Workshop to scope assessment methods to set thresholds (WKBENTH2)

Artigas, Miquel Canals; Baldrighi, Elisa ; Belin, Alice; Bell, James ; Bendraoui, Abdeladim ; Beukhof, Esther D.; Blomqvist, Mats; Boyé, Aurélien; Di Lorenzo, Blanca; Di Bona, Gabriele

Total number of authors:
60

Link to article, DOI:
10.17895/ices.pub.20731537

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Artigas, M. C., Baldrighi, E., Belin, A., Bell, J., Bendraoui, A., Beukhof, E. D., Blomqvist, M., Boyé, A., Di Lorenzo, Bl., Di Bona, G., Dinesen, G. E., Downie, A., Drgas, A., Duncombe-Smith, S., Fernández, U., Gavazzi, G. M., Gutierrez, L., Hansen, F., Haubner, N., ... Wijnhoven, S. (2022). *Workshop to scope assessment methods to set thresholds (WKBENTH2)*. International Council for the Exploration of the Sea (ICES). ICES Scientific Report Vol. 4 No. 70 https://doi.org/10.17895/ices.pub.20731537

WORKSHOP TO SCOPE ASSESSMENT METHODS TO SET THRESHOLDS (WKBENTH2)

VOLUME 4 | ISSUE 70

ICES SCIENTIFIC REPORTS

RAPPORTS
SCIENTIFIQUES DU CIEM

ICES INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CIEM COUNSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H.C. Andersens Boulevard 44-46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

ISSN number: 2618-1371

This document has been produced under the auspices of an ICES Expert Group or Committee. The contents therein do not necessarily represent the view of the Council.

ICES Scientific Reports

Volume 4 | Issue 70

WORKSHOP TO SCOPE ASSESSMENT METHODS TO SET THRESHOLDS (WKBENTH2)

Recommended format for purpose of citation:

ICES. 2022. Workshop to scope assessment methods to set thresholds (WKBENTH2). ICES Scientific Reports. 4:70. 99 pp. http://doi.org/10.17895/ices.pub.20731537

Editors

Jan Geert Hiddink • David Reid • Daniel van Denderen

Authors

Miquel Canals Artigas • Elisa Baldrighi • Alice Belin • James Bell • Abdeladim Bendraoui • Esther Beukhof • Mats Blomqvist • Aurélien Boyé • Bianca di • Bianca di • Gabriele Di Bona • Grete Elisabeth Dinesen • Anna Downie • Aleksander Drgas • Stephen Duncombe-Smith • Ulla Fernández • Giacomo Montereale Gavazzi • Lina Gutierrez • Flemming Hansen • Norbert Haubner • Cristina Herbon • Jan Geert Hiddink • José Manuel González Irusta • Axel Kreutle • Despina Kyriakoudi • Ellen L. • Pascal Laffargue • Anna Luff • Tim Mackie • Silvia Maltese • Liam Matear • Marco Milardi • Alessandra Nguyen • Antonia Nystrom • Hatice Onay • Nadia Papadopoulou • Marina Penna • Andrea Pierucci • Maider Plaza • Marina Pulcini • Elisa Punzo • Saša Raicevich • David Reid • Sofia Reizopoulou • Giada Riva • Marie-Julie Roux • Owen Rowe • Marta mega Rufino • Angella Santelli • Hannah Schartmann • Petra Schmitt • Alexander Schröder • Marija Sciberras • Chris Smith • Murray Thompson • Sebastian Valanko • Daniel van Denderen • Karin van der Reijden • Gert Van Hoey • Sandrine Vaz • Sander Wijnhoven

ICES CIEM International Council for the Exploration of the Sea Conseil International pour l'Exploration de la Mer

Contents

i Executive summary

The Marine Strategy Framework Directive (MSFD) requires Member States to achieve good environmental status (GES) across their marine waters. The EU have requested ICES to advise on methods for assessing adverse effects on seabed habitats, through selection of relevant indicators for the assessment of benthic habitats and seafloor integrity and associated threshold values for GES in relation to Descriptor 6 – Seabed integrity under the MFSD.

Two sets of criteria were developed to evaluate indicators and thresholds respectively for evaluation of suitability for assessing GES. 16 indicator and 12 threshold criteria were compiled and weighted by importance. The criteria were designed for evaluation at a subregional or regional level. The scoring for these criteria is meant as a guidance when choosing indicators and thresholds, so failure to meet one criterion will not necessarily prevent the use of the indicator or threshold in an assessment. The framework was evaluated for 6 indicators and for 11 methods for setting thresholds. The criteria were found to be useful for evaluation both indicators and thresholds. The process works most consistently when there are experts in the group on both the criteria themselves and on the indicators and thresholds.

The MFSD Descriptor 6 determination of GES needs both a quality threshold (when are seabed habitats in a good state in a specific location) and an extent threshold (proportion of the assessment area that needs to have seabed habitats in good state). Eleven different methods for setting thresholds were identified, of which more are suitable for setting quality than for extent thresholds. Preferred methods identified an ecologically-motivated difference between a good and degraded state, rather than another transition. Quality thresholds based on the lower boundary of the range of natural variation were considered most promising. This approach can be used for most, but not all, indicators.

The WK collated a standardized dataset to test the specificity, sensitivity and/or responsiveness of sampling-based benthic indicators to pressure gradients for evaluation by WKBENTH3. Risk-based methods will be evaluated as maps and by scored sensitivity and impact score per MSFD habitat type and subdivision. Participants provided input into the selection of indicators for the compilation of indicators. A template was developed for documenting the characteristics of each indicator to facilitate the evaluation of the indicators.

ii Expert group information

Expert group name	**Workshop on assessment methods to set threshold and assess adverse effects on seabed habitats (WKBENTH2)**
Expert group cycle	Annual
Year cycle started	2022
Reporting year in cycle	1/1
Chairs	Jan-Geert Hiddink, UK
	David Reid, Ireland
	Daniel van Denderen, USA
Meeting venues and dates	24-26 May 2022, Copenhagen, Denmark (number of participants)
	8-9 June 2022, Copenhagen, Denmark (number of participants)

1 Introduction

Countries and Regional Sea Conventions are developing indicators of pressure and impact on benthic habitats, including from bottom-trawl fisheries (Figure 1.1). Such indicators are developed to support status assessments for the Marine Strategy Framework Directive (MSFD) and underpin the management needed to ensure that biodiversity, structure and function of benthic ecosystems are safeguarded, and fisheries production is sustained.

The EU (DG ENV) have requested ICES to "*advise on methods for assessing adverse effects on seabed habitats*". This workshop, WKBENTH2, is the first of two workshops, to review and develop the required technical work. Following a peer-review of the WKBENTH2 report and technical service, another workshop will be convened (WKBENTH3) to evaluate the proposed assessment methods and thresholds using agreed upon criteria, methods and analysis of their performance. Based on peer-review of this work, formal advice will be prepared by ICES Advisory Committee (ACOM) to be published as ICES Advice and delivered to the EU by December 2022.

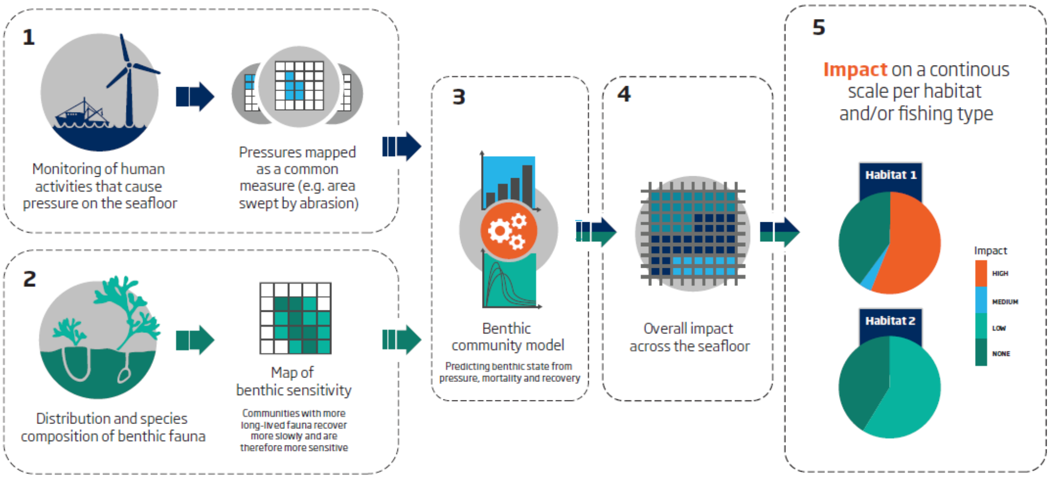

1
Monitoring of human activities that cause pressure on the seafloor
Pressures mapped as a common measure (e.g. area swept by abrasion)
2
Distribution and species composition of benthic fauna
Map of benthic sensitivity
Communities with more long-lived fauna recover more slowly and are therefore more sensitive
3
Benthic community model
Predicting benthic state from pressure, mortality and recovery
4
Overall impact across the seafloor
5
Impact on a continous scale per habitat and/or fishing type
Habitat 1
Habitat 2
Impact
HIGH
MEDIUM
LOW
NONE

Figure 1.1 Evaluating seafloor impact and benthic habitats that are at greatest risk from human activities disturbing the seafloor.

Structure of the workshop and report

The technical workshop was conducted as two hybrid meetings, each of 3 consecutive days. The work was organized around plenary sessions and breakout groups. WKBENTH2 had 64 participants, with an average of 25 active participants during any one day. Participants represented 40 different countries and included all EU waters (Iberian Coast, Celtic Sea, Bay of Biscay, North Sea, Baltic Sea, Mediterranean and the Black Sea). Benthic and policy experts from EU-funded projects, Regional Seas Conventions, and academia participated.

The structure of this report follows the four workshop resolutions (annex 2):

- Chapter 3 - Establish a set of criteria that can be used to evaluate suitability of regional indicators/assessment methods to assess adverse effects on seabed habitats for MSFD purposes (ToR A)

- Chapter 4 - Review methods and criteria to set thresholds adverse effects on seabed habitats, and suggest operational options that can be illustrated using worked examples (ToR B)

- Chapter 5 - Suggest quantitative and qualitative ways to evaluate and compare suitability and performance of indicators/assessment methods (ToR C)

- Chapter 6 - Provide input to a draft compilation of regional indicators/assessment methods to set threshold and assess adverse effects on seabed habitats (ToR D)

2 Policy context

Marine Strategy Framework Directive

The Marine Strategy Framework Directive (MSFD, 2008/56/EC) requires Member States to achieve and maintain good environmental status (GES) across their marine waters in relation to the eleven 'descriptors' set out in MSFD Annex I. Descriptor 1 (benthic habitats) and Descriptor 6 (sea-floor integrity) are the main descriptors for assessing the state of the seabed, while other descriptors address particular pressures and impacts on the seabed (e.g. D2 – non-indigenous species, D5 – eutrophication).

Commission Decision (EU) 2017/848 (the 'GES Decision') sets out criteria and methodological standards for determining GES and assessing the extent to which it has been achieved. It defines that benthic habitats (D1) and sea-floor integrity (D6) are to be addressed together via the assessment of 22 benthic 'broad habitat types' (BHTs) and at the scale of biogeographically relevant 'subdivisions' of each MSFD region or subregion. If wanted EU Member States can add so called other habitat types (OHTs) to their assessments. OHTs were not part of the workshop discussions.

The GES Decision sets out the following criteria for benthic habitats:

i. D6C1 Physical loss
ii. D6C2 Physical disturbance
iii. D6C3 Adverse effects of physical disturbance on habitats
iv. D6C4 Extent of habitat loss
v. D6C5 Extent of adverse effects on the condition of a habitat

As stated in the revised Art.8 guidance document, "The overall status is represented by the assessment of D6C5 per BHT, including the assessment of D6C3 and D6C4. GES of the BHT is achieved when these criteria have met the respective threshold values (extent threshold for D6C4, and quality and extent thresholds for D6C5). The extent of adverse effects from disturbance (D6C3) and the state (impact) assessment, and inputs from other descriptors (either as spatial impact analysis or qualitative description, as deemed appropriate) contribute to D6C5".

Biogeographically relevant 'subdivisions' for assessment were not fully defined by Member States in their 2018 MSFD Article 8 reports, although aspects of the assessments (e.g., for assessing D6C3) were done at subdivision level in the North-east Atlantic region (OSPAR Intermediate Assessment 2017). In the absence of subdivisions agreed by Member States, the ICES 2021 advice used an indicative set of 22 subdivisions covering the four regions/subregions addressed by the advice. Further work is needed by Member States to define operational subdivisions for these and other (sub)regions for use in the next (2024) updates of the MSFD Article 8 assessments; this definition should be achieved through regional cooperation, including via the preparation of quality status reports by the regional sea conventions.

The quality and extent threshold values and the method for assessing overall status of a habitat (integration of criteria D6C4 and D6C5) are being established through a Union-level process, considering regional or sub-regional specificities. The process will be overseen by the MSFD Common Implementation Strategy (CIS), particularly the Technical Group on seabed habitats and sea-floor integrity (TG Seabed), the Working Group on Good Environmental Status (WG GES) and the Marine Strategy Coordination Group (MSCG). TG Seabed was established in 2018 to develop a framework for assessing seabed habitats and to propose quality and extent threshold values for criteria D6C4 and D6C5 of Commission Decision (EU) 2017/848. The work of TG Seabed has been supported by ICES, through a number of ICES Advice documents.

Methods for assessing adverse effects on seabed habitats

In 2017, ICES held a series of workshops (WKBENTH, WKSTAKE, and WKTRADE) to address an advice request from the European Commission to evaluate indicators for assessing pressure and impact on the seafloor from one human pressure – mobile bottom-contacting fishing - and demonstrate trade-offs in catch/value of landings relative to impacts and recovery potential of the seafloor. Methods for assessing seafloor impact from mobile bottom-contacting fishing gears were developed and ICES advised on a set of indicators for assessing pressure and impact. These indicators were selected based on their ability to describe impacts on a continuous scale, because they could be used in the evaluation of trade-off between the fisheries and their impacts on the seabed, and applied to large sea areas and thus be suitable for MSFD assessment purposes.

Using some of the propose assessment methods, ICES ran a series of workshops (WKTRADE3) in 2021 to evaluate the suite of management options prioritized by stakeholders for different EU marine regions and analyse their consequences for the overall benefit to seabed habitats and loss of fisheries values.

HELCOM and OSPAR have also been developing indicators for assessing seabed habitats. These are based on models, quality sensitivity classification or impact-risk classification (e.g., BH3, CumI), empirical data of infauna (e.g., grab samples) and/or epifauna (e.g., bottom trawl hauls, under water images) (e.g., BH1-SoS, BH2, other benthic indicators) from different data sources (e.g., MEDITS, DATRAS) or drawing upon Water Framework Directive (WFD) and the Habitats Directive (HD) assessments (e.g., BQIs, other benthic indicators). First assessments were produced in 2017 (OSPAR Intermediate Assessment) and 2018 (HELCOM HOLAS II); further indicator development is underway within both RSCs for use in their forthcoming quality status assessments.

In addition to using RSC indicators and assessment results for 2018 MSFD reporting, Member States have used a variety of national indicators to assess GES of seabed habitats. Many of these are Water Framework Directive (WFD) and/or Habitats Directive (HD) indicators, applicable to coastal waters and focused on assessing nutrient and organic matter enrichment (i.e., eutrophication) or modified from WFD indicators for use beyond coastal waters. However, some Member States have also used or funded the development of other alternative D6 indicators (e.g., GBPI, mTDI, pTDI, AMBI, BENTIX; Labrune *et al*, 2021, Jac *et al*., 2020).

TG Seabed has prepared reports that outline the principles for assessing adverse effects on seabed habitats (MSCG_29-2021-05) and the standards that need to be met by indicators (SEABED_7-2021-10). Preparation of these reports has revealed the complexity of assessing adverse effects on seabed habitats, due to the wide range of seabed habitat types, to the variation in response to different pressures (e.g., physical, biological, chemical) and to the wide variety of available indicators. Consequently, further advice from ICES is sought on detailed aspects of assessment methodologies (i.e., for risk and state indicators), how the indicators respond to different pressures, how habitat quality should be assessed through use of threshold values, and how well the assessment results correspond between the different indicators. ICES will take account of TG Seabed reports and existing relevant literature when preparing its advice.

Lessons learnt from the Water Framework Directive intercalibration

The intercalibration process in the Water Framework Directive (WFD) was aimed at ensuring comparability of the classification results of the WFD assessment methods developed by the Member States for the biological quality elements. The essence of intercalibration is to ensure that the high-good and the good-moderate boundaries in all Member States' assessment methods correspond to comparable levels of ecosystem alteration. Therefore, the intercalibration exercise must establish values for the boundary between the classes of high and good status, and for the boundary between good and moderate status, which are consistent with the normative definitions of those class boundaries given in Annex V of the WFD. In the frame of the intercalibration exercise also the compliance of Member States assessment methods with the provisions of the Directive are checked.

To improve this process, an intercalibration procedure was developed defining more clearly the individual intercalibration steps and introducing several checking criteria. This was done in two steps, with the first guidance document published in 2005 for the first phase of the intercalibration (2004-2006) (CIS Guidance Document No. 14), showing several gaps and uncertainties as to the comparability of results. Therefore, an updated guidance was produced during the second phase of the intercalibration (2008-2011).

The technical intercalibration process consist of different steps and options and is presented in the flow chart (Figure 2.1). The questions that are asked in the flow chart serve the purpose of performing four basic checks for the identified necessary steps of the intercalibration exercise:

- Preconditions check: Check the compliance of national assessment methods with the WFD requirements with the help of WFD compliance criteria;

- Intercalibration feasibility check: Screening of Member States' assessment methods for acceptance in the current intercalibration exercise with the help of method acceptance criteria;

- Data set check: Evaluation of Member States' datasets for inclusion in common dataset / boundary calculations with the help of data acceptance criteria;

- Comparison of boundaries: Assess level of agreement of boundaries with the help of comparability criteria.

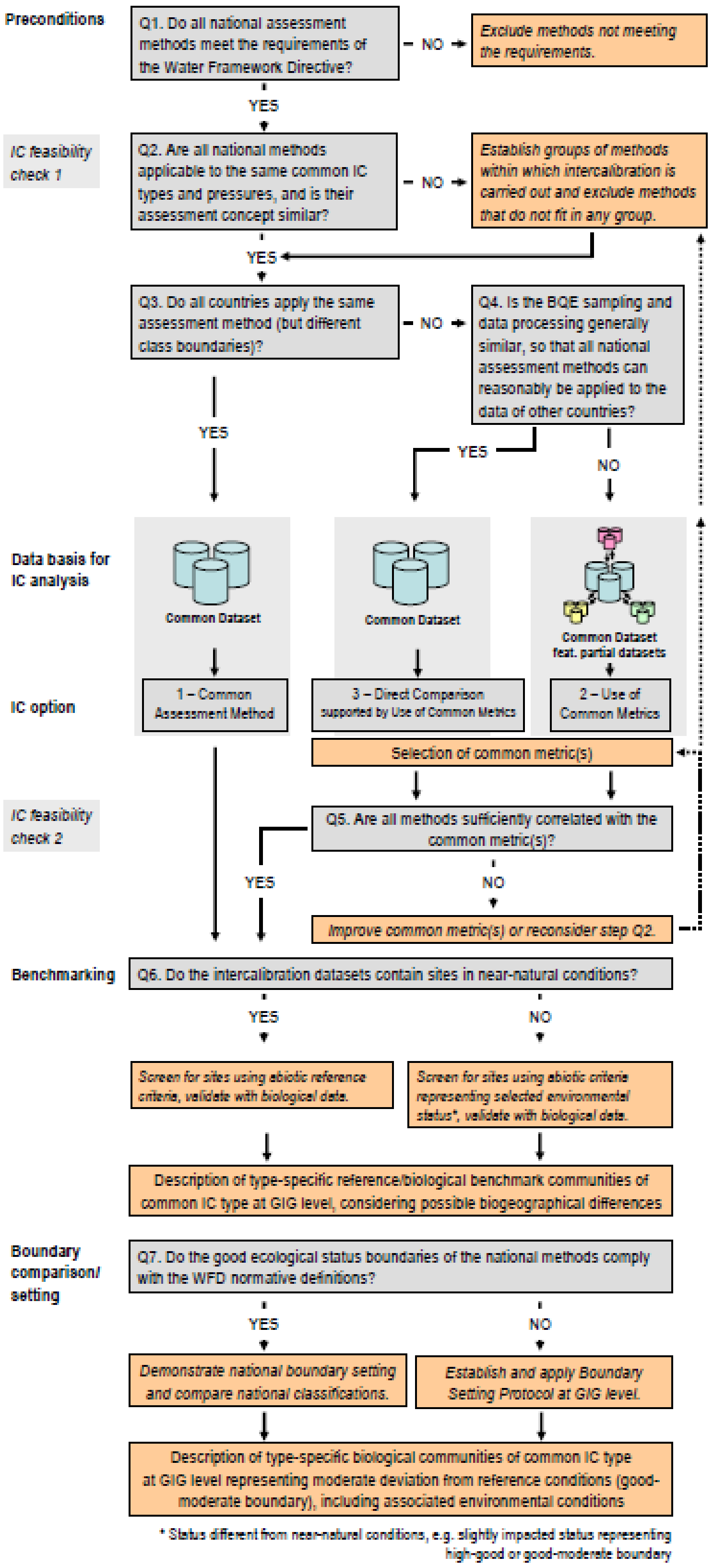

Preconditions
Q1. Do all national assessment methods meet the requirements of the Water Framework Directive?
NO
Exclude methods not meeting the requirements.
YES
IC feasibility check 1
Q2. Are all national methods applicable to the same common IC types and pressures, and is their assessment concept similar?
NO
Establish groups of methods within which intercalibration is carried out and exclude methods that do not fit in any group.
YES
Q3. Do all countries apply the same assessment method (but different class boundaries)?
NO
Q4. Is the BQE sampling and data processing generally similar, so that all national assessment methods can reasonably be applied to the data of other countries?
YES
YES
NO
Data basis for IC analysis
Common Dataset
Common Dataset
Common Dataset feat. partial datasets
IC option
1 – Common Assessment Method
3 – Direct Comparison supported by Use of Common Metrics
2 – Use of Common Metrics
Selection of common metric(s)
IC feasibility check 2
Q5. Are all methods sufficiently correlated with the common metric(s)?
YES
NO
Improve common metric(s) or reconsider step Q2.
Benchmarking
Q6. Do the intercalibration datasets contain sites in near-natural conditions?
YES
NO
Screen for sites using abiotic reference criteria, validate with biological data.
Screen for sites using abiotic criteria representing selected environmental status*, validate with biological data.
Description of type-specific reference/biological benchmark communities of common IC type at GIG level, considering possible biogeographical differences
Boundary comparison/ setting
Q7. Do the good ecological status boundaries of the national methods comply with the WFD normative definitions?
YES
NO
Demonstrate national boundary setting and compare national classifications.
Establish and apply Boundary Setting Protocol at GIG level.
Description of type-specific biological communities of common IC type at GIG level representing moderate deviation from reference conditions (good-moderate boundary), including associated environmental conditions
* Status different from near-natural conditions, e.g. slightly impacted status representing high-good or good-moderate boundary

Figure 2.1. Flow chart of the WFD intercalibration process.

WFD and MSFD

Some of the aspects of the WFD intercalibration guidance are potentially relevant to be used for checking comparability between the MSFD benthic assessment methods. Although the MSFD does not strive to have a WFD like intercalibration process, it is worthwhile to look for synergies and some of these are listed here:

- The MSFD is less strict in defining the indicator requirements compared to the WFD (Van Hoey *et al*., 2010) and a compliance check is not necessary. It is nevertheless advisable to outline for each MSFD assessment method how it concretely aligns to the MSFD descriptor criteria.
- The WFD intercalibration feasibility check is of relevance for the MSFD, as comparison of dissimilar methods ("apples and pears") has clearly to be avoided. In the WFD, the intercalibration exercise was focused on specific type / biological quality element / pressure combinations. The second step of the process introduced an "IC feasibility check" to restrict the actual intercalibration analysis to methods that address the same common type(s) and anthropogenic pressure(s), and follow a similar assessment concept. Different community characteristics - structural, functional or physiological - can be used in assessment methods which can render their comparison problematic. For example, biodiversity indices may give a different view on structural characteristics of the community compared to species composition indices. In several cases, the concept of the method required more specific typology issues to be taken into account to ensure comparability of results, e.g., the habitat typology for marine benthic fauna. This aspect accounts also for the evaluation of the MSFD assessment methods.
- The WFD intercalibration was rather complex (e.g. (pseudo)-common metrics; harmonization of reference conditions) and technical (boundary bias and class agreement). Whether such detailed comparability tests are needed for MSFD purposes need to be discussed, but some of the principles might be useful to evaluate MSFD assessment method comparability. WFD comparability was based on three different options depending on how comparable the approaches of the national methods were:

  Option 1 - same data acquisition and same numerical evaluation: Member States are using a common assessment method and intercalibration process can concentrate on the harmonisation of reference conditions and class boundary comparison/setting;

  Option 2 - different data acquisition and different numerical evaluation: requires the development of common metrics for intercalibration;

  Option 3 - similar data acquisition but different numerical evaluation: necessitates direct comparisons in which the pairwise differences of national assessment results are investigated. Common metrics are highly recommended as a supporting approach to evaluate the influences of biogeographical differences, the definition of reference conditions and the actual boundary setting.

  Comparability was always checked through the analysis of two components: boundary bias and class agreement (Figure 2.2). Sufficient comparability is reached when acceptability criteria on boundary bias are met, and class agreement has been checked.

Boundary bias is the deviation in the relative positioning of class boundaries and measured by the magnitude and direction of deviation by a class boundary of one national method relative to the common view of the Member States (i.e. defined by the common metric or by the global mean of all the methods = pseudo-common metric, for the high and good, and good and moderate status class boundary). This deviation is expressed in class equivalents. It reflects the level of ambition of different methods or how stringent Member States are in defining the good ecological status.

The value to meet the boundary bias criteria is that the different national boundaries should not differ more than 0.5 class from each other (= the maximum boundary deviation above or below each national boundary is a quarter of a class). If this is not the case, the member states need to adapt their boundaries until they are in line.

Class agreement is the confidence that two or more national methods will report the same class for a given site, as calculated by the average absolute class difference between all pairs of ecological quality ratio values across all participating Member States, the proportion of classifications differing by an agreed amount (half a class), and the multirater kappa coefficient. Class agreement depends a lot on how closely the methods are related.

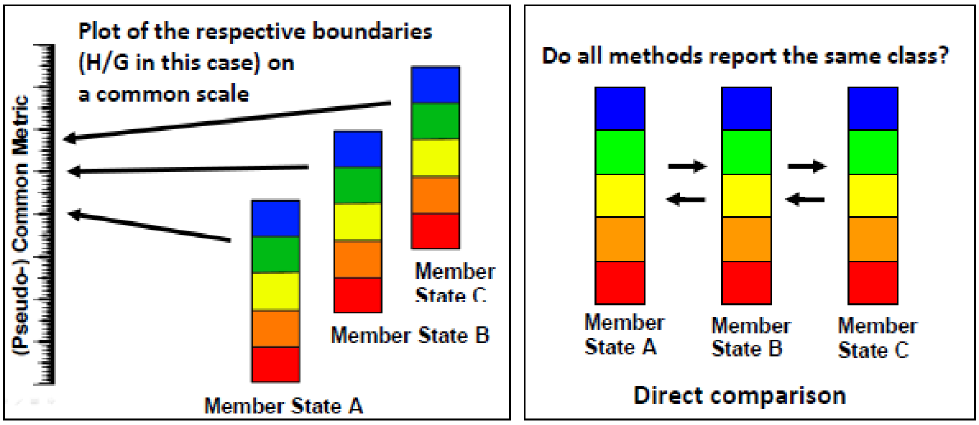


Figure 2.2. Illustrations of boundary bias (left) and class agreement (right) analyses.

WFD intercalibration example for benthic indicators in coastal Northeast Atlantic waters

The intercalibration of coastal waters in the Northeast Atlantic Geographical Intercalibration Group (NEA-GIG) 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 executed. 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). 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 was executed. In this phase, the benthic assessment approaches of nine European Member States (Belgium, Germany, Denmark, France, Ireland, the Netherlands, Portugal, Spain and the United Kingdom) and Norway were intercalibrated. The report Van Hoey *et al.* (2015) compiles all the latest information regarding the benthic WFD 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 biogeographical differences in the common dataset, perform an alternative benchmark delineation and the comparability analyses following the intercalibration guidelines. The result was that all benthic assessment approaches, except Benthic Opportunist Annelids Amphipods (BO2A) index, were finally meeting the comparability criteria of the intercalibration guidance, after raising the good/moderate boundary of Spain (m-AMBI) to a higher value.

References

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.

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

CIS Guidance Document No. 14: Common implementation strategy for the water framework directive (2000/60/EC); Technical Report - 2011 – 045

EU 2017/848 Commission decision (EU) 2017/848 of 17 May 2017 laying down criteria and methodological standards on good environmental status of marine waters and specifications and standardised methods for monitoring and assessment, and repealing Decision 2010/477/EU https://mcc.jrc.ec.europa.eu/documents/ComDec/Com_dec_GES_2017_848_EU.pdf

ICES. 2021. ICES advice to the EU on how management scenarios to reduce mobile bottom fishing disturbance on seafloor habitats affect fisheries landing and value. *In* Report of the ICES Advisory Committee, 2021. ICES Advice 2021. sr.2021.08. https://doi.org/10.17895/ices.advice.8191.

Jac, C., Desroy, N., Certain, G., Foveau, A., Labrune, C. and Vaz, S., 2020. Detecting adverse effect on seabed integrity. Part 2: How much of seabed habitats are left in good environmental status by fisheries?. Ecological Indicators, 117, p.106617

Labrune, C., Gauthier, O., Conde, A., Grall, J., Blomqvist, M., Bernard, G., Gallon, R., Dannheim, J., Van Hoey, G. and Grémare, A., 2021. A General-Purpose Biotic Index to Measure Changes in Benthic Habitat Quality across Several Pressure Gradients. Journal of Marine Science and Engineering, 9(6), p.654.

MSCG_29-2021-05 Marine Strategy Framework Directive (MSFD) Common Implementation Strategy 29th Meeting of the Marine Strategy Coordination Group (MSCG) Adverse effects on seabed habitats https://circabc.europa.eu/ui/group/326ae5ac-0419-4167-83ca-e3c210534a69/library/e07da833-e6a2-43ae-b916-6bddfdf998b8/details

MSFD, 2008/56/EC - Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0056

OSPAR (2017). Intermediate Assessment 2017. https://oap.ospar.org/en/ospar-assessments/intermediate-assessment-2017

SEABED_7-2021-10 7th meeting of the Technical Group on seabed habitats and sea-floor integrity (TG Seabed) - Draft text to address the ToR point G “Methods for assessing adverse effects, using point source and spatial data/models (D6C5)”https://circabc.europa.eu/ui/group/326ae5ac-0419-4167-83ca-e3c210534a69/library/12a14113-f484-472a-841f-8a703c51db8c/details

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, 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

3 Criteria to evaluate the suitability of indicators/assessment methods (ToR A)

Introduction

In a world where there are many indicators proposed for use in the MSFD, it is important to have a way of evaluating the relative merits and values of these indicators and threshold values for management of GES that can be derived from them. There are also a wide range of possible "criteria to assess the suitability of state indicators" (Kershner *et al*., 2011). The ICES working group on the Ecosystem Effects of Fishing Activities (WGECO) in 2012 (ICES 2012) was asked to evaluate approaches to developing criteria for adopting indicators for MSFD descriptors. This was largely based on the work of Rice & Rochet (2005). The WGECO approach was applied by WGBIODIV in 2013 (ICES 2013a). WGBIODIV adapted and revised the table of criteria from WGECO 2012 to evaluate the performance of "common indicators" proposed by OSPAR to support implementation of the MSFD at sub-regional and regional scales. The 16 criteria were grouped into five main categories, and the principle characteristic of each indicator's performance examined against each of the criteria. Each criterion was also given an importance weighting, so that those criteria considered most important were represented more strongly in the aggregate, additive, scoring of each indicator. Guidelines for assessing the level of compliance of each indicator against each criterion were also provided. This table from WGBIODIV was used as the basis for table 1 where the criteria have been adapted and expanded in the context of sea floor integrity indicators specifically.

WKBENTH 1 (ICES 2017), also developed more specific set of indicator evaluation criteria in the context of MSFD D6. This used many of the criteria from WGECO/WGBIODIV but also raised some new ones. Additionally, an extensive review was carried out for Department of Fisheries and Ocean Canada for indicator criteria (Bundy *et al* 2019). Both these reports were examined, and cross checked against the WGECO/WGBIODIV table, and modifications made as appropriate.

The ToR also asked for methods to evaluate the thresholds that could be derived for each of the indicators. The assumption would be that, in general, only the higher scoring indicators (or those retained for other reasons) would then be evaluated in terms of the methods and approaches to threshold development. Threshold evaluation was addressed by WGECO in 2013 (ICES 2013b), and a second table, adapted from the indicator table (ICES 2012) was produced. These were not given any weightings at the time, and these were developed at this workshop.

3.1 Criteria to evaluate the suitability of indicators

WKBENTH Indicator criteria

WKBENTH1 developed 11 evaluation criteria, and with some sub-criteria. These are summarised below taken from Table 8.1.1. in the WKBENTH report (ICES 2017). Each criterion was evaluated against the WGECO/WGBIODIV table (ICES 2013a):

1. Scientific evidence must provide a clear basis linking ecosystem features to impacts from bottom contacting fishing gears that are relevant to the achievement of objectives
   a. Does the indicator relate to features of the benthic community?
   b. Is there evidence linking pressure to ecosystem features?

   Both these criteria are addressed in WGECO/WGBIODIV Criterion 6

2. Trends in the indicator should be sensitive to changes in the pressure
   a. Is the indicator responsive to changes in pressure?
   b. Can the indicator be used to measure progress in time (e.g. GES in MSFD)
   c. Does the indicator represent tolerance/ resistance and recovery/resilience aspects?

   Criterion 2a is addressed in WGECO/WGBIODIV Criterion 6, and 2b in Criterion 2

3. Indicators should respond to the properties they are intended to measure (and it should be possible to disentangle the effects from other factors)
   a. Does the indicator solely relate to a single pressure (e.g. fishing disturbance or can effects from a single pressure be disentangled?
   b. Can the method quantify uncertainty?

   Criterion 3b is addressed in WGECO/WGBIODIV Criterion 3. Criterion 3a relates to specificity/sensitivity and is discussed below.

4. The underlying data layers should be adequate
   a. What is the type of data for indicator development?
   b. Is the spatial coverage of underlying data appropriate for indicator development?
   c. Are broad-scale habitat types represented?
   d. Are all relevant activities, e.g. all bottom-contacting fishing gears included?
   e. Is the output quantitative, semi-quantitative, or qualitative?

   Criteria 4a and 4e are addressed in WGECO/WGBIODIV Criterion 4. Criterion 4b is addressed in WGECO/WGBIODIV Criterion 5, and 4c is included implicitly as the criteria relate to region and sub-region.

5. An appropriate reference informs the model
   a. Is a reference state used to inform the indicator?
   b. Is an unimpacted reference state in relation to condition used to inform the indicator?

   Criteria 5a and 5b are addressed in WGECO/WGBIODIV Criterion 7

6. The indicator includes mechanisms that adhere to the precautionary principle

   a. Is a precautionary margin included in the indicator
   b. Is the indicator sensitive to keystone functions/species

Criterion 6a was considered as more appropriate for threshold evaluation. Criterion 6b was not included for this report as it would require a definition of what is "keystone" and would likely require a set of criteria for that definition.

7. The indicator is cost effective
   a. Is more empirical data required to apply the indicator to all broad-scale habitat types
   b. Is ongoing habitat monitoring required for indicator refinements?

Criteria 7a and 7b are addressed in WGECO/WGBIODIV Criterion 11

8. The indicator is able to include other pressures or cumulative effects
   a. Can cumulative physical abrasion be included?
   b. Can other pressures be included in the indicator?

These criteria were considered as relating to single v. multiple metric indicators, and this is now addressed in WGECO/WGBIODIV Criterion 6b

9. The indicator should be broadly applicable and comparable across regions
   a. Does the indicator cover all relevant broad-scale benthic habitats types? **C5**
   b. Does the indicator allow cross-regional comparison? **C5**

Criteria 9a and 9b are addressed in WGECO/WGBIODIV Criterion 5

Updating of the WGECO/WGBIODIV criteria

The WGBECO/WGBIODIV indicator evaluation criterion table was modified by WKBENTH 2 to make some of the criteria more appropriate to the seafloor integrity descriptor, and to clarify some of guidelines after reconsideration by the members of the current workshop. It was also updated to include considerations from WKBENTH 1 above. Consideration was also taken of more recent work on indicator evaluation criteria (Bundy *et al* 2019, Shin *et al* 2018. Most particularly this was with reference to the definitions and value of specificity and sensitivity. For sensitivity, WKBENTH 1 (ICES 2017) defined this as "Trends in the indicator should be sensitive to changes in the pressure". Houle *et al* (2012) define sensitivity as measuring how much an indicator would change if the community changed. For specificity, WKBENTH 1 (ICES 2017) said that "Indicators should respond to the properties they are intended to measure (and it should be possible to disentangle the effects from other factors)". Houle *et al* defined sensitivity as a measure of the proportion of change in the indicator attributed to fishing (or any other specific pressure) compared with other causes.

Criterion 1: State or pressure? This asked if the indicator was a "pressure" indicator being used for want of an appropriate "state" indicator? As this is a common approach in seafloor integrity indicators, for risk analysis, it was considered as redundant for WKBENTH 2.

Criterion 2: Existing and ongoing data. There was some uncertainty about this criterion as to whether it referred to widely available data or only in one region or sub-region. The guidelines were edited to add: "Indicator could be scored high even if only in one region. So the data support would have to include all the necessary data streams to calculate the indicator for that region"

Criterion 3 & 4: Minor language edits

Criterion 5: This was modified to refer to commercial gear as well as spatial coverage addressing WKBENTH1 criterion 4d. "Are all relevant bottom-contacting fishing gears included?" The text was modified to include that it should refer to a "a representative proportion of the MSFD sub-region (in terms of ecological and pressure gradients".

Criterion 6a: This criterion is where the concepts of sensitivity and specificity are included (Bundt *et al* 2019). The text has not been modified as these concepts were included, but perhaps needed emphasis.

Criterion 6b: This is a new criterion, and is designed to identify whether the indicator is specifically linked to a single (or predominant) pressure, making it "specific", or is not linked to a single (or predominant) pressure, but affected by a number of pressures, making it "non-specific". This criterion is not scored but represents a flag for attention. WKBENTH1 considered it desirable for an indicator to integrate multiple pressures, while WGECO/WGBIODIV felt that "specificity" was the critical factor. Both positions have merit, so this has been taken in a different approach, and the criterion can be used to separate the two indicator types, and allow the evaluator to determine which is most appropriate. WKBENTH2 participants concluded that both indicator types are equally important for the MSFD assessment of benthic habitats.

Criterion 7: Minor language edits

Criterion 8: The text was altered slightly to reflect that indicator links directly to management response should be included whether or not they are immediately operational.

Criterion 9-16: Minor edits

The final selection of criteria, rationale, weighting and guidance on scoring are presented in Table 3.1.1.

Table 3.1.1. Revised WGECO (ICES 2012) criteria used by WGBIODIV (ICES 2013) to evaluate the performance of "common indicators" proposed by OSPAR to support implementation of the MSFD at sub-regional and regional scale. The 16 criteria are grouped into five main categories, and the principle characteristic of each indicator's performance examined by each criterion is given. The importance weightings, and their associated scores, assigned by WGBIODIV to each criterion are shown, as are the guidelines for assessing the level of compliance of each indicator against each criterion. Pale blue cells indicate criteria not contributing to WGBIODIV's analytical assessment of the performance of the OSPAR "common indicators". In the compliance guidelines column, criteria automatically given a zero compliance score if the indicator was deemed to be a "pressure" indicator (criterion 1) are highlighted.

Criterion No.	Category	Characteristic	Criterion	*Importance* Weighting	*Importance* Score A	Guidelines for *Compliance* Assessment. Score B
1	Type of Indicator	State or pressure	Is indicator a "pressure" indicator being used for want of an appropriate "state" indicator?			Fully met (1): indicator is a "state" indicator; Not met (0): indicator is actually a "pressure" indicator. Not scored for WKBENTH2
2	Quality of underlying data	Existing and ongoing data	Indicators must be supported by current or planned monitoring programmes that provide the data necessary to derive the indicator. Ideal monitoring programmes should have a time series capable of supporting baselines and reference point setting. Data should be collected on multiple sequential occasions using consistent protocols, which account for spatial and temporal heterogeneity.	Core	3	Fully met (1): long-term and ongoing data from which historic reference levels can be derived and past and future trends determined; Partially met (0.5): no baseline information, but ongoing monitoring or historic data available, but monitoring programme discontinued, however potential to re-establish the programme exists; Not met (0): data sources are fragmented, no planned monitoring programme in the future. Indicator could be scored high even if only in one region. So the data support would have to include all the necessary data streams to calculate the indicator for that region
3	Quality of underlying data	Indicators should be concrete	Indicators should ideally be easily and accurately determined using technically feasible and quality assured methods, and have high signal to noise ratio, i.e. there is little variance in the calculation of the indicator, either from natural variability or sampling variability.	Core	3	Fully met (1): data and methods are technically feasible, widely adopted and quality assured in all aspects, signal to noise ratio is high; Partially met (0.5): potential issues with quality assurance, or methods not widely adopted, poor signal to noise ratio; Not met (0): indicator is not concrete or doubtful; noise excessively high due either to poor data quality or the indicator is unduly sensitive to environmental drivers

Criterion No.	Category	Characteristic	Criterion	*Importance* Weighting	*Importance* Score A	Guidelines for *Compliance* Assessment. Score B
4	Quality of underlying data	Quantitative versus qualitative	Quantitative measurements are preferred over qualitative, categorical measurements, which in turn are preferred over expert opinions and professional judgments.	desirable	2	Fully met (1): most data for the indicator are quantitative; Partially met (0.5): e,g. data for the indicator are semi-quantitative or largely qualitative; Not met (0): the indicator is largely based on expert judgement.
5	Quality of underlying data	Relevant spatial and gear_coverage	Data should be derived from a representative proportion of the MSFD sub-region (in terms of ecological and pressure gradients), at appropriate spatial resolution and sampling design, to which the indicator will apply. This should include if the indicator is capable of including different gears with different impacts on habitats or species, if this is relevant for the indicator and its application.	Core	3	Fully met (1): Representative monitoring is undertaken across the sub-region; Partially met (0.5): monitoring does not cover the full sub-region or the gears used and/or is not fully representative, but is considered adequate to assess status at sub-regional scale; Not met (0): monitoring is undertaken across a limited fraction of the sub-region or gears and is considered inadequate to assess status at sub-regional scale.
6	Quality of underlying data	Reflects changes in ecosystem component that are caused by variation in any specified manageable pressures SENSITIVITY and SPECIFICITY	The indicator reflects change in the state of an ecological component that is caused by specific significant manageable pressures (e.g. fishing mortality, habitat destruction). The indicator should therefore respond sensitively to particular changes in pressures. The response should be unambiguous and in a predictable direction, based on theoretical or empirical knowledge, thus reflecting the effect of change in pressures on the ecosystem component in question. Ideally the pressure-state relationship should be defined under both the disturbance and recovery phases.	Core	3	IF CRITERION 1 IS SCORED 0 THEN THE SCORE MUST BE 0. Otherwise: Fully met (1): the indicator is primarily responsive to a single or multiple pressures and all the pressure-state[1] relationships are fully understood and defined, both under the disturbance and recovery phases of the relationship; Partially met (0.5): the indicator's response to one or more pressures are understood, but the indicator is also likely to be significantly influenced by other non-anthropogenic (e.g. environmental) drivers, and perhaps additional pressures, in a way that is not clearly defined. Response under recovery conditions may not be well understood; Not met (0): no clear pressure-state relationship is evident.

[1] Here the term pressure-state relationship is used in the sense described by Piet *et al.* (2007): e.g. fishing ***pressure*** (fishing mortality rate [*F*]) – ***state*** of the stock (stock biomass [*B*]).

Criterion No.	Category	Characteristic	Criterion	*Importance* Weighting	*Importance* Score A	Guidelines for *Compliance* Assessment. Score B
6a	Quality of underlying data	Reflects changes in ecosystem component that are caused by variation in any specified manageable pressures SPECIFICITY	Details how specific the indicator is to the driver(s) of concern and whether the effects of one driver can be disentangled from other drivers. This criterion is not scored	NA	NA	The indicator is specifically linked to a single (or predominant) pressure, making it "specific". The indicator is not linked to a single (or predominant) pressure, but affected by a number of pressures, making it "non-specific".
7	Management	Relevant to MSFD management targets GES at criterion level	Clear targets that meet appropriate thresholds (absolute values or trend directions) for the indicator can be specified that reflect management objectives, such as achieving GES.	Core	3	Fully met (1): an absolute threshold value for the indicator is set; Partially met (0.5): no absolute threshold set for the indicator, but a threshold trend direction for the indicator is established; Not met (0): thresholds or trends unknown.
8	Management	Relevant to management measures	Indicator links directly to management response whether or not immediately operational. The relationship between human activity and resulting pressure on the ecological component is clearly understood.	Core	3	IF CRITERION 1 IS SCORED 0 THEN THE SCORE MUST BE 0. Otherwise: Fully met (1): both response-activity and activity-pressure relationships are well defined - advise can provided on both the direction AND extent of any change in human activity required and the precise management measures required to achieve this; Partially met (0.5): response-activity and activity pressure relationships are not well understood, or only one of the relationships is defined, but not the other, so that the precise changes in pressure resulting from particular management actions cannot be predicted with certainty; Not met (0): no clear understanding of either relationship, so that the link between management response and pressure is completely obscure.
9	Management	Comprehensible	Indicators should be interpretable and explainable in a way that is easily understandable by policy-makers and other non-scientists (e.g. stakeholders) alike, and the consequences of variation in the indicator should be easy to communicate.	Desirable	2	Fully met (1): the indicator is easy to understand and communicate; Partially met (0.5): a more complex and difficult to understand indicator, but one for which the meaning of change in the indicator value is easy to communicate; Not met (0): the indicator is neither easy to understand or communicable.

Criterion No.	Category	Characteristic	Criterion	*Importance* Weighting	*Importance* Score A	Guidelines for *Compliance* Assessment. Score B
10	Management	Established indicator	Indicators used in established management frameworks (e.g. EcoQO indicators) are preferred over novel indicators that perform the same role. Internationally used indicators should have preference over indicators used only at a national level.	Desirable	2	Fully met (1): the indicator is established and used in international policy frameworks; Partially met (0.5): the indicator is established as a national indicator; Not met (0): the indicator has not previously been used in a management framework.
11	Management	Cost-effectiveness	Sampling, measuring, processing, analysing indicator data, and reporting assessment outcomes, should make effective use of limited financial resources.	Desirable	2	Fully met (1): little additional costs (no additional sampling is needed); Partially met (0.5): new sampling on already existing programmes is required; Not met (0): new sampling on new monitoring programs are necessary.
12	Management	Early warning	Indicators that signal potential future change in an ecosystem attribute before actual harm is indicated are advantageous. These could facilitate preventive management, which could be less costly than restorative management.	Informative	1	IF CRITERION 1 IS SCORED 0 THEN THE SCORE MUST BE 0. Otherwise: Fully met (1): indicator provides early warning because of its high sensitivity to a pressure or environmental driver with short response time; Not met (0): relatively insensitive indicator that is slow to respond.
13	Conceptual	Scientific credibility	Scientific, peer-reviewed findings should underpin the assertion that the indicator provides a true representation of variation in the ecosystem attribute in question. Meets FAIR criteria	Core	3	IF CRITERION 1 IS SCORED 0 THEN THE SCORE MUST BE 0. Otherwise: Fully met (1): peer-reviewed literature; Partially met (0.5): documented but not peer-reviewed; Not met (0): not documented, or peer-reviewed literature is contradictory.
14	Conceptual	Metrics relevance to MSFD criteria	For D6, metrics should fit the indicator criteria stated in the 2017 MSFD Decision document.	Core	3	Fully met (1): the metric complies with the criteria; Not met (0): the metric does not comply with the criteria.
15	Conceptual	Cross-application	Metrics that are applicable to more than one MSFD descriptor are preferable. E.g. BH3 -> D1 benthic habitat and D6.	Desirable	2	Fully met (1): metric is applicable across several MSFD descriptors; Not met (0): no cross-application.
16	Indicator suites	Indicator correlation	Different indicators making up a suite of indicators should each reflect variation in different attributes of the ecosystem component and thus be	Desirable	2	Fully met (1): the indicators are un-correlated; Partially met (0.5): correlation between some indicators; Not met (0): all indicators are correlated.

Criterion No.	Category	Characteristic	Criterion	*Importance* Weighting	*Importance* Score A	Guidelines for *Compliance* Assessment. Score B
			complementary. Potential correlation between indicators should be avoided. UNIQUENESS			

Application of the criteria to evaluate indicators

Following the refinement of the criteria, these were tested on six indicators for which the information extraction tables had been completed. The following Table 3.1.2. summarises the scores given to each indicator under the criteria.

Table 3.1.2 Summary of the scores given to each indicator under the criteria

Criterion	Weighting	HELCOM BQI	SoS (BH1)	M-AMBI	TDI	PD2	DKI
2	3	1	1 needs time series, but not sure if this is the case in ALL sub regions where applied	1	0.5 relatively new index	1 not all regions e.g. deep water	1
3	3	1	1	1	1	1	1
4	2	1	1	1	1 but includes at least one categorical parameter	1 but includes at least one categorical parameter	1
5	3	0.5 Not really gear focused, and not necessarily fully representative of the MSFD sub region	1 but principally trawl metiers	1, but probably does not include different gear types	1 but principally trawl metiers	1	1, Denmark only and probably does not include different gear types
6	3	1	1	1	1, in France/Spain, but possibly not in e.g. Kattegat	1	1
6a	NA	B. Non specific	A. Specific	B. Non specific	A. Specific	A. Specific	B. Non specific
7	3	1	1, but methods for setting this are under development	1	1	0 not thresholds yet	1
8	3	1	1	0.5 But not clear if the pressures can be disentangled	1	1 but without TV is limited	0.5 But not clear if the pressures can be disentangled
9	2	0.5 calculations are more difficult to	0.5 calculations are more difficult to	0.5 calculations are more difficult to	1	1	0.5 calculations are more difficult to

Criterion	Weighting	HELCOM BQI	SoS (BH1)	M-AMBI	TDI	PD2	DKI
		explain but results are easy to communicate	explain but results are easy to communicate	explain but results are easy to communicate			explain but results are easy to communicate
10	2	1	1	1	0.5 France mainly	0 not yet used	0.5 Denmark mainly
11	2	1	1	1	1	1	1
12	2	0 designed as a state indicator	0 designed as a state indicator	0 designed as a state indicator	0 designed as a state indicator	0 designed as a state indicator	0
13	3	1	1	1	1	1	1
14	3	1	1	1 possibly also D1/D4	1	1	1
15	2	1	1	1	0 not applicable for other descriptors	0 not applicable for other descriptors	1
16	2	0.5 several other indicators exist for benthic communities	0 several other indicators exist for benthic communities and fishing pressure	0.5 several other indicators exist for benthic communities	0 several other indicators exist for benthic communities and fishing pressure	0 several other indicators exist for benthic communities and fishing pressure	0.5 several other indicators exist for benthic communities
SCORE		**32.5**	**33.0**	**32.5**	**29.5**	**27.0**	**31.5**

When the weighting factors were applied the results were broadly similar across the six indicators. All these indicators have been developed and used for some time, and have had extensive research carried out on them. They are all in common use in at least one region, so it is not surprising they all score well, although all had some weakness against the criteria – the maximum possible score would have been 37. There was no preference in the scoring for specific versus non-specific indicators.

A second evaluation was carried out by a different group and produced a different set of scores. This group noted that it was very important to have at least one person who was familiar with each indicator on the evaluation team. This group considered that an indicator would likely score lower without this, mainly due to uncertainty. It is also worth noting that the second team did not include anyone familiar with the evaluation criteria and who had only had a short introduction. So, for instance, they down scored indicators that were only applied in one region or subregion, and they also considered gear as referring to sampling gears rather than commercial gears. The evaluation exercise was intended mainly to test the working of the criteria rather than provide a definitive scoring, as this would have taken longer than the time available. For this reason, the report only included the evaluation where experts in the indicators AND the criteria were included.

Main conclusions on indicator evaluation criteria

- The criteria worked reasonably well although with some misunderstandings.
- The key conclusion is that evaluations in the future (i.e. WKBENTH3) should include people with knowledge of the indicators, and those with knowledge of the evaluation criteria (e.g., scientists, managers).

3.2 Criteria to evaluate the suitability of threshold values

As with the indicator evaluation criteria, a similar table was developed for evaluation of thresholds (Table 3.2.1). This was based on work carried out at WGECO in 2013 (ICES 2013b).

It was not possible to evaluate actual threshold vales until the indicators have been chosen and appropriate thresholds calculated. It was possible to evaluate a variety of methods or approaches to calculate thresholds, however. The approaches are detailed in chapter 4, and the results of the evaluation are presented in Table 3.2.1.

Table 3.2.2. Results of the evaluation of a range of different approaches to setting thresholds against the evaluation criteria.

Approach	Guidance	Score
Natural variation.	State is within the range of pressure-free variation	17.5
Statistically Detectable change.	State that is just statistically detectably different from the baseline.	11.5
Tipping point.	Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	15
Maintain function	Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	15.5
Trade-off.	Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	12.25
Avoid collapse.	Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. B*lim*	14.5
Zero pressure.	Any level of pressure results in a degraded state.	8
Distance to degradation.	Can be steep to flat or flat to steep	15

The rationale for the scores is presented in Table 3.2.3. When each method of developing threshold values was scored, the highest score was for a Natural Variation approach, although Tipping Points, Maintaining Function, Avoid Collapse, and Distance to Degradation also scored well.

Table 3.2.1. Criteria and guidance for evaluating thresholds

Criterion No.	Category	Characteristic	Evaluation criterion	Criterion specification	Weighting	Criterion levels
1	Overall evaluation	Method of derivation	Approach to define threshold given.	Rationale and methodological approach to define threshold should be given.	3	(1): Rationale for setting threshold fully documented (0): No scientific justification provided for the threshold chosen, and evaluation of the other criteria can only be based on expert judgement

2	Management evaluation	Framework consistency	Threshold consistency	Thresholds should not conflict across indicators within MSFD and with international policy frameworks	1	(1): No conflicts within MSFD and international legislation;
						(0.5): No conflicts within MSFD;
						(0): Conflicts within MSFD
						Where conflicts are identified the inconsistencies should be addressed, and justified or removed.

3	Management evaluation	Regional consistency	Level of regional coordination	Threshold should be coordinated on relevant regional scale for shared regions and sub-regions (?)	3	(1): Full coordination (0.5): Partial coordination (0): No coordination Where coordination is missing it should addressed
4	Management evaluation	Framework consistency	Preference for established thresholds	Thresholds already accepted and used by wider society as reliable and meaningful, should be preferred over novel thresholds that perform the same role.	1	(1):Yes. The threshold is already established and used in a relevant policy framework (0): The threshold has not previously been used in a management framework
5	Scientific evaluation	State of ecosystem	Integrity	To what level of integrity does the threshold refer (e.g. sustainable use)	3	(1): the threshold allows a sustainable use of marine resources; (0): the threshold allows human activities without reference to the concept of sustainability;
6	Scientific evaluation	State of ecosystem	Adaptability of threshold	The threshold should be assigned/allowed to change with (a) refined analyses and models of the indicator time series, and/or (b) change in ecosystem information	2	(1): Methods provided to establish threshold are adaptable to new evidence if available, and useful. (0): Methods or approaches for adaptability not provided

Criterion No.	**Category**	**Characteristic**	**Evaluation criterion**	Criterion **specification**	Weighting	**Criterion levels**
7	Scientific evaluation	Data quality	Uncertainty in threshold estimates	The statistical method used for thresholds setting should provide upper and lower confidence limits.	1	(1): Statistically sound estimate of confidence limits (0.5): Scientifically justified limits set without statistical certainty (0): No estimate of uncertainty
8	Scientific evaluation	Data quality	Derivation of threshold	Threshold should be based on analytical models and ecological theory. Empirical derivation based on time series or baseline data are preferred over expert judgement.	3	(1): Analytical and theoretical derivation based on data, and/or empirical setting with strong supporting theory; (0.5): Empirical derivation based on historical time series/baseline data only. This would include using data/models from other regions to set thresholds. (0): Expert judgement
9	Scientific evaluation	Data quality	Spatial extent (range)	Threshold should be based on data for the region for which is being applied and for the same spatial scale	2	(1): Threshold set based on data covering the same spatial extent as the spatial extent of the assessment area; (0.5): Threshold set based on a larger or smaller overlapping area. (0): Threshold set based on out of area.
10	Societal evaluation	Societal acceptance	Cross-sectoral integration	Thresholds should be informed by and subject to cross-sectoral public consultation to include social economic and ecological implications of targets for society	2	(1) Information published and cross-sectoral public consultation carried out; (0): cross-sectoral public consultation NOT carried out
11	Societal evaluation	Societal acceptance	Ease of understanding	Rationale for the threshold should be easily understandable by policy-makers and other non-scientists alike, and clear to communicate.	1	(1): Rationale behind the threshold easy to understand and clear to communicate and definable outcomes in terms of GES (0): The rationale behind the threshold is neither easy to understand nor to communicate
12	Management evaluation	Ecologically meaningful	GES good/degraded	Threshold should identify the separation between good and degraded environmental status based on established ecological principles and analysis	2	(1): Threshold is able to distinguish good v. degraded status -— quantitative and based on ecological difference between good and degraded. (0.5): cannot distinguish, but can suggest management action e.g. direction? (0): Arbitrary or based on statistical detection of differences that are not ecologically based/meaningful

Table 3.2.3. Rationale for the scores from table 3.2.2.

Criterion No.	1. Natural variation. State is within the range of pressure-free variation	2. Statistically Detectable change. State that is just statistically detectably different from the baseline.	3. Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	4. Maintain function. Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	5. Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	7. Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Blim	10. Zero pressure. Any level of pressure results in a degraded state.	11. Distance to degradation. Can be steep to flat or flat to steep.
1	1 - threshold is set based on data analysis	1 - threshold is set based on specified data analysis (but ecological reasoning is missing). Needs to be estimated as a state threshold, while often this approach is used to estimate a pressure threshold.	1 - mathematical methods for estimating where the line from flatter to steeper exist	1 - It is feasible, but creates the new problem that a threshold for EF needs to be set using one of the other methods. Note: some indicators may directly indicate functioning and then another approach for setting methods needs to be used	0.5 - This yields an extent threshold, which can only be used in combination with a quality threshold obtained through another method. (this does not define GES). The approach to plot the trade-off curves are established in ICES, but choosing a point on the curve as a threshold is likely to be driven by societal or management choices and cannot be chosen purely based on science.	1 - Evaluation at the whole population level is needed, which combines the quality in each cells to estimate total population size, from which an extent threshold can be obtained. It may still require a recruitment threshold, but if this is implemented in a full stock model, the recruitment threshold would be defined. Only works for indicators that relate to benthic species stock abundance. Needs a lot further developments.	1 - It is very easy to apply and easy to understand rationale.	1 - mathematical methods to identify this specific points on the curve exist
2	Very difficult to score. Natural variation is assessed for an undisturbed state, which is not what the MFSD requires. But this method defines the deviation from undisturbed that is acceptable for GES.							

Criterion No.	1. Natural variation. State is within the range of pressure-free variation	2. Statistically Detectable change. State that is just statistically detectably different from the baseline.	3. Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	4. Maintain function. Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	5. Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	7. Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Blim	10. Zero pressure. Any level of pressure results in a degraded state.	11. Distance to degradation. Can be steep to flat or flat to steep.
3				Can only be scored after indicators and thresholds have been agreed. It is a both a data issues (where data used to set the threshold needs to be consistent across regions) and a management issue. The method could be the same but the value could be different between different habitats and regions.				
4	1 - Similar methods have been used for the WFD (although not exactly for defining GES) (e.g. phytoplankton, saltmarsh, seagrass for quality. For extent also what fraction to protect under Convention for Biological Diversity)	1 - Several of the quality indicators used by different countries on the WFD use this approach (BQI etc.)	0 - No examples identified	0 - No examples identified	1- Marxan evaluations of MPA placement where the socio-economic value of cells is taken in to account.	1 - the use of Blim in fisheries management	1 - Management of MPAs, e.g. no take zone	0- The statistical method has been use in detection of spawning grounds, but not applied to management
5	1 - It is defined relative to an undisturbed state, but allows a deviation from the undisturbed state. The mean of the state under sustainable use would be	0 - does not link explicitly to sustainability, may be too strict or too lenient.	1 - yes, identifies a point where human use is possible without degrading the system. But it is dangerous to be too close to some tipping points,	1 - would allow sustainable use because some deviation of state from undisturbed possible.	0.5 - this identifies a point where the benefits to the ecosystem are maximized while minimizing the cost to human activities , but it is not assured that the resulting ecosystem state is good.	1 - would allow sustainable use because some deviation of state from undisturbed possible.	0 - as a quality threshold, does not allow any human use in the area without pressure (but human use possible in the extent of the area where pressure is	0.5 - It allows human use, and it is trying to find a balance between human use and seabed state, may be too strict or too lenient with regards to seabed state.

Criterion No.	1. Natural variation. State is within the range of pressure-free variation	2. Statistically Detectable change. State that is just statistically detectably different from the baseline.	3. Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	4. Maintain function. Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	5. Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	7. Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Blim	10. Zero pressure. Any level of pressure results in a degraded state.	11. Distance to degradation. Can be steep to flat or flat to steep.
	below the mean of the undisturbed state but this within the range of the undisturbed state.		depending state~pressure.				possible.) Also depends on how you define zero pressure, what pressures does it include.	
6	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.	0 - as a quality threshold, is fixed.	1 - Any method that uses data analysis to estimate the thresholds can be updated using more and better data.
7	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)	0 - uncertainty does not exist for this threshold because it is a fixed value	1 - the method should allow estimating of the confidence limit (depending on the quality of the data available to use the approach)
8	1- provided enough data is available for you habitat. In the practical implementation it may	0.5 - it is based on a statistical analysis, but not underpinned by a theoretical underpinning of why this	1 - It is based on a theory about how state and pressure relate and derived mathematically	1 - it is based on a knowing the EF-state relationships and the pressure state relationship, but relies on a	0.25 - The decision is support by empirical data analysis, but the choice of the threshold based on the	1 - Based on theory and existing models, even if the practical implementation will be hard.	0 - a zero pressure threshold can be derived using analytical methods using data analysis (using some of the	1 - It is based on a theory about how state and pressure relate and derived mathematically

Criterion No.	1. Natural variation. State is within the range of pressure-free variation	2. Statistically Detectable change. State that is just statistically detectably different from the baseline.	3. Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	4. Maintain function. Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	5. Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	7. Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Blim	10. Zero pressure. Any level of pressure results in a degraded state.	11. Distance to degradation. Can be steep to flat or flat to steep.
	be needed to use data from other regions, and the score would likely drop (here and for many other criteria and methods).	would differentiate between good and degraded.		threshold for EF which we don't have at the moment.	trade-off is an expert or societal judgement.		other approaches in this table), but a general zero pressure thresholds is not underpinned by theoretical derivation or data analysis and is based on expert judgement.	
9	0.5 - yes, this would be possible, although in practice it will be determined by data availability within the same ecoregion, and undisturbed areas are, by definition, not the same spatial extent as the full assessment area	1 - this would be possible, although in practice it will be determined by data availability within the same ecoregion, and it may be needed to use data from other regions. Where data availability is good, the state~pressure relationship can be fitted using data from the full extent of the region.	0.5 - this would be possible, although in practice it will be determined by data availability within the same ecoregion, and it may be needed to use data from other regions. Where data availability is good, the state~pressure relationship can be fitted using data from the full extent of the region.	0.5 - In theory, it would be possible, but it would be very unlikely that enough data would be available just from the local region	1 - this analysis is entirely based on data from the regions and habitats under consideration	0.5 - In theory, it would be possible, but it would be very unlikely that enough data would be available just from the local region	(1) Cannot be scored against this criteria because it is a fixed threshold rather than a method.	1 - this would be possible, although in practice it will be determined by data availability within the same ecoregion, and it may be needed to use data from other regions. Where data availability is good, the state~pressure relationship can be fitted using data from the full extent of the region.
10				We cannot score this because this is too early in the process, and therefore this has not been done yet. This can be done when a threshold for a specific indicator has been chosen relative to the GES quality and extent targets.				

Criterion No.	1. Natural variation. State is within the range of pressure-free variation	2. Statistically Detectable change. State that is just statistically detectably different from the baseline.	3. Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	4. Maintain function. Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	5. Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	7. Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Blim	10. Zero pressure. Any level of pressure results in a degraded state.	11. Distance to degradation. Can be steep to flat or flat to steep.
11	1 - Yes, the general concept is easy to understand	0 - Methodology is easy to explain, but the link to GES is not easy to explain because there is no direct link to GES	1 - The concept is easy to understand and linked to GES	1 - This concept of defined GES based on function is easy to understand, and function is one of the parameters that the MFSD aims the preserve, and more widely applied (but it will be difficult to implement and explaining the process may be much harder)	0.5 - Sustainable use is what MFSD wants to achieve, and this approach would clearly aim at sustainable use, but it is less clear that it achieves GES. It is however easy to communicate and likely to be an approach that is readily accepted.	0.5 - Avoiding collapse is obviously important to society, it is not clearly linked to GES and defines the difference between moderate and degraded.	1 - This is very easy to understand, but societal acceptance is going to be very variable. Technically easy to implement so may be popular with policy makers.	0.5 - Avoiding complete degradation is obviously important to society, it is not clearly linked to GES and defines the difference between moderate and degraded.
12	1 - Provided the correct indicator has been chosen to assess seabed integrity, and data quality and quantity to estimate the threshold were available. We can be quite certain that the state is good within the	0 - the detectable change is dependent on the magnitude of the change as well as the statistical power to detect the effect, and does therefore not define an ecological difference	0.5 - the tipping point is likely to represent an important change in the state-pressure relationship that is likely to coincide, but not necessarily coincides with a change from good to degraded	1 - Maintaining function at the a 'pristine' level is not defined as the level needed for sustainable use in the MFSD, so this still needs a threshold for what maintained EF is. Where EF drops and is not maintained, state	0.5 - it does not define good state, but it is a spatial management tool towards achieving a balance between state and human activities.	0.5 - it is clear that going below Blim is degraded, but the state may need to be higher to be defined as good	0 - It is not a method to derive a threshold value, but just a threshold value. Assumes that no deviation from the undisturbed state is compatible with good state, which is not generally accepted to be the	0.5 - It identifies a threshold beyond which the habitat is clearly degraded, but it does identify not identify a threshold between good and degraded.

Criterion No.	1. Natural variation. State is within the range of pressure-free variation	2. Statistically Detectable change. State that is just statistically detectably different from the baseline.	3. Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep.	4. Maintain function. Maintaining ecosystem function at levels without pressure (moves the problems along because it needs a threshold for good EF)	5. Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.	7. Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Blim	10. Zero pressure. Any level of pressure results in a degraded state.	11. Distance to degradation. Can be steep to flat or flat to steep.
	range of natural variation, although it is less certain that the state is degraded below the threshold.	between good and degraded		can be assumed to become degraded			distinction between good and degraded state. There however exist habitats where any level of disturbance will result in degradation.	

References

Bundy, A., Gomez, C. and Cook, A.M., 2019. Scrupulous proxies: Defining and applying a rigorous framework for the selection and evaluation of a suite of ecological indicators. *Ecological Indicators*, *104*, pp.737-754.

Houle, J.E., Farnsworth, K.D., Rossberg, A.G., Reid, D.G., 2012. Assessing the sensitivity and specificity of fish community indicators to management action. Can. J. Fish. Aquat. Sci. 69 (6), 1065–1079

ICES. 2012. Report of the Working Group on the Ecosystem Effects of Fishing Activities (WGECO), 11–18 April 2012, Copenhagen, Denmark. ICES CM 2012/ACOM:26. 192 pp.

ICES. 2013a. Report of the Working Group on Biodiversity Science (WGBIODIV). ICES Expert Group reports (until 2018). Report. https://doi.org/10.17895/ices.pub.8845

ICES. 2013b. Report of the Working Group on the Ecosystem Effects of Fishing Activities (WGECO), 1–8 May 2013, Copenhagen, Denmark. ICES CM 2013/ACOM:25. 117 pp.

ICES. 2017. Report of the Workshop to evaluate regional benthic pressure and impact indicator(s) from bottom fishing (WKBENTH), 28 February–3 March 2017, Copenhagen, Denmark. ICES CM 2017/ACOM:40. 233 pp.

Kershner, J., Samhouri, J.F., James, C.A. and Levin, P.S., 2011. Selecting indicator portfolios for marine species and food webs: a Puget Sound case study. *PLoS One*, *6*(10), p.e25248.

Rice, J.C., Rochet, M., 2005. A framework for selecting a suite of indicators for fisheries management. ICES J. Mar. Sci. 62, 516–527.

Shin, Y.J., Houle, J.E., Akoglu, E., Blanchard, J.L., Bundy, A., Coll, M., Demarcq, H., Fu, C., Fulton, E.A., Heymans, J.J., Salihoglu, B., 2018. The specificity of marine ecological indicators to fishing in the face of environmental change: a multi-model evaluation. Ecol. Ind. 89, 317–326.

4 Options for setting thresholds to evaluate adverse effects on seabed habitats (ToR B)

Assessing the state of ecosystems requires indicators, and thresholds above which the value of the indicator defines a good environmental state. Thresholds are defined here as the state at which an ecosystem transitions from a good to a degraded state. Existing thresholds of either ecosystem state ('quality') or spatial extent ('extent') have been chosen using a wide variety of approaches. Some of these approaches are criticized for being subjective and inconsistent, rather than being based on ecological principles and derived from the objective analysis of ecological data (Dorrough *et al.*, 2020)

Natural processes can result in fluctuations of the ecosystem state across space and time, and it is generally agreed that while some change in the state can be compatible with a system being in a good state, as well as some human use, larger change would lead to a degraded state (Folke *et al.*, 2003). Effective thresholds need to be ecologically meaningful, and therefore separate good and degraded state based on the characteristics of the ecosystem that management aims to conserve. Deciding how much change is compatible with a good state has proven difficult. Some participants made the point that only undisturbed areas contain key elements that ensure the biotic and abiotic structure of a habitat and its functions (e.g., its typical species composition and their relative abundance, presence of particularly sensitive or fragile species or species providing a key function, and the size structure of species) that guaranteed and in a good environmental status. Others have focused on thresholds that are either statistically detectable or where a small change in a driver causes a marked change in ecosystem condition (Groffman *et al.*, 2006). In other cases, sustainable use is equated with good state, accepting that sustainable use may result in reductions in abundance, biomass and even local extinction of some species. Additionally, it was also discussed if it can be called sustainable use if the total area of a habitat type is under anthropogenic pressure, and whether natural diversity, productivity, and ecological processes (functioning) are maintained and resilience of the habitat to environmental and climate change is secured. Nevertheless, many thresholds seem to have been chosen subjectively by experts or stakeholders (e.g. Muxika *et al.*, 2007).

The workshop reviewed the principles and criteria that should be used for setting thresholds for Good Environmental Status of seafloor habitats under TorA. Under TorB We identified and reviewed the different methods that can be used for setting environmental management thresholds. This chapter summarises and builds on work prepared by Hiddink *et al.* (In prep) and Nichols (2022).

4.1 What is good and what is degraded?

The challenge is to manage the ecosystem so that ecosystems/communities/habitats are at a sufficiently "good" state to ensure we sustain overall ecological integrity. The degradation from an undisturbed to a degraded and then lost ecosystem is described in Figure 4.1. Stage 1 & 2 both ensure biodiversity, structure, and function and can be considered 'good'. Most people would probably agree that stage 7 & 8 are degraded. Any changes from stage 3 to 6 may be considered as 'good enough' when part of a socio-economic trade-off and where a prioritization of the management actions is needed. We will therefore evaluate the different approaches that exists to define the transition from stage 1-2 to 3-8.

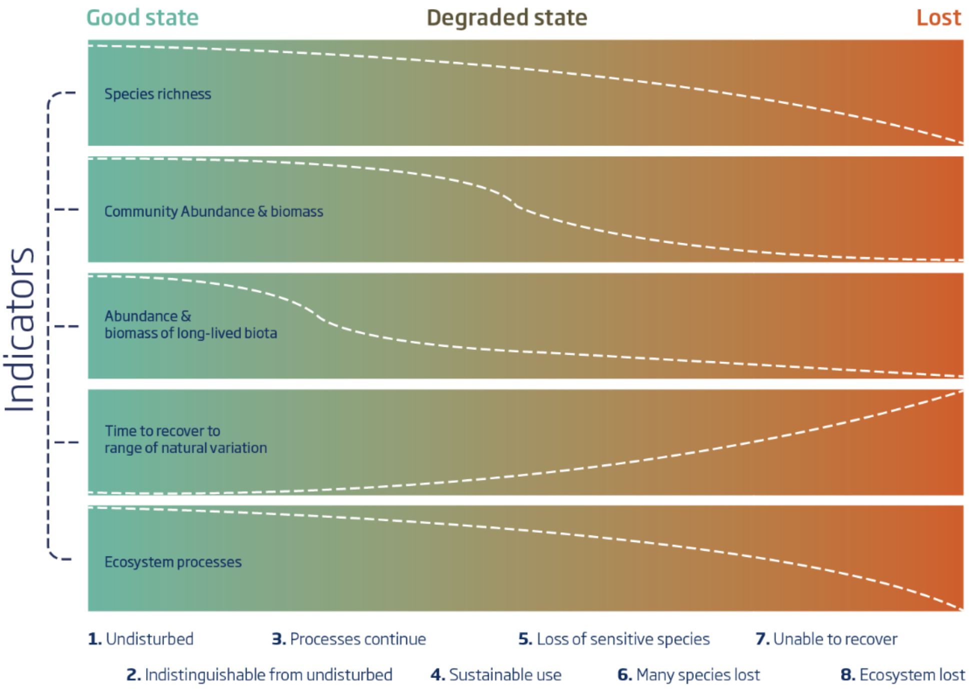


Figure 4.1. An undisturbed ecosystem is expected to have many species present, with each species having a natural distribution of abundance and biomass over the different age and size classes, with ecosystem processes at high rates (stage 1). Initially, when pressure from human activity is introduced, the ecosystem is indistinguishable from undisturbed in biodiversity, structure (age, size, species) and function because any changes fall within the range of natural variation (2). When the pressure increases further, it is expected that the largest and oldest individuals in the community will be lost, but all species will be present and ecosystem processes are likely to continue at rates that are near natural (3). Sustainable human use of the ecosystem can involve intense activities and is likely to result in widespread changes in size, age and species composition, with values generally outside the range of natural variation (4). Progressing pressure may result in the loss of the largest and most-long-lived species, resulting in large drops in the total biomass of the community, and large drops in the rates at which ecosystem processes occur (5). With further pressure, more species will be lost, and therefore overall species richness continues to drop, and all parameters are likely to be much lower than in undisturbed systems (6). At some level of pressure, the ecosystem would not be able to recover to its undisturbed state on human time-scales, even if the pressure was totally removed (7), and at the highest levels of pressure, the ecosystem can be considered lost and transformed into another ecosystem altogether (8) (Levin *et al.*, 2016). The indicator trends presented here assume a stochastic environment with no directional (and human-induced) environmental change

4.2 What are thresholds for good environmental state?

The threshold for good environmental state (GES) should identify the indicator value at which an ecosystem transitions from a good to a degraded state. Three different types of thresholds exist (quality, extent and connectivity), of which we will discuss quality and extent thresholds here in detail. Quality is defined by the indicator value on a local, point or cell, scale. A quality threshold defines at what value of the indicators the local quality can be considered to be 'good'. For consistency, we are discussing quality indicator values for quality as Ecological Quality Ratios, where the indicator values is scaled to range from 0 (fully degraded) to 1 (undisturbed). For example, a threshold for GES = 0.8 using community biomass as an indicator, would indicate that >80% of biomass needs to remain to achieve GES. The next step is to estimate the extent of the area (defined as a fraction of the total broad scale habitat (BSH) for the purposes of the MFSD) that is achieving this good quality. The extend threshold defines what fraction of the area of a BSH needs to be achieving the quality threshold for the whole BSH to be considered in a GES.

The aim of the TorB was to identify thresholds to distinguish between good and degraded, and identifying thresholds between degraded and lost was not considered to be part of the remit.

Finally, where indicators for the spatial coherence, configuration and/or connectivity of habitat patches in good quality exist, thresholds for what can be considered good can be developed too, although the current state of development of such indicators may make this premature.

Desirable characteristics of thresholds for good state are that they are habitat specific and estimated with their uncertainty. This means that a single approach for choosing thresholds that yields different thresholds for different habitats is desirable.

There was some discussion at the meeting about whether quality and extent threshold need to be interdependent, i.e. can a high quality threshold be combined with a lower extent threshold, and vice versa, for the achievement of GES? The consensus was that, although this may make ecological and intuitive sense, the knowledge base to implement such a trade-off between extent and quality does not currently exist.

4.3 Methods for estimating quality and extend thresholds

In a review of the literature and suggestions by workshop participants, we identified 11 approaches that have been proposed to set thresholds for good state (or to avoid an adverse effects or degraded state) in environmental management. Although we are evaluating methods for setting state thresholds, some of the reviewed methods are setting pressure thresholds as a method to achieving a good state. Figure 4.2 gives examples to illustrate each of the approaches.

Natural variation. State is within the range of pressure-free variation: quality threshold

The 'natural variation' threshold assumes that the quality is good if it is within the range of natural temporal variation, as it is therefore in effect indistinguishable from undisturbed (stage 1 and 2 in Figure 4.1). This threshold is conceptually easy to understand and defines good state in an ecologically meaningful way. Yet, the threshold relies on the availability of estimates of the natural variation in undisturbed systems, which may be hard to obtain. The natural range of variation can be defined as the 95% of values within which the indicator varies in undisturbed systems (Rossberg *et al.*, 2017, Östman *et al.*, 2020) or any other preferred quantile. This means that the threshold is a function of the natural variability of the indicator. If for example the abundance of species is used as the state indicator, a higher threshold will be needed for a long-lived species with a stable population size than for a short-lived species with large population fluctuations. This dependence on the life history of species is a desirable property, but the threshold may also depend on the magnitude sampling error in the dataset or natural asynchronous changes in abundance as a result of recruitment occurring at temporally different scales, which is not desirable. This approach assumes that the human pressure causes variation in abundance that are absorbed by natural variation. If human activities however cause variation that is additive to natural variations, even low levels of pressure may occasionally push the state outside the range of natural variation. Provided sufficiently long time series, quantitative estimates of the confidence of change can also be derived through this approach (Östman *et al.*, 2020).

Statistically **Detectable change.** State that is just statistically detectably different from the baseline. Quality threshold.

The 'detectable change' threshold sets a quality threshold for good state at the level that is just statistically detectable. Although this is an objective, data-driven method of defining a threshold, detectability is a function of the power to detect effects rather than a definition of ecological degradation of state, and a lack of detectability can just represent a lack of statistical power in a sampling design (i.e. the level of replication), or alternatively, given enough sampling effort, it may be possible to detect effects that are not ecologically meaningful. The threshold therefore does not necessarily equate to degradation of the state. Detectability is likely to decrease with the magnitude of natural variation, and the two concepts are therefore related. For seabed ecosystems, establishing a threshold for 'detectable change' hence mostly depends on the survey design rather than good state.

Tipping point. Breakpoint in statistical relationship between state and pressure, where the state-pressure relationship is going from flat to steep. Quality threshold**.**

'Tipping points' are thresholds at the level where the rate of harm per unit disturbance suddenly increases, and these could be indicative of regime shifts or alternative stable states (Folke *et al.*, 2003; Groffman *et al.*, 2006). These, usually abrupt, non-linear responses occur when the ecosystem no longer counters the effects of cumulative pressure via positive feedback loops, in other words, when the amount of disturbance exceeds its resilience. Identifying such tipping points is an objective data-driven approach to set thresholds, and although the tipping point does not necessarily coincide with a shift from good to degraded state, pressure beyond the tipping point is likely to lead to a degraded state (ranging from stage 4-8). A more serious problem is that tipping point thresholds are rarely detected for physical disturbance (Hillebrand *et al.*, 2020), but they are likely to be more commonly present for chemical and nutrient pollution. It also requires data across the full pressure gradient in order to derive meaningful relationships (Jac et al., 2020), and in the case of linear responses, there is no objective way to determine specific thresholds through this approach (Samhouri *et al.* 2012).

Maintain function. Maintaining ecosystem function at levels without pressure. Quality threshold.

Quality thresholds can be set to 'maintain the function' that is driven by the state rather than to maintain the state itself. This relies on identifying a relationship between state and ecosystem functioning and maintaining the ecosystem at the same levels that ecosystem functioning would be without pressure. This is an objective method of setting a threshold if the aim of the management is to maintain ecosystem functioning, but it still requires setting a threshold for good ecosystem functioning using one of the other methods for setting thresholds and therefore just moves the problem of setting a threshold one level further. For seabed ecosystems relevant functions have been suggested to be bioturbation, nutrient cycling and food provisioning for higher trophic levels (Rice *et al.*, 2012). However, each of these functions is likely to closely correlate with total community biomass, and therefore setting a threshold based on maintaining function is no easier than setting a threshold based on community biomass.

Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution. Extent threshold, and quality threshold in combination with extent

Rather than defining the threshold for a specific ecosystem, community or habitat based on ecological reasoning alone, an optimum state can be chosen at the wider landscape scale that achieves the best ecological state (generally estimated as the mean quality over the whole area) that can be achieved at low socio-economic cost. For example, when bottom trawling disturbs and degrades the seabed state, but also produces food, income and jobs, it may be possible to identify a situation where a balance between the seabed state and socio-economic benefits is found that is optimal from a societal point of view. Such a trade-off can be identified by plotting the relationship between the ecological state and the socio-economic benefits, and identifying the point on the curve that matches the societal preference (Lester *et al.*, 2013). ICES (2021) gives examples of such relationships. Identifying this point may be easy if there is a tipping point on a convex relationship and society prefers a fairly equal balance between conservation and human benefits, but will be hard to identify if there is a concave relationship. A disadvantage of the approach is that it does not identify a threshold that is necessarily ecologically good for a specific ecosystem, community or habitat, and that the threshold will vary with the changing intensity and distribution of pressure. This approach therefore sets a threshold for 'good enough' from a societal point of view across the wider landscape scale, rather than defining what 'good' is per se. This approach is therefore more of a management approach than a definition of a GES threshold.

Maximize ecosystem goods. Maximize the amount of goods that are produced for human use. Extent threshold.

'Maximizing the goods' that are produced for human use results in a quantitatively justified target and is, for example, used to set the biomass at which maximum sustainable yield of commercial fish populations is achieved. However, maximizing human benefits is not equivalent to achieving good environmental state, and the target is not a threshold above which the state is good, but a point estimate. Sustainable use should therefore not be conflated with good environmental state.

Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure, safeguarding future uses. Extent threshold

In fisheries management, thresholds have been chosen to avoid the collapse of fish stocks by maintaining the stock at a level where the reproductive capacity and recruitment is not impaired (B_{lim}, and this is similar to the concept of minimum viable population size which is used in conservation (Nunney & Campbell, 1993)). The total population size will be defined by the sum of quality indicators over the whole area, and can be implemented as an extent threshold. This is effectively the same as maintaining function, where the function to be maintained is reproductive output. However, the ability to maintain recruitment does not imply that the system is in a good state overall. This threshold offer information for defining the lower biological limit for sustainability and need for protection and/or management actions to maintain a viable population (avoiding stage 7 and 8), rather than defining good state. This reference point, often called B_{lim} in fisheries management, is around B/K=0.2 for many commercial fish species (Punt, 2010).

Recovery possible. Recovery of state possible within a specified time once the pressure is removed, safeguarding future use. Configuration, extent and quality threshold, because it depends on the total population reproductive potential and source-sink dynamics.

Rather than maintaining the current ecosystem in good state, thresholds can be chosen to keep the ecosystem in a state that would allow recovery within a specified time (e.g. years or generation time of species) once the pressure is removed, thereby safeguarding the option for future use (Rossberg *et al.*, 2017) and ensuring that the ecosystem remains resilient. For example, the FAO (2009) says that 'significant adverse impacts' on ecosystems will typically have recovery times exceeding 5–20 years. The state is already degraded at this level, so this threshold only preserves the ability to be good in the future. Recoverability may also be used as an extent threshold by evaluation at a broad habitat scale what level of abundance or biomass of benthic species needs to remain to allow recovery. This approach still requires setting a threshold for what is considered recovery and for the time to recovery that is acceptable. Note that in the MFSD, when recovery is expected to exceed a period of two reporting cycles (12 years), the area of habitat shall be considered as lost (COMMISSION DECISION (EU) 2017/848 L 125/43 of 17 May 2017).

Expert judgment. Expert elicitation to convert narrative description into quantitative description. Quality and/or Extent thresholds.

An alternative approach to setting thresholds involves 'expert judgment'. Such an approach uses expert elicitation to convert narrative description of good state into quantitative description, using any of the rationales of the approaches outlined above (e.g. Elliott *et al.*, 2018). Advantages of such an approach is the low demand for data, but this approach can be subjective, inconsistent and open to bias (Dorrough *et al.*, 2020).

Zero pressure. Any level of pressure results in a degraded state. Quality threshold

The 'zero pressure' threshold is effectively an expert-judgment approach where it is assumed that any location where human activity is not present is in a good state, and conversely that any location with activity is not in a good state. This has the advantage that data requirements for setting thresholds and evaluation state are minimal, but the disadvantage is that it ignores the fact that many systems can withstand some level of pressure without resulting in significant adverse impacts, and that this level varies between habitats and communities. An additional disadvantage is that locations with the same ecological state can be defined to be in either good or degraded state depending on their pressure level. It is important to note that the decision of aiming for a 'zero pressure' state can also be the outcome of the other approaches used, using a more data-driven and ecological threshold when the species and/or communities are highly vulnerable and cannot withstand any pressure intensity without being degraded (e.g. long-living and deep-sea corals (Clark *et al.*, 2016)). Studies have also shown that the establishment of no-take zones (zones without physical disturbance) can enhance resilience and recovery of benthic habitats and communities inside and outside these areas. Setting areas with zero-pressure can be a valid tool to achieve overall GES, but is not a method that distinguishes good from degraded. This approach is therefore a management approach rather than a definition of a GES threshold.

Distance to degradation. The point at which the habitat has already lost most of its quality (degradation point) and establish the condition threshold at a certain distance from it depending on habitat sensitivity (higher distances for higher sensitivities).Quality threshold.

The method consists of identifying the point at which the habitat has lost most of its quality (degradation point) and establishing the quality thresholds at different distances to this point depending on its sensitivity, giving the most sensitive habitats the highest distance to degradation. In this approach, the degradation point is defined as the point at which the pressure-state curves start to flatten out. Although several statistical tools are currently being explored to obtain this point, the method relies on the 45 degrees slope of the tangent to the curve, previously used in different works to determine the tipping point in aggregation curves (Colloca *et al.*, 2009; González-Irusta & Wright, 2017). Once this point has been computed, the condition threshold is established as a percentile of the distance between the origin of the curve and the degradation point. Currently, three potential distances are being explored (0.33 for habitats of sensitivity 4, 0.5 for habitats of sensitivity 3 and 0.66 for habitats of sensitivity 2), but further work is needed before these values can be considered final. The method used to establish habitat sensitivity (based on the pressure-state response curves) is explained in Serrano *et al.* (2022). In contrast to previous approaches, this is a specific method (currently under development in the frame of the NEA-PANACEA project and OSPAR OBHEG) and is not a general approach. The method has been developed to be applied to the OSPAR BH1 (also called SoS, Serrano *et al.*, 2022) and BH3 indicators and its application on other indicators may not be straightforward, and it may need further testing. Some members of the WK found it counter-intuitive that the less sensitive habitats reach the degradation point at higher state values than the sensitive habitats. However, this is an expected result since sensitive habitats will lose a higher proportion of their quality (e.g. proportion of sentinel species) compared to more resilient habitats before they stabilised. In fact, this is one of the main justifications (together with the precautionary approach) to establish a higher distance to this point in sensitive habitats than in the most resilient ones.

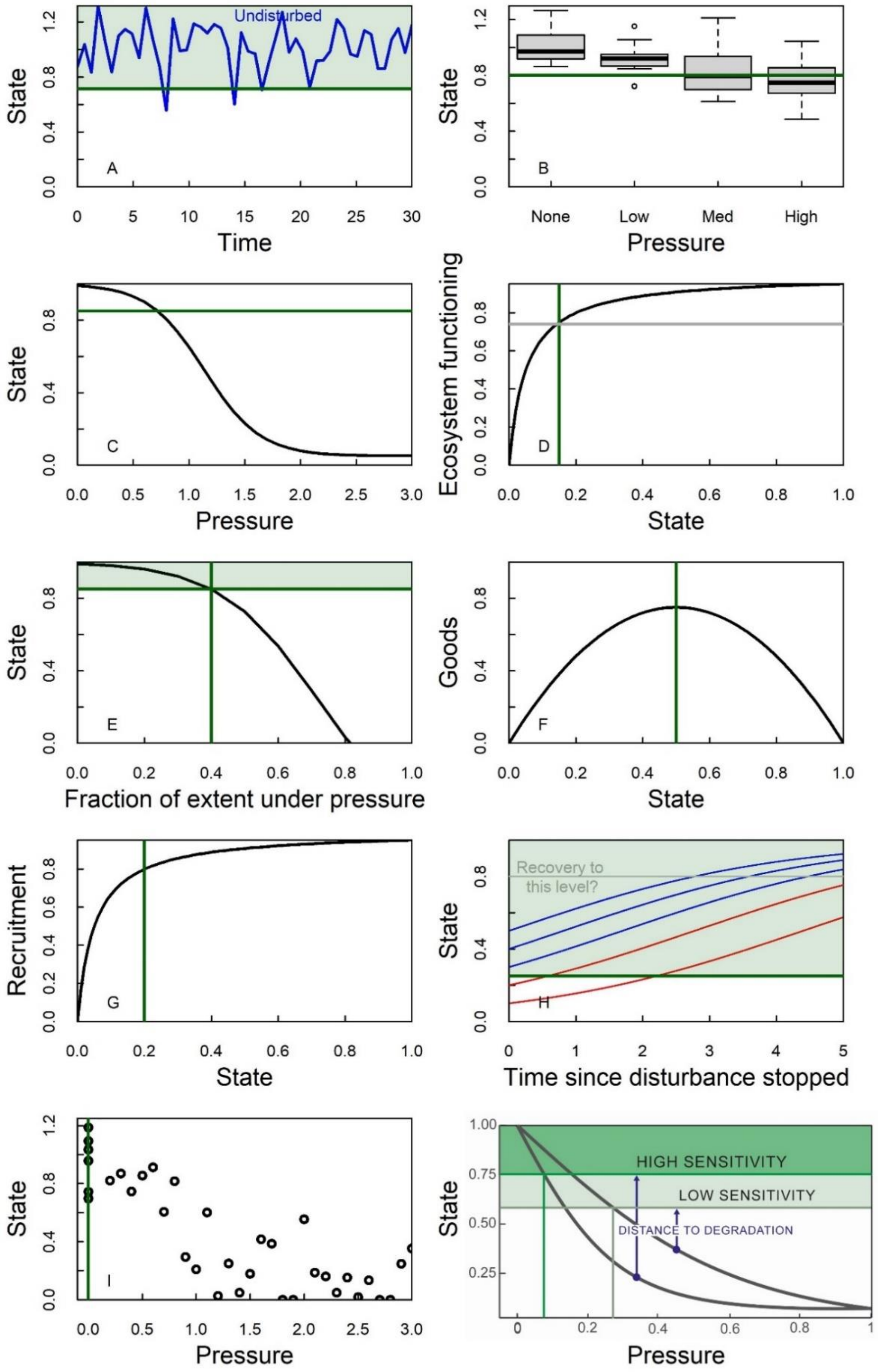


Figure 4.2. Illustrative examples of the derivation of different thresholds. The solid green line indicates the threshold between good state and degraded and the green polygon indicates the region above threshold (where present). Grey lines indicate other reference values used to derive the threshold.

Natural variation. State is within the range of pressure-free variation

Detectable change. State that is just statistically detectably different from the baseline.

Tipping point. Breakpoint in statistical relationship between state and pressure.

Maintain function. Maintaining ecosystem function at levels without pressure.

Trade-off. Find the point at which the increase in conservation benefits decreases relative to the decrease in the delivery of goods, as an optimal solution.

Maximize ecosystem goods. Maximize the amount of goods that are produced for human use.

Avoid collapse. Prevent collapse of ability to withstand (or recover from) pressure.

Recovery possible. Recovery of state possible within a specified time once the pressure is removed.

Zero pressure. Any level of pressure results in a degraded state.

Distance to degradation. State where the slope is 45 degrees plus a fraction of the distance to 1 depending on the sensitivity of the habitat.

4.4 Reflections of the WK participants on the suitability of different approaches

Each approach was formally evaluated against all criteria from TorA (presented in section 3), and in addition to this they were also discussed further by workshop participants.

General considerations

Firstly, there was agreement that a threshold is not a ‘target’ and that the term ‘target’ should be avoided when discussing points that describe the difference between good and degraded state.

It was considered a beneficial property of a threshold if it can be estimated as a probability distribution, so that it can be evaluated what the probability of exceeding the threshold is (i.e. threshold may be used in objectives-based risk assessment/analyses). This should be possible for any threshold that is estimated quantitatively from data.

There were some arguments about the need for two thresholds (e.g. one threshold at zero pressure where the EQR = 1 and another one at a lower EQR that separates good from degraded). Although there was strong support from a few participants to have both threshold values in the MSFD assessment, others argued that it is logically impossible to have two different thresholds that both separate good from degraded, and this suggestion was not taken forward. Some argued that the implementation of areas with zero pressure can nevertheless be useful as a management approach to achieve GES.

There was some discussion on whether there is a requirement to protect a higher extent of rarer or more vulnerable habitats. If they used to be more common and have been degraded because they are more sensitive, then this should be a characteristic of any threshold setting approach chosen. Other arguments given were that an area (in km^2) may be needed to maintain viability, rather than a fraction of the habitat as defined in the MSFD. This seems plausible based on general ecological insights.

The role of recoverability of the indicator after a management intervention and how this relates to the thresholds was mentioned several times, but no particular methods by which this could be achieved were suggested.

Prioritisation of approaches

The evaluation of the approaches would have ideally been performed after the selection of the high-priority indicators, because some approaches may work for some indicators but not for others. This was not feasible on the time-scale available for the workshop, and therefore the methods for setting thresholds were evaluated in a more generic way.

According to table 3.2.2. the workshop considered the 'natural variation' approach to estimating a quality threshold the most promising method, although in general most of the approached obtain similar values. The workshop considered that only two approaches (natural variation and maintaining function) clearly define an ecologically meaningful good state and estimate thresholds quantitatively from data. Both approaches were quality thresholds. Maintaining the state within the range of natural variation in undisturbed systems was considered a quantitative, objective and repeatable method that defines an ecologically meaningful good state. This approach may be applicable for most indicators although more work is needed before this can be established. The fact that the natural variation varies with environmental conditions and life-history of species was considered a desirable property as it would most likely result in higher thresholds for the most sensitive habitats (Nicholls, 2022). There are clear pathways towards making this approach operational, which would require the analysis of time-series from undisturbed locations. This requirement is unlikely to be satisfied for each broad scale habitat in each region, and therefore other approaches need to be taken to estimate natural variation for each habitat. The most problematic characteristic of this approach is that it is data hungry and finding data from undisturbed areas will be difficult a problem shared with other approaches (e.g. distance to degradation). Most time-series will also need to be detrended because long-term changes that are related to for example climate change will be causing long term increases or decreases. This was not considered a weakness of the approach, because it is a way of dealing with multiple pressures that are operating at different spatial scales.

The most promising approach would be to collate time-series from many different locations with a variety of environmental conditions, and use this dataset to predict the range of natural variation as a function of the environment. It may also be possible to use spatial rather than temporal variation when estimating natural variation. A similar approach has been taken to fit a sensitivity layer for the PD approach (Rijnsdorp *et al.*, 2018). One way of finding a quality threshold for "adversely affected" is to use areas without disturbance or least disturbed areas as a baseline. The distribution of indicator values calculated from the baseline data is considered to represent good environmental status and the threshold is a lower percentile calculated by bootstrapping from this distribution. This approach has been tested in Sweden for benthic grab data by Leonardsson *et al.* (2016). In their approach the threshold was set to the lower one-tailed 95% confidence limit of the mean indicator values based on five samples bootstrapped from the baseline data. One prerequisite for this method is that sufficient data from areas without or with little disturbance can be identified.

For extent, the WK agrees that in general, currently there is not enough scientific knowledge to provide an informed advice on extent thresholds able to distinguish good from degraded. From the two-extent threshold analysed the workshop considered that 'avoiding collapse' approach was the most promising method, because it defines an extent that is needed to maintain full reproductive potential of benthic invertebrates. This is most promising as an extent threshold, where the total population size of benthic invertebrates over the evaluated area is estimated as the sum of the quality of all cells. This is only sensible when using indicators of quality that directly relation to population size (biomass or abundance) and not for indicators such as M-AMBI or species richness. Operationalising this approach would require estimating the stock-recruitment relationship for the type of species that occur in a broad-scale habitat type. The majority of stock-recruitment relationships that exist have been derived for commercially exploited species of crustaceans and molluscs, and it is uncertain if they can be applied to seabed species in general. It is very unlikely that habitat specific relationships can be fitted. An exploration of this approach is presented below. For broadcast spawning species where Allee effects exist (depensation), because they need high densities to ensure external fertilisation (Gascoigne & Lipcius, 2004), there may also be potential for use 'avoiding collapse' as a quality threshold by identifying the abundance or biomass required for successful fertilisation at a local scale. However, it is unlikely that such relationships can be fitted for individual regions, and general Europe-wide or even global relationships may be need to be used.

We did not identify any methods to define extent thresholds that were considered optimal for distinguishing good from degraded. Two methods that were considered promising were the 'trade-off' and 'avoid collapse' approaches. The 'trade-off' approach instead is a spatial management tool towards achieving a balance between state and human activities, while the 'avoid collapse' method identifies a point below which the state is certainly degraded, but the state may need to be higher to be defined as good. Workshop participants were generally of the opinion that it may be very difficult to set scientifically justified extent thresholds, and expect that a trade-off approach may need to be used to set extent thresholds for GES.

Other approaches for distinguishing good and degraded

Other possibilities are using a qualitative description like in the WFD (Directive 2000/60/EC, Annex V) (no or minor deviation from reference conditions / no or minor anthropogenic alterations / abundance of characteristic species with no or slight deviation from reference condition), or by evaluating if the direction of the temporal change in the indicators is in a desirable direction. This was considered problematic, because it cannot be known if the direction is desirable if it is unknown if the current state is good or degraded. James Bell gave a presentation illustrating how thresholds for good state are derived by in the NAFO area, a summary is given in the appendix.

Thresholds for connectivity metrics

As stated in the previous, in addition to quality and extent thresholds, connectivity thresholds may be premature at its current state, or at least, it hasn't been thoroughly addressed in relation to the MSFD in general. Below is short update on the status of some of the current marine connectivity re-search activities, and suggestions on how connectivity may be incorporated into MSFD.

The fragmentation and deterioration of marine benthic habitats may disrupt dispersal pathways between habitats and patches increasing the vulnerability of meta-populations by reducing recruitment, decreasing the ability of population recovery and limiting gene transfer that may affect population resilience in time. This dispersal of marine organisms between seascape units are lately being referred to as marine functional connectivity, MFC (Darnaude *et al*. 2022), and MFC are typically inferred from studies using biophysical modelling that links predicted currents from oceanographic modelling with larval dispersal modelling.

While the importance of ocean currents for species dispersal has been generally accepted (e.g. Josefson and Hansen 2004, Cowen *et al*. 2006, Josefson 2016) recent years have provided a growing number of studies linking outcome of biophysical models with empirical data emphasizing the importance of MFC in both an evolutionary and demographic context. The former operating on multigenerational and long term time scales and the latter operating on year-to-year or ecological time scales (Lowe & Allendorf 2010, Marandel *et al*. 2017). As an example, a recent meta-study on coral reef fish found clear relationships between connectivity metrics and biodiversity indices and species abundances (Fontoura *et al*. 2022). Numerous other studies have found coincidence between empirical population genetic gradients and dispersal barriers inferred from biophysical model-ling (e.g. Mertens *et al*. 2018 and ref herein).

MFC is an evolving discipline and despite growing evidence on the importance MFC shaping marine benthic populations and communities, identification of indicators and thresholds in relation to management objectives (e.g. within the MFSD, WFD and the Habitat Directive) have not been fully developed. Methodologies on how to analyze MFC, however, are well established and based on network theories applied in other scientific disciplines studying complex networks, e.g. graph theory, information theory (e.g. Treml *et al*. 2008, Cecino *et al*. 2021).

A few examples of relevant metrics that can represent connectivity or configuration properties are shortly described below. Common for all, is that data analyses are based on biophysical modelling outputs that can incorporate any species-specific larval (or propagule) traits (or alternatively a more generic trait-based approach) and knowledge on habitat distribution and extent.

- Sink-Source dynamics: A habitat may serve primary as a source exporting propagules to other habitats, or as a sink, receiving propagules from other habitats. Pure source are-as may be particular vulnerable to habitat quality degradation due to limited recovery potential, and pure sink areas will not contribute to maintenance or recovery of other habitats. To optimize the configuration of habitats to meet quality thresholds of selected indicators, the fraction of habitats that serve as both sinks and sources should be maximized.
- Betweeness centrality: This is a metric in graph theory that identifies habitat which serve both as a source and as a sink, and that are particularly important as a link, via stepping stone dispersal, connecting different parts of a network which are otherwise less connected.
- Closeness centrality: This is somewhat supplementary to betweeness centrality a metric for detecting habitats that are able to spread propagules very efficiently, via stepping stone dispersal, through a habitat network
- Transitivity: This metric is also called “Cluster coefficient” and is a measure for how well the habitats in the neighbourhood of a given habitat is connected. Transitivity can be calculated for individual habitats (or patches) and for the whole network of habitats.
- Clustering: Clustering algorithms are often used in when analyzing MFC graphs to detect communities of habitats or patches, and particularly for detecting dispersal boundaries between otherwise well connected habitats. Dispersal boundaries from biophysical modelling are often found to coincide with population genetic gradient from empirical data.

A larger number of other graph metric exists that each represent distinct properties of the network.

While these metrics are relatively easy to calculate and compare for individual habitats and habitat networks, and for individual species, on a relative scale, absolute thresholds require state pressure relationships where pressures representing a level of fragmentation or modification of the habitat configuration, are included. This type of data may be difficult to obtain.

Instead, connectivity metrics could be addressed as quality criteria, where e.g. for any given extent threshold the configuration need to be optimized. The exact connectivity quality criteria would need to be developed and decided for. Examples could be:

- Criteria to minimize the number of isolated habitats (pure sources and/or pure sinks) in a current configuration setting or relative to a reference condition
- Criteria to avoid very limited number of habitats with high betweeness and closeness centrality, and instead favouring more habitats with intermediates values.
- Criteria to maximize the Transitivity of a given network

While connectivity metrics may be intuitively applicable to VME's and other habitats with limited spatial extent, connectivity metrics can be equally important for broad scale contiguous habitat types as e.g. as soft-bottom communities. Despite the contiguous properties of these habitats covering large areas, MFC analyses can identify areas that may be hydrographically isolated, marginally connected, or serving as a major stepping stone habitat linking habitats together which would otherwise be disconnected. Relevant studies for inspiration include Husebråten *et al*. (2018).

The future work on development of connectivity criteria to be cooperated into e.g. extent thresholds could benefit from combining the output of MFC metrics, e.g. sink and source dynamics, with demographic theory describing meta-population dynamics based on life history traits for individual species (fecundity, longevity, mortality, larval behaviour and dispersal, etc). This has been proposed for evaluating different resource and conservation management strategies (Puckett *et al* 2014, Theuerkauf *et al*. 2021). Despite obvious uncertainties in knowledge and presumptions required for such analysis, this type of exercise can be used to improve and test our understanding of the potential implications of habitat configurations on selected indicators/species in a more systematic, transparent and reproducible way, and thus, as a supplement to expert judgement.

4.5 Worked examples

Here we give an example of how the quality and extent thresholds can be estimated using the natural variation and avoid collapse approaches respectively for PD2 indicator (and not for L1), and show how they can be combined to evaluate if GES has been achieved. This illustration is partly based on work undertaken by Abigail Nichols as part of her MSc thesis at Bangor University.

Estimating a quality threshold using the natural variation approach for the PD method (Nichols, 2022)

The aim was to define GES by quantifying the annual range in natural variation (RNV) and its lower threshold of benthic invertebrate abundance in undisturbed seabed ecosystems; and to assess whether these measures covary with environmental variables.

A literature search was conducted to find benthic invertebrate abundance time series. The annual variation in abundance was used to calculate the RNV (defined as the 95% confidence interval) for each time series. It was hypothesised that the more stable the ecosystem, the smaller the RNV and therefore the higher the lower threshold (0.025 quantile) of benthic invertebrate abundance. Multiple linear regression analyses were performed with 52 studies; 402 studies were screened and examined against a set of eligibility criteria for inclusion in statistical analysis. Depth of the study site and benthic response (individual species or whole community abundance) were included as explanatory variables after backward model selection (Figure 4.3).

RNV significantly decreases and lower threshold significantly increases as depth increases, which is expected as ecosystems are typically more stable with increasing depth (Figure 4.4). This trend is seen across benthic responses, though the individual species studies have significantly higher RNVs and smaller lower thresholds than the whole community studies. Neither the RNV or lower threshold varied significantly with latitude, average species lifespan or substrate type.

On average the RNV of benthic invertebrate abundance was 1.06, translating to a lower threshold of conservatively estimated as 0.8. This means that the abundance can be reduced to around 80% of the mean abundance and still be considered as having a GES. There is potential to explore other environmental variables that may explain more of this variation.

The natural variation of the L1 indicator from time-series or spatially from undisturbed locations cannot be estimated because this indicator cannot be estimated without reference to a particular fishing intensity level (it is defined as the fraction of biomass with a longevity > 1 / swept-area-ratio).

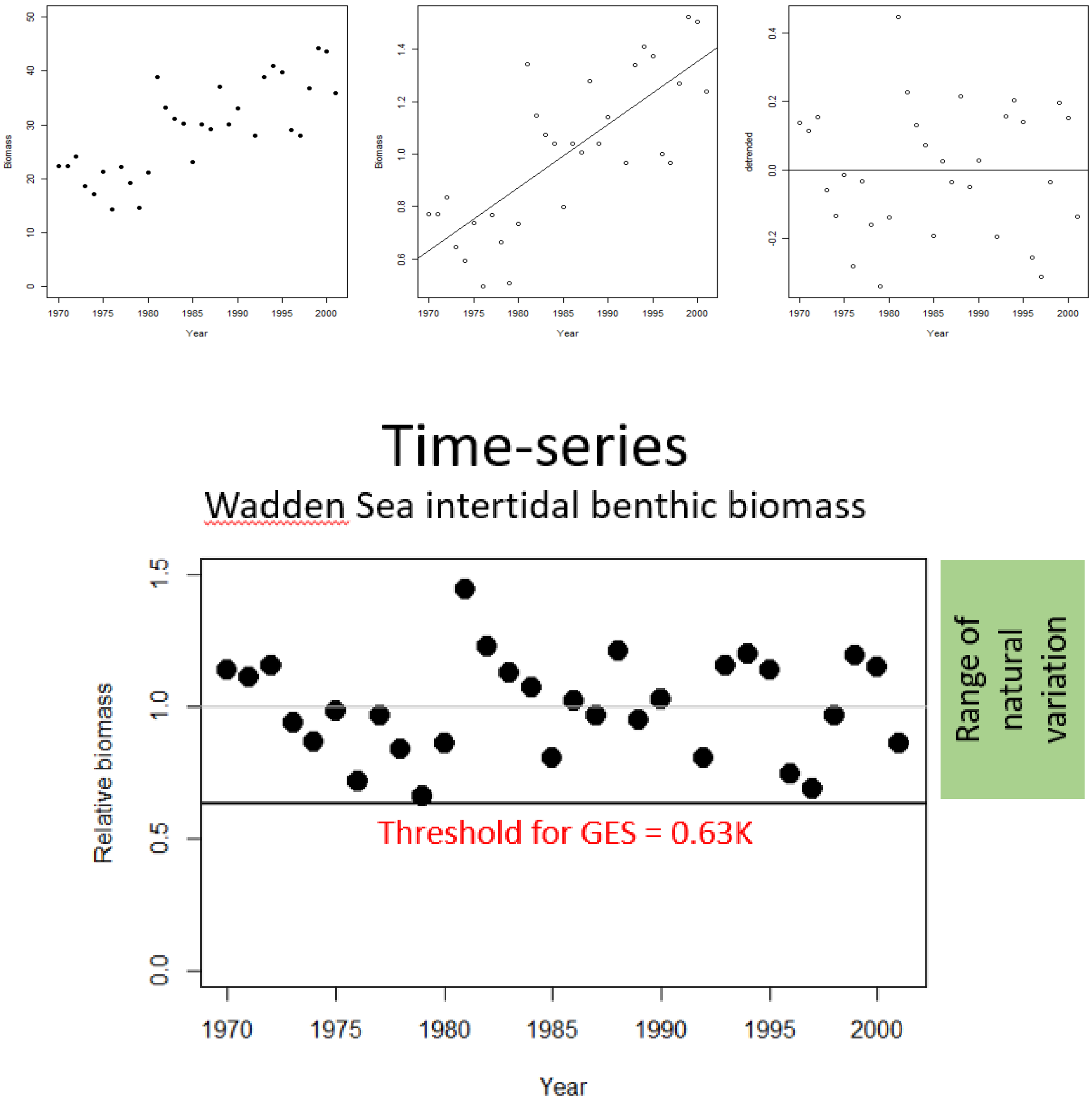


Figure 4.3. Example of the treatment of a time-series of intertidal community biomass. The raw data is scaled around 1 and a regression line fitted. The residuals around the regression line are then used to estimate the range of natural variation, which in this example would be at EQR = 0.63. Data from (Beukema & Cadee, 1997)

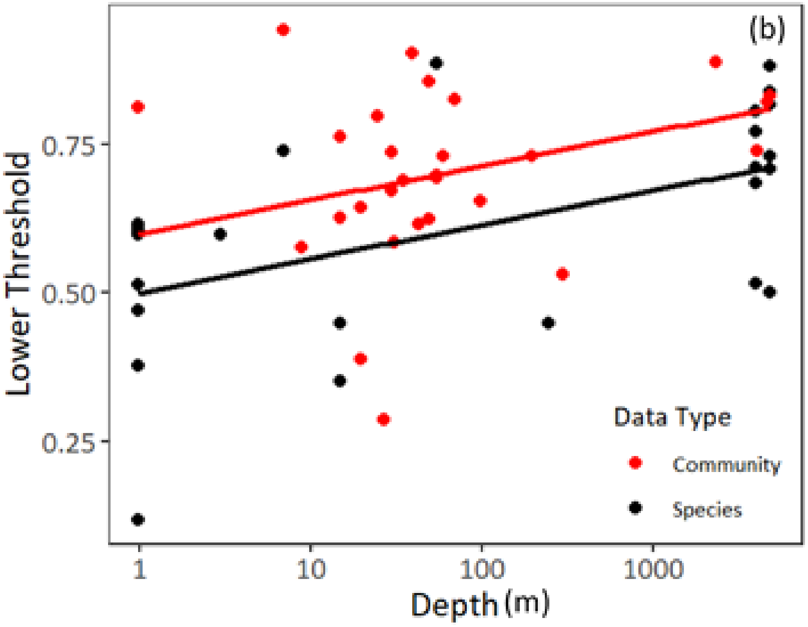


Figure 4.4. Results of a multiple linear regression analyses with lower threshold of abundance (numbers or biomass) as the response variable ($R^2 = 0.249$, $F(2, 49) = 8.105$, $p < 0.001$). Depth and benthic response were the explanatory variables in the models. Predicted equations for whole community: lower threshold = 0.598 + 0.057* $\log_{10}$(depth). Predicted equations for individual species: lower threshold = 0.499 + 0.057* $\log_{10}$(depth). Community: N = 28, Species: N = 24.

Estimating an extent threshold using the avoiding collapse approach for the PD method

This approach needs a stock-recruitment relationship that is relevant to the species that live in the habitats being evaluated, i.e. benthic invertebrates. The largest readily available dataset of stock-recruitment relationships comes from the RAMlegacy database (Ricard *et al.*, 2012). We fitted Ricker stock recruitment relationships for all species of invertebrates in the database, and for all data combined. The resulting dataset consisted mostly of shrimps and lobster, with a few mollusc stocks. Both stock abundance and recruitment were scaled for every stock so that the maximum was one.

This very preliminary analysis gives an indication that reproductive capacity of invertebrate species is not impaired when the stock is larger than 40% of its maximum observed value (Figure 4.5). There is obviously a lot of variation there and the fitted relationship is highly uncertain, but this example is meant to illustrate how this extent threshold could be made operational rather than provide a usable extent threshold.

Here we use outputs for PD and L1 indicators for the Baltic Sea and the North Sea that were created for WKTRADE3 to illustrate how the quality and extent thresholds can be combined to evaluate if the broadscale habitats are in GES. Here we show cumulative plots following the approach by Pitcher *et al.* (2022) (Figure 4.6). Lines that are further to the left on the plot are for habitats that are more impacted. The quality threshold is a value on the y-axis of these plots, while the extent threshold is a value on the x-axis of these plots.

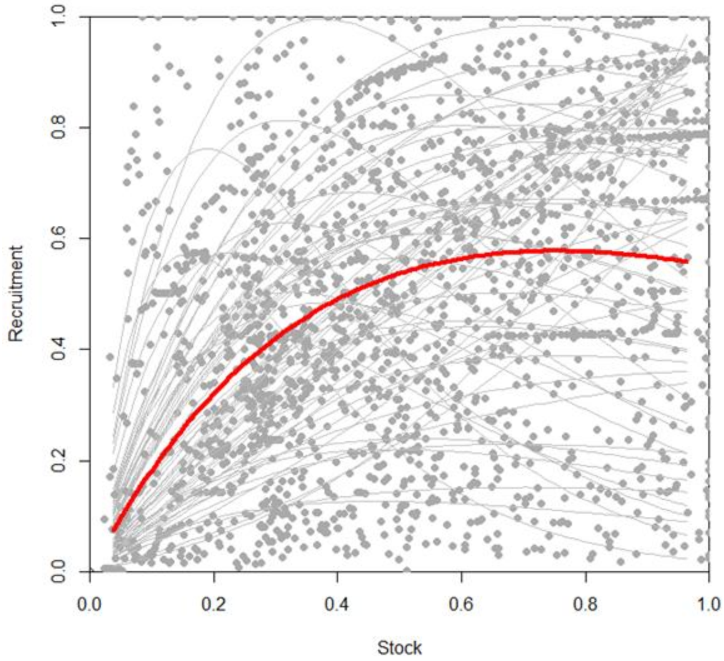


Figure 4.5. Stock-recruitment relationships for invertebrates from the RAMlegacy database. Grey lines indicate the fitted relationships for individual stocks, while the red line indicates the relationship fitted on all stocks together.

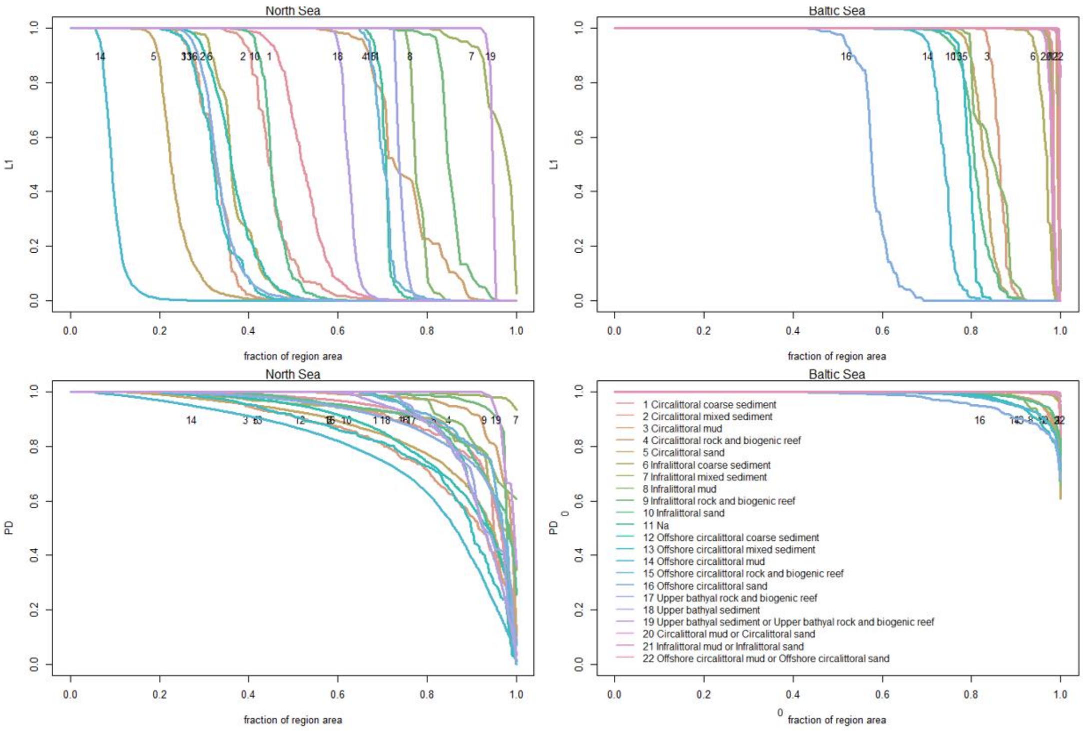


Figure 4.6. Distributions of grid cell indicator values (ordered 1 through 0) versus cumulative percentage of regional area. Where indicator = 1 at top/left indicates untrawled seabed, and indicator = 0 at bottom/right indicates depleted seabed.

Figure 4.7 illustrates how this would work out for the PD indicator, using the provisionally estimated quality (0.8) and extent thresholds (0.4). The quality threshold could have been made to vary with the depth of each of the habitats, so a different quality threshold was used for each habitat. In these plots, if the cumulative lines cross the pink box, that bounds the quality and extent threshold, a habitat would not be considered to be in GES. In this example all habitats would be in GES, with habitat 14 in the North Sea, offshore circalittoral mud, being closest to the thresholds. Given the preliminary nature of all analyses here, this example should be considered as an illustration of the process only.

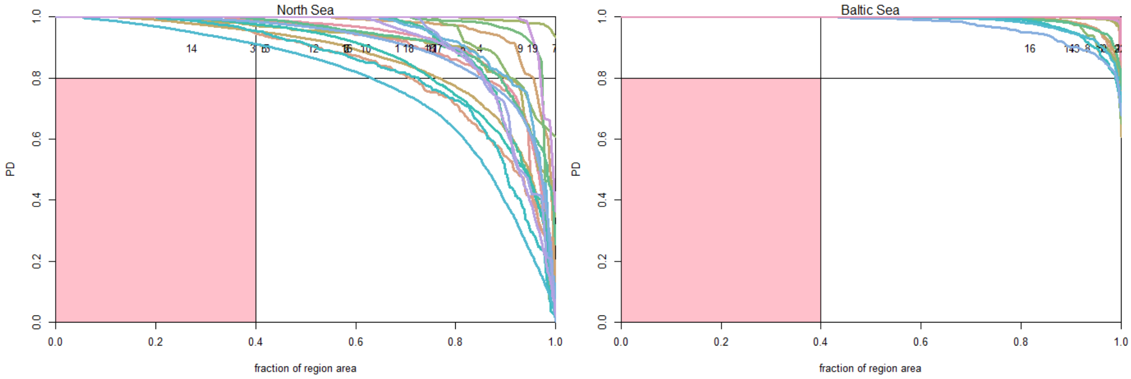


Figure 4.7. Distributions of grid cell indicator values (ordered 1 through 0) versus cumulative percentage of regional area for the PD indicator. The grey lines indicate the provisionally estimated quality and extent thresholds.

References

Beukema, J.J. & Cadee, G.C. (1997) Local differences in macrozoobenthic response to enhanced food supply caused by mild eutrophication in a Wadden Sea area: Food is only locally a limiting factor. Limnology and Oceanography, 42, 1424-1435.

Colloca, F., Bartolino, V., Lasinio, G. J., Sartor, P., & Ardizzone, G. (2009). Identifying fish nurseries using density and persistence measures. Marine Ecology Progress Series, 381(November 2015), 287–296. https://doi.org/10.3354/meps07942

Clark, M.R., Althaus, F., Schlacher, T.A., Williams, A., Bowden, D.A. & Rowden, A.A. (2016) The impacts of deep-sea fisheries on benthic communities: a review. ICES Journal of Marine Science, 73, i51–i69.

Dorrough, J., Watson, C., Martin, R., Smith, S., Eddy, D. & Farago, L. (2020) Identifying and testing conservation decision thresholds in temperate montane grasslands. Ecological Indicators, 118, 106710.

Cecino G., Valavi R., Treml E.A. (2021). Testing the Influence of Seascape Connectivity on Marine-Based Species Distribution Models. Frontiers in Marine Science 8

Cowen, R. K., Paris, C. B., & Srinivasan, A. (2006). Scaling of connectivity in marine populations. Science, 311(5760), 522–527

Darnaude, A., Arnaud-Haond, S., Hunter, E., Gaggiotti, O., Sturrock, A., Beger, M., Volckaert, F., Pérez-Ruzafa, A., López-López, L., Tanner, S. E., Turan, C., Ahmet Doğdu, S., Katsanevakis, S., & Costantini, F. (2022). Unifying approaches to Functional Marine Connectivity for improved marine resource management: the European SEA-UNICORN COST Action. Research Ideas and Outcomes, 8, 21

Elliott, S.A., Guérin, L., Pesch, R., Schmitt, P., Meakins, B., Vina-Herbon, C., González-Irusta, J.M., de la Torriente, A. & Serrano, A. (2018) Integrating benthic habitat indicators: working towards an ecosystem approach. Marine Policy, 90, 88-94.

FAO (2009) International guidelines for the management of deep-sea fisheries in the high seas. FAO.

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L. & Holling, C.S. (2003) Regime Shifts, Resilience, and Biodiversity in Ecosystem Management. Annual Review of Ecology, Evolution, and Systematics, 35, 557-581.

Fontoura L, D. S., Gamoyo, M., Barneche, D. R., Luiz, O. J., Madin, M. P., Eggertsen, L., & Maina, J. M. (2022). Protecting connectivity promotes successful biodiversity and fisheries conservation. Science, 375(6578), 336–340.

Gascoigne, J. & Lipcius, R.N. (2004) Allee effects in marine systems. Marine Ecology Progress Series, 269, 49-59.

González-Irusta, J. M., & Wright, P. J. (2017). Spawning grounds of whiting (Merlangius merlangus). Fisheries Research, 195, 141–151. https://doi.org/10.1016/j.fishres.2017.07.005

Groffman, P.M., Baron, J.S., Blett, T., Gold, A.J., Goodman, I., Gunderson, L.H., Levinson, B.M., Palmer, M.A., Paerl, H.W. & Peterson, G.D. (2006) Ecological thresholds: the key to successful environmental management or an important concept with no practical application? Ecosystems, 9, 1-13.

Hiddink, J.G., Van Denderen, D., Valanko, S. & Delargy, A. (In prep) Setting thresholds for good marine ecosystem state and significant adverse impacts. Manuscript.

Hillebrand, H., Donohue, I., Harpole, W.S., Hodapp, D., Kucera, M., Lewandowska, A.M., Merder, J., Montoya, J.M. & Freund, J.A. (2020) Thresholds for ecological responses to global change do not emerge from empirical data. Nature Ecology & Evolution, 4, 1502-1509.

Huserbråten, M., Moland, E., Jorde, P. E., Olsen, E. M., & Albretsen, J. (2018). Connectivity among marine protected areas, particularly valuable and vulnerable areas in the greater North Sea and Celtic Seas regions. In NorthSEE project report. North SEE - Interreg North Sea Region. Institute of Marine Research. https://northsearegion.eu/northsee/news/environmental-connectivity-study-published

ICES (2021) A series of two Workshops to develop a suite of management options to reduce the impacts of bottom fishing on seabed habitats and undertake analysis of the trade-offs between overall benefit to seabed habitats and loss of fisheries revenue/contribution margin for these options (WKTRADE3).

ICES Scientific Reports. 3:61. 100 pp. http://doi.org/10.17895/ices.pub.8206. In:

Jac, C., Desroy, N., Certain, G., Foveau, A., Labrune, C., & Vaz, S. (2020). Detecting adverse effect on seabed integrity. Part 2: How much of seabed habitats are left in good environmental status by fisheries?. Ecological Indicators, 117, 106617.

Leonardsson, K., Blomqvist, M. & Rosenberg, R. (2016) Reducing spatial variation in environmental assessment of marine benthic fauna. Mar Pollut Bull, 104, 129-38.

Josefson, A. B. (2016). Species Sorting of Benthic Invertebrates in a Salinity Gradient -- Importance of Dispersal Limitation. PLOS ONE, 11, 12.

Josefson, A. B., & Jls, H. (2004). Species richness of benthic macrofauna in Danish estuaries and coastal areas. Global Ecol Biogeogr, 13, 273–288

Lester, S.E., Costello, C., Halpern, B.S., Gaines, S.D., White, C. & Barth, J.A. (2013) Evaluating tradeoffs among ecosystem services to inform marine spatial planning. Marine Policy, 38, 80-89.

Levin, L.A., Mengerink, K., Gjerde, K.M., Rowden, A.A., Van Dover, C.L., Clark, M.R., Ramirez-Llodra, E., Currie, B., Smith, C.R., Sato, K.N., Gallo, N., Sweetman, A.K., Lily, H., Armstrong, C.W. & Brider, J. (2016) Defining "serious harm" to the marine environment in the context of deep-seabed mining. Marine Policy, 74, 245-259.

Lowe, W. H., & Allendorf, F. W. (2010). What can genetics tell us about population connectivity?: Genetic and demographic connectivity. Mol. Ecol, 19(15), 3038–3051.

Marandel, F., Lorance, P., Andrello, M., Charrier, G., Le Cam, S., Lehuta, S., & Trenkel, V. M. (2017). Insights from genetic and demographic connectivity for the management of rays and skates. Can J Fish Aquat Sci, 75(8), 1291–1302.

Mertens, L. E. A., Treml, E. A., & von der Heyden, S. (2018). Genetic and biophysical models help define marine conservation focus areas. Frontiers in Marine Science, 5(AUG).

Muxika, I., Borja, A. & Bald, J. (2007) Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. Mar Pollut Bull, 55, 16-29.

Nichols, A. (2022) How large are natural variations in the abundance of biota in seabed ecosystems? Bangor University,

Nunney, L. & Campbell, K.A. (1993) Assessing minimum viable population size: demography meets population genetics. Trends in Ecology & Evolution, 8, 234-239.

OSPAR (2021) OSPAR- Biodiversity Indicators. Common Indicator 1Sentinels of the Seabed (SoS). Generic guidelines for Coordinated Environmental Monitoring Programme (CEMP).

Östman, Ö., Bergström, L., Leonardsson, K., Gårdmark, A., Casini, M., Sjöblom, Y., ... & Olsson, J. (2020). Analyses of structural changes in ecological time series (ASCETS). Ecological Indicators, 116, 106469.

Pitcher, R., Hiddink, J.G., Jennings, S., Collie, J., Parma, A.M., Amoroso, R., Mazor, T., Sciberras, M., McConnaughey, R.A., Rijnsdorp, A.D., Kaiser, M.J., Suuronen, P. & Hilborn, R. (2022) Trawl impacts on the relative status of biotic communities of seabed sedimentary habitats in 24 regions worldwide. Proceedings of the National Academy of Sciences of the United States of America, 119, e2109449119.

Puckett, B. J., Eggleston, D. B., Kerr, P. C., & Luettich, R. A. (2014). Larval dispersal and population connectivity among a network of marine reserves. Fisheries Oceanography, 23(4), 342–361.

Punt, A.E. (2010) Harvest control rules and fisheries management. Handbook of marine fisheries conservation and management, 582-594.

Ricard, D., Minto, C., Jensen, O.P. & Baum, J.K. (2012) Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish and Fisheries, 13, 380-398.

Rice, J., Arvanitidis, C., Borja, A., Frid, C., Hiddink, J.G., Krause, J., Lorance, P., Ragnarsson, S.Á., Sköld, M., Trabucco, B., Enserink, L. & Norkko, A. (2012) Indicators for Sea-Floor Integrity under the European Marine Strategy Framework Directive. Ecological indicators, 12, 174–184.

Rijnsdorp, A.D., Bolam, S.G., Garcia, C., Hiddink, J.G., Hintzen, N., Kooten, T.v. & Denderen, P.D.v. (2018) Estimating the sensitivity seafloor habitats to disturbance by bottom trawling impacts based on the longevity of benthic fauna. Ecological Applications, 28, 1302-1312.

Rossberg, A.G., Uusitalo, L., Berg, T., Zaiko, A., Chenuil, A., Uyarra, M.C., Borja, A. & Lynam, C.P. (2017) Quantitative criteria for choosing targets and indicators for sustainable use of ecosystems. Ecological Indicators, 72, 215-224.

Samhouri, J. F., Lester, S. E., Selig, E. R., Halpern, B. S., Fogarty, M. J., Longo, C., & McLeod, K. L. (2012). Sea sick? Setting targets to assess ocean health and ecosystem services. Ecosphere, 3(5), 1-18.

Serrano, A., de la Torriente, A., Punzón, A., Blanco, M., Bellas, J., Durán-Muñoz, P., Murillo, F. J., Sacau, M., García-Alegre, A., Antolínez, A., Elliott, S., Guerin, L., Vina-Herbón, C., Marra, S., & González-Irusta, J. M. (2022). Sentinels of Seabed (SoS) indicator: Assessing benthic habitats condition using typical and sensitive species. Ecological Indicators, 140, 108979. https://doi.org/10.1016/j.ecolind.2022.108979

Theuerkauf, S. J., Puckett, B. J., & Eggleston, D. B. (2021). Metapopulation dynamics of oysters: sources, sinks, and implications for conservation and restoration. Ecosphere, 12(7), 7.

Treml, E. A., Halpin Urban, D. L., & Pratson, L. F. (2008). Modelling population connectivity by ocean currents, a graph-theoretic approach for marine conservation. Landscape Ecol, 23, 19–36.

Treml, E. A., Halpin, P. N., Urban, D. L., Pratson, L. F., Halpin Urban, D. L., & Pratson, L. F. (2008). Modeling population connectivity by ocean currents, a graph-theoretic approach for marine conservation. Landscape Ecol, 23(SUPPL. 1), 19–36

5 Quantitative and qualitative ways to compare indicators/assessment methods (ToR C)

Introduction

There is a wide variety of assessment methods being used to assess the state of seafloor habitats. These range from approaches with global implementation (e.g. Mazor *et al*. 2020, Pitcher *et al*. 2022, Borja *et al*. 2009), to approaches that are more regionally oriented (e.g. HELCOM's CumI or OSPAR's BH3) and/or (currently) developed for specific regions, e.g. SoS, GBPI, mTDI, pTDI (Serrano *et al*. 2022, Jac *et al*., 2020).

The assessment methods can roughly be classified into two different types. The first type are methods that make a risk assessment of impact from human pressure. This is typically based on an expert judgement and/or a quantitative meta-analysis how human pressure links to seafloor impact. These risk-based methods can help guide the choice of management measures needed to meet sustainability objectives and/or identify areas/seabed habitats that are most at risk (Pitcher *et al*. 2020). The second type are methods that evaluate the change in an area using empirical observations, typically using long-term monitoring data (e.g. Gislason *et al*. 2017). Both risk and sampling-based assessment methods can evaluate the same indicator, e.g. benthic species diversity, density or community biomass.

Each individual method, and associated indicator, is necessarily an incomplete simplification of the natural world. Methods may have been developed with different assumptions, structures, and processes, whereas indicators may focus on different components of the benthic ecosystem. It remains therefore unclear whether the different risk-based approaches currently available in EU waters find the same type of areas/seabed habitats most at risk from human activities such as bottom fishing. Additionally, it remains unclear whether the empirical observations support the risk-based findings.

This chapter will review and document options on how to evaluate and compare the suitability and performance of indicators/assessment methods. This will both include comparisons of the sampling-based methods as well as of seafloor sensitivity and impact/state. Worked examples for specific areas are included to ensure that the outcome can be used in WKBENTH3 for a further comparison and evaluation of indicators/assessment methods.

5.1 Evaluating sampling-based methods

Recent benthic indicator developments have resulted in several scientific papers where sampling-based methods have been compared against bottom trawling. We review three of these papers (section 5.1.1). Considering their results, WKBENTH2 decided to collate a standardized dataset to evaluate the specificity, sensitivity and/or responsiveness of the different sampling-based indicator methods to pressure gradients, predominantly bottom fishing disturbance, the dominant abrasion pressure in EU waters (ICES, 2019) (see section 5.1.2).

The collated dataset will be sent to experts end of June 2022. Experts will have until 31st of August to estimate indicator values. The outcomes will be quality checked by 20th of September and analysed at WKBENTH3 (see section 5.1.3).

5.1.1 Indicator evaluations in the literature

Serrano *et al*. 2022

Which indicators were evaluated?
In the work, we evaluated the SoS indicator across 6 case studies and we compared its performance with other indicators usually applied in the literature: Total Biomass or Total density when biomass was not available (case studies B and E), Shannon diversity index and Margalef index (in the case study D, Margalef index was replaced by Richness because the unavailability of density data).

What data?
We used a variety of datasets, depending on the case study:

1. Case studies A1 and A2 - We used data from the Spanish IBTS for the period 2013-2019 sampled using a scientific otter trawl (ICES, 2017). We also used the MSFD-Broad habitat map from Eunis (https://emodnet.ec.europa.eu/en/seabed-habitats) and VMS data from the Spanish government. A1 used samples from Offshore circalittoral sand whereas A2 used data sampled on Upper bathyal sediment.
2. Case study B, Ría de Vigo - We used data from endobenthos, sampled using a modified BOUMA box-corer with a sampling area of 0.0175 m^2. In addition, particle size, organic matter, heavy metals and other pollutants were also quantified. Details on the precise methods used can be consulted in Beiras *et al*. (2012). This case study was the only used to assess the performance of SoS measuring a pressure different to trawling impact (pollution).
3. Case study C: South-West Deeps, West Marine Conservation Zone - Data collected on the target MSFD broad habitat 'offshore circalittoral sand' was analysed using 101different box-corer samples distributed across a gradient of trawling effort across a narrow depth range, from 130 to 172 m depth. Biological communities were sampled using a mini Hamon grab, with a sampling area of 0.1 m^2. The biological data and the pressure associated to each habitat was provided by JNCC.
4. Case study D. Flemish Cap - Data from the Flemish Cap area (a high-seas zone off the Canadian coast) located at depths ranging from 600 to 1300 m (MSFD broad habitat 'mid bathyal sediments') was used to assess SoS performance. We used data from the 2007 EU Flemish Cap bottom-trawl research survey (Durán Muñoz, *et al*., 2020), using standardised sets of a Lofoten bottom trawl (with a swept area of ≈0.04 km^2 each) following a depth-stratified sampling design (see Murillo *et al*., 2016; 2020 for more information). Trawling effort was estimated using VMS data from international fisheries and calculated as the sum of pings by cell and year (can be translate into hous/km^2).
5. Case study E. Seco de los Olivos seamount - Species data were obtained from three ROV (Seaeye Falcon & FalconDR) surveys conducted by OCEANA on board the Oceana Ranger between 2010 and 2012 (for more information about the sampling area or method see de la Torriente *et al*., 2018, 2019, 2020). The sampling unit consisted of 1-minute continuous movement ROV tracks at a speed of 0.2–0.4 knots, covering an average distance of 13 m (mean = 13.16 ± 5.74 SD). The final data set selected for analysis was composed by 86 samples located in the target MSFD broad habitat (upper bathyal sediment) across a trawling effort gradient. For the trawling effort data, we used data supplied by the Spanish government.

What method was used for evaluation?
The SoS performance was compared against the other methods by computing the correlation between each indicator and the pressure values using Spearman correlation (Table 5.1).

Furthermore, the values of each indicator across the pressure gradient were visually shown using boxplots (Figure 5.1).

Outcome?

In each scenario, the SoS indicator was compared to the Shannon-Wiener diversity index, Margalef index and total biomass, being the only metric, which showed the expected significant negative response to pressure in all cases. Our results show that SoS was highly effective in assessing benthic habitats status under both physical and chemical pressures, regardless of the sampling gear, the habitat, or the case study, showing a great potential to be a useful tool in the management of marine ecosystems.

Table 5.1. Correlation values of the four tested metrics for each case study: proportion of sentinel species, total biomass (or total density when biomass not available), Shannon-Wiener index and Margalef index (replaced by species richness in Flemish Cap).

CASE STUDY/HABITAT	VARIABLE	rho	p-value
A1) DEMERSALES: Offshore Circalitoral Sand	Proportion of sentinel species	-0.24	0.006
	Total biomass(kg/km^2)	-0.25	0.003
	Shannon index	0.00	0.984
	Margalef index	0.09	0.293
A2) DEMERSALES: Upper Bathyal Sediment	Proportion of sentinel species	-0.58	<0.001
	Total biomass (kg/km^2)	0.10	0.061
	Shannon index	-0.44	<0.001
	Margalef index	-0.49	<0.001
B) Ría de Vigo: Infralitoral Mud	Proportion of sentinel species	-0.76	<0.001
	Total density (ind/km^2)	-0.72	<0.001
	Shannon index	-0.72	0.001
	Margalef index	-0.76	<0.001
C) South-West Deeps: Offshore Circalitoral Sand	Proportion of sentinel species	-0.22	0.036
	Total biomass (kg/km^2)	0.25	0.029
	Shannon index	0.22	0.026
	Margalef index	0.13	0.204
D) Flemish Cap: Mid Bathyal Sediment	Proportion of sentinel species	-0.49	0.011
	Total biomass (kg/km^2)	-0.60	0.001
	Species Richness	-0.61	0.001
	Shannon index	-0.46	0.018
E) Seco de los Olivos:	Proportion of sentinel species	-0.55	<0.001

Upper Bathyal Sediment	Total density(ind/km^2)	-0.66	<0.001
	Margalef index	-0.15	0.160
	Shannon index	-0.31	0.003

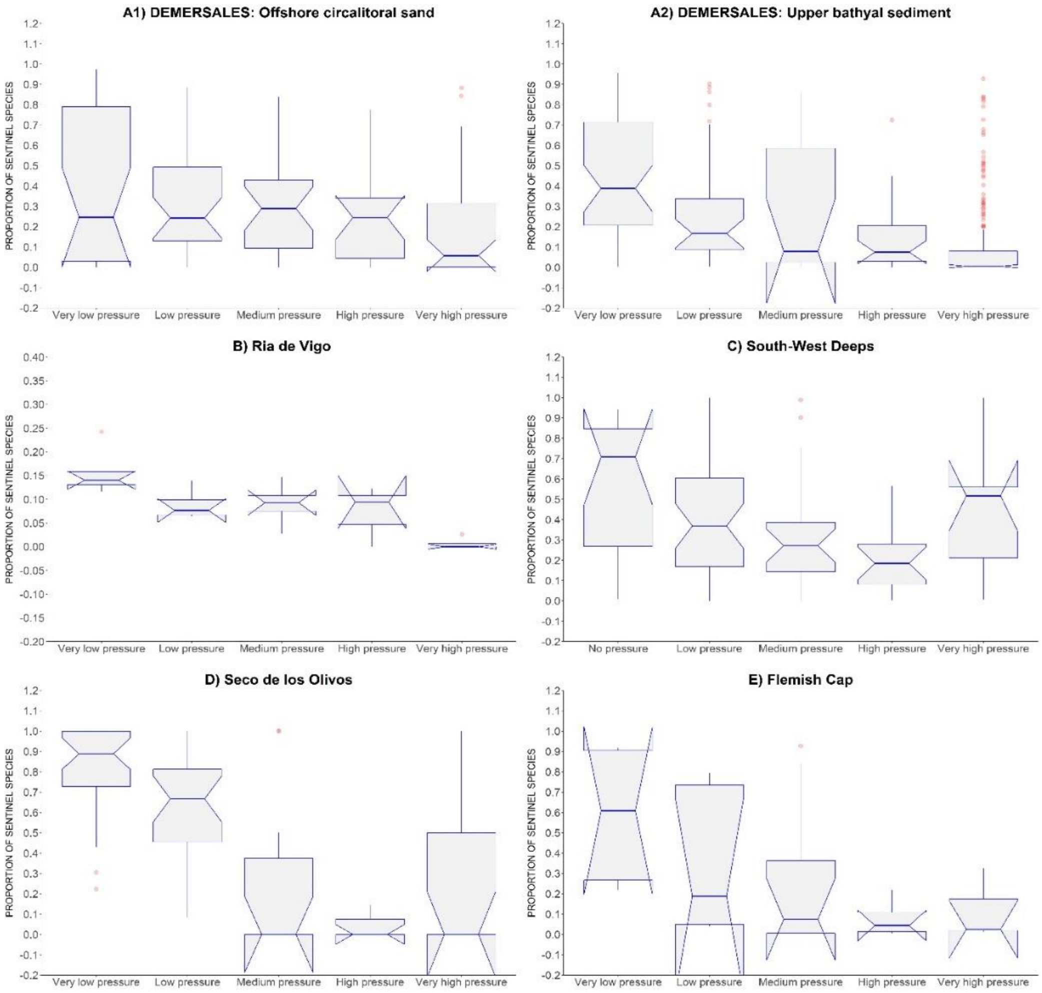


Figure 5.1. Proportion of sentinel species (y-axis) across the pressure gradient (x-axis) by case study A1) DEMERSALES offshore circalittoral sand, A2) DEMERSALES upper bathyal sediment, B) Ría de Vigo, infralittoral mud (Pollution),C) UK Waters, offshore circalittoral sand, D) Flemish Cap, mid bathyal sediment, and E) Seco de los Olivos seamount, upper bathyal sediment. The boxes represent the interquartile range (IQR), the line is the median and the notches are its confidence interval. The lines of the whiskers extend 1.5 IQR and outliers are identified as points beyond the whiskers.

Hiddink *et al*. 2020

Using a systematic review methodology, we collated data from 41 studies that compared the benthic biota in trawled areas with those in control locations in a meta-analysis (that were either not trawled or trawled infrequently), examining 7 potential indicators (numbers and biomass for individual taxa and whole communities, evenness, Shannon-Wiener diversity and species richness) to assess their performance against a set of 9 criteria (concreteness, theoretical basis, public awareness, cost, measurement, historical data, sensitivity, responsiveness, specificity).

The effects of trawling were stronger on whole-community numbers and biomass than for individual taxa. Species richness was also negatively affected by trawling but other measures of diversity were not. Community numbers and biomass met all criteria, taxa numbers and biomass and species richness satisfied a majority of criteria, but evenness and Shannon-Wiener diversity did not respond to trawling and only met few criteria, and hence are not suitable state indicators of the effect of bottom trawling. An evaluation of each candidate indicator against a commonly agreed suite of desirable properties coupled with the outputs of our meta-analysis showed that whole-community numbers of individuals and biomass are the most suitable indicators of trawling impacts as they performed well on all criteria. Particular strengths of these indicators are that they respond strongly to trawling, relate directly to ecosystem functioning, and are straightforward to measure. Evenness and Shannon-Wiener diversity are not responsive to trawling and unsuitable for the monitoring and assessment of bottom trawl impacts.

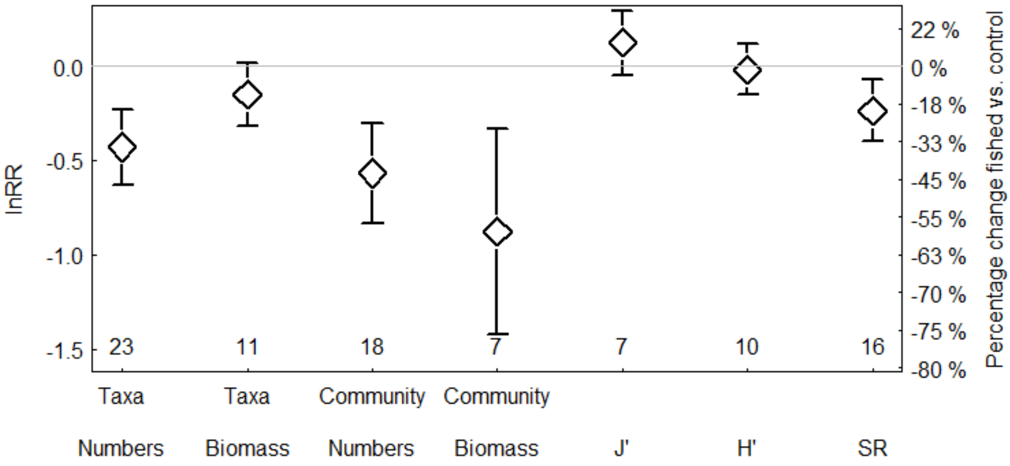


Figure 5.2. Mean response to trawling (lnRR) and 95% confidence intervals for the indicators. If the confidence interval overlaps 0 the effect was not significant. N (= number of studies reporting on each indicator) is given under each bar. The right-hand axis gives % changes for ease of interpretation. J': evenness, H': Shannon-Wiener diversity index, SR: species richness. Responses for taxa indicate the mean of the responses of the individual taxa that were reported in the studies.

Jac *et al.* 2020

Which indicators were evaluated?

Fifteen indices were investigated: taxonomic diversity metrics, functional diversity indices and functional indices, the two later based on sensitivity traits to physical abrasion (size, position, feeding, mobility, fragility).

Total community biomass and five common taxonomic diversity indices were calculated: species richness (S, the total number of taxon), Margalef index (Margalef, 1958), Shannon diversity (H', Shannon and Weaver, 1963), Pielou evenness (J', Pielou, 1969) and Simpson index (λ, Simpson, 1949). Functional Richness (FRic, Cornwell *et al.*, 2006; Villéger *et al.*, 2008), Functional Specialization (Fspe; Bellwood *et al.*, 2006, Villéger *et al.*, 2010), Functional Evenness (FEve; Mason *et al.*, 2005) and Functional Divergence (FDiv; Mason *et al.*, 2005) were investigated using the species-traits matrix mentioned above. Functional sensitivity indices, designed to detect particular impacts on communities were also computed. In contrast to functional diversity indices for which each trait level is given equal weight, semi-quantitative trait scoring indicates the potential sensitivity of each species to a given pressure. Functional sensitivity indices therefore integrate this scoring in their calculation. The tested indices were: AZTI Marine Biotic Index (AMBI; Borja *et* *al.*, 2000), Trawling Disturbance Index (TDI; de Juan and Demestre 2012), modified TDI (mTDI, Foveau *et al.*, 2017), partial TDI (pTDI) and the modified vulnerability Index (mT; modified from Certain *et al.*, 2015). TDI-derived indices were developed specifically to detect trawling impact, while mT is issued from a general framework allowing to address any pressure as long as specific sensitivity traits were available to detect it.

What data?
Benthic invertebrate fauna's samples, considered as bye-catch, were opportunistically collected and monitored during four scientific bottom trawl surveys. For each survey, species were sorted, identified, counted and weighed. Only the biomass data (g.km-2) were used in this approach to account for colonial species that could not be counted. Commercial species and cephalopods were removed from the analyses. A data filtering and aggregation procedure was proposed to avoid mis-identification errors and heterogeneous taxonomic expertise over the available serie.

1. Mediterranean - French MEDITS data for the period 2012-2018 were used, distinguishing two study areas, the Gulf of Lion and the Eastern shelf of Corsica. International SAR data (2009-2017) were computed from VMS data and vessel size. The 90th percentile value over the whole period was used as no significant change in the effort distribution could be detected.
2. English Channel - Three scientific trawling surveys were used: French North Sea IBTS (2009-2018), French Channel Ground Fish Survey (2008-2018) and CAMANOC survey (2014). International SAR data (2009-2017) were downloaded in March 2019 through OSPAR website (https://www.ospar.org). The 90th percentile value over the whole period was used as no significant change in the effort distribution and value could be detected.
3. Southern North Sea - The French North Sea IBTS (2009-2018) was used. International SAR data (2009-2017) were downloaded in March 2019 through OSPAR website (https://www.ospar.org). The 90th percentile value over the whole period was used as no significant change in the effort distribution and value could be detected.

What method was used for evaluation?
Within each of the four study areas, the properties of each indices, such as their capacity to detect trawling effect (spearman correlation, index of difference in spatial pattern between index and abrasion values), their statistical behaviour (skewness and kurtosis) or their ability to inform on community structure (percentage of variance of the community structure explained by RDA), were investigated. In order to simplify the assessment of all indices properties, a qualitative scoring scheme was used. Once a total score per index was computed, indices could be ranked according to their performance and those with the highest score were selected (Table 5.2). Boxplot of indices values across abrasion classes and interpolated maps of the best performing indices were also provided (Figure 5.3 and 5.4).

Outcome?
The evaluation of the efficiency of the 15 different indices showed the necessity to use indices specific to trawling to detect its effect on benthic habitat in these five very contrasted regions. Also, their detection power seemed limited in areas with low abrasion gradients (Corsica). Fours indices specific to fishery effect detection based on biological traits appeared to be the best performing benthic indices regarding these requirements: Trawling Disturbance Index (TDI), modified-Trawling Disturbance Index (mTDI), partial-Trawling Disturbance Index(pTDI), modified sensitivity index (mT). However, their detection power varied geographically and although closely related, it seems difficult to select and recommend only one of them. In conclusion, to monitor the effect of trawling on benthic communities in all European waters, these indices would need to be systematically screened and the locally most suitable one chosen for impact assessment.

Table 5.2. Results of spearman correlation tests and spatial correlation index for each index in the four studied areas.

Indices	Spearman correlation test					Spatial correlation (Lee index)				
	GoL	Corsica	IBTS	CGFS + CAM	Score	GoL	Corsica	IBTS	CGFS + CAM	Score
Community biomass	-0.15**	-0.05	0.001	0.26***	1.28	0.71	0.58	0.76	0.66	2.37
Species richness	0.13*	0.31**	0.10**	0.09**	1.77	0.31	0.44	0.49	0.42	3.82
Margalef index	0.13*	0.32**	0.10**	0.11**	1.86	0.31	0.44	0.49	0.42	3.82
Shannon index	0.24**	0.13	0.01	-0.02	1.10	0.32	0.47	0.52	0.43	3.65
Pielou Index	0.21**	0.07	-0.10**	-0.11**	1.40	0.32	0.48	0.53	0.43	3.62
Simpson index	0.19**	0.15	-0.01	-0.03	1.01	0.32	0.46	0.52	0.42	3.61
FRic	0.19**	0.36**	0.09*	0.15**	2.44	0.38	0.58	0.58	0.56	3.02
FDiv	0.08	-0.01	-0.14*	-0.21**	1.00	0.30	0.54	0.56	0.41	3.55
FEve	0.21**	-0.07	0.01	-0.11**	1.11	0.30	0.56	0.57	0.42	3.50
FSpe	0.14**	0.28*	0.03	0.05	1.41	0.30	0.53	0.54	0.41	3.60
AMBI	-0.26**	0.36**	0.003	-0.08**	1.99	0.42	0.57	0.58	0.52	3.02
TDI	-0.33**	0.07	-0.35***	-0.34***	3.08	0.32	0.56	0.57	0.47	3.34
mTDI	-0.31**	-0.08	-0.28**	-0.32***	2.80	0.31	0.52	0.53	0.42	3.59
pTDI	-0.26**	-0.09	-0.35***	-0.37***	3.01	0.30	0.72	0.72	0.61	2.87
mT	-0.34**	-0.01	-0.22**	-0.30***	2.47	0.29	0.50	0.50	0.37	3.86

GoL = Gulf of Lion. CGFS + CAM = CGFS and CAMANOC surveys * indicates that P < 0.05 ; ** indicates that P < 0.01 ; *** indicates that P < 0.001; ns indicates no significant difference. Grey shading indicates best scores

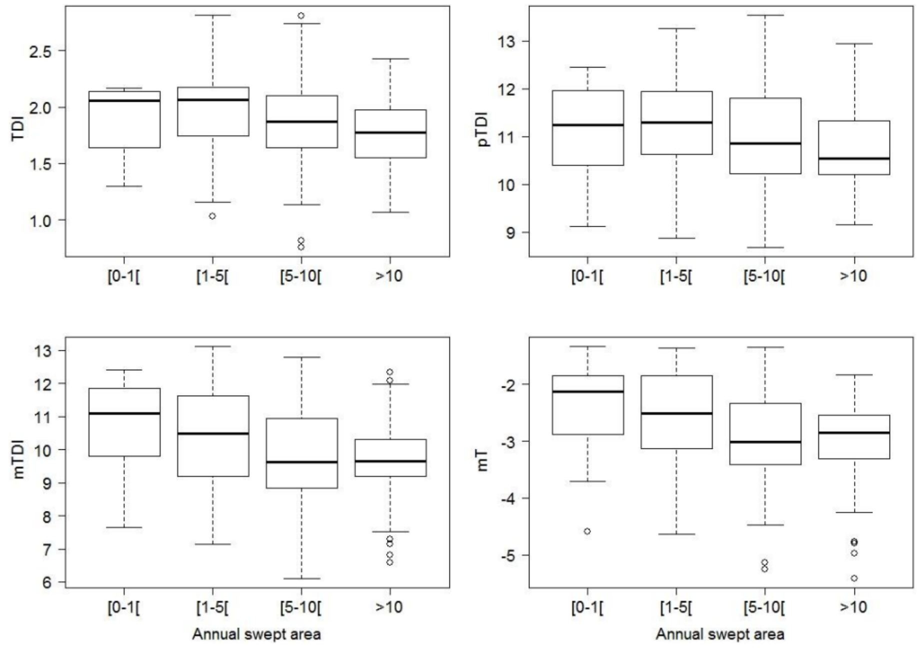


Figure 5.3. Values of the four selected indices by class of abrasion in the Gulf of Lion.

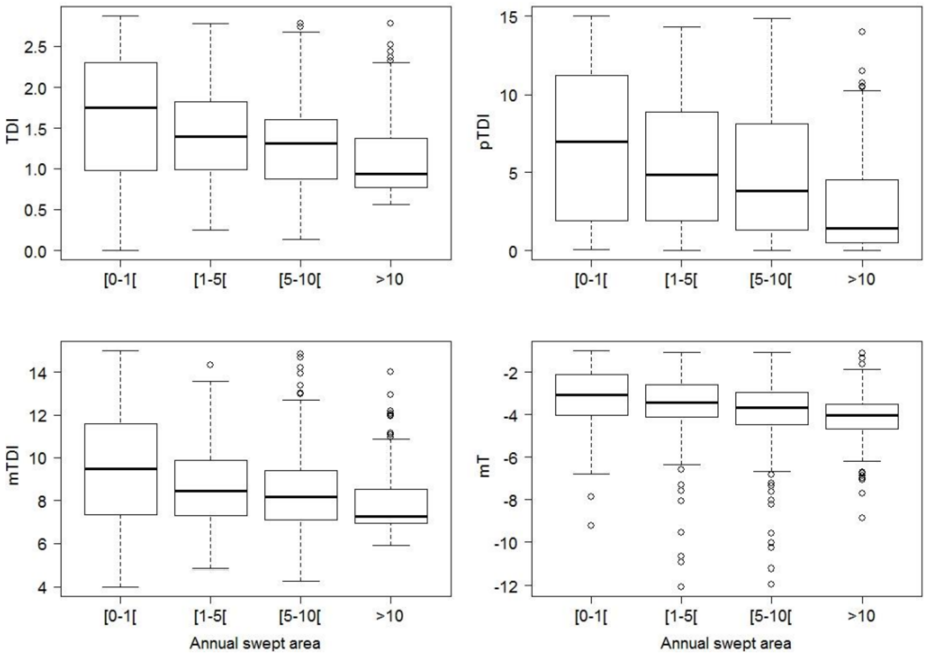


Figure 5.4. Values of the four selected indices by class of abrasion for the CGFS and CAMANOC data (English Channel).

5.1.2 Collating trawling gradients for a standardized comparison

WKBENTH2 identified 14 benthic datasets over gradients of commercial bottom trawling intensity, 2 benthic datasets over gradients of eutrophication and 1 over pollution. The benthic data consists of abundance and biomass per species (Table 5.3). This information will allow experts to calculate most of the currently available sampling-based indicators. One of the datasets is near the Flemish Cap (Canada). This dataset outside EU waters is included as most of the other datasets are in relatively shallow waters (Table 5.3).

The aim in WKBENTH3 is to evaluate change in benthic indicator values along the gradients and determine which indicators are most responsive to bottom trawling disturbance. Such a comparative analysis can result in differences in community composition that seem to be related to the pressure, while in fact the pressure gradient varies with environmental conditions. We therefore selected datasets that have limited variation in environmental conditions within each gradient. For each of these gradients, it is expected that the effects of trawling/eutrophication/pollution on benthic communities will have a larger impact than the environmental conditions.

5.1.3 Analysis in WKBENTH3

Indicator values will be compared by computing the correlation between each indicator and the (log-transformed) pressure values using Spearman correlation. The values of each indicator will also be plotted along the pressure gradients. Lastly, and only for all trawling gradients, we will calculate the mean response to trawling across all locations (similar to Figure 5.2). This response is estimated by calculating the change in indicator values from low versus high trawl disturbed stations at each location.

Table 5.3 Identified benthic datasets to be evaluated in WKBENTH3.

Ecoregion	Location	Sampling device	Pressure gradient	Depth (m)	Sediment type
Baltic Sea	Gotland region[1]	Van Veen grab	Trawling	37-59	Muddy sand
	Southern Baltic Sea[2]	Box core	Trawling	70-85	Sand
	Gulf of Finland**	xx	Eutrophication		
North Sea	Dutch EEZ coarse sediment[3]	Box corer	Trawling	22–36	Sand
	Thames[3]	Box corer	Trawling	16–40	Sand
	Silver Pit[3]	Box corer	Trawling	68–78	Muddy sand
	Fladen Ground[3]	Day grab	Trawling	143–153	Mud
	Dogger Bank[3]	Hamon grab	Trawling	25–30	Sand
	Long Forties[3]	Hamon grab	Trawling	74–83	Gravelly sand
Celtic Seas	Sellafield, Irish Sea[3]	Day grab	Trawling	21-42	Muddy sand
Iberian Coast	Offshore circalittoral sand[4]	Otter trawl	Trawling	71-202	Sand
	Upper bathyal sediment[4]	Otter trawl	Trawling	200-500	Mud
	Ria de Vigo[4,5]	Grab	Pollution	<30	Mud
Canada	Flemish Cap mid-bathyal sediment[4]	Otter trawl	Trawling	786-1236	-
Western Med Sea	Gulf of Lion**	Otter trawl	Trawling		
Adriatic Sea	To be determined**	Otter trawl	Trawling		
Ionian/ Aegean	Greek case study (coastal) **	xx	Eutrophication		

** to be confirmed / determined

1 see van Denderen *et al*. 2020 and references therein

2 see van Denderen *et al*. in press.

3 see van Denderen *et al*. 2015 and references therein

4 see Serrano *et al*. 2022 and references therein

5 see Beiras *et al*. 2012

5.2 Evaluating risk-based methods

Countries, Regional Sea Conventions and the ICES working group FBIT are developing risk-based approaches to assess the state of benthic habitats for MSFD D6 purposes.

Most risk-based methods that evaluate seafloor integrity have an underlying data layer that describes benthic sensitivity to bottom trawling (or any type of seafloor abrasion), where sensitivity varies with environmental conditions and/or habitat types, as well as a prediction of benthic impact (Figure 1.1). The WK decided that WKBENTH3 could evaluate these patterns of sensitivity and impact for regions where multiple methods are available (Table 5.4).

Table 5.4. Regions with overlap between risk-based approaches

Ecoregion	Spatial coverage	Indicator methods
North Sea	Entire North Sea	BH3, L1, L2, PD2, Margalef diversity (BH2 of OSPAR), BH4 of OSPAR
	Southern North Sea	TDI
	English Channel	TDI
Baltic Sea	Entire Baltic Sea	CumI, PD, L1
Bay of Biscay and the Iberian Coast	Iberian Coast	BH3, SOS (BH1 of OSPAR), PD
Ionian/ Aegean**	Greek EEZ	PD, WFD indicators
** analysis is underway and can potentially be included		

The WK suggested that an evaluation of risk-based approaches can be realised by visually comparing maps and through a ranked score per MSFD habitat type and subdivision (or EEZ). The ranking is helpful because it limits the need to standardize and/or calibrate outputs across approaches. The ranked score per habitat type will allow WKBENTH3 to examine if the risk-based approaches find the same habitat types most sensitive to physical disturbance and/or most at risk of adverse effects (section 5.2.1).

Since the approaches have different underlying units/ assumptions (e.g. sensitivity may be determined from benthic longevity, other biological traits and/or expert judgement), the WK decided not to compare the actual values across approaches or include an ensemble approach to combine the predictions of the methods.

For a standardized impact evaluation, it would have been preferable to re-run all methods with the same underlying fishing pressure layer ahead of WKBENTH3. However, this will be difficult for some methods given the available time. It was therefore decided to use the impact maps as they are, with the caveat that impact results may differ due to a different spatial distribution of abrasion pressures.

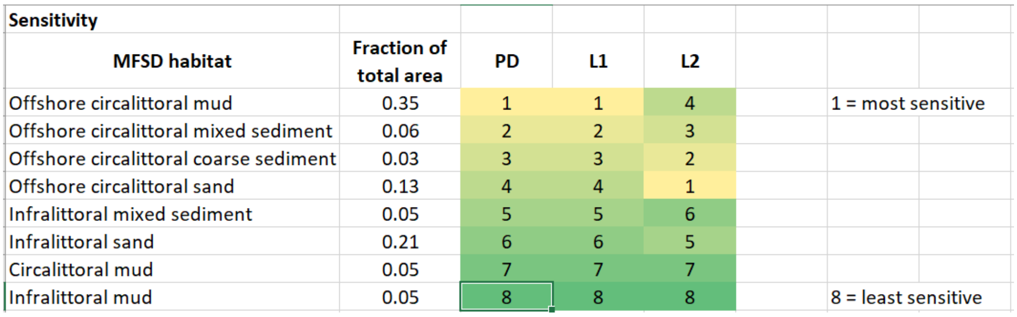
A more detailed comparison of several risk-based approaches is currently under development for the UK EEZ. This work was presented in the workshop and is summarized in section 5.2.2.

5.2.1 Evaluating sensitivity and impact

Evaluation of the different risk-based approaches will be done in WKBENTH3 by visually comparing the sensitivity/impact maps as well as by evaluating ranked sensitivity/impact scores per MSFD habitat type and subdivision (or EEZ). The ranked scores will show if the different risk-based approaches currently available in EU waters find the same type of areas/seabed habitats most sensitive to bottom fishing (see Figure 5.5) and/or most at risk of adverse effects (see Figure 5.6).

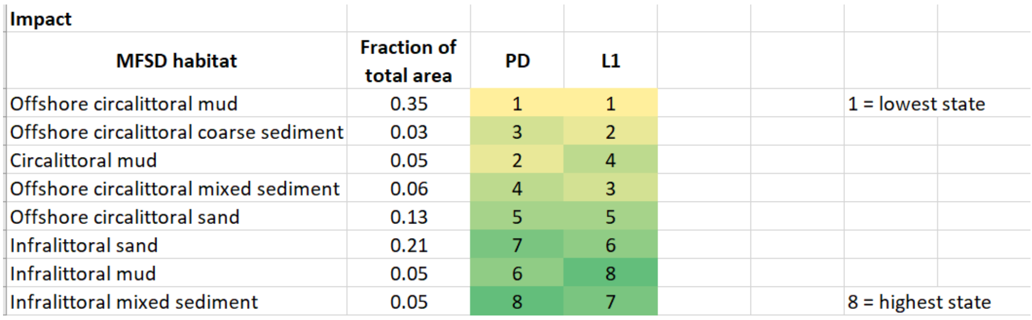
Experts will be contacted by mid-July and will be asked to provide spatial data layers and/or a scored sensitivity and impact per MSFD habitat type and subdivision (or EEZ). The outcomes will be quality checked by 20th of September and analysed at WKBENTH3.

The analysis will make use of the latest EMODNET EUSeaMap habitat data (Vasquez *et al*. 2021). In the absence of subdivisions agreed by Member States, the analysis will use an indicative set of 22 subdivisions following ICES advice (2021). Polygon shapefiles of these subdivisions are available: https://github.com/ices-eg/WKTRADE3/tree/main/5%20-%20Output/Subdivions%20shapefiles.



Sensitivity						
MFSD habitat	Fraction of total area	PD	L1	L2		
Offshore circalittoral mud	0.35	1	1	4		1 = most sensitive
Offshore circalittoral mixed sediment	0.06	2	2	3		
Offshore circalittoral coarse sediment	0.03	3	3	2		
Offshore circalittoral sand	0.13	4	4	1		
Infralittoral mixed sediment	0.05	5	5	6		
Infralittoral sand	0.21	6	6	5		
Circalittoral mud	0.05	7	7	7		
Infralittoral mud	0.05	8	8	8		8 = least sensitive

Figure 5.5 Example of sensitivity score per MSFD habitat type in the Kattegat subdivision for three different indicators (note that PD and L1 use the same underlying sensitivity layer).



Impact					
MFSD habitat	Fraction of total area	PD	L1		
Offshore circalittoral mud	0.35	1	1		1 = lowest state
Offshore circalittoral coarse sediment	0.03	3	2		
Circalittoral mud	0.05	2	4		
Offshore circalittoral mixed sediment	0.06	4	3		
Offshore circalittoral sand	0.13	5	5		
Infralittoral sand	0.21	7	6		
Infralittoral mud	0.05	6	8		
Infralittoral mixed sediment	0.05	8	7		8 = highest state

Figure 5.6 Example of impact score from bottom fishing disturbance per MSFD habitat type in the Kattegat subdivision for two different indicators. Fishing pressure layer used is the average annual based on VMS data for the years 2012-2018.

5.2.2 Example of a current project

Risk-based indicator evaluations in the UK EEZ - presented by Liam Matear (JNCC) during the workshop

The aims of the project are to test a suite of indicators, that have been used (or are planned to be used) for national and international reporting, using a common dataset, to evaluate data needs, to evaluate applicability at UK level, to compare results and help address some knowledge gaps identified through the UK Marine Strategy assessments. Within this project, we aim to improve the accuracy and confidence of the assessments of benthic habitats, building upon the lessons-learnt and knowledge gaps identified during the production of the OSPAR Intermediate Assessment, UK Marine Strategy and the Habitats Directive Article 17.

Several benthic indicators were chosen for testing and comparison within the scope of this project; Sentinels of the Seabed (SoS, previously named "Typical species composition BH1"), Benthic Indicator Species Index (BISI), Benthic multimetric index (BENMMI), Infaunal Quality Index (IQI), Population Dynamic 2 (PD2) and Extent of physical damage (BH3).

The project started in 2019 and includes the following tasks:

- Analyse and compare the results across a number of benthic indicators.
- Use multimetric indicators to calibrate and improve the data layers and methods underpinning the model-based indicator, extent of physical damage, exploring how the current method could be changed to improve the confidence of the results, and how additional pressures should be incorporated.
- Explore the proportion of available data which can be used by the different indicators.
- Specify data requirements and applicability of indicators across different habitat types.
- Compare indicator results at multiple scales.
- Test the indicators using additional activities and pressures.

Outputs of the project are forthcoming, with further testing planned to assess additional human activities (aggregate extraction via BH3) and new monitoring data from wider geographical areas and habitats. Further work will also include comparing indicator results at new scales of spatial resolution, including MPA and OSPAR Sub-Regional scales to better understand indicator performance.

References

Beiras, R., Durán, I., Parra, S., Urrutia, M.B., Besada, V., Bellas, J., Viñas, L., Sánchez-Marín, P., González-Quijano, A., Franco, M.A., Nieto, O., González, J.J., 2012. Linking chemical contamination to biological effects in coastal pollution monitoring. Ecotoxicology 21, 9–17, doi: 10.1007/s10646-011-0757-3

De la Torriente, A., Serrano, A., Fernández-Salas, L. M., García, M., Aguilar, R. 2018. Identifying epibenthic habitats on the Seco de los Olivos Seamount: species assemblages and environmental characteristics. Deep-Sea Res. Pt I 135, 9-22, doi: 10.1016/j.dsr.2018.03.015

De la Torriente, A., González-Irusta, J.M., Aguilar, R, Fernández-Salas, L.M., Punzón, A., Serrano, A. 2019. Benthic habitat modelling and mapping as a conservation tool for marine protected areas: A seamount in the western Mediterranean. Aquat. Conserv. 135, 9-22, doi: 10.1002/aqc.3075

De la Torriente, A., Aguilar, R., González-Irusta, J.M., Blanco, M., Serrano, A., 2020. Habitat forming species explain taxonomic and functional diversities in a Mediterranean seamount. Ecol. Indic. 118, 106747, doi = 10.1016/j.ecolind.2020.10674

Durán Muñoz, P., Sacau, M., García-Alegre, A., Román, E., 2020. Cold-water corals and deep-sea sponges by-catch mitigation: Dealing with groundfish survey data in the management of the northwest Atlantic Ocean high seas fisheries. Mar. Policy 116, 103712; doi: 10.1016/j.marpol.2019.103712

Hiddink, J.G., Kaiser, M.J., Sciberras, M., McConnaughey, R.A., Mazor, T., Hilborn, R., Collie, J.S., Pitcher, R., Parma, A.M., Suuronen, P., Rijnsdorp, A.D. & Jennings, S. (2020) Selection of indicators for assessing and managing the impacts of bottom trawling on seabed habitats. Journal of Applied Ecology, 57, 1199-1209.

ICES (2017) Manual of the IBTS North Eastern Atlantic Surveys.

ICES. 2019. EU request to advise on a seafloor assessment process for physical loss (D6C1, D6C4) and physical disturbance (D6C2) on benthic habitats. In Report of the ICES Advisory Committee, 2019. ICES Advice 2019, sr.2019.25, https://doi.org/10.17895/ices.advice.5742

Jac, C., Desroy, N., Certain, G., Foveau, A., Labrune, C. and Vaz, S., 2020. Detecting adverse effect on seabed integrity. Part 2: How much of seabed habitats are left in good environmental status by fisheries?. Ecological Indicators, 117, p.106617.

Murillo FJ, Kenchington E, Lawson JM, Li G, Piper DJW (2016) Ancient deep-sea sponge grounds on the Flemish Cap and Grand Bank, northwest Atlantic. Mar Biol 163:63

Murillo FJ, Weigel B, Bouchard Marmen M, Kenchington E (2020) Marine epibenthic functional diversity on Flemish Cap (north-west Atlantic)—Identifying trait responses to the environment and mapping ecosystem functions. Divers. Distrib:ddi.1302

Serrano, A., de la Torriente, A., Punzón, A., Blanco, M., Bellas, J., Durán-Muñoz, P., ... & González-Irusta, J. M. (In Press). Sentinels of Seabed (SoS) indicator: Assessing benthic habitats condition using typical and sensitive species. https://doi.org/10.1016/j.ecolind.2022.108979

Vasquez M, Allen H, Manca E, Castle L, Lillis H, Agnesi S, Al Hamdani Z, Annunziatellis A, Askew N, Bekkby T, Bentes L, Doncheva V, Drakopoulou V, Duncan G, Gonçalves J, Inghilesi R, Laamanen L, Loukaidi V, Martin S, McGrath F, Mo G, Monteiro P, Muresan M, Nikilova C, O'Keeffe E, Pesch R, Pinder J, Populus J, Ridgeway A, Sakellariou D, Teaca A, Tempera F, Todorova V, Tunesi L, Virtanen E (2021). EUSeaMap 2021. A European broad-scale seabed habitat map. D1.13 EASME/EMFF/2018/1.3.1.8/Lot2/SI2.810241– EMODnet Thematic Lot n° 2 – Seabed Habitats EUSeaMap 2021 - Technical Report. https://doi.org/10.13155/83528

Van Denderen, P.D., Bolam, S.G., Hiddink, J.G., Jennings, S., Kenny, A., Rijnsdorp, A.D. and Van Kooten, T., 2015. Similar effects of bottom trawling and natural disturbance on composition and function of benthic communities across habitats. Marine Ecology Progress Series, 541, pp.31-43.

van Denderen, P.D., Bolam, S.G., Friedland, R., Hiddink, J.G., Noren, K., Rijnsdorp, A.D., Sköld, M., Törnroos, A., Virtanen, E.A. and Valanko, S., 2020. Evaluating impacts of bottom trawling and hypoxia on benthic communities at the local, habitat, and regional scale using a modelling approach. ICES Journal of Marine Science, *77*(1), pp.278-289.

Van Denderen PD, Törnroos A, Sciberras M, Hinz H, Friedland R, Lasota R, Mangano MC, Robertson C, Valanko S, Hiddink JG (In Press). Effects of bottom trawling and hypoxia on benthic invertebrate communities https://doi.org/10.3354/meps14094

6 Compilation of assessment methods/indicators (ToR D)

ICES is requested to provide a detailed review of indicators used, or under development, by Regional Sea Conventions (RSCs), Member States and ICES, for assessing the state/condition of seabed habitats and relevant existing literature. The review should specify the input data, how data is processed, the parameters of habitat quality used, how quality is quantified, any threshold values used, the applicable seabed (habitat) and pressure types, how the output is expressed, and how confidence and uncertainty are handled.

The workshop will develop a framework to compile the required information on seabed indicators / assessment methods to enable their review and evaluation, and provides the opportunity to contribute to the subsequent compilation.

6.1 Development of information extraction table

To obtain information on benthic indicators and assessment methods in an objective and consistent manner, an information extraction table was developed. A first draft of such a framework, based on an initial draft of Daniel van Denderen and supplemented by Esther and Karin, was presented at the first plenary session of WKBENTH2. Subgroup D then adjusted the table based on recommendations from WKBENTH2 participants, after which all participants were given the opportunity to contribute outside of the (physical) meeting. A dedicated meeting with the lead of Subgroup A (David Reid) ensured that the established ‘criteria for indicators’ were integrated in the table. This then led to the final information extraction table shown in Table 6.1.

Table 6.1. The final information extraction table as used in the benthic indicator review.

Indicator description	**Indicator name**				
	Indicator description				
	Type of indicator	□ Model	□ Empirical-based		□ Pressure
	Pressure assessed				
	Human activity				
	MSFD criteria / descriptor				
	How does the indicator relate to benthic biological diversity?				
	How does the indicator relate to benthic community structure and function?				
	Indicator status	□ Under development	□ Applied for MSFD		□ Applied for other management, if so, for what:
	Regions with operational assessments				
Input data	Biological data input *(e.g. monitoring program, time series, sampling method)*				
	Targeted organisms	□ Infauna	□ Epifauna	□ Demersal fish	□ Other: ….
	Environmental data input *(e.g. empirical/modelled, source, time series)*				
	Pressure data input *(e.g. time series, empirical/modelled, source, national/international)*				
	Data availability				
Methodology	Parameters determined from biological data *(e.g. Species richness, abundance, biomass community, Shannon Weaver, Simpson, sensitivity classes)*				
	Parameters determined from pressure data *(e.g. total SAR, years not fished, trawling interval)*				
	Algorithm type				

Indicator description	**Indicator name**			
	Indicator description			
	Type of indicator	□ Model	□ Empirical-based	□ Pressure
	Pressure assessed			
	Human activity			
	MSFD criteria / descriptor			
	How does the indicator relate to benthic biological diversity?			
	How does the indicator relate to benthic community structure and function?			
	Indicator status	□ Under development	□ Applied for MSFD	□ Applied for other management, if so, for what:
	Regions with operational assessments			
	List of categorical information (Presence/Absence, …) *Direct measurements (counts, areas, concentrations, …)* *Single or multimetric indicators using basic arithmetics* *Indicators using multivariate and complex statistics* *Indicators derived from modelling approaches* *Indicators reporting on trends*			
	References for state - pressure relation			
	Uncertainty estimation methodology			
	Coding availability *(e.g. scripts, GitHub)*			
	Threshold present			
	Threshold methodology			
Output	Output variable type	□ Continuous	□ Categorical	□ Proportional
	Output variable range / classes			

Indicator description	**Indicator name**			
	Indicator description			
	Type of indicator	□ Model	□ Empirical-based	□ Pressure
	Pressure assessed			
	Human activity			
	MSFD criteria / descriptor			
	How does the indicator relate to benthic biological diversity?			
	How does the indicator relate to benthic community structure and function?			
	Indicator status	□ Under development	□ Applied for MSFD	□ Applied for other management, if so, for what:
	Regions with operational assessments			
	Output availability *(e.g. report, website, reference)*			
	Uncertainty handling *(e.g. present confidence interval)*			
	Spatial resolution *(e.g. grid cell size, habitat level)*			
	Temporal resolution			
	Seabed habitat levels presented?			
More info	Indicator lead person			
	Indicator data contact			
	References / Literature / Project websites			

6.2 Compiling a list of indicators / assessment methods

6.2.1 Existing compilations of benthic indicators

A suite of benthic habitat indicators was compiled by ICES in 2015 for the Workshop on guidance for the review of MSFD Descriptor 6 seafloor integrity II (WKGMSFDD6-II) from various sources (Member States, Regional Seas Conventions and projects) (ICES, 2015). A similar extensive suite of indicators was compiled by the EU project DEVOTES including for D6 and reported in Teixeira *et al*. 2016. Various tool databases are available[2,3] although their links to MSFD D6 criteria refer to the pre-MSFD 2017 revision. These resources include numerous proposed/non-operational and nationally used indicators originating from the Water Framework Directive (WFD). The actual list of the WFD indicators used by the Member states with their agreed quality thresholds-EQRs is available in the European Communication and Information Resource Centre for Administrations, Businesses and Citizens (CIRCABC) (Commission Decision (EU) 2018/229)[4]. An overview of indicators currently used in coastal waters to assess the Biological Quality Element "benthic invertebrate fauna" under the WFD can be found in Table 6.2. A number of these indicators, including for example AMBI, M-AMBI, are also currently used widely in the MSFD for D6.

Table 6.2. Indicators used under the WFD to assess the Biological Quality Element "benthic invertebrate fauna" in coastal waters.

Region	Country	Indicator
Baltic	Finland	BBI – Finnish Brackish water Benthic Index
	Sweden	BQI – Swedish multimetric biological quality index
	Estonia	ZKI – Estonian coastal water macrozoobenthos community index
	Latvia	BQI – Benthic quality index
	Lithuania	BQI – Lithuanian benthic quality index
	Denmark	DKIv2 – Danish Quality Index version 2
	Germany	MarBIT – Marine Biotic Index Tool
North East Atlantic	Belgium	BEQI – Benthic Ecosystem Quality Index
	Denmark	DKIv2 - Danish Quality Index version 2
	Germany	M-AMBI – Multivariate AZTI's Marine Biotic Index
	France	M-AMBI – Multivariate AZTI's Marine Biotic Index
	Ireland	IQI – Infaunal Quality Index
	Netherlands	BEQI2 – Benthic Ecosystem Quality Index 2

[2] https://mcc.jrc.ec.europa.eu/main/dev.py?N=simple&O=187&titre_page=DevoTool

[3] http://193.204.79.93:3838/SHINY/SHINY_SERVER/ACTIONMEDCATALOGUE/

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018D0229

Region	Country	Indicator
	Norway	NQI – Norwegian Quality Index
	Portugal	BAT – Benthic Assessment Tool
	Spain	M-AMBI – Multivariate AZTI's Marine Biotic Index
	Sweden	BQI – Swedish multimetric Biological Quality Index
	United Kingdom	IQI – Infaunal Quality Index
Mediterranean Sea	Italy	M-AMBI – Multivariate AZTI's Marine Biotic Index
	Slovenia	?
	Cyprus	BENTIX
	France	AMBI – AZTI's Marine Biotic Index
	Greece	BENTIX
	Spain	BOPA – Benthic Opportunistic Polychaetes Amphipods Index
	Spain	MEDOCC – MEDiterranean OCCidental
Black Sea	Bulgaria	M-AMBI(n) – Multivariate AZTI's Marine Biotic Index normalized
	Romania	M-AMBI(n) – Multivariate AZTI's Marine Biotic Index normalized

6.2.2 Compiling a final list for this review

The Technical Group on seabed habitats and seafloor integrity (TG Seabed) recently compiled an overview of quality and extent thresholds currently applied by Member States in relation to the benthic indicators used for the MSFD for D6 (TG 2022, 9th meeting). As such, TG Seabed members were requested to list indicators / assessment methods to be included in this review before WKBENTH2. These were compiled and distributed among the chairs of WKBENTH2 for potential supplements. Hereafter, the list was presented in the plenary session, where all participants were encouraged to contribute with any relevant indicators / assessment methods, and to provide their thoughts on the inclusion / exclusion of the listed ones. Here, it was mentioned that WFD indicators used in assessments of coastal waters may also be included, after which relevant WFD indicators (Table 6.2) not yet listed were added to the list. Additionally, Regional Seas Conventions (HELCOM, OSPAR, SPA/RAC) were contacted for any relevant indicator or assessment in use or currently under development that could be included in the review. Finally, existing compilations mentioned in the previous section were checked to ensure any relevant indicators not yet on the list of indicators to be reviewed were added. Note that the numerous indicators not operational or only weakly developed were not deemed relevant to be currently reviewed.

During WKBENTH2, it was suggested to also include an overview of 'simple' indicators (e.g. species richness, species abundance, the Simpons' index) as these often are incorporated within benthic indicator algorithms. Note however, that it has not been observed that any of these simple indicators has been used by Member States under the MSFD so far.

References

ICES. 2015. Report of the Workshop on guidance for the review of MSFD decision descriptor 6 – seafloor integrity II (WKGMSFDD6-II), 16-19 February 2015, ICES Headquarters, Denmark. ICES CM 2015\ACOM:50. 133 pp. https://doi.org/10.17895/ ices.pub.8500

Teixeira, H., Berg, T., Uusitalo, L., Fürhaupter, K., Heiskanen, A.-S., Mazik, K., *et al*. (2016). A Catalogue of marine biodiversity indicators. Front. Mar. Sci. 3:207. doi: 10.3389/fmars.2016.00207

7 Conclusions and next steps

7.1 Main findings

WKBENTH2 had 64 participants, with an average of 25 active participants during any one day. Participants represented 40 different countries and included all EU waters (Iberian Coast, Celtic Sea, Bay of Biscay, North Sea, Baltic Sea, Mediterranean and the Black Sea). Benthic and policy experts from EU-funded projects, Regional Seas Conventions, and academia.

Specific benthic indicator characteristics are required to evaluate the D6 criteria, D6C3 and D6C5, as they respectively focus more on the risk of adversely affecting the benthic habitat (based on sensitivity and/or evaluation) or determining the benthic habitat status under multiple pressures. Therefore, risk-based and empirical (sample-based) indicator approaches are needed to have an appropriate D6 assessment.

Annual benthic sampling across the EU is sporadic and lacking spatial and temporal coverage. It is difficult to standardize between countries. It is unlikely that sampling will ever be able to provide a representative picture of seafloor status at the regional scale, which the MSFD is required to do its assessment for D6. Therefore, modelling approaches have recently been developed that estimate benthic community sensitivity to various pressure types across regions. These risk-based approaches are able to deliver assessment for MSFD at the regional scale and per broad scale habitats type for D6. WKBENTH2 evaluated a combination of empirical (sampling-based) approaches and risk-based approaches.

The analysis focused on indicators relevant to D6C3 and C5.

ToR A - Criteria to evaluate suitability of indicators/assessment methods

- Two sets of criteria were developed to evaluate indicators and thresholds for suitability for D6 in the MFSD. These criteria were building on Rice and Rochet 2005, WGECO, WKBENTH and Bundy *et al*. 2019.
- A list of 16 **indicator criteria** based on 4 major categories (data quality, management, conceptual, uniqueness) were compiled. The indicator criteria were weighted by importance (3 level score based on Core, Desirable, Informative) and compliance scores were defined. The framework was evaluated and found suitable for 6 test indicators relevant to D6 (BQI, TDI, SoS, DKIv2, PD2, M-AMBI). These 6 indicators represented two major types of indicators: multivariate indicators of ecosystem state, and single pressure-state relationship.
- The expert group looked at developing a **threshold criteria** framework that will be purpose-fit for the MSFD revision 2017 and in particular D6. A list of 12 criteria based on 4 major categories (overall evaluation, management evaluation, scientific evaluation, societal evaluation) were compiled, based on WGECO 2013 and further added to by the

group. The threshold criteria were weighted by importance (3 level score) and criterion scores were defined.

- The criteria were designed to evaluate at a subregional or regional level, but not a cross-regional level. The scoring for these criteria are meant as a guidance when choosing indicators and thresholds, so failure to meet one criteria will not prevent the use of the indicator or threshold in an assessment as it may meet other criteria or regional specificities.
- The criteria were useful for evaluation both indicators and thresholds. The scoring process was sensitive to expert groups composition and works most consistently when there are experts in the group on both the criteria themselves and on the indicators and thresholds. When evaluating indicators or thresholds for a specific region or subregion it is important to have experts from that area.

ToR B - Options for setting thresholds to evaluate adverse effects

- In the MFSD assessment of benthic habitats we need both quality and extent threshold to achieve GES. The quality threshold sets what the local state is, extent how much needs to be in a good state. Eleven different methods for setting GES thresholds were identified, covering methods for both quality and extent thresholds. More options were identified that are suitable for setting scientifically-justified quality thresholds (7) than for scientifically-justified extent thresholds (5).
- The WK considered GES quality thresholds based on the lower boundary of the range of natural variation most promising. This approach can be used for many indicators, but not all.
- The WK considered that it is difficult to estimate extent thresholds for GES. Possible approaches for setting extent threshold are be about keeping the state of the ecosystem above that would result in impairment of recruitment of benthic species, or evaluating trade-offs.
- Methods for setting thresholds were prioritised for detailed evaluation against the criteria from TorA and scored. This resulted in a clear separation of favoured and less-favoured methods. A key distinction between methods are methods that aim to identify an ecologically motivated difference between a good and degraded state, vs. methods that identify a different transition on the state-pressure relationship.
- A preliminary worked example for estimating and evaluation of the quality and extent threshold for the PD indicator, using some very preliminary threshold estimates, for the North Sea and Baltic Sea is presented.

ToR C - Quantitative and qualitative ways to compare methods/indicators

- Datasets were identified to evaluate the specificity, sensitivity and/or responsiveness of the different sampling-based indicator methods. These will be used to calculate most of the currently available sampling-based indicators. The WK identified 14 benthic datasets over gradients of commercial bottom trawling intensity (from relatively undisturbed to adversely affected), and 3 gradients related to eutrophication/pollution. Output is evaluated in WKBENTH3.
- The WK suggests that an evaluation of risk-based approaches can be realised through a ranked sensitivity and impact score per MSFD habitat type and subdivision (or EEZ). The ranking is helpful because it limits the need to standardize and/or calibrate outputs

across approaches. The ranked score per habitat type will allow WKBENTH3 to examine if the risk-based approaches find the same habitat types most sensitive to physical disturbance and/or most at risk of adverse effects. A worked example of PD2 and L1 is provided.

ToR D - Compilation of assessment methods/indicators

- WKBENTH2 participants provided input into the selection of indicators for the compilation of indicators.
- A template was developed for documenting the characteristics of each indicator to facilitate the evaluation of the indicators.

7.2 Next Steps

The WK decided to collate a standardized dataset to test the specificity, sensitivity and/or responsiveness of benthic indicators to pressure gradients. The collated dataset will be prepared and sent to experts end of June 2022. Experts will have until 31st of August to estimate indicator values. The outcomes will be quality checked by 20th of September and analysed at WKBENTH3.

Risk-based methods will be evaluated as maps and by scored sensitivity and impact score per MSFD habitat type and subdivision. Risk-based methods will be selected using ToR D and experts will be contacted by mid-July. Experts will have until 31st of August to provide input data. The outcomes will be quality checked by 20th of September and analysed at WKBENTH3.

Ahead of WKBENTH3, the workshop report will be reviewed through a peer-review process. Additionally, TGSeabed and the upcoming chairs of WKBENTH3 may provide input in preparation to the next workshop.

Annex 1: List of participants

Member	Institute	Email	Country of Institute
Abdeladim Bendraoui	Unaffiliated	abdeladim.cap@icloud.com	Marocco
Aleksander Drgas	National Marine Fisheries Research Institute	olek@mir.gdynia.pl	Poland
Alessandra Nguyen Xuan	Institute for Environmental Protection and Research	alessandra.nguyenxuan@isprambiente.it	Italy
Alexander Schröder	Lower Saxony Water Management Coastal Defence and Nature Conservation Agency	Alexander.Schroeder@nlwkn-ol.niedersachsen.de	Germany
Alice Belin	European Commission Directorate-General for Environment A4	alice.belin@ec.europa.eu	Belgium
Andrea Pierucci	COISPA Tecnologia & Ricerca	pierucci@coispa.eu	Italy
Angella Santelli	Institute for Marine Biological Resources and Biotechnologies	angela.santelli@cnr.it	Italy
Anna Downie	Centre for Environment, Fisheries and Aquaculture Science	anna.downie@cefas.co.uk	United Kingdom
Anna Luff	GoBe Consultants	Annaluff@gobeconsultants.com	United Kingdom
Antonia Nystrom Sandman	AquaBiota Water Research	Antonia.Sandman@aquabiota.se	Sweden
Aurélien Boyé	Ifremer	Aurelien.Boye@ifremer.fr	France
Axel Kreutle	Federal Agency for Nature Conservation	axel.kreutle@bfn.de	Germany
Bianca di Lorenzo	Institute for Environmental Protection and Research	bianca.dilorenzo@isprambiente.it	Italy
Bianca di Lorenzo	Institute for Environmental Protection and Research	bianca.dilorenzo@isprambiente.it	Italy
Chris Smith	Hellenic Centre for Marine Research	csmith@hcmr.gr	Greece
Cristina Herbon	Joint Nature Conservation Committee	Cristina.Herbon@jncc.gov.uk	United Kingdom

Member	Institute	Email	Country of Institute
Daniel van Denderen	DTU Aqua, National Institute of Aquatic Resources	pdvd@aqua.dtu.dk	Denmark
David Reid	Marine Institute	david.reid@marine.ie	Ireland
Despina Kyriakoudi	Flanders Marine Institute	despina.kyriakoudi@vliz.be	Belgium
Elisa Baldrighi	Institute for Marine Biological Resources and Biotechnologies	elisa.baldrighi@irbim.cnr.it	Italy
Elisa Punzo	National Research Council	elisa.punzo@cnr.it	Italy
Ellen L. Kenchington	Fisheries and Oceans Canada	ellen.kenchington@dfo-mpo.gc.ca	Canada
Esther Beukhof	DTU Aqua, National Institute of Aquatic Resources	estb@dtu.dk	Denmark
Flemming Hansen	DTU Aqua, National Institute of Aquatic Resources	ftho@aqua.dtu.dk	Denmark
Gabriele Di Bona	University of Palermo	gabriele.dibona@you.unipa.it	Italy
Gert Van Hoey	The Flanders Research Institute for Agriculture, Fisheries and Food	Gert.vanhoey@ilvo.vlaanderen.be	Belgium
Giacomo Montereale Gavazzi	Royal Belgian Institute of Natural Sciences	gmonterealegavazzi@naturalsciences.be	Italy
Giada Riva	University of Padua	giada.riva@studenti.unipd.it	Italy
Grete Elisabeth Dinesen	DTU Aqua, National Institute of Aquatic Resources	gdi@aqua.dtu.dk	Denmark
Hannah Schartmann	Thünen-Institute of Baltic Sea Fisheries	hannah.schartmann@thuenen.de	Germany
Hatice Onay	Recep Tayyip Erdogan University	hatice.bal@erdogan.edu.tr	Turkey
James Bell	Centre for Environment, Fisheries and Aquaculture Science	james.bell@cefas.co.uk	United Kingdom
Jan Geert Hiddink	Bangor University School of Ocean Sciences	j.hiddink@bangor.ac.uk	United Kingdom
José Manuel González Irusta	Centro Oceanográfico de Santander	jmanuel.gonzalez@ieo.csic.es	Spain

Member	Institute	Email	Country of Institute
Karin van der Reijden	DTU Aqua, National Institute of Aquatic Resources	kjova@aqua.dtu.dk	Denmark
Liam Matear	JNCC Seabirds and Cetaceans	Liam.Matear@jncc.gov.uk	United Kingdom
Lina Gutierrez	Duke University	lina.gutierrez94@gmail.com	Colombia
Maider Plaza	The Spanish Institute of Oceanography	maider.plaza@ieo.csic.es	Spain
Marco Milardi	Unaffiliated	marco.milardi@gmail.com	France
Marie-Julie Roux	Fisheries and Oceans Canada	Marie-Julie.Roux@dfo-mpo.gc.ca	Canada
Marija Sciberras	Heriot-Watt University	M.Sciberras@hw.ac.uk	United Kingdom
Marina Penna	National Institute for Environmental Protection and Research - Branch Office Chioggia	marina.penna@isprambiente.it	Italy
Marina Pulcini	Institute for Environmental Protection and Research	marina.pulcini@isprambiente.it	Italy
Marta Mega Rufino	Portuguese Institute for the Sea and the Atmosphere	marta.rufino@ipma.pt	Portugal
Mats Blomqvist	Hafok AB	mb@hafok.se	Sweden
Miquel Canals Artigas	University of Barcelona	miquelcanals@ub.edu	Spain
Murray Thompson	Centre for Environment, Fisheries and Aquaculture Science	murray.thompson@cefas.co.uk	United Kingdom
Nadia Papadopoulou	Hellenic Centre for Marine Research	nadiapap@hcmr.gr	Greece
Norbert Haubner	Swedish Agency for Marine and Water Management	norbert.haubner@havochvatten.se	Sweden
Owen Rowe	Helsinki Commission (Baltic Marine Environment Protection Commission)	owen.Rowe@helcom.fi	Finland
Pascal Laffargue	Ifremer	Pascal.Laffargue@ifremer.fr	France
Petra Schmitt	BIOCONSULT Schuchardt & Scholle GbR	schmitt@bioconsult.de	Germany

Member	Institute	Email	Country of Institute
Sander Wijnhoven	Ecoauthor	sander.wijnhoven@ecoauthor.net	Netherlands
Sandrine Vaz	Ifremer	Sandrine.Vaz@ifremer.fr	France
Saša Raicevich	National Institute for Environmental Protection and Research - Branch Office Chioggia	sasa.raicevich@isprambiente.it	Italy
Sebastian Valanko	International Council for the Exploration of the Sea	sebastian.valanko@ices.dk	Other
Silvia Maltese	Institute for Environmental Protection and Research	silvia.maltese@isprambiente.it	Italy
Sofia Reizopoulou	Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC)	sreiz@hcmr.gr	Greece
Stephen Duncombe-Smith	JNCC Seabirds and Cetaceans	Stephen.Duncombe-Smith@jncc.gov.uk	United Kingdom
Tim Mackie	Department of the Environment of Northern Ireland	Tim.Mackie@doeni.gov.uk	United Kingdom
Ulla Fernández	The Spanish Institute of Oceanography	ulla.fernandez@ieo.csic.es	Spain

Annex 2: Resolutions

2022/WK/HAPISG The Workshop to scope assessment methods to set threshold and assess adverse effects on seabed habitats (WKBENTH2), chaired by Dave Reid (Ireland), Daniel van Denderen (USA), and Jan Geert Hiddink (UK), will be established and will meet in Copenhagen, Denmark, 24–26 May and 8-10 June 2022 to:

a) Establish a set of criteria that can be used to evaluate suitability of regional indicators/assessment methods to assess adverse effects on seabed habitats for MSFD purposes
b) Review methods and criteria to set thresholds adverse effects on seabed habitats, and suggest operational options that can be illustrated using worked examples
c) Suggest quantitative and qualitative ways to evaluate and compare suitability and performance of indicators/assessment methods
d) Provide input to a draft compilation of regional indicators/assessment methods to set threshold and assess adverse effects on seabed habitats

WKBENTH2 will report by the end of July 2022 for the attention of the Advisory Committee

Supporting information

Priority	High, in response to the stepwise process of delivering guidance on seafloor integrity for the Marine Strategy Framework Directive (MSFD). The workshop outputs will feed into ICES WGFBIT and the ongoing efforts to provide guidance o assessment methods to set threshold and assess adverse effects on seabed habitats i the operational implementation of the MSFD.
Scientific justification	Term of Reference a) ICES has previously produces criteria on what makes a good indicator, in general (e.g. WGECO, Rice and Rochet 2005) and specifically for assesing the seafloor habitats (WKBENTH 2017). Criteria should faciliate an evaluation on the suitability and shortcomings of any proposed indicators for MSFD assessment purposes, reflecting their performance to assess the parameters specified in Commission Decision (EU) 2017/848 on condition of seabed habitats and the adverse effects of key pressures. Criteria should take into account the indicators applicability across MSFD broad habitat types (or subtypes), their suitability for large sea areas (i.e., all marine waters of MS, marine regions or subregions). Term of Reference b) Options for setting thresholds should take into accouns as far as possible recent work by EU's TG SeaBed on threshold values for adverse effects on habitat condition (D6C5) and for the maximum allowable extent of habitat loss (D6C4) and of adverse effects (D6C5) Ref. document GES_26-2022-13. TOR b will suggest criteria on how to set thresholds and review potential methods that can be used to identify values (or ranges of values) for the indicators which would distinguish a habitat in good condition from the one which is adversely affected or lost (in general or by specific pressures) to set thresholds. This should, for example, reflect on whether there is a linear or non-linear response of the habita to particular pressures. Term of Reference c) Suggest options on how to quantitative and qualitative evaluate and compare suitability/performance of indicators/assessment methods. This may include

	identifying data sources (i.e. via TG Seabed), in order to evaluate the performance o selected (reviewed) benthic risk and state indicators, in relation to their ability to assess the state/condition of seabed habitats and adverse effects from specified pressures. Proposed analytical ways to compare methods should ensure WKBENTH3 (Sept/Oct 2022) can evaluate suitability and shortcomings of both risk and state indicators for MSFD assessment purposes at national and regional scales. Quantitative and qualitative analytical approaches should suitable for application using worked examples to demonstrate the suitable methods to set threshold and assess adverse effects on seabed habitats. Term of Reference d) ICES appointed experts will compile as a technical service information on suitable methods to set threshold and assess adverse effects on seabed habitat. TOR d gives WKBENTH2 the opportunity to provide input towards this compilation. The aim o ICES work is to produce as advice a detailed review of indicators used, or under development, by Regional Sea Conventions, Member States and ICES, for assessing the state/condition of seabed habitats suitable for MSFD assessments. The indicator considered can also include peer-reviewed indicators which have large-scale application. Provide a detailed review of indicators used, or under development, b Regional Sea Conventions (RSCs), Member States and ICES, for assessing the state/condition of seabed habitats and relevant existing literature. This should include indicators based on both direct observational data and on models. Relevant indicators to be reviewed include those of RSCs for quality status assessments, of Member States for MSFD purposes such as under the Water Framework Directive (WFD) and the Habitats Directive (HD), and those used by ICES. The review shoul specify the input data, how it is processed, the parameters of habitat quality used, how quality is quantified, any threshold values used, the applicable seabed (habita and pressure types, how the output is expressed, and how confidence and uncertainty are handled.
Resource requirements	ICES secretariat and advice process.
Participants	Workshop with researchers and RSCs investigators If requests to attend exceed the meeting space available ICES reserves the right to refuse participants. Choices will be based on the experts' relevant qualifications for the Workshop. Participants join the workshop at national expense.
Secretariat facilities	Data Centre, Secretariat support and meeting room.
Financial	Covered by DGENV special request.
Linkages to advisory committees	Direct link to ACOM.
Linkages to other committees or groups	Links to HAPISG and SCICOM.
Linkages to other organizations	Links to RSCs and EC.

Annex 3: Presentation abstract

Northwest Atlantic Fisheries Organisation – Working Group on Ecosystem Science and Assessment: Assessing SAI upon VMEs in the NAFO regulatory area

Following the FAO guidelines for Deep-Sea Fisheries in Areas Beyond National Jurisdiction (FAO, 209), the scientific committee of the Northwest Atlantic Fisheries Organisation (NAFO) have undertaken to complete an assessment of 'significant adverse impacts' (SAI) upon vulnerable marine ecosystems (VME) in its regulatory area, through its Working Group on Ecosystem Science and Assessment (WG-ESA). A brief presentation of this work was given at the meeting, and summarised here, specifically regarding recent efforts to establish reference points for SAI (NAFO, in prep.).

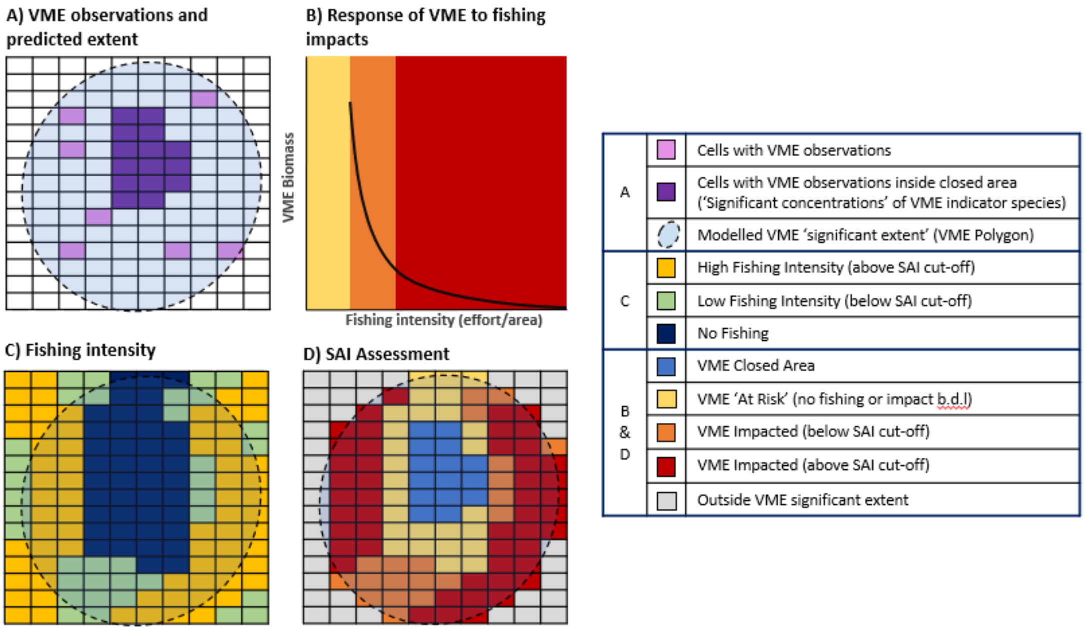

A) VME observations and predicted extent
B) Response of VME to fishing impacts
VME Biomass
Fishing intensity (effort/area)
C) Fishing intensity
D) SAI Assessment
A
Cells with VME observations
Cells with VME observations inside closed area ('Significant concentrations' of VME indicator species)
Modelled VME 'significant extent' (VME Polygon)
C
High Fishing Intensity (above SAI cut-off)
Low Fishing Intensity (below SAI cut-off)
No Fishing
B & D
VME Closed Area
VME 'At Risk' (no fishing or impact b.d.l)
VME Impacted (below SAI cut-off)
VME Impacted (above SAI cut-off)
Outside VME significant extent

Figure 1 – SAI assessment workflow developed in WG-ESA.

The WG-ESA workflow to assess SAI (Fig. 1D) is spatially explicit and takes account of differences in sensitivity between different taxa, and relies on three distinct data streams (Fig. 1A-C):

A. Determination of the 'significant extent' of selected VME indicator species (see Kenchington *et al*. 2014)

B. Estimation of taxa-specific responses to bottom trawl fishing intensity.

C. Analysis of vessel monitoring system records to determine the distribution of commercial fishing activity. In this case, fishing intensities are estimated from VMS data collected between 2010-19.

Determining the extent of SAI relies upon establishing a reference point at which SAI is considered to have occurred. The relationship between trawl fishing intensity and removal of VME indicator taxa (Fig 1B) is determined from VME bycatch volumes caught during annual NAFO scientific fishing surveys. Cumulative VME biomass is compared across a gradient of commercial trawl fishing intensity, with the assumption that VME indicators are decreasingly likely to be caught in survey hauls in cells with higher historic commercial fishing activity. The shape of this response curve for each taxon, coupled with a given percentage value of biomass loss taken to be SAI, is used to determine the fishing intensity equivalent of SAI, and completes the spatial assessment (Fig. 1D).

Individually determining the biomass loss that meets the FAO criteria of SAI for individual indicator species, e.g., through biomass distribution and larval dispersal modelling studies, would be considerably time-consuming and likely very context-specific. In the absence of such detailed information for all VME taxa, we examined the relationship between VME biomass loss and trawl fishing intensity (Figure 1B), using the spatial distribution of both. Excepting bryozoans, biomass loss rate demonstrated similar functional response curves, with rates of change typically highest above 80 % and below 20-30 %, with a generally much shallower, or in some cases almost zero, rate of change in between (Fig. 2).

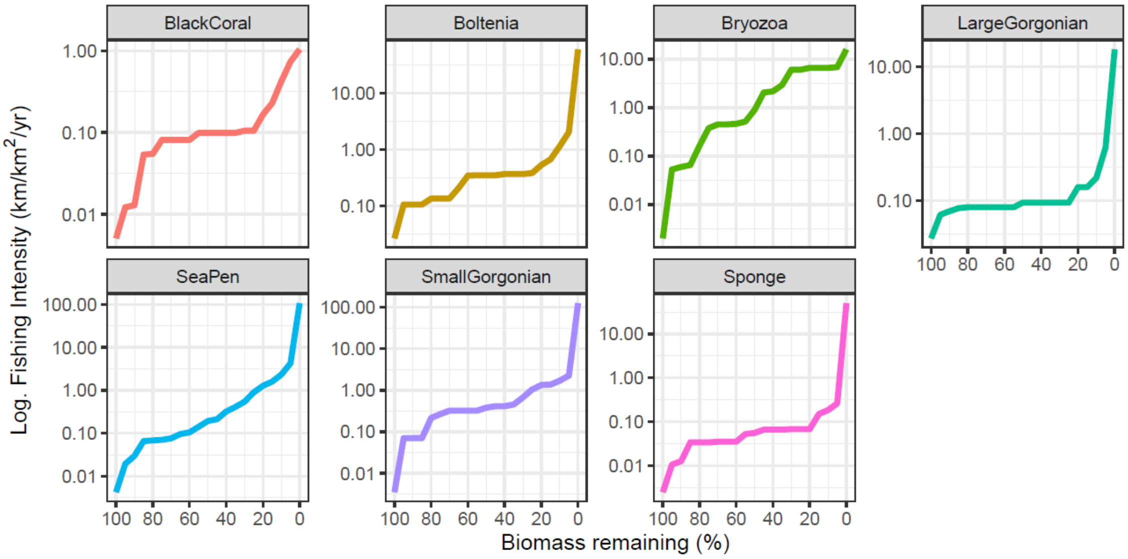


Figure 2 – Response curves (biomass loss for a given level of bottom trawl fishing intensity) of the seven VME taxa groups considered.

In practice, this means that for these VME taxa, *any level* of fishing intensity is expected to remove around 20 % of biomass. Beyond that, low fishing intensities (generally in the range of 0.1 – 1.0 km km^{-2} yr^{-1}) steadily removes the majority of the remaining biomass (Fig. 2). For these species, the majority of impacts occurr at low levels of fishing, areas that are actively fished less than once or twice a year, so potentially represent large areas where impacts are occurring, but that are of minimal importance to the industry.

The currently proposed reference points for SAI in NAFO are in the range of 20 – 35 % VME biomass remaining, based on the position of the second inflection point along the response curves (Fig. 2). It remains unclear however that these reference points are adequate to satisfy all of the SAI criteria set by FAO (2009), particularly for criteria that are more challenging to determine empiracally, such as the extent to which a given pressure alters or degrades ecosystem function.

Citations:

FAO, 2009. International Guidelines for the Management of Deep-sea Fisheries in the High Seas. 90 pp. Available online at https://www.fao.org/documents/card/en/c/b02fc35e-a0c4-545a-86fb-4fc340e13b52

Kenchington E, Murillo F J, Lirette C, Sacau M, Koen-Alonso M, Kenny A, Ollerhead N, Wareham V, Beazley L. 2014. Kernel Density surface modelling as a means to identify significant concentrations of Vulnerable Marine Ecosystem indicators. PLoS ONE 9(10): e109365. doi:10.1371/journal.pone.0109365

Kenny, A., Bell, J. B., Blasdale, T., Downie, A. *et al*. in prep. Grading on a curve: Determining Significant Adverse Impacts reference points for Vulnerable Marine Ecosystem indicator species in the northwest Atlantic

NAFO, in prep. Report of the 2021 meeting of the NAFO Working Group for Ecosystem Science and Assessment, November 2021. To be published online at https://www.nafo.int/Meetings/Past-Meetings/2021

Annex 4: Report from the Review Group to scope assessment methods to set thresholds and assess adverse effects on seabed habitats (RGBENTH2)

Participants: Sophie Mormede, Steven Degraer, Simon Jennings (chair)

Meeting: By correspondence July-September 2022

Request: Review group participants were asked to review two reports:

1. ICES Workshop on assessment methods to set thresholds and assess adverse effects on seabed habitats (WKBENTH2)

2. Technical service to produce a compilation of assessment methods and indicators that can be used to assess seabed habitats under D6/D1 for the MSFD.

And to assess whether,

a) The analyses were technically correct.

b) The scope and depth of the science were appropriate for the request.

c) The analyses contained the knowledge to answer the request for advice.

Background

ICES advised that the RGBENTH2 review of WKBENTH2 would be provided to WKBENTH3 as well as to the subsequent Advice Drafting Group. WKBENTH3 will have the task of evaluating proposed assessment methods and evaluating thresholds for assessing adverse effects on seabed habitats, using agreed upon criteria, methods, and analyses of their performance. Outcomes of WKBENTH3 will also contribute to the advice to DGENV.

Note on process

This report combines comments from the three reviewers: Sophie Mormede, Steven Degraer and Simon Jennings. The reviewers had different backgrounds and expertise and each reviewer conducted an individual review of the documents before meeting with other reviewers to agree responses and the structure of the combined report. Although the reviewers' comments focused on different aspects of the reports, the compiled comments were discussed and agreed collectively.

Two tables are used to summarise the reviewers' responses to questions 'a' to 'c' in the request, and these are followed by a section-by-section review of the WKBENTH2 report. An opening summary highlights the key messages from the review.

One reviewer also annotated the original WKBENTH2 report with smaller comments. The annotated report is available from ICES Secretariat.

Summary

Both the ICES Workshop on assessment methods to set thresholds and assess adverse effects on seabed habitats (WKBENTH2) and the ICES Technical Services team adopted a technically appropriate approach to fulfil their Terms of Reference and to provide inputs to WKBENTH3 and the ADG. The scale of their task was large given the amount of work on benthic indicators now being linked to the MSFD D6 (and noting that much of this work was not initiated for this purpose) and the range of interpretations of MSFD D6 processes that exist both nationally and internationally.

We note the substantial progress made with respect to seafloor integrity indicator and threshold evaluation. WKBENTH2 (feeding into WKBENTH3) is another milestone in the long history of development, evaluation, and selection of seafloor integrity indicators and thresholds. We encourage WKBENTH3 to focus on the steps needed to screen and condense available information and process and to draw strong and tractable conclusions that will underpin advice.

The scope and depth of the scientific treatment of the WKBENTH2 Terms of Reference and requirements for the Technical Service are largely appropriate. They do provide a common evaluation framework as requested. The main omission is that the focus on the requirements of Commission Decision (EU) 2017/848 is not strongly developed in the WKBENTH2 report and in the evaluation of indicators and thresholds. A stronger focus on 2017/848 would likely help ICES to provide clearer and more actionable information and advice on the value of proposed indicators and thresholds.

The work completed to date by WKBENTH2 and ICES Technical Services will go a long way towards guiding the expected tasks of WKBENTH3 and addressing the request for advice, although amendments to the summary tables developed by WKBENTH2 and used in the Technical Service task are recommended. Specifics are detailed in Table 1 and Table 2 and the body of this review. Parts of the WKBENTH2 report conflate indicators, methods and sometimes thresholds. Going forward, we suggest these should be explicitly split, defined, and criteria applied to each. An appropriate 'taxonomy' may be: 1- candidate indicator, 2- methodology to calculate the indicator, 3- application of the methodology, and 4- thresholds. There are also some inconsistencies and redundancies in the report, perhaps reflecting the drafting of sections by different subgroups. Specific examples are provided in the body of this review.

If an indicator and threshold can be defined, then many different methods may be used to assess the value of the indicator in relation to the threshold. Fisheries science provides a classic aquatic example, where one indicator that is broadly accepted internationally is biomass and thresholds are often set as a proportion of unimpacted (modelled) biomass. However, a very wide range of methods and models are developed, tested, and used to acquire data, and ultimately estimate the biomass. It would likely be easier for developers if a small suite of indicator(s) and thresholds was defined and the science was more strongly focused on the methods to estimate values and associated uncertainty for these indicators.

The reviewers were concerned about the large differences in scoring that were observed, and this implies that further work is needed to reduce ambiguity in the criteria and/ or to advance common understanding of the scoring process.

It is suggested to go beyond the weighting and scoring criteria presented to identify criteria that are essential for an indicator or threshold to meet the requirements of Commission Decision (EU) 2017/848 and the MSFD. If indicators or thresholds do not/ will not meet these criteria (now, or on some specified future time frame to be proposed/ decided by the group) then it is logical that they are not emphasised as potentially appropriate in the advice (at this time), even if they score highly on some criteria. If a proportion of the numerous indicators and thresholds are identified as inappropriate for use at this stage, this will ultimately contribute to stronger and more concise advice on (a) the suitability and shortcomings of both risk and state indicators for MSFD assessment purposes, and (b) on threshold values, at national and regional scales. More widely, with indicators and thresholds being both numerous and at varying stages of development, progress with any performance evaluation, intercomparison or intercalibration exercise would have to be very protracted, or simplified to the point of being uninformative, to accommodate all suggested indicators and thresholds irrespective of whether they meet essential criteria for the MSFD.

We appreciate, of course, that many groups developing or using specific indicators will strongly champion them, and that this can complicate selection exercises. For the purposes of moving the exercise forward a few indicators may be classified as mature (unconditional pass) and thus carried forward in the current advice and others, rather than being entirely dropped from the process, may be highlighted as 'conditional' passes with the necessary conditions for further development being clearly specified (tabulated). The conditional passes are the indicators that would then be flagged as inappropriate for use in the context of the MSFD at this time.

Two main types of indicators are listed in the Technical Services document: empirically-based methods and model-based methods, the former making the bulk of the indicators. Commission Decision 2017/848 states that physical loss shall be understood as a permanent change to the seabed which has lasted or is expected to last for a period of two reporting cycles (12 years) or more. This leads to a requirement to understand the rate of recovery, which clearly favours modelling options, as does the scale at which the assessment has to be applied compared to data availability.

Model-based methods can integrate spatial and temporal processes of impact and recovery and be calculated at the population or community scale. Caveats may include that they can become very complex, need to be well reviewed, tested, reproducible, and that initial state needs to be defined. Indicators BH3 and PD could be tested against each other, including in relation to recovery time, and with varying assumptions, data quality etc. BH4 and Cumul seem to need less data, and could potentially be used more widely, and could be correlated with more complex methods such as BH3 and PD. The uncertainty surrounding the simpler methods might be no more important than the combined effects of the assumptions in the more complex methods.

Many empirically-based methods (e.g. BISI, HELCOM etc) directly measure the current state, or the current pressure, and then compare with threshold values. There is merit in having direct measures of state and of pressure, particularly if they are to be monitored consistently over time in the same area. The risks of using such indicators include applying thresholds from other areas which might not be suitable, scale of sampling and monitoring etc. Further, we would caution against ensemble methods (such as NEAT) when there are very few well tested models to include in the ensemble. For example, if one result is correct and the other is not, the average will always be wrong.

The groups are encouraged to undertake a detailed review of 2017/848, especially the D6 annex, to discuss and determine the extent to which all requirements of this Decision lead to other relevant criteria for the selection of indicators and thresholds, and thus determine the appropriateness of these indicators and thresholds to support MSFD.

Much research effort has been focused on the impacts of active bottom fisheries on seafloor integrity. Many seafloor integrity indicators hence relate to the impact of bottom fisheries, which is visible in the WKBENTH2 report. The remit of WKBENTH2 however was to assess indicators for seafloor integrity. The WKBENTH2 resolution mentions "adverse effects on seabed habitats" and "condition of seabed habitats and the adverse effects of key pressures". There is no specific mentioning of bottom fisheries in the resolution. The report occasionally reads as if fishing pressure was the main topic. While we recognise the emphasis in Commission Decision 2007/848 that Member States "focus their efforts on the main anthropogenic pressures affecting their waters" and should “have sufficient flexibility, under specified conditions, to focus on the predominant pressures and their environmental impacts on the different ecosystem elements in each region or subregion….” some subtle reconsideration of the text may provide some more balance. For example, on lines 423-423 “…to impacts from bottom contacting fishing gears” is likely obsolete because the subcriteria do not refer to fishing pressure; Table 3.1.1, Criterion 5 “This should include if the indicator is capable of including different gears with different impacts on habitats or species, if this is relevant for the indicator and its application” gives a (presumably unintended) focus solely on fishing pressure; and in the case of lines 1174-1175 “The resulting dataset consisted mostly of shrimps and lobster, with a few mollusc stocks” the term “stock” typically refers to the population size of commercial species, while GES should not be restricted/related to commercial species. Better to use the term "population size". This would contribute to the general appreciation of the report as going beyond fisheries-related aspects of GES.

A final point relates to the role of ICES science in identifying both indicators and thresholds. A complexity of this process, and one that will be challenging for WKBENTH3, the ADG and ICES in general, is the absence of a stronger steer on thresholds and appropriate precaution from policy and policy-stakeholder dialogue. A science group would usually consider the consequences of setting different thresholds or adopting different levels of precaution, rather than advise on what the specific thresholds or the level of precaution should be. There is not much policy steer to help the group, but the little that has been agreed and published (primarily in Commission Decision 2017/848) should be directly addressed as a priority, especially the specific statement that physical loss shall be understood as a permanent change to the seabed which has lasted, or is expected to last, for a period of two reporting cycles (12 years) or more.

Tabulated review

Two tables with identical rows and columns were used to assess the contributions of WKBENTH2 and the ICES Technical Service to the DGENV advisory request. These are presented independently because one is focused primarily on interpretation of the Commission Decision 2017/848 and because the reviewers were familiar with different groups of indicators.

Table 1. Review 1 of contributions of WKBENTH2 and the ICES Technical Service to the DGENV advisory request.

Review question Advice request	a) Is the analysis technically correct?	b) Are the scope and depth of the science appropriate for the request?	c) Does the analysis contain the knowledge to answer the request for advice?
(i) A detailed review of indicators used, or under development, by Regional Sea Conventions, Member States and ICES, for assessing the state/condition of seabed habitats suitable for MSFD assessments. The indicators considered can also include peer-reviewed indicators which have large-scale application.	A detailed review is provided as a technical service, based on a template developed by WKBENTH2. The specifics of the review are dependent on specialists with knowledge of the individual indicators, although we do not find errors in assessments for the small number of indicators with which we are familiar. The analysis in relation to the properties/ criteria considered is technically thorough.	The process in general has been thorough, and information requested/ collected on the templates is relatively complete. We suggest the final review that goes into the advice should include additional criteria that link the properties of indicators explicitly to Commission Decision (EU) 2017/848 as well as to the generic properties of good indicators.	In part. The detail could be enhanced and more useful to the recipients if it included detail of links to Commission Decision (EU) 2017/848.

Review question Advice request	a) Is the analysis technically correct?	b) Are the scope and depth of the science appropriate for the request?	c) Does the analysis contain the knowledge to answer the request for advice?
(ii) Advise, using a set of agreed criteria, on a common framework to evaluate methods to assess benthic risk (model) and state (data) indicators, with respective threshold values.	The basis of the frameworks proposed is technically reasonable, except for the weighting and scoring processes where we suggest that indicators not meeting 'critical' criteria should be identified as unsuitable for MSFD support even if they score highly on other criteria. 'Critical' criteria will need to be identified. Note the suggestion that this may be handled by assigning an 'unconditional' and 'conditional' pass with additional (future) requirements clearly highlighted in the case of 'conditional' passes.	Appropriate (but incomplete) frameworks for evaluation were developed by WKBENTH2. As in (i)b we suggest the scientific evaluation should include an assessment of the properties of indicators and thresholds in relation to specific policy requirements. The treatment of uncertainty should be addressed beyond 'methodology' and 'output' (eg present confidence interval). This may require clarification from the requesters of the advice (and we note it does not appear in the ToR for the WKBENTH2) but 2017/848 states [threshold values should] "be set on the basis of the precautionary principle, reflecting the potential risks to the marine environment" so we interpret that uncertainty should also be assessed in this context when evaluating indicators and thresholds. In practice, this may mean a criterion that assesses whether the approaches adopted by the indicator developers enable an assessment that any given (calculated/ recorded) value of the indicator is consistent with avoiding a defined threshold (e.g. for loss as defined in 2017/848 in the extreme case) with a high probability. Note comments in the summary of this RGBENTH2 report on the distinction between indicators and methods.	"Yes" in general terms, but improvements to the criteria as described in ii(a) and ii(b) would improve the rigour of evaluation.
(iii) A targeted benthic data call (via TG Seabed), in order for ICES to evaluate the performance of selected (reviewed) benthic risk and state indicators, in relation to their ability to assess the state/condition of seabed habitats and adverse	Since this relates to a request via TG seabed it was not clear if the identification of benthic datasets in Section 5.1.2 of RGBENTH2 was relevant. The datasets identified in WKBENTH2 for use in WKBENTH3 would be suitable for assessing the effects of trawling disturbance.	See (iii)a.	See (iii)a.

Review question Advice request	a) Is the analysis technically correct?	b) Are the scope and depth of the science appropriate for the request?	c) Does the analysis contain the knowledge to answer the request for advice?
effects from specified pressures.			
(iv) Advice on threshold values to assess the quality of seabed habitats.	The basis for this advice is available in the WKBENTH2 report. There is much general text on thresholds, and this provides a technically appropriate review of the general topic, but the development of the specific links to MSFD could be more focused, especially in the section where workshop participants focus on the suitability of the approaches covered in the review. At least one threshold is defined in Commission Decision (EU) 2017/848 (as mentioned on line 850 of the WKBENTH2 report). This important point is not further developed, but our interpretation is that this already defines a threshold for loss ("Physical loss shall be understood as a permanent change to the seabed which has lasted or is expected to last for a period of two reporting cycles (12 years) or more") and the technical science question is whether recovery time can be determined for any proposed indicator given the defined threshold (ie. Does the scientific basis of this indicator provide for estimation of recovery time in years from the present state and therefore a determination of whether "loss" has occurred, as defined in 2017/848). If the science basis of an indicator does not allow this, then can it logically meet the needs of MSFD reporting at all? Note also the relevance of the 2017/848 text on precaution appears to apply "Threshold values should also be set on the basis of the precautionary principle, reflecting the potential risks to the marine environment".	Much of the material needed to provide advice is available in the WKBENTH2 report, but it needs to be significantly filtered to draw out material relevant to the MSFD and request. Note the guidance in the DGENV request to "Advise on values (or ranges of values) for the indicators which would distinguish a habitat in good condition from the one which is adversely affected or lost (in general or by specific pressures)". This also helps to guide the focus of the text. There is also a reference to the significance of loss to thresholds in the background to ToR 'b' for WKBENTH2.	Please note comments on precaution, uncertainty, and thresholds in (ii)b.

Review question Advice request	a) Is the analysis technically correct?	b) Are the scope and depth of the science appropriate for the request?	c) Does the analysis contain the knowledge to answer the request for advice?
(v) Advice on the suitability and shortcomings of both risk and state indicators for MSFD assessment purposes at national and regional scales.	The work that has been reported is technically appropriate to the extent we can judge, but is not sufficiently complete (in terms of criteria used and criticality of review) to make the full assessment as requested under (v)	Knowledge base could be strengthened, especially by assessing the evidence base related to the relationship between proposed indicators and thresholds and Commission Decision 2017/848.	"Yes" in general terms, but improvements to the criteria as described in ii(a) and ii(b) would improve the rigour of evaluation. Also applies to (ii).

Table 2. Review 2 of contributions of WKBENTH2 and the ICES Technical Service to the DGENV advisory request.

Review question Advice request	a) Is the analysis technically correct?	b) Are the scope and depth of the science appropriate for the request?	c) Does the analysis contain the knowledge to answer the request for advice?
TOR A: Establish a set of criteria that can be used to evaluate the suitability of regional indicators/assessment methods to assess adverse effects on seabed habitats for MSFD purposes	Yes – small comments. A precautionary margin should not be in the indicator, although the indicator should capture uncertainty. The precautionary margin should be explicit and included in the threshold only (or it could be double counted). Need to consider uncertainty in both indicators and thresholds. Suggest not to duplicate criteria in indicators and thresholds (e.g. spatial extent and analytical vs expert). All core criteria (for both indicators and thresholds) should have a fail if any essential criterion scores 0 but criterion 12 probably should not. A fail may be conditional (e.g. could be fixed with future work), with emphasis in the	Yes – minor comment. The process may be biased against new methodologies, whereas they should be encouraged (but shown to be better or complimentary to existing methodologies prior to adoption).	Partly. This analysis looked at indicators rather than assessment methods. The same indicators can be calculated using different methods or models. The split is not clear or explicit. Assessment methods require criteria too, such as peer review, agreed assumptions, tested sensitivities to assumptions, replicable, documented etc. Some of these are captured in the table of indicators.

Review question **Advice request**	**a) Is the analysis technically correct?**	**b) Are the scope and depth of the science appropriate for the request?**	**c) Does the analysis contain the knowledge to answer the request for advice?**
	future ICES advice on the 'unconditional' passes?		
TOR B: Review methods and criteria to set thresholds of adverse effects on seabed habitats, and suggest operational options that can be illustrated using worked examples	Some issues. 'Natural variation' assumes that there is enough comparable untouched habitat that has been surveyed to come up with values. It also assumes transferability between habitats (Yates *et al*. 2018), and also that the variability that arises from this assumption will somehow be smaller than the variability of a depleted state. It would have been useful to see the worked example used to calculate the values for impacted areas. There seems to be confusion in the worked example between extent and quality. We interpret the Worms analysis as treating 40%B0 as the lower limit (which is a maintain population size argument), not that a minimum of 40% of the population has to be above 80%B0. We suggest the two most promising thresholds relate to ecosystem state (at or above a specified threshold such as 40% B_0) and recovery time (which is specified in the Commission Decision 2017/848 anyway). . The lack of knowledge of stock-recruit relationship should not be a hindrance but used as sensitivity. If the thresholds are 40% of B0, or above for example, this value will have near no influence.	Some issues. There is some discussion about the importance of connectivity and the indicators and thresholds should probably be calculated at the metapopulation scale. Yet there is no discussion on the scale at which the analysis is to be carried out, and it is applied at the grid size in the worked example. Some thresholds will only be available for specific indicators. For example, population thresholds will only work for those indicators related to the populations while 'natural variation' (or some other level) will work for a wider range of indicators (e.g. including species richness).	Unsure. There are no operational options proposed, particularly with regards to the scale at which to calculate and to apply these thresholds. One of the worked examples uses the Pitcher method. This highlights again the need to investigate the assessment models as well as the indicators and thresholds. There are many assumptions within this method (including no stock recruit relationship and the level of vulnerability of benthic animals).
TOR C: Suggest quantitative and qualitative ways to evaluate and compare the suitability and performance	Difficult to tell. This is mostly a meta-analysis with little information so it is hard	Probably not. This section provides examples of indicators and different values at different pressures. But it does not apply	See other boxes. There does not seem to be clear guidance coming out of this section.

Review question Advice request	a) Is the analysis technically correct?	b) Are the scope and depth of the science appropriate for the request?	c) Does the analysis contain the knowledge to answer the request for advice?
of indicators/assessment methods	to tell if the comparison is like for like. The risk-based impact score is summarised over MFSD habitat, with no explanation if this is an average or other metric. It might be more transparent to report the proportion of each habitat which qualifies for each level of impact. It is also unclear which “extent” rule and “quality” rule have been applied.	the process of applying thresholds such as 95%ile of ‘natural variation’ and test if the 0.8 rule holds. For those examples the entire process should have been carried out: scoring the indicators, applying the thresholds and scoring the thresholds. None of the examples looked at the 12-year recovery period in Commission Decision 2017/848.	
TOR D: Provide input to a draft compilation of regional indicators/assessment methods to set threshold and assess adverse effects on seabed habitats	The authors commented how people familiar with specific methods are required to be able to score them. The same applies to the content of those tables.	This exercise is a balance between too much detail and not enough information. These summaries seem adequate to provide an idea of what is available. It would be worth cross checking that all criteria are covered in the tables (e.g. to capture the likelihood of future data availability). Major assumptions would be an informative extra category.	It appears that only indicators used in Europe were considered. Other work could have been considered such as what is done in SPRFMO for example, limiting to well-developed methods. It is good to see spatial resolution and uncertainty covered as they are in this table. Consideration of these issues may be developed and added to the criteria for assessing indicators.

Section reviews of WKBENTH2 report

Section 2.

Some of the information in Section 2 does not seem to link to the ToR and request. This distracts the reader from the core purpose, business, and conclusions of WKBENTH2 and we suggest this information may be removed from the report or included as Annexes.

Lines 144-152: It is unclear why the report elaborates on the challenges related to the identification of biogeographically relevant subdivisions, which is not part of the request. The valid advice for regional coordination to sort this out goes beyond the remits of WKBENTH2.

Lines 205- "Lessons learnt from the Water Framework Directive intercalibration", the relevance of this elaborate section is unclear. We suggest to either clarify its relevance (relative to the remits of WKBENTH2) at the start of the section or to consider moving the section to an Annex.

Section 3.

Previous WGECO and WGBIODIV work largely focused on generic properties of indicators, so to address the ToR it was necessary for WKBENTH2 to extend their approaches and to develop a criteria list suited to the specific requirements of the MSFD. The cross checking conducted by WKBENTH2 is a reassuring process in the context of completeness of the work. WKBENTH2 could have gone further in developing the specificity of the criteria to MSFD, especially in relation to the context provided by Commission Decision (EU) 2017/848 and what has already stated in the same Decision about appropriate scales of reporting (existing definition of units).

For criterion 2 in Table 3.1.1. there is emphasis on a monitoring time series to establish baselines and reference levels. Our reading of 2017/848 is that this would not be an essential prerequisite. While Article 4, 1h does state "be based on long time-series data, where available, to help determine the most appropriate value" this is not necessarily consistent with the preamble "marine ecosystems may recover, if deteriorated, to a state that reflects prevailing physiographic, geographic, climatic and biological conditions, rather than return to a specific state of the past." The latter implies model-based estimates of (current) baseline state may be required, unless there are comparator areas with comparable "prevailing physiographic, geographic, climatic and biological conditions" and where pressures are low enough to provide confidence that the state can be treated as 'baseline'.

For threshold values it was not clear why Article 4 of Commission Decision 2017/848 was not addressed directly as part of the WKBENTH2 work, especially in the cases where values will be established through Union, regional or subregional cooperation. There is some correspondence between 2017/848 and Table 3.2.1., and it is appropriate to identify additional criteria for evaluating thresholds (such as those previously considered by WGECO), but it would clarify the development of subsequent advice to work with those criteria mentioned in 2017/848 directly (and provide an operational interpretation of them).

For indicators, Commission Decision 2017/848 also suggests the need for a criterion to assess whether an indicator is responsive to a (known) main pressure (in the region where it is used), consistent with "As a result, the number of criteria that Member States need to monitor and assess should be reduced, applying a risk-based approach to those which are retained in order to allow Member States to focus their efforts on the main anthropogenic pressures affecting their waters" and "Member States should have sufficient flexibility, under specified conditions, to focus on the predominant pressures and their environmental impacts on the different ecosystem elements in each region or subregion in order to monitor and assess their marine waters in an efficient and effective manner and to facilitate prioritisation of actions to be taken to achieve good environmental status."

Overall, and for the purposes of this request, it is suggested to also have a set of criteria for indicators and thresholds that are clearly linked to the requirements of 2017/848. For transparency it is likely best to list these explicitly (rather than melding them into more general criteria from WGECO and elsewhere).

There are three specifics we would highlight in 2017/848 that would also be usefully considered and treated as criteria.

First, in the Annex for D6, it is stated that "Physical loss shall be understood as a permanent change to the seabed which has lasted or is expected to last for a period of two reporting cycles (12 years) or more." It follows that a valuable, and likely essential, property of an indicator is that it can be used to establish 'permanent' change, either through ongoing monitoring ("has lasted" ie shown to be below threshold for 12 years) or duration of recovery ("is expected to last" ie a quantitative prediction of recovery rate, ideally addressing uncertainty, shows the indicator will not meet the threshold after 12 years). Understanding of recovery is also emphasised in "Physical disturbance shall be understood as a change to the seabed from which it can recover if the activity causing the disturbance pressure ceases".

Second, area criteria for the threshold are highlighted in WKTRADE2, but is it necessary to consider these for the indicator too? It is clear from 2017/848 that indicators need to enable reporting of areas lost/ disturbed/ affected in units of km^2, so spatial coverage and resolution of application are both relevant (eg. to what extent can sampling be extrapolated to appropriate scales, what is the resolution of modelling).

Third, the treatment of uncertainty should be more explicit. 2017/848 states [threshold values should] "be set on the basis of the precautionary principle, reflecting the potential risks to the marine environment" so we interpret that uncertainty should also be assessed in this context when evaluating indicators and thresholds. In practice, this may mean a criterion that assesses whether the approaches adopted by the indicator developers enable an assessment that any given (calculated/ recorded) value of the indicator is consistent with avoiding a defined threshold (e.g. for loss as defined in 2017/848 in the extreme case) with a high probability. General provision of a confidence interval for the indicator may not enable this, depending on what the confidence interval represents. Note to avoid the risk of double counting by having precaution in the estimate of the indicator values as well as being addressed in the threshold. It would be most transparent to associate the precaution with the threshold (to define the required probability of avoiding an unwanted state) rather than the indicator. This would also be consistent with the recognition that some precaution is expected as defined in 2017/848 Article 4 Para 1(e) [thresholds shall] "(e) be set on the basis of the precautionary principle, reflecting the potential risks to the marine environment."

Having considered the need for more criteria in WKBENTH3, our next suggestion is to consider whether scoring and weighting alone will be sufficient to address the request for advice (the request relating to the suitability and shortcomings of indicators and not to the general review of indicators used). In the case of some criteria, and especially those linked to 2017/848, a pass/ fail approach would usefully be introduced for some criteria (perhaps attached to a timescale to reflect when a 'fail' may be converted to a 'pass' in the longer term eg. following further R&D). With this approach, and if an indicator or threshold fails on one of the key criteria, it would not be carried forward regardless of scores on other criteria. We have provided comments in the earlier parts of this report on how this may be handled as 'unconditional' and 'conditional' pass if necessary, where the 'conditions' to be addressed would be clearly listed.

Lines 406-408 "Threshold evaluation was addressed by WGECO in 2013 (ICES 2013b), and a second table, adapted from the indicator table (ICES 2012) was produced. These were not given any weightings at the time, and these were developed at this workshop". It is unclear what exactly has been done during WKBENTH2 and how this has been done, relative to what was already available.

Lines 421-422 "Each criterion was evaluated against the WGECO/WGBIODIV table (ICES 2013a)". It would clarify to add the conclusion from this comparison. Some bullets read as mere cross-check; other bullets read as an evaluation of the suitability of the criteria (for further uptake in the analysis).

Lines 496-503. Somewhat unclear and most likely incomplete statement. Suggestion to delete because of its low relevance to the exercise.

Lines 591-597. It is difficult to fully understand what has been done here, e.g. where does table 3.2.2 come from and what process was used link 3.2.2 "approaches" to 3.2.1 "evaluation criteria".

Weassume this will be less of an issue for the WKBENTH2 participants that will also contribute to WKBENTH3, but it may need some further explanation particularly for potentially new participants. Reference to Section 4.3 could be made because this is where the missing piece of the puzzle isfound.

Line 592 "This was based on work carried out at WGECO in 2013 (ICES 2013b)": Has there been any adaptation to what was reported by WGECO (2013)? If yes, it would be good to have that elaborated in the report (cf. to provide maximum clarity).

Lines 529-531 "WKBENTH1 considered it desirable for an indicator to integrate multiple pressures, while WGECO/WGBIODIV felt that "specificity" was the critical factor. Both positions have merit,…". We agree with this point of view. It is suggested to cross-check this decision with the MSFD expectations, where "specificity" may be an explicit requirement.

Table 3.1.1, Criterion 8: If the answer to 6a is "B", then (most likely?) criterion 8 should be scored <1. The scoring of criterion 8 seems to contradict the flexibility inherently adopted by criterion 6a.

Table 3.1.1, Criterion 9 "the indicator is easy to understand and communicate": While (the concept of) the indicator needs to be easy to understand and communicate, this does not necessarily hold true for the algorithm (or method). Given the somewhat interchangeable use of "indicator" and "method" throughout the document, this may need to be elaborated in the table.

Lines 573-580 "…For this reason, the report only included the evaluation where experts in the indicators AND the criteria were included" : The need for an expert in the evaluation criteria to be applied seems to be problematic. This issue could potentially (partly) be solved by further elaborating the text explaining each of the criteria, so they become unambiguous to non-experts as well.

Table 3.2.1: Would there be any value in normalizing the scoring for each of the categories; this to equalize the different aspects of a good indicator (rather than to put more emphasis on those aspects for which more criteria have been defined)?

Section 4

The range of options for setting thresholds and their pros and cons are well covered. As the text states at line 848 one threshold, notably loss, has already been defined as a policy decision and states what is effectively a limit reference point. It would be extremely helpful to develop this further in the context of the WKBENTH3 activity and the drafting of advice.

Key questions related to threshold setting are whether the point at which recovery time will exceed 12 years (and therefore the point at which habitat is defined as 'lost') can be determined with an indicator (and associated methods) selected and how precaution is introduced (including 'how much precaution')? Some precaution is expected as defined in 2017/848 Article 4 Para 1(e) [thresholds shall] "(e) be set on the basis of the precautionary principle, reflecting the potential risks to the marine environment." Other thresholds may be needed, but one for loss and one for loss plus precaution would seem to be a minimum set already defined by the MSFD for DC61.

If the seabed is not in the unimpacted state, and recovery time is less than 12 years, then the seabed will be classified as disturbed (DC62): "Physical disturbance shall be understood as a change to the seabed from which it can recover if the activity causing the disturbance pressure ceases".

There is still a question 'recovery time to what' that does not appear to be explicitly addressed in MSFD or 2017/848 and would be needed to set the threshold, though expected options are assumed to be close to the unimpacted state given this is necessarily the seabed that would not be classified as 'disturbed' or 'lost'.

We suggest the above line of reasoning is much more prominent in the next steps of this work. Such an approach may also help with intercomparison, collation, MSFD reporting and so on, because different indicators may be used regionally to determine locations of habitat loss and disturbance based on the same threshold for recovery time plus a defined uncertainty buffer.

If a threshold for loss plus precaution were identified (e.g. threshold indicator value associated with 95% probability that recovery time does not exceed 12 years), an important question in relation to the thresholds being discussed in Section 4 is whether the resulting value of the threshold would also be close to a target associated with other ‘desirable’ properties of the seabed (such as given in the example sections covering the extent of natural variation). If this were the case and shown with evidence, then less resources may need to be invested in proposing options for more complex targets and managing the complex debate about what they should be. Although WKBENTH2 give some emphasis to the extent of natural variation as a means of defining thresholds, this type of approach is very monitoring intensive at the scales considered for MSFD. An alternate option would be a threshold set on the basis of a defined probability of avoiding loss, coupled with case studies to understand the relationship between this threshold and the values of the indicator associated with natural variation.

The preceding comments relate to a ‘quality’ threshold of course, and do not address the setting of thresholds for ‘extent’. There appear to be no policy decisions thus far to guide progress on extent. Types of evidence sought are likely to be similar to those that have been used, in some cases controversially, to define area targets for MPA coverage. The use of arguments about connectivity, as highlighted in WKBENTH2, provide a science base that could lead to a presentation of options, though applying these in general terms will be a significant challenge. An obvious scientific point is whether the distribution of a given percentage loss will determine its implications and how this should be handled (eg. patches as opposed to one contiguous area in the assessment unit).

Lines 642-643 “Thresholds are defined here as the state at which an ecosystem transitions from a good to a degraded state”: Most likely "and/or extent" is to be added because the threshold(s) for both state and extent are to be considered. In general, “state” has sometimes been used in its widest sense (i.e. including also extent), sometimes in its more narrow sense (i.e. excluding extent) which complicates a correct understanding of the text.

Lines 653-665 and 788-791: The delivery of ecosystem services has not been listed as another concept to think about in relation to thresholds. There is an opportunity here to link to ecosystem services, with reference to how ecosystem services loss link to ecosystem function loss. The science is not quite there but efforts to link ecosystem services to ecosystem function are ongoing. May be useful to keep this in mind as the knowledge base continues to grow.

Lines 813-815: We could argue that this approach does define "good enough" relative to societal costs but not to the societal benefits delivered by well-functioning ecosystems. Here is where the ecosystem services approach ("how much do we need?") may come in.

Lines 819-824 “Sustainable use should therefore not be conflated with good environmental state”: Point taken. You may however argue that maximum sustainable yield as used in a fisheries context is different from "how much do we need" embracing "all" ecosystem services.

Lines 857-859 on expert judgement in threshold setting… “Advantages of such an approach is the low demand for data, but this approach can be subjective, inconsistent and open to bias (Dorrough *et al.*, 2020)”. The process to get to the expert judgement (consensus) may considerably help its objectification. This may be elaborated here.

Lines 969-972 “Most time-series will also need to be detrended because long-term changes that are related to for example climate change will be causing long term increases or decreases. This was not considered a weakness of the approach, because it is a way of dealing with multiple pressures that are operating at different spatial scales". This statement hints that GES is to be related only to regionally manageable pressure. It hence ignores that anthropogenic effects playing at scales larger-than-regional scales that may also change the ecological state of the marine environment. While there indeed is some logic behind this statement, it would be helpful to assess if this follows the MSFD philosophy.

Lines 1139-1140. It would be informative to also have some figures on variation around the average range in natural variation in the text.

Section 5

In general, the preparation of the datasets described is appropriate to support the next steps of the work and a reasonably broad range of geographies and depths are identified (although not quite as diverse as the range of depths and pressures proposed, perhaps optimistically, in the request for advice to ICES). The datasets identified in WKBENTH2 for use in WKBENTH3 would be suitable for assessing the effects of trawling disturbance (or comparable forms of abrasion), rather than a wider range of pressures discussed in the request. Trawling is, however, a good example for the testing in most regions and subregions given emphasis in 2007/848 that Member States "focus their efforts on the main anthropogenic pressures affecting their waters" and should "have sufficient flexibility, under specified conditions, to focus on the predominant pressures and their environmental impacts on the different ecosystem elements in each region or subregion…."

Section 6

The template provided the basis of a good technical service, although we lacked expertise in many of the indicators reviewed in the technical service document. Note comments on treatment of uncertainty under the review of Section 3. The request does ask ICES to consider "how confidence and uncertainty are handled" without further discussion, but it is important whether this handling of uncertainty is appropriate to what the MSFD seeks to achieve (as in previous comments we link the consideration of uncertainty primarily to the threshold). The work that follows from the provision of this template does address the request, although there would be added utility from the WKBENTH2 and Technical Service work as a whole if criteria scorings for the indicators and thresholds (when proposed) for all the same indicators were included in Section 3 (rather than a subset).

The process for collation of indicators as described seems thorough, though we do not have the combined expertise to comment reliably on completeness.

Section 7

It is pragmatic and reasonable for WKBENTH2 to conclude that it is unlikely that sampling will ever provide a representative picture of seafloor status at the regional scale where required for D6, so it is necessary to have an emphasis on models and extrapolation and addressing the uncertainties associated with this. Otherwise, Section 7 summarises topics already addressed in the review of preceding sections.

Annex A. Examples of the distinction between indicator, method, application, and threshold.

Example A – biomass-based example:

- Indicator: The biomass of a sentinel species might be a candidate indicator. Criteria that apply to the indicator might include if it is suitable to represent the health of the benthic ecosystem.

- Method: Many methods could be used to calculate the biomass of this sentinel species. The method itself will have many criteria including is it peer reviewed, what are the assumptions, has sensitivities to assumptions been carried out, is it replicable, documented etc.

- Application: The application of the method deals with what data are used in the application and at what scale the calculation is made for example. Criteria might include if the data are representative of the underlying processes, or if the calculation is made at a fine-enough temporal and spatial scale to be meaningful.

- Thresholds: Finally, thresholds need to deal with the spatial and temporal scale of that calculation. For example, a threshold might be that the biomass of that sentinel species at the scale of the population stays above 50% of a reference biomass with a 10% risk of dropping below 20% of some defined reference biomass. Another threshold could be that the relative benthic status does not drop below 20% in more than 20% of the entire range of the indicator species, or that it does not drop below 20% in more than 50% of the fished area in each habitat type. Another threshold could be the biomass that recovers to unimpacted biomass if all impacts are stopped for 12 years.

Example B – Some questions that arise from the example in Figure 5.1 in the report

- Indicator: proportion of sentinel species. Is this indicator suitable over multiple habitat types for example? Does it capture degradation of habitat or ecosystem processes adequately?

- Method: what species are counted and ignored? At what identification level are these required and does that level of identification need to be constant over time or over different areas to make the indicator comparable? Are there different ways to compute that indicator? What is the scale of the calculation?

- Application: what sampling regime is adequate spatially (does a sampling regime of 0.0175m^2 really capture biological processes) and can you compare between different sampling scales?

- Threshold: Over what scale is the threshold applied and how are the results of the calculation over that scale summarised? How is the variability captured in the threshold? Is the threshold value transferable between different habitats or even between different studies?

Annex B. How thresholds perform for selected indicators

This Annex provides a worked example of applying a 'natural variation' threshold, with reference to Figure 5.1 in the WKBENTH2 report (reproduced below).

An approximation of the 'natural variation' thresholds has been added to the figure. Green = 75%ile of very low pressure (we did not use the 95%ile as suggested in the text because it is difficult to figure out where it is in the figure) and red is 0.8 times that value (at a scale which starts at -0.2).

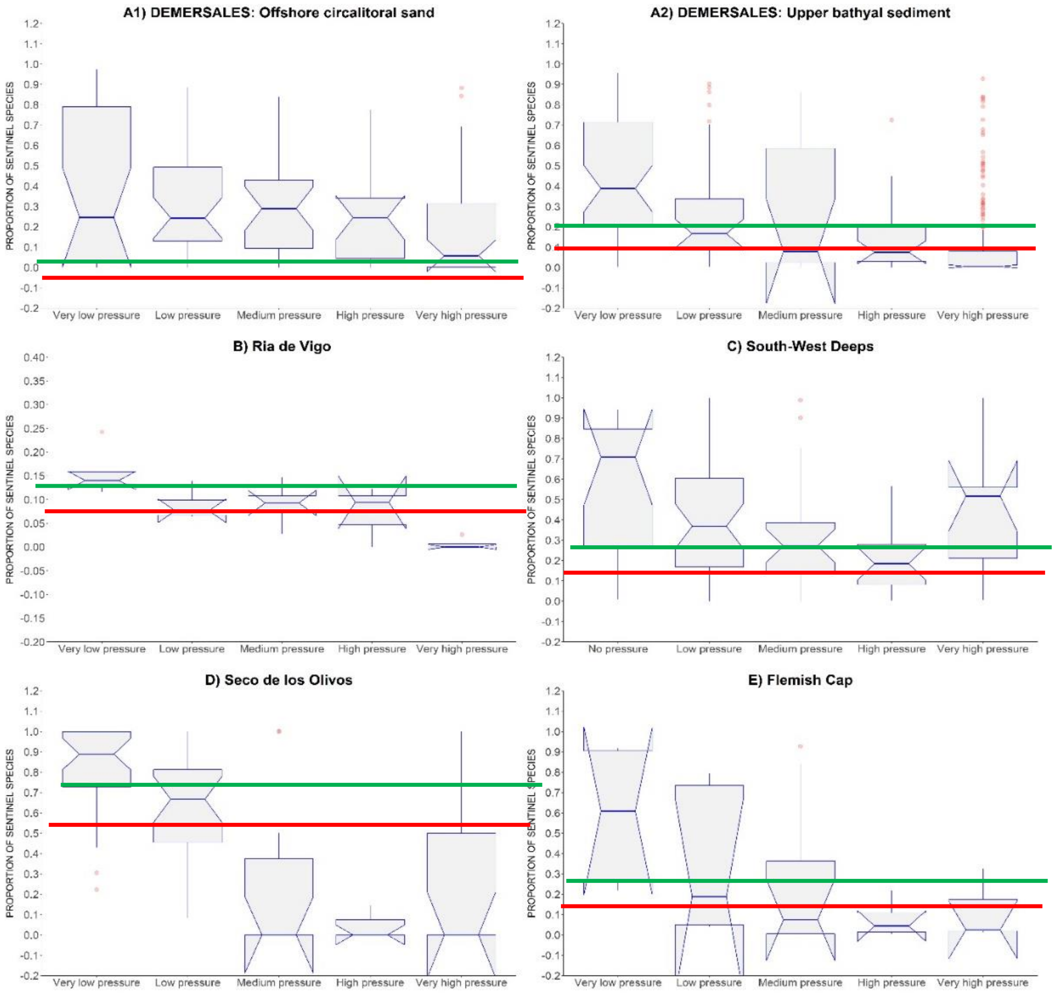
Panel A1 in Figure 5.1 shows no results below the threshold. Does that mean that even high pressure in this environment has limited effect or that the indicator is not suitable? Or is sampling inadequate to measure this change?

Panel D1 has all but low pressure below the threshold. Does that mean that only low pressure should be allowed? Is this environment more susceptible to trawling impacts than A1? Or is it that sampling captures change better?

In all other panels, results are partly above and partly below the red line. What constitutes a fail? Is it any point or a proportion of points or some other rule?

The area sampled varied between 0.0175m^2 and 0.04km^2. What effect have the sampled area and sample size (not reported) on the results and outcome? Can a sampling regime of 0.0175m^2 represent the biological processes? Should the thresholds be linked to the sampling regime?

Do we conclude that proportion of sentinel species is not a good indicator? Or just not a good indicator with regards to some specific habitat, level of habitat degradation, sampling type, some other reason, or a combination of all?



Reference

Yates, K.L.; Bouchet, P.J.; Caley, M.J.; Mengersen, K.; Randin, C.F.; Parnell, S.; Fielding, A.H.; Bamford, A.J.; Ban, S.; Barbosa, A.M.; Dormann, C.F.; Elith, J.; Embling, C.B.; Ervin, G.N.; Fisher, R.; Gould, S.; Graf, R.F.; Gregr, E.J.; Halpin, P.N.; Heikkinen, R.K.; Heinänen, S.; Jones, A.R.; Krishnakumar, P.K.; Lauria, V.; Lozano-Montes, H.; Mannocci, L.; Mellin, C.; Mesgaran, M.B.; Moreno-Amat, E.; Mormede, S.; Novaczek, E.; Oppel, S.; Ortuño Crespo, G.; Peterson, A.T.; Rapacciuolo, G.; Roberts, J.J.; Ross, R.E.; Scales, K.L.; Schoeman, D.; Snelgrove, P.; Sundblad, G.; Thuiller, W.; Torres, L.G.; Verbruggen, H.; Wang, L.; Wenger, S.; Whittingham, M.J.; Zharikov, Y.; Zurell, D.; Sequeira, A.M.M. (2018). Outstanding Challenges in the Transferability of Ecological Models. Trends in Ecology and Evolution 33, 790–802. https://doi.org/10.1016/j.tree.2018.08.001