Vlaanderen
is landbouw & visserij

ILVO Mededeling 191

May 2015

Intercalibration report for benthic invertebrate fauna of the North East Atlantic Geographical intercalibration group for Coastal Waters (NEA 1/26)

ILVO
Instituut voor landbouw-
en visserijonderzoek

www.ilvo.vlaanderen.be

Intercalibration report for benthic invertebrate fauna of the North East Atlantic Geographical intercalibration group for Coastal Waters (NEA 1/26)

ILVO MEDEDELING 191

May 2015

ISSN 1784-3197

Wettelijk Depot: D/2015/10.970/191

Gert Van Hoey

Wendy Bonne

Fuensanta Salas Herrero

Intercalibration report for benthic invertebrate fauna of the North East Atlantic Geographical Intercalibration Group for Coastal Waters (NEA 1/26)

- Final report-

Gert Van Hoey[1], Wendy Bonne[2], Fuensanta Salas Herrero[3]

Contributers:
Denmark: Alf Josefson
France: Stanislas Dubois, Nicolas Desroy, Rèmi Buchet, Marie Claude ximenes
Ireland: Francis O'Beirn
Germany: Jan Witt, Karin Heyer
Netherlands: Willem van Loon, Hans Ruiter
Portugal: Joao Carlos Marques, Joao Neto
Spain: Angel Borja, Emilio Garcia, Araceli Puente
Norway: Gunhild Borgerson, Brage Ryge
UK: Graham Phillips, Alison Miles
Ireland: Francis O'Beirn

[1] ILVO Instituut voor Landbouw- en Visserijonderzoek [2] JPI OCEANS [3] JRC EUROPEAN COMMISSION

Acknowledgment

This work for the WFD intercalibration obligation of the Member States has received funding within the framework of JPI Oceans through the JPI Oceans pilot action “Joint funding of the Scientific Intercalibration exercise for the EU Water Framework Directive (WFD) coastal and transitional waters in the North-East Atlantic”. This work is the effort of many benthic experts of the Member States, which deliver the necessary information and feedback. Thanks to all contributors during the three phases of the intercalibration. Recording all contributors to this intercalibration work in the NEA-GIG is impossible, but at least the current responsible persons for each Member State are mentioned.

Intercalibration of biological elements for transitional and coastal water bodies.

North East Atlantic Geographical Intercalibration Group (NEA-GIG):
Coastal Waters NEA 1/26– Benthic Invertebrate fauna

Contents

1 Summary

This report gives a description of the intercalibration of the different benthic assessment approaches for soft sediment habitats (mud to muddy sands) in coastal waters in the North East Atlantic Geographical Intercalibration Group (NEA-GIG) for type NEA 1/26 (Exposed or sheltered, euhaline, shallow waters). The benthic assessment approaches of nine European Member States (Belgium, Germany, Denmark, France, Ireland, the Netherlands, Portugal, Spain and the United Kingdom) and Norway are intercalibrated. This process is executed under the form of a JPI oceans pilot action. A benthic assessment approach consists of an indicator algorithm, boundary settings and a reference setting approach. All necessary information on these aspects are summarized in this report based on the WISER database. The common benthic dataset constructed in phase I was used and contain data of all Member States. The EQR values determined for those samples are re-calculated based on the latest versions of the indicator algorithms and reference settings.

The benthic assessment approaches of Belgium, Germany, Denmark, France, United Kingdom/Ireland, the Netherlands, Norway, Portugal and Spain (m-AMBI) meet all WFD compliance criteria. Only, the benthic assessment approach BO2A of Spain does not meet the requirements of compliance criteria N°3, due to the lack of a diversity parameter within their approach (scientific justification available).

All methods can show in one or another way, a certain response to certain pressures. A uniform comparison of the responses of the different benthic assessment approaches is done on the Garroch Head sewage sludge disposal ground data set of the UK, instead of different independent comparisons available in the literature (e.g. Borja et al., 2009; Josefson et al., 2009; Fitch et al., 2014; and others). This is a very large dataset (180 samples) that is already standardized for IC purposes and with accompanying quantitative pressure information (organic and metal pollution concentrations) available. All benthic assessment approaches shows a clear and similar response (non-linear) to the pressure (copper).

The characteristics of the common dataset were analyzed and revealed that the datasets from the different Member States could be discriminated separately, with the North Sea datasets showing the highest similarity. The factor depth seems to delimit two different type of habitats within the common dataset were considered as two sub-types within the common dataset for the comparability analysis.

An important aspect for the intercalibration is the benchmarking. Due to the absence of reference sites and (semi-) quantitative pressure information per sample, the benchmark samples (least disturbed sites) were determined by expert judgment of the Member States.

All benthic assessment approaches, except BO2A, are finally meeting the comparability criteria of the intercalibration guidance, after raising the good/moderate boundary of Spain (m-AMBI) to a value of 0.63 (0.53).

2 Introduction

This report gives a technical description of the intercalibration of the different benthic assessment approaches for soft sediment habitats (mud to muddy sands) in coastal waters in the North East Atlantic Geographical Intercalibration Group (NEA-GIG) for type NEA 1/26 (Exposed or sheltered, euhaline, shallow waters). Nine European Member States (Belgium, Germany, Denmark, France, Ireland, the Netherlands, Portugal, Spain and the United Kingdom) and Norway (further in the text considered as Member State) are involved.

The intercalibration in the NEA-GIG region for coastal waters has a long history. In the first phase, a pioneering intercalibration exercise was executed, which showed a high consistency between the different benthic assessment approaches of United Kingdom, Spain (m-AMBI), Denmark and Norway on a common benthos dataset (Borja et al., 2007). In the second phase, when the intercalibration guidelines were developed, a re-run of the analyses of the coastal waters of phase I following the new comparability criteria was expected. However, this process could not be completed in phase II for several reasons. The main recommendation from the Review Panel on the intercalibration exercise for the coastal waters in the NEA-GIG region was that additional analyses should be done (including all methods and all Member States) to further refine the comparability (Davies, 2012). Currently, further clarifications/justification should be compiled to confirm the comparability of the NEA-GIG benthic assessment approaches. Therefore, in phase III, under the form of a JPI oceans pilot action (http://www.jpi-oceans.eu/intercalibration-eu-water-framework-directive), this process has been executed. The objectives of this action are:

- WFD method compliance documentation check, explanations of the justifications for assessment methods including specific parameters, reference conditions and the boundary setting procedure. Also to check or improve pressure-response relationships (1st and 2nd phase results are available).
- Provide an alternative benchmarking clarification, trying to take regional biological differences and sampling protocol differences into account, based on already available data or validated expert judgment.
- Check and improve comparability analysis (1st and 2nd phase results are available).
- Prepare and compile finalized intercalibration technical report from the several existing current reports.

This report compiles all the latest information regarding the benthic assessment approaches, boundary- and reference settings for each Member State and common dataset characteristics. Specific analyses were conducted to demonstrate the pressure-response relationships of the benthic assessment approaches, detect possible bio-geographical differences in the common dataset, perform an alternative benchmark delineation and the comparability analyses following the intercalibration guidelines (Guidance document 14: guidance document on the intercalibration process 2008-2011).

3 Description of national assessment methods

Within the NEA-GIG region for coastal waters, 10 benthic assessment approaches were defined (Table 1). A benthic assessment approach consists of an indicator algorithm, boundary settings and a reference setting approach. Some Member States used the same indicator algorithm (e.g. m-AMBI), but were considered as a separate benthic assessment approach due to different boundary or reference settings. Only United Kingdom and Ireland share the same benthic assessment approach, the IQI. Each benthic assessment approach is considered as a separate method in the intercalibration exercise.

Table 1. Overview of the national assessment methods

Member State		Method		WISER database	Included in this IC exercise
Belgium	BE	Benthic Ecosystem Quality Index	BEQI	X	Yes
Germany	DE	Multivariate AZTI's Marine Biotic Index	m-AMBI[1]	X	Yes
Denmark	DK	Danish Quality Index	DKI	X	Yes
France	FR	Multivariate AZTI's Marine Biotic Index	m-AMBI[2]	X	Yes
Ireland	ROI	Infaunal Quality Index	IQI	X	Yes
Netherlands	NL	Benthic Ecosystem Quality Index 2	BEQI2	X	Yes
Norway	NO	Norwegian Quality Index	NQI	X	Yes
Portugal	PT	Benthic Assessment Tool	BAT	X	Yes
Spain (m-AMBI)	ES-BC/C	Multivariate AZTI's Marine Biotic Index	m-AMBI	X	Yes
Spain (BO2A)	ES-A	Benthic Opportunistic polychaetes/amphipods index	BO2A	x	yes
United Kingdom	UK	Infaunal Quality Index	IQI	X	yes

[1] m-AMBI method, but other reference and boundary settings.
[2] m-AMBI method, but other reference settings.

3.1 Methods and required BQE parameters

The current intercalibration exercise is based on the latest versions of the indicator algorithms (Table 2). The EQR values determined for the samples within the common dataset are re-calculated based on those algorithms. The metric values (e.g. Shannon diversity, AMBI, S, etc.) were determined based on the latest version of the common benthic dataset, which was made available by Angel Borja (the NEA-GIG benthos lead in phase II). The metric AMBI is now determined in the same way for all benthic assessment approaches, which was not the case for the previous exercises (Borja et al., 2007). These recalculations have led to slightly different EQR values for the samples of the common dataset compared to the previous analyses. The advantage of this is that the analyses were standardized, transparent and are repeatable in time. The WFD requires the inclusion of certain metrics within the national assessment method for benthic invertebrates, which are summarized for each Member State in Table 3.

Table 2. Overview of the algorithms of the NEA-GIG benthic invertebrate indicators for intercalibration. H': Shannon wiener diversity; N: abundance; S: Number of species; AMBI: AZTI Marine Biotic Index.

	MULTIMETRIC	
BEQI (Belgium)	EQR=average (EQR species+ EQR density+ EQR similarity)	(Van Hoey et al., 2007) http://www.beqi.eu
DKI[1] (Denmark)	$\left(\frac{1-\frac{AMBI}{7}+\left(\frac{H'}{Hmax}\right)}{2}\right)*\left(\frac{\left(1-\frac{1}{N}\right)+\left(1-\frac{1}{S}\right)}{2}\right)$	(Borja et al., 2007)
NQI[2] (Norwegian)	$(0.5*\left(1-\frac{AMBI}{7}\right)+\left(0.5\frac{SN}{2.7}*\frac{N}{N+5}\right)$	(Rygg, 1985 and 2002)
IQI (UK, Ireland)	$IQI_{v.IV}=\left(\left(0.38\times\left(\frac{1-(AMBI/7)}{1-(AMBI_{\mathrm{Ref}}/7)}\right)\right)+\left(0.08\times\left(\frac{1-\lambda'}{1-\lambda'_{\mathrm{Ref}}}\right)\right)+\left(0.54\times\left(\frac{S}{S_{\mathrm{Ref}}}\right)^{0.1}\right)-0.4\right)\Big/0.6$	Philips et al., 2014
BEQI2[1] (The Netherlands)	EQR (ecotope) = 1/3 * [Sass / Sref] + 1/3 * [H'ass / H'ref] + 1/3 * [(6 – AMBI$_{ass}$)/(6-AMBI$_{ref}$)]	Van Loon et al., 2015
BO2A (Spain)	*BO2A=log10((fAO^3/(fA^3+1))+1) //* *EQR BO2A=(log(2)-BO2Ameasured)/(log(2)-BO2Areference).*	Dauvin & Ruellet, 2007
	MULTIVARIATE	
M-AMBI (Spain, France, Germany)	Factor analysis: S, AMBI, Shannon diversity index[1]	(Borja et al., 2004 and Muxika et al., 2007) http://ambi.azti.es
BAT (Portugal)	Factor analysis[4]: AMBI, Margaleff diversity index, Shannon diversity index[1]	Teixeira et al., 2009; Marques et al., 2009

[1] DKI, BEQI2, m-AMBI, BAT: Shannon diversity: log base 2.
[2] NQI: SN= LN(S)/LN(LN(N)); 2.7 is the ref value for SN
[3] fAO = frequency opportunistic annelid (fpo = frequency opportunistic polychaeta + fo = frequency oligochaeta) and fA = frequency amphipods
[4] Factor analysis BAT in *Statgraphics Plus 5.1 (rotation=varimax)*

The BEQI assessment approach does not allow a calculation of EQR values at sample level, due to the fact that it acts on habitat or water body level (Van Hoey et al., 2007; 2013). For the calculation of BEQI EQR values, a set of samples need to be considered for the assessment. Therefore, a separate comparison of the BEQI approach at higher level with the other benthic assessment approaches is executed (see separate intercalibration in phase I). Therefore, the samples of the other Member States are grouped per 10 (ideally), but at least to 5, to allow a BEQI calculation. The grouping of the samples is done, based on the fact that they are from the same site and same time (or time period). The EQR values of the pooled samples are based on the average value of the individual sample EQR's. The BEQI assessment approach determines the difference between a set of assessment and reference samples and classifies this according to the five classes of the WFD. The set of reference samples needs to be country/area/habitat specific; for this reason, the set of benchmark samples per country out of the common dataset is used as the set of reference samples. In this way, the principle of the BEQI approach is intercalibrated with the other benthic assessment approaches.

Table 3. Overview of the metrics included in the national assessment methods

Member state	Full BQE method	Taxonomic composition	Abundance	Disturbance sensitive taxa	Diversity	Bio-mass	Taxa indicative of pollution	Combination rule of metrics
Belgium	Yes	Yes, species composition by means of Bray Curtis similarity	yes	As species composition without pre-classifying species in classes.	Yes, number of species	Yes	Specific opportunistic species	Average of the four parameters
Denmark	Yes	Not strictly – only as groups (5) of different sensitivity	Abundance is included as correction factor and relative abundance of different sensitivity groups and proportional abundance in Shannon Wiener index	5 sensitivity classes (AMBI)	Yes, number of species and Shannon wiener index	No	Specific opportunistic species	Algorithm
Netherlands	Yes	Not strictly – only as groups (5) of different sensitivity	As relative abundance of different sensitivity groups and proportional abundance in Shannon Wiener index	5 sensitivity classes (AMBI)	Yes, number of species and Shannon Wiener index	No	Specific opportunistic species	Average of 3 univariately normalized indicator EQR scores
Norway	Yes	Not strictly – only as groups (5) of different sensitivity	Species abundance as correction factor (Ntot/Ntot+5) and relative abundance of different sensitivity groups	5 sensitivity classes (AMBI)	Yes, number of species	No	Specific opportunistic species	Weighted algorithm: 50% AMBI and 50% number of species-abundance
Portugal	Yes	Not strictly – only as groups (5) of different sensitivity	As relative abundance of different sensitivity groups and proportional abundance in Shannon Wiener index and Margalef index	5 sensitivity classes (AMBI)	Yes, Shannon Wiener index and Margalef index	No	Specific opportunistic species	Factorial analyses, calculating vectorial distances to reference conditions

Member state	Full BQE method	Taxonomic composition	Abundance	Disturbance sensitive taxa	Diversity	Bio-mass	Taxa indicative of pollution	Combination rule of metrics
Spain (m-AMBI), France, Germany	Yes	Not strictly – only as groups (5) of different sensitivity	As relative abundance of different sensitivity groups and proportional abundance in Shannon Wiener index	5 sensitivity classes (AMBI)	Yes, number of species and Shannon Wiener index	No	Specific opportunistic species	Factorial analyses, calculating vectorial distances to reference conditions
Spain (BO2A)	No	Not strictly – only as groups (2) of different sensitivity	As relative abundance of opportunistic polychaetes and amphipods	2 sensitivity classes (sensitive or tolerant)	No	No	Specific opportunistic species	No combination
United Kingdom / Ireland	Yes	Not strictly – only as groups (5) of different sensitivity	As relative abundance of different sensitivity groups and proportional abundance in Simpson index	5 sensitivity classes (AMBI)	Yes, number of species and Simpson index	No	Specific opportunistic species	Weighted algorithm: AMBI for 38%; Simpson for 8% and number of species 54%

3.2 Sampling and data processing

The method of taking the benthic samples and processing for the WFD Monitoring within the different Member States is outlined in detail in annex 1. The information is extracted from the online WISER project database, which compiles all information regarding WFD assessment methods (version of Birk et al., 2010; http://www.wiser.eu/results/method-database/). This database is subjected to change: an update will probably be made in the near future.

3.3 National reference conditions

The determination of the reference conditions is a delicate subject (Van Hoey et al., 2010; Birk et al., 2013). The ecological status in the WFD has to be measured as a deviation from a reference condition. These reference conditions need to correspond to largely undisturbed (='near-pristine') conditions (no or minor impact from human activities). Indeed, the lack of appropriate reference sites or robust historical datasets is one of the major problems addressed in the intercalibration exercises and in setting the good ecological status boundaries (Borja et al., 2007; 2009). Scientists are faced with virtual lack of undisturbed sites along the European coasts and estuaries, and historical data are not easily accessible (Borja et al., 2004). Reference settings will need to be based on clear stressor-response relationships, a knowledge of the 'naturalness' of the system; and expert judgment may also have a role to play (Van Hoey et al., 2010). As summarized in Table 4, all Member States used the best available information (e.g. areas with least disturbed conditions) and their expert judgment to delineate appropriate reference values for their metrics.

The reference values used to calculate the EQR values for each sample in the common dataset are listed in Table 5. Those reference values were considered appropriate values for the samples of the subtidal soft-sediment habitats within the common dataset by each Member State. Those values were applied per benthic assessment approach on the entire common dataset.

Table 4. Overview of the methodologies used to derive the reference conditions for the national assessment methods included in the IC exercise

Member State	Type and period of reference conditions	Number of reference sites	Location of reference sites	Reference criteria used for selection of reference sites
Belgium[1]	Expert knowledge, Historical data, Least Disturbed Conditions. Data period 1994-2012 Habitat-specific	No reference sites; the reference data per habitat is selected out of the available benthos data collected over the period 1994-2012.	No reference sites	The most appropriate data for each benthic habitat of the BPNS as reference is based on the following selection criteria: - The data must be collected in the period 1994-2012 on the BPNS. - Data collected in areas where a certain human activity (dredge disposal, sand extraction, wind-farm construction) can disturb the natural variability of the benthic characteristics were excluded. - To have a good temporal and spatial coverage of samples within the reference dataset, we tried to have a balanced sampling (similar number of samples) over the years and within the areas of the BPNS.
Germany	Expert knowledge, Historical data, Least Disturbed Conditions; reference time: 1959 up to now. Habitat-specific	subtidal coast: 17	different sites Wadden Sea of Lower Saxony	The communities at the sites had to correspond with description of the reference community description referring to a certain habitat.
Denmark	Least Disturbed Conditions (Sites the least impacted - farthest from impact source); Recent data from least impacted sites. Surface water type-specific	Depends on type, but typically 5-50 sites	n.a.	Reference community and impact factor close to background.
France	Expert knowledge, Historical data, Least Disturbed Conditions. Data period : 1995-2006 Habitat-specific	Bretignolles_S Morlaix1_S SSMF06_S (Rade de Cherbourg)	Channel & Atlantic	Expert knowledge and least disturbed conditions. The reference conditions for 3 metric component M-AMBI were defined by habitat type, based on recent data (last decade) collated on sites of French Atlantic and English Channel coasts, in particular data collected as part of the French benthic network (REBENT: http://www.rebent.org/).

Netherlands	Historical data for 1991-2006; (b) AMBI(ref) estimated as the 1 percentile value; theoretical bad values: S(bad) = 0; H'(bad) = 0; AMBI(bad) = 6. (c) Statistical modelling for S(ref) and H'(ref): 99 percentile of S and H' for large ecotope dataset (highest indicator value which is robustly not an outlier).	Not true reference sites, but least disturbed sites can be selected if necessary.	The Wadden Coast and Wadden Sea are less impacted areas, compared to the Dutch Coast and Voordelta coastal zones.	Not applicable because coastal waters in The Netherlands are always subject to at least some level of anthropogenic impact. However, least disturbed samples from distinct sampling locations can be selected based on expert judgment using information on pressures at the sampling locations.
Norway	Recent data from least impacted sites	n/a	Outer coast of Skagerrak, southern Norway.	Reference sites were selected by the following criteria: Deeper than 5m, limited fresh water influence (> 1km from nearest estuary) and of sufficient distance (based on expert judgment) from any known pollution sources, such as large cities or industrial activity.
Portugal	Existing near-natural reference sites, Expert knowledge, Historical data, Least Disturbed Conditions; Data period Outer Minho (CW NEA1) – Feb and Jul 2006; Praia do Garrao (CW NEA26) - Apr and Nov 2006. Habitat-specific	14 sites (7 H/G and G/M sites, 2 historical data sites, 5 outfalls data)	Outer Minho (CW NEA1) – Reference site High-Good; Praia do Garrao (CW NEA26) - Reference site High-Good	Reference condition samples were identified as being from least disturbed conditions, selected on the basis of a) unimpacted sites; and b) from impact gradient study control sites. Reference condition values for Margalef, Shannon-Wiener and AMBI were identified from the data. Data was used from sites with low levels of natural disturbance and outliers (e.g., those with anomalously high taxa numbers in contrast to the remaining data) were identified according to expert judgment and excluded.
Spain (m-AMBI)	Expert knowledge, Historical data, Modeling (extrapolation of model results); period 1995-2005. Habitat-specific	no specific number	Basque Country	Virtual locations, see: Muxika, I., A. Borja, J. Bald, 2007. Using historical data, expert judgment and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. Marine Pollution Bulletin, 55: 16-29.
Spain (BO2A)	Least disturbed conditions. Habitat-specific	No real reference sites, only a benchmark site	In front of the Doñana National Reserve (site code: 51C0090, water	Lowest impact of urban and industrial sewage and lowest amount of agriculture and urban land use.

			body wise code: 510001, aprox. Coordinates (DD ETRS89: longitude -6.601, latitude 36.879	
United Kingdom/ Ireland	Expert knowledge, Least Disturbed Conditions and Modeling (extrapolation of model results); Data from 1979 to 2003. Habitat-specific	No reference sites; >1000 sites from UK and Ireland are used for setting reference conditions		Reference condition samples were identified as being from least disturbed conditions, selected on the basis of a) expert judgement and b) from impact gradient study control sites. Reference condition values for AMBI, Simpsons and taxa number were identified from the data. Data was used from sites with low levels of natural disturbance and outliers (e.g., those with anomalously high taxa numbers in contrast to the remaining data) were identified according to expert judgement and excluded.

[1] Changed compared to the WISER input, based on Van Hoey et al., 2014 report.

Table 5. Overview of the reference values per benthic characteristics used in the intercalibration exercise.

		Parameters									
REFERENCE VALUES	Sample surface (m^2)	Number of taxa	Shannon (H' log2)	SN	Simpson	Margaleff	AMBI	Density (Ind/m^2)	Biomass (gWW/m^2)	Bray Curtis similarity	Combined ref value
Belgium	1[7]	153						2517.8	642.7	1	
Germany	0.1	34	3.65				0.597				
Denmark	0.1[2]		5				0				
France	0.9[4]	58	4				1				
Ireland	0.1	68			0.97		0.96[3]				
Netherlands	0.078	31[1]	3.8[1]				0.01				
Norway	0.1			0.27							
Portugal	0.1		4.1			5	0				
Spain (m-AMBI)	0.3[5]	42	4				1				
Spain (BO2A)											0
United Kingdom[6]	0.1	68			0.97		0.96[3]				

[1] The values for the Netherlands are based on the combined reference value for the three Dutch coastal zones together.

[2] It is from circa 0.1m^2 obtained by pooling 6-7 smaller (ca 0.013m^2)

[3] (1-(AMBI/7))= 0,96

[4] 9 replicates of 0.1 m^2

[5] 3 replicates of 0.1m^2

[6] these values are specifically set for fully marine subtidal muddy sand/sandy mud sediments from 0.1 m^2 grabs, sieved at 1 mm and using 2004 truncation rules

[7] The reference values are generated for each sample surface (from 0.1m^2 to max reference sample surface) based on a randomization procedure of the reference dataset for each boundary. The values shown in the table are those that generate an EQR value of 1. The values for the good/moderate boundary are 56 for number of taxa, 0.48 for Bray Curtis similarity, 179.9 and 6089.8 for biomass and 1182 and 7835 for density. A higher and a lower boundary value is used for biomass and density, due to the fact that two high values were also seen as a disturbance of the benthic system. In this report, the values for the muddy fine sand habitat for a sample surface of 1m^2 are reported.

3.4 National boundary setting

The reported information in WISER regarding the boundary setting procedure for each Member State is summarized in Table 7. Most Member States reported that they take the boundaries from phase I intercalibration (Borja et al., 2007; 2009) and no specific approach for H/G or G/M boundary was reported in WISER. The boundary values used in the intercalibration for the different assessment approaches were summarized in Table 6.

Table 6. The boundary values (High/good and Good/moderate) for the different assessment approaches as used in the intercalibration exercise. BC: Basque Country, C: Cantabria, A: Andalusia.

National Method	Denmark	UK/ROI	Spain (m-AMBI)	Norway	Portugal	the Netherlands	Germany	France	Spain (BO2A)	Belgium
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770	0,830	0,800
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530	0,500	0,600
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380	0,400	0,400
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200	0,200	0,200

Table 7. Explanations for national boundary setting of the national methods included in the IC exercise

Member State	Type of boundary setting	Specific approach for H/G boundary	Specific approach for G/M boundary	BSP: method tested against pressure
Belgium	Equidistant division of the EQR gradient. The moderate/poor and poor/bad reference value were determined by equal scaling (respectively 2/3 and 1/3 of the good/moderate reference value).	The boundary setting procedure is based on the output of the randomization procedure of the reference dataset. The reference value for the high/good boundary is determined based on the median value (number of species, similarity) or the 25th and 75th percentile (density, biomass) out of the permutation distribution.	The boundary setting procedure is based on the output of the randomization procedure of the reference dataset. The reference value of the good/moderate boundary is determined based on the 5th percentile (number of species, similarity) or on the 2.5th and 97.5th percentile (density, biomass) out of the permutation distribution of each parameter of the reference dataset.	

Germany	Boundaries taken over from the intercalibration exercise (Borja et al., 2007[1]). Calibrated against pre-classified sampling sites. The boundary setting procedure is in line with the WFD's normative definitions.			The boundaries were additionally adjusted by the assessment of expert judgment (Heyer 2007). The m-AMBI relates to pressures of sediment enrichment, eutrophication and hazardous substances (Muxika et al. 2007).
Denmark	Equidistant division of the EQR gradient. Using discontinuities in the relationship of anthropogenic pressure and the biological response.		Usually, the border between good and moderate EcoQS (G/M) is determined as some deviation from a reference situation. Reference data, however, are difficult to find. An alternative procedure is described to estimate the G/M border, not requiring reference data. Threshold values, where faunal structure deterioration commences, were identified from non-linear regressions between indices and impact factors. Index values from the less impacted side of the thresholds were assumed to come from environments of Good and High EcoQS, and the 5th percentile of these data was defined as the G/M border.	
France	Boundaries taken over from the intercalibration exercise (Borja et al., 2009) and calibrated against pre-classified sampling sites			See: Borja et al., 2009.
Netherlands		The Good/Moderate boundary of 0.58 is primarily derived from the initial G/M boundary for sheltered coastal waters		

		(Wadden Sea), which was estimated using expert judgment and set at 0.58.		
Norway	National boundaries (Molvær et al., 1993) adjusted following the intercalibration exercise (Borja et al., 2007)			
Portugal	Boundaries taken over from the intercalibration exercise.			AMBI ecological group proportions were established for samples over a pressure gradient (urban treated outfall). Initially, equidistant class boundaries were set and each AMBI EG proportion was calculated for i) the overall status and ii) the lower and upper quartiles of the data in each status. Where the AMBI EG proportions did not conform to those interpreted from the WFD Normative Definitions, the status boundary was adjusted towards the quartile that gave a more accurate representation. Boundaries were further optimized during Intercalibration Phase I.
Spain (m-AMBI)	Boundaries taken over from the intercalibration exercise (Borja et al., 2007)			Borja et al., 2009 & others.
Spain (BO2A)	Using paired metrics approach, using the frequency of opportunistic annelid and the frequency of amphipods as metrics. Moderate/Status: amphipod frequency (except Jassa) less than 0.45, and opportunistic	Dauvin & Ruellet (2007) use the limits of the AMBI index (Borja et al., 2000) proposed by Borja et al.(2004) to theoretically calibrate BOPA limits: High/Good Status: amphipod frequency (except Jassa) between 1 and 0.80, and	Good/Moderate Status: amphipod frequency (except Jassa) less than 0.80, and opportunistic polychaete frequency higher than 0.20.	Yes, quantitative; The methods relates to a pressure gradient of eutrophication (nutrient and organic matter enrichment and discharges).

	polychaete frequency higher than 0.55 - Poor/bad Status: amphipod frequency (except Jassa) less than 0.28, and opportunistic polychaete frequency higher than 0.72.	opportunistic polychaete frequency between 0 and 0.20.		
United Kingdom/Ireland	Boundaries taken over from the intercalibration exercise (Borja et al., 2007[1]).			AMBI ecological group proportions were established for samples over a sewage sludge disposal pressure gradient. Initially, equidistant class boundaries were set and each AMBI EG proportion was calculated for i) the overall status and ii) the lower and upper quartiles of the data in each status. Where the AMBI EG proportions did not conform to those interpreted from the WFD Normative Definitions, the status boundary was adjusted towards the quartile that gave a more accurate representation. Boundaries were further optimized during Intercalibration Phase I.

3.5 Results of WFD compliance checking

Table 8. WFD compliance checking criteria.

Compliance criteria	Compliance checking conclusions
1. Ecological status is classified by one of **five classes** (high, good, moderate, poor and bad).	Yes, for all benthic assessment approaches
2. High, good and moderate ecological status are set in line with the WFD's **normative definitions** (**Boundary setting procedure**)	Yes, for all benthic assessment approaches (see Table 24 and Table 7).
3. **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.	All Member States included the relevant parameters (see Table 3), except Spain (BO2A). They do not include a diversity parameter (2011-12-16technical_report_ NEA_CW_invertebrates_ES(AN)_Dec2011). A combination rule to combine parameter assessment is defined by all benthic assessment approaches.
4. 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, for all Member States (see Table 9 and Table 10)
5. The water body is assessed against **type-specific near-natural reference conditions**	No (see Table 4). Alternative benchmark conditions (based on a "least disturbed condition" criteria) had to be defined due to the absence of near-natural reference conditions in the intercalibrated type.
6. Assessment results are expressed as **EQRs**	Yes, for all benthic assessment approaches (see Table 3).
7. Sampling procedure allows for **represent-tative** information about water body quality/ecological status **in space and time**	In most cases, the monitoring is considered as representative by the Member State itself (see annex 1). This aspect is not confirmed by specific, standardized analyses to test their representativeness. Sampling procedures are outlined in general, but not linked with the running WFD monitoring programs.
8. All data relevant for assessing the biological **parameters** specified in the WFD's normative definitions are covered by the **sampling procedure**	Yes, for all benthic assessment approaches. The sampling procedure defined by each Member State allows the collection of species-abundance data (see annex 1), which is necessary to calculate all metrics of the different benthic assessment approaches.
9. Selected taxonomic level achieves adequate **confidence and precision** in classification	Yes, for all benthic assessment approaches, with some difference in taxonomic detail per Member State, but sufficient comparability (see annex 1). Taxonomy between Member States datasets is standardized for intercalibration purposes.

There can be concluded that all compliance criteria were met for the benthic assessment approaches of Belgium, Germany, Denmark, France, United Kingdom/Ireland, the Netherlands, Norway, Portugal and Spain (Basque and Cantabria region) (Table 8). Only, the benthic assessment approach BO2A of Spain does not meet the requirements of compliance criteria N°3, due to the lack of a diversity parameter within their approach. However, a scientific justification for this is presented in their separate intercalibration document (2011-12-16technical_report_NEA_CW_invertebrates_ES(AN)_Dec2011).

4 Results intercalibration feasibility checking

4.1 Typology

In the NE Atlantic, seven basic intercalibration types have been agreed upon. In this report the type NEA1/26 is taken into account (see outline of characteristics in Table 9).

Table 9. NEA GIG Intercalibration Type NEA1/26

New Type ID	***Name***	***Salinity***	***Tidal range (m)***	***Depth (m)***	***Current velocity (knots)***	***Exposure***	***Mixing***	***Residence time***
CW – NEA1/26	Exposed or sheltered, euhaline, shallow	Fully saline (> 30)	Mesotidal (1 - 5)	Shallow (< 30)	Medium (1 - 3)	Exposed or sheltered	Fully mixed	Days

The types above occur in Member States' waters as detailed below in Table 10, and compromise all NEA-GIG countries except Sweden.

Table 10. Member States sharing types

Type	***BE***	***DE***	***DK***	***ES***	***FR***	***IE***	***NL***	***NO***	***PT***	***SE***	***UK***
CW – NEA1/26	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**		**X**

For benthic invertebrates, all classification schemes intercalibrated relate only to the soft sediment infauna component. Differences occur in the reference conditions for the types; these are specific for the habitat type, and for some Member States (NL and DE), sometimes even specific for the water body. However, the basic metrics in each country's benthic assessment approach remain the same.

4.2 Pressures addressed

4.2.1 Sample level

All methods can show in one or another way, a certain response to certain pressures (Table 7). For benthic indicators also an abundant number of papers and reports are available that shows their pressure-response relation (e.g. Borja et al., 2009; Josefson et al., 2009; Fitch et al., 2014; and others). Therefore, it can be concluded that the response of a certain benthic assessment approach is slightly different from pressure to pressure type and from area to area. Unfortunately, no combined analyses has been made regarding the pressure-response relationship of the 10 benthic assessment approaches of the NEA-GIG region on a certain pressure dataset. Therefore, rather than summarizing the available literature regarding this subject, the pressure-response of the different benthic assessment approaches is tested on a large pressure dataset out of the common dataset. This allows to a uniform comparison of the responses of the different benthic assessment approaches, instead of different independent comparisons.

An appropriate dataset for this exercise was the Garroch Head sewage sludge disposal ground data set of the UK (provided by Marine Scotland), which is a very large dataset (180 samples) that is already standardized for IC purposes and with accompanying quantitative pressure information (organic and metal pollution concentrations) available. The elements (nitrogen, carbon, copper, zinc, lead and chromium) are correlated with each other and are the explanatory variables for the pollution gradient at Garroch Head. In the further analyses and figures, Copper is used as proxy for the pollution gradient at Garroch Head, due to the fact that it shows the highest correlation with the benthic assessment approaches (Table 11).

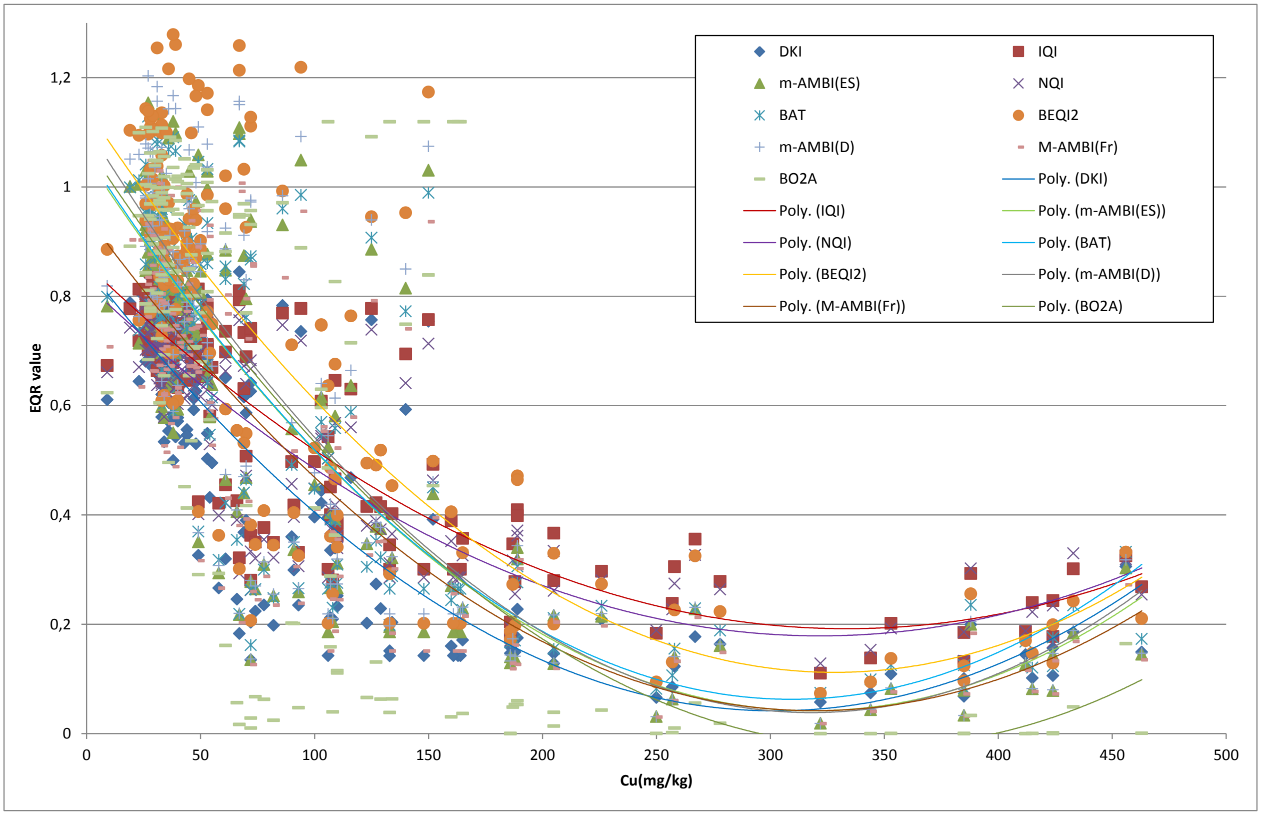


Figure 1. Correlation plot with trend line (polynomial 2nd order) between the different assessment approaches and Cu(mg/kg).

The different benthic assessment approaches shows no linear relation with the pollution gradient (copper), but a shift in benthic characteristics from 50-150mg/kg Cu (Figure 1). All benthic assessment approaches shows a clear and similar response to the pressure. Same, non-linear patterns in benthic characteristics against a metal pollution gradient were shown in the study of Josefson et al. (2009). All benthic assessments show a very similar correlation value with the pressure (Table 11). The highest correlation (cf Draftmans; Primer software) value is obtained with the IQI (UK/ROI) and the lowest with the BO2A (Spain, Andalusia).

Table 11. Draftmans plot correlation factors between benthic assessment approaches and organic and metal pollution parameters.

	Denmark	UK/ROI	Spain (m-AMBI)	Norway	Portugal	the Netherlands	Germany	France	Spain (BO2A)
N (%)	-0,681	-0,728	-0,692	-0,717	-0,684	-0,686	-0,693	-0,691	-0,580
Cu (mg/kg)	-0,729	-0,787	-0,732	-0,777	-0,728	-0,720	-0,735	-0,729	-0,672
Zn (mg/kg)	-0,704	-0,754	-0,710	-0,743	-0,704	-0,699	-0,712	-0,707	-0,632
Pb (mg/kg)	-0,621	-0,660	-0,636	-0,656	-0,633	-0,630	-0,638	-0,635	-0,572
C (%)	-0,701	-0,768	-0,719	-0,754	-0,708	-0,717	-0,720	-0,718	-0,628
Cr (mg/kg)	-0,692	-0,729	-0,696	-0,723	-0,694	-0,685	-0,699	-0,694	-0,624

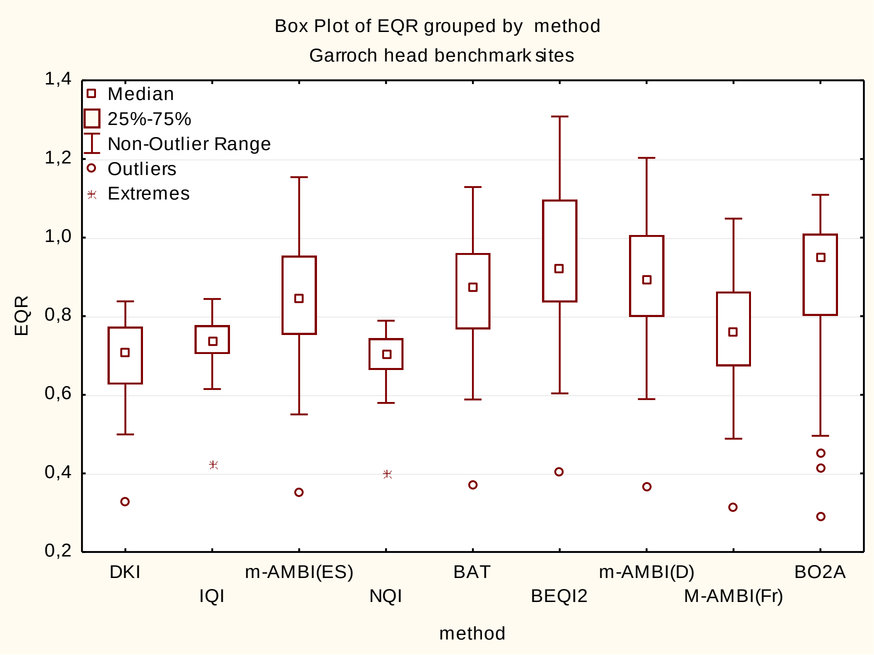


Figure 2. Box-Whisker plot of the EQR values of the benthic assessment approaches for the classification of the Garroch head benchmark sites.

The samples with a copper concentration less than 50mg/kg seem to represent non-disturbed conditions and could be used as benchmark sites (least disturbed samples). The box-whisker plot (Figure 2) gives a distinct visualization of the differences between the EQR values of the different benthic assessment approaches for these benchmark sites. Some approaches were more similar to each other than others. The median EQR values of the benchmark sites were a little bit lower for the DKI, IQI and NQI, which can be related to their higher reference values compared to the other approaches (Table 5). The BO2A shows the highest median EQR values for the benchmark sites. The values of m-AMBI (Fr) are in between. The m-AMBI (ES), BAT, BEQI 2 and m-AMBI(DE) EQR values were more or less similar for these benchmark sites. The differences of the EQR values of the benchmark sites were significantly different between the m-AMBI (ES), BAT, BEQI 2, m-AMBI(DE) and DKI, IQI, NQI and m-AMBI(Fr) (Kruskal-Wallis mean rank test) (Table 12). The IQI was not significantly different with the NQI, DKI and m-AMBI(Fr). The NQI was significantly different with all other approaches, except the IQI and DKI. The DKI is also significantly different with all other approaches, except the IQI and NQI. This to illustrate that there were differences in the benthic assessment approaches in the classification of the samples under similar pressure conditions. This benchmark aspect is further analyzed in point 4.3 below.

Table 12. Kruskal-Wallis p levels (multiple comparisons of mean ranks) by comparison the EQR values of each approach for the Garroch head benchmark sites.

	DKI	IQI	m-AMBI(ES)	NQI	BAT	BEQI2	m-AMBI(DE)	m-AMBI(Fr)	BO2A
DKI		1,000	0,000	1,000	0,000	0,000	0,000	0,009	0,000
IQI	1,000		0,000	1,000	0,000	0,000	0,000	1,000	0,000
m-AMBI(ES)	0,000	0,000		0,000	1,000	0,160	1,000	0,004	1,000
NQI	1,000	1,000	0,000		0,000	0,000	0,000	0,003	0,000
BAT	0,000	0,000	1,000	0,000		0,658	1,000	0,000	1,000
BEQI2	0,000	0,000	0,160	0,000	0,658		1,000	0,000	1,000
m-AMBI(DE)	0,000	0,000	1,000	0,000	1,000	1,000		0,000	1,000
m-AMBI(Fr)	0,009	1,000	0,004	0,003	0,000	0,000	0,000		0,000
BO2A	0,000	0,000	1,000	0,000	1,000	1,000	1,000	0,000	

4.2.2 Higher level comparison

The samples of the Garroch head are grouped in sets of samples from the same location and same time period to allow a BEQI comparison. The reference dataset are the samples which are characterized by a copper content of less than 50 mg/kg. A similar trend of the benthic assessment approaches in relation to copper is found as on sample level (Figure 3). The EQR values decreased with increasing copper value. The BEQI approach shows a similar pattern as the other approaches.

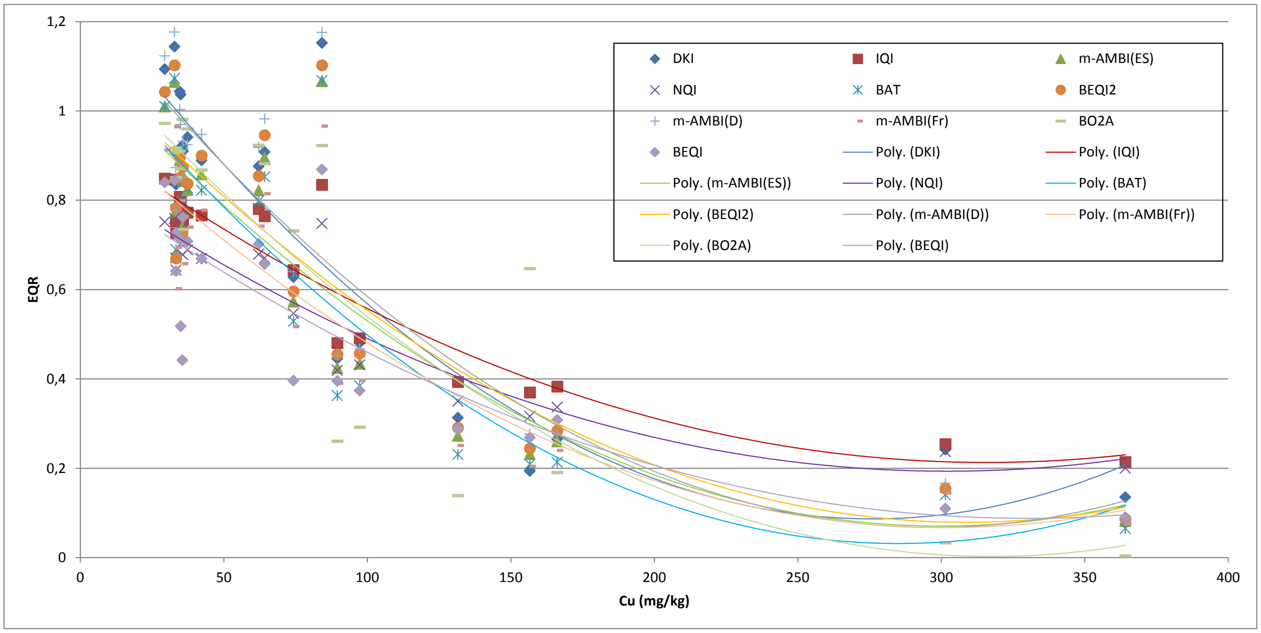


Figure 3. Correlation plot with trend line (polynomial 2nd order) between the different assessment approaches and Cu(mg/kg) for the set of pooled samples.

Table 13. Draftmans plot correlation factors between benthic assessment approaches and copper for the pooled samples.

	DKI	IQI	m-AMBI(ES)	NQI	BAT	BEQI2	m-AMBI(D)	m-AMBI(Fr)	BO2A	BEQI
Cu	-0,810	-0,886	-0,817	-0,875	-0,808	-0,813	-0,823	-0,814	-0,828	-0,805

The correlation between the copper concentration and the EQR values of the benthic assessment approaches are all high and comparable (Table 13). The BEQI shows the lowest correlation; the IQI the highest.

4.3 Assessment concept

Do all national methods follow a similar assessment concept?

The benthic assessment approaches within the NEA-GIG region are very similar, except the BEQI (Belgium). Based on included metrics (parameters) and algorithms, those benthic assessment approaches can be grouped in 4 groups, as outlined in Table 14. The difference in the methodology of calculation of the BEQI (sample aggregation a prior to assessment), compared to the others (at samples level), led to the need for a separate comparability test. This comparability test is executed on aggregated set of samples out of the common dataset.

Table 14. The different types of benthic assessment approaches.

Method	Assessment concept	Remarks
Method group A: m-AMBI, BEQI2	These approaches consist of similar parameters (AMBI, number of species and Shannon wiener), but a different algorithm (factorial analyses [m-AMBI] versus simple algorithm [BEQI2]. The assessment is performed on sample level.	
Method B: IQI, DKI, NQI, BAT	These approaches consist of different parameters (AMBI, number of species, Shannon wiener, Simpson, Margaleff or abundance) and a different algorithm (factorial or simple algorithm). The assessment is performed on sample level.	The simple algorithm differences are based on a different weighing of the parameters or using it as a correction factor (e.g. abundance)
Method C: BEQI	Algorithm including number of species, abundance, (biomass), species composition (Bray-Curtis Similarity) The assessment is performed on habitat level (sample are aggregate prior to assessment).	Difference in community characteristics, use of species composition index instead of a sensitive taxa proportion index.
Method D: BO2A	Based on the abundance of opportunistic polychaetes and amphipods; no diversity parameter.	Not fully WFD compliant

Conclusion

Is the Intercalibration feasible in terms of **assessment concepts?**
No identical approaches for the assessment, because they differs in their parameters or algorithm. The majority of benthic assessment approaches (method type A, B and D) can be intercalibrated on sample level. The BEQI approach (Method type C) needs to be intercalibrated separately on an aggregated set of samples (habitat/ water body level), because this approach does not generate EQR values per sample. Therefore, this method is compared separately with the other assessment approaches on a higher level (see 2.1).

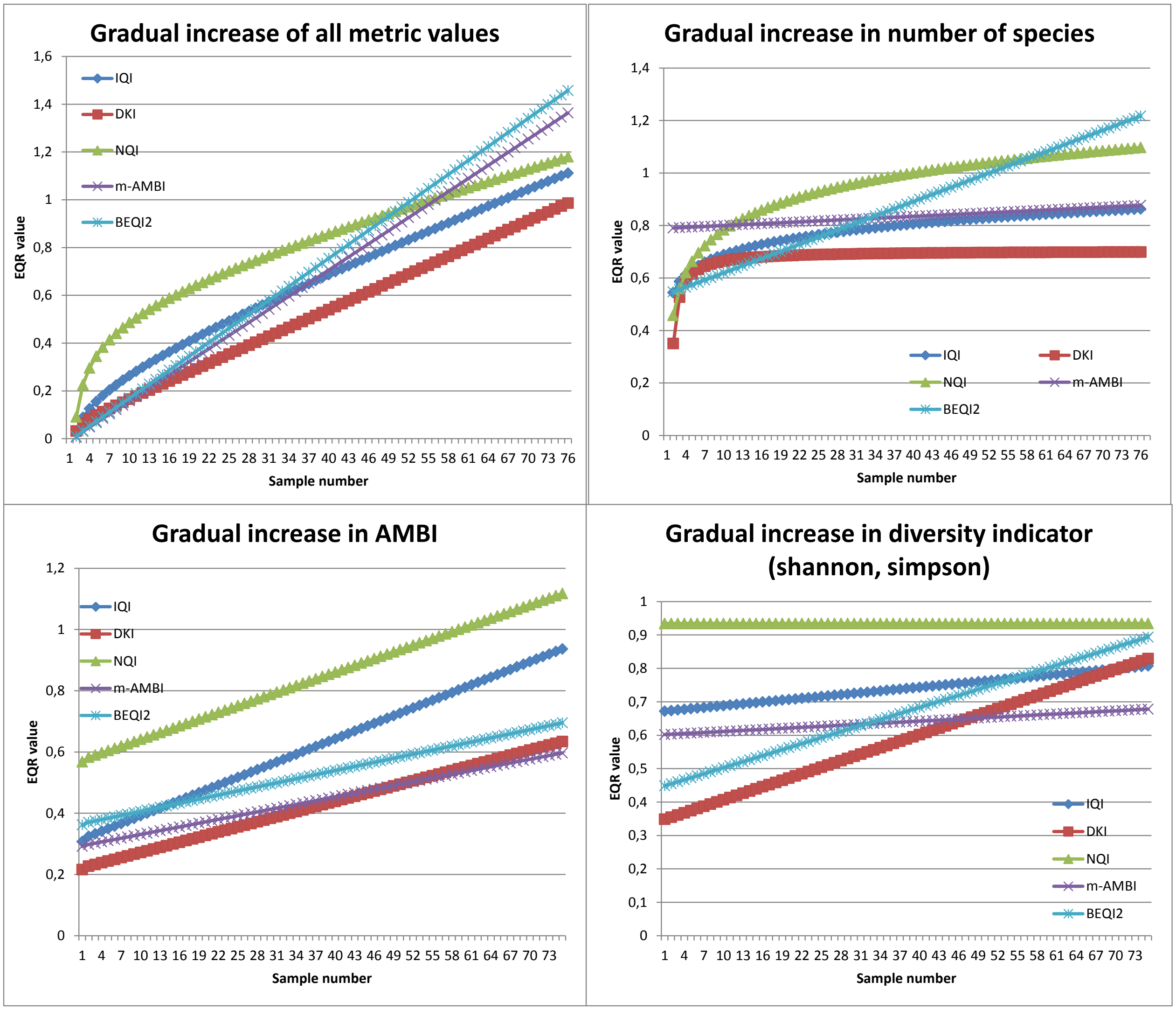
Theoretical behavior of the different benthic assessment approaches

To better understand and illustrate the differences between the different assessment approaches, a test was run to show the dependency of the metrics (parameters) within each algorithm on the overall EQR score and the behavioral response of the different algorithms. This was done by running analyses on a fictive benthic dataset, where some metrics were gradually changed and others were kept fixed. Some of those theoretical samples do not occur in nature, but this exercise was intended to increase the insights into the different algorithms of the benthic assessment approaches. The BO2A is not included, because it has no similar metrics compared to the other approaches.

As visualized in Figure 4, the different concepts show each some particularities, which can be summarized as follows:

- The approaches DKI, m-AMBI and BEQI2 shows a linear trend, when all metric values were slowly increased, whereas the NQI and IQI shows a more parabolic trend (decrease in EQR more strongly when low metric values were obtained). This type of pattern is related to the metric 'number of species' in both approaches.
- The behavior of the IQI is more complex. A decrease in number of species is buffered due to the transformation of the metric within the IQI, because the EQR values tend to decrease very slowly, except when low species numbers were reached. The IQI shows the highest dependency from the AMBI and the lowest for the Simpson.
- The DKI approach shows a linear pattern with increasing parameters, except for number of species (parabolic trend). This can be related to the correction factor (1-1/S) in the algorithm, when the number of species (5-10) are low.
- The EQR values obtained by the m-AMBI approach seem to be most influenced by changes in the metric AMBI and less by the diversity parameters (number of species, Shannon wiener).
- The BEQI2 approach is equally dependent on the metrics, which is related to the equal weight that is given to those metrics within the algorithm.

It is obvious that those differences between the algorithms of the benthic assessment approaches are partly responsible for the variation in the scoring of the samples in the common dataset.


Gradual increase of all metric values

Gradual increase in number of species

Gradual increase in AMBI

Gradual increase in diversity indicator (shannon, simpson)

Figure 4. Changes in EQR values on fictive datasets, to show the metric dependency and behavioral response of the algorithm.

5 Collection of intercalibration dataset and benchmarking

5.1 Dataset description

The benthic dataset of phase I is used for the intercalibration, because there was no time foreseen in this action for collecting new data. Data from Portugal, the Netherlands and France was added in a later stage (Borja et al., 2009) (not included in the publication of Borja et al., 2007). The Danish data set was not from the NEA1/26 type, but the data came from the Kattegat but with rather similar physical characteristics. Nevertheless, the methods used within this type (NEA8a/9/10) were already intercalibrated. Therefore, they supplied new data, which include some NEA1/26 type data. According to the advice of JRC (Fuensanta Salas Herrero), only data of the NEA 1/26 type will be used for the further analyses. A part of the samples of Ireland were excluded (e.g., Clew Bay), due the incomparable sampling size [small]. These were the small modifications done on the common dataset in comparison to phase I. An overview of the metadata information of the final common NEA-GIG, type 1/26 benthic dataset is given in Table 15.

The NEA-GIG intercalibration dataset consists of 656 samples taken from Portugal to Norway. Most of the data originates from time series (samples at certain station sampled in time) or some from spatial monitoring (mainly the Belgian ones). There were 838 different taxa recorded in the entire database, which were constructed based on the 2004 UK taxonomical truncation rules.

5.2 Data acceptance criteria

All NEA-GIG Member States have delivered data for the intercalibration exercise. Nevertheless, the Spanish data is only from the Basque Country, because no data from the regions Andalusia and Cantabria was immediately available.

To explore the common intercalibration dataset for benthic macro-invertebrates, we performed some standard multivariate analyses. This to evaluate the following aspects:

- to check for outliers (samples very different from the rest and showing a problem)
- If there were regional or sub-regional differences between the samples
- If different benthic communities could be detected, which can be related to different physical habitats (sedimentology).
- If there is any pattern in the data that justifies the delineation of sub-types for benchmarking

Table 15. Sample description of data submitted by Member States, from the NEA-GIG for the intercalibration exercise. VV=van Veen grab; HC=Haps core; DG= Day grab; BC=Box core; SMI=Smith-McIntyre

Country		Location Code		Sample method	Sample size	Number of stations	Period	Replicates per station	Samples submitted	Depth (m)	Sediment
B	Belgium	BGP	Station P2	VV	0,1026	1	1995	1	1	6,7	Sand (97%)
B	Belgium	BHA	Stations Habitat1999	VV	0,1026	37	1999	1	37	5-15	
B	Belgium	BHA	Stations Habitat2000	VV	0,1026	12	2000	1	12	5-15	Sand (85%)-Mud(15%)
B	Belgium	BMA	Stations Marebass	VV	0,1026	1	2000	1	1	13,8	Sand(30%)-Mud(70%)
B	Belgium	BMO	Stations M&OD	VV	0,125	6	1996	1	6	14,2	Sand(>99%)
B	Belgium	BOP	Station O&P	VV	0,125	17	1994,1997	1	17	3,3	Sand(>97%)
B	Belgium	BSU	Subtidale stations	VV	0,1026	58	2002	1	58	5-10	Sand(>93%)
DK(NS)	Denmark	Jammerb	Jammerbugten	HC	0,1*	3	1995	3	3	4-10	Fine sand
DK(NS)	Denmark	Skagerra	Skagerrak	HC	0,1*	3	2004	3	15	8-20	Fine sand
D	Germany	VOR	NS2 Vortrapptief	VV	0,1	1	1987-2004	3-5	64	13	Sand (94%)
NL	the	Ems-	Ems-Wadden coast	BC	0,078	6	2000-2003	1	24	<20	Muddy sand
NL	the	Holland	Holland coast	BC	0,078	5	2000-2003	1	20	<20	Muddy sand
NL	the	Voordelt	Voordelta	BC	0,078	4	2000-2003	1	16	<20	Muddy sand
PT	Portugal	E	Ericeira	SMI	0,1	9	2001	1	9	10-30	Very fine sand
PT	Portugal	FF	Figueira da Foz	SMI	0,1	3	2002	1	3	10-30	Very fine sand
Fr	France	MORWI	Bay of Vilaine	SMI	0,1	5	1992	3	15	<30	muddy fine sand
Fr	France	QUIW	Bay of Quiberon	SMI	0,1	8	2004	3-5	34	<30	muddy fine sand
UK	UK-England	HAR	Harwich	DG	0,1	3	2004	5	15	6,4	Mud(85,3%)
UK	UK-England	LIV	Liverpool Bay	DG	0,1	3	2004	5	15	5,7	Sand(70%)-Mud(30%)
UK	UK-Wales	MIL	Milford Haven	DG	0,1	3	2004	5	15	4,6	Mud(78,8%)
UK	UK-England	TRB	Torbay	DG	0,1	3	2004	5	15	13,7	Muddy sand
UK	UK-Scotland	KIL	Kibrannan Sound	DG	0,1	1	2004	10	10	50	soft muds
UK	UK-Scotland	GRK	Garroch Head	VV	0,1	10	1979-1998	1	180	69-180	Silt/Clay
E	Spain	SSO	San Sebastian-Pasaia	BC	0,186	9	2000-2004	3 (combined)	45	33-61	Sand(90%)-Mud(10%)
N	Norway	STA	Stavanger(S5A)	VV	0,1	1	1995	4	4	93	Mud(83%)
N	Norway	TRO	Trondheimsfjord (RAH1)	VV	0,1	1	2001	4	4	50	Mud(88%)
N	Norway	UTN	Utnes (U10)	VV	0,1	1	2001	4	4	38	Sand(89°%)
ROI	R. of Ireland	GRE	Greatmans Bay	DG	0,1	1	2003	2	2	40,1	Muddy sand
ROI	R. of Ireland	KEN	Kenmmare River	DG	0,1	3	2003	4	12	45,9	Muddy sand

5.2.1 General multivariate analyses

For the purpose of the multivariate analyses, the common dataset is fourth root transformed to reduce the effect of very abundant species on the overall pattern. Beside this, the rare species (in less than 1% of the samples and with a maximum of 3 individuals) were excluded from these analyses to reduce the effect of rare species on the overall pattern. This lead to a reduced dataset with 576 taxa. The similarity between samples is determined by the Bray-Curtis similarity. The sample groups were determined based on a cluster analyses, with cut-off level at certain similarity level. Multidimensional scaling (MDS) is used to visualize the cluster groups. The analyses were executed in PRIMER6.

The first analyses revealed no obvious rarities, but only some outlier samples. Those samples were excluded for all further analyses.

- The samples of station 3 in the Voordelta (the Netherlands) show an inconsistent pattern (two of them show the lowest similarity in comparison with the rest (outliers); the other two were classified in different cluster groups, depending on the analyses. This rare pattern indicates a problem at this location.
- Station Marebass from Belgium was also directly classified separately from the rest. Also the HA99-93 sample from the Belgian dataset classified different from the related samples and can be considered as outlier.

The general multivariate analyses show the following patterns (Figure 5; Table 16):

- All data clearly grouped per Member State and even data region (North Sea, , when the cluster analyses were sliced at a similarity level of 11. Even if when slicing it further at similarity 15, the grouped data were further split per Member State .
- The North Sea area forms one cluster of samples (cluster h in Table 16), with the samples of Belgium, the Netherlands, Germany and Denmark. The Liverpool Bay samples shows a high similarity with those North Sea samples. Another large cluster group contains a part of the UK data (Garroch Head), the Spanish and Norwegian data. The other Member States (France, Portugal, Ireland) datasets form separate clusters (Table 16; Figure 5).
- The data of most Member States clustered more or less together in the MDS plot, except the Portuguese data (cluster G), which were more scattered.
- A few samples of the Garroch Head dataset (cluster C) were also split from the others and were very similar. This because those samples contain very high densities of only one species (*Mediomastus fragilis)*.

Table 16. Number of samples of each Member State in each cluster group (slice at similarity level 11).

slice11	B	D	DK(NS)	E	Fr	N	NL	PT	ROI	UK	MS/regio
a					34						Fr (QUIW)
b										55	UK(Har, Kill, MILl, TRB)
c										6	UK(GRK)
d				45		12				174	Spain, UK(GRK), N
e									14		ROI
f					15						Fr (MORWI)
g								11			PT
h	130	64	18				56	1		15	UK(liv), NL, DK(NS), D, B

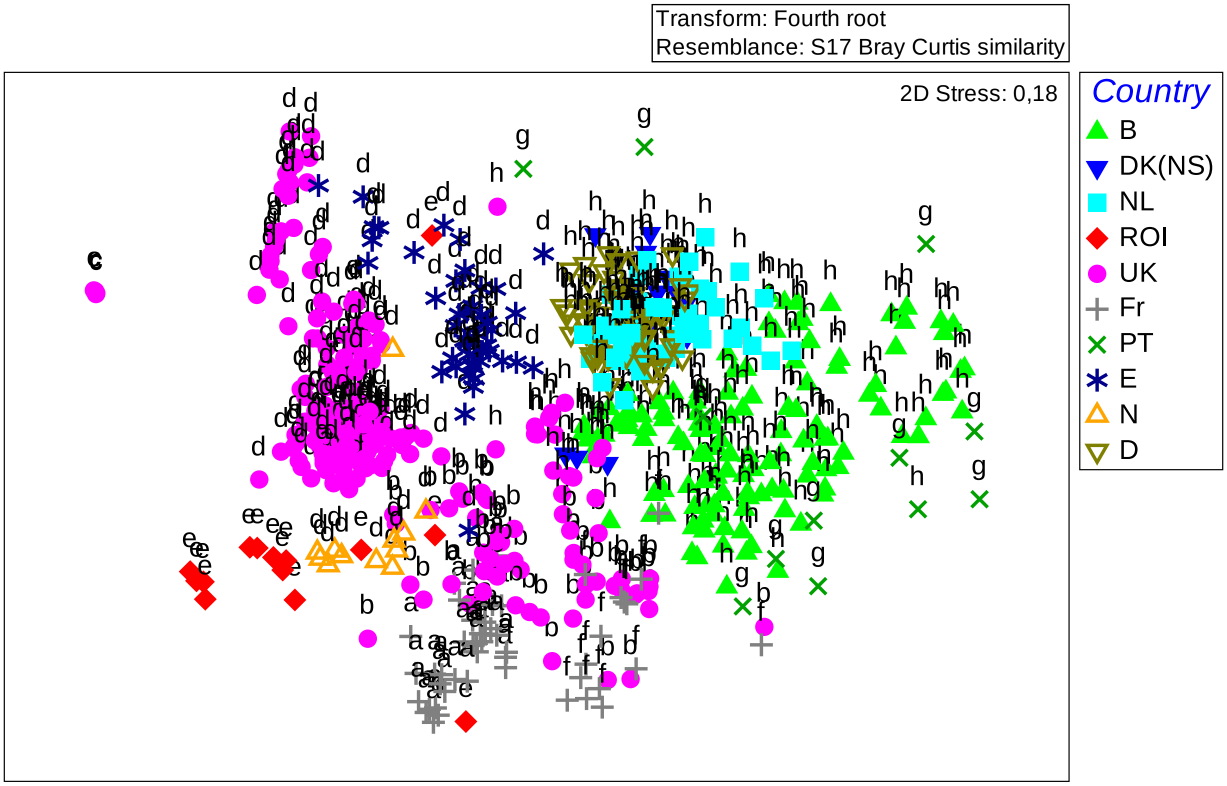


Figure 5. MDS plot of the intercalibration data with indication of the Member States (colored symbols) and the cluster groups (slice at similarity of 11).

It can be concluded that the different datasets show a low similarity with each other, because they are clearly split as separate identities at low similarity level. There is no clear grouping of the data in relation to a South-North gradient within the NEA-GIG region. The data seemed to be grouped in a group with the North Sea related datasets and Portugal; a group with the datasets from shallow areas in the UK and France and a group with samples from less shallow areas (>30m depth) of UK, Spain, Norway and Ireland (Table 15). As the analyses show, every region has its own benthic species composition, with commonalities over the NEA-GIG region. The main difference in species composition between the NEA-GIG samples seems in first instance to be related to depth, which can be used as a factor to delineate sub-regions in the intercalibration. The delineation of sub-regions based on bio-geographical reasons (North-South) seems not to be appropriate.

5.2.2 Multivariate analyses of the benthic univariate parameters

Species composition on its own is not a parameter that is included in the benthic assessment algorithms. The algorithms are constructed from diversity and species tolerance/sensitivity classification metrics. In these analysis, it is investigated if those parameters are different among the Member States' datasets.

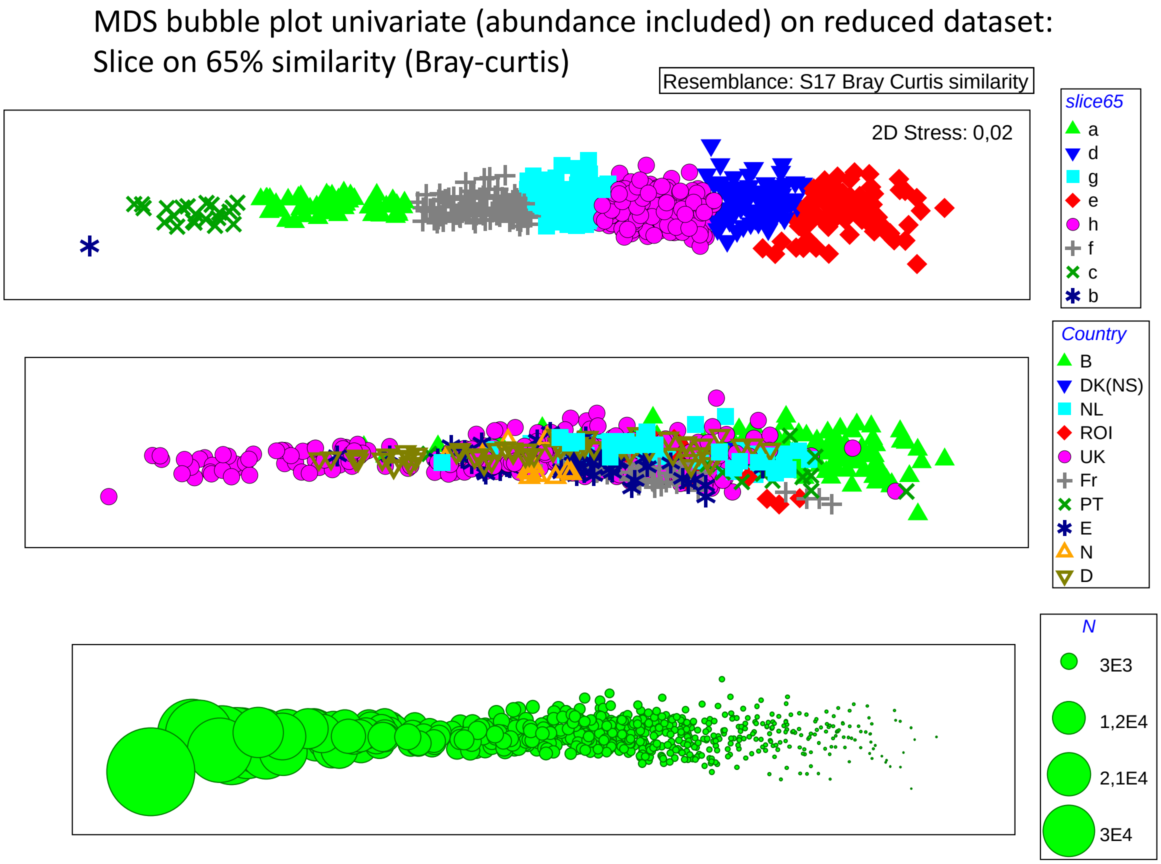

MDS bubble plot univariate (abundance included) on reduced dataset:
Slice on 65% similarity (Bray-curtis)

Figure 6. MDS plot of the univariate variables (inclusive abundance), with indication of the cluster groups at slice 65 (upper figure) and the behavior of the dataset of the different Member States (center figure). The figure below shows the pattern of abundance in the dataset.

The aim of this analysis, is to confirm if it is necessary to define sub-regions for the intercalibration by testing if there are differences between the samples in their univariate parameters/metrics (e.g., Shannon diversity [logbase2], Margalef, Simpson, number of species, SN [ln(S)/ln(ln(N)))], abundance, AMBI). These are all the parameters by which the benthic assessment approaches are constructed.

The multivariate pattern is firstly strongly influenced by the parameter abundance. There is no obvious difference between countries and the samples are spread along the univariate gradients. It seems that many of the samples of the Belgian dataset are characterized by low abundances as compared to the other datasets. When abundance is excluded from the analyses, because it is in most approaches only relatively taken into account, the multivariate pattern is different. The gradient is dominated by number of species, and the deviation (at lower number of species) at one end is related to the difference in AMBI (very high values in the upwards gradient) (Figure 7). These analyses in the univariate parameters shows that there is a gradient within the dataset based on the univariate parameters from samples with a higher diversity to samples characterized by low diversity (Table 17). The data of the Member States seems to be spread over this gradient. This pattern in univariate parameters seems to correspond with a possible pressure gradient on the benthic data, which cannot be quantified (due to the lack of pressure data). The upper gradient shows the gradient in benthic characteristics, related to the disposal pressure (Garroch head, Spain), whereas the lower diversity gradient can be related to physical pressures (natural, anthropogenic).

MDS bubble plot univariate on reduced dataset: Slice on 75% similarity (Bray-curtis)

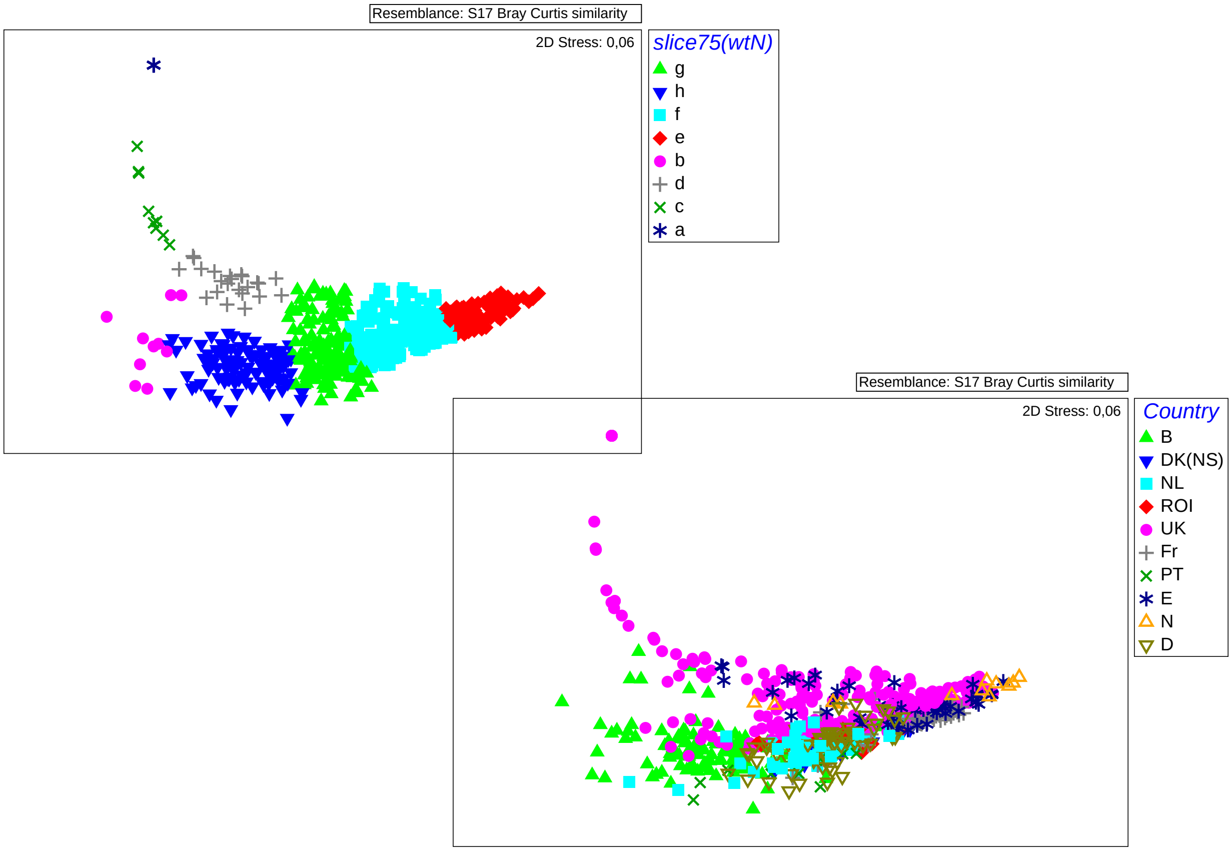


Figure 7. MDS plot of the univariate parameters (exclusive abundance) and indication of the cluster groups (slice 75) (upper figure) and the behavior of the datasets of the different Member States (lower figure).

Table 17. Average values of the benthic parameters for each cluster group and their standard deviation.

Group	S		d		H'(log2)		1-Lambda'		AMBI		SN	
a	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	3,000	0,000	0,000	0,000
b	3,100	0,568	0,783	0,161	1,183	0,256	0,860	0,041	1,476	1,066	1,201	0,245
c	3,100	0,876	0,270	0,097	0,532	0,420	0,609	0,112	6,000	0,000	0,537	0,131
d	8,043	1,894	1,027	0,319	0,895	0,493	0,858	0,051	5,441	0,753	1,076	0,167
e	**48,933**	8,160	**8,211**	1,355	**4,168**	0,658	**0,992**	0,002	2,238	0,556	**2,203**	0,126
f	**27,567**	5,816	**4,817**	1,263	**3,150**	0,948	**0,982**	0,010	2,214	1,198	**1,939**	0,227
g	15,513	2,466	2,810	0,673	2,222	0,904	0,963	0,021	2,221	1,793	1,693	0,272
h	7,423	2,168	1,842	0,426	2,032	0,578	0,951	0,023	1,313	0,574	1,667	0,365

5.2.3 Overall conclusions:

All data are suited for the analysis, except the few outline samples discriminated. Based on the multivariate analyses on the species-abundance data, we could discriminate the datasets from the different Member States, where the North Sea datasets show most similarity. The samples taken in less shallow regions (>30m) seem to be different regarding species composition compared to the samples taken in the more shallow regions. When this pattern is analyzed based on the metrics of the benthic assessment approaches, all datasets of the Member States are clustered together, but along a gradient. Therefore, no sub-regions based on biogeographical reasons can be discriminated. Only the factor depth seems to delimit two different type of habitats within the common dataset and can be considered as a relevant factor to distinguish between both dataset parts in the intercalibration. The review panel and JRC advise to distinguish this as two sub-types within the common dataset for the comparability analysis.

5.3 Common benchmark

An alternative procedure for the selection of benchmark sites need to be used in this intercalibration, because we cannot fulfill the guidance principle using this common dataset: “The benchmarking process must use harmonized criteria independent of national classifications (i.e., countries cannot simply nominate the sites they classify as high status as being their benchmark sites without further checking).” The following approaches could be used for benchmarking, but does not make it within the NEA-GIG NEA1/26 intercalibration exercise:

- The absence of qualitative or quantitative pressure data (and it was not the task to collect this, which is an impossible exercise),
- no reference sites for each Member State /region (this approach was tried by Angel with sites from Spain and Norway),
- indirect pressure quantification not appropriate (e.g., LUSI index), due to the selection of data away from point sources (rivers, harbors, etc.) and the majority of the data is time series data from one location.

- An approach that estimates the benthic conditions under least disturbed circumstances could be the selection of samples with the highest diversity characteristics (response variables), which show a theoretical relation with changes in the abiotic environment due to pressures (see Annex 2). This procedure to determine the benchmark samples out of the common dataset is not accepted by JRC. The main reasons argued are, as stated in the IC Guidance, selection of benchmark sites should be done by screening for sites meeting abiotic criteria that represent a similar low level of impairment. The option proposed by the BQE lead for selecting benchmark sites is not acceptable because is based on the diversity, a biotic parameter included in most of the methods to be intercalibrated, and therefore the method values are influenced by this parameter. Moreover, in basis on the Pearson & Rosenberg model, diversity is a critical parameter, as it does not show a monotonic trend along both spatial and temporal gradients of pollution (Subida et al, 2013). When moving away from the source of pollution, the peak of opportunists is often followed by a maximum value in diversity, which then stabilizes at a slightly lower level. This means that, in a gradient of pollution, the highest values for the diversity index may be recorded when the number of species is still low and the community is still in an early stage of recovery (Pearson & Rosenberg, 1978). So, a diversity parameter, in some situations, could indicate high values in moderately disturbed areas.

A review panel argued that from a scientific perspective, the approach is not convincing and that the group should collect pressure data to do the benchmark standardization properly. JRC remained to the review panel the necessity to provide solutions in basis on the available data set. In this sense, JRC proposed to select benchmark sites in basis on the expert judgment.

Based on the knowledge of the coastal areas and the stations included in the dataset, they could indicate the stations that were under minor pressures (or with more distance from the focus of main disturbances). Therefore, the Member States indicate on basis of their opinion (and not based on the methods results), the stations with minor pressures. For Spain and Norway, the benchmark sites selected during phase II were used: In the case of Norway because they have reference sites, and in the case of Spain because they already selected in the previous phase less disturbed sites.

The review panel accepted this proposal.

5.4 Benchmark standardization

The principal aim of benchmarking in intercalibration is to identify and remove differences among national assessment methods that are not caused by anthropogenic pressure but rather by systematic discrepancies (due to different methodology, biogeography, typology etc.) (Annex V, IC Guidance).

Benchmark standardization will correct for differences in median EQR values between the Member States' benchmark sites obtained by certain assessment approaches. Those median values will be corrected by the benchmark standardization procedure; this correction will be more obvious for cases where the medians are significantly different.

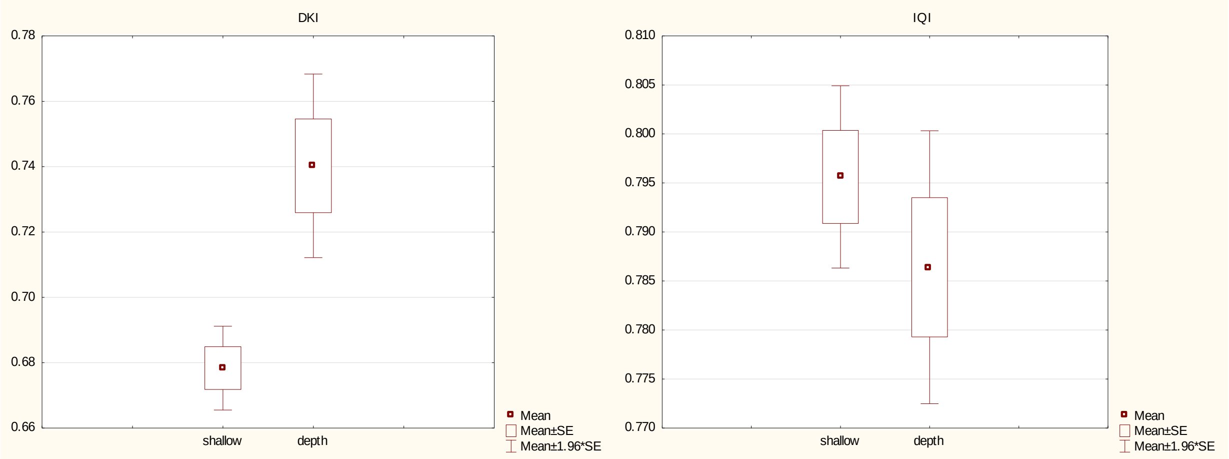
We tested whether benchmark standardization was necessary. Student's sT was used to compare the benchmark sites values for the two subtypes (shallow/depth) and the national methods.

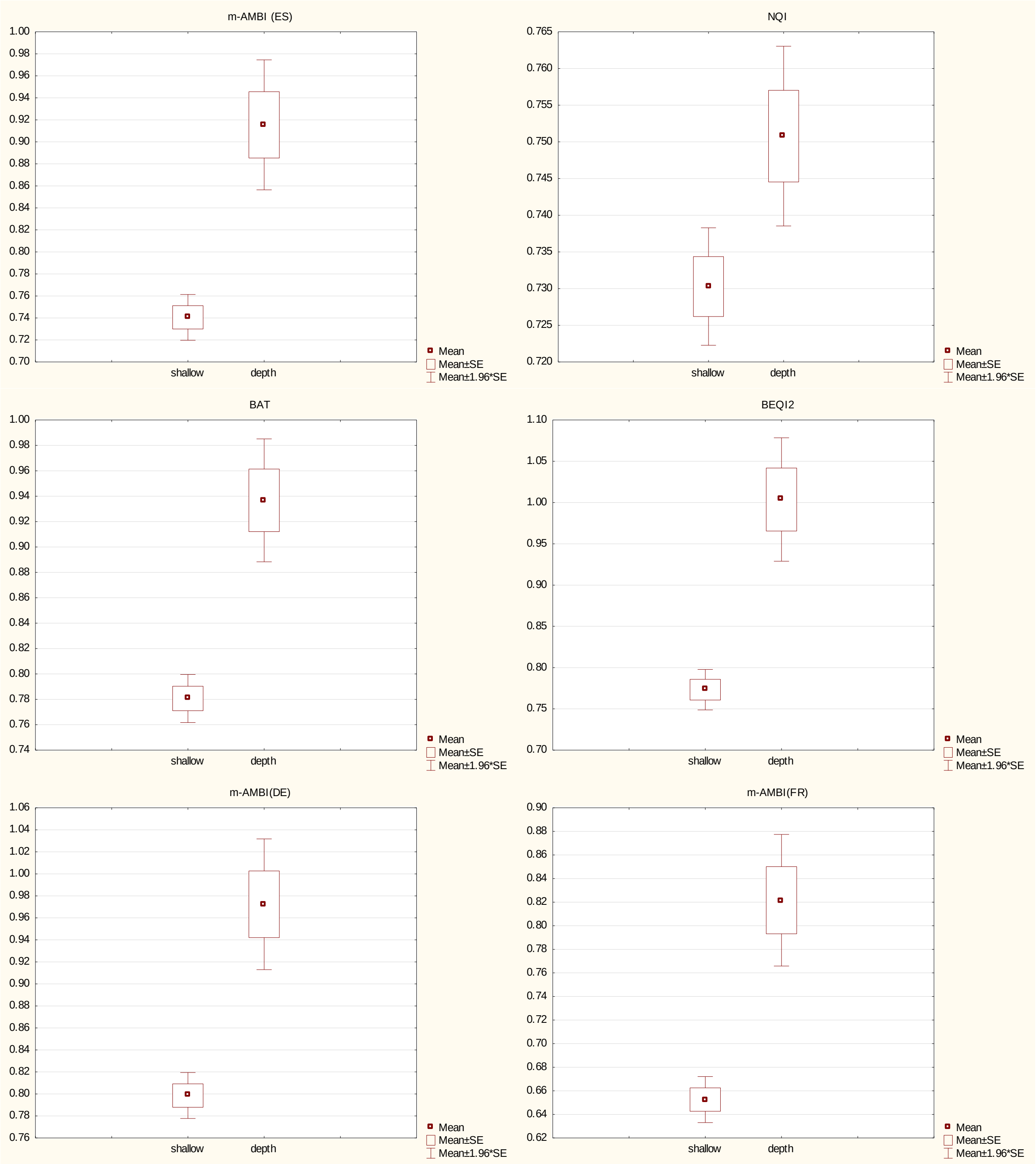
There were statistical difference (P<0.05) between both subtypes for all the methods, except for the IQI (UK/ROI method) (Table 18; Figure 8). Because of this, benchmark standardization was applied using the Excel sheet for option 3. The correlation between the average value of all national EQRs per survey in the full dataset was significantly correlated (P<0.01) with its standard deviation, thus national EQRs converge towards the bad end of the quality gradient, and therefore, division was used for the standardization.

Benchmark samples where more than three national methods show EQR values less than good status (in accordance to the national boundaries) were excluded. This criteria was used in the previous phase by several MED GIG BQE groups (Fuensanta personal communication). This were 8 samples of the Belgian dataset (station HA99-117; HA99-77, HA00-1; HAA00-11; HA00-21; HA00-3; HA00-4; HA00-5) and 3 samples of the German dataset (VORWI0700B [replica E]; VORWI0897B; VORWI0897B).

Table 18. Student's sT – P values

Method/Member State	P values
DKI (DK)	0.000046
IQI (UK/ROI)	0.28
m-AMBI(ES)	9.409E-07
NQI (NO)	0.0072
BAT (PT)	1.717E-07
BEQI2 (NL)	3.696E-07
m-AMBI(DE)	1.230E-06
m-AMBI(FR)	5.806E-07
BO2A (ES)	7.884E-09




m-AMBI (ES)
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
shallow
depth
Mean
Mean±SE
Mean±1.96*SE
NQI
0.765
0.760
0.755
0.750
0.745
0.740
0.735
0.730
0.725
0.720
shallow
depth
Mean
Mean±SE
Mean±1.96*SE
BAT
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
shallow
depth
Mean
Mean±SE
Mean±1.96*SE
BEQI2
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
shallow
depth
Mean
Mean±SE
Mean±1.96*SE
m-AMBI(DE)
1.06
1.04
1.02
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
shallow
depth
Mean
Mean±SE
Mean±1.96*SE
m-AMBI(FR)
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0.62
shallow
depth
Mean
Mean±SE
Mean±1.96*SE

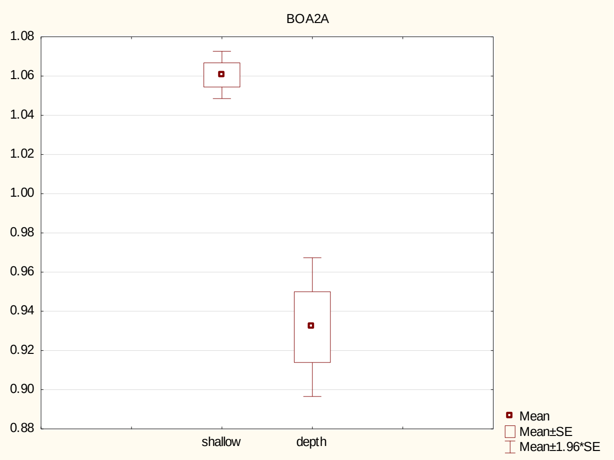


Figure 8. Box-whisker plot median, percentile values and no-outlier range of the EQR values at the Member States' benchmark sites with the national methods for the two subtypes (shallowness and depth).

6 Comparison of methods and boundaries

6.1 Intercalibration option and common metrics

Option 3a. Intercalibration can be performed based on commonly assessed sites and whether the ecological quality gradient is sufficiently covered. More than three methods are used for this exercise. Following the advice of JRC and the review panel following intercalibration aspects need to be taken into account:

- The benchmark sites selected by the experts and following the review panel recommendations (see 5.3)
- As benchmark standardization procedure, the division options is the appropriate one (see 5.4).
- Two sub-types, based on depth, need to be distinguished (see 4.2.3) .
- Due to the fact that the BO2A method does not meet the criteria in the previous comparisons (see 6.3), this method can be excluded in the final calculations.

The intercalibration excel sheet IC_Opt3_Div_v1.24.xlsx is used for executing the comparisons.

Because the BEQI assessment approach does not allow the calculation of EQR values on samples level (see 3.1 methods and 4.3 assessment concepts), a separate intercalibration on higher level (set of grouped samples) is executed. This separate intercalibration to analyze if the BEQI assessment approach meets the intercalibration criteria compared to the other assessment approaches. This separate comparability check on higher level implies that there no boundary adjustment could be suggested for the other assessment approaches based on those outcomes.

An intercalibration on sample level and higher level (to include the BEQI approach) was executed, with the benchmark samples selected based on expert judgment.

6.2 History

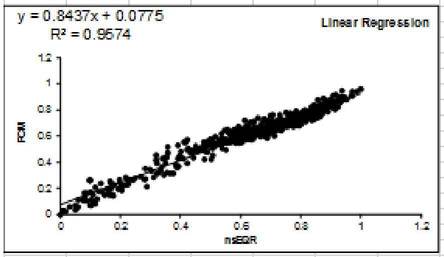
A set of comparisons between the benthic assessment approaches are executed during this third intercalibration phase. To keep record of it and to allow for checking which options were tested, this information is included in annex 3 of this report. This were all intermediate comparability analyses to explore the intercalibration and to guide towards the selection of the comparison most in line with the intercalibration guidelines and acceptable for JRC and the review panel.

Different outcomes were obtained, based on the different options of benchmarking (biotic or expert judgment), standardization (subtraction or division), inclusion of methods (with or without BO2A), sub-regions (yes or no) and level of comparison (sample or higher). The use of these different options in the comparison lead to difference in the comparability criteria results and the need for boundary adjustments (or not). But the options selected for the final comparability analyses (section 6.3 and 6.4), seems to be the most appropriate regarding the intercalibration guidelines.

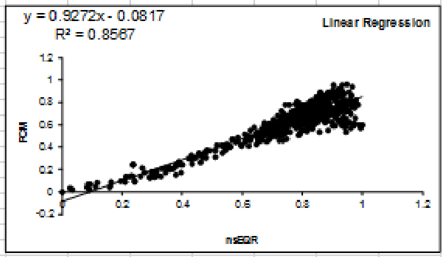
6.3 Results of the regression comparison

6.3.1 Sample level comparison

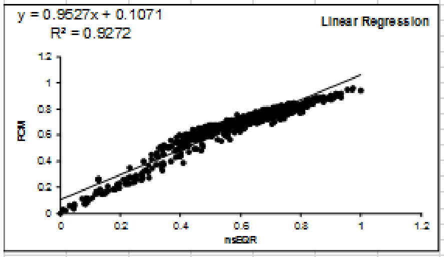
Denmark



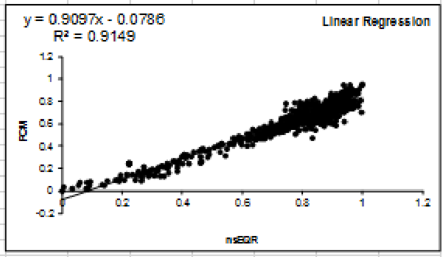
UK/ROI



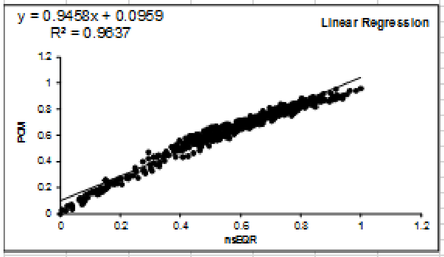
Spain (BQ, CQ)



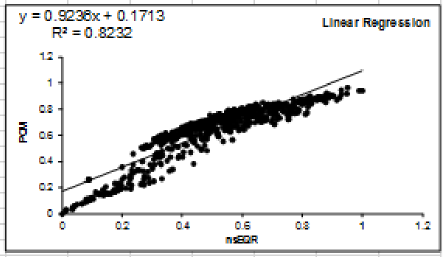
Norway



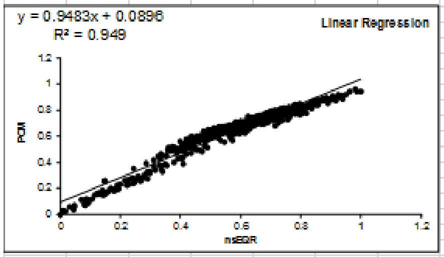
Portugal



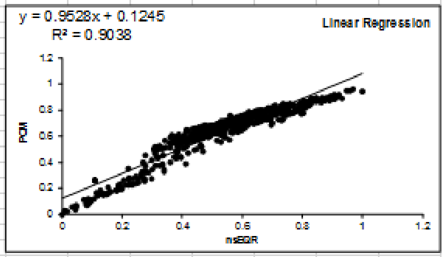
the Netherlands



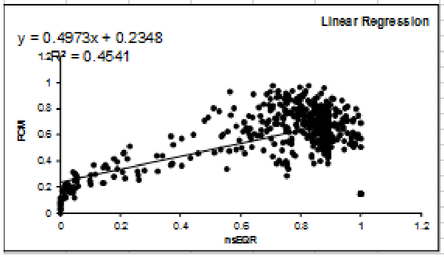
Germany



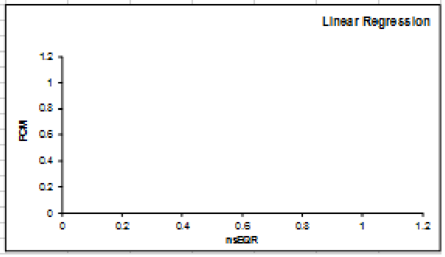
France



Spain (AC)

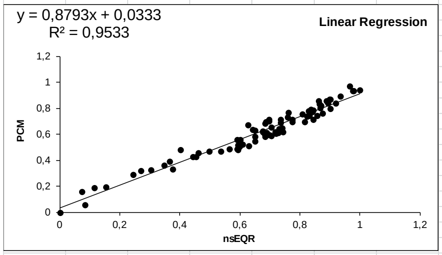


0

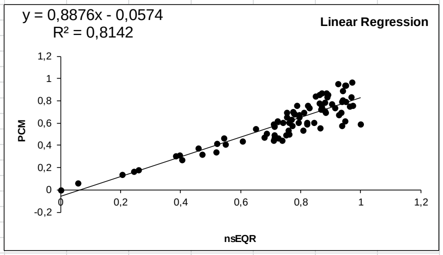


6.3.2 Higher level comparison (+ BEQI, Belgium)

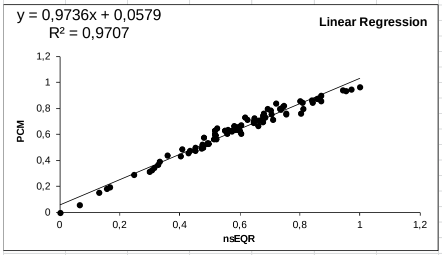
Denmark



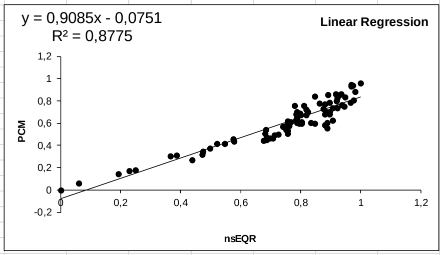
UK/ROI



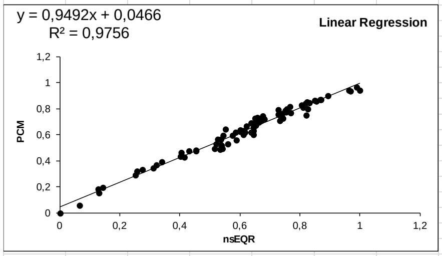
Spain (BC, C)



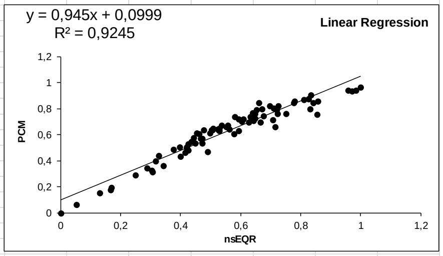
Norway



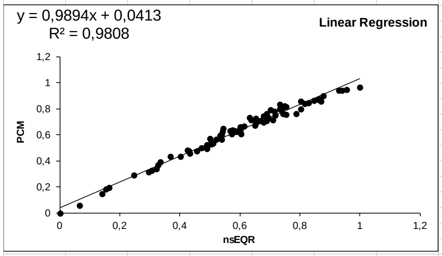
Portugal



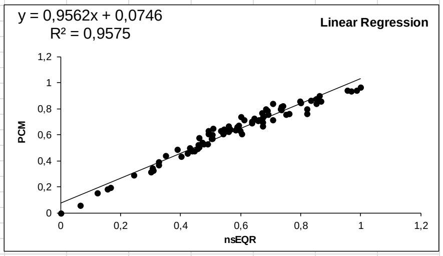
the Netherlands



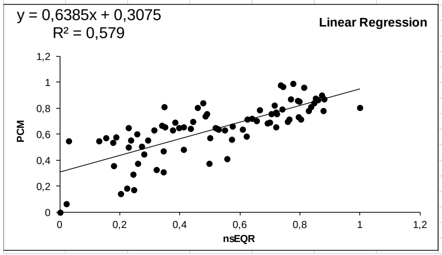
Germany



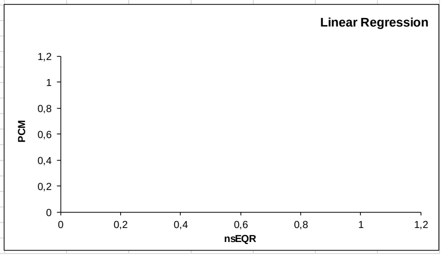
France



Belgium



0



6.3.3 Summary

The correlation between the metrics is determined in the intercalibration excel sheet. For all the intercalibration comparisons, the benthic assessment approaches fulfill the criteria (R^2 > ½ maxR^2) of the regression comparison, except for BO2A (Table 19). The BO2A of Spain shows the lowest correlation with the pseudo-common metric. For the IQI and the NQI, the samples were less equally spread over the linear regression line (dominance in upper part) in comparison to the other approaches, as was the case in the analyses on the theoretical behavior of the benthic assessment approaches.

Table 19. Summary of the correlation coefficient (R^2) of each approach with the common metric for the different intercalibration comparisons. Values outside the criteria were put in red.

Method	Sample level comparison	Higher level comparison
	Sub-region	Sub-region
Denmark	0.957	0.9533
UK/ROI	0.854	0.8142
Spain (m-AMBI)	0.927	0.9707
Norway	0.914	0.8875
Portugal	0.963	0.9756
The Netherlands	0.823	0.9245
Germany	0.949	0.9808
France	0.903	0.9575
Spain (BO2A)	0.452	/
Belgium	/	0.579

The Spanish method (BO2A) had to be excluded from the comparability analysis due to its low correlation with the PCM (r^2=0.452).

6.4 Comparability criteria

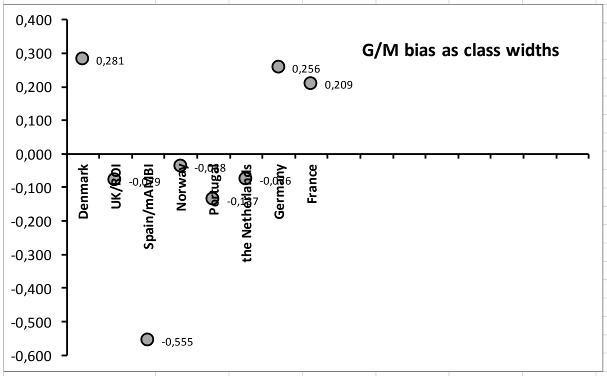
6.4.1 Sample level comparison

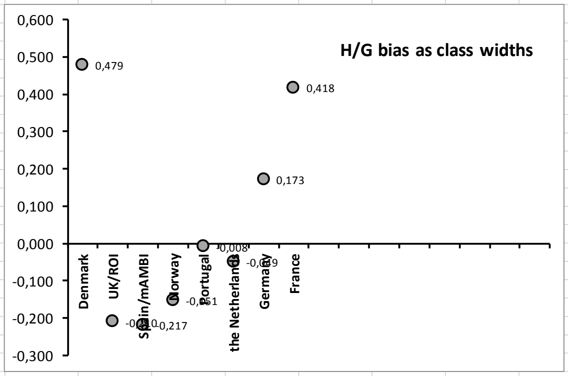
Table 20. Summary of the boundary bias and class differences analyses following the division benchmark standardization, with discrimination of the sub-regions.

	Denmark	UK/ROI	Spain/mAMB	Norway	Portugal	e Netherland	Germany	France
Max	1,000	1,000	1,292	1,000	1,130	1,270	1,189	1,027
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200

H/G bias_CW	0,479	-0,210	-0,217	-0,151	-0,008	-0,049	0,173	0,418
G/M bias_CW	0,281	-0,079	-0,555	-0,038	-0,137	-0,076	0,256	0,209

	Denmark	UK/ROI	Spain/mAMB	Norway	Portugal	e Netherland	Germany	France
Count	4445	4445	4445	4445	4445	4445	4445	4445
Absolute Class Difference	0,4189	0,3735	0,2650	0,3582	0,2731	0,3028	0,3024	0,3042





For certain national methods do not comply with the comparability criteria. Boundary bias is exceeded by the methods of
- Denmark – Boundaries HG and GM too stringent
- Germany - Boundaries GM too stringent
- France - Boundaries too stringent
- Spain (m-AMBI)- Boundaries GM too relaxed

The average absolute class difference after boundary harmonization meets the comparability criteria for all national methods.

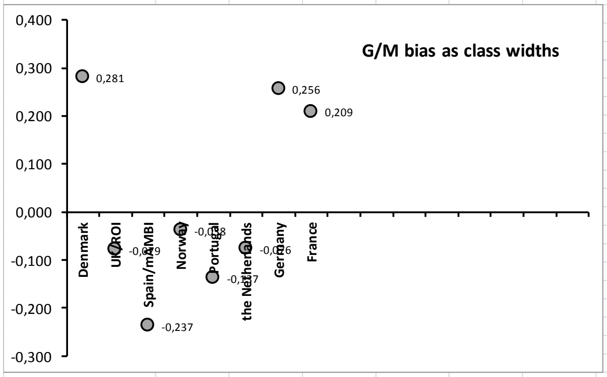
Spain is requested to adjust the boundaries to allow for completing the intercalibration exercise by raising its Good/moderate boundary to a value of 0.63.

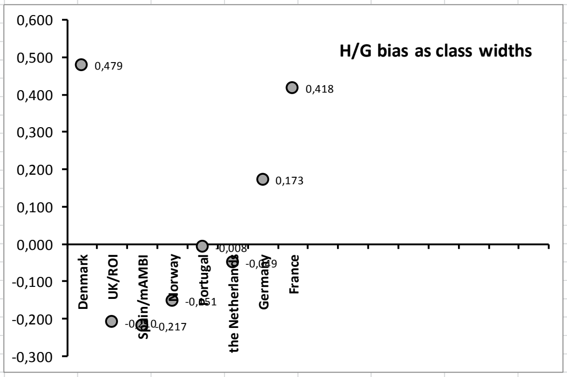
Germany, Denmark and France are not obliged to lower the boundaries that have been identified as being too stringent. The intercalibration criteria values after boundary harmonization are given in Table 21.

Table 21. Summary of the boundary bias and class differences analyses following the division benchmark standardization, with discrimination of the sub-regions, after harmonization of the boundaries.

	Denmark	UK/ROI	Spain/mAMB	Norway	Portugal	e Netherland	Germany	France
Max	1,000	1,000	1,292	1,000	1,130	1,270	1,189	1,027
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770
G/M	0,600	0,640	0,630	0,630	0,580	0,580	0,700	0,530
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200
H/G bias_CW	0,479	-0,210	-0,217	-0,151	-0,008	-0,049	0,173	0,418
G/M bias_CW	0,281	-0,079	-0,237	-0,038	-0,137	-0,076	0,256	0,209

	Denmark	UK/ROI	Spain/mAMB	Norway	Portugal	e Netherland	Germany	France
Count	4445	4445	4445	4445	4445	4445	4445	4445
Absolute Class Difference	0,4072	0,3874	0,2713	0,3748	0,2947	0,3087	0,2772	0,2893





6.4.2 Higher level comparison (BEQI, Belgium)

This higher level comparison is to test the comparability of the BEQI method with the other assessment approaches. Not meeting certain comparability criteria by these other assessment approaches has no consequence for the boundary harmonization (at sample level). The BEQI EQR values are determined on a set of stations (instead of one station) (see section 2.1 for more detail).

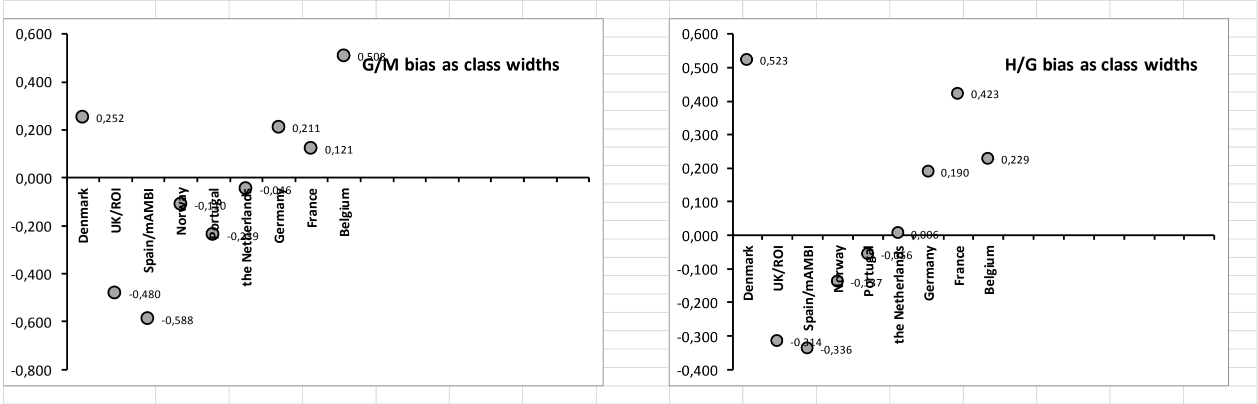
The boundary bias (<0.25) in this analysis is too high for the good/moderate and high/good boundary for the m-AMBI (BC, C) and IQI (Table 22). The DKI and BEQI (Belgium) are more stringent for the good/moderate boundary. The French and Danish approach is also more stringent for the high/good boundary. The class difference (<0.5 class) is below the criteria level for all benthic assessment approaches, except for the IQI where it is at criteria level. The BEQI assessment approach meet the comparability criteria in comparison with the other approaches. Further boundary adjustment cannot be suggested, as this is a comparability check on higher level than sample level; in most assessment approaches, their boundaries were based on a sample level evaluation. Besides this, the BEQI is comparable with all methods applied in sub-region A (very shallow) type - all Belgian coastal waters belong to sub-region A.

Table 22. Summary of the boundary bias and class differences analyses following the division benchmark standardization, with discrimination of the sub-regions.

	Denmark	UK/ROI	Spain/mAMBI	Norway	Portugal	e Netherland	Germany	France	Belgium
Max	1,000	1,000	1,229	1,000	1,016	1,049	1,040	1,000	1,000
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770	0,800
G/M	0,600	0,640	0,630	0,630	0,580	0,580	0,700	0,530	0,600
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380	0,400
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200	0,200

H/G bias_CW	0,523	-0,314	-0,336	-0,137	-0,056	0,006	0,190	0,423	0,229
G/M bias_CW	0,252	-0,480	-0,588	-0,110	-0,239	-0,046	0,211	0,121	0,508

	Denmark	UK/ROI	Spain/mAMBI	Norway	Portugal	e Netherland	Germany	France	Belgium
Count	648	648	648	648	648	648	648	648	648
Absolute Class Difference	0,4136	0,5278	0,3194	0,4228	0,3843	0,3611	0,3256	0,3210	0,4799



7 Final results to be included in the EC

7.1 Table with EQRs

After the boundary harmonization, the final boundaries for the benthic assessment approaches for coastal waters in the Northeast Atlantic were given in Table 23. The comparability of the BEQI boundary values are tested in a separate comparability analyses (see above) and were comparable with the other methods. For the moment, only the BO2A approach does not meet the comparability criteria and is not included in the final boundary list (Table 23).

Table 23. Boundary values of the different benthic assessment approaches after intercalibration. The boundaries in red are those changed after boundary harmonization.

		Ecological quality ratios			
Country	Benthic assessment	High-good boundary	Good-moderate	Moderate-poor boundary	Poor-bad boundary

	approach		boundary		
Denmark	DKI	0.8	0.6	0.4	0.2
France	m-AMBI	0.77	0.53	0.38	0.2
Germany	m-AMBI	0.85	0.70	0.4	0.2
Netherlands	BEQI2	0.78	0.58	0.38	0.18
Norway	NQI	0.72	0.63	0.4	0.2
Portugal	BAT	0.79	0.58	0.44	0.27
Spain (m-AMBI)	m-AMBI	0.77	0.63	0.38	0.2
United Kingdom / Ireland	IQI	0.75	0.64	0.44	0.24
Belgium	BEQI	0.8	0.6	0.4	0.2

7.2 Correspondence common types versus national types

The common type (NEA1-26) is recognized as type in every Member State and is related to the national types.

7.3 Gaps of the current intercalibration

The current intercalibration exercise and this technical report summarize all required information to finalize the intercalibration of the benthic assessment approaches for coastal waters in the NEA-GIG region for type NEA1/26. One aspect of the intercalibration guidance that cannot be met completely is the real quantification of the pressure gradient in the common dataset to avoid expert judgment. This is the best that can be obtained with the available information. Otherwise the intercalibration exercise has to restart with new data collection, which contains ecological data accompanied with quantitative pressure info.

8 Ecological characteristics

8.1 Description of reference or alternative benchmark communities

The description of the benthic community characteristics at reference or alternative benchmark is summarized in Table 24. This information is generated from the WISER database. Only for France, Norway and Spain (Andalusia) this information is not available.

8.2 Description of good status communities

The description of the benthic community characteristics at good status is summarized in Table 24. This information is generated from the WISER database. Only for Norway and Spain (Andalusia) this information is not delivered.

Table 24. Overview of the description by the Member States of the macro-invertebrate reference community and good status community

Member State	Description of reference community	Description of good status community
Belgium	The reference benthic characteristics of each habitat were defined on the randomization of a reference dataset, reflecting the spatial and temporal variability expected in that habitat, based on existing data and knowledge.	Is not defined textually.
Germany	Benthic communities, species numbers, diversity typically for the habitat (sediment, salinity, exposure)- low number of opportunistic species.	High portion of sensitive taxa, complex communities, low number of opportunists, high species number and high diversity assemblages.
Denmark	High diversity (H and richness). Dominance of sensitive species sensu Borja et al. 2000.	High diversity (H and richness). Dominance of sensitive species *sensu* Borja et al. (2000).
France	High diversity (H and richness). Dominance of pollution sensitive taxa *sensu* Borja *et al,*. 2000.	Richness and diversity are slightly reduced in comparison to values under reference conditions, while variables according to habitat (community abundance as assessed by AMBI) are slightly unbalanced: sensitive taxa (EG I) abundance may range from high sub-dominant to absent; indifferent taxa (EG II) are of low sub-dominant abundance; tolerant taxa (EG III) of dominant abundance; abundance of opportunistic (EG IV) and indicator taxa (EG V) may range from negligible or low to comparable abundance with indifferent taxa (EG II).
Netherlands	level 3: reference community description is specific for each individual water body. Reference conditions based on historical data from 1970's.<br>Furthermore a general description is given (in Dutch) in:<br>STOWA (2009) Referenties en maatlatten voor natuurlijke watertypen. report 2007-32	n.a.
Norway	n.a.	n.a.
Portugal	Reference condition macrobenthic communities are dominated by pollution sensitive taxa (AMBI Ecological Group (EG) I taxa), have low relative abundance of indifferent (EG II) and tolerant (EG III) taxa and negligible relative abundance of opportunist (EG IV) and pollution indicator (EG V) taxa. High numbers of taxa with an even abundance distribution throughout the community is also indicative of reference conditions.	Community species richness (Margalef) and equitability (Shannon-Wiener) values are slightly reduced in comparison to values under reference conditions. While variable according to habitat, community composition (as assessed by AMBI) is slightly unbalanced. Community composition still dominated by EG I and II taxa. Slight reduction of sensitive taxa (EG I), and slight increase on tolerant taxa (EG III).
Spain (m-AMBI)	See: Borja, A., F. Aguirrezabalaga, J. Martinez, J.C. Sola, L. Garciaarberas &amp; J.M. Gorostiaga, 2003. Benthic	Borja, A., A.B. Josefson, A. Miles, I. Muxika, F. Olsgard, G. Phillips, J.G. Rodríguez & B. Rygg, 2007. An approach to the

	communities, biogeography and resources management. In: Borja, A. &amp; M. Collins, (Ed.). Ocenaography and Marine Environment of the Basque Country, Elsevier Oceanography Series n. 70: 27-50.	intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. Marine Pollution Bulletin 55: 42-52.
Spain (BO2A)	n.a.	n.a.
United Kingdom/Ireland	Reference condition macrobenthic communities are dominated by pollution sensitive taxa (AMBI Ecological Group (EG) I taxa), have low relative abundance of indifferent (EG II) and tolerant (EG III) taxa and negligible relative abundance of opportunist (EG IV) and pollution indicator (EG V) taxa. High numbers of taxa with an even abundance distribution throughout the community is also indicative of reference conditions.	Taxa number and Simpsons evenness are slightly reduced in comparison to values under reference conditions, while variables according to habitat (community abundance as assessed by AMBI) are slightly unbalanced: sensitive taxa (EG I) abundance may range from high sub-dominant to absent; indifferent taxa (EG II) are of low sub-dominant abundance; tolerant taxa (EG III) of dominant abundance; abundance of opportunistic (EG IV) and indicator taxa (EG V) may range from negligible or low to comparable abundance with indifferent taxa (EG II).

9 References

Birk, S., Strackbein, J. & Hering, D., 2010. WISER methods database.
Version: March 2011. Available at http://www.wiser.eu/results/method-database/.

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

Borja, A., Franco, F., Valencia, V., Bald, J., Muxika, I., Belzunce, M.J., et al., 2004. Implementation of the European Water Framework Directive from the Basque country (northern Spain): a methodological approach. Marine Pollution Bulletin 48 (3-4), 209-218.

Borja, A., Josefson, A.B., Miles, A., Muxika, I., Olsgard, F., Phillips, G., Rodriguez, J.G., Rygg, B., 2007. An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. Marine Pollution Bulletin 55, 42–52.

Borja, A., Muxika, I., Rodriguez, J.G., 2009. Paradigmatic responses of marine benthic communities to different anthropogenic pressures, using M-AMBI, within the European Water Framework Directive. Marine Ecology – An Evolutionary Perspective 30, 214–227

Borja et al., A., A. Miles, A. Occhipinti-Ambrogi, T. Berg, 2009. Current status of macroinvertebrate methods used for assessing the quality of European marine waters: implementing the Water Framework Directive. Hydrobiologia, 633: 181-196.

Davies, Susan P., 2012. Peer review of the intercalibration exercise phase II: European water framework directive.

Fitch, J.E., Cooper, K.M., Crowe, T.P., Hall-Spencer, J.M., Philips, G., 2014. Response of multi-metric indices to anthropogenic pressures in distinct marine habitats: The need for recalibration to allow wider applicability. Marine Pollution Bulletin dx.doi.org/10.1016/j.marpolbul.2014.07.056.

Josefson, A.B., Blomqvist, M., Hansen, J.L.S., Rosenberg, R., Rygg, B., 2009. Assessment of marine benthic quality change in gradients of disturbance: comparison of different Scandinavian multi-metric indices. Marine Pollution Bulletin 58, 1263-1277

Marques J.C., F. Salas, J. Patrício, H. Teixeira & J.M. Neto. 2009. Ecological Indicators for Coastal and Estuarine Environmental Assessment. A user guide. Ed. 00, ISBN: 978-1-84564-209-9. UK: WIT Press.

Phillips, G.R., Anwar, A., Brooks, L., Martina, L.J., Miles, A.C., Prior, A., 2014. Infaunal Quality Index: Water Framework Directive Classification Scheme for Marine Benthic Invertebrates Environment Agency (UK) R&D Technical Report, No SC080016."

Teixeira, H., Neto, J.M., Patrício, J., Veríssimo, H., Pinto, R., Salas, F. & Marques, J.C. 2009. Quality assessment of benthic macroinvertebrates under the scope of WFD using BAT, the Benthic Assessment Tool. Marine Pollution Bulletin, 58: 1477-1486. (doi:10.1016/j.marpolbul.2009.06.006).

Van Hoey, Gert; Borja, Angel; Birchenough, Silvana; Degraer, Steven; Fleischer, Dirk; Kerckhof, Francis; Magni, Paolo; Buhl-Mortensen, Lene; Muxika, Iñigo; Reiss, Henning; Schröder, Alexander; Zettler, Michael, 2010. The use of benthic indicators in Europe: from the Water Framework Directive to the Marine Strategy Framework Directive. Marine Pollution Bulletin 60: 2187-2196

Van Hoey, Gert; David Cabana Permuy; Magda Vincx ; Kris Hostens, 2013. An Ecological Quality Status assessment procedure for soft-sediment benthic habitats: Weighing alternative approaches. Ecological Indicators 25, 266-278

Van Hoey, G., Vanaverbeke, J., Degraer, S., 2014. Study related to the realization of the Water Framework Directive intercalibration for the Belgian Coastal waters, to design the descriptive elements 1 and 6 of the Marine Strategy Framework Directive and the nature objectives of the Habitat Directive for invertebrate bottom fauna of soft substrates. ILVO-mededeling 170.

van Loon, W.M.G.M., Boon A.R., Giitenberger, A., Walvoort, D.J.J., Lavaleye, M., Duineveld, G.C.A. and Verschoor A.J., 2015. Application of the Benthic Ecosystem Quality Index 2 to benthos in Dutch transitional and coastal waters. Journal of Sea Research 103, 1-13

10 Annex 1: Sampling and data processing information

	Denmark	Belgium	United Kingdom / Ireland	Germany
Sampling guideline	Holme, N.A. & A.D. McIntyre, 1984. Methods for the study of marine benthos. IBP Handbook 16, Blackwell, Oxford.	ISO standard (ISO 16665:2005(E)) "Water quality – Guidelines for quantitative sampling and sample processing of marine soft-bottom macrofauna".	ISO standard (ISO 16665:2005(E)) "Water quality – Guidelines for quantitative sampling and sample processing of marine soft-bottom macrofauna".	Muster-Standardarbeitsanweisung für Laboratorien des Bund/Länder-Messprogramms Prüfverfahren-SOP: Makrozoobenthos-Untersuchungen in marinen Sedimenten (Weichboden)
Sampling description	Three to six Van Veen are taken (blindly) at a site or area using ships. Alternatively 40 Haps are taken, one at each geographical position, mostly regularly spaced within an area. For the case of point sites, 5-10 Haps are taken blindly at each site and sampling occasion.	Habitat approach, the main habitat types within a water body were sampled in such way to get a confident ecological quality classification (enough samples, spatially and eventually temporal distributed within a habitat). In such way, the amount of samples per habitat can vary between 10 to 40 samples. The samples were taken randomly within the habitat area.	Sampling design variable according to UK and Ireland monitoring authority. Samples taken from soft bottom habitats, either i) spread as single samples or ii) taken as replicates at one or more stations. Surveys are undertaken either i) annually or ii) once in a reporting cycle according to monitoring authority. Biological samples require an associated sediment field sample for particle size analysis and supporting depth and salinity information.	5 to 20 sediment samples are taken from 1 ecotope. Each sample is sieved separately (1mm, 0,5mm mud) and residue is stored and transferred to the laboratory. Benthic species are separated and identified to the lowest taxonomic level.

	Spain (m-AMBI)	Netherlands	Portugal
Sampling guideline	ISO standard (ISO 16665:2005(E)) “Water quality – Guidelines for quantitative sampling and sample processing of marine soft-bottom macrofauna”.	STOWA, 2009. Instructie; Richtlijn Monitoring Oppervlaktewater en Protocol Toetsen en Beoordelen (28 april 2009); STOWA, NN. Quality Handbook Hydrobiology (in prep).	ISO standard (ISO 16665:2005(E)) “Water quality – Guidelines for quantitative sampling and sample processing of marine soft-bottom macrofauna”.
Sampling description	2-6 sampling locations are visited per water body once a year in winter. At each location 3 van Veen grab replicates are taken (0.1 square-metres each), and sieved on board by 1 mm mesh.	Normally sediment cores are collected at sampling stations with a device like the Reineck Box corer operated from a ship for subtidal stations. The sediment is washed through a 1mm mesh. Specimens are sorted form the residue, identified to the species level, counted and weighed. Biomass is most accurately measured by the difference between dry weight and ash weight, the ash free dry weight AFDW.	Biological samples are collected from soft bottom habitats, by using a 0.1 m^2 sampling area Van Veen Grab (or equivalent). Sampling stations are placed at representative sites of water bodies, and in sufficient number to cover natural variations, according to monitoring authority. A minimum of 3 replicates per sampling station are collected. Biological samples require an associated sediment field sample for particle size and organic matter content analysis, and supporting depth, salinity, temperature, and chemical parameters information (bottom water).

	Spain (BO2A)	France	Norway
Sampling guideline	ISO standard (ISO 16665:2005(E)) “Water quality – Guidelines for quantitative sampling and sample processing of marine soft-bottom macrofauna”.	ISO standard (ISO 16665 :2005(E)) “ Water Quality – Guidelines for quantitative sampling and sample processing of marine soft-bottom macrofauna “	ISO standard (ISO 16665:2005)) Water quality – “Guidelines for quantitative sampling and sample processing of marine sotf-bottom macrofauna”.
Sampling description	Overall, one sampling station was defined for each water body, provided it was considered representative of the whole water body. Soft-bottom sampling is carried out, in broad daylight, with the vessels owned by the Regional Agency of Environment (Regional Government of Andalucía), except in shallower areas where it may be carried out by direct sampling or with small auxiliary vessels. A sample corresponds to the average of 3 sampling units. The sediment collected in each sampling unit is posteriorly sieved through a 0.5 mm mesh.	Above all, a monitoring location is defined on the basis of its representativeness across the whole WB. In order to consider the intra-stational variability, it was decided that each location will be studied in 3 points (3 replicates per point), bringing to 9 the number of replicates for each monitoring locations. In subtidal areas, the sampling (one replicate) is carried out by the mean of a grab (area=0.1 m^2) and sieved on board by a 1mm mesh. In intertidal areas, the sampling (one replicate) is carried out by the mean of a hand corer (area = .029 m^2) and sieved by a 1mm mesh. Biological samples require an associated sediment field sample (each of the 3 points constituting the monitoring location), for analysis of particle size and organic matter.	Samples collected by using a 0.1m^2 van Veen grab, and sieved on board by 1 mm mesh. 4 replicates per station. An associated sediment field sample taken for grain size and TOC.

	Denmark	Belgium	United Kingdom / Ireland	Germany
Method to select the survey site	Expert knowledge, Random sampling/surveying	Stratified Random sampling/surveying	Stratified Random sampling/surveying	Expert knowledge, Random sampling/surveying
Sampling Device	Corer, Grab	Grab	Corer, Grab	Corer, Grab
Specification of sampling device	0.1 m² Van Veen Grab, 0.0143 m² Haps-corer	Van Veen Grab (0.1m²)	Van Veen Grab (0.1m²), Day Grab (0.1m²), Hand Core (0.01m²)	Van Veen-grab (0.1m²), corers with 9-15cm diameter
Sampled habitat	Single habitat(s)	All available habitats per site (Multi-habitat)	Single habitat(s)	Single habitat(s)
Specification of sampled habitat	Soft bottom (sand - mud)	soft bottom sediments (muddy sediments [Macoma balthica habitat], fine muddy sand [Abra alba habitat], clean sands [Nephtys cirrosa habitat])	Soft bottom	Soft bottom
Sampled zones in tidal areas	Subtidal zone	Subtidal zone	Both tidal zones	Both tidal zones
Sampling months	April to June	October	February to May (current recommended target months)	May or September/October
Number of sampling occasions in time	One per year	One occasion per year (preferential autumn)	Minimum of one occasion for classification (varies between 1-3 for UK and Ireland monitoring authorities)	One occasion per sampling season

	Spain (m-AMBI)	Netherlands	Portugal
Method to select the survey site	Expert knowledge; Fixed sampling stations, representative of the water body	Fixed locations	Expert knowledge
Sampling Device	Grab	Corer	Grab
Specification of sampling device	Van Veen Grab	corer tube; box corer (e.g. Reineck Box corer), flushing sampler (only in saline lakes 0-2 m)	Van Veen Grab (0.1 m²) or equivalent
Sampled habitat	Single habitat(s)	Single habitat(s)	Single habitat(s)
Specification of sampled habitat	Soft bottom	All present habitats in the water body.	Soft bottom (sandy-mud)
Sampled zones in tidal areas	Both tidal zones	Both tidal zones	Subtidal zone
Sampling months	Winter (Basque country); Summer (Cantabria)	Coastal water types (NEA1, NEA3): March 1st to June 15th	February - March
Number of sampling occasions in time	Once a year	Minimum one survey per year (preferably fall), and scores and classification preferably averaged over three years.	Minimum of one occasion per the chosen sampling season

	Spain (BO2A)	**France**	**Norway**
Method to select the survey site	Expert knowledge	Expert knowledge, Fixed sampling stations representative of the WB	Expert knowledge
Sampling Device	Grab	grab	grab
Specification of sampling device	Van Veen Grab	Van Veen Grab or Day Grab or Smith-McIntyre Grab	Van Veen grab (0.1 m^2)
Sampled habitat	Single habitat(s)	Single habitat(s)	Single habitat(s)
Specification of sampled habitat	Soft bottom	Soft bottom	Soft bottom
Sampled zones in tidal areas	Subtidal zone	Subtidal and intertidal zone	Subtidal zone
Sampling months	Summer: June - August	From February to April	May, August, September
Number of sampling occasions in time	One occasion per sampling season	One occasion per sampling season	one per year

	Denmark	Belgium	United Kingdom / Ireland	Germany
Number of spatial sampling replicates	Six 0.1 m² Van Veen, or 40 Haps samples	Depends on habitat type samples (18 for Macoma balthica habitat, 20 for Abra alba habitat and 18 for Nephtys cirrosa habitat)	Variable according to habitat, number of years/ stations, methodology and required confidence.	6-10 replicates per ecotope
Total sampled area or duration	0.6 m²	Depends on habitat type samples (1.8 m² for Macoma balthica habitat, 2.0 m² for Abra alba habitat and 1.8 m² for Nephtys cirrosa habitat)	Variable according to habitats, number of years/ stations, methodology and required confidence.	1 m² per ecotope, 2-4 ecotopes per water body, average of several years
Minimum size of sampled organisms	1 mm (mesh-size of sieve)	1 mm	1000 µm (Coastal Waters)	1000 µm, 500 µm in mud sediments
Sample treatment	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.
Level of taxonomic identification	Other, Species/species groups	Family, Genus, Other, Species/species groups	Species/species groups	Genus, Species/species groups

	Spain (m-AMBI)	Netherlands	Portugal
Number of spatial sampling replicates	3 replicates per station (2-6 stations per water body)		Variable according to habitat, number of years/stations, and required confidence.
Total sampled area or duration	0.3 m² (each replicate has 0.1 m²)		Variable according to habitat, number of years/stations, and required confidence.
Minimum size of sampled organisms	1 mm mesh	1 mm	1000 µm for Coastal Waters
Sample treatment	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.
Level of taxonomic identification	Species/species groups	Species/species groups	Other, Species/species groups

	Spain (BO2A)	France	Norway
Number of spatial sampling replicates	3	3	4
Total sampled area or duration	0.025 m² (average of 3 spatial replicates)	0,9m² (3 locations, 3 replicates per location)	0,4m²
Minimum size of sampled organisms	0.5 mm mesh size	1 mm	1 mm
Sample treatment	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.	Organisms of the complete sample are identified.
Level of taxonomic identification	Family, Other, Species/species groups	Species/species groups	Species/species groups

	Denmark	**Belgium**	**United Kingdom / Ireland**	**Germany**
Specification of level of determination	Species level (or if not possible to determine, genus or family level): Echinodermata, Polychaeta, Crustacea, Mollusca; Higher Group level: Nemertea, Nematoda, Turbellaria	Determination to the lowest level possible. Oligochaeta to level of order. Some Polychaeta to the level of family (Cirratulidae). Taxonomy between assessment and reference data were set consistently.	n.a.	All to species level except some Oligochaeta, Diptera, Priapulida,...
Determination of abundance	Individual counts	Individual counts	Individual counts	Individual counts
Abundance is related to	Area	Area	Area	Area
Unit of the record of abundance	individuals per m²	Number of individuals per one square-metre	Number of individuals per area of sample	Number of individuals per one m²
Quantification of biomass	n.a.	Wet weight	n.a.	n.a.
Other biological data	none	none	none	none

	Spain (m-AMBI)	**Netherlands**	**Portugal**
Specification of level of determination	Some groups can be indentified to higher taxonomical levels.	n.a.	Truncation rules (Borja et al., 2007)
Determination of abundance	Individual counts	Individual counts	Individual counts
Abundance is related to	Area	Area	Area
Unit of the record of abundance	Number of individuals per one m²	Number of individuals per one m²	Number of individuals per sampling area
Quantification of biomass	n.a.	n.a.	n.a.
Other biological data	none		none

	Spain (BO2A)	France	Norway
Specification of level of determination	Plathelminthes, Nemertina and Nematoda to phylum level; oligochaetes to sub-class level; harpacticoid copepods to order level; insects to class level, except chironomids; chironomids to family level; hemichordates to phylum level.	Species level, except for the following groups: *Echiura, Hemichordata, Hydrozoa, Insecta, Nemertea, Oligochaeta, Phoronida, Platyhelminthes* et *Priapulida*	Species level or lowest level possible
Determination of abundance	Individual counts	Individual counts	Individual counts
Abundance is related to	Area	Area	Area
Unit of the record of abundance	Number of individuals per one m^2	Number of individuals per 0,1 m^2	Number of individuals per 0,1 m^2
Quantification of biomass	n.a.	n.a.	n.a.
Other biological data	none	none	none

	Denmark	Belgium	United Kingdom / Ireland	Germany
Special cases or additions of sampling	none	none	Presence/absence recorded where taxa are unsuitable for quantification (e.g. colonial taxa). Truncation rules are applied to the data to exclude non-benthic and non-invertebrate fauna from the IQI assessment.	none
Comments on 'data acquisition' part	The DKI is applied on 0.1 m² samples and therefore Haps samples are pooled to this sample size (6-7 Haps)	none	none	none

	Spain (m-AMBI)	**Netherlands**	**Portugal**
Special cases or additions of sampling	none		Presence/absence recorded where taxa are unsuit
Comments on 'data acquisition' part	none	The present Dutch surveillance monitoring (BIOMON program) can be split up in 3 areas,with differences in sampling strategy, namely (1) the Delta area, (2) the Dutch coast and (3) the Waddenzee; Eems-Dollard. The macrobenthic fauna monitoring activities are all under the responsibility of one agency (but different offices) could lead to some small taxonomic differences in the methodology. Since these differences also exist in the reference data sets, it is expected that the impact on the EQR-scores are very small.	none

	Spain (BO2A)	**France**	**Norway**
Special cases or additions of sampling	none	none	none
Comments on 'data acquisition' part	none	none	none

11 Annex 2: Alternative benchmark approach (based on biotic variables)

This procedure to determine the benchmark samples out of the common dataset is not accepted by JRC and the review panel (personal communication, Fuensanta). In the authors' point of view, this gives a reliable, objective alternative for the determination of the benchmark samples, which is explained in this annex.

An alternative procedure for the selection of benchmark sites can be used in this intercalibration due to the absence of quantitative and even qualitative pressure data of each sample within the common dataset. The collection of such information on sample level in a standard way is rather impossible (except for some sub-data sets, e.g., the Garroch Head analyses), due to the absence of such information at the Member State level. Alternative pressure quantifications, as general pressure index, distance from the coast, are not appropriate for this NEA-GIG dataset due to the type of data (many samples from the same location), indirect influence of harbors and rivers being rather low for the majority of samples, other pressures being probably more important (local pollution [such as dumping activity at Garroch Head dataset, Basque Country dataset is at a submarine outfall], fishery, and the like). Besides this, the variation in pressure quantification will be low and many samples will be cataloged within the same pressure status, due to the absence of detailed pressure information. Such a general pressure index approach was tested for the intercalibration of transitional waters within the NEA-GIG region in phase II and was inadequate.

For the dataset, where some pressure information was available (see Garroch Head dataset), we could objectively distinguish least disturbed samples (lower copper concentration), and showed that there is some variability in the classification of those samples by the different benthic assessment approaches (see section 3.2). Unfortunately this does not meet the set-up of the benchmarking in the intercalibration guidance (benchmark sites in each Member State are necessary).

An approach that estimates the benthic conditions under least disturbed circumstances could be the selection of samples with the highest diversity characteristics. In theory, areas characterized by samples with a high diversity (expressed as any type of diversity indices) are less subjected to pressures on the system than areas characterized by lower diversity (Pearson & Rosenberg, 1978) (Figure 9). This relationship is not linear, but a clear gradient exists. The multivariate analysis on the common dataset (see higher) show the benthic variability within the data, but also a clear gradient in benthic characteristics (Figure 7). The gradient within these benthic univariate parameters can be used as a proxy for the pressures on the samples of the NEA-GIG common dataset.

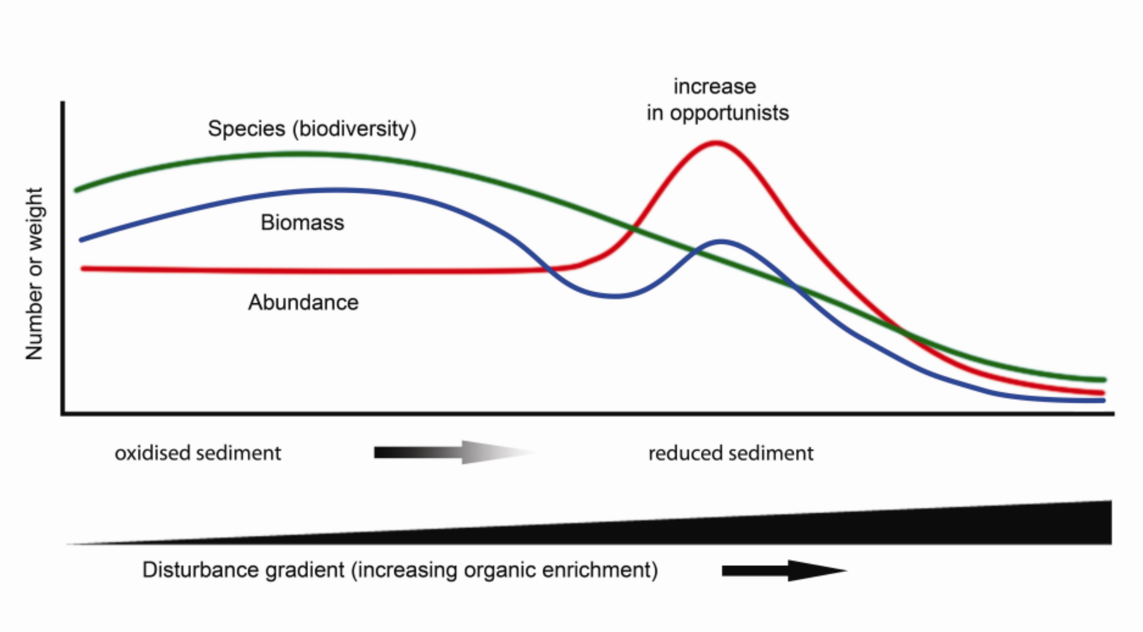


Figure 9. Pearson & Rosenberg relation between the benthic characteristics and a disturbance gradient (organic pollution)

Therefore, the X-axis of Figure 7 can be used as a proxy for the pressure gradient within the NEA-GIG benthic coastal dataset (or the first principle component of the multivariate analysis). Along this gradient, the samples clustered in group E and F can be considered as alternative benchmark sites, because they are characterized by similar diversity characteristics. These diversity characteristics should reflect the status of benthos under least disturbed conditions. The amount of samples in group E and F is high, which allows a good characterization of the natural variability of the benthos within the NEA-GIG region under least disturbed conditions and covered the upper part of the theoretical gradient in benthic characteristics along a disturbance gradient (Figure 9).

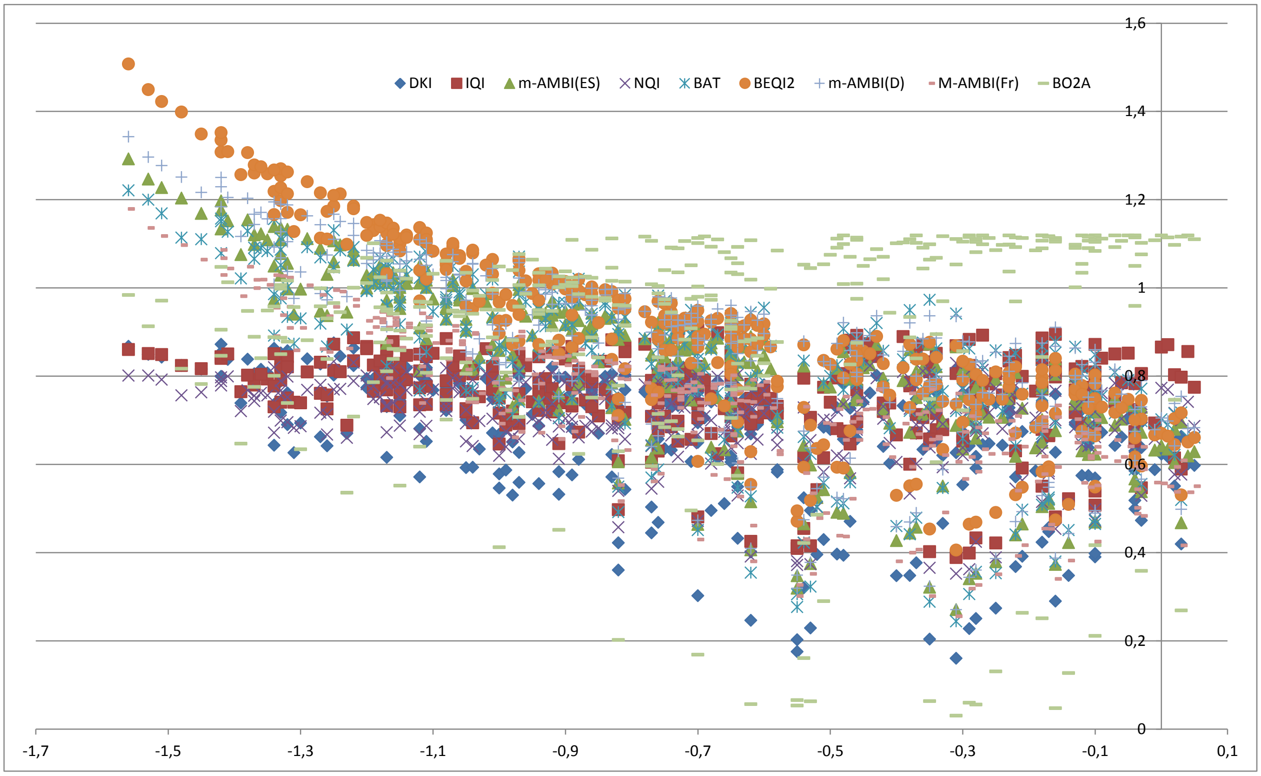


Figure 10. EQR values of the assessment approaches for the benchmark samples.

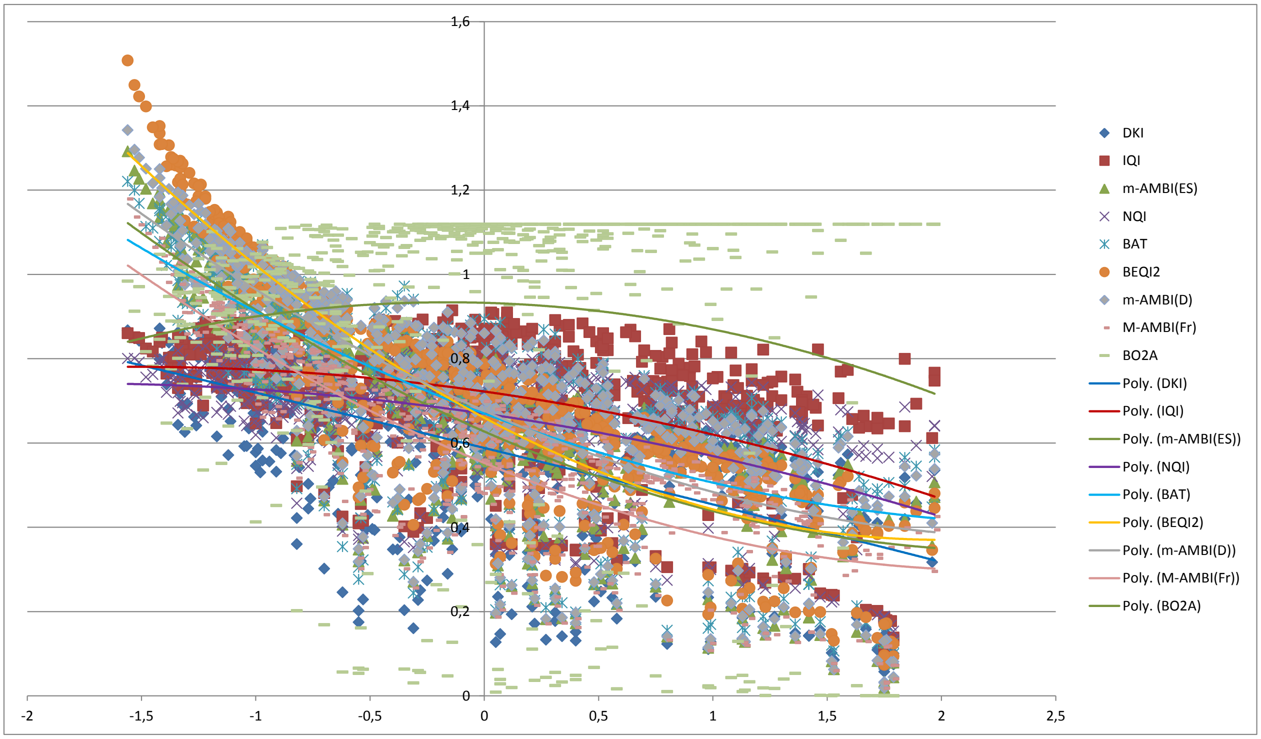


Figure 11. EQR values of the different methods with trend line (2nd order polynomial trend line) along the pressure gradient (X-axis values of MDS).

The analysis of those benchmark sites (Figure 10) and gradients (Figure 11) show that most benthic assessment approaches have a high variability along the gradient, but were more or less in line with each other. The BO2A shows the lowest affinity with this gradient and the highest variability in EQR values for the benchmark sites. The trend line of the BO2A is not in line with the others. Beside this, the M-AMBI approaches, BAT and BEQI2 show the same trend line, whereas the NQI and IQI deviate a little bit from this. They show a more buffered pattern, characterized by less variability at high status, which is related to their algorithm (see 3.3 assessment concept).

11.1 Benchmark standardization?

11.1.1 General pattern

In general, significant differences between the different assessment approaches were observed on the benchmark sites within the common dataset (Figure 12, Table 25), except in a few cases (DKI and NQI; IQI and m-AMBI (Fr); BAT and m-AMBI (ES &D); BEQI2 and m-AMBI (D)).

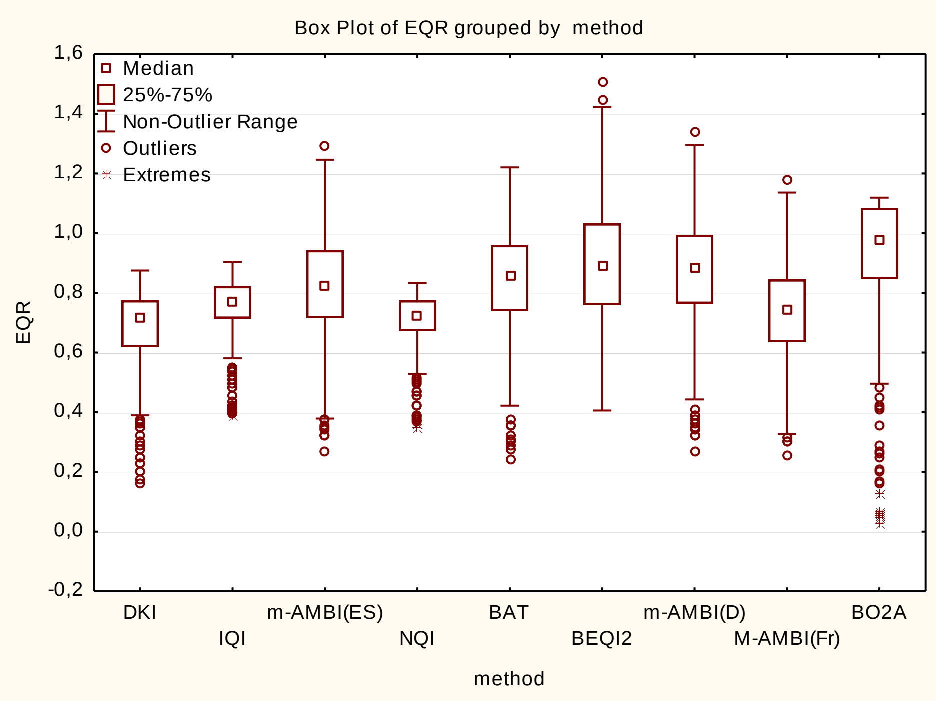


Figure 12. Box-whisker plot of the EQR values at the benchmark sites for the different benthic assessment methods, with indication of the outlier values.

Table 25. Kruskal-Wallis p levels by comparison the EQR values of each approach for the benchmark sites (samples of cluster group E and F).

	DKI	IQI	m-AMBI(E	NQI	BAT	BEQI2	m-AMBI(D)	M-AMBI(Fr)	BO2A
DKI		0,000000	0,000000	1,000000	0,000000	0,000000	0,000000	0,000304	0,000000
IQI	0,000000		0,000005	0,000059	0,000000	0,000000	0,000000	1,000000	0,000000
m-AMBI(ES)	0,000000	0,000005		0,000000	1,000000	0,000360	0,021182	0,000000	0,000000
NQI	1,000000	0,000059	0,000000		0,000000	0,000000	0,000000	0,013701	0,000000
BAT	0,000000	0,000000	1,000000	0,000000		0,041123	0,829322	0,000000	0,000000
BEQI2	0,000000	0,000000	0,000360	0,000000	0,041123		1,000000	0,000000	0,032067
m-AMBI(D)	0,000000	0,000000	0,021182	0,000000	0,829322	1,000000		0,000000	0,000607
M-AMBI(Fr)	0,000304	1,000000	0,000000	0,013701	0,000000	0,000000	0,000000		0,000000
BO2A	0,000000	0,000000	0,000000	0,000000	0,000000	0,032067	0,000607	0,000000	

Benchmark standardization will correct for differences in median EQR values between the Member States benchmark sites obtained by certain assessment approaches. Therefore, we analyze the median EQR values of the Member States (per type [<30m and >30m]) benchmark sites for each of the different assessment approaches separately. Those median values will be corrected by the benchmark standardization procedure and this correction will be more obvious for cases where the medians are significantly different.

11.1.2 Benthic assessment approaches at the Member States benchmark sites

1) M-AMBI (Germany)

The EQR values at the benchmark sites of Spain, France and Norway are significantly different from the German and UK-type 1 benchmark sites by the m-AMBI (Germany) approach (Figure 13, Table 26). UK-type 2, the Dutch and UK-type 2 are also significantly different with the French benchmark sites.

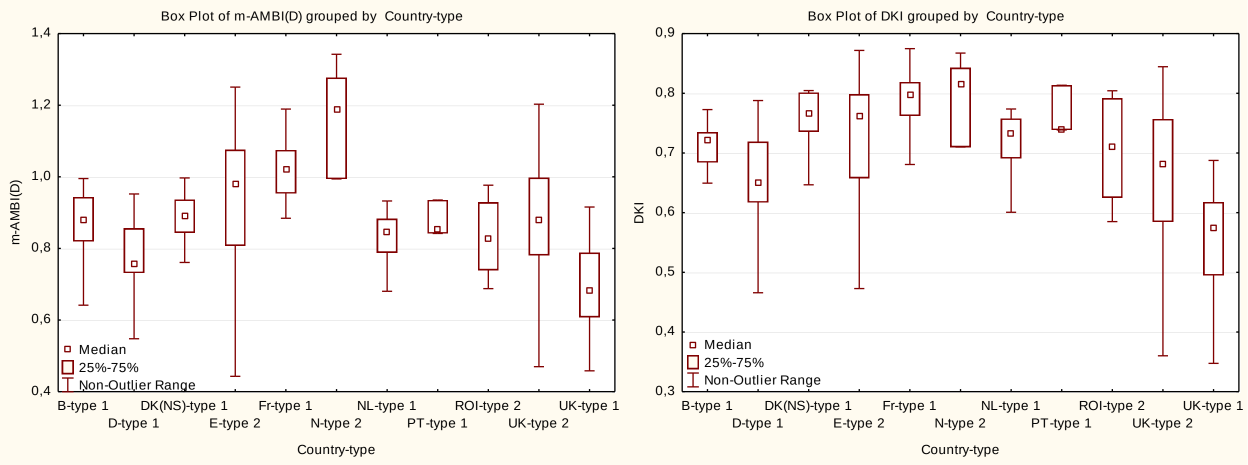


Figure 13. Box-whisker plot (median, percentile values and no-outlier range) of the EQR values at the Member States benchmark sites with the m-AMBI (Germany) (left) and the DKI (Denmark) (right).

Table 26. Kruskal-Wallis p levels (multiple comparison of mean ranks for all groups) by comparison the EQR values of each Member State benchmark sites with the m-AMBI (Germany) (white fields) and the DKI (Denmark) (grey fields).

DKI ➔ m-AMBI(D) ↓	B-type 1	D-type 1	DK(NS)-type 1	E-type 2	Fr-type 1	N-type 2	NL-type 1	PT-type 1	ROI-type 2	UK-type 2	UK-type 1
B-type 1		1,000	1,000	1,000	0,021	1,000	1,000	1,000	1,000	1,000	0,017
D-type 1	1,000		0,219	0,065	0,000	0,107	1,000	1,000	1,000	1,000	0,692
DK(NS)-type 1	1,000	1,000		1,000	1,000	1,000	1,000	1,000	1,000	0,784	0,000
E-type 2	1,000	0,000	1,000		0,377	1,000	1,000	1,000	1,000	0,219	0,000
Fr-type 1	0,069	0,000	1,000	1,000		1,000	0,173	1,000	1,000	0,000	0,000
N-type 2	0,504	0,000	1,000	1,000	1,000		1,000	1,000	1,000	0,391	0,000
NL-type 1	1,000	1,000	1,000	0,405	0,002	0,064		1,000	1,000	1,000	0,002
PT-type 1	1,000	1,000	1,000	1,000	1,000	1,000	1,000		1,000	1,000	0,207
ROI-type 2	1,000	1,000	1,000	1,000	0,205	0,449	1,000	1,000		1,000	0,156
UK-type 2	1,000	0,013	1,000	1,000	0,000	0,230	1,000	1,000	1,000		0,003
UK-type 1	0,117	1,000	0,350	0,000	0,000	0,000	1,000	1,000	1,000	0,000	

2) DKI

The EQR values at the UK-type1 are significantly different (lower) from most other benchmark sites, except the Portuguese, Irish and German sites (Figure 13,Table 26). The French and UK-type 1, Belgian and German sites are also significantly different to the DKI benthic assessment approach.

3) M-AMBI of France

The EQR values at the benchmark sites of Spain, France and Norway are significantly different from the German and UK-type 1 benchmark sites by the m-AMBI (France) approach (Figure 14,Table 27). UK-type 2, the Dutch and UK-type 2 are also significantly different from the French benchmark sites. The benchmark sites of the Member States which are significantly different from each other are the same as with the m-AMBI approach of Germany and Spain.

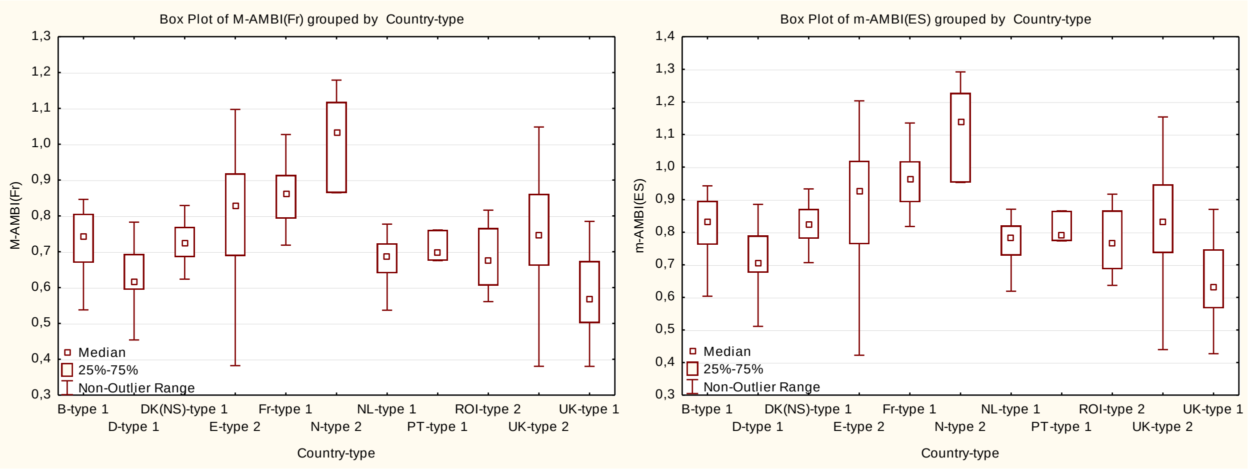


Figure 14. Box-whisker plot (median, percentile values and no-outlier range) of the EQR values at the Member States benchmark sites with the m-AMBI (France) (left) and the m-AMBI (Basque Country, Cantabria region) (right).

Table 27. Kruskal-Wallis p levels (multiple comparison of mean ranks for all groups) by comparison the EQR values of each Member State benchmark sites with the m-AMBI (France) (white fields) and the m-AMBI (Basque Country; Cantabria) (grey fields).

BC/CR ➔ France ↓	B-type 1	D-type 1	DK(NS)-type 1	E-type 2	Fr-type 1	N-type 2	NL-type 1	PT-type 1	ROI-type 2	UK-type 2	UK-type 1
B-type 1		0,840	1,000	1,000	0,102	0,471	1,000	1,000	1,000	1,000	0,142
D-type 1	0,687		1,000	0,000	0,000	0,000	1,000	1,000	1,000	0,002	1,000
DK(NS)-type 1	1,000	1,000		1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,777
E-type 2	1,000	0,000	1,000		1,000	1,000	0,138	1,000	1,000	1,000	0,000
Fr-type 1	0,128	0,000	0,729	1,000		1,000	0,001	1,000	0,163	0,002	0,000
N-type 2	0,439	0,000	0,894	1,000	1,000		0,026	1,000	0,292	0,309	0,000
NL-type 1	1,000	1,000	1,000	0,066	0,000	0,014		1,000	1,000	1,000	1,000
PT-type 1	1,000	1,000	1,000	1,000	1,000	1,000	1,000		1,000	1,000	1,000
ROI-type 2	1,000	1,000	1,000	1,000	0,141	0,219	1,000	1,000		1,000	1,000
UK-type 2	1,000	0,001	1,000	1,000	0,005	0,384	1,000	1,000	1,000		0,000
UK-type 1	0,190	1,000	1,000	0,000	0,000	0,000	1,000	1,000	1,000	0,000	

4) M-AMBI of Spain

The EQR values at the benchmark sites of Spain, France and Norway are significantly different from the German and UK-type 1 benchmark sites by the m-AMBI (Spain) approach (Figure 14, Table 27). UK-type 2, the Dutch and UK-type 2 are also significantly different from the French benchmark sites. The benchmark sites of the Member States which were significantly different from each other are the same as with the m-AMBI approach of Germany and France.

5) BEQI2 of the Netherlands

The EQR values at the benchmark sites of Spain, France and Norway are significantly different from the German, Dutch and UK-type 1 benchmark sites by the BEQI2 approach (Figure 15, Table 28). The EQR values of the benchmark site of UK-type 2 and UK-type 1 and Germany are also significantly different. The benchmark sites of the Member States which were significantly different from each other are the same as with the m-AMBI approach of Germany, Spain and France.

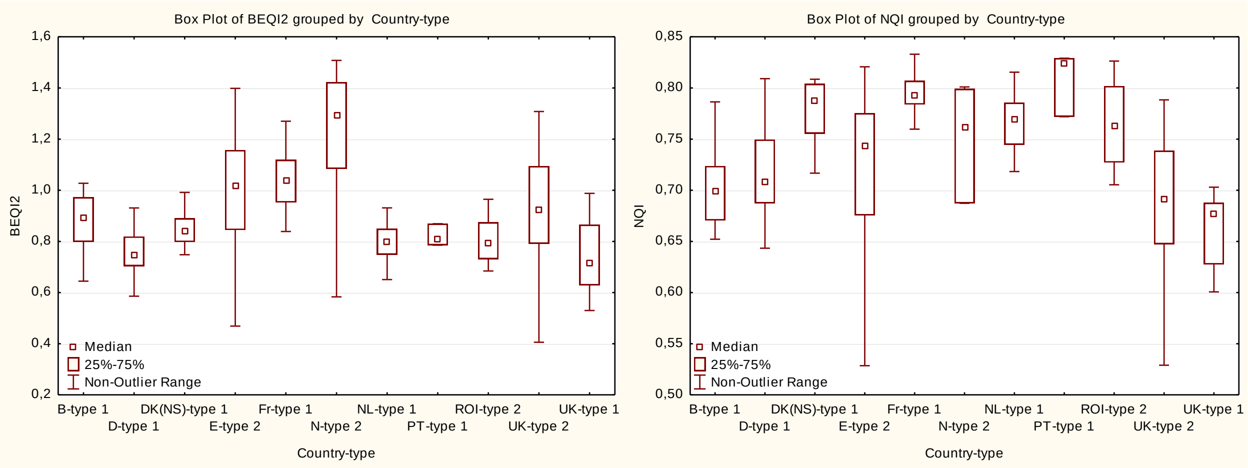


Figure 15. Box-whisker plot (median, percentile values and no-outlier range) of the EQR values at the Member States benchmark sites with the BEQI2 (the Netherlands) (left) and the NQI (Norway) (right).

Table 28. Kruskal-Wallis p levels (multiple comparison of mean ranks for all groups) by comparison the EQR values of each Member State benchmark sites with the BEQI2 (the Netherlands) (white fields) and the NQI (Norway) (grey fields).

NQI ➔ BEQI2 ⬇	B-type 1	D-type 1	DK(NS)-type 1	E-type 2	Fr-type 1	N-type 2	NL-type 1	PT-type 1	ROI-type 2	UK-type 2	UK-type 1
B-type 1		1,000	0,047	1,000	0,000	1,000	0,023	0,260	0,531	1,000	1,000
D-type 1	0,609		0,221	1,000	0,000	1,000	0,123	0,727	1,000	1,000	0,198
DK(NS)-type 1	1,000	1,000		1,000	1,000	1,000	1,000	1,000	1,000	0,001	0,000
E-type 2	1,000	0,000	1,000		0,001	1,000	0,916	1,000	1,000	0,199	0,012
Fr-type 1	0,190	0,000	0,449	1,000		1,000	1,000	1,000	1,000	0,000	0,000
N-type 2	0,324	0,000	0,386	1,000	1,000		1,000	1,000	1,000	1,000	0,124
NL-type 1	1,000	1,000	1,000	0,014	0,000	0,004		1,000	1,000	0,000	0,000
PT-type 1	1,000	1,000	1,000	1,000	1,000	1,000	1,000		1,000	0,085	0,012
ROI-type 2	1,000	1,000	1,000	0,572	0,064	0,071	1,000	1,000		0,065	0,005
UK-type 2	1,000	0,000	1,000	1,000	0,089	0,641	0,166	1,000	1,000		1,000
UK-type 1	0,749	1,000	1,000	0,000	0,000	0,000	1,000	1,000	1,000	0,000	

6) NQI of Norway

The EQR values at the benchmark sites of UK-type 1 are significantly (lower) different from most other benchmark sites by the NQI, except for the Belgian, German, Norwegian and UK-type 2 benchmark sites (Figure 15, Table 28). The French benchmark sites are also significantly different from many other sites (Belgium, Germany, Spain, UK- type 1 and UK-type 2). There are also significant differences between the Belgian and Danish and Dutch benchmark sites with the NQI approach.

7) BAT of Portugal

The EQR values at the benchmark sites of UK-type 1 are significant different from Spain, France, Norway and UK-type 2 benchmark sites by the BAT benthic assessment approach (Figure 16, Table 29). The German benchmark sites are significant different with Spain, France, Norway and UK type 2. Significant difference are also observed between the Dutch and French and Belgian and French benchmark sites and the French and the UK-type 2 sites.

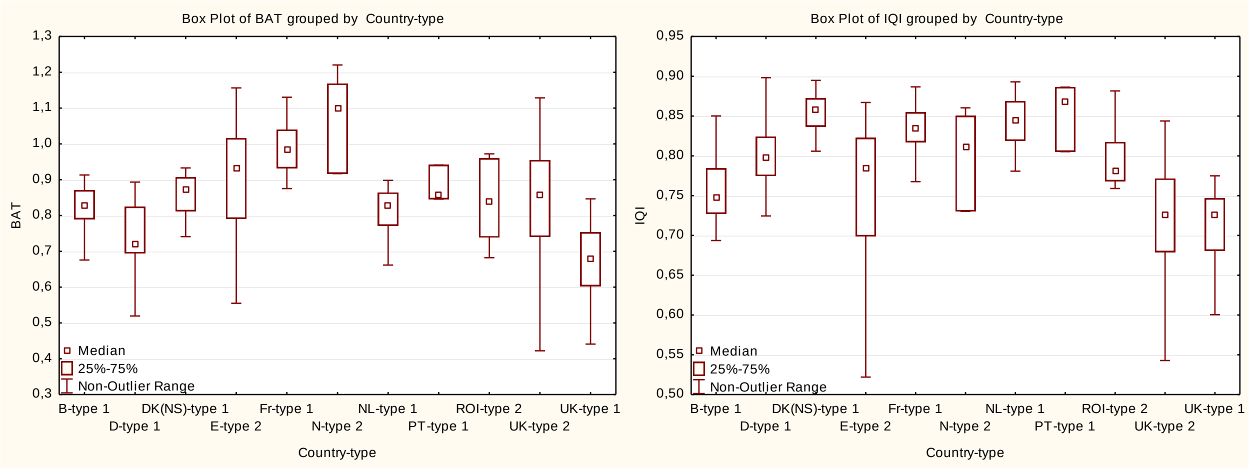


Figure 16. Box-whisker plot (median, percentile values and no-outlier range) of the EQR values at the Member States benchmark sites with the BAT (Portugal) (left) and the IQI (United Kingdom, Ireland) (right).

Table 29. Kruskal-Wallis p levels (multiple comparison of mean ranks for all groups) by comparison the EQR values of each Member State benchmark sites with the BAT (Portugal) (white fields) and the IQI (UK and Ireland) (grey fields).

IQI → BAT ↓	B-type 1	D-type 1	DK(NS)-type 1	E-type 2	Fr-type 1	N-type 2	NL-type 1	PT-type 1	ROI-type 2	UK-type 2	UK-type 1
B-type 1		1,000	0,011	1,000	0,002	1,000	0,001	1,000	1,000	1,000	1,000
D-type 1	1,000		1,000	1,000	1,000	1,000	0,539	1,000	1,000	0,000	0,002
DK(NS)-type 1	1,000	1,000		0,034	1,000	1,000	1,000	1,000	1,000	0,000	0,000
E-type 2	1,000	0,000	1,000		0,004	1,000	0,003	1,000	1,000	0,148	0,394
Fr-type 1	0,004	0,000	1,000	1,000		1,000	1,000	1,000	1,000	0,000	0,000
N-type 2	0,321	0,000	1,000	1,000	1,000		1,000	1,000	1,000	0,622	0,545
NL-type 1	1,000	1,000	1,000	0,791	0,001	0,181		1,000	1,000	0,000	0,000
PT-type 1	1,000	1,000	1,000	1,000	1,000	1,000	1,000		1,000	0,104	0,079
ROI-type 2	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000		0,400	0,342
UK-type 2	1,000	0,013	1,000	0,893	0,000	0,307	1,000	1,000	1,000		1,000
UK-type 1	0,291	1,000	0,189	0,000	0,000	0,000	0,590	1,000	0,755	0,000	

8) IQI of UK/Ireland

The classification of the benchmark sites of the different Member States by the IQI leads also to some significant differences (Figure 16, Table 29). The Danish sites are significantly different from the Belgian, Spanish, UK-type 2 and UK-type 1 sites. The French sites are significantly different from the Belgian, Spanish, Dutch, UK-type 2 and UK-type 1 sites. The Dutch sites are also significant different from the Spanish, UK-type 2 and UK-type 1 sites.

9) BO2A of Spain

From Andalusia region, no benthic data was included in the common dataset. Therefore, no benchmark sites were delimitated for this region of this Member State. The median values of the benchmark sites of the different Member States, evaluated with the BO2A are also different in some cases (Figure 17). The sites of UK-type 2, Spain, France and Norway have lower values than the others.

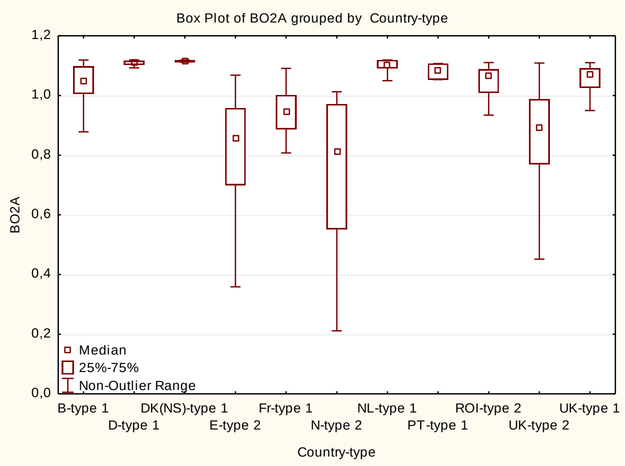


Figure 17. Box-whisker plot (median, percentile values and no-outlier range) of the EQR values at the Member States benchmark sites with the BO2A (Spain, Andalusia region).

12 Annex 3: Comparison of methods and boundaries

12.1 Benchmark sites, based on biotic criteria

12.2 Results of the regression comparison

For all the intercalibration comparisons, the benthic assessment approaches fulfill the criteria ($R^2 > ½ maxR^2$) of the regression comparison (Table 19), except BO2A. The BO2A of Spain shows the lowest correlation with the pseudo-common metric. For the IQI and the NQI, the samples were less equally spread over the linear regression line (dominance in upper part) in comparison to the other approaches, as was the case in the analyses on the theoretical behavior of the benthic assessment approaches.

Table 30. Summary of the correlation coefficient (R^2) of each approach with the common metric for the different intercalibration comparisons. Values outside the criteria were put in red.

Method	Subtraction standardization		Division standardization	
	No sub-region	Sub-region	No sub-region	Sub-region
Denmark	0.9553	0.9549	0.9536	0.9566
UK/ROI	0.8402	0.8692	0.8267	0.8105
Spain (m-AMBI)	0.8864	0.9406	0.8887	0.9261
Norway	0.8965	0.9148	0.8911	0.9141
Portugal	0.9477	0.9671	0.9465	0.9621
The Netherlands	0.7869	0.8762	0.7923	0.8503
Germany	0.9121	0.9569	0.9227	0.9475
France	0.8514	0.9224	0.8542	0.9026
Spain (BO2A)	0.3599	0.4315	0.3546	0.4508

12.3 Comparability criteria

12.3.1 Subtraction benchmark standardization

1) no sub-regions (deep/shallow areas) within NEA 1/26 type.

The boundary bias (<0.25) is too high for the BO2A (Table 31). Denmark, France and Germany are more stringent than the other approaches, especially for the good/moderate boundary. The class differences (<0.5 class) is too high for the BO2A and around criteria level for the DKI.

Table 31. Summary of the boundary bias and class differences analyses following the subtraction benchmark standardization, no discrimination of sub-regions.

Boundary bias	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (AC)
Max	1,000	1,000	1,292	1,000	1,220	1,508	1,342	1,179	1,119
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770	0,830
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530	0,500
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380	0,400
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200	0,200
H/G bias_CW	0,602	-0,149	0,021	-0,043	-0,022	-0,035	0,217	0,374	-1,355
G/M bias_CW	0,351	-0,172	-0,151	0,061	-0,171	0,042	0,288	0,227	-0,906
	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (A)
Absolute Class Difference	0,512	0,469	0,339	0,413	0,342	0,369	0,411	0,402	0,789

2) Sub-regions (deep/shallow areas) within NEA 1/26 type

The boundary bias (<0.25) is in this analysis is too high for BO2A and slightly too high for the m-AMBI (Spain) (Table 32). The DKI, m-AMBI (Germany & France) are more stringent for the good/moderate boundary and the high/good boundary compared to the others. The class difference (<0.5 class) is too high for the BO2A and at criteria level for the DKI. The m-AMBI can meet the criteria by elevating the good/moderate boundary value to 0.56.

Table 32. Summary of the boundary bias and class differences analyses following the subtraction benchmark standardization, with discrimination of the sub-regions.

Boundary bias	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (AC)
Max	1,000	1,000	1,292	1,000	1,220	1,508	1,342	1,179	1,119
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770	0,830
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530	0,500
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380	0,400
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200	0,200
H/G bias_CW	0,546	-0,129	-0,073	0,082	-0,004	0,010	0,275	0,454	-1,310
G/M bias_CW	0,331	-0,132	-0,313	0,156	-0,085	0,158	0,340	0,363	-1,106
	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (A)
Absolute Class Difference	0,510	0,449	0,326	0,394	0,327	0,356	0,389	0,387	0,749

12.3.2 Division benchmark standardization

1) No sub-regions (deep/shallow areas) within NEA 1/26 type

The boundary bias (<0.25) is in this analysis is too high for BO2A (Table 33). The DKI is more stringent for the good/moderate and high/good boundary, in France for the high/good and Germany for the good/moderate boundary. The class difference (<0.5 class) is too high for the BO2A approach and at criteria level for the DKI.

Table 33. Summary of the boundary bias and class differences analyses following the division benchmark standardization, no discrimination of the sub-regions.

Boundary bias	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (AC)
Max	1,000	1,000	1,292	1,000	1,220	1,508	1,342	1,179	1,119
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770	0,830
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530	0,500
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380	0,400
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200	0,200
H/G bias_CW	0,585	-0,143	0,016	-0,043	-0,033	-0,032	0,204	0,371	-1,242
G/M bias_CW	0,339	-0,158	-0,156	0,059	-0,186	0,046	0,282	0,222	-0,836
	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (A)
Absolute Class Difference	0,512	0,469	0,339	0,413	0,342	0,369	0,411	0,402	0,789

2) Sub-regions (deep/shallow areas) within NEA 1/26 type

The boundary bias (<0.25) is in this analysis is too high for BO2A and slightly too high for the m-AMBI (Spain). The bias for DKI, Germany and France is more stringent for the good/moderate and high/good boundary. The class difference (<0.5 class) is too high for the BO2A approach and at criteria level for the DKI. The m-AMBI can meet the criteria by elevating the good/moderate boundary value to 0.55.

Table 34. Summary of the boundary bias and class differences analyses following the division benchmark standardization, with discrimination of the sub-regions.

Boundary bias	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (AC)
Max	1,000	1,000	1,292	1,000	1,220	1,508	1,342	1,179	1,119
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770	0,830
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530	0,500
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380	0,400
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200	0,200
H/G bias_CW	0,539	-0,210	-0,069	0,066	-0,008	0,000	0,263	0,448	-1,367
G/M bias_CW	0,327	-0,209	-0,290	0,137	-0,131	0,081	0,305	0,278	-0,978
	Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	Netherlands	Germany	France	Spain (A)
Absolute Class Difference	0,510	0,449	0,326	0,394	0,327	0,356	0,389	0,387	0,749

12.3.3 Conclusion

The intercalibration of the benthic assessment approaches within the NEA-GIG region can be executed following the intercalibration guidelines. As shown in the analysis, the benthic assessment approaches are very comparable (some after a small adaptation of their boundaries) and meet the intercalibration criteria, except for the BO2A. The subtraction and division standardization delivers the same results regarding the acceptability of the criteria. The subtraction standardization only delivers slightly higher values compared to the division standardization

The BO2A does not meet the criteria of boundary bias and class difference in any intercalibration comparison option. The adaptation of the boundaries to meet the criteria is rather impossible for this approach, because the tests to change the boundary levels of the BO2A do not lead to any situation that meets the criteria. They even influenced the criteria levels of the other approaches, mostly in a negative way. The application of the BO2A on this common NEA-GIG dataset seems to be more complicated and different from the results of the own intercalibration analyses of the Andalusia region (see separate document (2011-12-16technical_report_NEA_CW_invertebrates_ES(AN)_Dec2011)).

The DKI and m-AMBI (Germany & France) show in all intercalibration comparison options a more stringent evaluation than the other approaches. Therefore, those boundary values can even be lowered to be more comparable with the other methods, but this is not required.

The m-AMBI (Spain) shows in the intercalibration comparison option, where sub-regions are distinguished, a slightly too high boundary bias. This approach scored not in correspondence with the other approaches (IQI, NQI) for samples typical for less shallow areas. This approach can easily meet the criteria, as the good/moderate boundary is slightly increased (+0.02 or 0.03).

All other benthic assessment approaches (BAT, BEQI2, IQI) meet the comparability criteria.

Based on the analyses and the experience with the data and the assessment approaches, the intercalibration comparison with the division benchmark standardization and no discrimination of the sub-regions should be most appropriate. This because, the approaches show similar trend lines, but there are differences between them along the pressure gradient (some of them vanish). Besides this, there was no hard evidence to discriminate sub-regions, and the reference settings for these soft sediment habitats are similarly set by the Member States for this type.

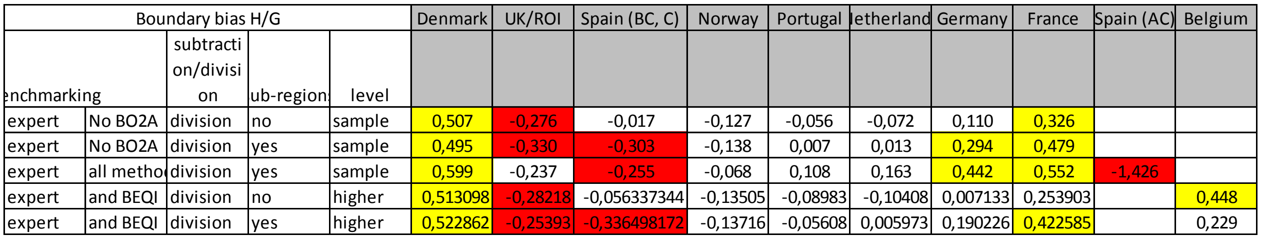
12.4 Benchmark selection based on expert judgment

The comparison is executed based on certain conditions, but the selection of those conditions has its effect on the boundary bias values. In section 12.3, the results of the biotic benchmark are shown, which reveals no fail in the boundary bias criteria for IQI, whereas on expert judgment it does (Table 35). Also the inclusion or exclusion of sub-regions, regardless of the benchmarking, has an effect on the boundary bias, especially for Spain. When no sub-region is recognized, no boundary harmonization is necessary, whereas this is necessary when sub-regions are recognized. This could be related to inappropriate reference values for this sub-region type in Spain, but this seems not to be the case (see below).

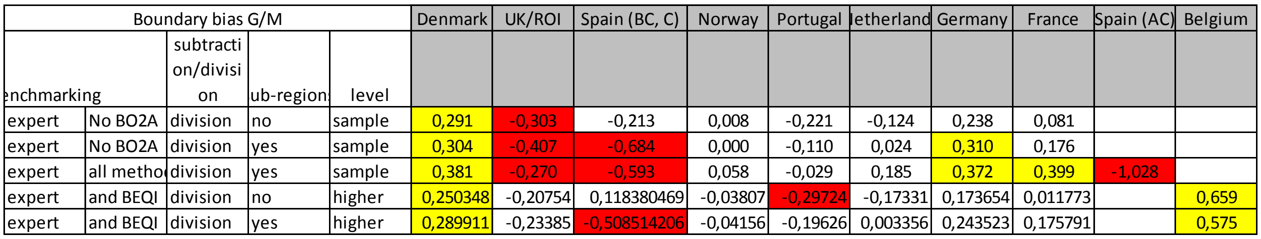
Further, the inclusion or exclusion of a method has its consequence on the boundary bias values, which became slightly lower. This happens because the BO2A assessment approach is not comparable with the others. This aspect is worth mentioning, because adding or changing a method has consequences on the obtained comparability results.

This were all intermediate comparability analyses to explore the intercalibration and to move towards the selection of the comparison most in line with the intercalibration guidelines.

Table 35. Summary of the boundary bias for H/G and G/M following different conditions regarding discrimination of sub-region or not or including/excluding certain methods.



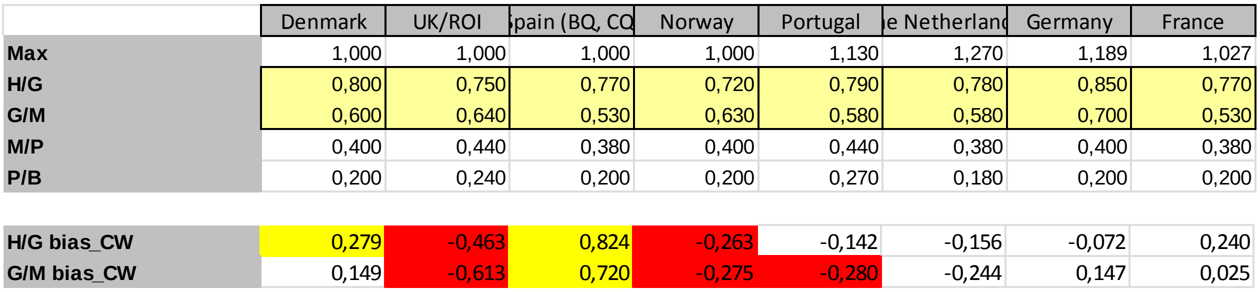
Boundary bias H/G					Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	letherland	Germany	France	Spain (AC)	Belgium
enchmarking		subtraction/division	ub-region	level										
expert	No BO2A	division	no	sample	0,507	-0,276	-0,017	-0,127	-0,056	-0,072	0,110	0,326		
expert	No BO2A	division	yes	sample	0,495	-0,330	-0,303	-0,138	0,007	0,013	0,294	0,479		
expert	all metho	division	yes	sample	0,599	-0,237	-0,255	-0,068	0,108	0,163	0,442	0,552	-1,426	
expert	and BEQI	division	no	higher	0,513098	-0,28218	-0,056337344	-0,13505	-0,08983	-0,10408	0,007133	0,253903		0,448
expert	and BEQI	division	yes	higher	0,522862	-0,25393	-0,336498172	-0,13716	-0,05608	0,005973	0,190226	0,422585		0,229



Boundary bias G/M					Denmark	UK/ROI	Spain (BC, C)	Norway	Portugal	letherland	Germany	France	Spain (AC)	Belgium
enchmarking		subtraction/division	ub-region	level										
expert	No BO2A	division	no	sample	0,291	-0,303	-0,213	0,008	-0,221	-0,124	0,238	0,081		
expert	No BO2A	division	yes	sample	0,304	-0,407	-0,684	0,000	-0,110	0,024	0,310	0,176		
expert	all metho	division	yes	sample	0,381	-0,270	-0,593	0,058	-0,029	0,185	0,372	0,399	-1,028	
expert	and BEQI	division	no	higher	0,250348	-0,20754	0,118380469	-0,03807	-0,29724	-0,17331	0,173654	0,011773		0,659
expert	and BEQI	division	yes	higher	0,289911	-0,23385	-0,508514206	-0,04156	-0,19626	0,003356	0,243523	0,175791		0,575

Test for changing the reference values of Spain for sub-type 2.

If the m-AMBI reference values for the deeper samples (AMBI: 1, Diversity: 5.7, Richness: 130) are applied, it seems that they were too high. This is because there is no sample in high status for this sub-type in the common dataset, which is not true (some stations has no pressures, such as Norway). Therefore, Spain became too stringent in their assessment, whereas the other countries of type 2 does not meet the boundary bias criteria at all. Spain will thus therefore accept the boundary harmonisation (0.63 for G/M).



	Denmark	UK/ROI	pain (BQ, CQ	Norway	Portugal	e Netherland	Germany	France
Max	1,000	1,000	1,000	1,000	1,130	1,270	1,189	1,027
H/G	0,800	0,750	0,770	0,720	0,790	0,780	0,850	0,770
G/M	0,600	0,640	0,530	0,630	0,580	0,580	0,700	0,530
M/P	0,400	0,440	0,380	0,400	0,440	0,380	0,400	0,380
P/B	0,200	0,240	0,200	0,200	0,270	0,180	0,200	0,200
H/G bias_CW	0,279	-0,463	0,824	-0,263	-0,142	-0,156	-0,072	0,240
G/M bias_CW	0,149	-0,613	0,720	-0,275	-0,280	-0,244	0,147	0,025

Contact

Gert Van Hoey, Wetenschappelijk onderzoeker
Instituut voor Landbouw- en Visserijonderzoek
Dier
Ankerstraat 1
8400 Oostende
T +32 59 56 98 47
gert.vanhoey@ilvo.vlaanderen.be

Deze publicatie kan ook geraadpleegd worden op:
www.ilvo.vlaanderen.be/pers en media/ILVO mededelingen

Te citeren als: Van Hoey Gert, Bonnen Wendy, Fuensanta Salas Herrero, 2015. Intercalibration report for benthic invertebrate fauna of the North East Atlantic Geographical intercalibration group for Coastal Waters (NEA 1/26). ILVO-mededeling 191. 79pp

ILVO

Aansprakelijkheidsbeperking

Deze publicatie werd door ILVO met de meeste zorg en nauwkeurigheid opgesteld. Er wordt evenwel geen enkele garantie gegeven omtrent de juistheid of de volledigheid van de informatie in deze publicatie. De gebruiker van deze publicatie ziet af van elke klacht tegen ILVO of zijn ambtenaren, van welke aard ook, met betrekking tot het gebruik van de via deze publicatie beschikbaar gestelde informatie.
In geen geval zal ILVO of zijn ambtenaren aansprakelijk gesteld kunnen worden voor eventuele nadelige gevolgen die voortvloeien uit het gebruik van de via deze publicatie beschikbaar gestelde informatie.

ILVO

Instituut voor Landbouw- en Visserijonderzoek
Burg. Van Gansberghelaan 92
9820 Merelbeke - België

T +32 9 272 25 00
ilvo@ilvo.vlaanderen.be
www.ilvo.vlaanderen.be