Fitting the revised assessment method for rivers in The Netherlands using macrophytes to the results of the completed Central-Baltic rivers' intercalibration exercise.

commissioned by Rijkswaterstaat Water Verkeer en Leefomgeving, Lelystad

Authors:

Roelf Pot
Sebastian Birk

Oosterhesselen, December 2015

Roelf Pot
research and consultancy

Fitting the revised assessment method for rivers in The Netherlands using macrophytes to the results of the completed Central-Baltic rivers' intercalibration exercise.

Colofon

authors: Roelf Pot
Sebastian Birk

date: december 2015 (final adjustments 29-2-2016)

contact: Roelf Pot
Pandijk 2, 7861 TE Oosterhesselen, The Netherlands
e-mail: roelfpot@wxs.nl
internet: www.roelfpot.nl

commissioned by: Rijkswaterstaat Water Verkeer en Leefomgeving, Lelystad

refer to: Pot, R. & S. Birk, 2015: Fitting the revised assessment method for rivers in The Netherlands using macrophytes to the results of the completed Central-Baltic rivers' intercalibration exercise. Roelf Pot, Oosterhesselen, Rijkswaterstaat, Lelystad

Table of contents

1 Introduction

The Netherlands have previously participated in the intercalibration exercise of the European assessment methods for water quality in rivers using macrophytes. It appeared that the original Dutch assessment method (Stowa, 2007) did not show sufficient correlation with the Pseudo-Common Metric for the rivers in the Central-Baltic Geographical Intercalibration Group and therefore could not complete the intercalibration (Birk & Willby, 2011).

National evaluation of the Dutch metric showed several flaws within the method and suggestions for improvements. In 2012 a revised assessment method was published (Stowa, 2012) and since then used in The Netherlands. The revision was largely inspired by the intercalibration process and results, and the method was also partly calibrated using the assessment results of the Pseudo-Common Metric for the dataset of macrophyte samples used in the intercalibration exercise. Completion of the development of the metric was nevertheless not in time to be included in the completion of the intercalibration process.

In 2013 a procedure was developed to fit in new or revised assessment methods in the completed intercalibration exercise (Birk et al., 2013; Willby et al., 2014). The document at hand describes the process of carrying out this procedure and is used to intercalibrate the revised Dutch assessment method for macrophytes in rivers.

2 The revised method

2.1 Short description

The Dutch assessment method is a revised version of the original method (Stowa, 2007). It combines two metrics, one based on species composition and the other based on total abundance of a selection of growth forms. Originally the metric was focused on assessing whole waterbodies by combining all data from a standard number of samples in a waterbody. The metric on species composition correlated too well to species number, and therefore, undesirably, to monitoring intensity as well. For small and medium sized rivers, samples of six sites were obligatory to standardize the monitoring intensity; the assessment of single sites was not possible.

The most important changes in the revision are the formulas used to assess the species composition of the samples. They make the method almost independent from species number in the samples and therefore single sites also can be evaluated. In all other aspects there was no change, apart from a few details regarding the status of some species.

For the fitting procedure the method is regarded as a new method since the initial method was not intercalibrated. According to European Commission (2011) only methods meeting the requirements of the Water Framework Directive (WFD) can be intercalibrated. The features of the initial method were tested satisfactory in the intercalibration exercise on both compliance and feasibility (Birk & Willby, 2011). These have not been changed in the revision but will, nevertheless, be discussed briefly in the next sections.

2.2 Checking of compliance with the WFD requirements

The WFD compliance criteria are specified in the reporting template for milestone reports (Annex VI of European Commission, 2011). We used this template to document the compliance of the Dutch method (Table 1). All compliance criteria are met for the revised method.

Table 1. Compliance criteria and compliance checking conclusions

Compliance criteria	Compliance checking conclusions
Ecological status is classified by one of five classes (high, good, moderate, poor and bad).	Yes – the method classifies ecological status by one of five classes (in Dutch: Zeer goed, Goed, Matig, Ontoereikend, Slecht).
High, good and moderate ecological status are set in line with the WFD's normative definitions (Boundary setting procedure).	Following the WFD's normative definitions for macrophytes, high ecological status is defined by a taxonomic composition and total abundance of selected growth forms that corresponds totally or nearly totally to undisturbed conditions. Good status exhibits slight changes in the species composition and no indication of accelerated plant growth as indicated by the total abundance of growth forms. At moderate status, taxonomic composition and

	abundance differ moderately from the type specific community that can be recognized as due to human impact.
All relevant parameters indicative of the biological quality element are covered (see Table 1 in the IC Guidance). A combination rule to combine parameter assessment into BQE assessment has to be defined. If parameters are missing, Member States need to demonstrate that the method is sufficiently indicative of the status of the QE as a whole.	The BQE "Macrophytes and Phytobenthos" is assessed with three metrics, from which macrophytes are involved in two of these: species composition of macrophytes and total abundance of growth forms. The growth forms include macrophytes as well as filamentous algae, which is regarded as a metric for abundance of phytobenthos. Diatoms species composition is the third metric. This makes that all relevant parameters of the BQE are covered.
Assessment is adapted to intercalibration common types that are defined in line with the typological requirements of the WFD Annex II and approved by WG ECOSTAT.	Yes, the typological features of all Dutch types correspond with the requirement of the WFD; the two most common NL-types of small and medium sized rivers correspond fully with two of the intercalibration common types.
The water body is assessed against type-specific near-natural reference conditions.	Yes, the method assesses against type-specific near-natural reference conditions.
Assessment results are expressed as EQRs.	Yes, the methods expresses the assessment results as normalised EQRs.
Sampling procedure allows for representative information about water body quality/ ecological status in space and time.	Yes, the sampling procedure follows the European Standard CEN 14614 and allows for representative information about the ecological status.
All data relevant for assessing the biological parameters specified in the WFD's normative definitions are covered by the sampling procedure.	Yes, all relevant data are covered by the sampling procedures.
Selected taxonomic level achieves adequate confidence and precision in classification.	Yes, the species-level used guarantees adequate confidence and precision in classification.

2.3 Intercalibration feasibility check

The intercalibration feasibility check evaluates whether the new method considers the same common intercalibration types and pressures as addressed in the completed intercalibration exercise. Furthermore, the check examines whether the assessment concept of the method is similar to the concept of the methods intercalibrated in the completed exercise.

Typology

The common types in the intercalibration process R-C1 and R-C4 (Birk & Willby, 2011) closely correspond with the Dutch types R5 and R6, respectively (see Table 2). N.B. All waterbodies in The Netherlands have an altitude < 200 m and there are no rivers with low alkalinity.

Table 2. Comparison of typology

Types	Type characteristics	Common type	NL type
Sandy lowland brooks Common type: R-C1 NL type: R5	Catchment area:	10 - 100 km^2	10 - 100 km^2
	Altitude:	<200 m	-
	Geology:	Siliceous	Siliceous
	Channel substrate:	Sand	Sand
	Alkalinity:	>1 meq/l	-
	Slope:	-	< 1 m/km
	Stream velocity:	-	< 0,5 m/sec
Medium-sized lowland streams Common type: R-C4 NL type: R6	Catchment area:	100 - 1000 km^2	100 - 200 km^2
	Altitude:	<200 m	-
	Geology:	Mixed	Mixed
	Channel substrate:	Gravel and sand	Clay and sand
	Alkalinity:	>2 meq/l	-
	Slope:	-	< 1 m/km
	Stream velocity:	-	< 0,5 m/sec

Assessment concept

The assessment concept of the Dutch method is more or less similar to those of the intercalibrated methods. All classifications are based on indicator species responding to anthropogenic stress, both eutrophication and hydromorphological degradation of the river. All assessment methods are also focussing on the same community structures, vegetation zones and (within the common types) habitat characteristics and life forms.

As discussed Birk & Willby (2011), there are differences in the concepts of the national methods that have effect on the feasibility of the intercalibration. For that reason some methods had to be (partly) excluded from the intercalibration process in this round.

The first version of the Dutch method differed from all other concepts in the aspect that it was designed to indicate any loss of species and structures, even though it differentiated between sensitive and tolerant species. After revision the species composition metric became more similar to the concept of the other metrics in which an index is calculated based on positive, negative and indifferent indicator species.

The Dutch slowly running lowland rivers are, like in other lowlands (e.g. Denmark, Flanders, Northern Germany), (assumed to be) typically mesotrophic at reference conditions, and loss of biological quality is more related to hydromorphological degradation than eutrophication and is less readily assessable with metrics based on trophic status.

Despite the differences in concepts, the species composition metrics in the national methods could be intercalibrated for most of the countries anyway. This was because the results of the metrics showed that they are similar enough to compare (Birk & Willby, 2011). The original Dutch method did not correlate significantly and intercalibration was therefore regarded as not feasible in this phase. The revised method showed a much higher correlation with the Pseudo-Common Metric, and is therefore more likely to meet the correlation criteria.

Features specific to the Dutch method comprise the assessment of the total cover of growth forms: submerged, floating leaved, emergent, filamentous algae and lemnids, as well as the tree cover at the banks. The assessment of the growth forms cover contributes 50% of the macrophyte assessment result. Nevertheless, these features still allow for a high correlation with other methods already intercalibrated, as demonstrated below. The correlation of the compete method is even higher than when using species composition alone.

2.4 Selecting fitting case

The completed river macrophyte intercalibration exercise used Option 3 to compare the results of the national assessment methods. This means that not a common biological metric was used in the completed exercise, but a so-called 'pseudo-common metric' (PCM), which is a way to facilitate direct comparisons of the methods.

Benchmark standardisation was used to identify and remove differences among national assessment methods that are not caused by anthropogenic pressure but by systematic discrepancies (due to different methodology, biogeography, typology etc.) (Birk & Willby, 2011).

Key to successful fitting in the intercalibration is the identification of a BRINC, i.e. the best-related and intercalibrated national classification method. This BRINC can be found by comparing the assessment results of all intercalibrated methods with the assessment results of the revised Dutch method, using Dutch macrophyte- and pressure data. The next chapter is about data collecting and choosing the BRINC.

Birk & Willby (2011) and also Birk & Van de Weyer (2015) report that the benchmark standardisation approach used was the 'continuous benchmarking', but in fact this showed not to be successful due to lack of sufficient samples in all ranges of pressure state and assessment results. 'Alternative benchmarking' was eventually used with a selection of benchmark sites within a 'window' of comparable status. The criteria used for selecting sites within this window was a good status, or moderate but close to good status, on the national assessment method, and none of the pressures having a high level of impact.

Therefore the correct fitting procedure is identified as Case B1 (Birk et al, 2013; Willby et al, 2014).

3 Data

Data were initially collected and used in the intercalibration exercise (Birk et al., 2007; Birk & Willby, 2011). The data were assessed by all involved assessment methods. The revised Dutch method was calibrated using the outcomes of these assessment results. For the fitting procedure, however, the number of samples was too low. There were 14 samples of type R5 (R-C1) and eight samples of the type R6 (R-C4), and no pressure data were collected for benchmarking.

A new set of biological and chemical data was derived from the national database for water samples Limnodata Neerlandica (PBL, 2013). A total of 3353 macrophyte samples in small rivers (NL type R5 and R6) taken between 1980 and 2013 were available. 89 samples of the type R5 (R-C1) and 55 samples of the type R6 (R-C4) were selected based of a number of criteria:

- Samples were taken following a procedure compatible with the European Standard CEN 14614.
- Data for all relevant chemical parameters (orthophosphate, nitrate, ammonium, biological oxygen demand, conductivity) were also available for the sample sites in the same year of macrophyte sampling.
- Samples without indicator species were excluded.
- A maximum of two samples were used in the same waterbody, and if two were used then as far apart as possible in space or time.
- Maximum spread in quality classes.
- Maximum geographical spread.

Biological data consisted of a list of species found at the sampling sites with an indication of abundance expressed as cover percentage or as abundance class 1-9 according to Stowa (Table 3). The Dutch metric on species composition can be calculated with these data only after conversion of the abundance scale to a three classes scale as is indicated in Table 3. Also some taxonomical issues had to be resolved with older data.

Table 3. Stowa classes for species abundance in macrophyte samples, with Tansley and Braun-Blanquet code equivalents, mean cover percentage for conversion to growth form cover and scale 1-3 conversion for the species composition metric.

Class	Abundance	Tansley	Braun Blanquet	Cover (%)	Metric
1	Rare	R	r	1	1
2	Occasionally	O	+	2	1
3	Locally frequent	LF	1	4	1
4	Frequent	F	2a	8	2
5	Locally abundant	LA	2b	18	2
6	Abundant	A	2m	25	2
7	Locally dominant	LD	3	35	2
8	Co-dominant	CD	4	60	3
9	Dominant	D	5	85	3

The Dutch metric on abundance of growth forms needs data on total cover of all species in the specific growth forms (submerged, emergent, floating not lemnids nor algae, lemnids, floating filamentous algae, shading tree cover at the banks). These data are collected only since the publication of the metric and are not available in Limnodata Neerlandica. For this study these abundance data of five of the six growth forms were estimated from the abundance data of the species using the mean cover percentage per scale unit (Table 3) and nominal growth form for each species. Species that are often found in different growth forms (such as *Nuphar lutea, Sagittaria sagittifolia* etc.) were added proportionally in all relevant growth forms. Tree cover was estimated by expert judgement (in most sites completely absent).

Land use in the river catchment above each sampling site was acquired from the CORINE-database, update 2006 (EEA, 2006). All Corine Land Cover (CLC) categories were combined into four as has been done in the intercalibration exercise: urban (all CLC categories class 1), agricultural with high impact (CLC codes 2.1, 2.2, 2.4.1, 2.4.2), agricultural with low impact (CLC codes 2.3.1, 2.4.3, 2.4.4), natural (CLC codes 3.1.1, 3.1.2, 3.1.3, 3.2, 3.3, 4 and 5).

The downloaded vector files were combined in GIS software (ARCMAP) with both vector files of the sampling sites, constructed from the coordinates given in Limnodata, and vector files of the waterbodies, collected by Informatiehuis Water (2014). Percentages of land use within one of these four categories of relevant area upstream of the sampling point were estimated manually and rounded to 10 % accuracy (or 5 % when low or high).

Data on orthophosphate, nitrate, ammonium, biological oxygen demand and conductivity were usually stored in the database with different site identification codes than biological data. Matching of these data was done by comparing geographical reference coordinates. In many cases these were identical or almost identical. If almost identical the sites were still assumed to be identical when the description of the sites were the same. In other cases the samples were not taken at exactly the same, but still comparable, locations. The latter samples had to be matched manually. Samples were matched if they were in the same waterbody, the distance between the sites did not exceed 500 m and the site was in the same section of the river. Chemical data were usually collected several times during the year; we averaged data from samples taken in the summer half year (1 April- 31 October) in line with the water quality assessment methods for these parameters.

3.1 Assessment and comparison

All biological data were assessed with the Dutch assessment tool QBWat (Pot, 2014). The data were assessed with the intercalibrated methods of UK, FR, FL, DE and PL using a tool specifically developed by Sebastian Birk for the intercalibration exercise. Some samples were omitted from further testing when at least one of the intercalibrated methods was not able to give a result because of lacking relevant indicators. 73 samples of the type R5 (R-C1) and 24 samples of the type R6 (R-C4) got a result with all metrics.

Table 4. Abbreviations used for metrics used in this report.

Metric	Meaning
UK	British Ecological Classification of Rivers using Macrophytes, LEAFPACS
FR	French Biological Macrophyte Index for Rivers, IBMR
PL	Polish Macrophyte Index for Rivers, MIR
FL(KKB)	Flemish macrophyte assessment system (for small brooks in The Kempen)
FL(KB)	Flemish macrophyte assessment system (for small brooks)
FL(GKB)	Flemish macrophyte assessment system (for large brooks in The Kempen)
FL(GB)	Flemish macrophyte assessment system (for large brooks)
DE(o)	German assessment system, PHYLIB (for brooks < 30 cm deep)
DE(d)	German assessment system, PHYLIB (for brooks > 30 cm deep)
NL	Dutch revised metric
NLab	Dutch revised metric partly, only regarding abundance of growth forms
NLsp	Dutch revised metric partly, only regarding species composition

The assessment results correlated remarkably low with the results of the intercalibrated metrics for the R5 (R-C1) type compared to the results for the R6 (R-C4) (Tables 5 and 6). For acceptability, r ≥ 0.50 is required. With these samples not only the Dutch metrics, but also the German metric correlates insufficiently with the other intercalibrated metrics. The correlation is much lower than in the assessment results for the 22 samples used in the earlier intercalibration exercise (Table 7).

Table 5. Pearson correlation coefficients (r) between assessment results of the various metrics for type R5 (R-C1), italics: unacceptable

	UK	FR	PL	FL(KKB)	FL(KB)	DE(o)	NL	NLab	NLsp
UK	1	0.65	0.65	0.68	0.59	*0.44*	*0.27*	*0.02*	*0.39*
FR	0.65	1	*0.49*	0.61	0.50	*0.38*	*0.17*	*0.02*	*0.24*
PL	0.65	*0.49*	1	0.58	0.69	0.50	*0.37*	*0.08*	0.50
FL(KKB)	0.68	0.61	0.58	1	0.80	*0.38*	*0.33*	*0.10*	*0.42*
FL(KB)	0.59	0.50	0.69	0.80	1	*0.27*	*0.29*	*0.00*	*0.44*
DE(o)	*0.44*	*0.38*	0.50	*0.38*	*0.27*	1	*0.26*	*0.13*	*0.27*
NL	*0.27*	*0.17*	*0.37*	*0.33*	*0.29*	*0.26*	1	0.78	0.81
NLab	*0.02*	*0.02*	*0.08*	*0.10*	*0.00*	*0.13*	0.78	1	*0.27*
NLsp	*0.39*	*0.24*	0.50	*0.42*	*0.44*	*0.27*	0.81	*0.27*	1

Table 6. Pearson correlation coefficients (r) between assessment results of the various metrics for type R6 (R-C4), italics: unacceptable

	UK	FR	PL	FL(GKB)	FL(GB)	DE(d)	NL	NLab	NLsp
UK	1	0.68	0.70	0.66	0.70	0.62	0.60	0.45	0.52
FR	0.68	1	0.81	0.59	0.64	0.53	0.57	0.39	0.53
PL	0.70	0.81	1	0.49	0.54	0.59	0.72	0.71	*0.47*
FL(GKB)	0.66	0.59	0.49	1	0.97	*0.46*	0.50	0.27	0.53
FL(GB)	0.70	0.64	0.54	0.97	1	*0.45*	*0.46*	*0.29*	*0.46*
DE(d)	0.62	0.53	0.59	*0.46*	*0.45*	1	0.52	*0.48*	*0.36*
NL	0.60	0.57	0.72	0.50	*0.46*	0.52	1	0.79	0.83
NLab	*0.45*	*0.39*	0.71	*0.27*	*0.29*	*0.48*	0.79	1	*0.31*
NLsrt	0.52	0.53	*0.47*	0.53	*0.46*	*0.36*	0.83	*0.31*	1

Table 7. Pearson correlation coefficients (r) between assessment results of the various metrics for all 22 Dutch samples originally used in the intercalibration exercise.

	UK	FR	PL	FL(KKB)	FL(KB)	DE(o)	DE(d)	NL
UK	1	0.71	0.55	0.83	*0.49*	0.55	0.51	0.65
FR	0.71	1	0.58	0.53	*0.33*	0.52	0.70	0.52
PL	0.55	0.58	1	0.51	0.63	0.63	0.53	0.53
FL(KKB)	0.83	0.53	0.51	1	0.68	0.67	*0.42*	0.77
FL(KB)	*0.49*	*0.33*	0.63	0.68	1	0.53	*0.22*	0.46
DE(o)	0.55	0.52	0.63	0.67	0.53	1	*0.47*	0.71
DE(d)	0.51	0.70	0.53	*0.42*	*0.22*	*0.47*	1	0.54
NL	0.65	0.52	0.53	0.77	0.46	0.71	0.54	1

When assessment results of the Dutch samples were compared with pressure values of the associate samples, again a remarkably low correlation was found, especially for type R5 (R-C1). The low correlation was not only found with the Dutch metric, but also with the other metrics (Tables 8 and 9). These findings suggest that something with the selected samples is wrong. Non-significant values (p>0.05) are indicated in italics.

Table 8. Pearson correlation coefficients (r) between assessment results of the various metrics and pressure values for type R5 (R-C1); italics: not significant (p >0.05)

	NH4	NO3	PO4	BOD5	COND	CLC-U	CLC-AH	CLC-AL	CLC-N
UK	*-0.10*	*-0.09*	*-0.13*	*0.07*	*0.00*	*-0.04*	*0.05*	*-0.03*	*0.00*
FR	*-0.13*	*-0.10*	-0.24	*-0.04*	*0.06*	*-0.14*	*0.02*	*0.05*	*-0.04*
PL	-0.31	*-0.11*	*-0.10*	*-0.10*	*-0.04*	*-0.19*	*-0.07*	*0.13*	*0.05*
FL(KKB)	-0.33	*-0.11*	*-0.12*	*-0.07*	*0.11*	*-0.19*	*-0.13*	0.21	*0.03*
FL(KB)	-0.28	*-0.16*	*-0.01*	*-0.06*	*0.02*	-0.24	*-0.06*	*0.13*	*0.09*
DE(o)	*-0.17*	-0.23	-0.22	*0.05*	*-0.08*	-0.27	*0.08*	*0.04*	*-0.02*
NL	*-0.19*	*0.00*	*0.05*	*-0.01*	*-0.14*	-0.21	*0.09*	*-0.05*	*0.13*
NLab	*-0.09*	*-0.07*	*0.14*	*0.10*	*-0.04*	*-0.07*	*0.15*	*-0.11*	*-0.02*
NLsp	-0.21	*0.07*	*-0.04*	*-0.10*	*-0.17*	-0.25	*0.00*	*0.03*	0.22

Table 9. Pearson correlation coefficients (r) between assessment results of the various metrics and pressor values for type R6 (R-C4); italics: not significant (p >0.05)

	NH4	NO3	PO4	BOD5	COND	CLC-U	CLC-AH	CLC-AL	CLC-N
UK	*-0.17*	*-0.02*	-0.48	*-0.10*	*-0.07*	*0.00*	*-0.05*	*0.07*	-0.41
FR	*-0.13*	*0.00*	-0.51	*-0.18*	*-0.01*	-0.33	*0.10*	*0.13*	-0.21
PL	*-0.13*	*-0.03*	-0.40	*0.03*	*0.05*	*-0.14*	*-0.06*	*0.16*	-0.36
FL(GKB)	-0.37	*0.12*	-0.25	-0.30	*-0.09*	*-0.10*	*0.04*	*0.06*	-0.41
FL(GB)	-0.34	*0.11*	-0.29	-0.24	-0.20	*-0.10*	*0.03*	*0.06*	-0.42
DE(d)	*0.04*	*-0.18*	*-0.13*	*0.01*	*0.00*	*-0.10*	-0.22	0.27	-0.43
NL	-0.34	*-0.05*	-0.50	*-0.10*	*0.03*	*-0.08*	*0.01*	*0.05*	*-0.18*
NLab	*-0.13*	*0.06*	*-0.14*	*0.17*	*-0.16*	*0.18*	*-0.03*	*-0.07*	-0.20
NLsp	-0.40	*-0.13*	-0.65	-0.31	0.20	-0.29	*0.06*	*0.14*	*-0.09*

3.2 Further selection of samples for type R5 (R-C1)

The metrics should have a reasonable correlation with pressure parameters. The already intercalibrated metrics showed this, as well as a reasonable correlation with each other. Low correlation of the intercalibrated metrics assessment results and the pressure values when applied to Dutch data can only mean that something is wrong with the Dutch data (for the R5 (R-C1) type).

We looked for inconsistencies in the Dutch data, especially for unexpected results of the Dutch metric and at least one of the other metrics at high or low values of the pressure parameters. Since the ammonium concentration (NH_4) correlated best (negatively) with the assessment results of all metrics, we looked particularly at this parameter. Second best would be orthophosphate concentration (PO_4) or urban land use in the catchment (CLC-U) but the values of these parameters are rather skewed, making finding of unexpected values difficult. Samples were omitted when they met one of the following criteria:

- high NH_4 and high assessment result with both NL and other metrics;
- low NH_4 and low assessment result with both NL and other metrics.

The remaining selection of 39 samples showed a good correlation between the Dutch metric and most of the other metrics.

Table 10. Pearson correlation coefficients (r) between assessment results of the various metrics for type R5 (R-C1), without doubtful samples, italics: unacceptable

	UK	FR	PL	FL(KKB)	FL(KB)	DE(o)	NL	NLab	NLsp
UK	1	0.62	0.67	0.64	0.60	*0.35*	*0.28*	*-0.04*	*0.40*
FR	0.62	1	0.56	*0.49*	*0.49*	*0.33*	*0.39*	*0.11*	*0.41*
PL	0.67	0.56	1	0.70	0.77	*0.49*	0.62	*0.16*	0.66
FL(KKB)	0.64	*0.49*	0.70	1	0.82	*0.31*	0.70	*0.29*	0.66
FL(KB)	0.60	*0.49*	0.77	0.82	1	*0.20*	0.61	*0.22*	0.61
DE(o)	*0.35*	*0.33*	*0.49*	*0.31*	*0.20*	1	*0.34*	*0.19*	*0.29*
NL	*0.28*	*0.39*	0.62	0.70	0.61	*0.34*	1	0.63	0.77
NLab	*-0.04*	*0.11*	*0.16*	*0.29*	*0.22*	*0.19*	0.63	1	*-0.02*
NLsp	*0.40*	*0.41*	0.66	0.66	0.61	*0.29*	0.77	*-0.02*	1

Table 11. Pearson correlation coefficients (r) between assessment results and pressure values for type R5 (R-C1), without doubtful samples; italics: p >0.05

	NH4	NO3	PO4	BOD5	COND	CLC-U	CLC-AH	CLC-AL	CLC-N
UK	-0.28	-0.23	-0.22	*-0.04*	*-0.10*	*-0.04*	*0.06*	*-0.06*	*0.07*
FR	-0.20	*-0.07*	-0.31	*-0.18*	*-0.08*	*-0.15*	*0.15*	*-0.09*	*0.04*
PL	-0.47	-0.28	*-0.13*	-0.24	-0.24	*-0.19*	*-0.19*	0.23	0.22
FL(KKB)	-0.63	-0.26	*-0.18*	-0.35	*-0.18*	-0.24	*-0.12*	0.21	*0.14*
FL(KB)	-0.43	-0.29	*0.00*	*-0.09*	*-0.10*	-0.23	*-0.10*	*0.14*	0.26
DE(o)	-0.28	-0.35	-0.32	*-0.17*	*-0.19*	-0.29	*0.10*	*0.10*	*-0.12*
NL	-0.65	*-0.15*	*-0.16*	-0.33	-0.23	-0.42	*-0.04*	0.24	*0.09*
NLab	-0.33	*-0.16*	*0.04*	*-0.16*	*-0.08*	*-0.16*	*0.06*	*0.10*	-0.23
NLsp	-0.56	*-0.07*	-0.24	-0.29	-0.22	-0.42	*-0.10*	0.23	0.30

The inconsistencies of the samples that are omitted could neither be explained by mistakes in the data collection process, nor by errors in the database Limnodata Neerlandica. If one of these would be the cause of the inconsistencies, it cannot be explained that all metrics, both the Dutch and the intercalibrated metrics, showed low correlation with each other and with pressure values.

A better explanation can be found in temporal asynchrony or rapidly changing conditions. Part of the sites show a much better or worse biological condition than chemical condition. Change in macrophyte composition and abundance takes some years after change in chemical condition. In many brooks in The Netherlands the chemical condition has changed in the period 1980-2010 and the specific samples most likely show a biological condition that did not fully respond to this yet.

After the final selection of data the remaining sites were still distributed reasonably well across the country (Figure 1). Rivers of these types are hardly found in the areas without samples in this selection.

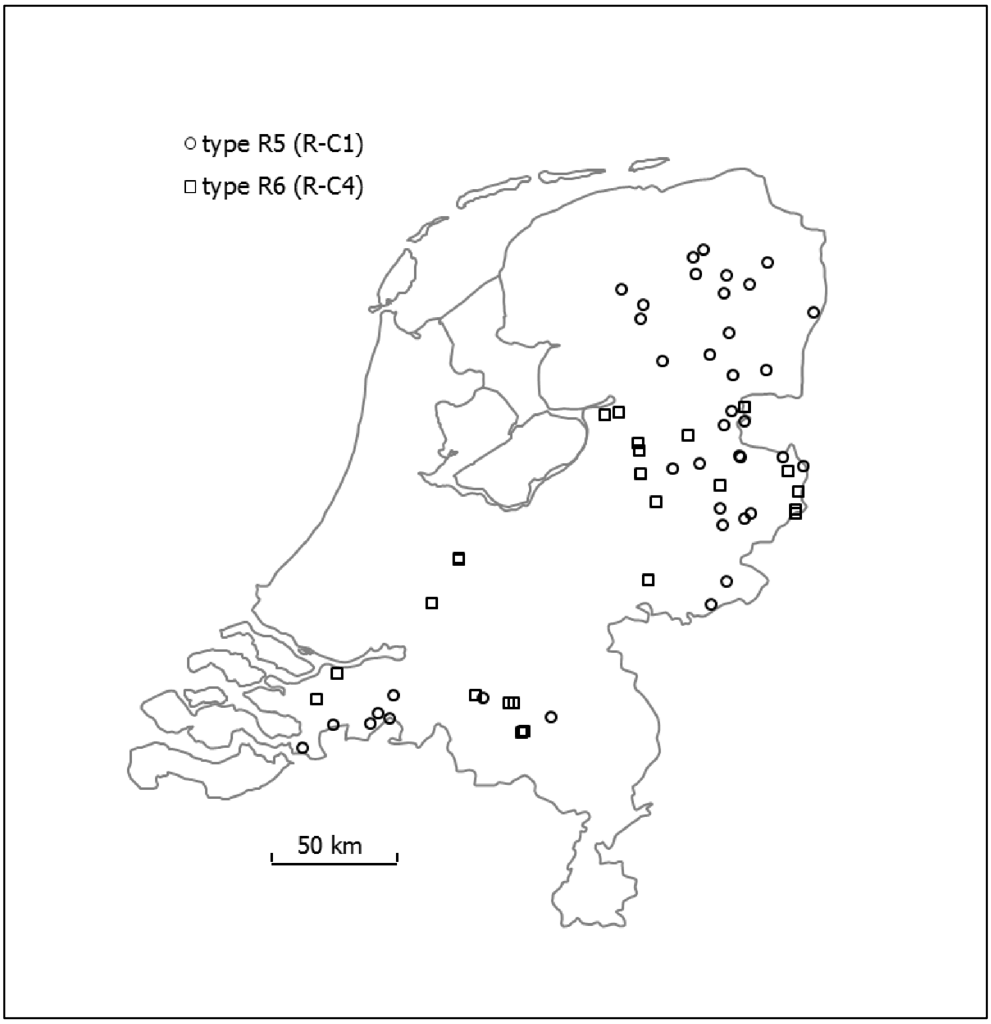

Figure 1. Location of the selected sites.

4 Fitting into intercalibration

4.1 Identification of the BRINC

As shown in Tables 6 and 10, the best-related and intercalibrated national classification methods (BRINC) are the Flemish (KKB) method for type R5 (R-C1) with r = 0.70 and the Polish method for type R6 (R-C4) with r = 0.72.

Methodological differences with the Flemish metric are relatively small and survey protocols are very similar. Biogeographical differences are small, because of the low distance, similar climate and rather similar geology and landscape. Like the Dutch metric, the Flemish metric uses abundance of certain growth forms in addition to species composition. This aspect however was excluded from the intercalibration for technical reasons and it will be excluded from this fitting process as well.

Methodological differences with the Polish metric are not larger than with most other metrics. The Polish metric only works on species composition. Abundance is only regarded on species level. Partial comparison with the Dutch species composition metric only could be considered, but as Table 6 shows, the correlation with this partial metric is much lower than with the full metric, as it is with most of the metrics. Biogeographical differences have to be dealt with through benchmarking.

4.2 Selecting benchmarking sites and criteria

Benchmarking in the completed intercalibration exercise was performed with a selection of benchmark sites within a 'window' of comparable status, being good status, or close to good status, on the national assessment method, and none of the pressures having a high level of impact. Approximately half of the selected Dutch samples fit in this window, being classified good with both the Dutch metric and the BRINC.

4.3 Benchmarking standardisation

The fitting procedure for case B1 only involves benchmark standardisation of the BRINC. The new metric is supposed to be fitted without consideration of further benchmarking, apparently because the selected sites already meet the criteria for benchmark sites of the BRINC. However, although the samples are accessed similarly with both the Dutch metrics and the BRINC, and pressure impact of the benchmark sites are low, this does not mean that there is no difference in the assessment results caused by other effects than anthropogenic pressure.

"The principle aim of benchmarking in intercalibration is to identify and remove differences among national assessment methods that are not caused by anthropogenic pressure but by systematic discrepancies (due to different methodology, biogeography, typology etc.). If such differences are ignored they may have an overriding effect on the comparability exercise. Therefore, the pressure

effects on the assessment scores (i.e. national EQRs) have to be controlled to disclose any remaining discrepancies." (Birk & Willby, 2011)

In this section we investigated comparability of the pressure effects by comparing the response of the BRINC to the most relevant pressure in two datasets; the first being the one used in the intercalibration exercise from the country of the BRINC, the other being the new Dutch dataset used in this fitting procedure. The most relevant pressure in this investigation is the best-correlating pressure, shown in Table 12.

Table 12. Pressure parameters correlating best with assessment results as shown in Table 9 and Table 11.

NL type	IC type	Pressure	r (NL)	r (BRINC)
R5	R-C1	NH4	-0.65	-0.63
R6	R-C4	PO4	-0.50	-0.40

Type R5 (R-C1)

It can be assumed that benchmarking is not needed because both metrics are developed for the same bioregion and with similar sampling procedure. However this assumption is tested anyway.

The best-correlating pressure with both Dutch metric and BRINC is Ammonium (NH_4). Disregarding two Flemish sites with exceptionally high values for NH_4 with high EQR, the BRINC responses to NH_4 do not differ too much in both datasets (Figure 2). The exceptionally high values could be the result of temporal asynchrony or rapidly changing conditions, like in some Dutch samples discussed in section 3.2

In general, both sets have comparable responses for the benchmark sites (all sites in the IC dataset, half of the sites in the NL dataset) and further Benchmark standardisation is therefore not required.

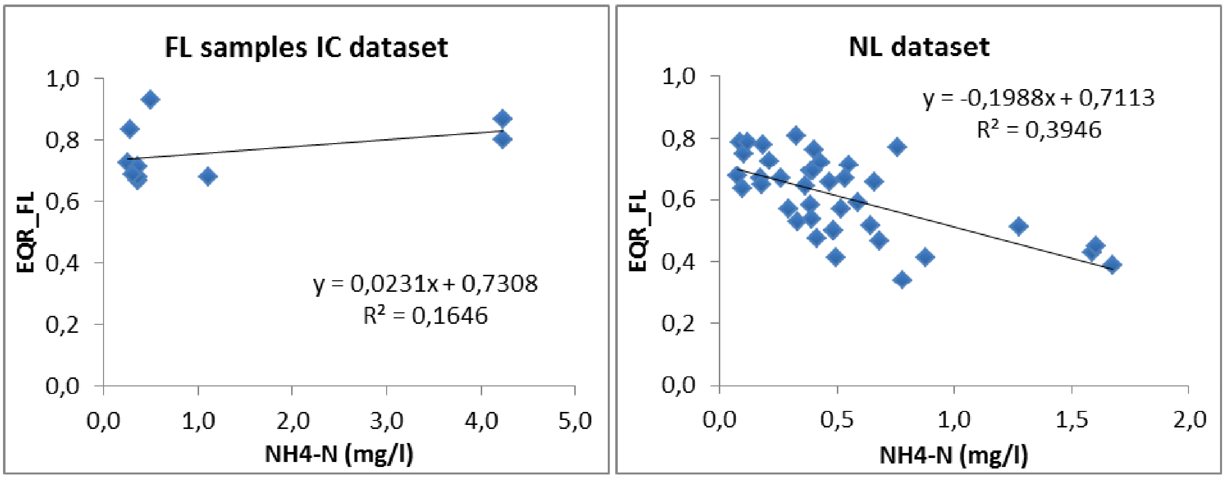

Figure 2. OLS regression between Ammonium-N and the FL metric using FL samples from the original intercalibration dataset (left) and using NL samples (right). The two high values in the IC dataset (all benchmark sites) may be caused by unusual Ammonium contents in the samples, since there is no effect on the EQR.

Type R6 (R-C4)

The best-correlating pressure with both the Dutch metric and the BRINC is orthophosphate (PO_4). Figure 3 shows that there is a huge difference between the response of the Polish metric to PO_4 in Polish samples and in Dutch samples.

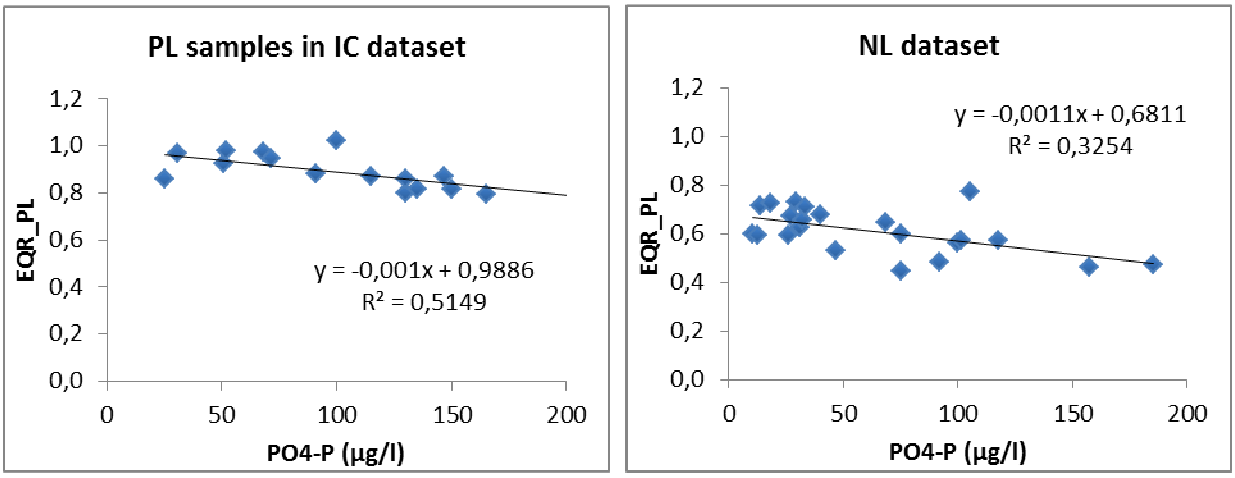

Figure 3. OLS regression between orthophosphate-P and the PL metric using PL samples from the original intercalibration dataset (left) and using NL samples (right).

This difference suggests that the Polish metric indicates near-reference conditions at low orthophosphate values in Poland with EQR=1.0, but with EQR=0.7 at low orthophosphate values in The Netherlands. Increasing pressure values result in lower EQR in both metrics but the differences remain.

If we fully use this difference in response to orthophosphate for the benchmark standardisation it would result in an inexplicable and unrealistical over-compensation. It would mean that EQR=0.6 on the Dutch metric, which is supposed to be the Good/Moderate boundary, would be translated to EQR=0.9 on the Polish metric, which is even higher than the Global Mean View of the High/Good boundary (Birk & Willby, 2011, Table 8.1).

Apart from biogeographical and possibly methodological differences between Poland and The Netherlands, there is also an alternative explanation for the huge difference in response to orthophosphate. Most probably a pressure other than orthophosphate, in particular hydromorphological degradation, is much more responsible for the decrease of biological quality in The Netherlands than in Poland. All brooks and small rivers have been straightened and channelized to improve water management, resulting in a big loss of habitat diversity. There is some relation between these pressures: improvement of water management was meant to facilitate improvement of land use and therefor the phosphate release into the rivers from agriculture increased. On the other hand, urban phosphate releases into river systems in The Netherlands have been reduced very successfully in the last decades and phosphate application in agriculture is very much optimized. Actual response to phosphate as indicated by macrophytes, if specifically recognizable, might even be the result of historical impact.

Recognizing the difference in response to orthophosphate, but mostly explained by the effect of other pressures, we propose to apply only one third of the effect of the difference to the benchmark standardisation. It is most likely that two third of the difference can be explained by other pressures. This results in a factor 0.9 to apply to the formula that predicts the Dutch metric class boundaries on the BRINC scale in following sections.

4.4 Global mean translated to the BRINC

The BRINC's both show some class boundary bias in the intercalibration exercise. These are summarized in Table 8.1 of Birk & Willby (2011). A bias up to 0.25 is accepted in the intercalibration exercise, but in this fitting process we need to compare with the Global Mean View without bias. Global Mean View (GMV) of the BRINC expressed on the scale of the BRINC is calculated from boundaries and class width of the Dutch method and de Bias of the BRINC as follows:

GMV = Boundary – (Class width * Bias)

As example for the High/Good boundary of type R5/R-C1:

GMV = 0.800 – ((1.000 – 0.800) * - 0.26) = 0.852

Table 13. Boundaries, bias and Global Mean View of the BRINC expressed on the scale of the BRINC (using the information given in Table 8.1 of Birk & Willby. 2011).

NL type	IC type	BRINC		H/G	G/M
R5	R-C1	FL (KKB)	Boundary	0.800	0.600
			Bias	-0.26	-0.31
			GMV	0.852	0.662
R6	R-C4	PL	Boundary	0.900	0.650
			Bias	0.05	-0.21
			GMV	0.888	0.703

4.5 Predicting the Dutch metric class boundaries on the BRINC scale

The next step is to calculate OLS regression to determine the relation between the Dutch metric and the BRINC. The result of this, using the assessment results of the selected samples as graphically represented in Figure 4, is:

- EQR (FL) = 0.6916 * EQR (NL) + 0.2087
- EQR (PL) = 0.6823 * EQR (NL) + 0.2210

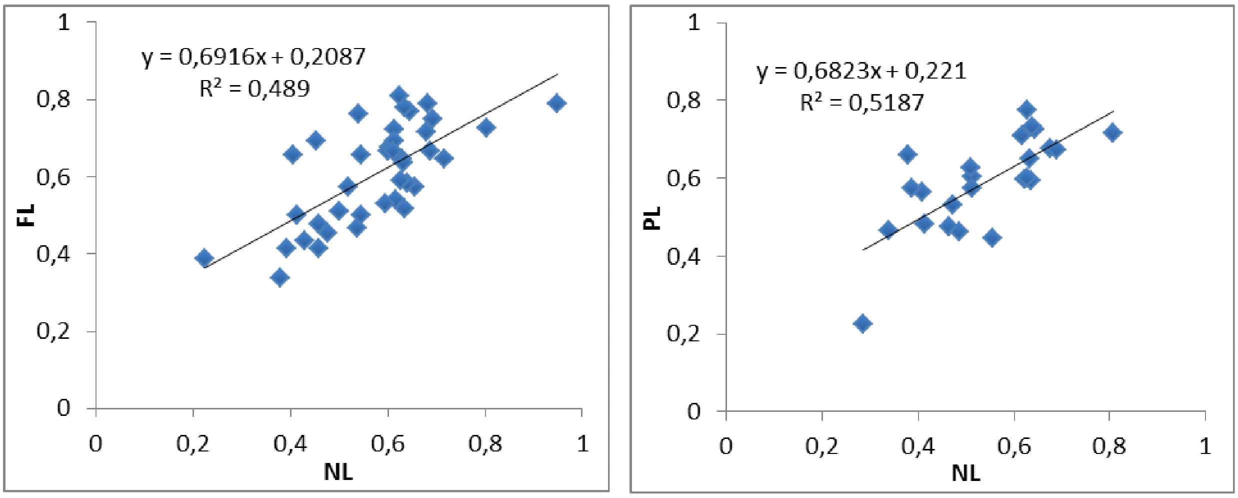

Figure 4. OLS regression between the Dutch metric and the BRINC for type R5 (R-C1) left and R6 (R-C4) right.

The OLS regression formulas are used to project the class boundaries of the Dutch metric onto the BRINC. As example for the Flemish metric for the High/Good boundary:

EQR = 0.6916 * 0.8 + 0.2087 = 0.76

As a result of benchmark standardisation we apply an extra factor 0.9 for the R6 (R-C4) type to predict the boundaries on the Polish metric for sites in The Netherlands (PL_bm).

Table 14. NL boundaries projected on the scale of the BRINC

NL type	IC type	scale	Reference	H/G	G/M	M/P
R5	R-C1	NL	1.00	0.80	0.60	0.40
		FL (KKB)	0.90	0.76	0.62	0.49
R6	R-C4	NL	1.00	0.80	0.60	0.40
		PL	0.90	0.77	0.63	0.49
		PL_bm	1.00	0.86	0.70	0.54

4.6 Calculating the class boundary bias of the Dutch method

The class boundaries of the Dutch method, projected on the scale of the BRINC (Table 14) are then compared with the Global Mean View.

- EQR (BRINC) is the translated NL boundary on the BRINC Scale
- GWV (BRINC) is the Global Mean View of the boundary expressed on the BRINC Scale
- Bias is calculated as the difference between these two, expressed as a fraction of the class width in EQR (BRINC)

According to the intercalibration comparability criteria the difference should not exceed a quarter of the class width. This means that the bias should be between 0.25 and - 0.25.

Table 15. NL boundaries projected on the scale of the BRINC for R5 (R-C1)

Boundary	EQR NL	EQR FL	GMV FL	Differs	Class width	Bias
Reference	1	0.90				
H/G	0.8	0.76	0.852	-0.09	0.14	-0.651
G/M	0.6	0.62	0.662	-0.04	0.14	-0.277
M/P	0.4	0.49				

Table 16. NL boundaries projected on the scale of the BRINC for R6 (R-C4) after benchmark standardisation

Boundary	EQR NL	EQR PL	GMV PL	Differs	Class width	Bias
Reference	1	1.00				
H/G	0.8	0.85	0.888	-0.04	0.15	-0.234
G/M	0.6	0.70	0.703	-0.00	0.15	-0.014
M/P	0.4	0.55				

As shown in Tables 15 and 16 then Dutch metric for type R5 (R-C1) is too relaxed, having a bias exceeding the |0.25| limit, and for type R6 (R-C4) no adjustment is needed.

4.7 Adjusting the class boundaries of the Dutch method

The adjustment step in the procedure is only applied to type R5 (R-C1). Firstly, the Good/Moderate class boundary has to be adjusted until |bias| <0.25 and then the High/Good class boundary has be to adjust likewise.

1. Adjusting the class boundary for Good/Moderate to EQR = 0.61 on the NL scale is enough to raise the value of the bias above -0.25.
2. Class boundary for High/Good has to be adjusted to EQR = 0.88 to reduce bias until |bias| <0.25, while Reference on the NL metric is increased to EQR=1.15 to keep equal class width. This implies that the original Reference value was chosen too relaxed and also should be redefined.
3. Increasing the Good class width by raising the High/Good boundary results in a reduction of the Good/Moderate boundary bias because of the increase of the class width. Therefore the adjustment can be reversed: EQR=0.6 then shows a bias of -0.198.
4. As a result of the reversal even the High/Good boundary can be lowered to EQR=0.87 and Reference can be lowered to EQR=1.14 to keep equal class width. Bias for Good/Moderate becomes -0.205 and bias for High/Good is also just above -0.25.

The final results of the adjustments are shown in table 17.

Table 17. NL adjusted boundaries projected on the scale of the BRINC for R5 (R-C1)

Boundary	EQR NL	EQR FL	GMV FL	Differs	Class width	Bias
Reference	1.14	1.00				
H/G	0.87	0.81	0.852	-0.04	0.19	-0.223
G/M	0.60	0.62	0.662	-0.04	0.19	-0.205
M/P	0.33	0.44			0.19	

To meet the intercalibration comparability criteria for type R5 (R-C1) the class boundaries of the Dutch metric have to be adjusted to:

- High/Good: EQR = 0.87 on the NL scale
- Good/Moderate: EQR = 0.60 on the NL scale
- EQR = 1.0 on the NL scale does not reflect Reference conditions

The Dutch method uses a formula for assessment of species composition that can result in a value higher than 1.0, but higher values are levelled down to 1.0. Adjustment of the constant that determines slope in the formula serves the adjustment. This constant is defined differently for every type and can be adjusted for R5 and comparable Dutch types without affecting the assessment of other Dutch types.

5 Summary

In this report we documented the fitting procedure of the revised Dutch assessment method using macrophytes to the results of the completed Central-Baltic rivers' intercalibration exercise. The method successfully passed the tests of intercalibration feasibility and WFD compliance.

A set of biological and chemical data from the national database for water samples Limnodata Neerlandica was used to compare the quality class boundaries of the method with the global mean view of the completed exercise. There were almost no samples assessed as High quality involved in this exercise because there are no (nearby) reference sites found in The Netherlands. Some samples had to be excluded because of inexplicable inconsistences that were most probably caused by temporal asynchrony or rapidly changing conditions when sampling.

We demonstrated that the Dutch metric should be subject to benchmark standardisation because the classification method that was used for comparison showed different assessment results with Dutch data than with data from its own country with the same known pressure impact. We made plausible that the remaining differences are caused by anthropogenic pressures that are less accounted for, if at all, in other countries.

The analyses showed the necessity to adjust the High/Good boundary and the view on reference conditions of the method for the sandy lowland brooks (intercalibration type R-C1). This can be achieved by adjusting the formulas to calculate EQR for Dutch type R5 (and comparable) in such a way that:

- Higher response of the indicator species metric should result in EQR=1
- The slope of the relation between response of the indicator species and metric result has to be adjusted in such a way that the assessment results of sites at the Good/Moderate boundary do not change.

6 Literature

Birk, S. & K. van de Weyer (1015). Fitting the Assessment System for Rivers in Northrhine-Westphalia (Germany) using Macrophytes to the results of the completed Central-Baltic rivers' intercalibration exercise. Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen (LANUV NRW), Essen.

Birk, S. & N. Willby (2011). CBrivGIG Intercalibration Exercise "Macrophytes" – WFD Intercalibration Phase 2: Milestone 6 Report. Joint Research Institute, Ispra (IT): 41 pp.

Birk, S., N. Willby, C. Chauvin, H. Coops, L. Denys, D. Galoux, A. Kolada, K. Pall, I. Pardo, R. Pot, D. Stelzer (2007). Report on the Central Baltic River GIG Macrophyte Intercalibration Exercise.

Birk, S., N. Willby, S. Poikane & W. van de Bund (2013a). Procedure to fit new or updated national classification of ecological status to the results of a completed intercalibration exercise. WG Ecostat. Version 3.0

Birk, S., N.J. Willby, M.G. Kelly, W. Bonne, A. Borja, S. Poikane & W. van de Bund (2013b). Intercalibrating classifications of ecological status: Europe's quest for common management objectives for aquatic ecosystems. Science of the Total Environment 454-455 (2013) 490–499.

EEA (2006). Corine Land Cover 2006 seamless vector data, European Environment Agency.(http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-3, downloaded 13-11-2014)

European Commission (2011). Guidance document on the intercalibration process 2008–2011. Guidance Document No. 14. Implementation strategy for the Water Framework Directive (2000/60/EC). Technical report-2011-045.

Informatiehuis Water (2014). Shape files of linear waterbodies, owm_SGBP2_20140507_lijn, http://www.waterkwaliteitsportaal.nl/Beheer/Rapportage/Publiek?viewName=Bronbestanden&jaar=2014&maand=Mei (downloaded 30-10-2014).

PBL (2013). Limnodata Neerlandica, database for biological and chemical monitoring data in The Netherlands since 1980 (received 11-10-2013).

Pot, R. (2014). QBWat, Version 5.31. http://www.roelfpot.nl/qbwat

Willby, N., S. Birk, S. Poikane & W. van de Bund (2014). Water Framework Directive Intercalibration Manual. Procedure to fit new or updated classification methods to the results of a completed intercalibration. Joint Research Institute, Ispra (IT): 33 pp. http://publications.jrc.ec.europa.eu/repository/handle/JRC89002