Evaluation of benthic diatom classification in UK rivers using LM and NGS methods

Report number E18-56 (PO 4057194)

Martyn Kelly[1], Steve Juggins[2], Geoff Phillips[3] and Nigel Willby[3]

[1] Bowburn Consultancy, 11 Monteigne Drive, Bowburn, Durham DH6 5QB,
[2] School of Geography, Politics and Sociology, Newcastle University, Newcastle NE1 7RU,
[3] Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA

September 2018

Executive summary

Two earlier projects, funded by the Environment Agency (SC140024 and SC160014) demonstrated the potential for using next generation sequencing (NGS) for the analysis of the composition of benthic diatom assemblages in rivers. This opened the possibility that, for the first time, ecological assessment of an element of the freshwater biota required under the Water Framework Directive could be performed using molecular, rather than traditional morphology-based, taxonomy. This report describes a further series of evaluations of this NGS method, including data from all parts of the United Kingdom, and also describes an improved reference model for ecological status assessments.

A further 381 samples from Northern Ireland, Scotland and Wales, each analysed by both light microscopy (LM) and NGS, have been added to the project database, giving a total of 1728 paired samples, of which 1518 have been linked to chemical data. This new dataset was used to recalibrate TDI5LM (the optimized version of the Trophic Diatom Index, TDI, described in SC160014) and this, in turn, was used to derive a new version of TDI5NGS, using the same approach as described in earlier reports. This new version of TDI5NGS has very similar characteristics to the version delivered in the previous phase of the project and, using the current reference model, is 5.3% less stringent than TDI4 but only 0.4% less stringent than TDI5LM.

A new reference model is proposed, using the 10th percentile of the relationship between TDI and alkalinity as an estimate of the “best available” TDI at any given alkalinity. Nitrate and season are also included in this equation, the latter as a means of separating the effect of land use from the alkalinity gradient. The resulting reference model removes the effect of alkalinity on EQR (Ecological Quality Ratio) but yields classifications that are substantially more stringent than those obtained using the current reference model, and which are also more stringent than those obtained using macrophytes.

In order to overcome this, a new combination rule, based on the average, rather than lowest, of the two sub-elements of the “macrophytes and phytobenthos” biological quality element is suggested. This yields classifications for the combined BQE that are on average 10% more stringent than those derived from the present diatom and macrophyte models; however, a final adjustment could be applied to bring the stringency in line with the present method. Whilst the nationwide picture will not change, individual site classifications should be more accurate using this approach.

Finally, options for evaluating risk of misclassification and confidence of class using the new methods are described, with an approach based on a combination of observed and modelled standard deviations being recommended.

Scope of report

This report provides an evaluation of benthic diatom classification in UK rivers using the light microscopy and next gene sequencing methods, and provides options for an appropriate assessment of WFD status results.

Acknowledgements

Special thanks are extended to Dr Kerry Walsh (Environment Agency). The report was funded through SEPA and managed by Dr Jan Krokowski.

1 Introduction

Environment Agency science reports SC140024 (Kelly et al, 2018a) and SC160014 (Kelly et al., 2018b) described the development of a metabarcoding approach for the use of benthic diatoms to assess ecological status in rivers, leading to a prototype metric that showed good agreement with the current analytical method based on light microscopy. This, for the first time, offers the potential for a nationwide application of Next Generation Sequencing (NGS) technologies for routine assessments compatible with the requirements of the Water Framework Directive.

Although these earlier reports used data from throughout the UK, the majority of samples (90%) came from England, with limited sampling in Northern Ireland (61 sites), Scotland (67 sites) and Wales (11 sites), and these mostly from putative "reference sites". Whilst the dataset spanned a wide environmental gradient, concerns were expressed that there were insufficient data to allow the administrations in Northern Ireland, Scotland and Wales to evaluate performance of the new approach in their territory. Work presented here was commissioned in order to allow thorough evaluation in each of the constituent parts of the United Kingdom using spring and autumn samples from one year. At the same time, the addition of new data, potentially including habitats not sampled during earlier phases, has provided an opportunity for a final recalibration of the models.

The development of the NGS-based metric took place at the same time as the reference model that underpinned DARLEQ2 (Diatoms Assessment of River and Lake Ecological Quality - the current tool for determining phytobenthos status in the UK) was being evaluated. In SC160014, the consequences of using an alternative reference model were included in order to put the changes that would result from a switch from analysis by light microscopy to NGS into perspective. This showed that changing to a plausible alternative reference model had a greater effect on final classifications of ecological status than the switch to NGS. Therefore, in light of the limitations of the current reference model, this report also includes a proposal for an alternative model, and an evaluation of the consequences of this for ecological status classification across the UK.

2 Recalibration of TDI and finalisation of TDI5NGS

2.1 Dataset summary

The datasets used in this report consist of 381 new light microscopy (LM) and NGS samples collected from rivers in Northern Ireland, Scotland and Wales, together with an existing dataset of 1362 samples reported in SC160014. These two datasets are referred to here as Phase 3 (new samples collected in this study) and Phase 2 (existing samples) respectively. LM and NGS taxon counts have been harmonised against the DARLEQ master taxon dictionary which has been updated to reflect new taxa recorded in the Phase 3 data and screened to remove any NGS samples with fewer than 500 reads of non-planktonic taxa.

Phase 3 diatom counts have been matched to hydrochemistry data from Wales, Scotland and Northern Ireland, and all samples (Phase 2 and 3) have been matched to associated macrophyte data where available. Environmental variables included are PO_4-P, NO_3-N, alkalinity, conductivity and pH. Hydrochemistry data are expressed as annual means using either the arithmetic mean (alkalinity and pH) or geometric mean (all other variables) of all available data for the period 2012-2017).

Determinations less than the detection limit were taken as half the detection limit. This may overestimate actual values at low concentrations but water chemistry data was used primarily to validate diatom metrics and only used as a guide to modify the indicator values of a few, rare taxa (Section 2.2).

Table 2.1 gives a breakdown of the numbers of samples available with matching LM and NGS counts, and for these, with matching water chemistry (PO_4-P and NO_3-N), and macrophyte survey data. The merged dataset contains a total of 1714 samples with matching LM and NGS counts and this is used to refine the TDI5NGS metric (Section 2.3). A smaller set of 1505 samples was used to validate and update TDI5LM indicator values (Section 2.2), to compare the revised diatom metrics with the pressure gradient (Section 2.4), and to explore alternative diatom reference models (Section 3). The subset of samples with matching macrophyte data is used to explore phytobenthos and macrophyte combination rules (Section 3.5). Summary statistics and figures are split by Phase 2 & 3 datasets, with the latter further split by UK region. Cross-tabulations showing the implications of the various metrics and reference models for final status classification are split by UK region alone.

Table 2.1: Number of samples in Phase 2 and Phase 3 datasets.

	Total LM	Total NGS	Matching LM & NGS	With chemistry	With macrophyte data
SEPA	206	199	198	194	139
NRW	149	149	137	127	47
NIEA	26	23	23	23	23
Phase 3	381	371	358	344	209
Phase 2			1356	1161	234
Total			1714	1505	443

2.1.1 Species profiles

After taxonomic harmonisation the combined Phase 2 & 3 LM and NGS datasets contains a total of 525 and 307 non-planktonic taxa respectively. This includes 34 new taxa unique to the Phase 3 LM dataset and 3 new taxa unique to the Phase 3 NGS dataset (Table 2.2). Most of the new taxa in the LM dataset are rare, with low numbers of occurrences and low maximum abundance (Table 2.3), with only one taxon, the softwater *Eunotia naegelii* locally abundant. For the NGS data, all new taxa are very rare, with maximum abundance less than 0.1%.

Table 2.2: Number of non-planktonic taxa in light microscopy and NGS samples.

	Light Microsc.	eDNA
Total number of taxa (Phase 2)	491	304
Total number of taxa (Phase 3)	350	269
Total number of taxa (Combined)	525	307
Number of taxa unique to Phase 3	34	3

Table 2.3: Taxa unique to the Phase 3 LM dataset, showing number of occurrences (N), occurrences expressed as Hill's N2 diversity, and maximum abundance (Max) for a subset of taxa with maximum abundance > 1.0%.

	TaxonId	N	N2	Max
Eunotia naegelii	EU048A	2	1.0	28.1
Nitzschia soratensis	NITZ-03	5	3.2	3.4
Gomphonema auritum	GO030A	1	1.0	3.3
Fragilaria radians	FR059A	4	2.7	3.3
Fragilaria pectinalis	FRAG-04	8	4.2	2.9
Nitzschia abbreviata	NITZ-04	4	2.7	2.5
Fragilaria pararumpens	FRAG-03	8	4.8	1.9
Encyonema brehmii	ENCY-04	2	1.6	1.7
Encyonema hebridicum	EY003A	5	3.7	1.3

Figure 2.1 shows the total benthic diatom read count (after excluding samples with less than 500 reads) and the percentage of unassigned reads ("No blast hit") by dataset. The profiles for Phase 3 datasets are very similar to those from Phase 2, with an average read count of 33800 and an average of 40% unassigned reads. Samples from Northern Ireland had fewer unassigned reads than from other parts of the UK but there is no obvious reason for this.

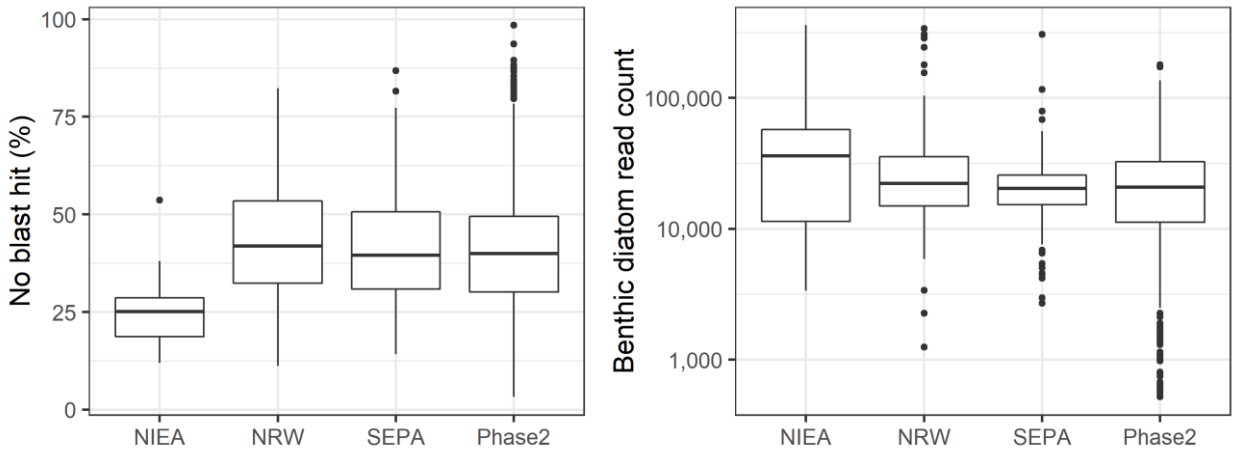

Figure 2.1: Distribution of % no blast hits (left) and total benthic diatom read counts (right) by dataset.

2.1.2 Environmental data

Over 1500 of the paired LM/NGS samples could be matched to water chemistry data (Table 2.4). Figures 2.2 and 2.3 summarise the coverage of key variables by dataset. Phase 3 samples provide good coverage of the low to moderate PO_4-P and NO_3-N gradients, with median values somewhat lower than the much larger Phase 2 dataset (Table 2.4). Similarly, alkalinity and conductivity values for Phase 3 have slightly lower values overall than Phase 2 but also cover these gradients well, with no bias towards softwaters. The distribution of PO_4-P values reflects the use of different limits of detection for routine analyses within the different agencies (0.02, 0.01 and 0.001 mg L^{-1}).

Table 2.4: Summary statistics of selected environmental variables for the combined LM/NGS dataset.

Variable	Dataset	N	Mean	Median	Min	Max
PO_4-P (ug/L)	Phase 2	1163	96.60	34.46	1.00	3600.00
	Phase 3	344	28.64	16.01	4.00	430.12
NO_3-N (mg/L)	Phase 2	1161	2.69	1.76	0.05	27.25
	Phase 3	348	1.75	0.93	0.07	18.32
Conductivity (uS/cm)	Phase 2	1016	358.39	273.92	32.21	2162.27
	Phase 3	348	206.71	168.82	25.91	969.29
Alkalinity (mg/L)	Phase 2	1211	85.82	58.70	1.70	381.65
	Phase 3	348	59.22	46.54	3.10	255.40
pH	Phase 2	1032	7.76	7.81	5.77	8.44
	Phase 3	348	7.51	7.60	5.98	8.38

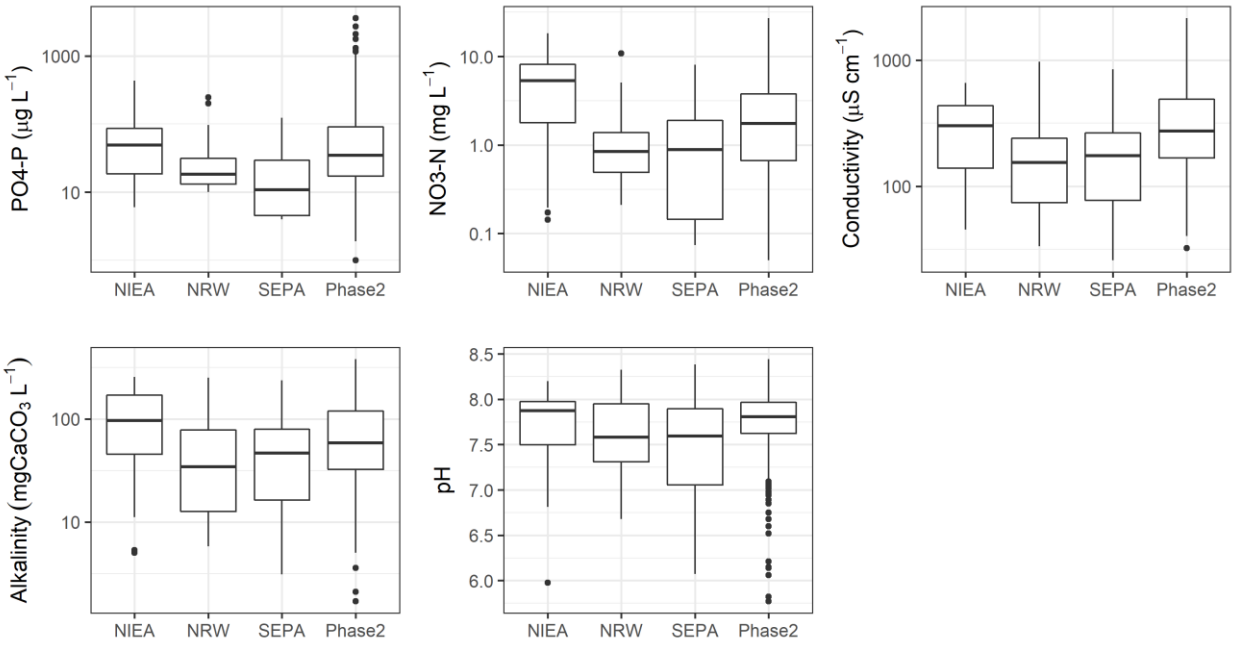

Figure 2.2: Summary distributions of select hydrochemistry by dataset.

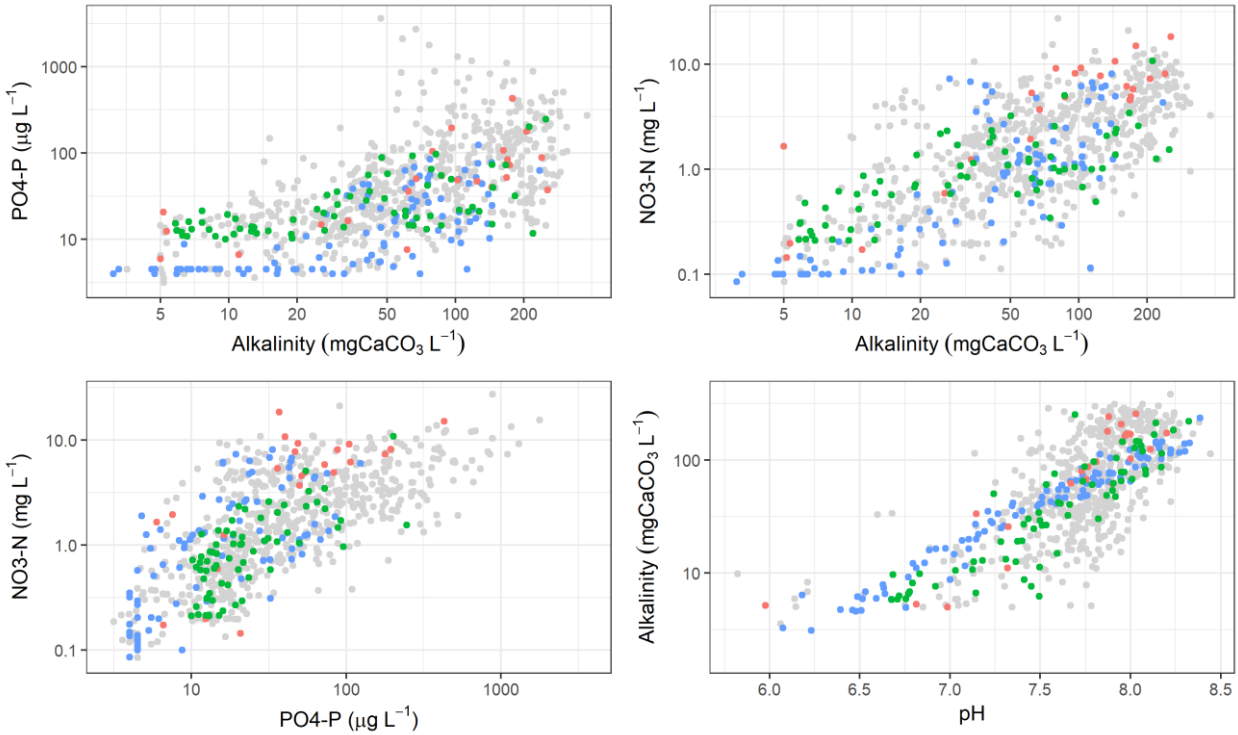

Figure 2.3: Relationships between hydrochemical variables by dataset (red = NIEA, green = NRW, blue = SEPA, grey = Phase 2).

2.2 Optimising the LM metric

In SC140024 and SC160014 the large combined LM and chemistry dataset ("Phase 2") was used to test the efficiency of the TDI4 metric as an indicator of the nutrient pressure gradient, and to derive a new metric, TDI5LM. The majority of taxon indicator values in TDI5LM were the same as those in TDI4 but TDI5LM did include improvements over TDI4. These were firstly, the large Phase 2 LM dataset allowed the addition of new taxa into the metric that were not included in TDI4. Secondly, analysis of the combined LM / chemistry dataset allowed the validation of taxon indicator values, and in some cases a revision of these values to better reflect the taxon's distribution along the nutrient gradient. Finally, the analysis in Phase 2 also presented an opportunity to update the master taxon list to reflect recent changes in diatom nomenclature.

In this section we analyse the combined Phase 2 and 3 dataset to provide additional, minor, updates to TDI5LM. In the text below DARLEQ3 refers to versions of the metrics in DARLEQ3 Version 0.8.4 (i.e. the version produced as an output of SC140016). DARLEQ3.1 refers to new versions of the metrics updated during the work reported here and implemented in DARLEQ3 software version 1.0.0.

2.2.1 TDI5LM vs. TDI4

Figure 2.4 shows the relationship between TDI4 and TDI5LM using the DARLEQ3 version of the metrics. As reported in SC140016, the TDI4 and TDI5LM sample metric scores for Phase 2 data show very close agreement (r = 0.993) (Figure 2.4, right). However, for Phase 3 data (Figure 2.4, left) TDI4 values are systematically

higher for some samples, especially those from Wales. These samples have moderate to high relative abundance (RA) of *Gomphonema intricatum* type. The entity *Gomphonema intricatum* type derives from an earlier project where a number of species aggregates were adopted in response to requests to make identification of taxa for the TDI more straightforward (Kelly & Yallop, 2012). One of these amalgamated several taxa in the *Gomphonema pumilum/angustum* complex. The name given to this aggregate was "*Gomphonema intricatum* type", as this would have been the name used for these species in older floras (e.g. Hustedt, 1930). The name *G. intricatum* is, in fact, no longer used. Most records for this complex probably refer to G. pumilum.

Gomphonema intricatum type has a TDI4 indicator value of 3.6. It is relatively uncommon in Phase 2 data where it is rarely found above 20% RA (N=5). On the basis of analysis in SC160014 this taxon was given a lower revised indicator value of 2.0 in TDI5LM. However, the re-analysis with the additional Phase 3 samples, where *Gomphonema intricatum* type is more frequent and abundant, suggests that the indicator value of 2 is too low for this taxon (Figure 2.5). The larger dataset therefore gives us an opportunity to re-evaluate the indicator values for this, and other taxa, and this is explored in the next section.

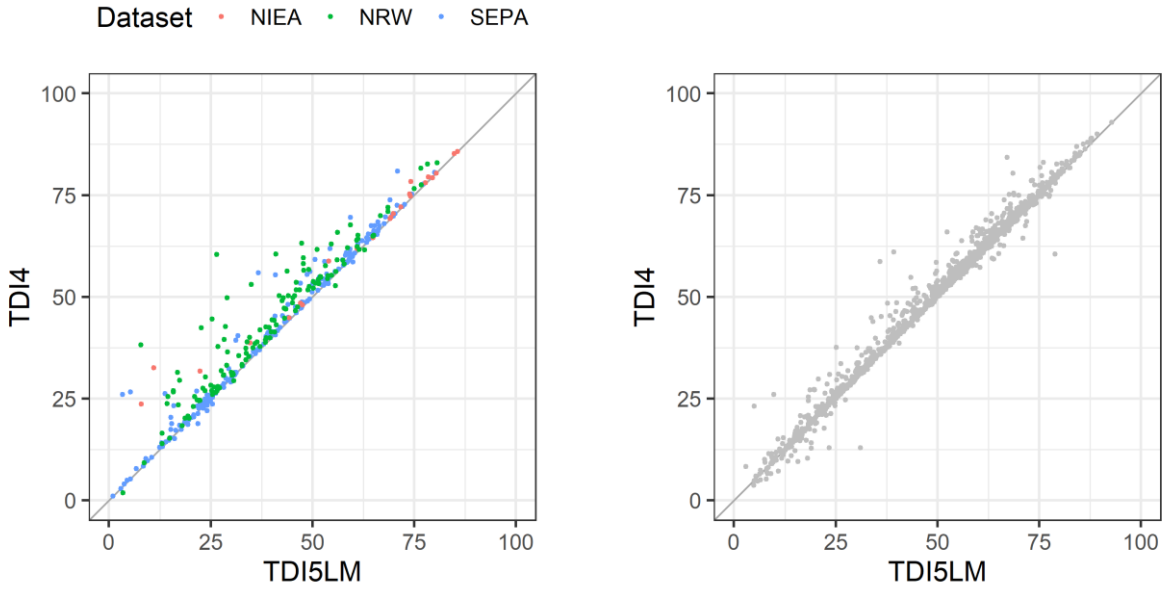

Figure 2.4: Scatter plot of TDI4 vs. TDI5LM scores for Phase 3 (left) and Phase 2 (right) datasets using the original DARLEQ3 metrics.

2.2.2 Updating TDI5LM indicator values

Figure 2.5 shows the relationships between TDI5LM indicator values and the weighted average (WA) indicator values of a model of taxon response along the first component of a principal components analysis of PO_4-P and NO_3-N (PC1). This component effectively combines the phosphorus (P) and nitrogen (N) gradients into a single pressure variable, and the WA indicator values give the centroid of a taxon's distribution along this gradient in the combined Phase 2 & 3 dataset. Note that most taxa have TDI4 and TDI5LM indicator values that are integers but a few have indicator values that are decimals as a result of an earlier Environment Agency funded project in which groups of taxa that proved challenging to analysts were amalgamated into categories that were given the weighted mean indicator value of the constituent species (Kelly & Yallop, 2012).

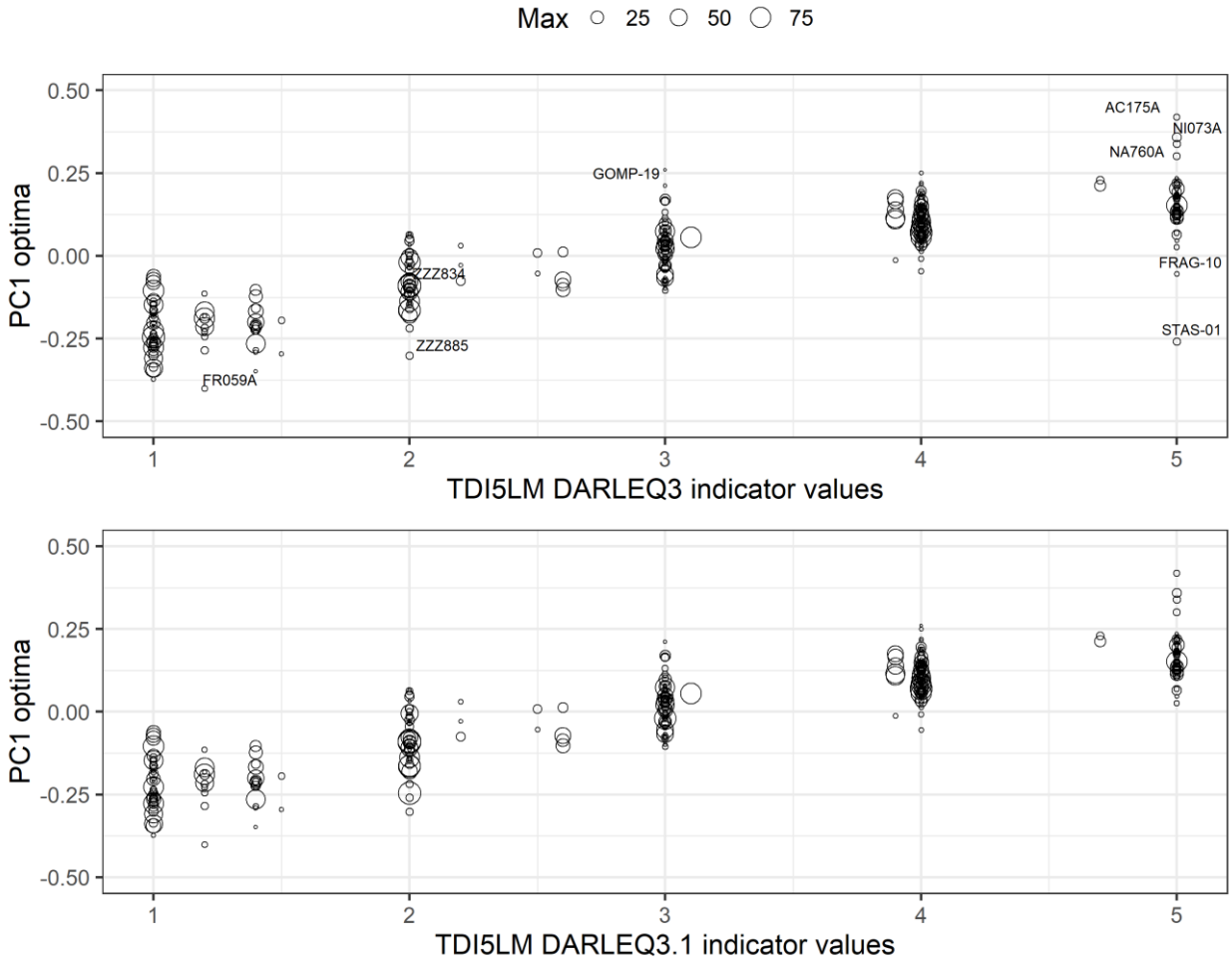

Figure 2.5: TDI5LM indicator values plotted against WA optima derived from the PC1 nutrient pressure gradient. Top = original (DARLEQ3) indicator values, bottom = updated (DARLEQ3.1) TDI5LM indicator values. See table 2.5 for taxon names.

Figure 2.5 shows that there is a range of PC1 optima for any given TDI5LM indicator value, as a result of the granular nature of TDI5LM indicator values. However, several taxa (labeled), appear to be mis-classified by TDI5LM according to their distribution in the Phase 2 & 3 dataset. These apparent outliers are listed in Table 2.20. The indicator values of these taxa have been updated in light of the new data (Table 2.5) and the updated values incorporated into the revised version of TDI5LM in DARLEQ3.1. A list of all taxa included in the DARLEQ TDI metrics, along with their TDI4, TDI5LM and TDI5NGS indicator values, is available in the R package DARLEQ3 available at https://github.com/nsj3/darleq3.

Table 2.5: List of taxa with updated indicator values for TDI5LM

Taxon Code	Taxon Name	Original Ind. Val.	New Ind. Val
AD9999	Achnanthidium sp.	1	2
FRAG-10	Fragilaria famelica	5	4
ZZZ834	Gomphonema “intricatum” type	2	3
GOMP-19	Gomphonema productum	3	4
SR9999	Staurosira sp.	4	3
STAS-01	Staurosirella lapponica	5	2
STAS-02	Staurosirella martyi	5	3

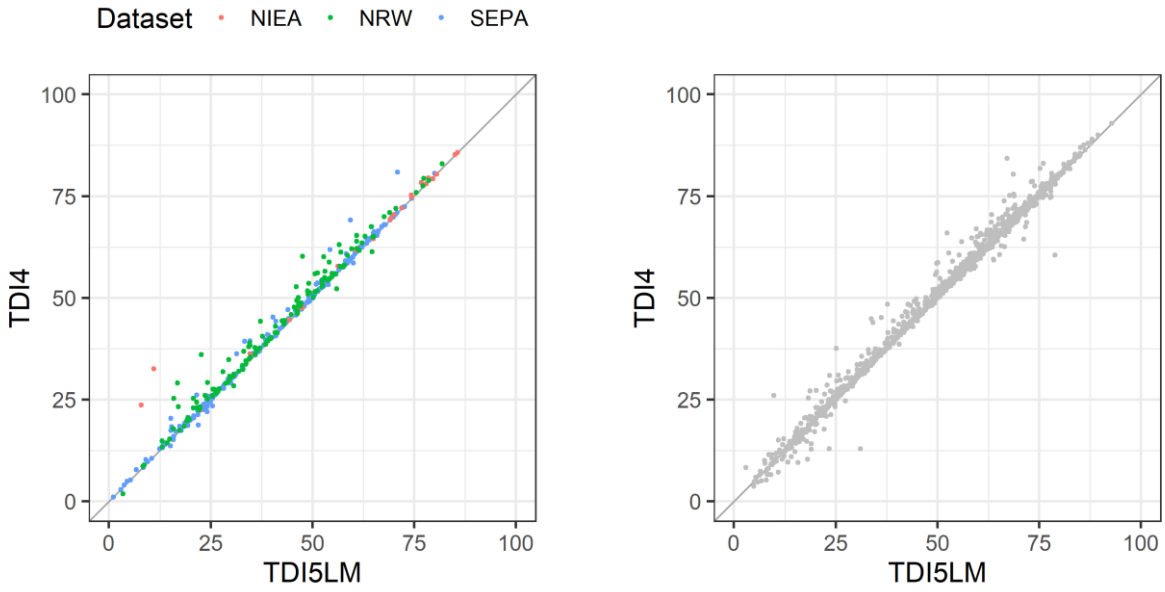

Figure 2.6: Scatter plot of TDI4 vs. updated TDI5LM scores for Phase 3 (left) and Phase 2 (right) datasets.

Table 2.6: Correlations between TDI4 and original (left) and updated (right) TDI5LM scores.

	Correlation	Concordance correlation
Phase 2	0.995	0.995
Phase 3	0.992	0.990

Figure 2.6 shows the relationship between TDI4 and TDI5LM using the DARLEQ3.1 (i.e. updated) version of TDI5LM, and Table 2.6 lists the associated correlation coefficients. Although the difference in correlation between the old and updated scores is small (r = 0.995 and 0.992 for Phase 3 data with original and updated metrics respectively), the outliers caused by the mis-classified *G intricatum type* and other taxa have been reduced. The small differences now observed between TDI4 and TDI5LM in Figure 2.6 are the result of differences in indicator values between these two metrics which we believe reflect more accurately the response of taxa to the nutrient gradient.

Tables 2.7 and 2.8 show effect the updates to TDI5LM have on classification. The number of samples that change class is small, with the main differences affecting samples containing *G. intricatum* type, which move to a higher class. This is the result of the lower indicator value now assigned to this taxon (Table 2.5) and more accurately reflects the status of these samples.

Table 2.7: Comparison between ecological status classes for samples computed by original TDI5LM (columns) and updated TDI5LM (rows). Green shading: identical classification for both metrics; yellow shading: agreement to within one class.

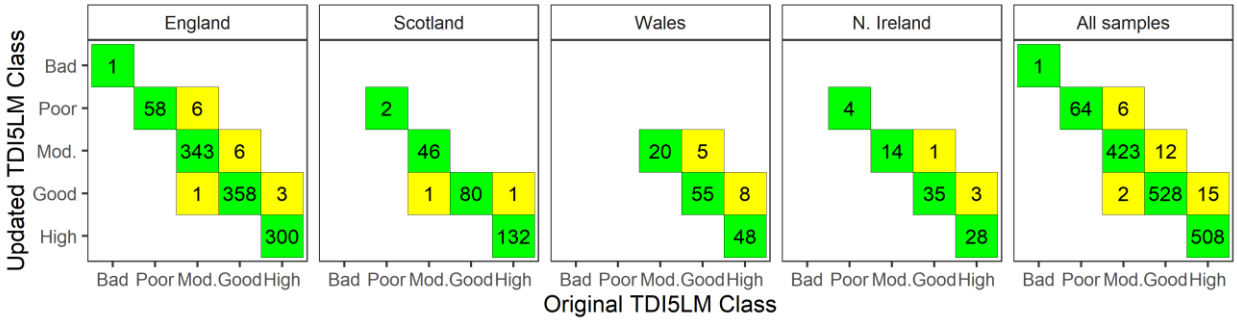

Table 2.8: Summary statistics for classifications presented in Table 2.7.

	England	Scotland	Wales	N. Ireland	All samples
N	1076.0	262.0	136.0	85.0	1559.0
% Agree	98.5	99.2	90.4	95.3	97.8
% Agree within one class	100.0	100.0	100.0	100.0	100.0
% Bias	1.3	0.0	9.6	4.7	2.0

2.3 Updating TDI5NGS

This section uses the new, larger Phase 2 & 3 datasets and revised TDI5LM scores to derive new TDI5NGS indicator values using the methods described in SC160014.

2.3.1 Updating TDI5NGS indicator values

Figure 2.7 summarises the changes to TDI5NGS. Of the 304 taxa in the original TDI5NGS metric, the indicator values of only 23 change by more than 0.5 units (Figure 2.7, right). The majority of the changes involve rare taxa, whose distribution is poorly constrained in the NGS data. The revision for *Gomphonema productum* (GOMP-19) is the result of changes in the LM indicator value. Changes to the other more abundant taxa (*Gomphonema subclavatum* (GOMP-17) and *Nitzschia fonticola* (NI002A)) are the result of unusually high abundances in a few Phase 3 NGS samples and require further investigation.

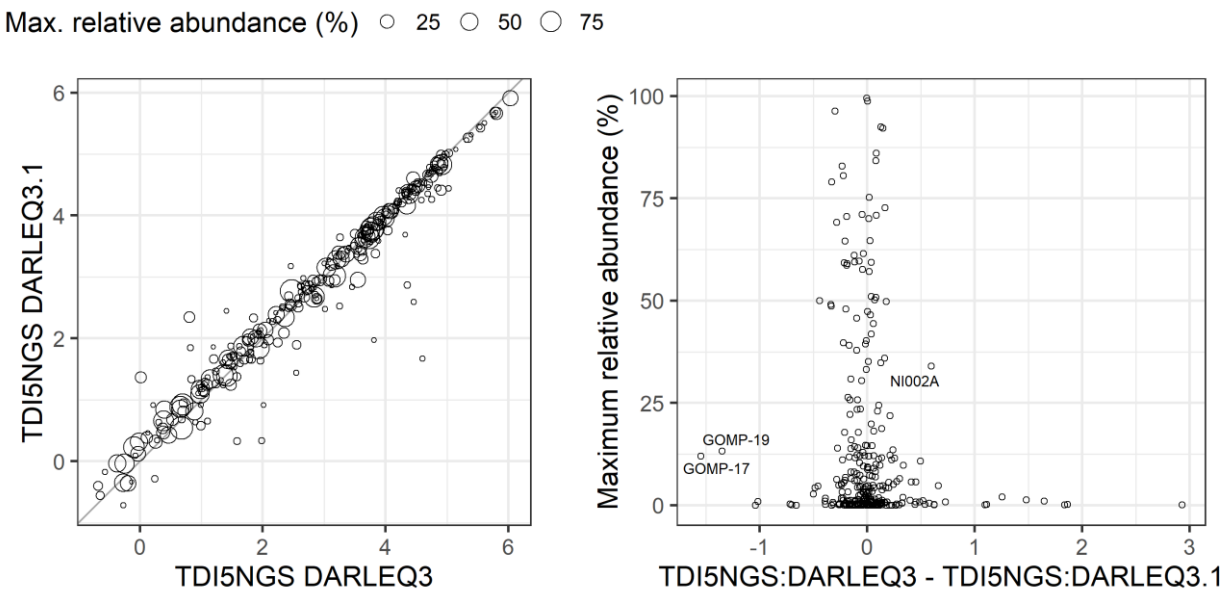

Figure 2.7: Scatter plot of TDI5NGS taxon indicator values for original (DARLEQ3) and updated (DARLEQ3.1) versions of the metric (left) and plot of difference in indicator values vs. maximum relative abundance.

2.3.2 TDI5NGS Phase 3 vs. Phase 2

The effect of the updates to TDI5NGS indicator values on the TDI metric scores is shown in Figure 2.8 and Table 2.9. Overall there is very little difference between the two versions of the metric.

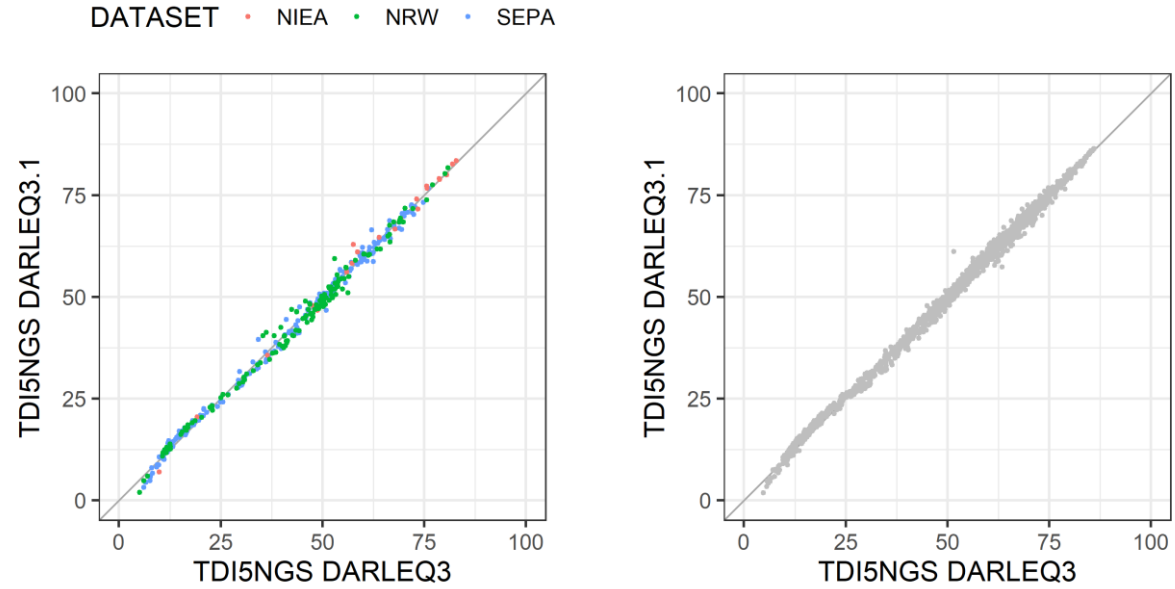

Figure 2.8: Scatter plot of TDI5NGS sample scores for original (DARLEQ3) and revised (DARLES3.1) versions of the metric for Phase 3 (left) and Phase 2 (right) datasets.

Table 2.9: Correlations between TDI5NGS sample scores for original (DARLEQ3) and updated (DARLEQ3.1) versions of the metric.

	Correlation	Concordance correlation
Phase 2	0.998	0.998
Phase 3	0.997	0.997

Although there is little change in TDI5NGS values between the original and update metrics, approximately six percent of samples do change status class (Table 2.10). However, there is little bias in the changes (Table 2.11). Such changes in status are almost inevitable with even small changes to the metric, as it only takes a very small change in a metric to move a sample that is very close to boundary.

Table 2.10: Comparison between ecological status classes for samples computed by original TDI5NGS (columns) and updated TDI5NGS (rows). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes.

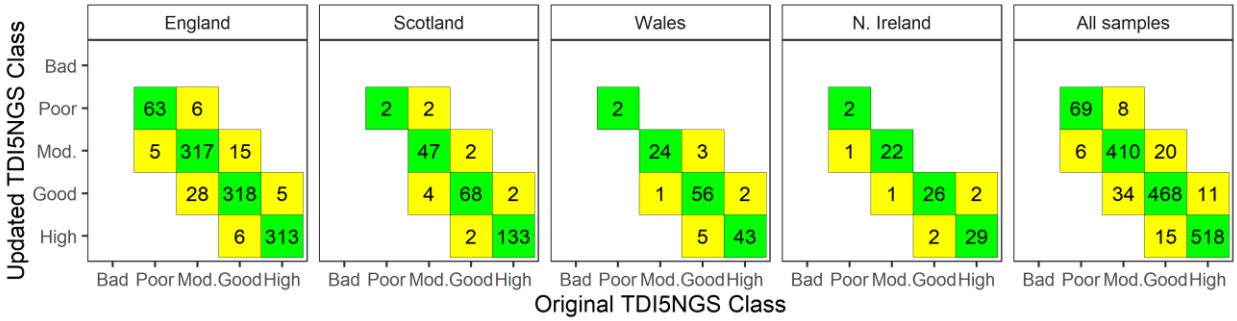

Table 2.11: Summary statistics for classifications presented in Table 2.10.

	England	Scotland	Wales	N. Ireland	All samples
N	1076.0	262.0	136.0	85.0	1559
% Agree	94.0	95.4	91.9	92.9	94
% Agree within one class	100.0	100.0	100.0	100.0	100
% Bias	-1.2	0.0	-0.7	-2.4	-1

This “boundary” effect can be seen in Figure 2.9, which shows the distance from a class boundary for those samples that move class in Table 2.10. The majority of samples that change class were originally less than 0.02 EQR units from a boundary, so would be classified with a high degree of uncertainty.

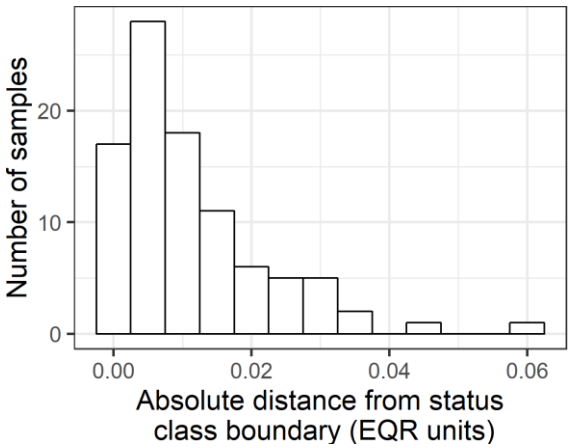

Figure 2.9: Histogram showing, for those samples that change class between original (DARLEQ3) and updated (DARLEQ3.1) TDI5NGS, the distribution of absolute distance to a WFD class boundary in EQR units. Note that 0.1 EQR units is equivalent to half a status class.

2.4 Relationship between TDI5 (LM & NGS) and the nutrient pressure gradient

The ability of TDI4, TDI5LM and TDI5NGS to reflect the nutrient pressure gradient was evaluated using correlations of each model to PO_4-P, NO_3-N, and the first component (PC1) of a principal components analysis of PO_4-P and NO_3-N (PC1). This, in effect, combines the phosphorus and nitrogen gradients into a single pressure variable. Figure 2.10 shows the relationship between these models and pressure gradients whilst Table 2.12 gives the Pearson correlation coefficients for each of these relationships. As shown in the previous report, correlations against PC1 are greater than against nitrogen or phosphorus separately but there is little difference between TDI4 and TDI5LM (~ 1%) and a slight decrease in performance (~6%) when TDI5NGS is used. Table 2.12 also includes correlation coefficients for maximum likelihood response curve (MLRC) models derived using the LM and NGS data. These models reflect a “best-possible” pressure-gradient predictions for the data, assuming taxa follow a unimodal model (ter Braak & Barendregt, 1986). For both LM and NGS data the MLRC models do show a small increase in performance over their TDI counterparts (compare Table 2.12 row 4 with row 2 and row 5 with row 3). However, the improvement of the more complex and less tractable MLRC models is small, indicating that the TDI metrics do faithfully capture most of the statistically explainable variation in diatom turnover along the nutrient pressure gradient in a simple metric.

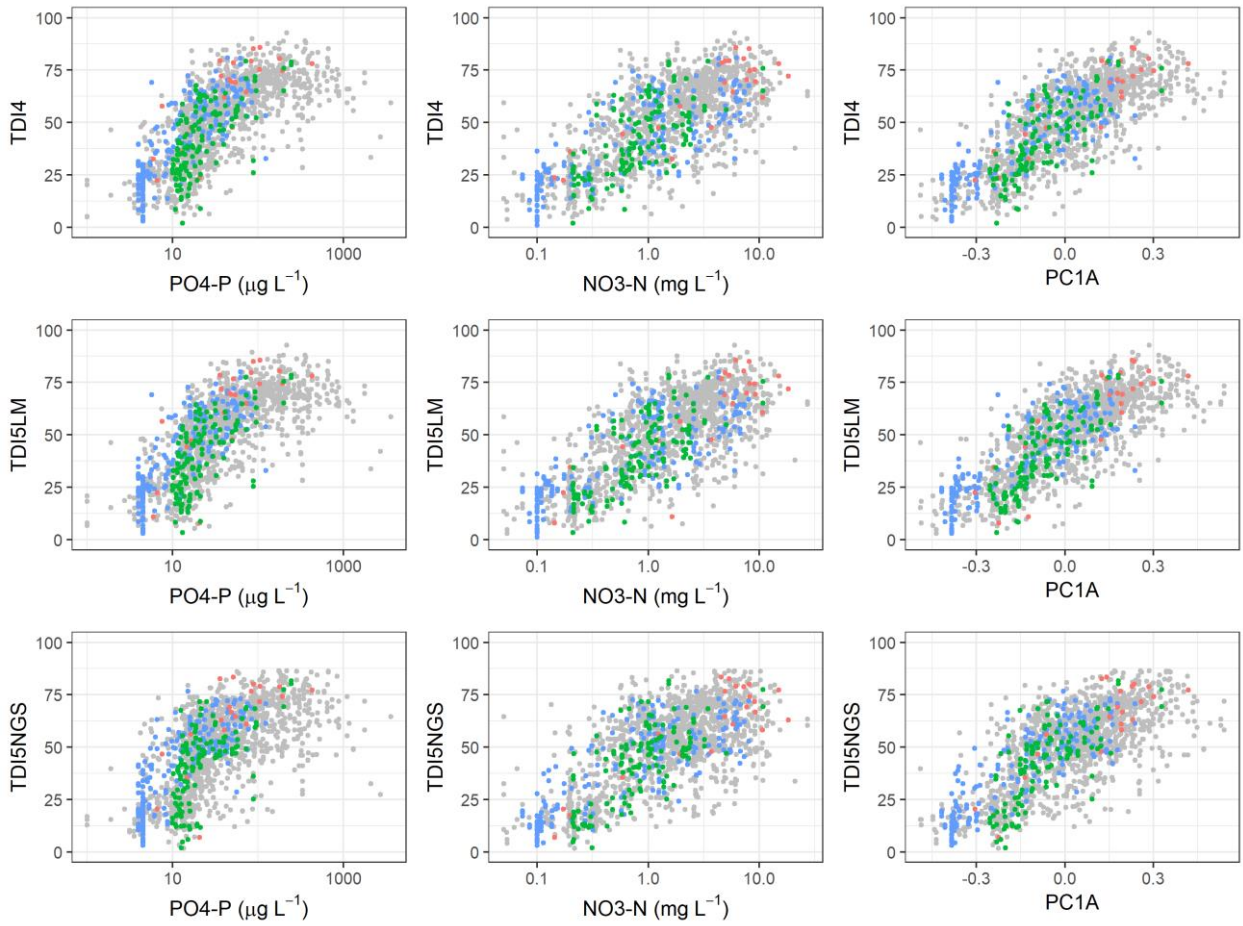

Figure 2.10: Relationship between TDI4 (top), TDI5LM (middle) and TDI5NGS (bottom) and the three nutrient pressure variables. Key to datasets: red=NIEA, green=NRW, blue=SEPA, grey=Phase 2.

Table 2.12: Pearson product-moment correlation coefficients (r) for the relationships shown in Figure 2.10(rows 1-3). Rows 4-5 show correlations between the pressure gradients and diatom-based sample scores using a maximum likelihood response curve (MLRC) model using light (row 4) and NGS (row 5) diatom data. See text for details.

	PO4-P	NO3-N	PC1
TDI4	0.71	0.71	0.77
TDI5LM	0.72	0.72	0.78
TDI5NGS	0.65	0.68	0.72
MLRC_LM	0.78	0.75	0.83
MLRC_NGS	0.71	0.70	0.76

2.5 Implications for classification

There is, now, a strong relationship between the recalibrated TDI5NGS and the two metrics based on LM data, TDI4 and TDI5LM (Figure 2.11, Table 2.13). The next step is to examine the implications of the remaining differences between the metrics on classifications.

Tables 2.14 to 2.19 show the effect on sample level classifications of adopting TDI5LM and TDI5NGS. About two thirds of the samples are classified as high or good status, probably reflecting issues with the current reference model (see below). Overall, the bias between ecological status calculated with the current metric ("TDI4") and TDI5LM or TDI5NGS is slightly negative (-5.6% for TDI5LM, -6.1% for TDI5NGS), indicating that TDI5 classifications are slightly less stringent for both LM and NGS methods. Comparison between TDI5LM and TDI5NGS, by contrast, have a very low level of bias (-1.2%), suggesting that these may be more interchangeable than TDI4 and TDI5NGS. Some of the differences may be related to the shortcomings of the present reference model (discussed in section 3).

Table 2.13: Correlations between TDI5NGS, and TDI4 and TDI5LM sample scores, for Phase 2 and Phase 3 datasets.

	Correlation	Concordance correlation
TDI4 (Phase 2)	0.872	0.872
TDI4 (Phase 3)	0.897	0.895
TDI5LM (Phase 2)	0.877	0.877
TDI5LM (Phase 3)	0.909	0.908

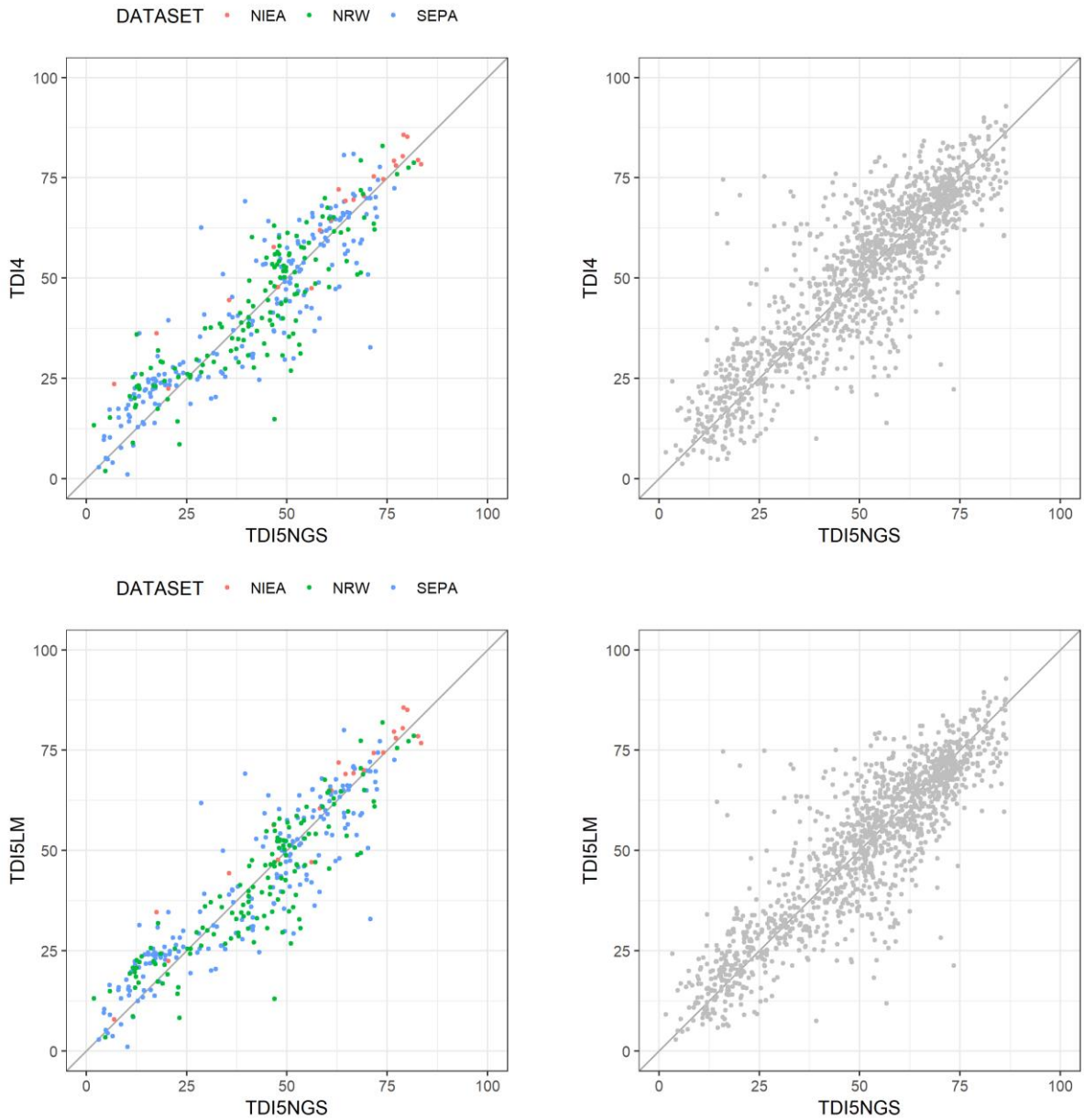

Figure 2.11: Scatter plot of updated TDI5NGS sample score vs. TDI4 scores (top) and updated TDI5LM scores (bottom), for Phase 3 (left) and Phase 2 (right) datasets.

Table 2.14: Comparison between ecological status classes for samples computed by TDI5LM (rows) and TDI4 (columns). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes; red shading: greater than two class difference between methods. Top row: all data; bottom row: samples < 120 mgL^{-1} $CaCO_3$ only.

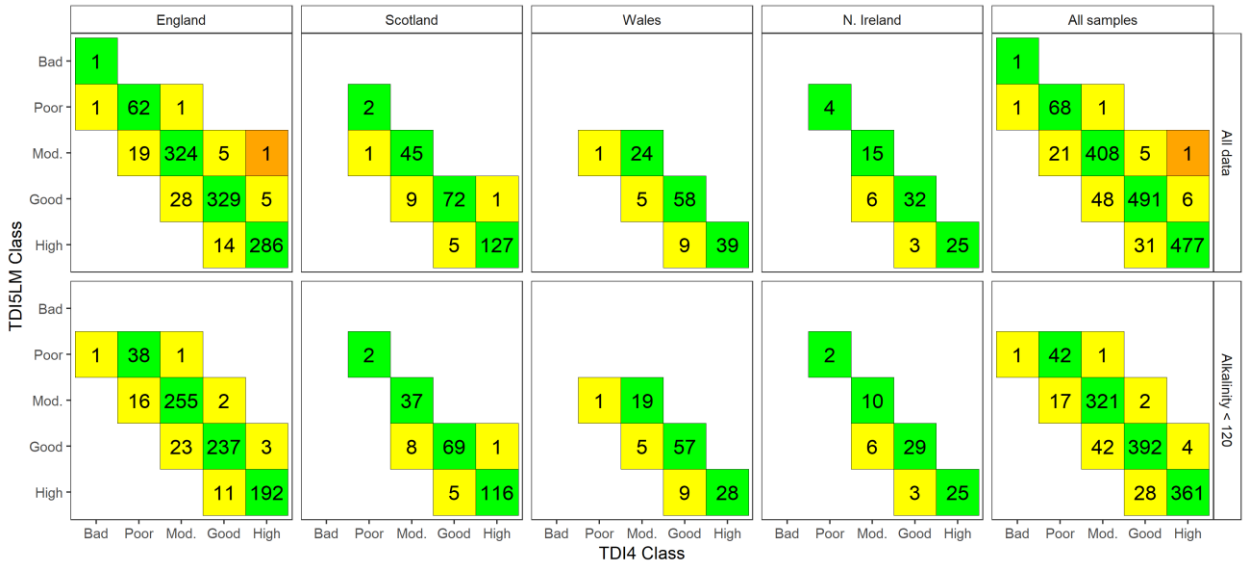

Table 2.15: Summary statistics for classifications presented in Table 2.14.

	England	Scotland	Wales	N. Ireland	All samples
All data					
N	1076.0	262.0	136.0	85.0	1559.0
% Agree	93.1	93.9	89.0	89.4	92.7
% Agree within one class	99.9	100.0	100.0	100.0	99.9
% Bias	-4.6	-5.3	-11.0	-10.6	-5.6
Samples < 120 mgL^{-1} $CaCO_3$					
N	779.0	238.0	119.0	75.0	1211.0
% Agree	92.7	94.1	87.4	88.0	92.2
% Agree within one class	100.0	100.0	100.0	100.0	100.0
% Bias	-5.8	-5.0	-12.6	-12.0	-6.7

Table 2.16: Comparison between ecological status classes for samples computed by TDI5NGS (rows) and TDI4 (columns). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes. Top row: all data; bottom row: samples < 120 mgL^{-1} $CaCO_3$ only.

Table 2.17: Summary statistics for classifications presented in Table 2.16.

	England	Scotland	Wales	N. Ireland	All samples
All data					
N	1076.0	262.0	136.0	85.0	1559.0
% Agree	58.2	67.9	58.8	61.2	60.0
% Agree within one class	95.5	96.9	97.8	98.8	96.2
% Bias	-6.1	-4.6	-7.4	-8.2	-6.1
Samples < 120 mgL^{-1} $CaCO_3$					
N	779.0	238.0	119.0	75.0	1211.0
% Agree	59.3	69.3	54.6	61.3	60.9
% Agree within one class	96.3	97.9	97.5	98.7	96.9
% Bias	-4.5	-4.6	-8.4	-6.7	-5.0

Table 2.18: Comparison between ecological status classes for samples computed by TDI5NGS (rows) and TDI5LM (columns). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes; red shading: greater than two class difference between methods. Top row: all data; bottom row: samples < 120 mgL^{-1} $CaCO_3$ only.

Table 2.19: Summary statistics for classifications presented in Table 2.18.

	England	Scotland	Wales	N. Ireland	All samples
All data					
N	1076.0	262.0	136.0	85.0	1559.0
% Agree	60.0	66.8	64.0	63.5	61.7
% Agree within one class	96.1	96.6	97.8	100.0	96.5
% Bias	-2.2	0.4	2.2	1.2	-1.2
Samples < 120 mgL^{-1} $CaCO_3$					
N	779.0	238.0	119.0	75.0	1211.0
% Agree	61.6	68.1	60.5	64.0	62.9
% Agree within one class	96.4	97.5	97.5	100.0	96.9
% Bias	0.6	0.0	2.5	4.0	0.9

3 Development of an alternative reference model

3.1 Why is a new reference model and combination rule needed?

The understanding of the Macrophyte and Phytobenthos biological quality element has evolved considerably during the lifetime of the Water Framework Directive, with an inevitable tension between science and pragmatism as two quite different components of the aquatic biota are evaluated and combined to produce the final assessment. Alongside a growing understanding of how each sub-element responds to nutrients, recognition that one sub-element is better suited to some types of river than the other has, along with financial constraints, led in many cases to the state of the overall BQE (as defined by the WFD) being inferred from just one sub-element. Whilst there is now general recognition that the current diatom reference model does not provide accurate assessments at high alkalinity, practical considerations mean that a proposal that necessitated the use of both sub-elements at all sites would not be welcome. Such considerations need to be recognised, but our work raises issues

about the interpretation of WFD status as assessed by the diatom and macrophyte tools which may require a re-evaluation of the way these are combined, if not now in the longer term.

Weaknesses with the current reference model are described below (3.2). In brief, the lack of suitable reference sites at high alkalinity led to adoption of an alternative view of the reference condition that is, as far as we can tell, effective for macrophytes but has proved to be less so for diatoms. This led us to explore different reference concepts for diatoms, leading to the new reference model, described in 3.3 and explained in greater detail in 3.5. However, this new reference model was not only more stringent at high alkalinity than the current model, but it was also more stringent than the macrophyte reference model. This means that, using the current practice of defining status as the worst of the two sub-elements, not only would the proportion of sites that achieve at least good status for macrophytes and phytobenthos be lower but the contribution that macrophytes made to these status assessments would be much reduced, despite their significant contributions to the structure and functioning of river ecosystems. Put simply, the new diatom reference model provides more reliable evidence of nutrient enrichment, but if the macrophyte classification does not corroborate this we should perhaps interpret this as less certain evidence of a failure to achieve good status of the overall macrophyte and phytobenthos quality element. The rapidly responding algal assemblage has changed sufficiently to indicate an issue, but the slower structural component has not. If we consider both tools to provide evidence of equal value and we assume similar levels of uncertainty, combining their results by averaging should provide a more robust assessment or ecological status.

If this interpretation of ecological status is correct then shifting from the use of the worst of the two sub-elements to their average is the obvious conclusion. It is important to recognize that the “one out, all out” rule is a requirement of the WFD when comparing BQEs but is not obligatory when combining sub-elements within a BQE. There are, indeed, precedents within the UK monitoring toolkit (averaging is used in the lake phytoplankton tool, for example) and we believe that there are good reasons why averaging offers a more realistic view of the condition of the macrophyte and phytobenthos BQE than use of the worst of the two sub-elements.

However, in practice, both sub-elements are not assessed at all sites and, for this reason, a final adjustment needs to be included which allows the value of the combined BQE to be predicted from either one of the sub-elements. Whilst differences between macrophytes and phytobenthos at individual sites may convey extra information on nutrient loads that is missed by routine chemical monitoring, the reality is that, with current levels of resources, the possibility of inferring the condition of the BQE from a single sub-element will give managers more choices. We have, therefore, considered these options in 3.6.

Changing combination rules should provide a more realistic overall assessment of status. However, to remain within the spirit of the intercalibration process we suggest a final subtraction of 0.05 EQR units to ensure that the *overall* level of precaution does not change. While this final subtraction gives the impression of a “fiddle factor”, our argument is that at the *water body* level averaging of metrics is the most robust approach. Since this amounts, *on average*, to a slight reduction in stringency we advocate compensating for this by subtracting the 0.05 EQR unit (or by increasing the boundary values of both sub-metrics by 0.05). This will ensure that the UK approach has a similar level of ambition to that of methods from elsewhere in Europe.

3.2 Performance of the current diatom reference model

The current method (implemented in DARLEQ3 and DARLEQ2 software) uses a model to determine reference TDI which was developed from a set of reference sites (Kelly *et al*., 2013). Many of the reference sites at low and moderate alkalinity fulfilled ECOSTAT screening criteria (Pardo et al., 2012); however, no high alkalinity sites fulfilled these criteria and thus the reference TDI values at higher alkalinities were an extrapolation. Expected TDI values, particularly at high alkalinity, were higher than those computed using the original TDI reference model (Kelly *et al*., 2008) but were often associated with sites that had rich macrophyte floras corresponding to high ecological status. Both models were derived from alkalinity which is considered to represent natural fertility and the modelled reference TDI values are illustrated in Figure 3.1. At higher alkalinity values the current reference TDI value is greater than many of the observed points, suggesting that many of these sites would have a diatom EQR value exceeding 1 and thus be at high status. This effect can clearly be seen when the TDI EQR values are modelled using a multivariate GAM including soluble nutrients (PO_4-P and NO_3-N) and alkalinity. TDI EQR has a clear response to phosphorus and a weaker response to nitrate-nitrogen but there is a marked increase of EQR at higher alkalinity values (Figure 3.2). By comparison, the same relationships for macrophytes (Figure 3.3) show a much smaller EQR response to alkalinity. The consequence of this on EQR values for these two BQEs is shown in Figure 3.4, with diatoms providing more stringent classifications at lower alkalinity (< 50 mg $CaCO_3$ L^{-1}), but less stringent at high alkalinity (> 125 mg $CaCO_3L^{-1}$) (Figure 3.5).

These differences in classification were recognised by Kelly *et. al*. (2013) and were the reason that it was recommended that the minimum of these two EQRs were used for classification when referring to the combined BQE. Where only a single component of the BQE was used it was suggested that this should be diatoms in low alkalinity rivers, but macrophytes in very high alkalinity rivers and that both should be used when the alkalinity was moderate (75-125 mg $CaCO_3$ L^{-1}). However, macrophyte assessment is not appropriate or possible in some high alkalinity rivers (small streams, ditches, canalized rivers etc.) and in these situations diatom assessments could be useful. This suggests that a further review of the reference TDI would be beneficial.

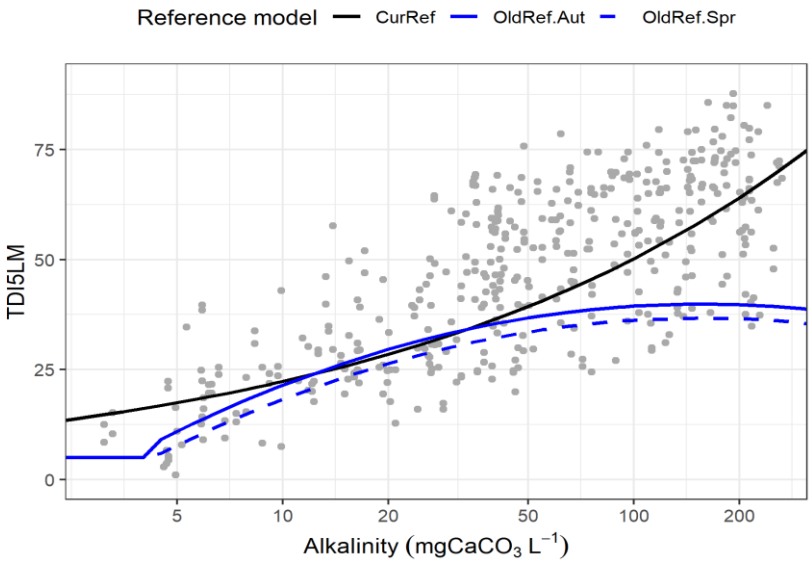

Figure 3.1: Relationship between observed TDI5LM (light microscope) and alkalinity, with the current (DARLEQ2 & 3 software: black line) and original (DARLEQ 1

software: reference models superimposed (blue line spring = dashed, autumn = solid).

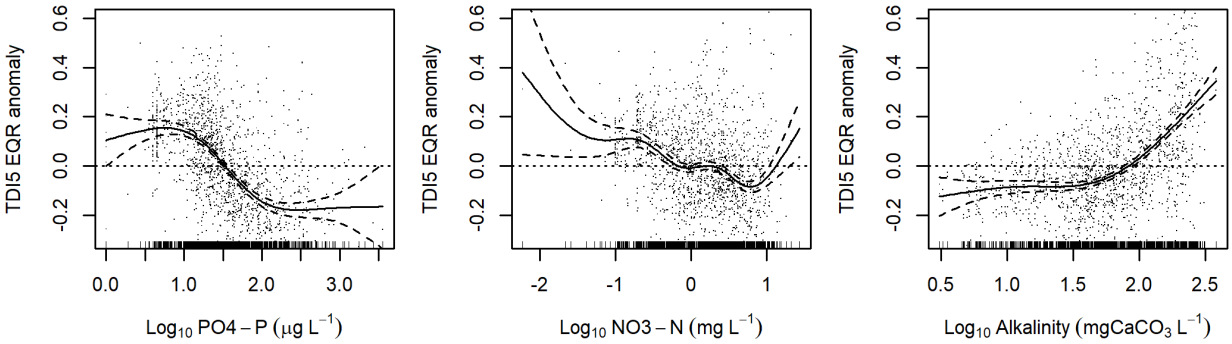

Figure 3.2: Relationship between TDI5LM EQR (light microscope) and soluble reactive phosphorus, nitrate-nitrogen and alkalinity, showing GAM smooths. EQR values are relative to the overall mean, points show distribution of residuals.

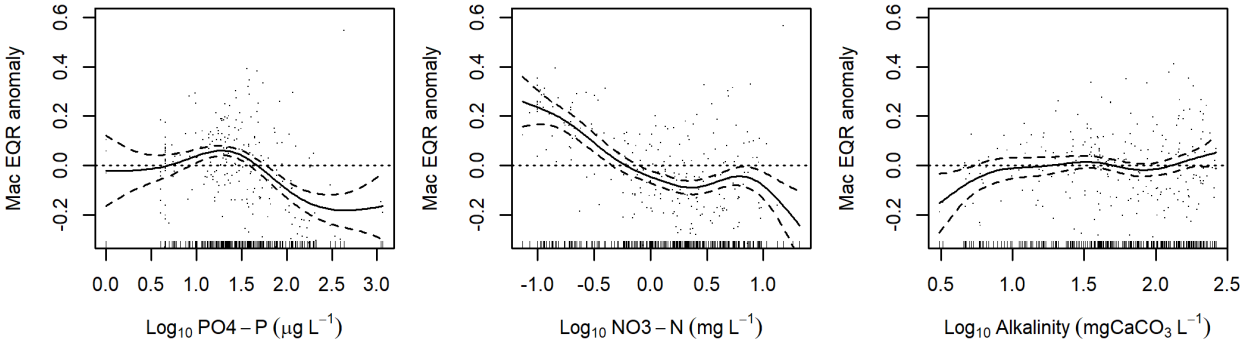

Figure 3.3: Relationship between LEAFPAC macrophyte final EQR and soluble reactive phosphorus, nitrate-nitrogen and alkalinity, showing GAM smooths. EQR values are relative to the overall mean, points show distribution of residuals.

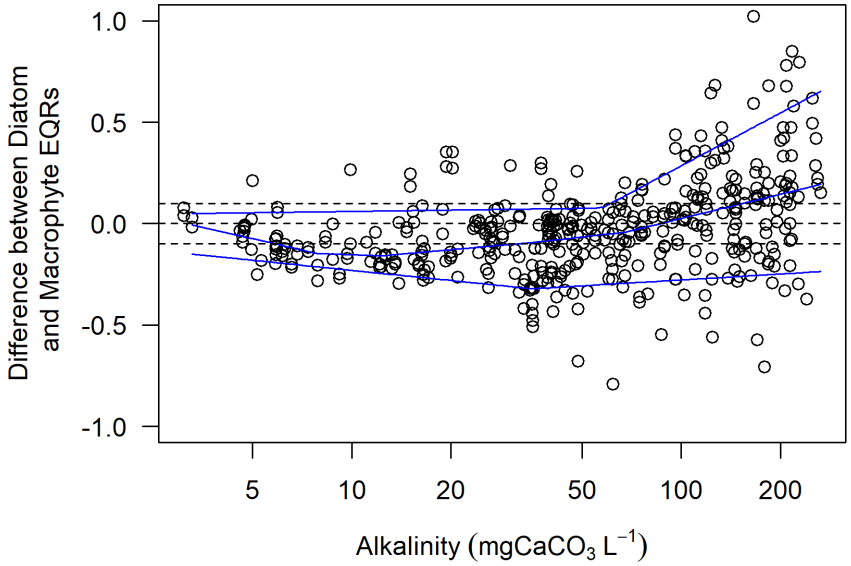

Figure 3.4: Difference between diatom (TDI5) and macrophyte EQR values plotted against alkalinity. Blue lines show regressions fitted to 90th, 50th and 10th quantiles, dotted lines mark an EQR of ±0.1 (a WFD class) and the zero value (no class difference).

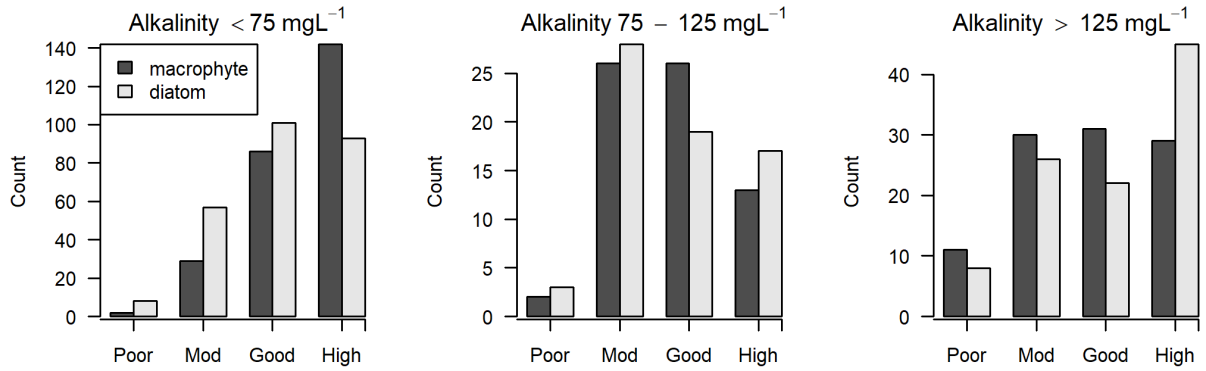

Figure 3.5: Comparison of the number of samples classified by macrophyte and diatom at low, moderate and high alkalinity, comparing individual diatom sample classifications with macrophyte survery classification.

3.3 An alternative reference model for TDI

It is clear that selecting reference sites in rivers is a difficult process and inevitably results in few sites from lowland higher alkalinity rivers. An alternative strategy is to fit regression models to a sub-set of sites that have the lowest observed TDI values for a given alkalinity. This can be done by fitting a regression to a lower quantile (e.g. 10th quantile) of the relationship between observed TDI and alkalinity. Alkalinity is correlated to the soluble nutrient concentration (Figure 3.6) and by fitting a regression between alkalinity and a lower quantile of TDI we allow for the effect of an increasing background (natural) phosphorus. However, alkalinity is also correlated with nitrate-nitrogen (Figure 3.6 right). Although background phosphorus is likely to be correlated with alkalinity as sources of both are related to catchment geology this is unlikely to be true for background nitrogen, which is likely to be low across the range of alkalinity. To allow for this we include nitrate-nitrogen concentration as a predictor variable in a quantile regression.

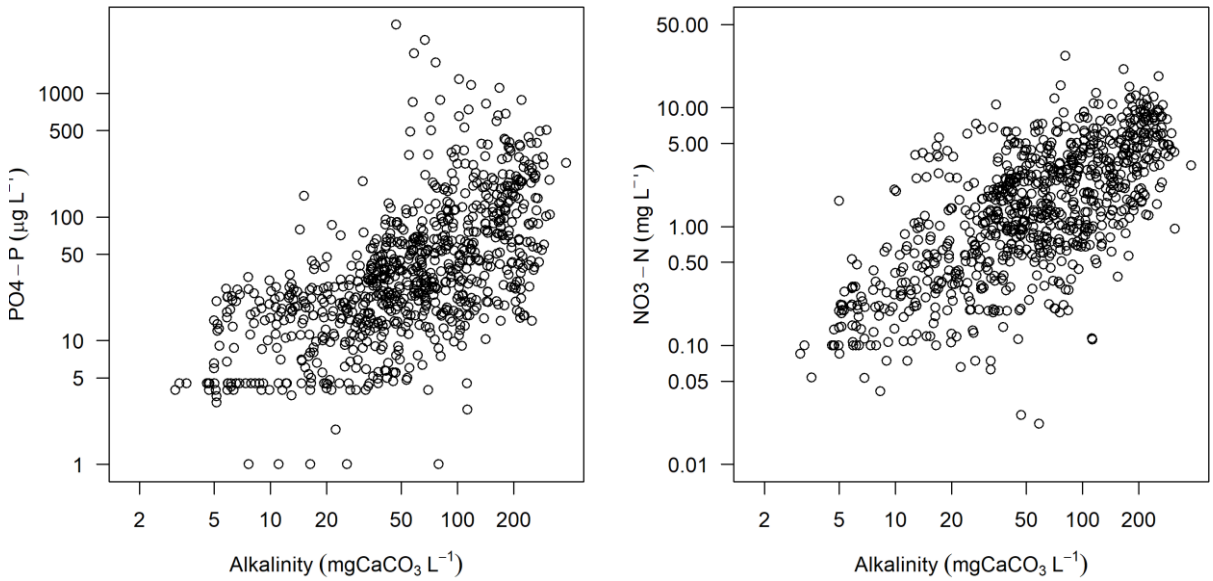

Figure 3.6: Relationship of (left) soluble reactive phosphorus and (right) nitrate with alkalinity.

A quantile regression was fitted using the R package quantreg (Koenker, 2017) for the 25th quantile using the log$_{10}$ of alkalinity and nitrate nitrogen as predictor variables

(Model 1) and, additionally, including sample season as a categorical variable (spring = 0, autumn = 1) with season split before/after July (Model 2), as this was found to be a significant variable in the original diatom reference model (Kelly *et al*., 2008).

Both models show highly significant effects of alkalinity and nitrate-nitrogen, and model 2 showed a just significant effect of season ($p = 0.03$) (Table 3.1). The resulting models are shown in Table 3.1 and Fig. 3.7, together with the current reference model and the original seasonal reference models. These parameters were then used to predict reference TDI values, taking a nitrate-nitrogen concentration of 0.5 mgL^{-1} a value assumed to be consistent with reference conditions across the range of alkalinity (Pardo *et al*., 2012).

Table 3.1: Quantile regression model outputs for revised reference model.

Model 1

term	estimate	std.error	statistic	p.value
(Intercept)	7.216	2.025	3.563	0.0003782
Log10 Alkalinity (mg/L CaCO3)	18.9	1.231	15.35	0
Log10 NO3-N (mg/L)	15.15	1.013	14.95	0

Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	5.061	2.105	2.404	0.01633
Log10 Alkalinity (mg/L CaCO3)	19.69	1.239	15.89	0
Log10 NO3-N (mg/L)	14.95	1.008	14.83	0
Season	1.856	0.8629	2.151	0.03165

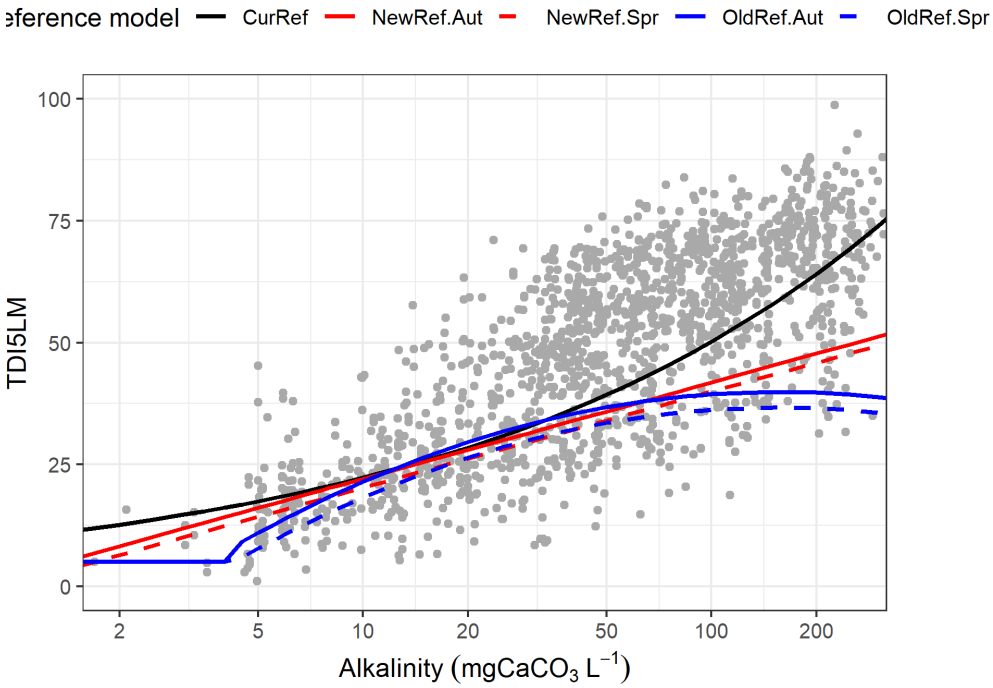

Figure 3.7: Modelled reference TDI values overlain on scatter plot of observed TDI. Black line = current reference model, red line = new reference model (spring dashed, autumn solid, blue line = original reference model (spring dashed, autumn solid).

3.4 EQR using the new reference model

The EQR for TDI5 was calculated using the new reference model, including season as a predictor, values were normalised as for current EQRs by multiplication by 0.8. Fitting a GAM model including nutrients and alkalinity demonstrates that the new reference model has removed the effect of alkalinity on EQR (Figure 3.8, Table 3.2) and should thus be a more reliable value.

However, the overall result on classification is that the diatom method would now be the most stringent sub-element across all alkalinity classes, while using the present method this is only likely to be the case for low alkalinity. This is considered in more detail in 3.7.

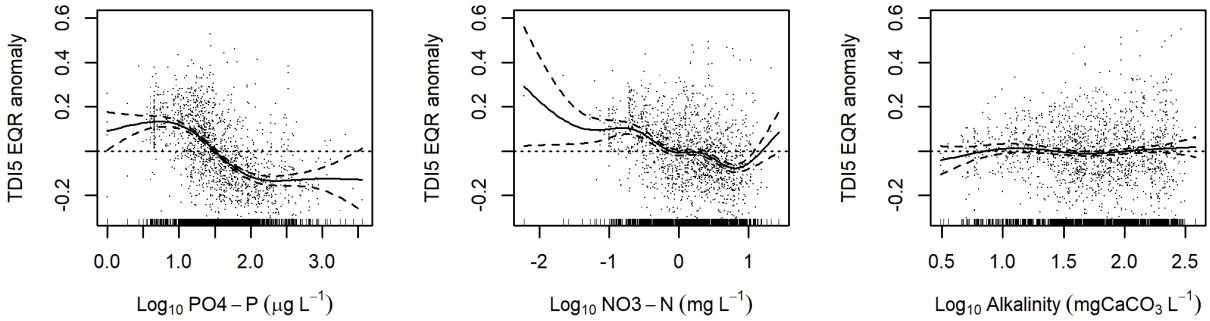

Figure 3.8: Relationship between TDI5LM EQR (light microscope) using the new reference TDI model (with season) and soluble reactive phosphorus, nitrate nitrogen and alkalinity, showing GAM smooths. EQR values are relative to the overall mean, points show distribution of residuals.

Table 3.2: GAM model for TDI5 EQR using new reference TDI model against nutrients and alkalinity.

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6162191	0.0034404	179.1151	0

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(LogP)	4.760637	5.493949	56.823940	0.0000000
s(LogN)	5.820506	5.983477	17.999707	0.0000000
s(LogAlk)	3.839166	4.632530	1.721545	0.1217385

3.5 Justification for new diatom reference model

“Macrophytes and Phytobenthos” includes photosynthetic organisms with a wide range of growth strategies. Trying to reconcile differences in classifications produced by macrophytes and phytobenthos needs some recognition of how these respond at different spatial and temporal scales.

“Macrophytes” encompass a range of growth forms, including filamentous algae, mosses, free floating and rooted vascular plants, the latter including species that are wholly-submerged and emergent. There is also a range of sizes, from a few millimetres to greater than a metre. Macrophytes exploit a range of habitats within a stream, some growing directly on rocks, whilst others are rooted in fine or coarse sediments. Life-cycles range from a few weeks (in the case of some of the algae) to a year or longer, in the case of vascular plants. This means that the macrophyte assemblage as a whole is exposed to a variety of sediment and water column nutrient pools, and respond to change at different temporal scales.

“Phytobenthos”, on the other hand, is sampled from a single habitat (biofilms on rocks and/or plant surfaces). The assemblage is dependent primarily on water column nutrients, and individual organisms are smaller and shorter-lived. Studies have shown that the diatom assemblage is shaped by in-stream nutrient and hydrology conditions over the preceding two to three weeks (Lavoie *et al*., 2008; Snell *et al*., 2014).

It is important to acknowledge these differences in order to develop a robust approach to dealing with the combined “Macrophyte and Phytobenthos” BQE. They also help to explain the problems encountered with the present approach, in which

high alkalinity reference sites for phytobenthos were selected using expert judgement based on an understanding of the macrophyte communities. Rich macrophyte communities will be better-buffered against consequences of occasional nutrient pulses than phytobenthos and we believe that using phytobenthos from such sites led to inflated predictions of expected TDI in high alkalinity rivers.

This leaves the problem of how reference conditions for phytobenthos should be set in high alkalinity rivers. Having exhausted other options, we have adopted a new approach based on the “best available” results obtained from ongoing monitoring. The lower edge of the data cloud produced when TDI is plotted against alkalinity, regardless of stressor state, should indicate the best possible conditions that are encountered. That the current reference model follows a line closer to the median, particularly at high alkalinity, suggests a problem with this model. We have, therefore, fitted a new relationship to this data cloud using quantile regression. The result is a model that is more stringent than the current one, particularly at high alkalinity but it is a better reflection of the state of the data.

However, this means that we now have different reference concepts for the two sub-elements within a single BQE. Can this be justified? Given the differences between macrophytes and phytobenthos, different responses to pressure are to be expected and this will extend to the appropriate variables used to screen reference sites. In particular, the sensitivity of phytobenthos to nutrients at a temporal scale finer than that used for routine monitoring raises issues about the use of a chemical screening threshold that cannot be supported by land use screening criteria.

All of our work to date suggests that a significant change in community composition occurs at lower nutrient concentrations for phytobenthos than it does for macrophytes. Therefore, the way in which the two sub-elements are combined into the final BQE is critical. Using the “one out all out” rule **within** the macrophytes and phytobenthos BQE with a stringent diatom model will lead to some high alkalinity sites (such as chalk streams) failing to achieve GES despite other, more conspicuous elements of the biota (invertebrates, fish, macrophytes) being at high or good status. In particular, this does not acknowledge the health of half of the BQE or recognize the basic biological differences between the sub-elements. Averaging the sub-elements means that information from both sub-elements contributes to the final decision, and “one out, all out” still applies **between** BQEs. These possibilities are developed in the next sections.

3.6 Implications for phytobenthos classification

The following tables show the effect on classifications of revising the reference model on classifications produced using TDI4, TDI5LM and TDI5NGS. In all cases, the bias from changing the reference model is greater than that between the different variants of TDI (summarized in section 2.5) with about a third of samples moving to a lower status class (Tables 3.3 to 3.8). This effect is most pronounced in high alkalinity waters and, consequently, has a greater effect on classifications in England than in other parts of the UK.

Table 3.3: Comparison between ecological status classes for samples computed by TDI4 with current reference model (rows) and TDI4 using the revised reference model (columns). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes.

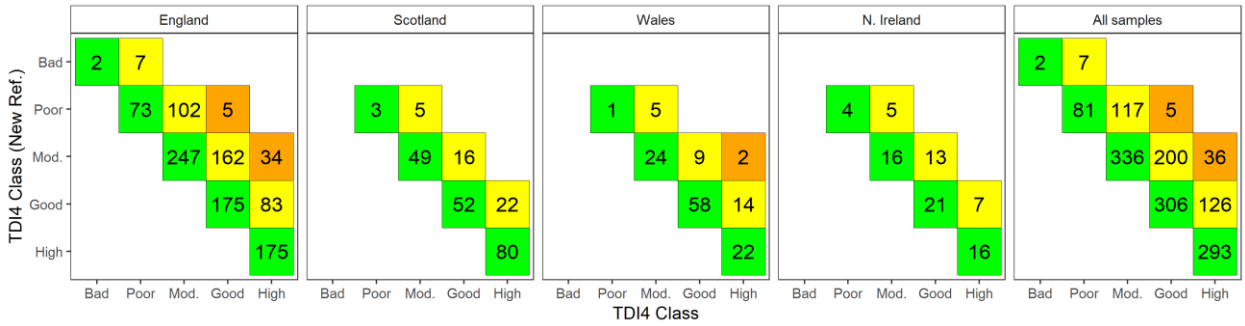

Table 3.4: Summary statistics for classifications presented in Table 3.3.

	England	Scotland	Wales	N. Ireland	All samples
N	1065.0	227.0	135.0	82.0	1509.0
% Agree	63.1	81.1	77.8	69.5	67.5
% Agree within one class	96.3	100.0	98.5	100.0	97.3
% Bias	-36.9	-18.9	-22.2	-30.5	-32.5

Table 3.5: Comparison between ecological status classes for samples computed by TDI5LM with current reference model (rows) and TDI5LM using the revised reference model (columns). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes.

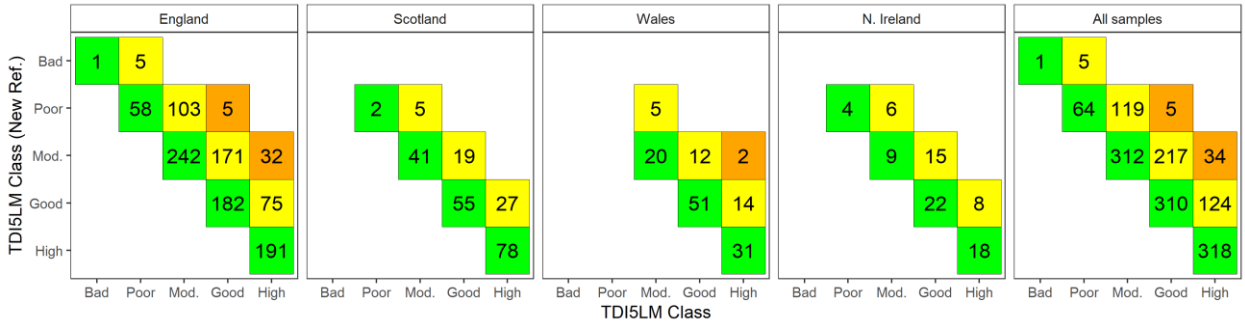

Table 3.6: Summary statistics for classifications presented in Table 3.5.

	England	Scotland	Wales	N. Ireland	All samples
N	1065.0	227.0	135.0	82.0	1509.0
% Agree	63.3	77.5	75.6	64.6	66.6
% Agree within one class	96.5	100.0	98.5	100.0	97.4
% Bias	-36.7	-22.5	-24.4	-35.4	-33.4

Table 3.7: Comparison between ecological status classes for samples computed by TDI5NGS with current reference model (rows) and TDI5NGS using the revised reference model (columns). Green shading: identical classification for both metrics; yellow shading: agreement to within one class; orange shading: agreement to within two classes.

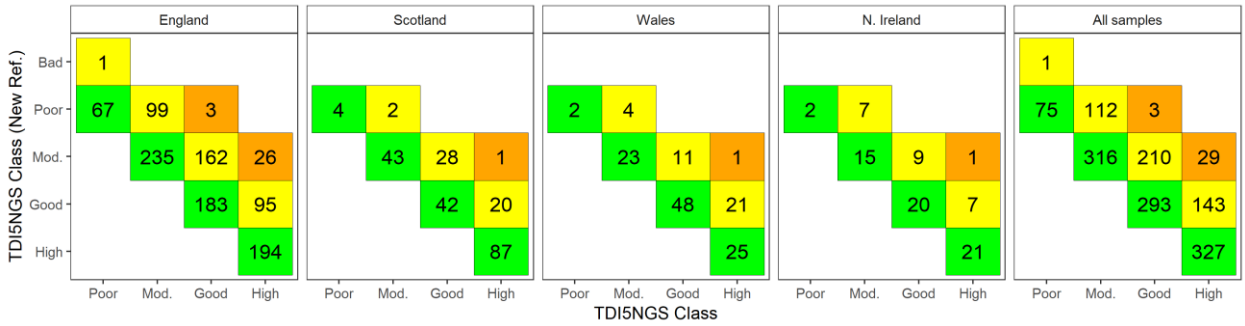

Table 3.8: Summary statistics for classifications presented in Table 3.7.

	England	Scotland	Wales	N. Ireland	All samples
N	1065.0	227.0	135.0	82.0	1509.0
% Agree	63.8	77.5	72.6	70.7	67.0
% Agree within one class	97.3	99.6	99.3	98.8	97.9
% Bias	-36.2	-22.5	-27.4	-29.3	-33.0

3.7 Implications for phytobenthos / macrophyte combination

3.7.1 Combination rules

The current combination rule for macrophytes and diatoms is to take the worst of the two EQR values to determine the overall macrophyte and phytobenthos BQE classification. This was the only logical approach, given the different relative levels of their EQRs. In practice, diatoms tended to be more stringent at low alkalinity and macrophytes at high alkalinity (Figure 3.9 left). However, the new diatom reference model shifts this balance, leading consistently more stringent classifications being obtained using diatoms across the entire alkalinity range (Figure 3.9 right). This means that, in effect, macrophytes will rarely determine final classifications and, in theory, have less direct relevance to the river basin management process, if one continues to apply the current combination rule. There is a case, therefore, for re-examining the manner in which results from macrophytes and diatoms are combined and, in particular, to consider whether averaging the metrics might provide a better approach.

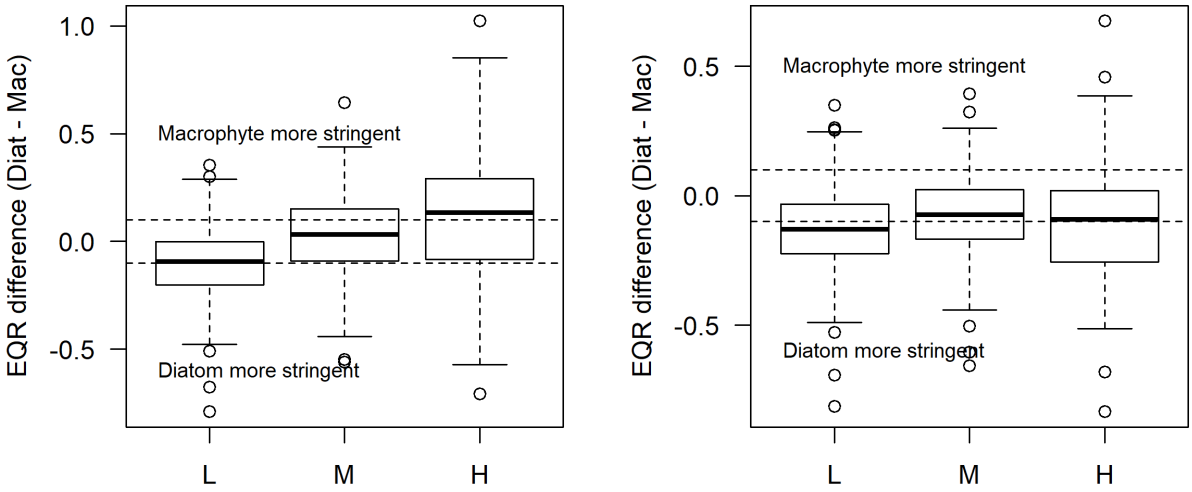

Figure 3.9: Difference between diatom and macrophyte EQR values using (left) current TDI reference and (right) new TDI reference, split by alkalinity type (L = <75 $mgCaCO_3L^{-1}$, M = 75-125 $mgCaCO_3L^{-1}$, >125 $mgCaCO_3L^{-1}$). Horizontal lines mark ±0.1 EQR units i.e. 1 WFD class.

Comparing the relationships between each of the metric EQRs, the average and the worst of either metric EQR with PO_4-P concentration (Figure 3.10) and P EQR (Figure 3.11) clearly demonstrates that the new diatom reference model has a better relationship with phosphorus gradient than the current model (linear model r^2 0.301 compared to r^2=0.107) whilst the average of the new diatom EQR and macrophyte EQR gives the strongest relationship of all (r^2=0.351), although this is only slightly different to that obtained from the worst of diatom and macrophyte EQRs (r^2=0.326). [P EQR is defined as the ratio between "observed" annual mean phosphorus and the phosphorus concentration expected in the absence of human alteration to a catchment, modelled from site altitude and alkalinity. This approach was used to derive the present river phosphorus standards for the UK.]

Table 3.9: Linear model predicting Average EQR from Diatom TDI5LM EQR and log_{10} Alkalinity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3426	0.02585	13.25	8.829e-34
TDI5LM_EQR	0.6988	0.02326	30.05	4.336e-107
Log10_Alkalinity	-0.05446	0.009041	-6.024	3.689e-09

Table 3.10: Linear model predicting Average EQR from Macrophyte final EQR and log_{10} Alkalinity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2694	0.0266	10.13	9.684e-22
Macrophyte_EQR	0.6849	0.02162	31.67	8.36e-114
Log10_Alkalinity	-0.05329	0.008706	-6.122	2.111e-09

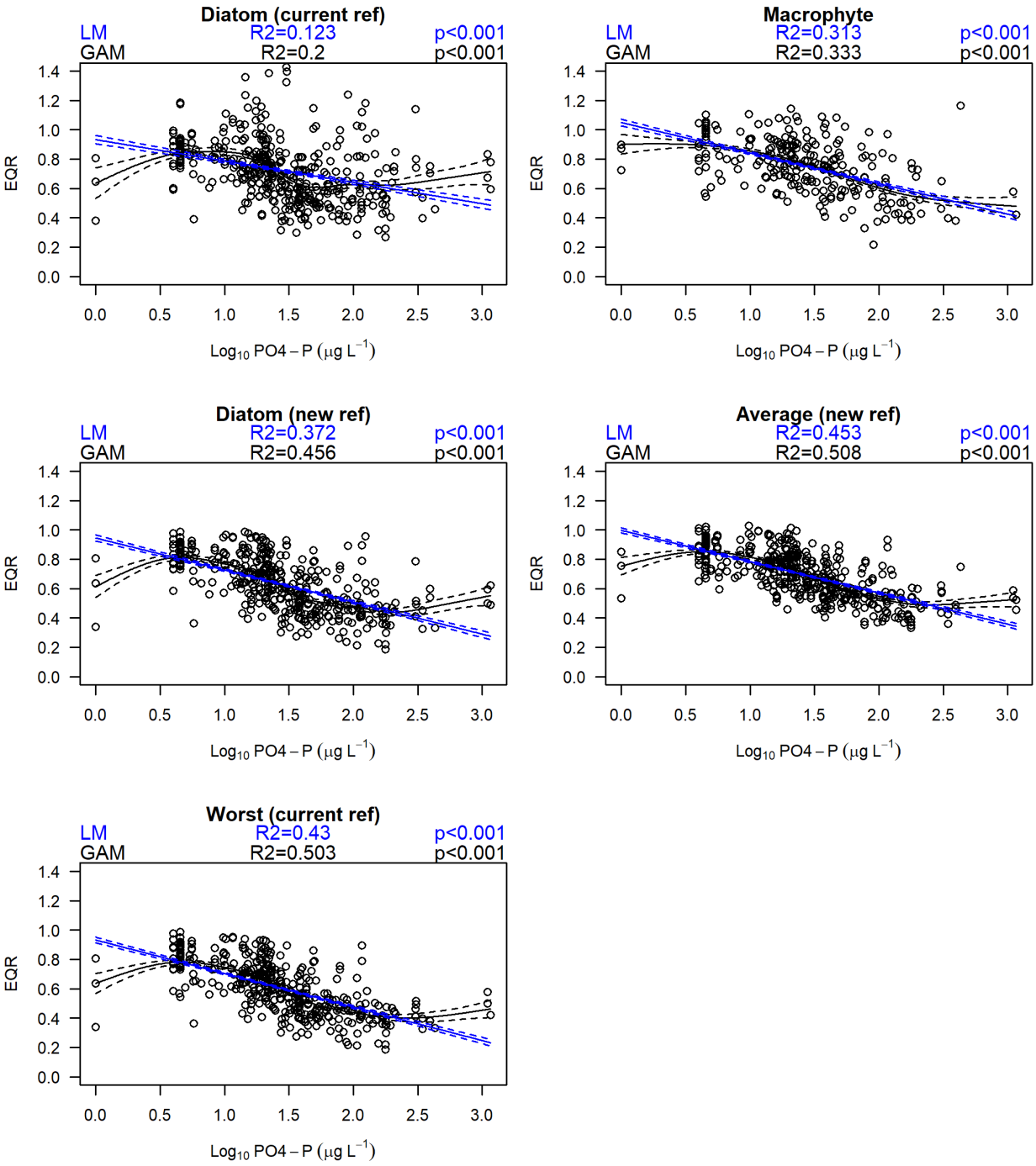

Figure 3.10: Relationship between EQR and SRP concentration, with GAM model fit.

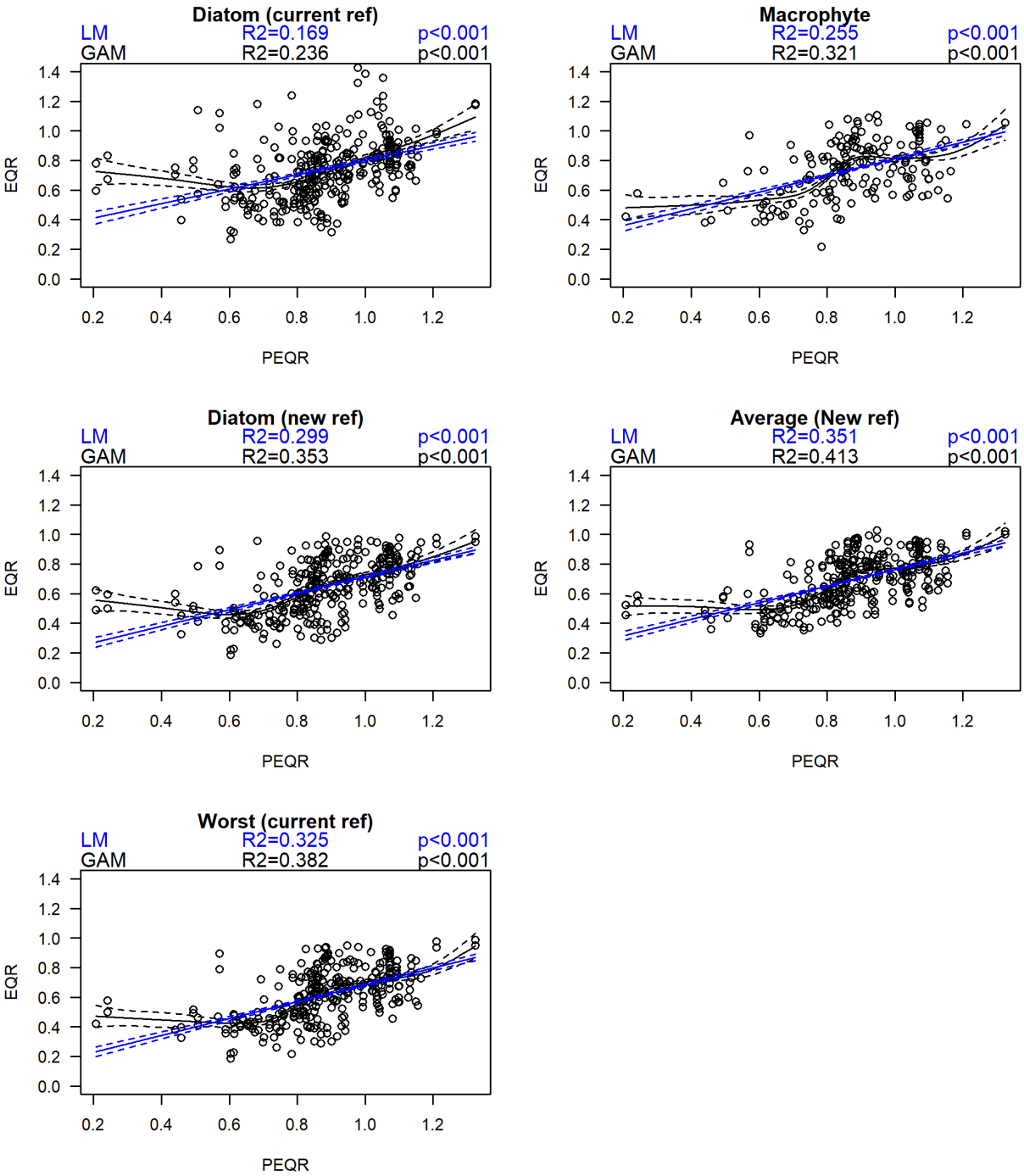

Figure 3.11: Relationship between EQR and P-EQR concentration, with GAM model fit.

3.7.2 Prediction of average EQR from single metrics

The average EQR of the new diatom and macrophyte EQRs will be lower than one of the two individual metrics, typically macrophytes as diatoms are, on average, now more stringent. The extent that either sub-element departs from the average of the two can be estimated by modelling the average EQR from the individual metric EQRs and alkalinity (Table 3.9, Table 3.10, and Figure 3.12). On average the Diatom EQR values are increased by 0.06 EQR units and the macrophyte EQR values are

decreased by 0.06 EQR units (slightly more than a quarter of a class), though the extent of the change depends upon the position along the gradient (Figure 3.13).

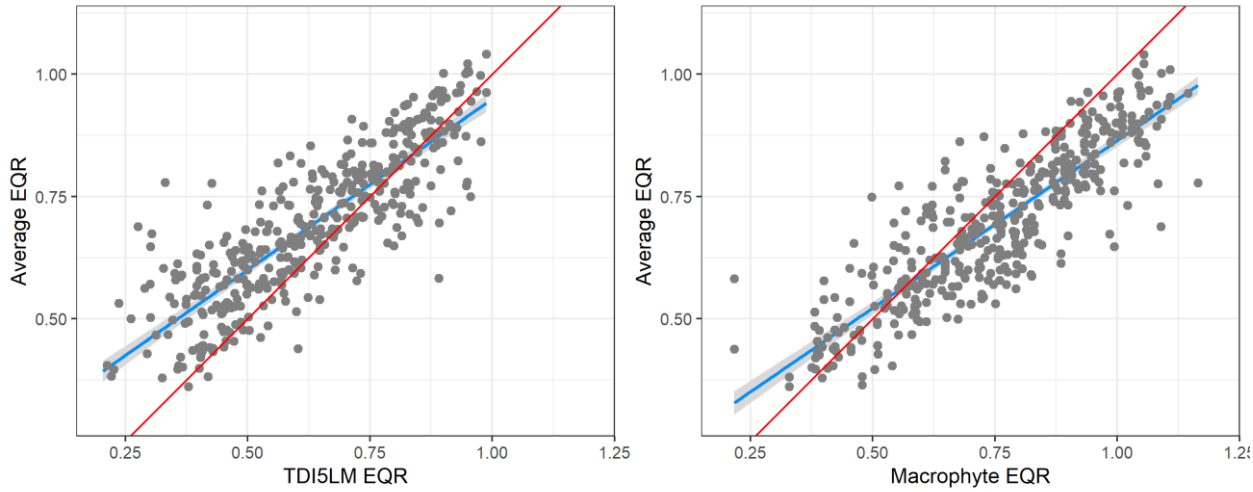

Figure 3.12: Conditional regression plots, showing relationship between Average EQR and TDI5LM-EQR (left) and Macrophyte-EQR (right)) for models listed in Tables 3.9 and 3.10 at median alkalinity.

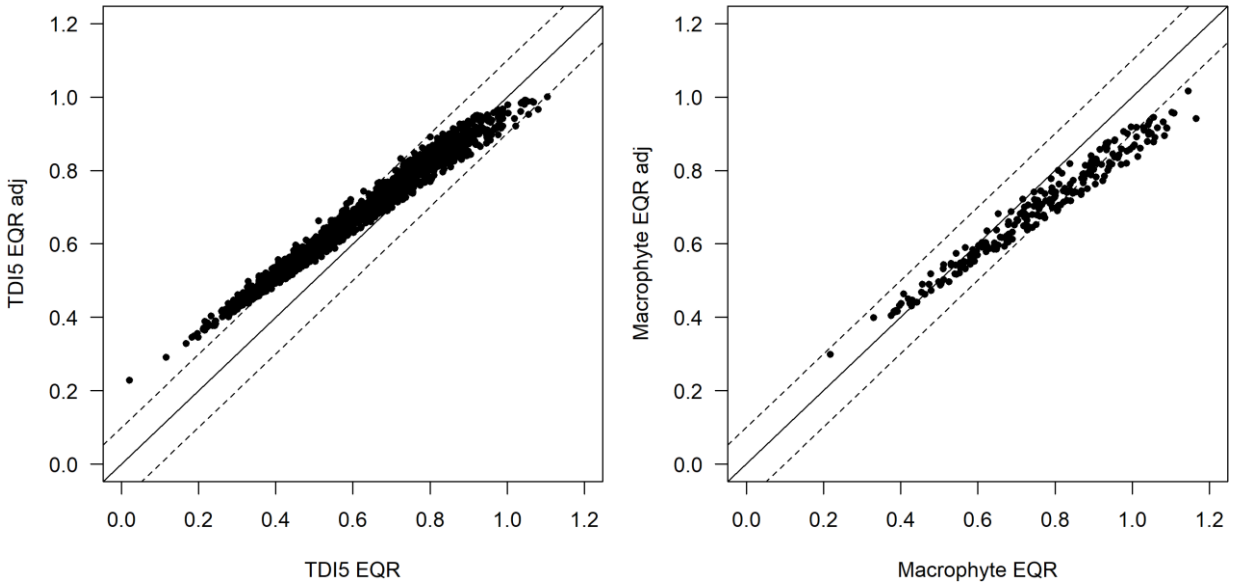

Figure 3.13: Relationship between single metric EQR and adjusted EQR to allow for effect of combining metrics by averaging when only a single metric is available. Lines show 1:1 line ± 0.1 EQR (WFD class).

3.8 Effect on classification

By changing the reference model for diatoms there will clearly be an increase in stringency in the diatom classification. However, the effect of this on the classification of the full BQE could be mitigated by changing the combination rule for the combined macrophyte and phytobenthos metric from the worst of either to the average of both.

The effect of this on classification can be seen from the distribution of the combined EQRs for the macrophyte and phytobenthos shown by probability density plots (Figures 3.14- 3.16). In comparison with the current method, taking the worst of either macrophytes or diatoms using the new reference model results in a decrease in EQR, while taking the average causes an increase. The effect on classification is shown in Table 3.11 and Figures 3.17- 3.18; typically the new reference model and the existing “worst of either” combination rule would increase the percentage of sites less than good by 10%, while averaging would decrease this by 10%.

One approach to minimising the overall change in classification would be to make an allowance for the decreased stringency associated with averaging classifications. The difference between the worst of either and average has already been made in the approach used for setting the UK river phosphorus standards, where an allowance was made for the increased stringency of the worst of either classification. Deducting 0.05 from the EQR of the averaged metrics, effectively a tightening of the normalised boundary values by 0.25 of a class, would result in no significant overall change in the classifications of the combined macrophyte and phytobenthos metric (Figure 3.16 and Table 3.11).

A possible problem with averaging is that this will reduce the likelihood of detecting ecological impacts in situations where macrophyte status is lower than that of diatoms due to non-nutrient pressures. A final possibility (not considered here) would be to introduce a more complex rule e.g. to use the average of the two sub-elements in cases where macrophyte EQR > Diatom EQR, but, otherwise, to use the worst case. This type of rule is already in use in the lake phytoplankton tool (“PLUTO”) where cyanobacteria abundance are combined with the other constituent metrics by averaging when they are worse than the other metrics but are ignored when they are better. Applying a similar rule to macrophytes and phytobenthos could be considered but would additional testing.

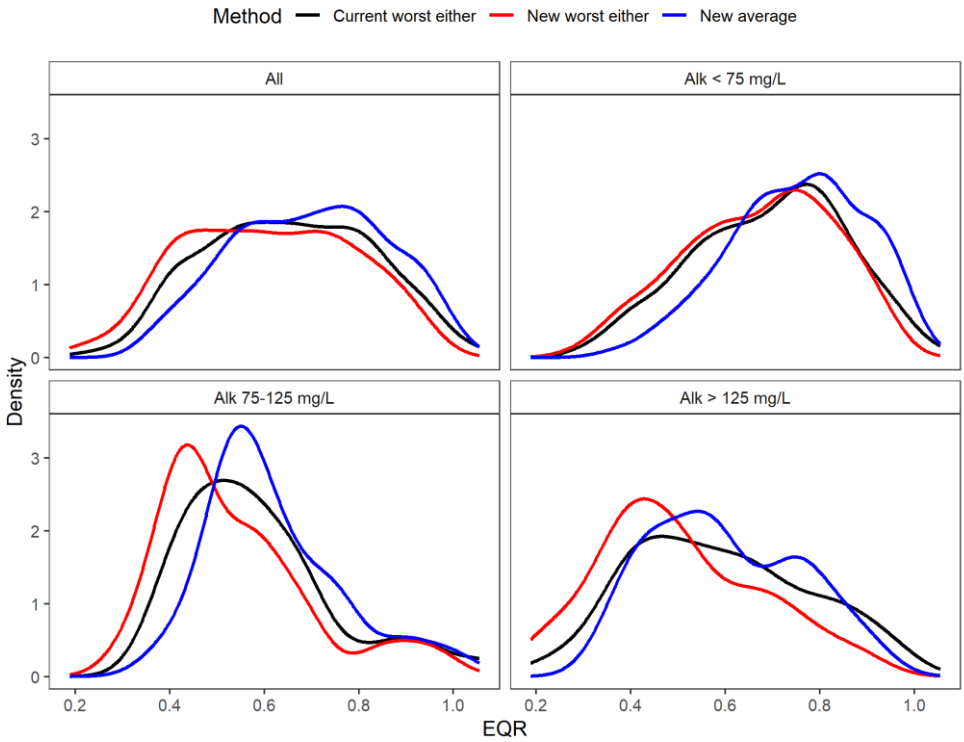

Figure 3.14: Probability densities of the combined macrophyte and diatom EQRs showing distributions for the worst of either current method (black line), the worst of either the new diatom and current macrophyte method (red line), and the average of

the new diatom and current macrophyte methods (blue line). Plots split by alkalinity range.

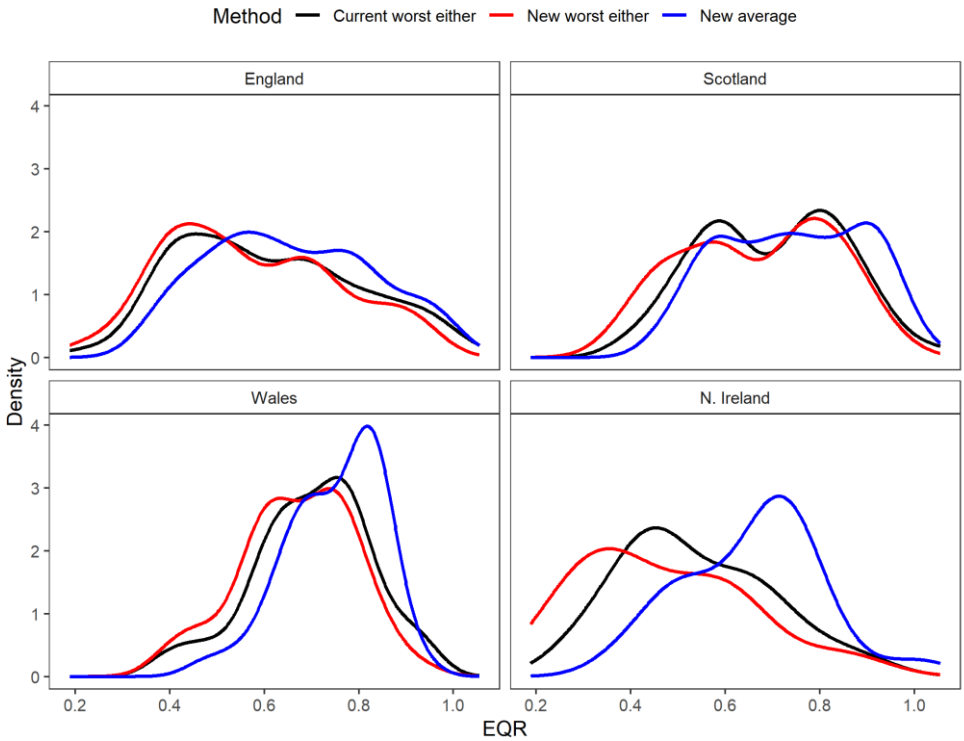

Figure 3.15: Probability densities of the combined macrophyte and diatom EQRs showing distributions for the worst of either current method (black line), the worst of either the new diatom and current macrophyte method (red line), and the average of the new diatom and current macrophyte methods (blue line). Plots split by administration.

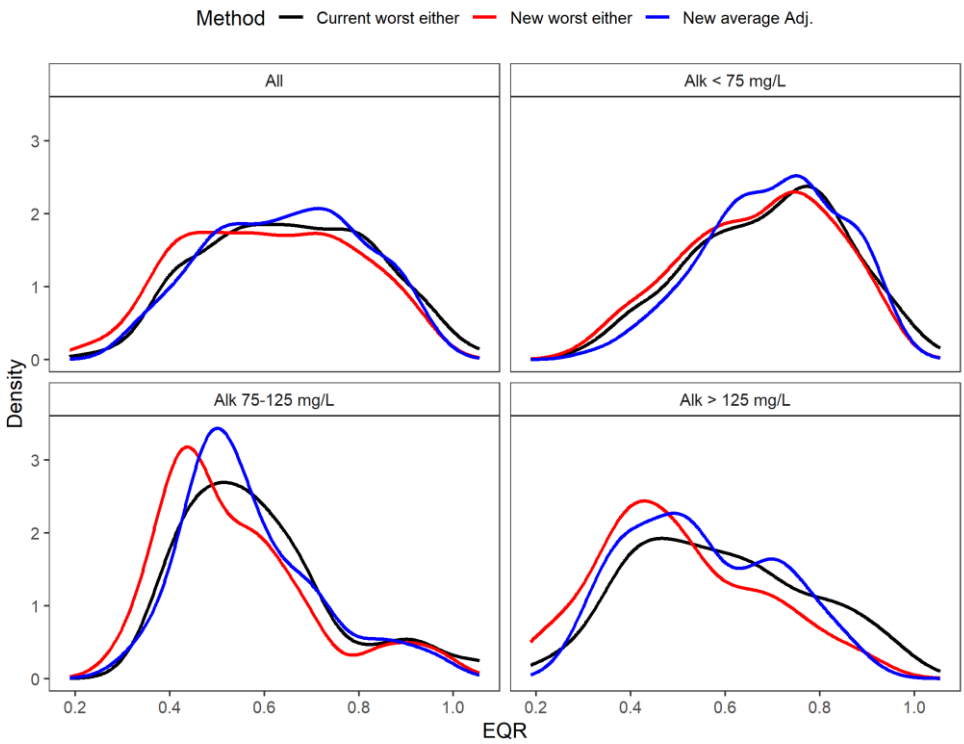

Figure 3.16: Probability densities of the combined macrophyte and diatom EQRs showing distributions for the worst of either current method (black line), the worst of either the new diatom and current macrophyte method (red line), and the average of the new diatom and current macrophyte methods, *but with a shift of –0.05 EQR units* (blue line). Plots split by alkalinity range.

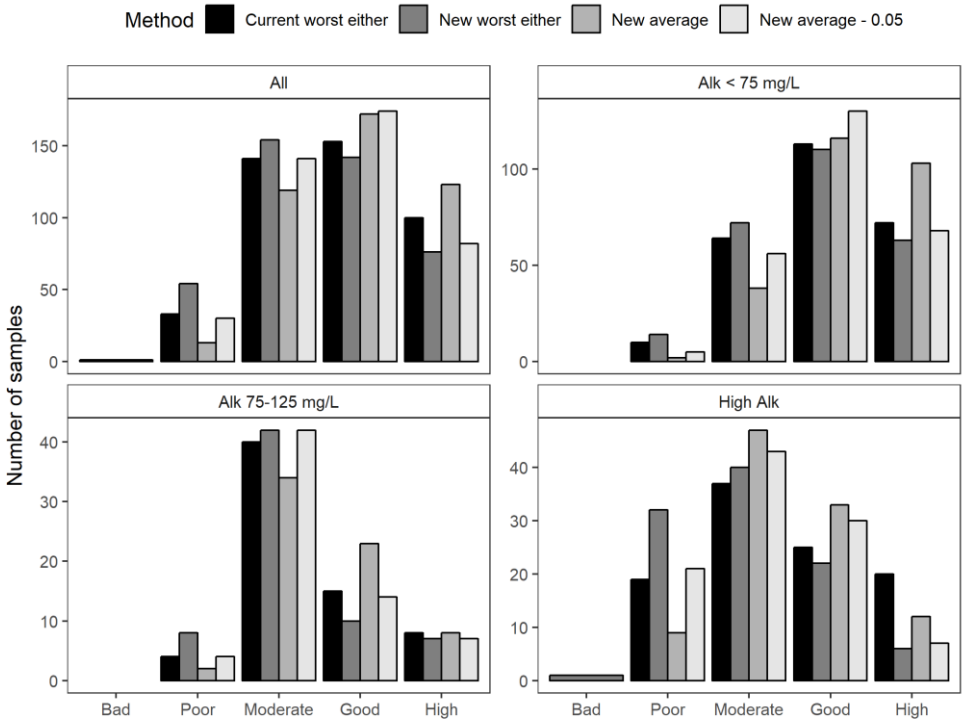

Figure 3.17: Number of samples classified by the different approaches split by alkalinity type, using individual diatom sample classifications paired with macrophyte survery classification.

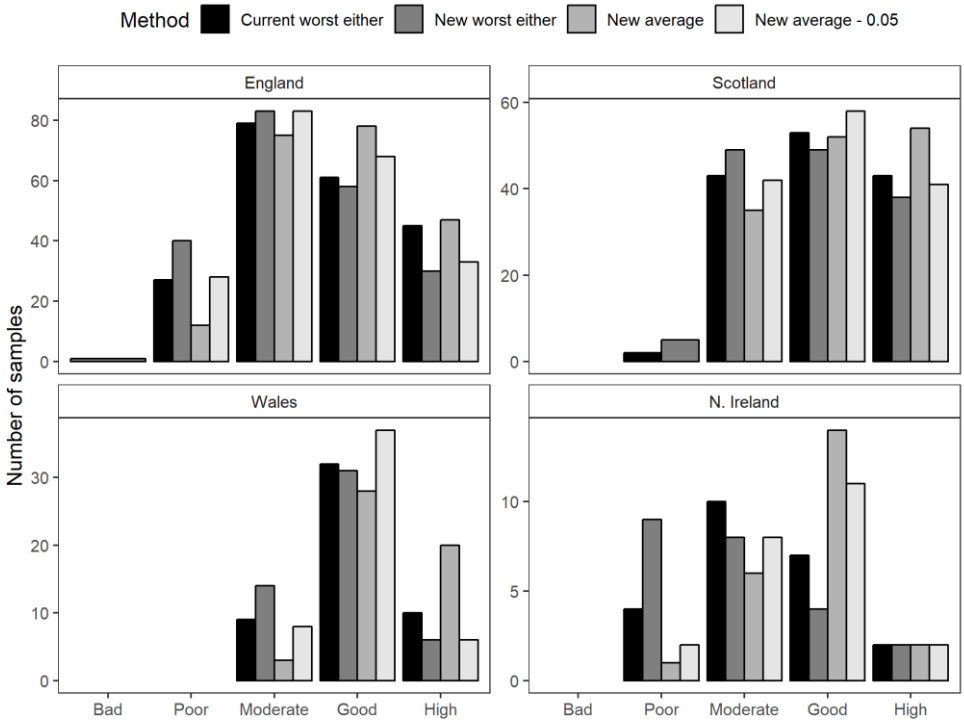

Figure 3.18: Number of samples classified by the different approaches split by country, using individual diatom sample classifications paired with macrophyte survery classification.

Table 3.11: Comparison of percentage of samples in each class using different methods, split by country.

	Current worst	New worst	Ave. New	Ave. New ± Adj.
England				
Bad	0	0	0	0
Poor	13	19	6	13
Moderate	37	39	35	39
Good	29	27	37	32
High	21	14	22	16
Less than Good	50	58	41	52
Scotland				
Bad	0	0	0	0
Poor	1	4	0	0
Moderate	30	35	25	30
Good	38	35	37	41
High	30	27	38	29
Less than Good	31	39	25	30
Wales				
Bad	0	0	0	0
Poor	0	0	0	0
Moderate	18	27	6	16
Good	63	61	55	73
High	20	12	39	12
Less than Good	18	27	6	16
N. Ireland				
Bad	0	0	0	0
Poor	17	39	4	9
Moderate	43	35	26	35
Good	30	17	61	48
High	9	9	9	9
Less than Good	60	74	30	44

3.9 Case studies

The potential implications are further explored in Fig. 3.19, which takes data from three water bodies where the spatial and temporal sampling intensity is greater than that typically used for classification. This gives a better indication of the scale of within-water body variability against which the consequences of changes can be judged.

For the Ehen (upper including Liza) water body in Cumbria, six samples were collected from four locations in the River Ehen in a 5 km stretch immediately downstream from Ennerdale Water during 2014. The status for Macrophytes and Phytobenthos for this low alkakinity river is presently”good" (note that the water body extends for some distance downstream from the sampled stretches, with influences from richer farmland and some village sewage works). Phosphorus status is “high”. The current DARLEQ tool would result in a classification of high status and this will drop slightly using the new reference model, though not below high status. The predicted EQR for the full BQE with and without adjustment are also both high status.

The upper stretches of the River Wear were chosen as an example of a moderate alkalinity water body (Wear from Middlehope Burn to Houselop Beck). Three sites, each sampled on four occasions during 2014 and early 2015, yielded a classification of “high status” using the current reference model, but this will drop to good status (probably with low confidence due to overlap with moderate status) using the new reference model. The median of the predicted EQR for the full BQE, however, would be more securely within good status. Macrophytes and phytobenthos were not

assessed by the EA in 2014; however, evaluations in 2015 and 2016 were “high status”, with overall status determined by invertebrates which were “good status”. Once again, phosphorus was at “high status” but zinc was at “moderate status” (due to historic mining in the upper catchment) and this may have had consequences for the biota.

Finally, the headwaters of the River Wylye show the likely consequences of proposed changes for a chalk stream. Six samples from five locations in this water body were sampled in 2010 and 2011. These yield a classification of “good status” using the present model, but this would drop to “poor status” using the revised model, although the prediction for the combined BQE would be “moderate status”. Macrophytes and phytobenthos were not formally assessed by the EA during this period, but are presently classified as “moderate status”, with overall status determined by fish, which are “poor" . Phosphorus is currently classified as “moderate”, with recent evidence for episodic delivery from phosphorus-saturated soils in the riparian zone along with flushing of bankside septic tanks (Lloyd *et al*., 2018).

Whilst only a limited study, these three examples suggest relatively minor effects at low alkalinity but more pronounced effects at moderate and high alkalinity. These changes are mostly due to the change in the reference model, with the subsequent prediction of the combined BQE and the final 0.05 EQR adjustment each having a relatively minor effect. At high alkalinity, in particular, the change between the current and proposed models needs to be interpreted in light of the known shortcomings of the current model.

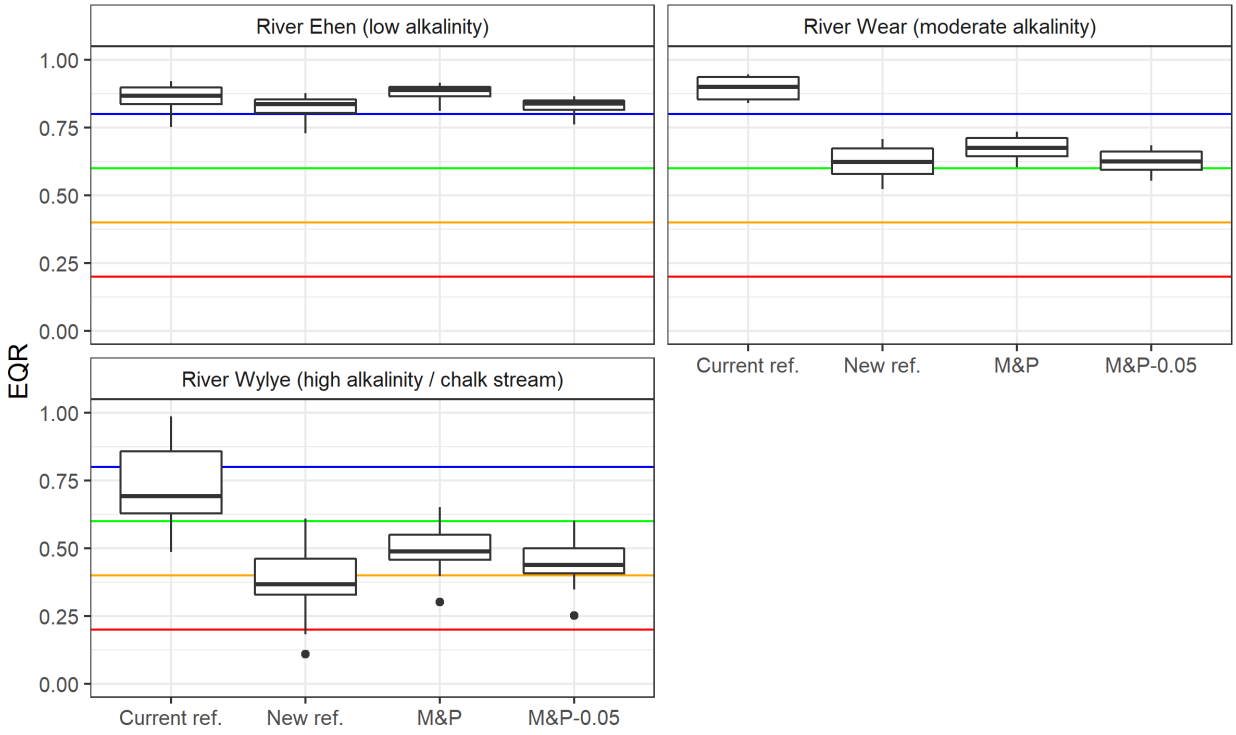

Figure 3.19: Effect of changing reference model and combination rule on phytobenthos-derived status for three water bodies of contrasting alkalinity. Current ref. = EQR calculated using the present (DARLEQ2) reference model; New ref. = EQR calculated using the reference model proposed in 3.4; M&P = EQR for combined BQE, calculated using formula in 3.7; M&P-0.05 = as M&P but with final 0.05 EQR adjustment included. All calculations are based on TDI4.

4 Calculation of confidence of class for TDI5NGS

4.1 Confidence of Class (CoC) calculations for TDI4

Confidence of Class (CoC) estimates in DARLEQ are based on estimates of site-level temporal variation using the framework developed by Ellis & Adriaenssens (2006) and developed for TDI metrics in Kelly *et al*. (2009). Briefly, if we assume that the uncertainty, or 'confidence distribution', associated with an EQR follows a normal distribution with known standard deviation, the probability of observing an EQR of X or better if the true EQR u were on a class boundary is given by:

$$p_i = 1 - \Phi(x - u_i)/s_i$$

where Φ denotes the cumulative normal probability, u_i denotes the EQR of class boundary i, and s_i denotes the standard deviation of the EQR at class boundary i. Computing the p_i for each class boundary enables us to calculate the confidence of class for each status class:

$$Confidence\ of\ bad\ status = 100p_B$$
$$Confidence\ of\ poor\ status = 100(p_P - p_B)$$

$$Confidence\ of\ moderate\ status = 100(p_M - p_P)$$
$$Confidence\ of\ good\ status = 100(p_G - p_M)$$
$$Confidence\ of\ high\ status = 100(1 - p_G)$$

The risk of misclassification (ROM) is then given by the sum of CoC values for all classes except the observed class. Confidence of class will be close 50% when an EQR falls close to a very class boundary and will fall towards the middle of the class at a rate depending on the standard error of the mean EQR (Figure 4.1). The standard error decreases in proportion to the square root of the number of samples, so CoC increases and ROM decreases as more samples become available to calculate a mean EQR for a site.

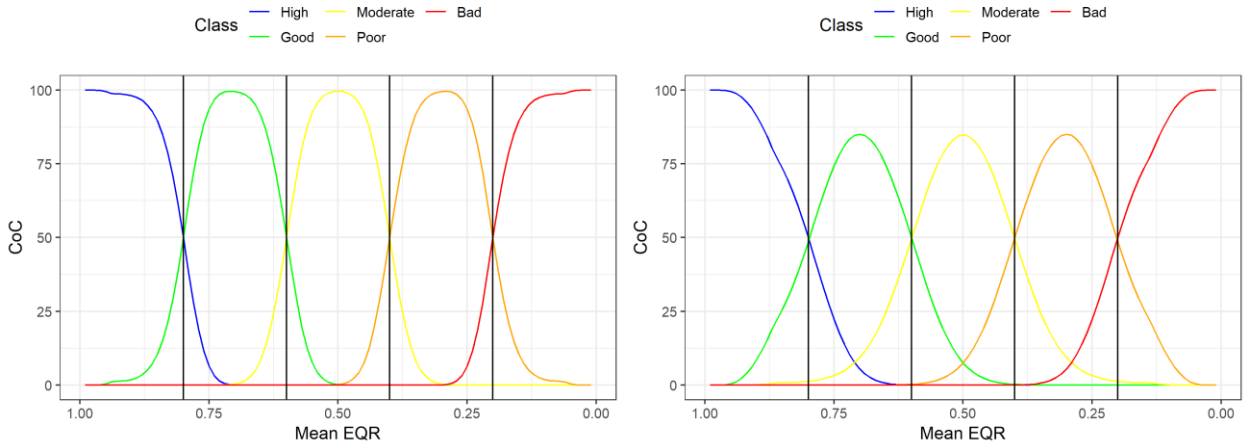

Figure 4.1: Example of confidence of class (CoC) for TDI4 as a function of mean EQR based on 2 samples with a standard deviation of 0.05 TDI units (left) and 0.1 (right).

The approach described above was developed primarily to estimate the Confidence of Class for a single site sampled on a single occasion. In this situation we will not usually have a direct measure of the uncertainty associated with an individual sample EQR. In order to develop a general approach for determining CoC, we therefore use a model relating typical EQR standard deviation to mean EQR, calibrated using data from sites with multiple sample EQR measurements. Because EQRs are constrained to fall between 0 and 1, we would also expect the EQR standard deviation to approach zero at the ends of the EQR gradient. The approach used therefore fits polynomial curve through the data, with the additional constraints that the curve passes through two 'anchor' points at EQR=0 and EQR=1. Figure 4.2 shows the relationship between EQR standard deviation and mean EQR for the calibration dataset used in Kelly *et al*. (2009), along with the fitted curve. This model is used to estimate the likely standard deviation for a given EQR in the above equation.

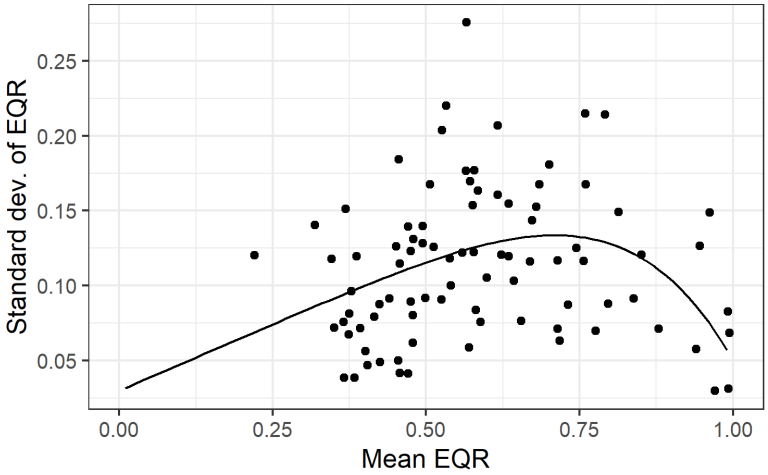

Figure 4.2: Within-site variability of TDI4 EQRs for UK rivers with six or more samples from Kelly et al. (2009). The fitted polynomial function is used in DARLEQ TDI4 confidence of class calculations to predict site-level uncertainty at a given mean EQR.

The relationship above was derived for TDI3. For the Phase 2 and 3 datasets we have 688 sites with multiple temporal samples (658 sites with N=2 (usually spring/autumn); 15 sites with N=3 and 15 sites with N=4). Figure 4.3 shows the distribution of standard deviations for these data for TDI4, TDI5LM and TDI5NGS and Figure 4.4 shows the relations between the relationships between EQR standard deviations for TDI5LM (left) and TDI5NGS (right) against TDI4 standard deviations. The distributions of TDI4 and TDI5LM standard deviations are very similar, which is not surprising given the very close agreement between the samples scores for these two metrics. The close agreement between TDI4 and TDI5LM EQR standard deviations suggests that it is appropriate to use the polynomial within-site EQR model developed for TDI4 with TDI5LM assessments.

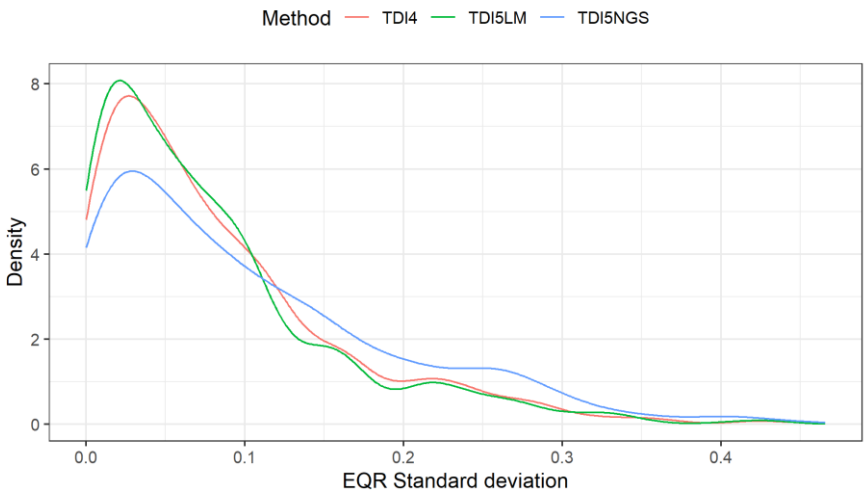

Figure 4.3: Density plots showing the distributions of EQR standard deviations for TDI4, TDI5LM, and TDINGS for Phase 2 & 3 site-aggregated data with two or more temporal samples.

The pattern of standard deviations for TDI5NGS is similar to TDI4 and TDI5LM (Figure 4.3) but shows a general shift to slight higher values than TDI4 or TDI5LM (mean SD for TDI4 = 0.083, mean SD for TDI5NGS = 0.102), reflecting a greater amount of temporal variation in NGS samples. More surprising is that there is only a moderate correlation between TDI5NGS and TDI4 EQR standard deviations (r=0.4, p<0.001; Figure 4.4, right) whereas standard deviations for TDI4 and TDI5LM are closely related (Fig. 4.4, left).

SC160014 discussed some possible reasons for the difference between TDI5LM and TDINGS scores. However, the reasons for the greater temporal variation in the NGS data as compared to LM assessments made on the same sample are not clear and requires more work to quantify the sources of variability in NGS determinations.

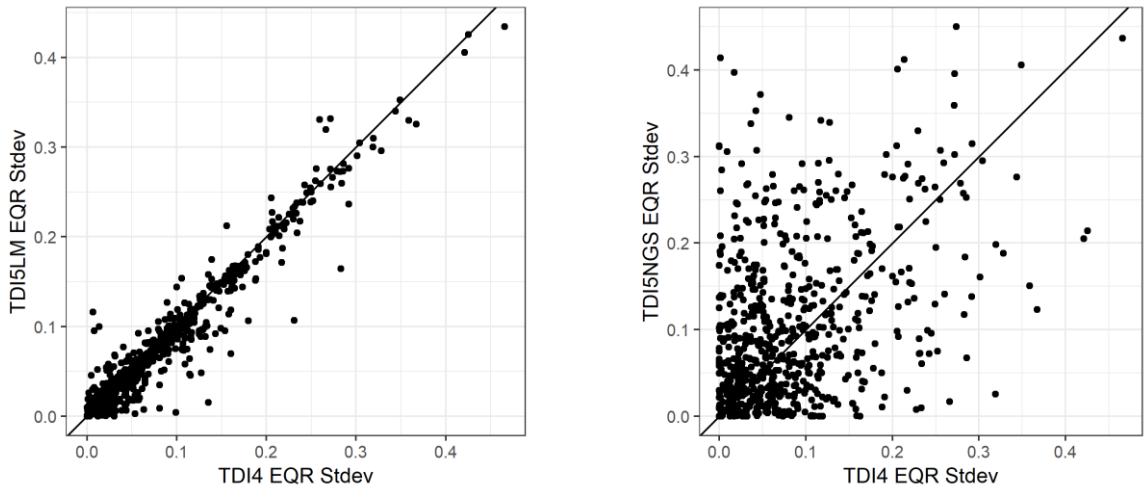

Figure 4.4: Relationship between EQR standard deviations for TDI4 and TDI5LM (left) and TDI5NGS (right) for Phase 2 & 3 site-aggregated data with two or more temporal samples.

The relationship shown in Figure 4.4 suggests that the within-site EQR model developed for TDI4 could be used for TDI5NGS EQRs. Figure 4.5 shows the model superimposed over a plot of with Phase 2 & 3 site-aggregated mean and SD EQRs for TDI4, TDI5LM and TDI5NGS. Also shown on the plots is the fitted polynomial re-calibrated using the Phase 2 & 3 data. For TDI4 and TDI5LM the re-calibrated

models are similar and both predict slightly lower standard deviations that the original model. Thus, the original within-site EQR model is more conservative and will yield slightly lower CoC estimates. The model fitted to within-site TDI5NGS data closely tracks the original TDI4 model.

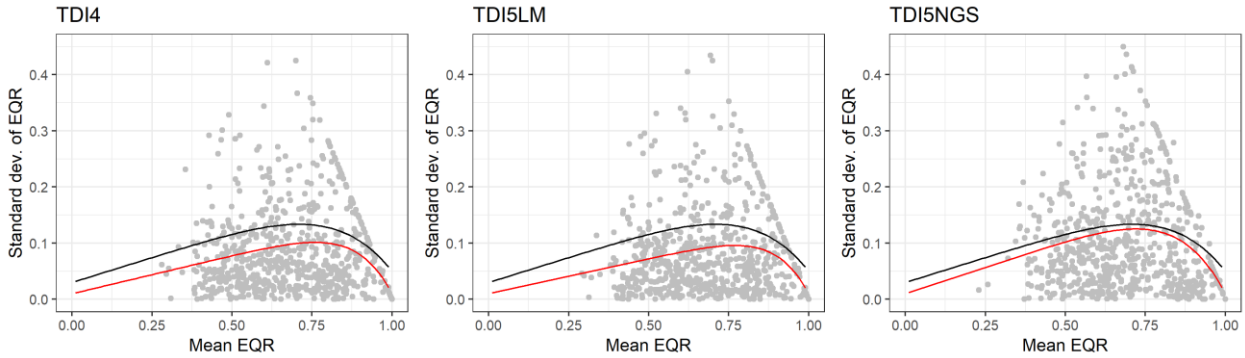

Figure 4.5: Mean EQRs and associated standard deviations for Phase 2 and 3 sites with 2 or more samples. Lines show original TDI CoC model (black) and model fitted to site-aggregated Phase 2 & 3 data, using constrained polynomial regression (red).

4.2 Comparison with VISCOUS

VISCOUS (Davey, 2009) is an extension of the approach developed by Ellis & Adriaenssens (2006) to account for within-waterbody spatial variability. The tool differs from the approach describes in 4.1 and used in DARLEQ3 in that it is designed for classifying water bodies, rather than individual sites. Input to VISCOUS is an EQR for a site, along with an associated standard deviation that quantifies the temporal variation in the EQR for that site along with any measurement error. If there is only one site in the waterbody the site EQR is used to classify the waterbody. Where there are multiple sites in a waterbody, and where the waterbody is spatially homogeneous, a mean EQR is used and sites can be optionally weighted to derived a weighted-mean EQR to reflect the size or areal representation of individual sites within the waterbody. For heterogeneous waterbodies that have discontinuities in conditions, individual sites can be allocated to different strata, and the between-site (within-strata) and between-strata variability estimated and included in a pooled estimate of uncertainty for the waterbody.

Currently DARLEQ3 can classify samples containing one or more samples and calculate Confidence of Class estimates using an estimate of the temporal uncertainty. DARLEQ3 does not have an option to classify waterbodies using multiple sites and extension to this scenario is beyond the scope of this report. To classify a single site, VISCOUS requires an *a-priori* estimate of the temporal uncertainty. This can be derived in three ways:

1. Use the observed standard deviation of multiple temporal samples at the site,
2. Use a modelled standard deviation. This could be a constant value or derived from a polynomial within-site model of EQR variation as in Kelly et al. (2009) and the current DARLEQ3 tool.
3. Use a pooled estimate of 1 and 2.

Where only one sample is available for a site options 1 and 3 are not possible and a modelled EQR standard deviation must be used. Option 1 uses the observed standard deviation, so requires a good estimate of this value. This may be available in some situations but in many cases N may be low, and the observed standard deviation may not be a good estimate of the real temporal variation at a site. Option 2

is implemented in DARLEQ2 and DARLEQ3 software and reflects the average temporal variation at a given EQR but may under or overestimate this error at any particular sites (e.g. Figure 4.2). Option 3 may be a good compromise and allows the adjustment of the average standard deviation calculated by option 2 using empirical evidence of temporal variation from each site. For the pooled estimate a decision has to be made on the relative weighting of the modelled and observed standard deviation. In the examples below we follow VISCOUS and weight by degrees of freedom, using a notional value of 5 for the modelled standard deviation, and N-1 for the observed value.

Figure 4.6 shows the relationship between Confidence of Class estimates calculated using these three approaches for each TDI metric and split be status class, with CoC based on the EQR within-site model on the x-axis and CoC based on observed and pooled standard deviation on the y-axis. Option 1 yields generally higher CoC than option 2, and in many cases unrealistically high confidence, as a result of the low observed EQR standard deviation at many sites. Option 3 yields almost identical CoCs for high and poor status sites but adjusts the modelled version by up to 10% for good and moderate status sites. The magnitude of this adjustment will increase with the number of temporal samples: for N=5, the adjustment is up to c. 20%.

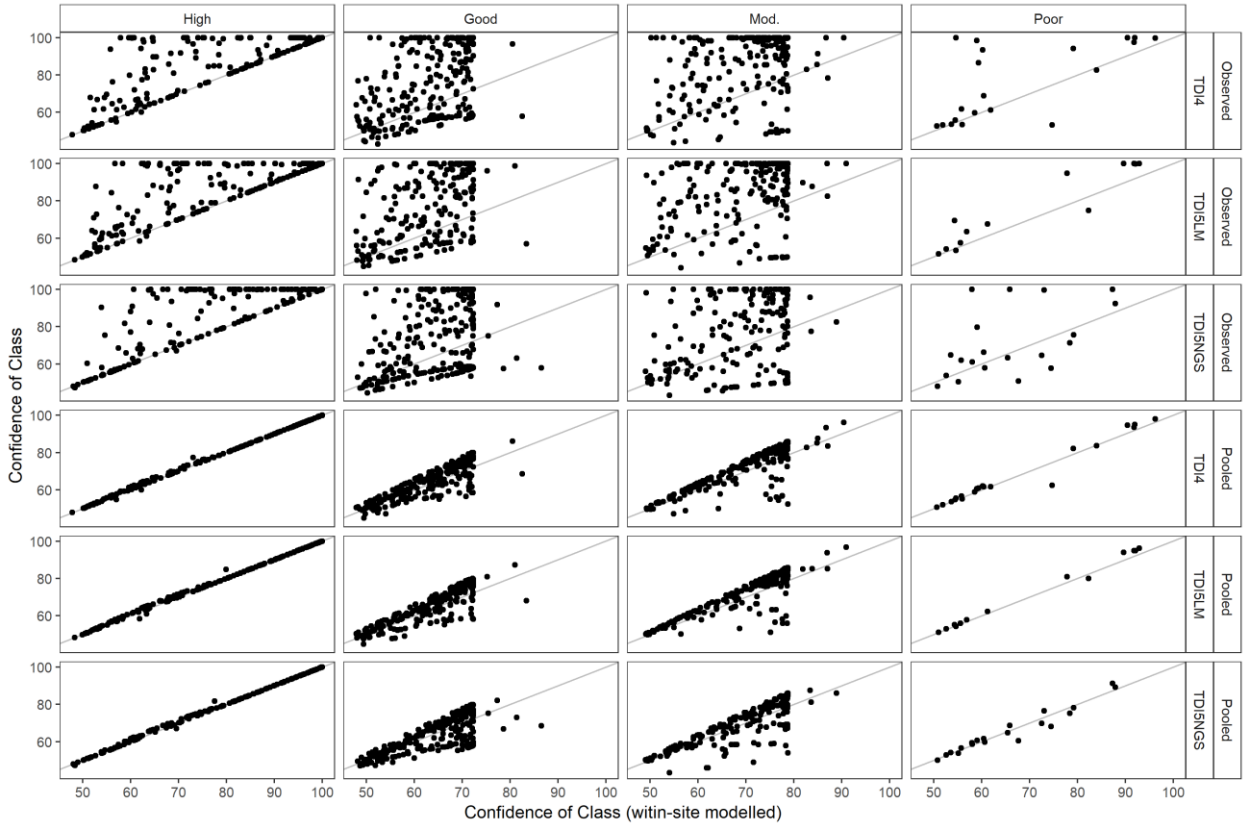

Figure 4.6: Confidence of Class for Phase 2 & 3 site-aggregated data, showing relationship between COC based on within-site EQR model (x-axis), and (y-axis) CoC based on observed EQR standard deviation (top three rows) and pooled (observed + modelled) standard deviation (bottom three rows), for each metric (rows), split by status class (columns).

4.3 Discussion

The original requirement in DARLEQ was to provide estimates for Confidence of Class for site-level assessments with one or more temporal samples per site. In the absence of an independent measure of within-site variation when N=1, a within-site

EQR model was developed to predict likely EQR standard deviation at a given EQR. The TDI4 within-site model was based on data from 106 sites, with at least six temporal samples per site collected over a period of at least 3 years. The majority of the site-aggregated Phase 2 and 3 data only have two samples per site collected during the same year and do not allow a detailed or accurate assessment of within-site variability. Given the lack of an appropriate dataset to recalculate the within-site EQR model using NGS data, and the observation that the original model is more conservative than models fitted using the Phase 2 and 3 dataset of site EQRs, an argument can be made for using the existing TDI4 model to estimate EQR standard deviations for all metrics, that is, TDI4, TDI5LM and TDI5NGS.

The within-site model of EQR temporal variability represents a pragmatic solution for the problem of estimating standard deviation when the number of temporal samples is low. Natural systems do however display considerable variation in temporal variability in EQR (e.g. Figure 4.2) and there is an argument for incorporating empirical estimates of this variability into the CoC calculation when multiple temporal samples are available using a pooled estimate of the standard deviation. By weighting the contribution of the observed standard deviation by N-1 the influence of the observed standard deviation will increase as the estimate of it improves. An approach based on a pooled estimate of the EQR variability has the advantage of primarily reflecting a modelled standard deviation when N is low but increasingly us site-specific information on temporal variability as N increases. It has the additional advantage that it can also incorporate the apparent greater temporal variability observed in the NGS-based metric.

5 Options analysis

This report outlines different options for both the TDI and its conversion to an EQR. The implications of these have been evaluated separately, with the underlying assumptions that a consistency of approach across the UK is desirable, but that the decision to adopt NGS is, in part, driven by internal factors and economies of scales, which will differ between the UK's regions. The previous sections have demonstrated that the effect of changing the reference model will have a greater impact than the change in metric, so the consequences of changes to the base metric and changes to the reference model have been evaluated separately.

Intercalibration needs are described as "minor" if the agreement between current and new methods is high ($r^2 > 0.8$ - see CIS Guidance 30) and reference model is unchanged; otherwise, as "significant". All options (apart from "do nothing") will require an intercalibration report of some kind.

5.1 Alternative versions of the TDI

5.1.1 Do nothing: continue using TDI4

- Only suitable for samples analysed by LM
- Intercalibration needs: none

5.1.2 Adopt TDI5LM as a replacement for TDI4

- Very similar strength of relationship to nutrients compared to TDI4

- Classifications 5.7% less stringent than TDI4 (assuming no change to reference model)
- Classifications are very similar to those based on TDI5NGS (0.4% bias)
- Fully backwardly compatible with TDI4
- Intercalibration needs: minor (if no change to reference model; otherwise, see below)

5.1.3 Adopt TDI5NGS as a replacement for TDI4

- Slightly weaker relationship with nutrients, compared to TDI4
- Classifications 5.3% less stringent than TDI4 (assuming no change to reference model)
- Classifications are very similar to those based on TDI5LM (0.4% bias)
- Results show consistent within-site trends as TDI4 (but limited case studies so far)
- Intercalibration needs: significant (revise compliance criteria and demonstrate that boundaries are no more relaxed than at present)

5.1.4 Recommendation

In effect, TDI5LM is a minor upgrade to TDI4 that offers similar performance and greater consistency with TDI5NGS. Administrations can base decision on whether to use LM or NGS on practical considerations.

5.2 New reference model

5.2.1 Do nothing: continue with DARLEQ2 reference model

- Cannot be used in high alkalinity streams and rivers, limiting capacity to evaluate ecological status in situations where macrophytes cannot be used.
- Intercalibration needs: none

5.2.2 Update diatom reference model, continue using lowest of macrophyte and diatom classifications

- Improved science, compared to DARLEQ2: model is based on more data and is less dependent on extrapolation and assumptions
- Classification principle remains unchanged
- New diatom reference model can be used across the entire alkalinity gradient.
- Diatom-based classifications are at least 30% more stringent than those using DARLEQ2 reference model
- Diatom-based classifications are more stringent than those based on macrophytes, so will tend to determine combined result.

- Little justification for using macrophytes for classification in future
- Overall, about 10% more sites will be classified as less than good status.
- Intercalibration needs: significant (revise compliance criteria, and demonstrate that boundaries are no more relaxed than at present)

5.2.3 Update diatom reference model, classify using average of macrophyte and diatom classifications

- Provides a more ecologically appropriate overall assessment of the impact of nutrients in rivers
- Will need to justify decision to alter combination rule to stakeholders
- Overall, about 10% fewer sites will be classified as less than good status
- Value of combined macrophyte/diatom EQR can be inferred from either sub-element, offering the choice of monitoring either or both sub-elements in any situation
- Intercalibration needs: as previous (consequences of change apply only to combined macrophyte/phytobenthos BQE, which is not subject to a separate intercalibration)

5.2.4 Update diatom reference model, classify using average of macrophyte and diatom classifications, but with final classification adjusted to ensure consistency with current "worst of either" approach

- Has the advantages of averaging metrics whilst still retaining the stringency of the existing classification
- Will need to justify decision to alter combination rule to stakeholders
- Intercalibration needs: as previous (consequences of change apply only to combined macrophyte/phytobenthos BQE, which is not subject to a separate intercalibration): required

5.2.5 Recommendation

We recommend adoption of the revised reference model for diatoms, and that to provide a more ecologically appropriate assessment of WFD status results of macrophyte and diatom classifications are now averaged. The revised reference model is an improvement over the current model, particularly for high alkalinity rivers and streams but the shift to the revised reference model will necessitate adoption of an alternative combination rule in order to ensure the continued relevance of macrophytes and this, in turn, will require a further adjustment in order to ensure that the revised approach is at least as stringent as the current approach. In practice, the consequences the shift to the revised model will be felt mostly in England (which has the highest proportion of high alkalinity sites).

6 References

Davey, A. 2009. VISCOUS: Taking account of spatial variability in waterbody classification. Science Report SC080051/SR1, Environment Agency, Bristol.

Ellis J., Adriaenssens V. 2006. Uncertainty estimation for monitoring results by the WFD biological classification tools. WFD Report: GEHO1006BLOR-E-P. Environment Agency (UK).

Hustedt, F., 1930. Susswasserflora von Mitteleuropa 10: Bacillariophyceae. Gustav Fischer, Jena.

Kelly, M., Juggins, S., Guthrie, R., Pritchard, S., Jamieson, J., Rippey, B., Hirst, H., Yallop, M., 2008. Assessment of ecological status in U.K. rivers using diatoms. Freshw. Biol. 53, 403-422.

Kelly, M., H. Bennion, A. Burgess, J. Ellis, S. Juggins, R. Guthrie, J. Jamieson, V. Adriaenssens and M. Yallop (2009). Uncertainty in ecological status assessments of lakes and rivers using diatoms. Hydrobiologia 633(1): 5-15.

Kelly, M. & Yallop, M. 2012. A streamlined taxonomy for the Trophic Diatom Index. Report SC070034/TR1. Bristol: Environment Agency.

Kelly, M., Willby, N., Phillips, G., Benstead, R., 2013, The integration of macrophyte and phytobenthos surveys as a single biological quality element for the Water Framework Directive. Bristol, p. 28.

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann. D., Pass, D., Sapp, M., Sato, S., Glover, R., 2018a. A DNA based diatom metabarcoding approach for classification of rivers. Science Report SC140024/R, Environment Agency, Bristol.

Kelly, M., Boonham, N., Juggins, S., Mann. D., Glover, R., 2018b. A DNA based diatom metabarcoding approach for classification of rivers: Further development. Science Report SC160014/SR, Environment Agency, Bristol.

Koenker, R., 2017, quantreg: Quantile Regression. R package version 5.36. http://CRAN.R-project.org/package=quantreg.

Lavoie, I., S. Campeau, F. Darchambeau, G. Cabana and P. J. Dillon, 2008. Are diatoms good integrators of temporal variability in stream water quality? Freshwater Biology. 53, 827-841.

Lloyd, C.E.M., Johnes, P.J., Freer, J.E., Carswell, A.M., Jones, J.I., Stirling, M.W., Hodgkinson, R.A., Richmond, C. & Collins, A.L. (2018). Determining the sources of nutrient flux to water in headwater catchments: Examining the speciation balance to inform the targeting of mitigation measures. Science of the Total Environment 648: 1179-1200.

Pardo, I., Gómez-Rodríguez, C., Wasson, J.-G., Owen, R., van de Bund, W., Kelly, M., Bennett, C., Birk, S., Buffagni, A., Erba, S., Mengin, N., Murray-Bligh, J., Ofenböeck, G., 2012. The European reference condition concept: A scientific and technical approach to identify minimally-impacted river ecosystems. Sci. Total Environ. 420, 33-42.

Snell, M. A., P. A. Barker, B. W. Surridge, A. R. Large, J. Jonczyk, C. M. Benskin, S. Reaney, M. T. Perks, G. J. Owen, W. Cleasby, C. Deasy, S. Burke and P. M. Haygarth, 2014. High frequency variability of environmental drivers determining benthic community dynamics in headwater streams. Environ. Sci. Processes Impacts 16, 1629-1636.

ter Braak, C. J. F. and L. G. Barendregt, 1986. Weighted averaging of species indicator values: Its efficiency in environmental calibration. Mathematical Biosciences 78: 57-72.