altitude, slope and discharge category were now taken from their GIS values within the new CEH-GIS-RICT Database. Using the GIS values of these four variables rather than the original manually-measured map variables gave about the same overall precision of predicted expected (E) values in terms of standard deviation of O/E values and squared correlation n between the observed (O) and E values amongst the 685 GB reference sites.
- Three new models M44, M45 and M46, were fitted which used the newer BGS version 5 geological classes GB map rather than the BGS version 4 geology classes used in model M37 and M43. All three models used the new GIS values of the site distance from source, altitude, slope and discharge category and upstream catchment area and mean altitude, but model M44 used major geology classes which tried to reproduce those used in Clarke et al (2011) and model M37, while models M45 and M46 used a slightly different set of major geology classes which allowed some detailed version 5 classes to be classed as new mixed-type classes (e.g. shale/limestone) in additions to the previous major classes.
- The new RIVPACS predictive models which involved the geological classification involving mixed geology classes (models 45 and 46) did not give any improvement over model 44, the latter being equivalent to model M37 but with all possible variables' values taken from the new CEH-GIS-RICT database, as described above.
- Our recommendation is to use new model M44 for use in making predictions of expected index values at stream sites which might already be subject to hydro-morphological stress. Model M44 is based on the CEH-GIS-RICT Database values of both the original time-invariant RIVPACS variables (distance from source, site altitude and slope, discharge category), upstream catchment area and mean altitude, together with the upstream catchment percentage cover of each of 'peat', 'clay', 'chalk', 'limestone' and 'hard rocks' based on BGS version v5 equivalent major geological classes (akin to those used in previous model development projects WFD119 (Clarke et al., 2011) and model M37(Clarke and Davy-Bowker, 2017a)).
- New model M44 is the first to base RIVPACS-model predictions of expected fauna and expected biotic index values on the new CEH-GIS-RICT database of GIS-based stream site and upstream catchment environment predictor variable. It will enable RIVPACS-type predictions of expected values to be made automatically, without any site visit for almost any river site in Britain.
- New model M44 is the best model available to make predictions for sites potentially subject to hydromorphological stress. However, it may over-predict expected values and thus under-estimate O/E values for some deep river sites dominated by fine sediment substratum.

This report also includes:

- Discriminant Functions file for new Model M44 for use in updated RICT software
- Test Input Dataset of 12 sites and corresponding Test Dataset Outputs from new model M44, including (i) Probabilities of End-Group Membership (ii) Expected values for each of a range of indices for spring, summer and autumn samples

An updated RIVPACS Database which includes the new CEH-GIS-RICT Database values for the new environmental predictor variables involved in model M44 will be added to the FBA website

Table of Contents

1. BACKGROUND AND SPECIFIC AIMS

The River InVertebrate Prediction and Classification System (RIVPACS) model, now incorporated into the River Invertebrate Classification Tool (RICT), allows prediction of the river invertebrate fauna expected under conditions of minimal anthropogenic pressure. Calculation of resulting scores for biological metrics allows comparison with observed results, and this forms the basis of river-invertebrate-based classification for Water Framework Directive (WFD) purposes.

RIVPACS models use a set of predictive variables to predict reference values for biotic indices and predicted faunal lists at test sites. These predictor variables can be divided into two groups: time variant (different on each sampling occasion) and time-invariant predictors (measured from maps). For example:

Time Invariant	Time variant (measured at site)
Altitude	Substrate composition
Slope	Width
Discharge category	Depth
Distance from source	
Latitude	Alkalinity
Longitude	
Recent historical Air Temperature	

Time invariant predictors are generally derived from maps or GIS layers, and represent gradients such as altitude, distance from source, or mean air temperature. These can be regarded as not being affected by any of the stressors that need to be assessed. Time variant predictors are recorded at the time a test site is sampled (width, depth, and substrate composition), or, in the case of alkalinity, over a recent period of time. Time variant variables are more prone to being altered by stressors. For example, sedimentation, abstraction, hydromorphological alteration and acidification can all affect one or more of the time variant variables. This can have consequences for predictions, because the values of one or more time-variant predictor variables measured in the field for a site may already have had its ‘natural’ value (or range of values) altered by one or more of the stresses who biological effects it is hoped to assess through the calculation of RIVPACS O/E ratios. Using altered predictor variable values might lead to incorrect and/or inappropriate RIVPACS model predictions of the expected fauna and expected (E) values of biotic indices.

Examples can be imagined for fine sediment stress affecting substrate composition, and thereby causing predictions of biotic index reference values to be distorted towards end groups that naturally have finer substrata. Similarly, acidification stress may cause biotic index predictions to be distorted towards end groups that have naturally lower alkalinities.

The problem of stressors affecting RIVPACS variables has been less of an issue in the past, when most water pollution problems arose from organic pollution (since this stressor does not affect any predictor variables); moreover the previous BMWP and newer WHPT indices were originally intended to primarily detect organic stress. However, as more and more stress types now need to be assessed, and some of these are physical in nature (or alkalinity related), there has been a growing need to examine the issue of stressors affecting the RIVPACS predictor variables.

To get round this problem alternative variables are needed that are not affected by stress. For the RIVPACS variables this means removing the time variant variables: substrate, width and depth, all of which are affected by physical modifications to test sites. It may also be necessary to remove alkalinity as a predictor variable because its measured values at test sites may be modified by acidification (and potentially sewage and industrial discharges that can add excess base thereby increasing alkalinity. The other variables are regarded as being robust with respect to stressors.

To get round this issue, in SNIFFER project WFD119, Clarke *et al.,* (2011) investigated a wide range of new models which excluded the either (i) flow/sediment related predictor variables (stream width, depth and substrate composition) (ii) acidity related variables (alkalinity) or (iii) both sets of time-variant variables. In addition, a whole new range of potential RIVPACS predictor models were considered by including a replacement set of time-invariant pressure-insensitive variables derived from geographical Information systems (GIS), many involving measures of site and upstream catchment physical and geological characteristics.

In their Executive Summary, Clarke *et al*., (2011) recommended the following predictive models for assessing watercourses affected by flow/hydromorphological and/or acidity stress:

- For flow/hydromorphological stressors that may have modified width, depth and/or substrate in GB, it is suggested that a new '**RIVPACS IV – Hydromorphology Independent'** model (Model 24) is used (this does not use the predictor variables width, depth and substratum, but includes a suite of new stressor-independent variables).
- For acidity related stressors in GB, it is suggested that a new '**RIVPACS IV – Alkalinity Independent'** model (Model 35) is used (this does not use the predictor variable alkalinity, but includes new stressor-independent variables).
- For flow/hydromorphological stressors *and* acidity related stressors in GB, it is suggested that a new '**RIVPACS IV – Hydromorphology & Alkalinity Independent'** model (Model 13) is used (this does not use the predictor variables width, depth, substratum and alkalinity, but includes a suite of new stressor-independent variables)."

The best 'Hydromorphology Independent' model (Model 24) excluded stream width, depth and substrate composition from the original RIVPACS default model (Model 1), but included GIS-based measures of the upstream catchment area, average upstream altitude and percentage cover of specific drift and solid geological types in the upstream catchment of a site.

However, Model 24 also involved a new variable called PROPWET, which is an estimated measure of the proportion of time the upstream catchment soils are wet and is derived from the CEH Flood Estimation Handbook (FEH). As there may be intellectual property rights restrictions with deriving estimates of PROPWET for any site and any RIVPACS/RICT software user (i.e. outside of the UK environment agencies), a new model (Model 37) was developed which involved all of the variables in best model 24 except PROPWET.

A difficulty for RICT users is that the remaining replacement GIS-based variables (upstream catchment area, mean upstream catchment altitude, percent cover of key geological features of the catchment) are not readily obtainable. To address this issue, SEPA commissioned the Centre for Ecology and Hydrology (CEH) to re-calculate these variables, together with the existing predictor variables distance to source, altitude, slope and discharge category, at a 50m resolution across the UK rivers network and assemble the results into a geo-referenced database (Kral *et al.* (2017).

During the database development work of Kral et al. (2017), a few apparent errors in the geology cover data came to light. These arise from errors during the aggregation of geological sub-categories into the higher-level categories used for the geology variables. As a result, and in preparation for this current project, the geology variable data were re-examined to re-check the aggregation process and identify and correct these, and any other, errors. Additionally, the British Geological Survey source dataset has been updated (from version 4 to 5) since the original pressure-independent model development work of Clarke *et al*. (2011).

The variables data in the new database described above will be used as input data for predictions by the new RIVPACS model. Although the data matches the original RIVPACS reference site data well, there are discrepancies. The new data is seen as being more accurate as it was derived from more up to date GIS models, and updated geology source data. There is now a need to recalibrate the recent flow-stress-independent RIVPACS Model 37 using data from the new database, which will hopefully improve the model's accuracy and align it with data that will be used for input to make RICT predictions and assessments for other (non-reference) stream sites. This is the purpose of this project.

Aim of the current work:

- To derive and assess new models, akin to model 37 which derives predictions that are independent of flow-related stress, but which use the new CEH GIS-derived database of the RIVPACS predictor variables involved in Model 37, namely:

  At the Site :
  Distance from Source, Altitude, Slope and Discharge category

  Upstream catchment:
  Upstream catchment Area and Upstream catchment Mean Altitude
  %surface geology cover of peat
  %cover of each of the following major solid geology classes:
  Clay, Chalk, Limestone and Hard Rocks (and optional mixed classes)

- Report on model performance compared to previous models

- Deliver the best new GIS-based model with files for incorporation into the RICT software

- Provide a test dataset with new model predictions for future software validation and an updated RIVPACS database

2. PROVISION OF CEH DATABASE OF GIS-BASED PREDICTOR VARIABLES

CEH (with BGS) were recently commissioned by SEPA to develop a GIS-based version of the environmental variables for use in RIVPACS predictions of site-specific expected taxonomic composition and expected values of biotic indices (Kral *et al*. 2017). The Kral *et al*. (2017) report gives details of all the various methods and algorithms used to automate the capture and estimation of each of the existing and new potential RIVPACS environmental predictor variables used within this current RIVPACS/RICT model-building project.

This data is referred to here as the 'CEH-GIS-RICT database'. The new environmental data values for the 685 RIVPACS IV GB reference sites were provided to us by Cedric Laize (CEH) in January 2018.

2.1 Unaltered original RIVPACS predictor variables

The original RIVPACS Reference sites' values for the following variables were retained, as they were not re-estimated within the CEH-GIS-RICT database:

● air temperature mean ● air temperature range

● latitude ● longitude ● alkalinity

2.2 Re-estimated original RIVPACS predictor variables

The CEH-GIS-RICT database provided revised estimates of the following original RIVPACS predictor variables:

- Distance from source (km) ● Altitude (m) ● Slope (m/km)
- (Mean) Discharge category (1-10)

2.3 Estimation of upstream catchment variables

The CEH-GIS-RICT database also provided estimates of the upstream catchment of each RIVPACS reference site, from which the following variables were derived for each site:

- (Log_{10})Upstream catchment area (km^2)
- (log_{10}) Upstream catchment mean altitude (m)
- Upstream catchment percentage cover of key geological types (see section below)

2.4 Major Geology classes (Option 1) and mixed geology classes (Option 2)

Kral *et al.* (2017) used the latest version 5 of the BGS geological classification, which contained more and some different sub-divisions of the UK geology than the previous version 4 used to form the major solid geology classes that were eventually used to derive the RIVPACS predictive models in SNIFFER project WFD119 (Clarke *et al*. 2011) and the later model M37 (model M24 without the variable PROPWET) in Clarke and Davy-Bowker (2017a).

2.5 Estimation of missing slope values for a few problem sites

The only unsolvable problems with using the CEH GIS algorithms to derive estimates for the required new RIVPACS predictor variables were for site slope for 12 of the 685 GB Reference sites. For these 12 sites, we used their original RIVPACS values for slope that were manually-derived from printed OS maps, details are given in Table 1.

Table 1 Original RIVPACS slope estimates used for 12 Reference sites with no CEH GIS values for slope at site

RICT_ID	Original RIVPACS Slope value (m/km)	Reason for missing slope value (-9) from CEH GIS
1011	0.9	Last cell at end of river (no upstream cells to calculate slope from)
1409	0.7	Site in flat area; US and DS cells have same elevation so 50m elevation difference threshold does not operate, therefore the maximum 10 cell-move should take over but yields same elevations; -9 might be understood as 0 in this case
1411	0.7	Site in flat area
4311	2.0	Site in flat area
5509	0.3	Last cell at end of river (no upstream cells to calculate slope from)
7145	5.5	Site in quite flat area but problem looks more like a flow routing issue (complex local network)
7417	4.8	Last cell at end of river (no upstream cells to calculate slope from)
NH07	33.3	Site is on a lake (as far as the modelled river network is concerned) so it is similar to being on flat area
SEPA_N03	16.7	Last cell at end of river (no upstream cells to calculate slope from)
SEPA_N05	16.2	Last cell at end of river (no upstream cells to calculate slope from)
SEPA_N08	0.1	Last cell at end of river (no upstream cells to calculate slope from)
SEPA_N28	13.0	Last cell at end of river (no upstream cells to calculate slope from)

3. METHODS TO COMPARE EFFECTIVENESS OF PREDICTIVE MODELS

The RIVPACS bioassessment system is based on comparing the ratio (O/E) of the observed (O) values of biotic indices to the site- and season-specific expected (E) values of the indices. The expected values are based on a statistical predictive model of the relationship between the macroinvertebrate sample composition of a set of reference sites and their environmental characteristics.

RIVPACS model development involves three main stages:

- (i) Biological classification (using TWINSPAN) of the reference sites into end-groups based on their (three-season combined) sample macroinvertebrate composition
- (ii) Multivariate discrimination of the end-groups based on a suite of environmental predictor variables
- (iii) Deriving site-specific expected (E) values of biotic indices from end-group means of observed values of indices for the reference sites, weighting end-groups by discriminant-based probabilities of the site belonging to each end-group
- (iv) Assessing model effectiveness by comparing the strength of the relationship between O and E values amongst the reference sites.

3.1 Percentage of reference sites correctly allocated to biological end-group

The most common method of measuring statistical discrimination success as a whole is to calculate the percentage of sites discriminated to their correct group, here their TWINSPAN end-group. This can be calculated using either (i) the re-substitution method (ReSub) whereby all sites are used to fit the model and test it or (ii) the cross-validation or leave-one-out method (XVal) for which the fit to each site in turn is based on the model fitted to all other sites. The percentage correct statistics (ReSub and XVal) have been used to select environmental predictor variables in all previous developments of RIVPACS (Moss *et al*., 1999, Clarke *et al*., 2003). Their advantage is that they generate overall measures of fit which are independent of any biological index. However, the RIVPACS predictive models do not allocate sites to the most probable group but calculate expected index values using the probabilities of a site belonging to each end-group.

3.2 Percentage variation explained (R^2_{OE}) and SD(O/E)

The ultimate aim is to assess site ecological status using O/E ratios. A major aim of the modelling is therefore for the predicted expected (E) values of biotic indices to agree with the observed values for the reference sites as closely as possible. The level of agreement can be measured by the statistic R^2 (denoted R^2_{OE} here) measuring the percentage of the total variation in observed (O) index values amongst the reference sites explained by the site-specific expected (E) values; the higher the better. An alternative measure of model effectiveness for any particular index is the standard deviation of the O/E ratios for the reference sites (denoted SD(O/E)); the lower the better.

The aim of the modelling is for the O/E values amongst the reference sites for any particular index to vary as little as possible (i.e. have low SD) about the overall average value of approximately one. There should then be more opportunity and statistical power to detect departures from reference condition with low O/E values resulting from the impact of anthropogenic and other stresses.

The statistic SD(O/E) has been used in this study to measure and compare the effectiveness of the various trial discrimination models and recommend the best to carry forward.

However, the various trial models are also summarised and compared by their R^2_{OE} values as a check.

3.3 Comparison with Null Model SD(O/E)

If there was no predictive model for the expected values, or none of the trial models had any real discriminatory power, there would be no reliable information to set different “target” expected E values for an index for the different types of site. In such cases it would only be possible to use the average of the observed values of an index across all reference sites as the single ‘target’ expected E value for all sites. This is termed a ‘Null Model’ because there are no predictor variables involved. It is akin to a regression model with no explanatory X variables and just an intercept term (which is then estimated as the overall average of the dependent Y variable).

The SD(O/E) for the Null Model, termed SD_0(O/E), is simply the SD of the O values for all of the reference sites divided by their mean value (which is equivalent to the coefficient of variation (CV =SD/Mean) of the observed index values for the reference sites (Van Sickle *et al*., 2005). The effectiveness of any predictive model for any one index can be compared both to other models and to the Null Model by comparing their SD(O/E) for the same biotic index. The lower the value, the better the non-null model is at predicting observed values for the reference sites, and thus the site-specific ‘target’ expected (E) for other sites.

It is important to understand that some biotic indices are inherently more variable in orders of magnitude than others, partly because of how the indices were invented and defined. This is represented by their CV amongst reference sites which, as mentioned above, is equal to the SD(O/E) for a Null Model. Therefore although the observed (O) values of each biotic index are ‘standardised’ by dividing by the site-specific expected E values to give O/E with an average value across all reference sites of around one, in practice, the O/E values are inherently more variable for some indices than others. For example, the average of the single season Null Model SD_0(O/E) for WHPT number of families (WHPT NTAXA) is much higher (i.e. 0.281) than that for abundance-weighted WHPT ASPT (i.e. 0.159). However, as mentioned above, for a given index, any reduction in SD(O/E) obtained by using a different predictive model indicates an improvement in overall predictive ability for that index.

3.4 Summary of effectiveness measures

In comparing the effectiveness of different trial models, the main emphasis should be on comparing the SD(O/E) values and R^2_{OE} values separately within indices. In particular, it is useful to compare the SD(O/E) with the Null Model SD_0(O/E) and also the R^2_{OE} with a null model R^2 value of zero.

4. EFFECTIVENESS OF NEW GIS-BASED RIVPACS MODELS

4.1 Effectiveness of previous models, including model M37

As a means of introduction, and for completeness, the summary of the relative effectiveness of model M37, developed by Clarke and Davy-Bowker (2017a) is repeated here.

The environmental predictor variables involved in each of the four RIVPACS models that we previously compared are given in Table 2:

- Model 1 is the original RIVPACS IV (and RICT software) default predictive model
- Model 2 is based on the original suite of variables but excludes width, depth and substrate composition
- Model 24 also excludes stream width, depth and substrate composition but involves measures of upstream catchment area, average altitude, drift and solid geological types and PROPWET
- Model 37 is the same as Model 24 but excludes PROPWET

Table 2 Environmental predictor variables used (X) in RIVPACS IV models 1, 2, 24 and new model 37.

Variable name	Variable description	Derived from	Model			
			1	2	24	37
LAT	Latitude	RIVPACS	X	X	X	X
LONG	Longitude	RIVPACS	X	X	X	X
ATEMPMEAN	Mean Air Temp	RIVPACS	X	X	X	X
ATEMPRANGE	Air Temp Range	RIVPACS	X	X	X	X
DISCHARGE	Discharge Category (historical long-term average)	RIVPACS	X	X	X	X
LOGDFS	Log$_{10}$ Distance From Source	RIVPACS	X	X	X	X
LOGALT	Log$_{10}$ Altitude	RIVPACS	X	X	X	X
LOGSLOPE	Log$_{10}$ Slope (at site)	RIVPACS	X	X	X	X
ALK	Alkalinity	RIVPACS	X	X	X	X
LOGALK	Log$_{10}$ Alkalinity	RIVPACS	X	X	X	X
LOGWIDTH	Log$_{10}$ Water Width	RIVPACS	X			
LOGDEPTH	Log$_{10}$ Water Depth	RIVPACS	X			
MSUBST	Mean Substratum (phi units)	RIVPACS	X			
%DRIFT1-PEAT	%Drift Geology Class 1 – Peat in upstream catchment	IRN+BGS			X	X
%SOLID3-CLAY	%Solid Geology Class 3 – Clay in upstream catchment	IRN+BGS			X	X
%SOLID6-CHALK	%Solid Geology Class 6 - Chalk	IRN+BGS			X	X
%SOLID7-LIMESTONE	%Solid Geology Class 7 - Limestone	IRN+BGS			X	X
%SOLID8-HARDROCKS	%Solid Geology Class 8 - Hard Rocks	IRN+BGS			X	X
LOGAREA	Log$_{10}$ Upstream catchment Area (from DTMGEN)	DTMGEN			X	X
LOGALTBAR	Log$_{10}$(ALTBAR) Upstream catchment mean Altitude	FEH			X	X
PROPWET	Proportion of time upstream catchment soils are wet	FEH			X	

The relative effectiveness of these four models for each the WHPT and LIFE$_{(fam)}$ indices of immediate interest here are summarised in Table 3.

Table 3 Summary of Model 1 (default RIVPACS/RICT predictor model) with three models which exclude (stream width, depth and substrate composition: Model 2 (no new variables), Model 24 (both adding GIS-based geological cover variables and upstream catchment LOGAREA, LOGALTBAR and PROPWET variables and new Model 37 (same as Model 24 except excludes PROPWET); based on average single season SD(O/E) and R^2_{OE} , together with the discrimination % correctly allocated (ReSub and XVal) to biological end-group.

	Null	Model			
Model	Model	1	2	24	37
RIVPACS IV default minus -->	All	None	-Flow	-Flow	-Flow
+ new GIS variables -->				Y	Y
+ PROPWET variable				Y	N
%Correct (ReSub)	6.3	51.7	47.3	49.9	49.8
%Correct (XVal)	6.3	38.7	36.4	36.8	37.2
		SD(O/E) (lower is better)			
TL1 BMWP NTAXA	0.268	0.200	0.203	0.196	0.197
TL1 BMWP ASPT	0.120	0.076	0.079	0.079	0.079
TL2 WHPT NTAXA	0.281	0.206	0.208	0.202	0.203
TL2 WHPT ASPT	0.159	0.088	0.092	0.092	0.092
TL2 $LIFE_{(fam)}$ (DistFam)	0.085	0.053	0.055	0.055	0.055
TL4 $LIFE_{(sp)}$	0.099	0.057	0.061	0.061	0.061
		R^2_{OE} (higher is better)			
TL1 NTAXA	0.0	44.3	43.1	46.9	46.2
TL1 ASPT	0.0	61.8	58.8	59.2	59.3
TL2 WHPT NTAXA	0.0	45.7	44.7	48.2	47.7
TL2 WHPT ASPT	0.0	71.8	68.7	68.6	68.5
TL2 $LIFE_{(fam)}$ (DistFam)	0.0	63.1	58.9	59.8	59.7
TL4 $LIFE_{(sp)}$	0.0	68.8	65.0	64.8	64.8

Summary:
The model 37 (which excludes the use of the variable PROPWET in the predictions) gives almost the same accuracy of predictions as the previous best model 24 found by Clarke *et al.,* (2011) for predicting biotic fauna and biotic indices without using flow-dependent predictor variables (Table 3).

4.2 Explanation of new models M41-M46 involving CEH-GIS-RICT Database values of predictor variables

Six new models, denoted M41-M46, were assessed (Table 4).

All existing models (e.g. model M1, M2 and M37) and new models M41-M46) involved the same original RIVPACS Reference site values for the following environmental predictor variables:

Latitude
Longitude
Mean Air Temperature
Air Temperature Range
Alkalinity (and Log_{10} Alkalinity)

Table 4 Environmental predictor variables used in RIVPACS IV models M41-M46 and their source (R = Original manual RIVPACS value, 119 = geological values used in WFD119 report, C = CEH-GIS-RICT database value, G1 and G2 = CEH-GIS-RICT Database values for upstream geology based on geological super-classes options G1 and G2 respectively.

Variable name	Variable description	Model					
		M41	M42	M43	M44	M45	M46
LAT	Latitude	R	R	R	R	R	R
LONG	Longitude	R	R	R	R	R	R
ATEMPMEAN	Mean Air Temp	R	R	R	R	R	R
ATEMPRANGE	Air Temp Range	R	R	R	R	R	R
ALK	Alkalinity	R	R	R	R	R	R
LOGALK	Log_{10} Alkalinity	R	R	R	R	R	R
DISCHARGE	Discharge Category	C	C	C	C	C	C
LOGDFS	Log_{10} Distance From Source	C	C	C	C	C	C
LOGALT	Log_{10} Altitude	C	C	C	C	C	C
LOGSLOPE	Log_{10} Slope (at site)	C	C	C	C	C	C
LOGAREA	Log_{10} Upstream catchment Area (from DTMGEN)	C	C	C	C	C	C
LOGALTBAR	Log_{10}(ALTBAR) Upstream catchment mean Altitude	C	C	C	C	C	C
%PEAT	%Peat in upstream catchment (Drift class 1)			119	G1	G2	G2
%CLAY	% Clay in upstream catchment (Solid class 3)			119	G1	G2	G2
%CHALK	%Chalk in upstream catchment (Solid class 6)			119	G1	G2	G2
%LIMESTONE	%Limestone in upstream catchment (Solid class 7)			119	G1	G2	G2
%HARDROCKS	%Hard Rocks in upstream catchment (Solid class 8)			119	G1	G2	G2
%LISHASAND	%Limestone, shale and sandstone mix (Solid class 9)						G2
%SHALI	%shale and limestone mix (Solid class 11)						G2

However, all the new models assessed here (models M41-M46) involved the newly-derived CEH-GIS-RICT Database replacement values (Kral et al 2017) for the following RIVPACS time-invariant predictor variables that were previously derived manually from OS maps and elsewhere:

Log_{10} Distance from Source
Log_{10} Altitude
Log_{10} Slope
Discharge category (log-term average 1961-90)

Model M41, M42 and M43 are the same as existing models M1, M2 and M37 respectively (summarised in Table 2 and Table 3), except that the values for the above four original manually-derived map variables (Distance from source, Altitude, Slope and Discharge category) are replaced by their GIS values from the CEH-GIS-RICT Database for each of the

685 GB reference sites. These three models are included to assess whether using the semi-automated GIS-derived values of these original RIVPACS variables makes much difference to the overall predictive fit of the models in terms of SD(O/E) and R^2_{OE}.

Models M44, M45 and M46 are the same as model M43, but use the new CEH_GIS-RICT Database values for the upstream geology variables, as detailed in Table 4.

Geology super classes option G1 attempted to force all BGS version v5 detailed geological classes in the same geological super-classes (1-8) as those used by both Clarke et al (2011) in their original development of model M24 and then more recently by Clarke and Davy-Bowker (2017a) in their development of model M37.

Geological super-classes option G2 allowed for some of the BGS version v5 detailed geological classes to be considered as of mixed geology in terms of the super classes. Thus in option G2, detailed geological classes were assigned to either one of the original super classes (1-8) or to new mixed super-classes (9-12):

Mixed class 9:	lishasand’ = mix of limestone, shale and sand
Mixed class 10:	‘shasand’ = mix of shale and sand
Mixed class 11:	‘shali’ = mix of shale and limestone
Mixed class 12:	‘unconssand’ = unconsolidated

Model M44 uses geological option G1 to provide new values for the original upstream geology super-classes used previously in models M24 and M37. Model M45 and M46 use geological option G2 to provide values for the previous super-classes and model M46 also includes two extra variables representing the upstream catchment percentage cover of the two main mixed geology classes 9 and 11 (Table 4).

4.3 Effectiveness of new models M41-M46

The relative overall effectiveness of each new model M41-M46 is assessed in terms of their average single-season standard deviation in O/E values (SD(O/E) and squared correlation (R^2_{OE}) between observed (O) and predicted expected (E) index values for both the standard BMWP and WHPT indices in RICT, but crucially also for the LIFE, PSI and E-PSI indices which were designed to respond to differences or changes in hydromorpological conditions and stress (Table 5). The main purpose of these new model developments is to derive predictions of expected fauna and expected values of biotic indices which are not dependent on the levels of flow- and sediment- related stress at the site at the time of biological sampling.

Table 5 Summary of new GIS-based RIVPACS-RICT predictive models M41-M46 relative to the default RIVPACS/RICT model M1, model M2 (as per model M1 but excluding stream width, depth and substrate composition) and model M37; based on average single season SD(O/E) and R^2_{OE}, together with the discrimination % correctly allocated (ReSub and XVal) to biological end-group (see section 4.2 for explanation of upstream geology options).

	Model								
	M1	M2	M37	M41	M42	M43	M44	M45	M46
GIS site variables				Y	Y	Y	Y	Y	Y
Upstream Geology			119			119	G1	G2	G2 +mix
%Correct (ReSub)	51.7	47.3	49.8	49.1	44.5	48.9	49.2	48.6	48.9
%Correct (XVal)	38.7	36.4	37.2	38.0	34.7	35.5	36.8	35.9	35.9
				SD(O/E) (lower is better)					
TL1 NTAXA	0.200	0.203	0.197	0.202	0.205	0.200	0.199	0.198	0.198
TL1 ASPT	0.076	0.079	0.079	0.077	0.080	0.079	0.078	0.078	0.078
TL2 WHPT NTAXA	0.206	0.208	0.203	0.208	0.211	0.205	0.204	0.204	0.204
TL2 WHPT ASPT	0.088	0.092	0.092	0.089	0.094	0.093	0.092	0.092	0.093
TL2 LIFE(DistFam)	0.053	0.055	0.055	0.053	0.056	0.055	0.055	0.055	0.056
TL4 LIFE(Sp)	0.057	0.061	0.061	0.058	0.062	0.061	0.061	0.061	0.061
TL3 PSI(Fam)	0.213	0.209	0.214	0.216	0.215	0.212	0.206	0.206	0.207
TL4 PSI(Sp)	0.271	0.256	0.289	0.272	0.256	0.253	0.248	0.248	0.249
TL3 E-PSI(fam69)	0.173	0.178	0.183	0.174	0.180	0.179	0.176	0.176	0.177
TL4 E-PSI	0.199	0.199	0.224	0.196	0.196	0.192	0.189	0.189	0.190
TL5 E-PSI	0.200	0.198	0.224	0.197	0.196	0.192	0.189	0.189	0.190
				R^2_{OE} (higher is better)					
TL1 NTAXA	44.3	43.1	46.2	43.6	42.0	45.0	45.2	45.2	45.3
TL1 ASPT	61.8	58.8	59.3	60.9	58.0	58.8	59.9	60.0	59.4
TL2 WHPT NTAXA	45.7	44.7	47.7	45.0	43.8	46.6	46.5	46.6	46.7
TL2 WHPT ASPT	71.8	68.7	68.5	70.7	67.5	67.8	68.3	68.5	67.9
TL2 LIFE(DistFam)	63.1	58.9	59.7	62.2	57.7	58.8	58.5	58.6	58.1
TL4 LIFE(Sp)	68.8	65.0	64.8	68.0	63.4	64.2	64.3	64.4	63.9
TL3 PSI(Fam)	70.1	66.3	66.4	69.2	64.5	65.2	65.9	66.1	65.5
TL4 PSI(Sp)	74.8	70.7	70.8	73.8	69.1	69.7	70.0	70.3	69.7
TL3 E-PSI(fam69)	77.2	71.0	70.9	76.9	70.3	70.6	70.5	70.8	70.2
TL4 E-PSI	78.0	71.1	69.7	77.6	70.0	69.8	70.2	70.4	69.9
TL5 E-PSI	77.4	70.6	69.2	76.9	69.5	69.4	69.7	69.9	69.4

As found previously, leaving out the on-site measured variables of stream width, depth and estimated substratum composition from the model predictions (i.e. model M2 versus model M1) tends to lead to higher SD(O/E) and lower R^2_{OE} for nearly all indices. This is not surprising as the observed conditions at a river site at the time of sampling are expected to have influence on the biota currently present. However, we are looking for the best model 'fit for purpose', which is to predict what macroinvertebrate biota should be at any site in the absence of any stress and alteration to the hydromorpological conditions at a site.

Previous models M1, M2 and M37 are not generally improved by replacing the manually-derived time-invariant map variables with their semi-automated CEH-GIS-RICT Database equivalents to give corresponding models M41, M42 and M43 respectively; but this was not the primary aim of the task. In general both SD(O/E) and R^2_{OE} are about the same for both methods of obtaining these variables, although SD(O/E) tends to be lower for model M43 than model M37 for the PSI and E-PSI indices (Table 5).

Also, model M43 is better than model M42, indicating that involving the percentage cover of major geological classes in the upstream catchment improves overall predictions of biological end group and biotic indices (as was found previously by Clarke et al (2011) when the original map-based variables were still based on their original RIVPACS manually-measured values).

Models M44-M46 only differ in the form of upstream catchment geology variables that they use. Models M45 and M46 both use the geological classification option denoted G2, which allows for some detailed BGS version v5 classes to be classed as mixtures of two or more major geological types; model M46 which adds in extra variables based on the percentage cover of two mixtures of the major types does not give any improvement in fit over model M45.

In a comparison of models M44 and M45, based on the major geological types under geological grouping option G1 and G2, model M44 appears to be slightly better based on percentage of reference sites allocated to correct biological end-group, while model M45 is slightly better based on the R-squared between observed and predicted expected values for many biotic indices. Overall, we recommend using model M44 which is a direct replacement for model M37 using the new CEH-GIS-RICT Database values for the RIVPACS time-invariant predictor variables and for the upstream catchment major geology classes.

Summary:

Our recommendation is to use new model M44. Model M44 is based on the CEH-GIS-RICT Database values of both the original time-invariant RIVPACS variables (distance from source, site altitude and slope, discharge category), upstream catchment area and mean altitude, together with the upstream catchment percentage cover of each of 'peat', 'clay', 'chalk', 'limestone' and 'hard rocks' based on geological major categories option G1 (which attempted to assign the detailed BGS version v5 geological classes to the same major classes as used in previous model development projects WFD119 (Clarke et al., 2011) and model M37(Clarke and Davy-Bowker, 2017a).

4.4 Summary and discussion of new model M44

4.4.1 Relative accuracy of model M44

Although new model M44 is not as precise as the current RIVPACS/RICT default predictor variable model M1 in terms of either percentage of reference sites allocated to the correct biological end-group, the SD of the reference sites O/E values for each index or the squared correlation between observed and predicted expected index values, model M44 is still an effective predictive model in that it is much better than the null model with no predictors, and more importantly is an improvement over model M2 (which is model M1 but without using the site's stream width, depth and substratum composition at time of sampling) (Table 6).

Table 6 Summary of new model M44 relative to a null model (with no predictors), the default RIVPACS/RICT predictor model M1 and model M2 which excludes stream width, depth and substrate composition but include no new variables; based on average single season SD(O/E) and R^2_{OE} , together with the discrimination % correctly allocated (ReSub and XVal) to biological end-group.

	Null Model	Model		
		M1	M2	M44
%Correct (ReSub)	6.3	51.7	47.3	49.2
%Correct (XVal)	6.3	38.7	36.4	36.8
	SD(O/E) (lower is better)			
TL1 NTAXA	0.268	0.200	0.203	0.199
TL1 ASPT	0.121	0.076	0.079	0.078
TL2 WHPT NTAXA	0.281	0.206	0.208	0.204
TL2 WHPT ASPT	0.159	0.088	0.092	0.092
TL2 LIFE(DistFam)	0.085	0.053	0.055	0.055
TL4 LIFE(Sp)	0.099	0.057	0.061	0.061
TL3 PSI(Fam)	0.303	0.213	0.209	0.206
TL4 PSI(Sp)	0.346	0.271	0.256	0.248
TL3 E-PSI(fam69)	0.260	0.173	0.178	0.176
TL4 E-PSI	0.263	0.199	0.199	0.189
TL5 E-PSI	0.262	0.200	0.198	0.189
	R^2_{OE} (higher is better)			
TL1 NTAXA	0.0	44.3	43.1	45.2
TL1 ASPT	0.0	61.8	58.8	59.9
TL2 WHPT NTAXA	0.0	45.7	44.7	46.5
TL2 WHPT ASPT	0.0	71.8	68.7	68.3
TL2 LIFE(DistFam)	0.0	63.1	58.9	58.5
TL4 LIFE(Sp)	0.0	68.8	65.0	64.3
TL3 PSI(Fam)	0.0	70.1	66.3	65.9
TL4 PSI(Sp)	0.0	74.8	70.7	70.0
TL3 E-PSI(fam69)	0.0	77.2	71.0	70.5
TL4 E-PSI	0.0	78.0	71.1	70.2
TL5 E-PSI	0.0	77.4	70.6	69.7

To aid interpretation of the general effectiveness of new model M44 predictions, plots of the spring sample O/E values for the 685 GB reference sites for E values based on the new model M44 are given for the original BMWP indices (Figure 1), the abundance-weighted WHPT indices (Figure 2), the family and species level LIFE indices (Figure 3), family and mixed level PSI (Figure 4) and the family and mixed level E-PSI indices (Figure 5).

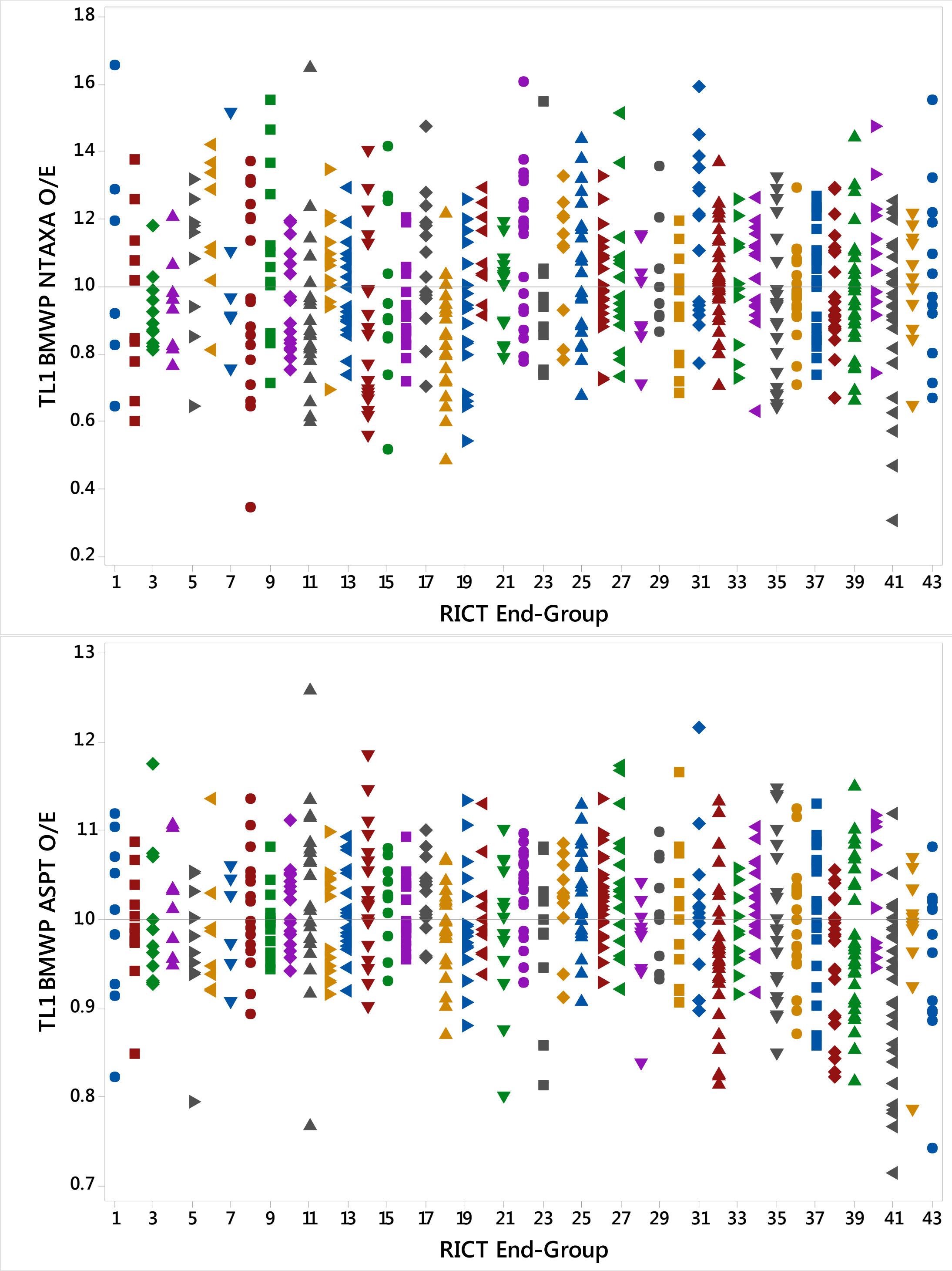

Figure 1 Plot of individual GB reference site spring sample O/E values by end-group (spring samples) for (a) TL1 BMWP NTAXA and (b) TL1 BMWP ASPT using predictions of expected (E) values based on Model 44, which excludes flow-related variables but includes new CEH-GIS-RICT predictor variables.

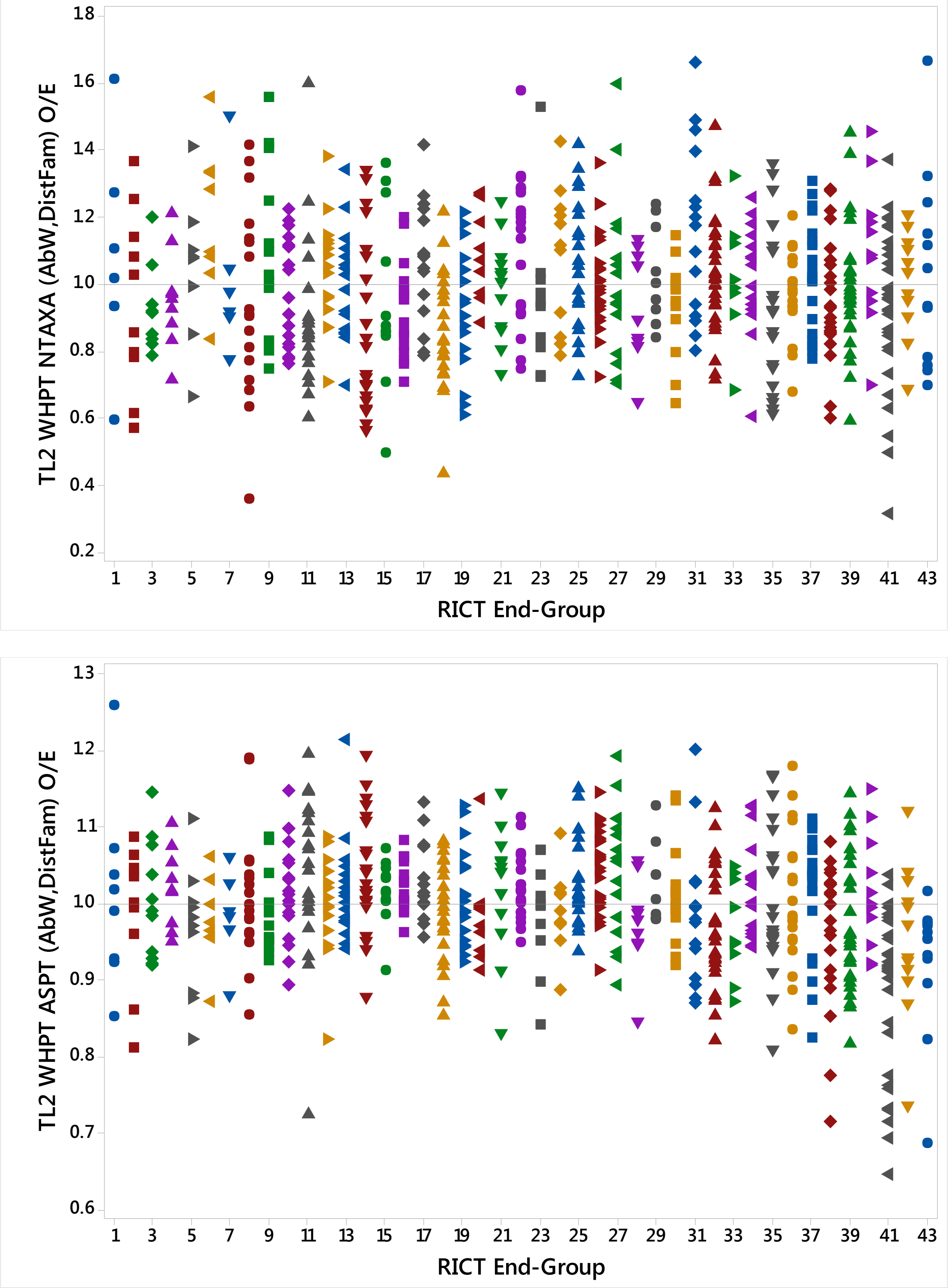

Figure 2 Plot of individual GB reference site spring sample O/E values by end-group (spring samples) for abundance-weighted (a) TL2 WHPT NTAXA and (b) TL2 WHPT ASPT using predictions of expected (E) values based on Model 44, which excludes flow-related variables but includes new CEH-GIS-RICT predictor variables.

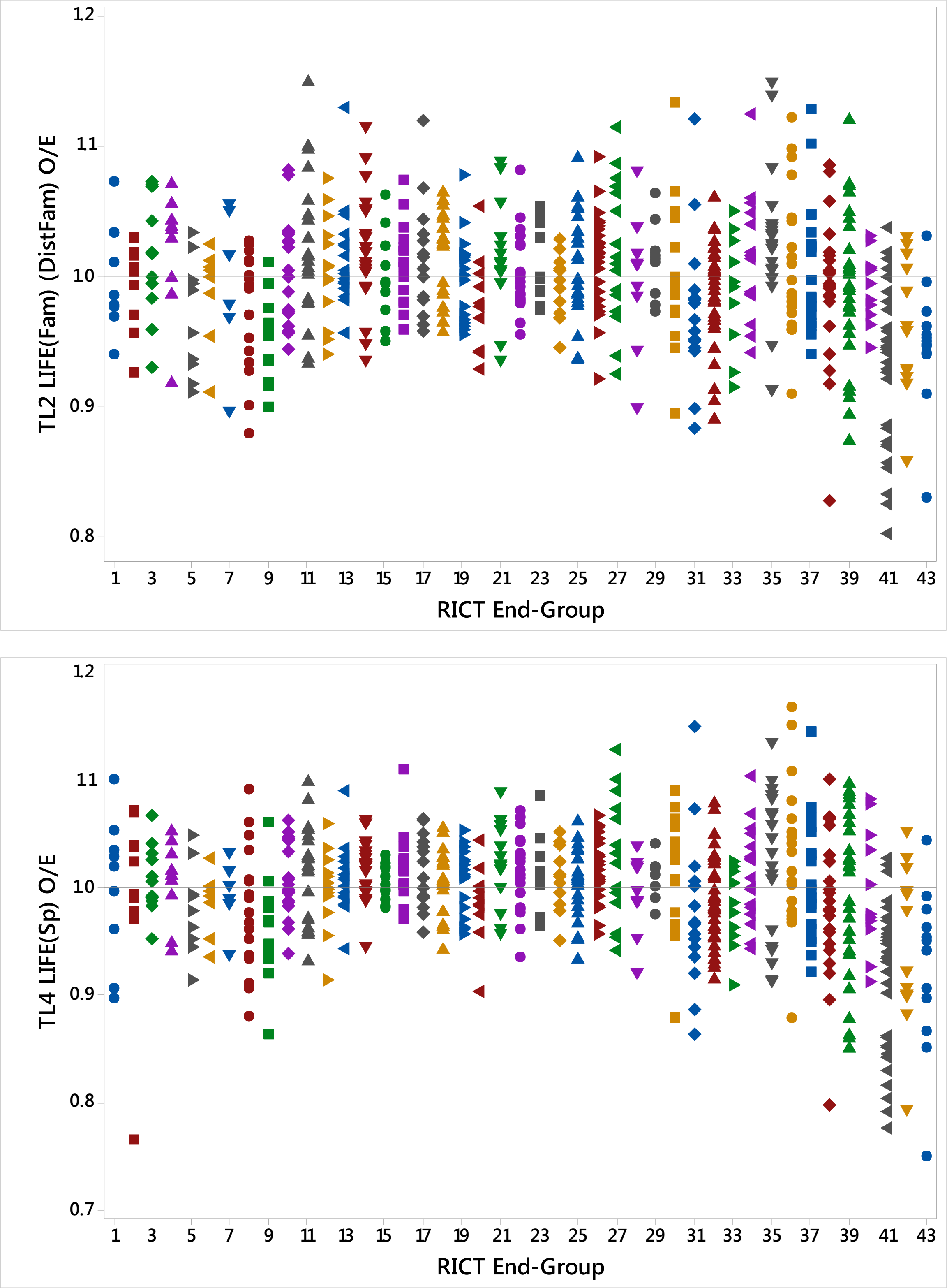

Figure 3 Plot of individual GB reference site spring sample O/E values by end-group (spring samples) for abundance-weighted (a) TL2 LIFE$_{(fam)}$ and (b) TL4 LIFE$_{(sp)}$ using predictions of expected (E) values based on Model 44, which excludes flow-related variables but includes new CEH-GIS-RICT predictor variables.

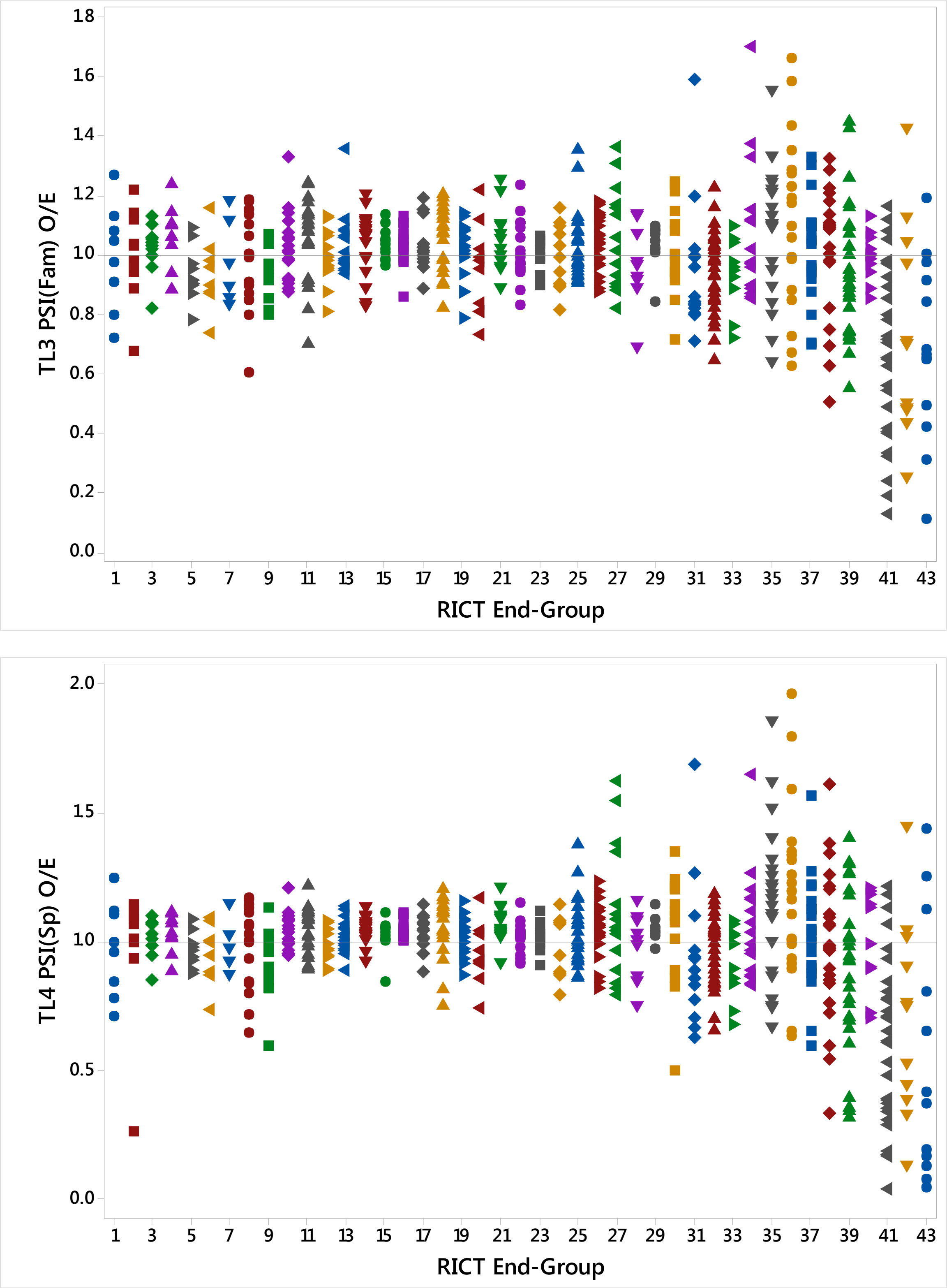

Figure 4 Plot of individual GB reference site spring sample O/E values by end-group (spring samples) for abundance-weighted (a) TL3 $PSI_{(fam)}$ and (b) TL4 $PSI_{(sp)}$ using predictions of expected (E) values based on Model 44, which excludes flow-related variables but includes new CEH-GIS-RICT predictor variables.

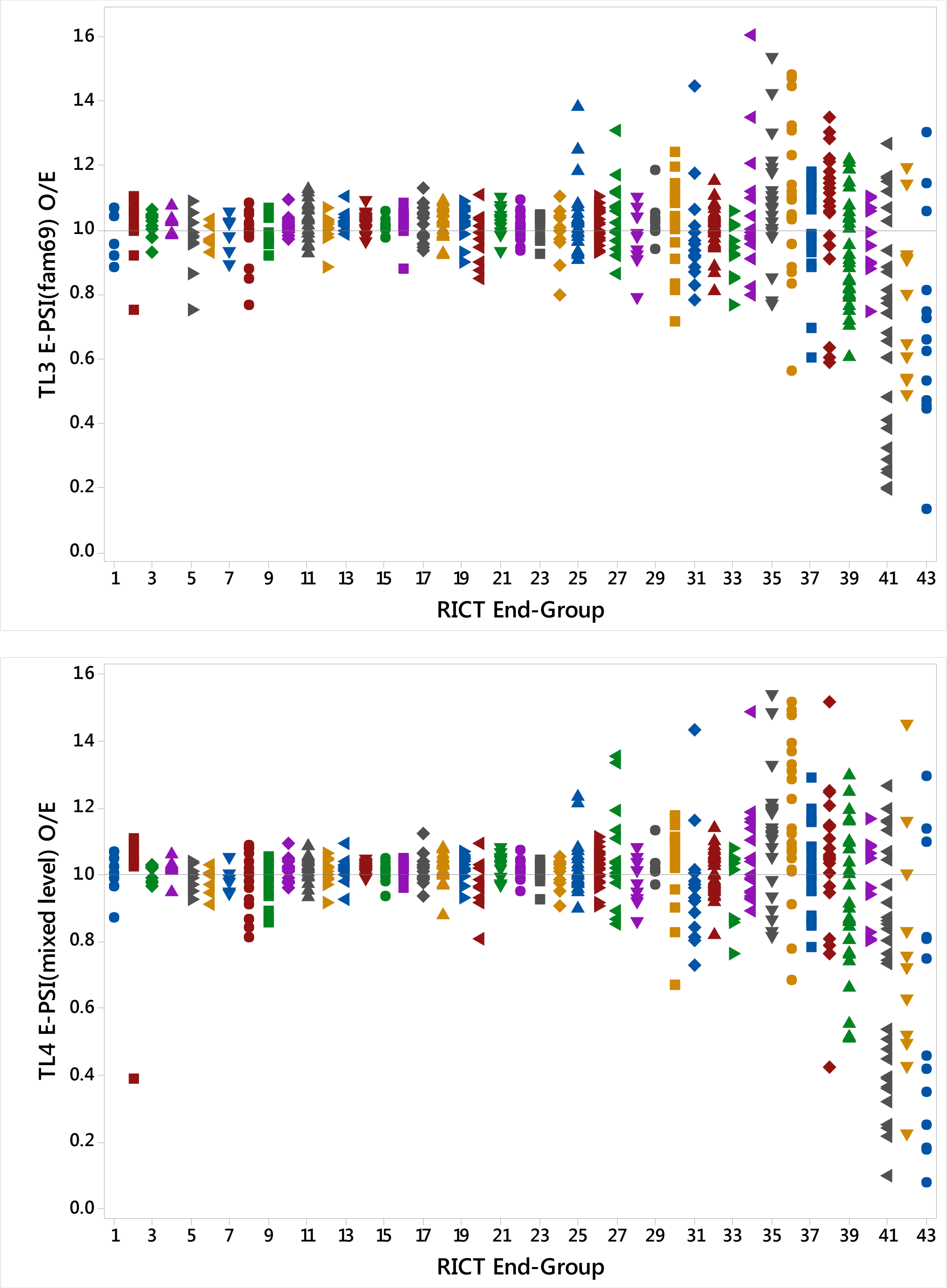

Figure 5 Plot of individual GB reference site spring sample O/E values by end-group (spring samples) for abundance-weighted (a) TL3 $\text{PSI}_{\text{(fam)}}$ and (b) TL4 E-$\text{PSI}_{\text{(mixed level)}}$ using predictions of expected (E) values based on Model 44, which excludes flow-related variables but includes new CEH-GIS-RICT predictor variables.

4.4.2 Over-estimation of expected values for end-groups 41-43

For the two taxonomic richness indices BMWP NTAXA (Figure 1(a)) and WHPT NTAXA (Figure 2(a)), the O/E values for the reference sites are broadly centred about one across the whole range of biological types of stream site from end-group 1 to 43. However, for the other indices, which are all based on abundance-weighted taxon-scoring systems, under new model M44 there is a general tendency for the majority of O/E values and the average O/E value to be less than one for reference sites in the last three biological end-groups 41-43 (Figures 1-5). End-groups 41-43 tend to be sites on large rivers; they have the deepest average water depth (89,154 and 145 cm respectively) and the greatest average percentage substratum cover of 'silt and clay' (38%, 38% and 91 % respectively) amongst all 43 RIVPACS End-groups.

This problem has been increasingly apparent to us since our initial development back in 2011 of predictive models which excluded the use of site stream width, depth and substratum composition. However, given the potential use of new model M44 in a future version of the RICT software for stream site assessments for sites potentially subject to flow- and sediment- related stress, it is useful for us to investigate and explain the reasons why this tendency for over-estimation of expected values and thus under-estimation of true O/E values may be a problems for these types of sites – which may also be amongst those types of sites most prone to such stresses.

The problem is that the biota present at a stream site at the time of sampling is influenced by the actual hydro-morphological conditions present at the site at that time. It is precisely because these physical conditions at the time of sampling may have been altered by human impacts that we ideally would not use physical variables (i.e. stream width depth and substratum composition) representing these conditions in our model predictions of what biota to be expected at a site in the absence of any such anthropogenic stress. However, the inevitable consequence is that our surrogate correlative predictive measures of the conditions at a site, such as upstream catchment geology, area and mean altitude, are unlikely to be as good predictors of the biota at a site.

The problem is especially acute for predictions of types of river site that have the most extreme values of biotic indices under reference conditions. This will be explained below.

RIVPACS model predictions of the expected index value for a site are a weighted average of the end-group mean observed reference site values of that index, where the weight given to each end-group in the site's prediction is based on the multiple discriminant analysis (MDA) using the selected environmental predictor variables. Thus, sites in end-groups with the lowest mean observed reference site values for any index can be predicted by the MDA to either (i) belong entirely (i.e. with probability one) to that end-group and hence have average expected values roughly equal to average observed values and average O/E values around one or (ii) to belong partly to their 'correct' end-group and partly to other end-groups (which must have higher mean observed values) and thus the prediction of the expected value will on average be higher than the mean observed values of sites in that end-group and in a sense an over-estimate, so that the average O/E value will tend to less than one and hence an under-estimate. Similarly, reference sites in the End-group with the highest mean observed index value can either be roughly predicted as definitely belonging to that group and hence have average O/E values around one, or they can be predicted to at least partly belong to other end-groups with lower mean observed value and hence have relatively lower predictions of expected values and thus average O/E values which are greater than one.

Table 7 and Table 8 give the average observed (O) value, average predicted expected (E) value and average O/E value for the reference sites in each RIVPACS biological End-group

(1-43) for four indices: WHPT NTAXA, LIFE, PSI and E-PSI, all at family level. Spring samples are used for illustration but the same features occur for other seasons.

Table 7 Comparison of new model M44 with current model M1 in terms of end-group (1-43) spring sample average observed (O), average predicted expected (E) and average O/E for (a) WHPT NTAXA (TL2 AbW, DistFam) and (b) LIFE (TL2, DistFam)

End-Group	(a) TL2 WHPT NTAXA (AbW,DistFam)					(b) TL2 LIFE(Fam) (DistFam)				
	O	E	E	O/E	O/E	O	E	E	O/E	O/E
		M1	M44	M1	M44		M1	M44	M1	M44
1	11.78	11.78	11.78	1.00	1.00	7.35	7.35	7.35	1.00	1.00
2	17.55	18.29	18.53	0.97	0.96	7.45	7.47	7.49	1.00	1.00
3	13.73	14.73	14.96	0.93	0.92	7.92	7.83	7.81	1.01	1.01
4	16.11	17.25	17.05	0.94	0.95	8.00	7.81	7.85	1.03	1.02
5	18.60	17.82	17.36	1.05	1.07	7.30	7.47	7.54	0.98	0.97
6	25.25	20.81	21.56	1.22	1.19	7.62	7.76	7.72	0.98	0.99
7	19.33	18.89	18.93	1.02	1.02	7.81	7.84	7.85	1.00	1.00
8	17.53	19.56	18.78	0.92	0.95	7.53	7.74	7.72	0.97	0.98
9	23.67	22.43	21.95	1.08	1.10	7.41	7.72	7.79	0.96	0.95
10	17.78	19.04	18.42	0.94	0.97	7.84	7.80	7.81	1.00	1.00
11	18.86	21.08	21.22	0.91	0.90	7.94	7.82	7.79	1.02	1.02
12	23.50	22.31	22.58	1.07	1.05	7.82	7.80	7.78	1.00	1.01
13	18.59	18.89	18.86	0.98	0.99	7.91	7.82	7.83	1.01	1.01
14	15.33	17.98	17.74	0.87	0.88	8.01	7.84	7.85	1.02	1.02
15	16.18	17.08	16.94	0.95	0.96	7.88	7.88	7.87	1.00	1.00
16	21.29	23.59	23.2	0.91	0.93	7.90	7.82	7.83	1.01	1.01
17	28.87	26.86	27.28	1.08	1.07	7.79	7.70	7.68	1.01	1.01
18	21.09	24.32	24.68	0.87	0.85	7.90	7.78	7.76	1.02	1.02
19	23.11	24.89	24.76	0.93	0.93	7.68	7.69	7.68	1.00	1.00
20	28.50	26.38	25.9	1.08	1.10	7.52	7.68	7.68	0.98	0.98
21	27.62	27.97	27.8	0.99	0.99	7.81	7.64	7.69	1.02	1.02
22	27.40	24.53	24.66	1.12	1.11	7.78	7.75	7.75	1.01	1.00
23	23.30	23.78	24.57	1.00	0.95	7.83	7.78	7.76	1.01	1.01
24	30.36	28.40	28.3	1.07	1.08	7.52	7.52	7.58	1.00	0.99
25	31.96	30.33	30.53	1.06	1.05	7.72	7.66	7.65	1.01	1.01
26	26.89	26.54	26.75	1.02	1.01	7.88	7.74	7.78	1.02	1.01
27	25.19	25.34	25.23	1.01	1.01	7.61	7.50	7.48	1.02	1.02
28	22.56	24.64	24.02	0.91	0.94	7.73	7.76	7.75	1.00	1.00
29	27.00	26.80	26.43	1.01	1.03	7.74	7.71	7.65	1.00	1.01
30	20.07	20.67	21.28	0.97	0.95	7.34	7.31	7.34	1.00	1.00
31	33.33	29.39	28.82	1.14	1.16	7.24	7.41	7.44	0.98	0.98
32	27.72	27.25	27.02	1.02	1.03	7.35	7.47	7.48	0.99	0.98
33	27.70	28.11	27.8	0.99	1.00	7.15	7.23	7.23	0.99	0.99
34	33.41	30.55	32.3	1.10	1.03	7.23	7.03	7.14	1.03	1.01
35	25.43	26.92	27.42	0.94	0.93	7.25	7.15	7.08	1.01	1.03
36	25.95	26.78	26.74	0.97	0.97	6.83	6.84	6.75	1.00	1.01
37	28.90	27.10	28.2	1.07	1.03	7.07	6.96	7.01	1.02	1.01
38	25.52	26.61	26.58	0.96	0.97	6.86	6.93	6.93	0.99	0.99
39	25.37	25.98	26.22	0.98	0.97	7.09	7.13	7.15	1.00	0.99
40	31.73	28.41	28.99	1.12	1.10	7.16	7.17	7.22	1.00	0.99
41	24.34	26.47	26.09	0.92	0.94	6.24	6.60	6.75	0.95	0.93
42	29.00	28.23	28.32	1.03	1.02	6.19	6.32	6.44	0.98	0.96
43	26.92	26.89	26.36	1.00	1.02	5.77	5.78	6.07	1.00	0.95

Table 8 Comparison of new model M44 with current model M1 in terms of end-group (1-43) spring sample average observed (O), average predicted expected (E) and average O/E for (a) PSI (TL3, Fam) and (b) E-PSI (TL3, Fam69)

End-Group	(a) TL3 PSI(Fam)					(b) TL3 E-PSI(Fam69)				
	O	E	E	O/E	O/E	O	E	E	O/E	O/E
		M1	M44	M1	M44		M1	M44	M1	M44
1	69.5	69.5	69.5	1.00	1.00	93.5	93.5	93.5	1.00	1.00
2	66.5	67.1	67.0	0.99	1.00	90.8	91.0	90.4	1.00	1.01
3	78.1	76.9	76.5	1.02	1.02	96.7	95.7	95.4	1.01	1.01
4	82.1	76.7	77.7	1.07	1.06	98.6	95.8	96.6	1.03	1.02
5	66.9	70.2	71.7	0.96	0.94	88.7	91.0	92.0	0.97	0.96
6	67.7	74.0	72.4	0.92	0.94	91.7	94.9	94.0	0.97	0.98
7	75.0	76.6	76.7	0.98	0.98	94.5	95.8	95.9	0.99	0.99
8	72.8	74.6	74.1	0.98	0.98	92.6	93.9	93.5	0.99	0.99
9	67.5	72.6	74.0	0.93	0.92	92.3	94.1	94.8	0.98	0.98
10	77.4	76.1	76.1	1.02	1.02	96.9	96.0	96.0	1.01	1.01
11	76.4	73.8	73.4	1.04	1.04	97.3	95.0	94.8	1.02	1.03
12	72.1	73.0	72.7	0.99	0.99	94.3	94.2	94.0	1.00	1.00
13	79.4	76.5	76.6	1.04	1.04	98.6	96.3	96.4	1.02	1.02
14	81.2	77.3	77.5	1.05	1.05	98.5	96.2	96.4	1.03	1.02
15	81.2	78.5	78.2	1.03	1.04	98.1	97.0	96.9	1.01	1.01
16	77.3	74.2	74.6	1.04	1.04	96.4	93.7	94.2	1.03	1.02
17	68.8	67.5	66.6	1.02	1.03	90.7	89.8	89.0	1.01	1.02
18	73.3	70.6	69.9	1.04	1.05	93.8	92.8	92.3	1.01	1.02
19	68.5	68.4	68.4	1.00	1.00	94.5	91.9	91.8	1.03	1.03
20	64.9	68.7	69.2	0.95	0.95	87.3	89.3	90.1	0.98	0.97
21	71.9	66.9	68.3	1.08	1.06	91.8	88.4	89.4	1.04	1.03
22	71.0	69.8	70.1	1.02	1.01	94.0	92.8	93.0	1.01	1.01
23	70.4	71.1	69.9	0.99	1.01	91.8	91.4	91.0	1.00	1.01
24	64.9	63.8	65.6	1.02	0.99	85.9	85.2	87.0	1.01	0.99
25	69.2	68.1	67.7	1.02	1.03	89.0	87.3	86.8	1.02	1.03
26	75.1	71.0	71.6	1.06	1.05	92.5	89.9	90.4	1.03	1.03
27	68.0	64.7	63.4	1.06	1.08	83.1	82.0	80.1	1.02	1.04
28	67.4	69.9	69.5	0.96	0.97	84.4	88.1	86.7	0.96	0.97
29	70.8	70.1	68.7	1.01	1.03	87.5	86.8	85.2	1.01	1.03
30	62.8	61.4	61.6	1.03	1.03	76.8	75.6	76.0	1.02	1.01
31	56.2	60.2	61.0	0.95	0.94	80.3	82.6	83.3	0.99	0.98
32	58.0	61.6	61.7	0.95	0.94	85.4	85.4	85.4	1.00	1.00
33	50.7	54.0	53.7	0.94	0.94	69.5	74.6	74.4	0.93	0.93
34	53.4	48.0	50.9	1.13	1.07	71.4	65.4	68.5	1.10	1.05
35	53.7	52.2	50.3	1.03	1.09	73.4	70.9	68.1	1.04	1.09
36	44.3	43.4	40.9	1.08	1.11	62.3	60.7	57.8	1.06	1.10
37	49.6	47.2	48.4	1.05	1.02	65.2	64.5	65.1	1.01	1.00
38	47.1	46.8	46.9	1.02	1.02	67.2	64.2	64.1	1.05	1.06
39	49.1	52.0	52.9	0.98	0.94	63.7	67.7	69.1	0.96	0.93
40	55.2	54.0	55.7	1.03	0.99	69.5	69.5	70.9	1.00	0.98
41	26.2	36.3	40.9	0.76	0.68	40.3	51.6	57.8	0.78	0.73
42	18.5	24.6	27.7	0.83	0.74	31.7	38.6	42.5	0.88	0.79
43	12.9	13.1	22.3	0.99	0.68	20.6	20.9	33.6	0.99	0.70

Firstly, the observed number of WPHT taxa is an example of an index with no such prediction problem. It varies between end-groups, it is by far the lowest for end-group 1, which are reference sites in the Shetlands, but the model predictions are able to always correctly allocate these sites (i.e. with probability one) to end-group 1 and hence the expected values are on average equal to the observed values and the average O/E value is one (Table 7(a)).

However, the other three indices illustrated, LIFE, PSI and E-PSI, all have the lowest mean observed index values for the three extreme end-groups 41-43 mentioned above. This means that if the reference sites in these groups cannot be confidently (i.e. with high probability) predicted to belong to the correct or other of these groups, then the predicted expected values will tend to be higher than the average observed and the O/E values will tend to be less than one for the majority of reference sites in these end-groups. Model M1 (involving stream width depth and substratum composition) is able to predict with greater confidence (i.e. higher probabilities) than new GIS-based model M44 the sites that belong to these extreme end-groups, which leads to less ‘over-estimation’ of expected LIFE values and less ‘under-estimation’ of LIFE O/E values for all three end-groups 41-43 (Table 7(b)). The average O/E amongst the reference sites for end-groups 41-43 is 0.95, 0.98 and 1.00 under model M1 and lower at 0.93, 0.96 and 0.95 under model M44.

However, the problem is less acute for LIFE (Table 7(b)) than for PSI or E-PSI (Table 8(a) and (b)). This is due to the fact that end-groups 41-43 have much lower average observed values for PSI and E-PSI than for any other end-groups. This is because these three end-groups are the only types of reference sites which have relatively few ‘fine-sediment-sensitive taxa present (as defined by the PSI (taxa sensitivity groups A and B) and E_PSI taxa sensitivity-weights in their index definitions) (see Clarke and Davy-Bowker 2017b Appendices 2 and 3 for details). Thus if the MDA model predictions for such sites are not largely correct, then expected E values will be major over-estimates and O/E values will be under-estimates.

For PSI, the average observed spring sample reference site values for end-groups 41-43 are 26.2, 18.5 and 12.9, whereas all of the end-groups have average observed values of between 44.3 and 81.2. The mean expected PSI for end-groups 41-43 are higher at 36.3, 24.6 and 13.1 under model M1 but even higher under new model 44, especially for extreme end-group 43, at 40.89 27.7 and 22.3 respectively. Thus model M1 can ‘correctly’ estimate sites in end-group 43 using their actual on-site features (average O/E = 0.99) but model M44 cannot from using their GIS-based catchment features (average O/E = 0.68) (Table 8(a)).

A similar problem occurs with the use of E-PSI in that although model M1 ‘over-estimates’ E values for end-groups 41 and 42, model M44 ‘over-estimates’ E and ‘under-estimates’ O/E for end-groups 41-43 to a greater extent than model M1 (Table 8(b)).

These problems are not specific to model M44, but apply to any RIVPACS-type model which cannot correctly predict the biological taxonomic composition of particular sites or types of site from the chosen suite of environmental predictor variables. In fact it applies to any type of statistical prediction model for biological types of sites whose environment features to be used in the taxonomic and index predictions do not confidently distinguish them from other types of sites with significantly different taxonomic composition and/or biotic index values.

4.4.3 Summary

Summary:
New model M44 is the first to base RIVPACS-model predictions of expected fauna and expected biotic index values on the new CEH-GIS-RICT database of GIS-based stream site and upstream catchment environment predictor variable. It will enable RIVPACS-type predictions of expected values to be made automatically, without any site visit for almost any river site in Britain.
It is the best model available to make predictions for sites potentially subject to hydromorphological stress. It may over-predict expected values and under-estimate O/E values some deep river sites dominated by fine sediment substratum.

5. TEST SITES INPUT DATA AND RICT OUTPUT DATA BASED ON MODEL M44

The intention is for the new GIS-based predictive model M44 to be incorporated into an upgraded or new version of the RICT software.

The mathematical algorithms to derive expected values using the MDA discriminant functions are all detailed in the original RICT development SNIFFER project WFD72C (Davy-Bowker et al. 2008).

To help both the RICT software programmers and maybe also RICT Users, we have provided critical data files to both help implement model M44 in RICT and to test the new software using a standard test dataset of 12 sites we have developed and provided.

The model M44 discriminant functions, together with the necessary Test sites dataset input and RICT output results are provided in the following separate Excel workbook file accompanying this report:

‘Model M44 Input Env Data and RICT Output Check Results for the 12 Test sites.xlsx’

This is referred to below as the ‘Model M44 Test Excel file’.

5.1 Discrimination functions for model M44

To aid the implementation of new model M44, we have provided an Excel file with the MDA discrimination functions needed to use with the model M44 environment predictor variables values of any new site to calculate its probability of belonging to each of the 43 end-groups (from which expected taxonomic composition and expected values of biotic indices can then be calculated).

The discriminant functions are provided as spreadsheet:

‘MDA DiscFunctions’ in the Model 44 Test Excel file.

5.2 Test dataset of 12 sites

As part of a previous project to test the current RICT software (Davy-Bowker and Clarke, May 2016), we developed a test dataset of 12 GB reference sites chosen to cover a range of stream types, sizes and geographic locations. We then added a copy of these 12 sites which we artificially degraded by altering their actual observed values of the biotic indices, giving 12 reference and 12 degraded test sites. As the expected values of the 12 reference sites and the corresponding 12 degraded sites are the same, here we merely give the details for the 12 reference sites as the environmental input data values, probability of end-group and expected values of biotic indices are the same for the 12 matching degraded test sites as for the 12 test reference sites.

The 12 test sites are listed in
The actual variables used in the predictions, where necessary in the log-transformed form, are given in the first set of columns in the spreadsheet and then to the right we give the untransformed values of the transformed variables:
Upstream catchment Area and mean altitude,
Site Distance from source, altitude and slope

The RICT software requires the untransformed predictor variable values as input and internally log-transforms the required variables to make the site predictions.

Table 9.

The values of the 17 environmental predictor variables used in model M44 predictions are provided as spreadsheet:

‘EnvData’ in the Model 44 Test Excel file.

The actual variables used in the predictions, where necessary in the log-transformed form, are given in the first set of columns in the spreadsheet and then to the right we give the untransformed values of the transformed variables:

Upstream catchment Area and mean altitude,
Site Distance from source, altitude and slope

The RICT software requires the untransformed predictor variable values as input and internally log-transforms the required variables to make the site predictions.

Table 9 List of 12 Test sites

Test Data Site Number	England/ Scotland/ Wales*	TWINSPAN End Group (1-43)	Site Code	River Name	Site Name
TST-01-R	Eng.	20	3101	Derwent	Langdale End
TST-02-R	Eng.	24	9581	Lathkill	Alport
TST-03-R	Eng.	28	8805	Coombevalley Stream	Kilkhampton
TST-04-R	Eng.	32	2007	Blithe	Newton
TST-05-R	Eng.	36	2307	Colne	Fordstreet Bridge
TST-06-R	Eng.	40	7145	Ed	Pains Moor
TST-07-R	Eng.	43	6111	Ouse/Cam	Hilgay Bridge
TST-08-R	Scot.	1	SEPA_N06	Shetland: Burn of Laxdale	North Voxter
TST-09-R	Scot.	4	SEPA_W05	Islay: Duich/Torra	Torra Bridge
TST-10-R	Scot.	8	3785	Green Burn	Dalmary
TST-11-R	Scot.	12	NE01	Lossie	Cloddach
TST-12-R	Wales	16	WE03	Afon Caseg	Braichmelyn

5.3 Probabilities of End-group under model M44 for the 12 Test sites

To provide a check that the model M44 discriminant functions have been used correctly to calculate probabilities of end-group for a site in the RICT software, the predicted probabilities of belonging to each of the 43 End-groups for each of the 12 test sites are provided as spreadsheet:

‘ProbEndGroup’ in the Model 44 Test Excel file.

5.4 Expected values of biotic indices under model M44 for the 12 Test sites

The RIVPACS expected (E) values based on new predictive model M44 of the following biotic indices:

TL1 NTAXA
TL1 ASPT
TL2 WHPT NTAXA (AbW,DistFam)
TL2 WHPT ASPT (AbW,DistFam)
TL2 LIFE(Fam) (DistFam)

TL4 LIFE(Sp)
TL3 PSI(Fam)
TL4 PSI(Sp)
TL3 E-PSI(fam69)
TL4 E-PSI(mixed level)
TL5 E-PSI(mixed level)

for each of the 12 test sites are provided as spreadsheet:

‘ExpIndValues’ in the Model 44 Test Excel file.

6. REFERENCES

Clarke R. T., Davy-Bowker J., Dunbar M., Laize C., Scarlett P.M. & Murphy J.F. (2011). *Enhancement of the River Invertebrate Classification Tool (RICT)*. A report to the Scotland and Northern Ireland Forum for Environmental Research. [SNIFFER project WFD119].

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. *Ecological Modelling* 160: 219–233.

Clarke R.T. & Davy-Bowker J. (2017a) A new hydromorphology-independent RIVPACS predictive model (Model 37). A report to the Environment Agency.

Clarke R.T. & Davy-Bowker J. (2017b) River Invertebrate Classification Tool Science Development Project: Assessing uncertainty in E-PSI$_{(fam69)}$ and WFD-AWIC$_{(sp)}$. A report to the Scottish Environment Protection Agency (SEPA).

Davy-Bowker J., Clarke R., Corbin T., Vincent H., Pretty J., Hawczak A., Blackburn J., Murphy J. & Jones I. (2008). *River Invertebrate Classification Tool*. Scotland & Northern Ireland Forum for Environmental Research. Edinburgh, Scotland, UK. (SNIFFER project WFD72C).

Kral F., Robertson O., Fry M & Laizé C. (2017) River Invertebrate Classification Tool Database and Delivery System. A report to the Scottish Environment Protection Agency (SEPA Report R15056PUR 13June2017)

Moss D., Wright J.F., Furse M.T. & Clarke R.T. (1999). A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. *Freshwater Biology* 41: 167–181.

Van Sickle J., Hawkins C.P., Larsen D.P. & Herlihy A.T. (2005) A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society* 24: 178–191.

Acknowledgements:

We thank Ian Milne and John Murray-Bligh for their support to this project and comments on the draft report.

This page intentionally blank

www.fba.org.uk

The Freshwater Biological Association
The Ferry Landing
Far Sawrey, Ambleside
Cumbria, LA22 0LP, United Kingdom

The Freshwater Biological Association
River Laboratory
East Stoke, Wareham
Dorset, BH20 6BB, United Kingdom

Freshwater Biological Association