Manuscript Number: ECOLIND-5882R1

Title: Confidence in ecological indicators: A framework for quantifying uncertainty components from monitoring data

Article Type: Research Paper

Keywords: Compliance probability; Indicator standardization; Marine Strategy Framework Directive; Monitoring programs; Status assessment; Water Framework Directive

Corresponding Author: Dr. Jacob Carstensen, Ph.D.

Corresponding Author's Institution: Aarhus University

First Author: Jacob Carstensen, Ph.D.

Order of Authors: Jacob Carstensen, Ph.D.; Mats Lindegarth

Abstract: The value of an ecological indicator is no better than the uncertainty associated with its estimate. Nevertheless, indicator uncertainty is seldom estimated, even though legislative frameworks such as the European Water Framework Directive stress that the confidence of an assessment should be quantified. We introduce a general framework for quantifying uncertainties associated with indicators employed to assess ecological status in waterbodies. The framework is illustrated with two examples: eelgrass shoot density and chlorophyll a in coastal ecosystems. Aquatic monitoring data vary over time and space; variations that can only partially be described using fixed parameters, and remaining variations are deemed random. These spatial and temporal variations can be partitioned into uncertainty components operating at different scales. Furthermore, different methods of sampling and analysis as well as people involved in the monitoring introduce additional uncertainty. We have outlined 18 different sources of variation that affect monitoring data to a varying degree and are relevant to consider when quantifying the uncertainty of an indicator calculated from monitoring data. However, in most cases it is not possible to estimate all relevant sources of uncertainty from monitoring data from a single ecosystem, and those uncertainty components that can be quantified will not be well determined due to the lack of replication at different levels of the random variations (e.g. number of stations, number of years, and number of people). For example, spatial variations cannot be determined from datasets with just one station. Therefore, we recommend that random variations are estimated from a larger dataset, by pooling observations from multiple ecosystems with similar characteristics. We also recommend accounting for predictable patterns in time and space using parametric approaches in order to reduce the magnitude of the unpredictable random components and reduce potential bias introduced by heterogeneous monitoring across time. We propose to use robust parameter estimates for both fixed and random variations, determined from a large pooled dataset and assumed common across the range of ecosystems, and estimate a limited subset of parameters from ecosystem-specific data. Partitioning the random variation onto multiple uncertainty components is important to

obtain correct estimates of the ecological indicator variance, and the magnitude of the different components provide useful information for improving methods applied and design of monitoring programs. The proposed framework allows comparing different indicators based on their precision relative to the cost of monitoring.

Response to Reviewers: Reviewers' comments are in black and our responses in blue.
Reviewer #1:
The central message of this manuscript is that it is important to adequately quantify the different major uncertainty components in ecological monitoring data. The manuscript discusses the inherent challenges and some solutions and why this issue matters. Although resolving contributions from different uncertainty components is not novel in itself this is a useful and very well written manuscript that develops a number of lines of thinking and is probably one of the most comprehensive and helpful studies addressing this subject. The chosen case studies are ideal because between them they capture many of the key elements of pelagic (chlorophyll) and benthic (seagrass) monitoring.

Thank you.

The only significant point that I feel is neglected is an acknowledgment that uncertainty in monitoring data is not the only form of uncertainty that contributes to uncertainty in status classifications. Even if monitoring data could be obtained, theoretically, without error, there would be uncertainty in status classifications due to uncertainty in reference conditions (e.g. associated with the models used to predict reference conditions and the measurement error associated with model predictors). Many compositional metrics also use some form of species score, which though not specifically relevant here, introduces a further source of uncertainty to information derived from monitoring data as these species scores themselves are inherently uncertain. Perhaps a comment that addresses these points would be appropriate in section 2.

We acknowledge that uncertainty associated with reference conditions and class boundaries will also affect the confidence of a status assessment from a scientific point-of-view. In practice, these boundaries are treated as fixed values since they are typically incorporated into the legislative framework. Most legislative frameworks introduce fixed target values (speed limits, tolerances in manufacturing, etc.) although the scientific results underpinning these values are associated with uncertainty. We have added a paragraph in the introduction to address this issue. It reads:
"Boundaries or target values defining different status classes are typically derived from model results, historical data or expert judgement and are, as such, inherently uncertain. In theory, this uncertainty associated with boundary or target values will also affect the confidence of the status assessment. In practice, however, these boundaries, based on best scientific knowledge, are typically formulated as part of the legislative framework and for the purpose of status assessment considered fixed values, despite the inherent uncertainty. "

Page 3, line 30 - should be installed

Corrected.

Page 3, line 56 etc. There are some additional parallel analyses for river macrophytes in Davey & Garrow (2009) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291734/scho1109brhq-e-e.pdf that it would be useful to refer to.

Thank you for the reference. We have added a sentence with reference to this study.

Page 4, line 13, allowing temporal variations to be quantified

Changed.

Page 4, line 30, in order to avoid biased assessments

Corrected.

Section 2.4 line 8, …and analysis also introduce uncertainty…. Corrected.

Section 2.4, line 12 - different personnel - whether performing the sampling (also relevant), or analysing samples

Amended the suggested change.

Section 2.4, line 15 - the assumption that methodological uncertainty is independent of season and location should probably be discussed. Some conditions are unquestionably easier to survey than others with a given method, and surveyor motivation tends to be reduced in the cold, dark and wet!

We believe it is fair to assume that the errors introduced by different methods are independent of season and location, e.g. the errors are not systematically negative or positive for certain months, years or specific stations. Monitoring methods are generally designed to be independent of the given monitoring conditions. However, it is possible that the magnitude of the methodological errors could increase under specific conditions, but we are not aware of any studies documenting such claims. We have commented on this issue in Section 2.4.

Section 2.4 line 24 - I suspect the act of sampling could be better described (at least statistically) as a device x operator interaction

True, this interaction describes the combination of devices and operators and hence includes the simpler model assuming additive effects of devices and operators. However, we maintain these two sources as independent main factors for simplicity and keeping the degrees of freedom low, and because we believe the interaction of device and operator will be marginal in most cases.

Section 2.4 line 38 - efforts have been made in the past to correct for the systematic recording bias of some personnel

We agree that this issue is (or should be) continuously addressed by most monitoring programs and we have added a sentence to recognize these efforts.

Generally I think that the transition from section 2 to section 3 would be easier for the reader to make if you had introduced some of the notation from Table 1 into section 2.

Point taken. We have now included the notation from Table 1 in section 2 as the different terms are being introduced.

Page 9, line 19, shouldn't it be gr

We have corrected it.

Section 3 develops a framework for estimating different sources of uncertainty. Does this allow for the fact that the variance in an indicator is unlikely to be randomly distributed over the gradient of that indicator (e.g. because it is likely to be bounded - certainly at zero and potentially at some upper value). Perhaps relevant to discussion at end of section 5.1.

We have added text in both Section 3 and Section 5.1, describing that the basic assumption of the model is that all parameters (including variance parameters) are constant, and this assumption can be justified through appropriate selection of data and transformation of these data.

Page 12, line 29, may potentially exhibit trends

Corrected.

Page 14: In terms of the analysis of spatial variation in eelgrass stem density some of this variation can be accounted for by depth. I suspect that substrate and fetch could also account for other variation that if not 'extracted' will be treated as random. Maybe it is worth emphasising that other easily measured environmental predictors should, where realistic, be integrated as fixed effects. Also relevant to section 5.2 on page 18.

We have addressed this in Section 5.2. We believe the case study should only relate to the available information, whereas in Section 5.2 we have discussed the possibility of reducing the random spatial variation further by including fetch and substrate as explanatory variables.

Page 16 line 51 - should be made (or obtained) rather than estimated

We have replaced ‘estimated’ with ‘obtained’.

Page 17, line 34 - state that you are referring to the eelgrass data here

Done.

Page 17, line 49, dependence

Corrected.

Page 18, line - desired properties are low bias and high precision

Corrected.

Table 1

Line 7 - assume should be gr (not g), Line 18 assume should be yr (not y) Generally the tables are not standalone as the codes for the variance components require reference to Table 1 or equations within the text.

Corrected gr and yr. In Section 2 we have introduced all the terms in the text as they appear with reference to Table 1. See also our response above.

Figure 1 and 2 - dots indicating individual sampling sites are hard to discern and may need to be shown larger and as different symbol types rather than colours

In preparation of this figure we carefully considered different symbol sizes and colors. Larger symbols would imply a much higher degree of overlapping. Since the symbols cannot be larger we do not believe that choosing different symbols would make it easier to distinguish sampling areas. We are confident that the symbols in the figure represent the best compromise between symbol, size and color.

Figure 5 - this would be better as a histogram with paired bars for estuaries and coasts for the different variance components. I'm not not sure what value the 'combined' line has.

We have now displayed variances in a bar plot as suggested. 'combined' is actually the entire dataset and we have changed the legend to reflect this.

Reviewer #2:
Your paper covers an actual aspect of long-term data analysis. The proposed GLMM approach is quite very interesting. But, often such additive model is not so comfortable because of missing information on sources of uncertainty. Give a clear and comprehensive red line how you got the results. Therefore, you should add more details in chapter 3.

We think the reviewer has misunderstood the objectives of the manuscript. The manuscript is not about long-term data analysis, but the formulation of a statistical framework for assessing uncertainties associated with typical aquatic monitoring data. The objectives are clearly articulated in the last paragraph of the introduction. Overall, this apparent misunderstanding recurs in the additional comments below. Furthermore, we do not understand why the reviewer has reservations against GLMM – GLMM actually allows for including different sources of uncertainty.

Additional comments to the Author:

1. The highlights are too general and do not characterise the content of your paper. The same is valid for the abstract.

This comment is not intelligible. It is very unspecific and difficult to address. We believe the comment is based on an overall misunderstanding of the objectives of the study (see last paragraph in the introduction). See also our response to point 6 below.

2. In the cases of general statements you add examples (pp. 3,5, 7, 8, 9, 10, 17, 19). Readers which are familiar with the topic do not need such examples for explanations. For the other one, your example is too short. Therefore, you should delete such examples.

We would appreciate if this comment was made more specific, pointing out sentences where examples were not needed or needed to be elaborated. We believe there is an adequate balance between describing the uncertainty framework and using examples to support the theoretical descriptions.

3. What does it mean "XX" on page 4?Take it off.

We have removed the 'XX' from the parenthesis.

4. Chapter 2 is too general and not really an overview on sources of uncertainty.

Table 1 lists the sources of uncertainty and these are described in Section 2. We have included the notation from Table 1 throughout the text to make this linkage clearer (see also our response to reviewer #1). Hopefully, this amendment will improve the readability to provide the overview.

5. In chapter 3 nothing is said the computational methods (hardware, software, computer time).

We do not find it relevant to include such information. Most statistical software packages can analyze GLMM (e.g. R, SAS, SPSS) and they run on an ordinary PC within a reasonable time. Computation time mostly depends on the number of observations and to a lesser degree on the complexity of the GLMM, but in our examples computation time was less than one minute.

6. Chapter 5 is too general and not an understandable discussion of your results. You stress more general aspects of quantifying uncertainty than a discussion of your experimental results. And what about sensitivity you mentioned on page 18? You should give some further comment concerning this essential aspect of long-term data analysis.

We believe this comment is based on a misunderstanding of our objective. The objective is to introduce a framework for quantification of different sources of uncertainty affecting typical aquatic monitoring data used for deriving indicators for status assessment. The two examples are included to demonstrate the usability of the framework. The intention is not to discuss the specific implications of the two examples regarding seasonal patterns or trends.

7. In section 5.2 you stress the modelling aspect. But, you should add some further comments.

We believe we have discussed and exemplified how modelling can be used to reduce uncertainty. If the comment was more specific, we would be happy to address it.

AARHUS UNIVERSITY

Ecological Indicators

Date: 2 March 2016

Journal no.:
Annual card no.:
CPR no..:
Receiver's CVR no.:
Sender's CVR no.:
Reference:

Resubmission of manuscript for Ecological Indicators

Dear Editor,

We hereby resubmit our manuscript ‘Confidence in ecological indicators: A framework for quantifying uncertainty components from monitoring data’ for consideration as an original research paper in Ecological Indicators.

Our resubmission includes the revised manuscript as well as a response letter addressing each comment from the two reviewers separately. We would like to emphasize that we found the comments from reviewer #2 unspecific and unsubstantial, and that the critique of this reviewer is most likely based on a misunderstanding of the objectives of the manuscript. The objectives are clearly stated in the last paragraph of the introduction. If the paper will be sent out for a second round of review, we recommend that an alternative reviewer is chosen or that the editor provides some guidance to reviewer #2.

On behalf of the authors,

Yours sincerely,

Jacob Carstensen
Dept. of Bioscience
Aarhus University
Frederiksborgvej 399
DK-4000 Roskilde
DENMARK
Phone: (+45) 46301345
Fax: (+45) 46301114
email: jac@bios.au.dk

Responses to reviewers' comments

Reviewers' comments are in black and our responses in blue.

Reviewer #1:

The central message of this manuscript is that it is important to adequately quantify the different major uncertainty components in ecological monitoring data. The manuscript discusses the inherent challenges and some solutions and why this issue matters. Although resolving contributions from different uncertainty components is not novel in itself this is a useful and very well written manuscript that develops a number of lines of thinking and is probably one of the most comprehensive and helpful studies addressing this subject. The chosen case studies are ideal because between them they capture many of the key elements of pelagic (chlorophyll) and benthic (seagrass) monitoring.

Thank you.

The only significant point that I feel is neglected is an acknowledgment that uncertainty in monitoring data is not the only form of uncertainty that contributes to uncertainty in status classifications. Even if monitoring data could be obtained, theoretically, without error, there would be uncertainty in status classifications due to uncertainty in reference conditions (e.g. associated with the models used to predict reference conditions and the measurement error associated with model predictors). Many compositional metrics also use some form of species score, which though not specifically relevant here, introduces a further source of uncertainty to information derived from monitoring data as these species scores themselves are inherently uncertain. Perhaps a comment that addresses these points would be appropriate in section 2.

We acknowledge that uncertainty associated with reference conditions and class boundaries will also affect the confidence of a status assessment from a scientific point-of-view. In practice, these boundaries are treated as fixed values since they are typically incorporated into the legislative framework. Most legislative frameworks introduce fixed target values (speed limits, tolerances in manufacturing, etc.) although the scientific results underpinning these values are associated with uncertainty. We have added a paragraph in the introduction to address this issue. It reads:
"Boundaries or target values defining different status classes are typically derived from model results, historical data or expert judgement and are, as such, inherently uncertain. In theory, this uncertainty associated with boundary or target values will also affect the confidence of the status assessment. In practice, however, these boundaries, based on best scientific knowledge, are typically formulated as part of the legislative framework and for the purpose of status assessment considered fixed values, despite the inherent uncertainty. "

Page 3, line 30 - should be installed

Corrected.

Page 3, line 56 etc. There are some additional parallel analyses for river macrophytes in Davey & Garrow (2009)
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291734/scho1109brhq-e-e.pdf
that it would be useful to refer to.

Thank you for the reference. We have added a sentence with reference to this study.

Page 4, line 13, allowing temporal variations to be quantified

Changed.

Page 4, line 30, in order to avoid biased assessments

Corrected.

Section 2.4 line 8, …and analysis also introduce uncertainty….
Corrected.

Section 2.4, line 12 - different personnel - whether performing the sampling (also relevant), or analysing samples

Amended the suggested change.

Section 2.4, line 15 - the assumption that methodological uncertainty is independent of season and location should probably be discussed. Some conditions are unquestionably easier to survey than others with a given method, and surveyor motivation tends to be reduced in the cold, dark and wet!

We believe it is fair to assume that the errors introduced by different methods are independent of season and location, e.g. the errors are not systematically negative or positive for certain months, years or specific stations. Monitoring methods are generally designed to be independent of the given monitoring conditions. However, it is possible that the magnitude of the methodological errors could increase under specific conditions, but we are not aware of any studies documenting such claims. We have commented on this issue in Section 2.4.

Section 2.4 line 24 - I suspect the act of sampling could be better described (at least statistically) as a device x operator interaction

True, this interaction describes the combination of devices and operators and hence includes the simpler model assuming additive effects of devices and operators. However, we maintain these two sources as independent main factors for simplicity and keeping the degrees of freedom low, and because we believe the interaction of device and operator will be marginal in most cases.

Section 2.4 line 38 - efforts have been made in the past to correct for the systematic recording bias of some personnel

We agree that this issue is (or should be) continuously addressed by most monitoring programs and we have added a sentence to recognize these efforts.

Generally I think that the transition from section 2 to section 3 would be easier for the reader to make if you had introduced some of the notation from Table 1 into section 2.

Point taken. We have now included the notation from Table 1 in section 2 as the different terms are being introduced.

Page 9, line 19, shouldn't it be gr

We have corrected it.

Section 3 develops a framework for estimating different sources of uncertainty. Does this allow for the fact that the variance in an indicator is unlikely to be randomly distributed over the gradient of that indicator (e.g. because it is likely to be bounded - certainly at zero and potentially at some upper value). Perhaps relevant to discussion at end of section 5.1.

We have added text in both Section 3 and Section 5.1, describing that the basic assumption of the model is that all parameters (including variance parameters) are constant, and this assumption can be justified through appropriate selection of data and transformation of these data.

Page 12, line 29, may potentially exhibit trends

Corrected.

Page 14: In terms of the analysis of spatial variation in eelgrass stem density some of this variation can be accounted for by depth. I suspect that substrate and fetch could also account for other variation that if not 'extracted' will be treated as random. Maybe it is worth emphasising that other easily measured environmental predictors should, where realistic, be integrated as fixed effects. Also relevant to section 5.2 on page 18.

We have addressed this in Section 5.2. We believe the case study should only relate to the available information, whereas in Section 5.2 we have discussed the possibility of reducing the random spatial variation further by including fetch and substrate as explanatory variables.

Page 16 line 51 - should be made (or obtained) rather than estimated

We have replaced ‘estimated’ with ‘obtained’.

Page 17, line 34 - state that you are referring to the eelgrass data here

Done.

Page 17, line 49, dependence

Corrected.

Page 18, line - desired properties are low bias and high precision

Corrected.

Table 1
Line 7 - assume should be gr (not g), Line 18 assume should be yr (not y) Generally the tables are not standalone as the codes for the variance components require reference to Table 1 or equations within the text.

Corrected gr and yr. In Section 2 we have introduced all the terms in the text as they appear with reference to Table 1. See also our response above.

Figure 1 and 2 - dots indicating individual sampling sites are hard to discern and may need to be shown larger and as different symbol types rather than colours

In preparation of this figure we carefully considered different symbol sizes and colors. Larger symbols would imply a much higher degree of overlapping. Since the symbols cannot be larger we do not believe that choosing different symbols would make it easier to distinguish sampling areas. We are confident that the symbols in the figure represent the best compromise between symbol, size and color.

Figure 5 - this would be better as a histogram with paired bars for estuaries and coasts for the different variance components. I'm not not sure what value the 'combined' line has.

We have now displayed variances in a bar plot as suggested. ‘combined’ is actually the entire dataset and we have changed the legend to reflect this.

Reviewer #2:

Your paper covers an actual aspect of long-term data analysis. The proposed GLMM approach is quite very interesting. But, often such additive model is not so comfortable because of missing information on sources of uncertainty. Give a clear and comprehensive red line how you got the results. Therefore, you should add more details in chapter 3.

We think the reviewer has misunderstood the objectives of the manuscript. The manuscript is not about long-term data analysis, but the formulation of a statistical framework for assessing uncertainties associated with typical aquatic monitoring data. The objectives are clearly articulated in the last paragraph of the introduction. Overall, this apparent misunderstanding recurs in the additional comments below. Furthermore, we do not understand why the reviewer has reservations against GLMM – GLMM actually allows for including different sources of uncertainty.

Additional comments to the Author:

1. The highlights are too general and do not characterise the content of your paper. The same is valid for the abstract.

This comment is not intelligible. It is very unspecific and difficult to address. We believe the comment is based on an overall misunderstanding of the objectives of the study (see last paragraph in the introduction). See also our response to point 6 below.

2. In the cases of general statements you add examples (pp. 3,5, 7, 8, 9, 10, 17, 19). Readers which are familiar with the topic do not need such examples for explanations. For the other one, your example is too short. Therefore, you should delete such examples.

We would appreciate if this comment was made more specific, pointing out sentences where examples were not needed or needed to be elaborated. We believe there is an adequate balance between describing the uncertainty framework and using examples to support the theoretical descriptions.

3. What does it mean "XX" on page 4?Take it off.

We have removed the ‘XX’ from the parenthesis.

4. Chapter 2 is too general and not really an overview on sources of uncertainty.

Table 1 lists the sources of uncertainty and these are described in Section 2. We have included the notation from Table 1 throughout the text to make this linkage clearer (see also our response to reviewer #1). Hopefully, this amendment will improve the readability to provide the overview.

5. In chapter 3 nothing is said the computational methods (hardware, software, computer time).

We do not find it relevant to include such information. Most statistical software packages can analyze GLMM (e.g. R, SAS, SPSS) and they run on an ordinary PC within a reasonable time. Computation time mostly depends on the number of observations and to a lesser degree on the complexity of the GLMM, but in our examples computation time was less than one minute.

6. Chapter 5 is too general and not an understandable discussion of your results. You stress more general aspects of quantifying uncertainty than a discussion of your experimental results. And what about sensitivity you mentioned on page 18? You should give some further comment concerning this essential aspect of long-term data analysis.

We believe this comment is based on a misunderstanding of our objective. The objective is to introduce a framework for quantification of different sources of uncertainty affecting typical aquatic monitoring data used for deriving indicators for status assessment. The two examples are included to demonstrate the usability of the framework. The intention is not to discuss the specific implications of the two examples regarding seasonal patterns or trends.

7. In section 5.2 you stress the modelling aspect. But, you should add some further comments.

We believe we have discussed and exemplified how modelling can be used to reduce uncertainty. If the comment was more specific, we would be happy to address it.

Resubmitted for Ecological Indicators

Confidence in ecological indicators: A framework for quantifying uncertainty components from monitoring data

Jacob Carstensen[a*], Mats Lindegarth[b]

[a]Department of Bioscience, Aarhus University, Frederiksborgvej 399, DK-4000 Roskilde, Denmark; Email: jac@bios.au.dk

[b]Department of Biology and Environmental Science, Gothenburg University, SE-45296 Strömstad, Sweden; Email: mats.lindegarth@marine.gu.se

[*]Corresponding author

Keywords: Compliance probability, Indicator standardization, Marine Strategy Framework Directive, Monitoring programs, Status assessment, Water Framework Directive

Highlights:

- Monitoring data are affected by many different sources of uncertainty
- Large datasets are needed to quantify all the different uncertainty components
- Indicator bias and uncertainty can be reduced through improved modeling
- Monitoring programs can be optimized to obtain more precise indicators

Abstract

The value of an ecological indicator is no better than the uncertainty associated with its estimate. Nevertheless, indicator uncertainty is seldom estimated, even though legislative frameworks such as the European Water Framework Directive stress that the confidence of an assessment should be quantified. We introduce a general framework for quantifying uncertainties associated with indicators employed to assess ecological status in waterbodies. The framework is illustrated with two examples: eelgrass shoot density and chlorophyll a in coastal ecosystems. Aquatic monitoring data vary over time and space; variations that can only partially be described using fixed parameters, and remaining variations are deemed random. These spatial and temporal variations can be partitioned into uncertainty components operating at different scales. Furthermore, different methods of sampling and analysis as well as people involved in the monitoring introduce additional uncertainty. We have outlined 18 different sources of variation that affect monitoring data to a varying degree and are relevant to consider when quantifying the uncertainty of an indicator calculated from monitoring data. However, in most cases it is not possible to estimate all relevant sources of uncertainty from monitoring data from a single ecosystem, and those uncertainty components that can be quantified will not be well determined due to the lack of replication at different levels of the random variations (e.g. number of stations, number of years, and number of people). For example, spatial variations cannot be determined from datasets with just one station. Therefore, we recommend that random variations are estimated from a larger dataset, by pooling observations from multiple ecosystems with similar characteristics. We also recommend accounting for predictable patterns in time and space using parametric approaches in order to reduce the magnitude of the unpredictable random components and reduce potential bias introduced by heterogeneous monitoring across time. We propose to use robust parameter estimates for both fixed and random variations, determined from a large pooled dataset and assumed common across the range of ecosystems, and estimate a limited subset of parameters from ecosystem-specific data. Partitioning the random variation onto multiple uncertainty components is important to obtain correct estimates of the ecological indicator variance, and the magnitude of the different components provide useful information for improving methods applied and design of monitoring programs. The proposed framework allows comparing different indicators based on their precision relative to the cost of monitoring.

1 INTRODUCTION

The growing human pressure on nature's ecosystems has prompted increasing monitoring efforts to assess consequences of human activities in order to protect these systems from degradation and potential collapse. Measurement programs for aquatic ecosystems were initiated throughout the western world in the 20th century to monitor pollution effects, starting with physical-chemical characteristics in urban rivers and streams (e.g. Delaware River near Philadelphia; Sharp 2010) and lakes (e.g. Great Lakes; Chapra et al. 2012). As downstream pollution effects became evident monitoring programs were similarly established in estuaries (e.g. Chesapeake Bay; Hagy et al. 2004) and coastal waters (e.g. Tampa Bay; Greening et al. 2014).Over the years the physical-chemical monitoring has increasingly been supplemented with biological measurements. These monitoring programs were originally initiated with the aim to assess changes over time in various indicators compiled from the monitoring data; often relying on a single monitoring point or station assuming the trends of this location to be representative of the larger system. However, the actual indicator level may not represent the mean of the entire system as temporal evolution was only considered.

In recent years the objective of many monitoring programs have changed from reporting trends to reporting status, and to assess the compliance of the status with established target values. For instance, in the European Water Framework Directive (WFD, European Commission 2000) and Marine Strategy Framework Directive (MSFD, European Commission 2008) the status during a 6-year assessment period should be compared to the boundary between "Good" and "Moderate" ecological status (WFD) or the boundary between "Good" and below "Good" environmental status (MSFD). Both directives mandate that management measures should be installed in case "Good" ecological/environmental status is not achieved. Thus, the two directives are unique by directly linking assessment of monitoring data and management. Since management measures can be quite costly and the failure of mitigating pollution in time can have severe ecological consequences, it is important that the confidence in the status assessment is sufficient (European Commission 2000). However, the confidence can only be assessed provided that the uncertainties of the indicator values used in the status assessment are quantified.

Boundaries or target values defining different status classes are typically derived from model results, historical data or expert judgement and are, as such, inherently uncertain. In theory, this uncertainty associated with boundary or target values will also affect the confidence of the status assessment. In practice, however, these boundaries, based on best scientific knowledge, are typically formulated as part of the legislative framework and for the purpose of status assessment considered fixed values, despite the inherent uncertainty.

It is well known that benthic and planktonic assemblages of animals and plants in aquatic environments vary at a broad range of spatial and temporal scales (e.g. Pinel-Alloul et al. 1988; Levin 1992; Morrisey et al. 1992; Downes et al. 1993; Norén and Lindegarth 2005). Despite this and considering the importance of knowing the confidence in ecological status assessments, it is surprising that the uncertainty of ecological indicators and the different sources contributing to this uncertainty are seldom quantified. Dromph et al. (2013) employed a specific sampling design to partition variation in measurements of phytoplankton density and pigments into variations among coastal waterbodies, among stations within waterbodies, among samples at each station, among sub-samples after splitting the samples, and among persons analyzing the sample (analysts). They found that spatial variations among stations contributed the most to the uncertainty within waterbodies; however, the study was carried out within two summer months implying that temporal variability was not truly addressed. In a similar study of lakes, Carvalho et al. (2013) and Thackeray et al. (2013) found that variations between stations within lakes was relatively small compared to variations between samples at the same station and variation between analysts. Dudley et al. (2013) examined four different macrophytes community indicators and found that the large-scale spatial variation among stations within lakes was generally larger than the small-scale spatial variation among transects within stations, but their study did not address temporal variations. Macrophyte communities in UK rivers also displayed large spatial variability and large operator variability, whereas temporal variations were relatively small (Davey and Garrow 2009). Using a macroinvertebrate dataset from the UK, Clarke et al. (2013) found that spatial, temporal and replicate variability were of same magnitude for two different community indicators.

In contrast to the abovementioned studies that were based on data specifically collected to quantify different uncertainty components, Balsby et al. (2013) used Danish monitoring data of eelgrass depth limits and reported that spatial variations (both large- and small-scale within coastal waterbodies) were largest, although interannual variation and variation between divers were also considerable. Monitoring data have the advantage of spanning multiple years (and different seasons for some measurements), allowing temporal variations to be quantified. Sampling programs specifically designed to estimate different uncertainty components often take excessive dimensions, if all relevant sources of uncertainty are included. For instance, interannual variation cannot be estimated from a single year of data (Carvalho et al. 2013; Dromph et al. 2013; Dudley et al. 2013; Thackeray et al. 2013), but this does not imply that this source of variation is negligible. In fact, omitting significant sources of variation will result in underestimating the overall indicator uncertainty. Thus, it is important to include all significant uncertainty components when assessing the uncertainty of an indicator.

One common feature of many contemporary assessment systems aimed at estimating and classifying the status of aquatic environments (e.g. the WFD and MSFD), is that they are intended to provide holistic, integrated assessments based on a number of ecological indicators (Borja et al. 2010). Depending on the exact context, this means that two or more indicators, each associated with particular uncertainties, are combined and aggregated and potentially contributing to uncertainty in the overall status classification (e.g. Caroni et al. 2013). In order to avoid biased assessments and to ensure that uncertainties can be estimated and assessed on a common scale, a unified quantitative approach accounting for spatial, temporal and methodological sources of variation is needed (e.g. Hering et al. 2010; Clarke 2013).

The objective of this study is to formulate a generic framework for assessing the uncertainty of an indicator compiled from aquatic monitoring data. We will describe a list of uncertainty components that are relevant to consider and show how subsets and/or combinations of these sources can be quantified. Furthermore, we will demonstrate how estimates of relevant variance components obtained from studies of larger datasets can be used to achieve more realistic estimates of uncertainty when only a small and potentially incomplete dataset is available for determining the status of an indicator. Finally, we will describe how the aggregation of different uncertainty components influences the overall indicator uncertainty. The framework will be exemplified and tested using both benthic and pelagic monitoring data.

2 SOURCES OF UNCERTAINTY IN MONITORING DATA

Characterizing an entire waterbody over a given assessment period (e.g. 6 years for WFD and MSFD) by means of an indicator compiled from a discrete number of samples is inherently uncertain, and the number of potential sources to this overall uncertainty can be large. However, whereas some sources of uncertainty are small and can be disregarded, other sources of uncertainty may contribute substantially to the overall indicator uncertainty (Table 1). Moreover, the different sources of variation can be described with a finite set of possible values (fixed effects with parameters for each possible outcome) or an infinite set of possible values (random effects described as stochastic distributions). Fixed effects do not contribute uncertainty to the indicator, but can, in general, be estimated and accounted for.

Any given measurement used for calculating an indicator is affected by uncertainties related to both sampling and analysis of the sample. Sampling produces a finite number of observations to characterize the infinite spatial-temporal distribution representing the entire waterbody over the assessment period. The key question is how well the data represent underlying spatial variations (large-scale gradients and small-scale patchiness), temporal variations (interannual, seasonal, diurnal), and the spatial-temporal interactions? After sampling, the analysis of the sample can introduce variations caused by using different sampling devices, instruments, analysts, sub-sampling and replication. Such differences may introduce uncertainty additional to that introduced by sampling.

2.1 Uncertainties associated with spatial variations

The ecological status assessment of a waterbody should apply to the entire waterbody and not just a few sampling locations. Since it is impossible to monitor every parcel of water or every square meter of the bottom, the ecological indicator is normally calculated using data from a few spatially distinctive monitoring locations. The water column is traditionally sampled at a number of stations scattered throughout the waterbody, whereas benthic monitoring often involves replicated sampling within different stations to account for the small-scale heterogeneity in benthic communities. Thus, spatial variations can be partitioned into large-scale gradients (gr and GR) within the waterbody (typically described by different stations) and small-scale fluctuations or patchiness (pa and $PA(GR)$) (typically described by replicated sampling within stations) (Table 1).

These small- and large-scale spatial patterns of variation can partly be explained by various environmental factors, assuming that the station-specific mean correlates with co-variables such as depth, substrate characteristics, and mean salinity. Such sources of variation are fixed (predictable) (gr and pa) and if they can be incorporated in the model, they do not add uncertainty to the indicator value. Small-scale patchiness can often be partitioned in fixed and random variation, provided that relevant co-variables are monitored together with the replicated samples. For example, variations in cover of macroalgae can often be explained by depth and the availability of hard substrate (e.g. Krause-Jensen et al. 2007, Svensson et al. 2013). However, it is unlikely that all large-scale variation can be explained and the remaining variation (GR) is considered random (unpredictable). Similarly, unexplained small-scale variation between replicated samples is considered random and described as a nested effect within the large-scale gradient ($PA(GR)$), since the exact location of replicated samples shifts between visits to the station.

2.2 Uncertainties associated with temporal variations

The ecological status assessment of a waterbody should also apply to the entire span of the assessment period and not just a finite number of sampling occasions in time. Most ecological variables display significant interannual and seasonal variation, and for some variables (e.g. phytoplankton) even diurnal variations (Table 1). The water column is typically sampled several times over the year due to the dynamic nature of pelagic processes, whereas benthic monitoring is typically carried out on an annual basis aiming to sample approximately the same period every year.

Interannual variation describes the variation between years of sampling and this variation is partly fixed (yr) and partly random (YR). The fixed variation can be described by external factors influencing the environmental time series such as temperature, freshwater discharge etc., whereas the random variation describes the remaining unexplained interannual variation.

Seasonal variation has a cyclic character repeating itself every year. It can also be partitioned into a fixed component (se) (typically discretized into monthly resolution), which is the mean seasonal variation, and a random component ($SE\times YR$), describing fluctuations around the mean seasonal variation. The fixed seasonal variation can be calculated if multiple years of data have been collected; otherwise the seasonal variation is entirely random. Accounting for the fixed seasonal variation substantially reduces the uncertainty of pelagic indicators (Carstensen 2007).

Similarly, diurnal variation has a cyclic character repeating itself every day. It can be partitioned into a fixed common variation (di) or fixed season-specific variation ($di\times se$). These fixed diurnal variations can be calculated if multiple years of data have been collected. Deviations from the fixed diurnal variations are considered random ($DI\times SE\times YR$). Co-variables such as daylight, temperature can also be used to describe the fixed diurnal variation. However, most monitoring variables are not sampled sufficiently frequent to resolve diurnal variations and therefore these temporal variations are seldom considered.

Irregular fluctuations (IR) are temporal variations that are not described by interannual, seasonal or diurnal patterns. For example, if a macroalgae transect is monitored twice within the same month and year the variation between the two dates is considered irregular fluctuations. Similarly, if the water column is sampled bi-weekly and the seasonality is described with a monthly resolution, the variation between samples from the same month and year is considered irregular fluctuations.

2.3 Uncertainties associated with spatial-temporal interactions

In addition to the pure spatial and temporal sources of variation, it is also relevant to consider their interactions, i.e. changes in spatial pattern over time or changes in temporal variations over space. In theory, there are many spatial-temporal interactions that could be considered (i.e. combinations of the spatial and temporal factors) but we have included only the two most relevant (Table 1), assuming that 1) small-scale patchiness is independent of all temporal variations, and 2) large-scale spatial gradients may change systematically between years and over the season but not over the course of the day.

The large-scale gradients can change between years ($GR \times YR$) due to various reasons, e.g. benthic vegetation in shallower waters might disappear for several years after a storm or intrusion of hypoxic water might eradicate deeper benthic communities. Interannual variations in freshwater inputs, temperature, wind-mixing and other external factors may also systematically affect the spatial gradients within a waterbody. Similarly, the large-scale gradients may also change seasonally ($GR \times SE$) in response to various external factors, e.g. the spatial distribution of phytoplankton biomass in an estuary varies seasonally.

2.4 Uncertainties associated with sampling and analysis methodology

The methods involved in sampling and analysis also introduce uncertainty to a varying degree depending on the specific type of measurement. Monitoring programs operating over several years typically involve different sampling devices, different instruments or procedures for the analysis of the sample, and different personnel - whether performing the sampling, or analyzing samples (Table 1). It is assumed that the uncertainties introduced by these different methodologies are independent of time and space, i.e. the methodological uncertainty is not changing systematically with year, season or spatial location of the sample. It is possible that the magnitude of methodological uncertainties could depend on time and space (e.g. measurement errors might be larger in areas with substantial wave movement of the research vessel), but such phenomena are most likely specific to certain types of measurements and therefore not included here.

Different sampling devices are typically employed in most monitoring program, since equipment is replaced over time and different institutions using different equipment can be involved in the monitoring. For example, integrated water samples can be collected using Niskin bottles (mixing discrete-depth samples) or a hose, cover of benthic vegetation can be assessed by video or diver recording, benthic fauna can be sampled with van Veen, Smith-McIntyre grabs or another similar device. The population of sampling devices is finite and differences are described using fixed factors (sd).

Similarly, different instruments (method, brand, model, etc.) can be employed for measuring the sample. For examples, chlorophyll can be measured using HPLC, spectrophotometry and fluorometry using equipment from different manufacturers having different models. Nevertheless, the population of analytical instruments used is finite and differences are described as fixed factors (ai).

Involving different people in the sample analysis is unavoidable and inflates the uncertainty to a varying degree. Whereas the variation between technicians analyzing hydrochemistry is considered relatively small, variation is considerable between taxonomists analyzing phytoplankton (Jakobsen et al. in press) or macroinvertebrates (Haase et al. 2006) as well between divers surveying eelgrass transects (Balsby et al. 2013). The number of analysts that have and will be involved in monitoring programs is essentially infinite and therefore this variation is random (AN). It should be stressed that monitoring programs continuously strive to keep this random variation low by training and inter-comparison among analysts.

Occasionally, duplicate or triplicate measurements of the sample are carried out or the sample is split and each sub-sample is analyzed. This procedure is normally pursued when the analytical uncertainty is relatively large. The variation between these replicated measurements of the same sample is random (RE).

3 UNCERTAINTY FRAMEWORK

An observation from a monitoring program (y) is potentially influenced by all the relevant sources of variation described in Section 2 (using the notation from Table 1):

$$\begin{aligned} y = \mu + & \underbrace{gr + GR + pa + PA(GR)}_{spatial\ sources\ of\ uncertainty} \\ + & \underbrace{yr + YR + se + SE \times YR + di + di \times se + DI \times SE \times YR + IR}_{temporal\ sources\ of\ uncertainty} \\ + & \underbrace{GR \times YR + GR \times SE}_{spatial-temporal\ interactions} \\ + & \underbrace{sd + ai + AN + RE}_{sampling\ and\ measurement\ uncertainties} \end{aligned} \qquad \text{Eq. (1)}$$

which in standard statistical matrix notation is formulated as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \qquad \text{Eq. (2)}$$

where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{X}$ is the design matrix for the fixed effects (in small letters), $\boldsymbol{\beta}$ is a vector of parameters for the fixed effects, $\boldsymbol{Z}$ is the design matrix for the errors ($\boldsymbol{\gamma}$) introduced by the various random effects (in capital letters) and $\boldsymbol{\varepsilon}$ is the residual variation (e.g. Searle et al. (1992) for more information on the matrix formulation). The variance of $\boldsymbol{y}$ is:

$$V[\boldsymbol{y}] = \boldsymbol{Z}V[\boldsymbol{\gamma}]\boldsymbol{Z}' + V[\boldsymbol{\varepsilon}] \qquad \text{Eq. (3)}$$

where $V[\boldsymbol{\gamma}]$ is a diagonal matrix with variances for all the random effect errors and $V[\boldsymbol{\varepsilon}]$ is a diagonal matrix with the residual variance in the diagonal. For example, if the random effects included are GR and YR with 4 and 3 levels, then $V[\boldsymbol{\gamma}]$ is a diagonal matrix (dimension 7×7) with the variance for GR in the first four diagonal elements and the variance for YR in the last three diagonal elements.

However, it is often difficult to quantify all these sources in Eq. (1) separately, since this would require an unrealistically large monitoring program with combinations of the different factors at different levels. In practice, it is only possible to estimate a few of these factors (or combinations of factors) from monitoring data, and those identifiable factors will be specific for each monitoring dataset (exemplified in Section 4). Another important issue is that several of the factors contribute relatively little variation to the observations and therefore cannot be quantified. Therefore, it is only relevant to partition the variation in $\boldsymbol{y}$ into those factors that significantly influence variations in the given monitoring data.

3.1 Obtaining parameters for indicator estimation

A statistical model including both fixed and random factors (Eq. 1) is termed a mixed model and is based on the assumption of normality. In many cases, observations can be normalized by choosing an appropriate transformation, although the distribution of y is not restricted to the normal distribution but can be selected within the exponential family (e.g. Poisson, binomial, multinomial, gamma, negative binomial) within the framework of generalized linear mixed models (GLMM). Here, we will consider the normal case only. An important assumption of the model is that all parameters (both fixed effect and variance parameters) are constant over the domain of observations. This assumption can be justified through careful selection of the spatial extent of data combined with appropriate transformation (see Section 5).

Different estimation methods exist (see discussion in Bolker et al. 2009) for quantifying the variation of fixed effects (coefficients/parameters for all levels) and random effects (variance parameters for their distributions). These parameters are more accurately estimated if a large monitoring dataset is used, since the number of parameters involved can be high (see examples in Section 4). This means that the variance parameters can be calculated from a large dataset covering several waterbodies and assessment periods, and these parameters can subsequently be used when estimating the uncertainty of an indicator for a subset or new dataset (denoted $\boldsymbol{y}^*$), i.e. data covering a specific waterbody during a given assessment period.

Given the estimated variance parameters, the covariance matrix for $\boldsymbol{y}^*$ (denoted $\boldsymbol{V}^*$**;** asterisks are used to describe vectors/matrices derived from $\boldsymbol{y}^*$) is determined from the structure of the dataset using the design matrix ($\boldsymbol{Z}^*$) for the random effects (cf. Eq. 2). Subsequently, parameters for the fixed effects can be computed by maximizing the likelihood or equivalently minimizing the generalized least squares, i.e. $(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)'\boldsymbol{V}^{*^{-1}}(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)$. Following the same argument as above, it can be convenient to also fix some of the fixed effect parameters in $\boldsymbol{\beta}^* = [\boldsymbol{\beta}^*_{est}; \boldsymbol{\beta}_{fix}]$ to the estimates obtained from the large data set and only estimate parameters that are considered specific to the waterbody and assessment period. For example, if the large-scale gradient *gr* is explained by depth with a well-established relationship from the entire data set, this parameter can be employed instead of using a more uncertain parameter estimated from a relatively small dataset. Similarly, parameters for *se*, *sd* and *ai* in Eq. (1) can be fixed to the estimates obtained from the entire dataset. Typically, only parameters for μ and *yr* will be estimated specifically with the dataset restricted to the given waterbody and assessment period.

Special attention should be given to the variance contribution of the random temporal factors when calculating $\boldsymbol{V}^*$. Within an assessment period the interannual variation (*YR*) is a finite population with levels corresponding to the number of years. If all years within the assessment period have been monitored then all levels of *YR* can be estimated and the variance contribution of *YR* is essentially zero, i.e. the entire finite population is estimated and *V[YR]=0*. However, if the number of years with monitoring data ($n_{YR,data}$) is less than the number of years in the assessment period ($N_{YR,period}$) then a finite population correction factor $fpc = \left(1 - \frac{n_{YR,data}}{N_{YR,period}}\right)$ (Cochran 1977) should be multiplied to *V[YR]* (see also Clarke and Hering 2006). A similar finite population correction factor $fpc = \left(1 - \frac{n_{SE\times YR,data}}{N_{SE\times YR,period}}\right)$ should be employed for the random seasonal variation among years (*SE*×*YR*), where the number of observed combinations of years and seasons is $n_{SE\times YR,data}$and $N_{SE\times YR,period}$ is the total number of combinations of years and seasons within the assessment period. Finally, the correction factor also applies to *DI*×*SE*×*YR* that has a finite population as well. Thus, for calculating $\boldsymbol{V}^*$ the correction factors should be multiplied to those diagonal elements in $V[\boldsymbol{\gamma}^*]$ associated with finite populations:

$$\boldsymbol{V}^* = V[\boldsymbol{y}^*] = \boldsymbol{Z}^* V[\boldsymbol{\gamma}^*]\boldsymbol{Z}^{*\prime} + V[\boldsymbol{\varepsilon}^*] \qquad \text{Eq. (4)}$$

3.2 Standardizing indicators from parameter estimates

Indicator values representing different waterbodies and assessment periods are calculated from the fixed effect parameter estimates obtained with the approach in Section 3.1, however, indicator calculations must be carried out in a standardized way to allow comparison between waterbodies and assessment periods. For instance, different sampling depths, months and devices may have been used in different waterbodies such that these cannot be compared unless differences between sampling depths, months and devices are resolved.

Fortunately, such differences are described by the parameter estimates obtained from the large dataset and can be accounted for by subtracting the fixed effects (i.e. deviations from the model expectations) or calculating the expected value of the indicator for a given combination of the fixed effects. In both cases, the standardized indicator can be expressed as a linear combination ($\boldsymbol{L}$) of the parameter estimates; i.e. $\boldsymbol{L\beta}^*$. As an example, consider an indicator based on observations with a fixed spatial gradient described by depth (linear gr relationship with slope parameter b_{gr}), fixed seasonal variation described by 4 seasonal coefficients ($se_i; i = 1-4$), and fixed variation between three different sampling devices ($sd_i; i = 1-3$). Parameters estimated from the large dataset (9 in total) are used to describe these variations ($\boldsymbol{\beta_{fix}} = [\mu, b_{gr}, se_1, \dots, se_4, sd_1, sd_2, sd_3]$), and deviations from model predictions using these parameters are modelled with a mean (μ^*) estimated from $\boldsymbol{y}^*$. In this case, $\boldsymbol{\beta}^* = [\boldsymbol{\beta}^*_{est}; \boldsymbol{\beta_{fix}}] = [\mu^*; \mu, b_{gr}, se_1, \dots, se_4, sd_1, sd_2, sd_3]$ is a vector of 10 parameter estimates. Deviations from model predictions (using the parameters from large dataset; $\boldsymbol{\beta_{fix}}$) have a mean of μ^*, which can be expressed by means of $\boldsymbol{L} = [1, 0, \dots., 0]$. However, it is also possible to use other combinations of $\boldsymbol{\beta}^*$ to express the expected indicator values that are more relevant for interpretation, e.g. the indicator value at a standard depth of 3 m, averaged over all 4 seasons and using the most recent sampling devices (given by sd_1). The value of the indicator for this standard set of fixed effects can be calculated using $\boldsymbol{L} = [1, 1, 3, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 1, 0, 0]$.

The variance of the standardized indicator, $V^*_{\boldsymbol{L\beta}}$, given by the linear combination $\boldsymbol{L\beta}^*$**,** is found as:

$$\boldsymbol{V}^*_{\boldsymbol{L\beta}} = \boldsymbol{LV}[\boldsymbol{\beta}^*]\boldsymbol{L}' \qquad \text{Eq. (5)}$$

where the covariance matrix for the parameter estimates is

$$V[\boldsymbol{\beta}^*] = \begin{bmatrix} V[\boldsymbol{\beta}^*_{est}] & 0 \\ 0 & V[\boldsymbol{\beta}_{fix}] \end{bmatrix} \qquad \text{Eq. (6)}$$

Similarly, a confidence interval for the indicator value can be calculated assuming that $\boldsymbol{L\beta}^*$ is approximately normal distributed, i.e. $[\boldsymbol{L\beta}^* - z_{\alpha/2} \cdot \sqrt{V^*_{\boldsymbol{L\beta}}}; \boldsymbol{L\beta}^* + z_{1-\alpha/2} \cdot \sqrt{V^*_{\boldsymbol{L\beta}}}]$. The assumption of normality is justified if the variances of the random effects have been estimated on a large dataset with >30 degrees of freedom used for their computation. Similarly, probabilities for the ecological indicator meeting target values (e.g. MSFD good environmental status) and probabilities for different intervals of the indicator (e.g. WFD ecological status classes) can be calculated.

3.3 Partitioning the variance contributions for an indicator

An important question arises when calculating an indicator: How much do the different random effects contribute to the overall indicator uncertainty? This question can be addressed by sequentially setting the different variance components to zero and calculate the reduction in the indicator variance, but such calculations can also be quite cumbersome, when assessing different combinations of sampling designs for a given waterbody.

However, to get a proxy measure of the importance of the various uncertainty components consider the simple case of averaging N observations from a waterbody during an assessment period. For each of the random factors affecting the observations the variance contribution to the average can be calculated. If a random factor X has been measured at k levels with n_i ($i=1,...,k$) observations for each level ($n_1+n_2+...+n_k=N$) then the variance contribution of the random factor to the average is

$$V\left[\frac{1}{N}\sum_{1}^{N} X_i\right] = V\left[\frac{1}{N}\left(\sum_{1}^{n_1} X_1 + \cdots + \sum_{1}^{n_k} X_k\right)\right] = V\left[\frac{n_1}{N}X_1 + \cdots + \frac{n_k}{N}X_k\right] = \frac{\sum_{i=1}^{k} n_k^2}{N^2}V[X] \qquad \text{Eq. (7)}$$

If observations are balanced between levels ($n_k=N/k$) the more common form of random factor variance partitioning is obtained, i.e. $V\left[\frac{1}{N}\sum_1^N X_i\right] = \frac{1}{k}V[X]$. The total variance of the average is calculated by summing the variance contributions of all relevant random effects (Table 1), providing a measure of the relative importance of the different sources of random variation.

4 CASE STUDIES

In Section 3 the uncertainty framework was presented in general terms that we submit can be applied to most monitoring data used to calculate ecological indicators. It is unlikely that all sources of variation can be determined with a given dataset due to spatial and temporal resolution of the data. In practice, a subset of the variations (Table 1) can be identified and used to calculate the indicator value and uncertainty. In this section we will exemplify the framework and show how the framework can be modified to three different types of indicators. The notation for the sources of variation (Table 1) is employed for all examples and additional variables used in the analyses are introduced as needed.

4.1 Eelgrass shoot density

Eelgrass has been monitored from 1995 to 2011 at a number of stations, covering eight WFD waterbodies, along the Swedish coast of the Sound (Fig. 1). Sampling was carried out using 6 or 12 frames with an area of 0.0625 m^2, which were randomly placed on top of representative meadows near the monitoring station. For each frame, the eelgrass shoot density (esd) was measured. For each station the depth (z; ranging from 1.4 to 5.6 m) and the diver placing the frames (AN) were recorded. Monitoring was conducted between July and October every year and with a few exceptions stations were sampled only once per year. The 17 years of data were divided into three 6-year assessment periods (ap) in order to obtain an estimate of the variation between years within assessment periods as opposed to a variation between years over longer periods, where time series may potentially exhibit trends.

Eelgrass samples were unevenly distributed over waterbodies and years (Table S1). Different divers carried out the sampling but the use of divers was non-systematic relative to monitored waterbodies and years. Hence, it was possible to determine variations caused by different divers. Eelgrass shoot density was approximated with a lognormal distribution as esd varied over a large span with a right-skewed distribution.

4.1.1 Estimating variance components from waterbodies separately

In order to illustrate the problem of identifying several variance components from small datasets, we partitioned variations in $\log(esd)$ for each waterbody separately replacing μ in Eq. (1) with means specific to each assessment period (ap). Moreover, large-scale spatial patterns in $\log(esd)$ were assumed linearly related to depth ($gr(z)$), i.e. shoot density was expected to decrease exponentially with depth.

$$\log(esd) = ap + gr(z) + GR + YR(ap) + se + GR \times YR(ap) + AN + PA(GR) \qquad \text{Eq. (8)}$$

Only three waterbodies had sufficient data to allow estimation of all variance components in Eq. (8) (Table 2). Except for the random spatial variation between stations (GR), variance estimates were comparable across waterbodies; particularly those waterbodies with more data. The variation between frames ($PA(GR)$) was determined as the residual variation from the analysis. The random spatial variation between stations was not well determined, since there were generally few stations within waterbodies (Table S1). Therefore, more precise estimates of the variance components can be obtained by pooling observations, assuming that variance components have same magnitude across waterbodies. Moreover, slopes for the depth relationship varied from -0.14 to -0.37 and could not be estimated for two waterbodies that did not have any variation in depth. Since a generic depth relationship would be expected, a more precise estimate of this parameter could also be obtained by pooling observations.

4.1.2 Estimating variance components from the entire dataset

Using the entire dataset, the model (Eq. 8) was expanded to incorporate variations between waterbodies (wb):

$$\log(esd) = wb + ap + wb \times ap + gr(z) + GR(wb) + YR(wb \times ap) + se + GR \times YR(wb \times ap) + AN + PA(GR(wb)) \qquad \text{Eq. (9)}$$

The first three fixed effects in Eq. (9) were used to describe the variations between combinations of waterbodies and assessment periods, i.e. each combination had its own mean value. The two other fixed effects, depth relationship (*gr(z)*) and seasonal variation (*se*), were assumed common to all waterbodies. The random variations in Eq. (9) had same magnitude for all combinations of waterbody and assessment period. *GR(wb)* described the variation between stations within waterbodies, *YR(wb×ap)* described the variation between years within an assessment period in a waterbody, *GR×YR(wb×ap)* described interannual changes in the variation between stations within an assessment period in a waterbody, *PA(GR(wb))* described the variation between frame replicates at a given sampling station and time, and *AN* described the variation between divers. As above, *PA(GR(wb))* was estimated as the residual variation.

Eelgrass shoot density decreased significantly with depth (common slope= -0.31) and there were significant variations between waterbodies, whereas log(*esd*) did not change significantly between the four months with monitoring data or between the three assessment periods, neither in general (*ap*) nor specific to waterbodies (*wb×ap*) (Table 3). Random variations were largest between divers and frame replicates, whereas interannual variations within assessment periods, variation between stations within waterbodies and interannual changes in this spatial pattern among years were smaller (Table 3). However, the variance estimate for the variation between divers was relatively uncertain, since the monitoring program was not specifically designed to capture this variation, e.g. by rotating divers between years and stations.

The log-transformation of *esd* implies that these variances for random variations can be interpreted as relative errors (using the transformation $\exp\left(\sqrt{V[\quad]}\right) - 1$, where $V[\quad]$ is the variance of the random factor). Using this formula we found that eelgrass shoot density typically varied by ±36% among frame replicates placed by the same diver, whereas different divers would increase the random variation additionally by ±%29. Consequently, variation between frame replicates at the same station but placed by different divers would be ±%49. The spatial variation between stations within a waterbody was smaller (±12%) than the patchiness between frame replicates, and changes in this spatial pattern over years had a similar magnitude (±18%). The overall variation between years within assessment periods was ±13%. The total variation accumulated to ±59% between two randomly selected frame samples from the entire population (eelgrass shoot density across all meadows in a waterbody over six years).

4.1.3 Calculating eelgrass shoot density indicator value and uncertainty

The calculation of indicator value ($\boldsymbol{L\beta}^*$) and its variance ($V_{\boldsymbol{L\beta}}^*$) can be illustrated using data from Lomma Bay. During the three assessment periods 288, 192 and 96 frames were sampled, distributed over 4, 10 and 4 stations, and sampled by 4, 2 and 1 divers, respectively (Table 4). Using the variance estimates (Table 3) for calculating the covariance matrix $\boldsymbol{V}$ as well as fixing parameters for the depth relationship (slope) and seasonal variation (four months' estimates) to those obtained from analyzing the large dataset, the mean $\log(esd)$ was estimated as 6.75, 6.48, and 6.54 for the three assessment periods (for a standard depth of 3 m and averaged over July-October, i.e. $L = [1, 1, 3, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$), respectively, corresponding to geometric means of 852, 653, and 694 shoots per m^2.

The standard errors of the of the mean $\log(esd)$ for the three assessment periods were 0.15, 0.21, and 0.27, corresponding to relative uncertainties for the geometric means of 17%, 23%, and 32%. For the eelgrass shoot density the 95% confidence interval for the geometric means were [636;1143], [442;965], and [408;1180]. If the boundary for good environmental status for this indicator was set at 500 shoots per m^2, then confidences of complying with this boundary were 99.98%, 90.98%, and 88.66% for the three assessment periods, respectively. Thus, the confidence of compliance was lower for 2007-2012 than for 2001-2006 even though the geometric mean was higher in the more recent assessment period. The reason is that the mean $\log(esd)$ was more uncertain for the 2007-2012 period, because it was based on fewer observations, and consequently, the distribution of the indicator had heavier tails.

However, it is not simply the number of observations that determines the uncertainty of the indicator; it is the actual distribution of samples across stations, years and divers. Using Eq. (7) proxy measures for the variance contribution of the different random effects to the average of $\log(esd)$ observations were calculated (Table 4). For all three assessment periods the variation between divers was the largest source of uncertainty, contributing between 78% and 93% of the total variance. Variation among stations was the second largest source of uncertainty. Moreover, the variance contribution from interannual variation was zero for the first two assessment periods, because all six years were sampled.

The variance contribution from divers and stations was not strongly reduced despite the many samples due to limited replication of these two factors. In fact, all samples during 2007-2012 were carried out by a single diver, implying that the variance of the mean $\log(esd)$ could not be less than 0.066 (Table 4), because the same diver-specific error affected all samples during this assessment period. Since this source of variation was quite dominant, it will be important to reduce this uncertainty in future monitoring. This can be done by 1) developing a more precise guidance for placing frames to reduce variation between divers and 2) using (rotating) more divers in the monitoring.

4.2 Chlorophyll a

The uncertainty components associated with assessing the surface (top 10 m) chlorophyll a ($chla$) concentration were examined using observations from the Danish and Swedish national monitoring programs covering the WFD type NEA8b (Northeast Atlantic 8b). The $chla$ observations were distributed over 89 monitoring stations representing 39 WFD waterbodies (wb) (Fig. 2) and spanned a 30-year period (1985-2014), which was divided into five assessment periods (ap) of six years each. As for most monitoring programs the distribution of samples among waterbodies was heterogeneous (Table S2), and many waterbodies were represented by a single station implying that sources of spatial variation could not be determined for such waterbodies. Hence, the random factors were determined from the combined dataset using the log-transform of $chla$ and assuming that random variations had the same magnitude across waterbodies on a relative scale.

4.2.1 Estimating variance components from the entire dataset

It was assumed that the uncertainty associated with sampling devices and measurement of the sample was small, and therefore uncertainties associated with sampling properties in time and space were considered only.

$$\begin{aligned}&\log(chla) = wb + ap + wb \times ap + GR(wb) + YR(wb \times ap) + se(wb) \qquad \text{Eq. (10)}\\ &+SE \times YR(wb \times ap) + GR \times YR(wb \times ap) + SE \times GR(wb \times ap) + IR\end{aligned}$$

The first three fixed effects in Eq. (10) were used to describe the variations between combinations of waterbodies and assessment periods, i.e. each combination had its own mean value (see Section 4.2.2). The remaining fixed effect (*se(wb)*) described the seasonal variation specific to each waterbody using a monthly resolution. $GR(wb)$ described the random variation among stations within waterbodies, $YR(wb \times ap)$ described the random variation between years within an assessment period in a waterbody, $SE \times YR(wb \times ap)$ described interannual changes in the seasonal pattern within the same waterbody and assessment period, $GR \times YR(wb \times ap)$ described the variation between stations among years within the same waterbody and assessment period, $SE \times GR(wb \times ap)$ described changes in the seasonal pattern among stations within the same waterbody and assessment period, and the residual variation described irregular temporal variation (IR) between $chla$ samples taken within the same month at the same station. The small-scale spatial variation (patchiness) could not be assessed from the data, since stations were not spatially subsampled, and it was therefore implicitly included as part of the large-scale spatial variation ($GR(wb)$).

Irregular variation was the largest source of uncertainty (Table 5), suggesting that variation between samples taken at the same station and month could vary by ±81% (using the transformation $\exp\left(\sqrt{V[\quad]}\right) - 1$). The second largest source of uncertainty was the interannual changes in the seasonal pattern within an assessment period, deviating from the common seasonal variation of the waterbody by ±49%. Random spatial variation among stations within waterbodies was ±30%, and the spatial variation changed between years by ±19% and between seasons by ±16%. Interannual variation was the smallest random variation changing ±12% between years. The total variation between two randomly selected $chla$ surface samples within the entire waterbodies and assessment period was ±124%, i.e. more than a factor two variation.

4.2.2 Calculating chlorophyll a indicator value and uncertainty

Three different waterbodies, each having five assessment periods, were used to show the calculation of the chlorophyll a indicator value ($\boldsymbol{L\beta}^*$) and variance ($V^*_{\boldsymbol{L\beta}}$). The mean $chla$ was calculated for each combination of waterbody and assessment period using fixed effect parameters for the mean and seasonal variation obtained from the large dataset, i.e. $\boldsymbol{\beta}^* = \left[\boldsymbol{\beta}^*_{\boldsymbol{est}}; \boldsymbol{\beta}_{\boldsymbol{fix}}\right] = [\mu^*; \mu, se_1, \dots, se_{12}]$. The corresponding linear combination for calculating the indicator was $\boldsymbol{L} = [1, 1, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}]$.

Temporal trends at each waterbody were described by five mean values representing assessment periods, and a fixed seasonal variation common to all years were estimated for each waterbody (Fig. 3). The seasonal variation changed from a bimodal pattern in the open waters to unimodal pattern in shallow estuaries more strongly influenced by land-based nutrient inputs. For the three sites, the number of samples available for calculating the chlorophyll a indicator varied from 11 to 765 for combinations of waterbody and assessment period. Variances of the indicator values were computed using the variance estimates of the random factors (Table 5) and the finite population correction factor fpc for $V[YR(wb \times ap)]$ and $V[SE \times YR(wb \times ap)]$. For most assessment periods at the three sites fpc=0, because all years and months had been monitored. Nevertheless, fpc>0 for the first assessment period in Central Sound coastal water and Hevring Bay as well as for the last assessment period in Hevring Bay, where only two years and 12 months of data were available (in the latter case fpc=0.67 and 0.83 for the two random factors with finite population correction factor). Overall, the confidence intervals of the mean $chla$ were relatively wider when the mean was based on fewer observations, whereas many observations resulted in higher precision of the mean estimate (Fig 3).

If the boundary for good ecological status was set at 1.5 μg L^{-1} for Central Sound coastal water, the geometric $chla$ mean would comply with good ecological status in four out of five of the assessment periods with probabilities of 84%, 87%, 51%, 57% and 83%, respectively. Comparing the first two assessment periods it is observed that the geometric $chla$ mean is lowest in the first assessment period, but interestingly the probability of achieving good ecological status is lower as well, because the assessment of the first period is based on fewer observations and hence more uncertain, i.e. the distribution of the indicator value is wider.

Irregular temporal variation (IR) was the largest source of uncertainty for individual observations (Table 5), but the contribution of this random variation to the chlorophyll a mean over an assessment period was strongly reduced due to many observations (replications for this factor, cf. Eq. 7) in each assessment period (Fig. 4). In most cases, this random variation contributed less than 10% to the total indicator variance, except for periods with few observations such as the first assessment period in the Central Sound coastal water (n=11) and last assessment period in Hevring Bay (n=21). Spatial variation between stations (GR) dominated the variance contribution to the mean $chla$, because this variation was relatively large (Table 5) and most assessment periods were monitored using a single station. For the three waterbodies used to exemplify the calculations, only Skive Fjord was monitored with more than one station. In fact, three stations were sampled at least biweekly in Skive Fjord until 2009, when the monitoring program was reduced to a single station. The consequence of this reduction in monitoring efforts was quite apparent in the $chla$ indicator uncertainty that almost doubled from 2003-2008 to 2009-2014 (Fig. 3). The $chla$ results clearly show that the spatial variations are under-sampled, given that an estimate representative of the entire waterbody should be obtained. Increasing the number of monitoring stations within waterbodies will have a large effect on the $chla$ indicator uncertainty whereas increasing the monitoring frequency will only marginally reduce the indicator uncertainty.

5 DISCUSSION

Monitoring data, which are used to assess the ecological status of a waterbody within the WFD-context, are influenced by many different sources of uncertainty. In order to evaluate and quantify the confidence any status classification and to minimize uncertainty, by modifying monitoring designs or accounting for predictable factors, it is necessary that the contribution of different sources of variability to the overall uncertainty can be understood and accounted for (Clarke and Hering 2006, Carstensen 2007, Clarke 2013, Dromph et al. 2013). Here we present a general framework that can be applied to a wide range of typical aquatic monitoring data, and we have illustrated the framework with two examples, using both benthic and pelagic monitoring data. In addition, the framework has been applied to a broad range of Swedish monitoring data, both freshwater and marine water, confirming the general applicability (www.waters.gu.se). The framework outlines 18 different sources of uncertainty that could be considered relevant for various types of monitoring data. Nevertheless, in practice only a subset of these uncertainty sources can be estimated from regular monitoring data that have not been specifically sampled to quantify specific sources. Some of the identifiable uncertainty components may combine several sources of uncertainty (e.g. for chlorophyll a sampling the variance between stations within waterbodies includes both large- and small-scale spatial variation), whereas some sources of random variation may not be relevant for specific types of monitoring data. Overall, we submit that the general framework can be adapted to analyze the most common types of monitoring data.

5.1 Why is it important to quantify the different uncertainty components?

Essentially, for most applications only the variance of an indicator values is of interest, but it is also important that the variance estimate is based on a correct interpretation of the different sources of uncertainty. An incorrect representation of the different sources of uncertainty may lead to a severe underestimation of the indicator variance. For example, using eelgrass data from Laholm Bay only the random variation between replicated frames could be estimated, due to lack of replication of other factors such as stations, years and divers (Table 2). Nevertheless, all the other unquantified sources of uncertainty affect the observations of eelgrass shoot density even if they cannot be estimated. This implies that the indicator variance will be grossly underestimated if it were calculated from observations in Laholm Bay only (cf. Table 3), as the variance of the indicator would not include random variation between stations, years and divers. However, even if all relevant sources of uncertainty were replicated (e.g. Lomma Bay in Table 2), pooling all these sources together to a single error term will also underestimate the indicator variance. If a single uncertainty component is employed it is implicitly assumed that all observation errors are independent, which never is the case in standard monitoring data. For example in Lomma Bay, the 576 observations were distributed over 20 stations, 17 years and 5 divers (Table S1), implying that the error terms associated with these factors cannot be independent for all 576 observations. If this spatial and temporal dependence would have been ignored and a single error term would have been employed with a variance for $\log(esd)$ of 0.2148 (Table 4) the variances of the average log(*esd*) for the three assessment periods would be 0.00075, 0.0521 and 0.01042 (corresponding to incorrect relative uncertainties for the geometric means of 3%, 7%, and 11%). This is much smaller than the variances accounting for the distribution of error terms between stations, years and divers and 2-5 times smaller than when properly accounting for the dependence according to the framework presented here, i.e. the average log(*esd*) is 0.15, 0.20 and 0.27 for the three periods (relative uncertainties for the geometric means of 16%, 22%, and

31%). Hence, this example clearly demonstrates the importance of quantifying all the major uncertainty components, and employing a correct representation of the associated error terms to calculate indicator uncertainty.

Furthermore, it is also important to have good estimates of the variances of the uncertainty components for calculating the indicator variance. In both examples above, the number of stations within waterbodies was relatively small and the number of divers monitoring eelgrass shoot density was even less (Table S1 and S2). Therefore, variance estimates for the random variation between stations and divers will not be well-determined if these are based on data from a single waterbody only. Consequently, more precise variance estimates can be obtained if the uncertainty components are estimated from a larger dataset, by pooling observations from several waterbodies. The consequence of doing so is that the random variations are assumed to have the same magnitude across all waterbodies included. Naturally, this assumption can be violated if data from many and highly diverse waterbodies are combined.

The waterbodies used in the two examples above were from restricted regions with somewhat similar characteristics. However, the chlorophyll a data spanned from enclosed estuaries to open coastal waterbodies, and it could be argued that variations are larger in those systems more strongly connected to land. On the other hand, chlorophyll a levels in such systems are also expectedly higher. Hence, the magnitude of random variations in chlorophyll a is presumably constant relative to the expected mean concentration, justifying that the analysis was carried out on log-transformed observations. In fact, similar variances for the random factors in Eq. (10) were obtained by splitting the large chlorophyll a dataset into estuarine and coastal stations (Fig. 5), supporting the estimation of variance components from a large dataset covering a broad gradient of waterbodies. Furthermore, appropriate transformation of observations can account for heteroscedasticity among waterbodies as well as within waterbodies.

5.2 How can we reduce indicator uncertainty?

Desired properties of ecological indicators are low bias and high precision. Unbiasedness implies that the calculated indicator is not sensitive to changes in the spatial and temporal distribution of samples, or the sampling and measurement equipment used. Precision implies that the calculated indicator provides an estimate close to the unknown “true indicator value”.

Monitoring data are commonly heterogeneously distributed in time and space for various reasons. Therefore, statistical analyses of such heterogeneous data should include a description of temporal and spatial variations, corresponding to the resolution of the data, to ensure that changes in sampling across assessment periods will not affect the derived indicator value, i.e. estimated indicators are unbiased by changes in the sampling program. Similarly, if different sampling or analytical methods as well as different people (e.g. divers) have been involved in the monitoring, these variations should be described as part of the statistical analyses to ensure comparability of indicator values across assessment periods.

These variations in spatial-temporal sampling as well as sampling and analysis methodology can be accounted for by a combination of fixed and random effects. Describing variations as fixed effects has the advantage of reducing the indicator uncertainty, since the variation is predictable and can be accounted for. However, only a subset of the sources of variation (Table 1) can be described as fixed effects. Importantly, the variance of the random factors decrease the better the model parameterization for the fixed effects is. Spatial variation was considerable in both examples above, but for eelgrass shoot density it was demonstrated that a large portion of the variation between stations could be explained by depth differences. Without the depth model for $\log(esd)$, the variance between stations (V[GR]=0.1138) would be almost 10 times larger, highlighting the importance of incorporating fixed factors to describe variations and reduce uncertainty. Furthermore, because the variability among stations was generally the largest among the spatial and temporal sources, accounting for depth also had a major influence on the overall uncertainty for a whole assessment period. As an example the overall indicator standard error for the different periods would have been 0.21, 0.22 and 0.32 if depth was not accounted for, compared to 0.15, 0.21 and 0.27 in the proposed model, and variation between stations dominated the indicator variance contribution in the first and third assessment period. It is possible that the random spatial variation could be further reduced by including fetch and substrate as explanatory variables, but such information was not available for the present study.

For chlorophyll a, corrections for monthly means were made, but no attempts were made to explain the spatial variation using other fixed, environmental factors. It is, however, likely that including salinity as covariate could have accounted for some spatial and temporal variation and thereby improve the indicator precision as was demonstrated for total nitrogen in Carstensen (2007). Moreover, it is also possible that coupled hydrodynamic-biogeochemical models can account for a larger fraction of the variation in chlorophyll a measurements and consequently reduce the remaining variations, spatial as well as temporal random variations.

5.3 How can uncertainty estimates improve monitoring networks?

The analyses of variance contributions of different random factors to the overall indicator uncertainty clearly identify the most critical sources of uncertainty. For eelgrass shoot density it was obvious that variation between divers has large consequences on the uncertainty of eelgrass indicator. This could be due to lack of clear guidelines on where to place the frames to be representative of the eelgrass meadow. For chlorophyll a spatial variation between stations was the largest contribution to indicator variance. A single station was used to characterize the chlorophyll a level in most waterbodies (Table S2), and even if the random variation between stations was not the largest source of uncertainty (Table 5) the few stations relative to the number of levels of other random factors resulted in variation between stations being the most important contribution to indicator variance (Fig. 4). Hence, the uncertainty of the chlorophyll a indicator can only be substantially reduced, provided that more stations within waterbodies are monitored or that a large part of the spatial variation between stations can be described using salinity as covariate or through coupled hydrodynamic-biogeochemical models.

Monitoring programs for assessing status of a waterbody can be designed to obtain the most precise indicators under a given financial frame. If costs associated with ship time, sampling and analysis are known then the optimal allocation of samples in time and space can be found as a minimization problem (minimum indicator variance) under the financial constraint. In the chlorophyll a example it is possible that shifting sampling efforts to less temporal and more spatial sampling will provide a more precise chlorophyll a indicator without being more costly. When the status assessment is based on multiple indicators the optimal monitoring program can be determined by minimizing a weighted average of the indicators' variances. Thus, quantification of the different uncertainty components affecting monitoring data used for calculating ecological indicators allows for optimal distribution of samples in time and space.

Acknowledgements

The present work is a contribution of the WATERS project funded by the Swedish Environmental Protection Agency and the DEVOTES project (ENV.2012.6.2-3, Grant agreement no. 308392) funded under the European Commission 7th framework program. We are grateful to Per Olsson, Toxicon AB, for providing the dataset on eelgrass shoot density and all the monitoring people in Denmark and Sweden who have collected and analyzed chlorophyll a samples as part of the national monitoring programs.

REFERENCES

Balsby, T.J.S., Carstensen, J., Krause-Jensen, D., 2013. Sources of uncertainty in estimation of eelgrass depth limits. Hydrobiologia 704, 311-323. doi: 10.1007/s10750-012-1374-8

Borja, A., Elliott, M., Carstensen, J., Heiskanen, A.S. van de Bund, W. 2010. Marine management – Towards an integrated implementation of the European Marine Strategy Framework and the Water Framework Directives. Mar. Poll. Bull. 60, 2175–2186

Carstensen, J., 2007. Statistical principles for ecological status classification of Water Framework Directive monitoring data. Mar. Pollut. Bull. 55, 3-15. doi: 10.1016/j.marpolbul.2006.08.016

Carvalho, L., Poikane, S., Solheim, A.L., Phillips, G., Borics, G., Catalan, J., de Hoyos, C., Drakare, S., Dudley, B.J., Järvinen, M., Laplace-Treyture, C., Maileht, K., McDonald, C., Mischke, U., Moe, J., Morabito, G., Nõges, P., Nõges, T., Ott, I., Pasztaleniec, A., Skjelbred, B., Thackeray, S.J., 2013. Strength and uncertainty of phytoplankton metrics for assessing eutrophication impacts in lakes. Hydrobiologia 704, 127-140. doi: 10.1007/s10750-012-1344-1

Chapra, S.C., Dove, A., Warren, G.J., 2012. Long-term trends of Great Lakes major ion chemistry. J. Great Lakes Res. 38, 550-560. doi:10.1016/j.jglr.2012.06.010

Clarke, R.T., 2013. Estimating confidence of European WFD ecological status class and WISER Bioassessment Uncertainty Guidance Software (WISERBUGS). Hydrobiologia 704, 39-56. doi: 10.1007/s10750-012-1245-3

Clarke, R.T., Hering, D., 2006. Errors and uncertainty in bioassessment methods – major results and conclusions from the STAR project and their application using STARBUGS. Hydrobiologia 566, 433-440. doi: 10.1007/s10750-006-0079-2

Cochran, W.G., 1977. Sampling techniques, Wiley, New York.

Davey, A., Garrow, D. 2009. Variability components for macrophytes communities in rivers: summary report. Environment Agency, Bristol, UK. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291734/scho1109brhq-e-e.pdf

Downes, B.J., Lake, P.S., Schreiber, E.S. 1993. Spatial variation in the distribution of stream invertebrates: implications of patchiness for models of community organization. Freshw. Biol. 30(1), 119–132

Dromph, K.M., Agusti, S., Basset, A., Franco, J., Henriksen, P., Icely, J., Lehtinen, S., Moncheva, S., Revilla, M., Roselli, L., Sørensen, K., 2013. Sources of uncertainty in assessment of marine phytoplankton communities. Hydrobiologia 704, 253-264. doi: 10.1007/s10750-012-1353-0

Dudley, B., Dunbar, M., Penning, E., Kolada, A., Hellsten, S., Oggioni, A., Bertrin, V., Ecke, F., Søndergaard, M., 2013. Measurements of uncertainty in macrophyte metrics used to assess European lake water quality. Hydrobiologia 704, 179-191. doi: 10.1007/s10750-012-1338-z

European Commission, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. Off. J. Eur. Union L327, 1–72.

European Commission, 2008. Directive 2008/56/EC of the European Parliament and of the Council establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). Off. J. Eur. Union L164, 19–40.

Greening, H., Janicki, A., Sherwood, E.T., Pribble, R., Johansson, J.O.R., 2014. Ecosystem responses to long--term nutrient management in an urban estuary: Tampa Bay, Florida, USA. Estuar. Coast. Shelf Sci. 151, A1--A16. doi: 10.1016/j.ecss.2014.10.003

Haase, P., Murray-Bligh, J., Lohse, S., Pauls, S., Sundermann, A., Gunn, R.,Clarke, R., 2006. Assessing the impact of errors in sorting and identifying macroinvertebrate samples. Hydrobiologia 566, 505-521. doi: 10.1007/s10750-006-0075-6

Hagy, J.D., Boynton, W.R., Keefe, C.W., Wood, K.V., 2004. Hypoxia in Chesapeake Bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. Estuaries 27, 634-658.

Hering D., Borja A., Carstensen J., Carvalho L., Elliott M., Feld C.K., Heiskanen A.S., Johnson R.K., Moe J., Pont D., Solheim A.L., van de Bund W. (2010) The European Water Framework Directive at the age of 10: a critical review of the achievements with recommendations for the future. Sci. Tot. Env., 408: 4007–4019.

Jakobsen, H., Carstensen, J., Harrisson, P.J., Zingone, A., in press. Estimating time series phytoplankton carbon biomass: Inter-lab comparison of species identification and comparison of volume-to-carbon scaling ratios. Estuar. Coast. Shelf Sci. doi: 10.1016/j.ecss.2015.05.006

Krause-Jensen, D., Middelboe, A.L., Carstensen, J., Dahl, K., 2007. Spatial patterns of macroalgal abundance in relation to eutrophication. Mar. Biol. 152, 25-36. doi: 10.1007/s00227-007-0676-2

Levin, S.A. 1992. The problem of pattern and scale in ecology. Ecology 73, 1943-1967

Morrisey, D.J., Howitt, L., Underwood, A.J., Stark, J.S., 1992. Spatial variation in soft-sediment benthos. Mar. Ecol. Prog. Ser. 81, 197-204

Norén K., Lindegarth M., 2005. Spatial, temporal and interactive variability of infauna in Swedish coastal sediments. J. Exp. Mar. Biol. Ecol. 317, 53-68

Pinel-Alloul, B., Downing, J.A., Perusse, M., Codin-Blumer G., 1988. Spatial Heterogeneity in Freshwater Zooplankton: Variation with Body Size, Depth, and Scale. Ecology. 69(5), 1393-1400

Searle, S. R., Casella, G., McCulloch, C. E., 1992. Variance Components, New York: John Wiley & Sons.

Sharp, J.H., 2010. Estuarine oxygen dynamics: What can we learn about hypoxia from long-time records in the Delaware Estuary? Limnol. Oceanogr. 55, 535-548.

Svensson, J. R., Jonsson, L., Lindegarth, M. 2013. Excessive spatial resolution decreases performance of quantitative models, contrary to expectations from error analyses. Mar. Ecol. Prog. Ser., 485: 57-73.

Thackeray, S.J, Nõges, P., Dunbar, M.J., Dudley, B.J., Skjelbrede, B., Morabito, G., Carvalho, L., Phillips, G., Mischke, U., Catalan, J., de Hoyos, C., Laplace, C., Austoni, M., Padedda, B.M.,Maileht, K., Pasztaleniec, A., Järvinen, M., Solheim, A.L., Clarke, R.T., 2013. Quantifying uncertainties in biologically-based water quality assessment: A pan-European analysis of lake phytoplankton community metrics. Ecol. Ind. 29, 34-47. doi: 10.1016/j.ecolind.2012.12.010

Table 1: Sources of variation in aquatic monitoring data. The table only includes the most relevant variations affecting traditional environmental/ecological monitoring data. Fixed effects are shown in small letters and random effects are shown in capital letters.

Type	Notation	Description
Spatial	gr	Fixed large-scale gradient within a waterbody that can be explained by other spatial co-variables.
	GR	Unexplained random large-scale gradient, i.e. variations unaccounted for by gr.
	pa	Fixed small-scale variation explained using co-variables such as depth and substrate.
	$PA(GR)$	Random variation between samples taken at different locations within a small area.
Temporal	yr	Fixed interannual variation that can be explained by explanatory temporal factors.
	YR	Unexplained random interannual variation, i.e. variations unaccounted for by yr.
	se	Fixed seasonal variation common to all years.
	$SE\times YR$	Random seasonal deviations between years from the common pattern (s).
	di	Fixed diurnal variation common to all days within the assessment period.
	$di\times se$	Fixed season-specific variation common to all days within the given seasons of the assessment period.
	$DI\times SE\times YR$	Random fluctuations in the diurnal pattern unaccounted for by d and $d\times s$.
	IR	Random irregular fluctuations between samples taken within a time interval shorter than other temporal variations used to describe observations.
Spatial-temporal	$GR\times YR$	Random large-scale variations in the spatial gradient between years.
	$GR\times SE$	Random large-scale variations in the spatial gradient between seasons.
Methodological	sd	Fixed variation between sampling devices used.
	ai	Fixed variation between analytical instruments used.
	AN	Random variation between analysts performing the measurement.
	RE	Variations between replicated measurements of the same sample.

Table 2: Variance components estimated from Eq. (8). All variance component could not be estimated for each site, as indicated with a '-', due to lack of replication for these components.

Waterbody	*V[GR]*	*V[YR(ap)]*	*V[GR×Y(ap)]*	*V[AN]*	*V[PA(GR)]*
Laholm Bay coastal water	-	-	-	-	0.0512
Northern Sound coastal water	0.7587	0.0120	0.0055	0.0076	0.1101
Central Sound coastal water	0.1418	0.0291	-	0.0577	0.0660
Lomma Bay	0.0020	0.0314	0.0205	0.0787	0.1014
Southern Sound coastal water	0.0121	0.0103	0.0451	0.0110	0.0850
Southern Sound offshore water	-	0.0541	0.0125	-	0.0376
Höllviken embayment	0.2340	-	-	-	0.0367
W South coast coastal water	-	0.0794	0.0042	-	0.0805

Table 3: Significance of fixed and random effects for eelgrass shoot density in Eq. (9). F-tests for fixed effects were calculated by imposing constraints on the fixed effect parameters. Wald's Z-tests for random effects were calculated using asymptotic standard errors for the variance estimates.

Fixed effects	Num DF	Den DF	*F*	*P (F>0)*
wb	7	30	17.73	<0.0001
ap	2	56	2.71	0.0755
wp×ap	9	56	0.91	0.5213
gr(z)	1	1736	2349.21	<0.0001
se	3	1736	1.51	0.2098
Random effects	Var. estimate	SE of estimate	*Z*	*P (Z>0)*
GR(wb)	0.01241	0.00758	1.64	0.0508
YR(wb×ap)	0.01529	0.00605	2.53	0.0057
GR×YR(wb×ap)	0.02761	0.00671	4.11	<0.0001
AN	0.06551	0.04045	1.62	0.0527
PA(GR(wb))	0.09394	0.00319	29.41	<0.0001

Table 4: The variance contribution of the different random effects to the average $\log(esd)$ in Lomma Bay for the three assessment periods calculated using Eq. (7). The variance estimate in the second column is from Table 3 and used for comparison.

Random effect	Variance	1996-2000			2001-2006			2007-2012		
	estimate	k=	Var. contribution		k =	Var. contribution		k =	Var. contribution	
$GR(wb)$	0.01241	4	0.00323	14%	10	0.00145	3%	4	0.00543	7%
$YR(wb\times ap)$	0.01529	6	0.00000	0%	6	0.00000	0%	5	0.00062	1%
$GR\times YR(wb\times ap)$	0.02761	19	0.00158	7%	25	0.00124	3%	11	0.00280	4%
AN	0.06551	4	0.01820	78%	2	0.04094	93%	1	0.06551	87%
$PA(GR(wb))$	0.09394	288	0.00033	1%	192	0.00049	1%	96	0.00098	1%
Total	0.21477		0.02334	100%		0.04413	100%		0.07534	100%

Table 5: Estimates and significance of random effects for chlorophyll a in Eq. (10). Wald's Z-tests for random effects were calculated using asymptotic standard errors for the variance estimates.

Random effects	Var. estimate	SE of estimate	*Z*	*P (Z>0)*
GR(wb)	0.06716	0.01946	3.45	0.0003
YR(wb×ap)	0.01386	0.00369	3.76	<0.0001
SE×YR(wb×ap)	0.16154	0.00523	30.88	<0.0001
GR×YR(wb×ap)	0.02999	0.00354	8.47	<0.0001
SE×GR(wb×ap)	0.02238	0.00289	7.73	<0.0001
IR	0.35203	0.00386	91.09	<0.0001

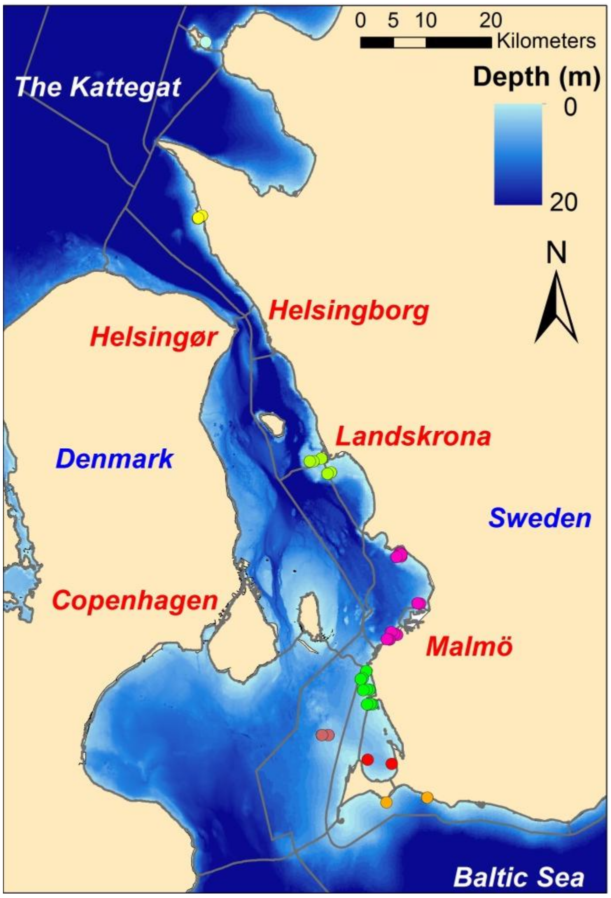

Fig. 1: Monitoring locations for eelgrass shoot density along the Swedish part of the Sound. WFD waterbodies are delineated with gray lines and eelgrass monitoring locations within a waterbody have same color. The monitoring locations around Landskrona represent a combination of three waterbodies in the central part of the Sound.

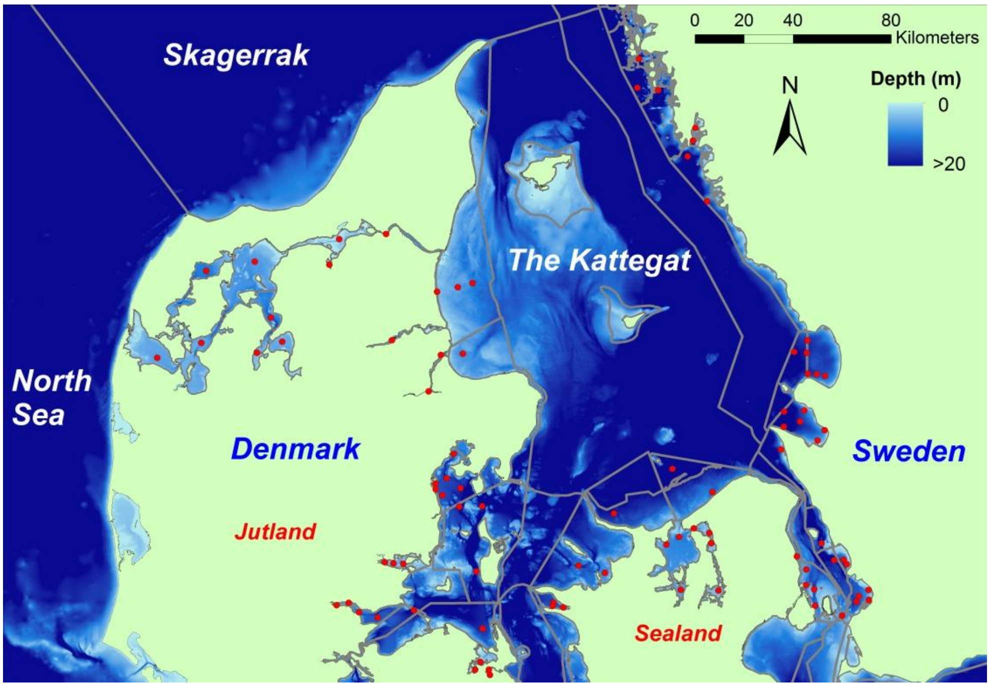

Fig. 2: Danish and Swedish water quality monitoring stations located within the WFD type NEA8b. WFD waterbodies are delineated with gray lines. The number of stations within each waterbody can be found in Table S2.

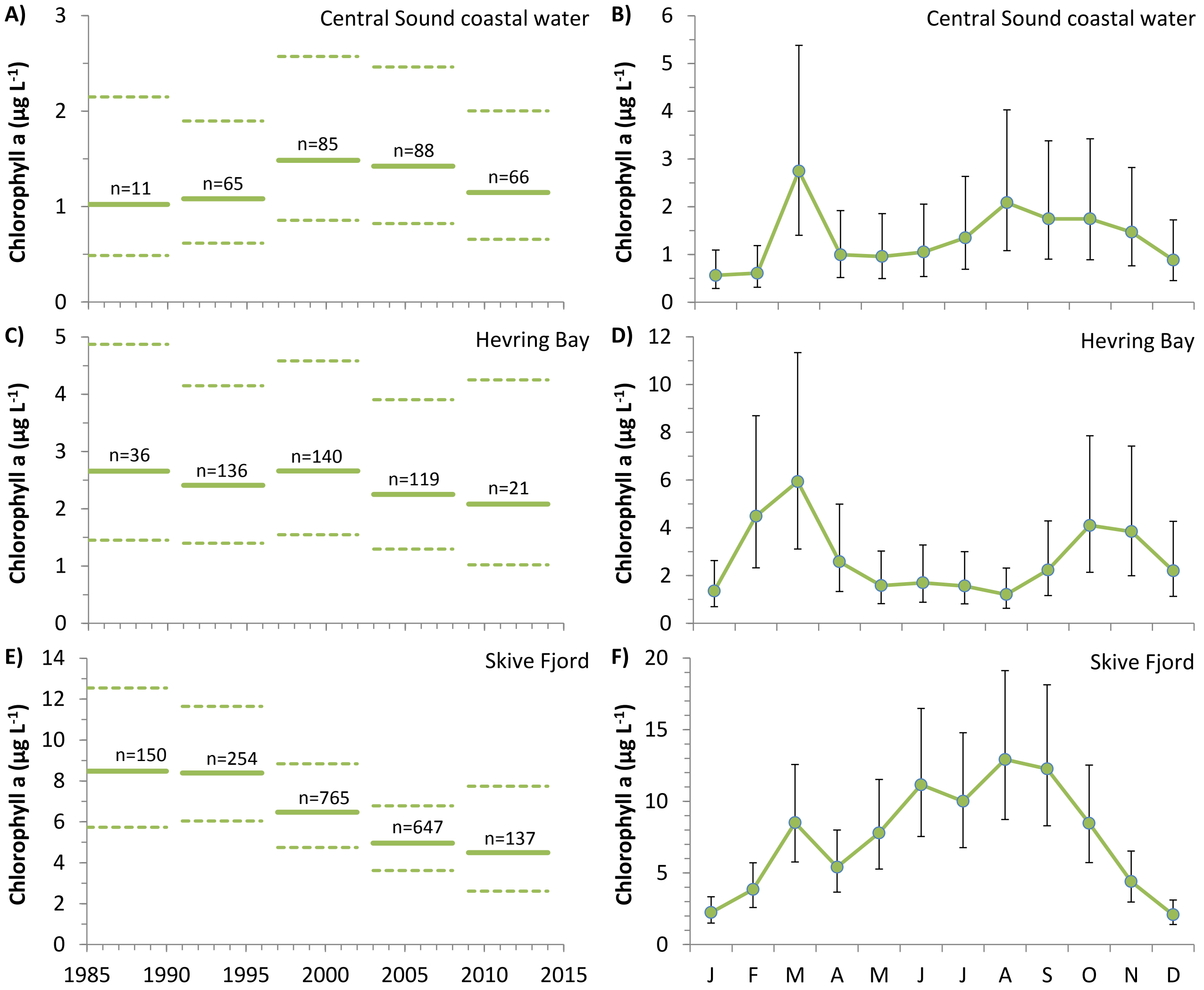

Fig. 3: Estimated mean chlorophyll a levels (Jan-Dec) for 6-year assessment periods (left panel) and the estimated fixed seasonal variation (right panel) for three different coastal waterbodies, ranging from open coastal, embayment and estuary. The dashed lines (left) and the error bars (right) show the 95% confidence interval for the means of the assessment period and the fixed seasonal pattern, respectively. The number of observations used to compute the means of the assessment period is listed above the mean.

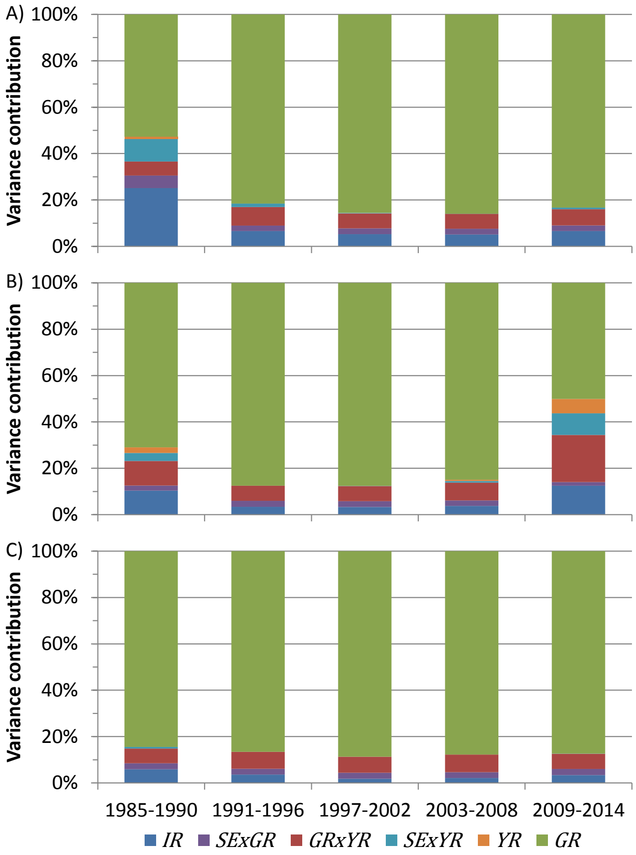

Fig. 4: The relative variance contributions (using Eq. 7) to the chlorophyll a indicator mean for the five assessment periods in A) Central Sound coastal water, B) Hevring Bay, and C) Skive Fjord.

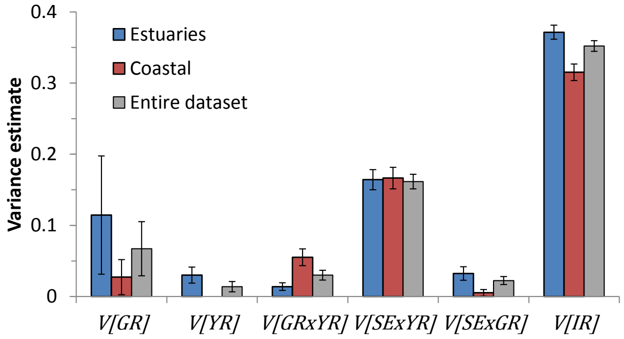

Fig. 5: Variance estimates for chlorophyll a random factors in the entire dataset and split between estuaries and coastal waterbodies. Error bars are the 95% confidence limits for the variance estimates.

SUPPLEMENTARY INFORMATION

Table S1: Distribution of eelgrass shoot samples (n=1913) between stations, years and divers for the nine waterbodies. The total number of divers was 6, and one diver conducted almost half of the samples. Waterbodies are listed from north to south (see Fig. 1).

Waterbody	# stations	# years	# divers	# observations
Laholm Bay coastal water	1	2	2	12
Northern Sound coastal water	5	12	2	138
Central Sound coastal water	6	17	4	288
Lomma Bay	20	17	5	576
Southern Sound coastal water	10	17	4	654
Southern Sound offshore water	2	11	2	167
Höllviken embayment	2	1	1	12
W South coast coastal water	4	10	3	66

Table S2: Distribution of chlorophyll a samples (n=25633) between years and stations for the 39 waterbodies belonging to the WFD type NEA 8b. Data include both Danish and Swedish monitoring stations (Fig. 2).

Waterbody	Period	# years	# stations	# observations
Århus Bugt inner	1985-2014	30	4	1733
Århus Bugt outer	1986-1997	10	2	165
Central Sound coastal water	1987-2014	28	1	315
Dana Fjord	1986-2014	29	1	355
Endelave coastal water	1985-2014	30	2	1349
Gothenburg archipelago	1989-1992	4	1	13
Hesselø coastal water	1988-2014	27	1	152
Hevring Bay	1985-2010	22	1	452
Horsens Fjord inner	1985-2014	29	3	1471
Isefjord	1989-2014	26	3	1225
Kalundborg Fjord	1989-2010	21	2	676
Kattegat	2002-2014	13	1	241
Kertinge Nor	1987-2013	20	2	873
Kungsbackafjorden inner	1993-2014	22	1	233
Kungsbackafjorden outer	1993-2014	22	1	243
Laholms Bay	1993-2014	22	5	483
Limfjorden	1985-2014	30	7	3167
Lomma Bay	1985-2014	22	4	344
Lundåkra Bay	1985-2012	23	2	274
Mariager Fjord inner	2002-2014	13	1	294
NE Sealand coastal water	1989-2006	16	1	124
Northern Belt Sea	1985-2014	30	1	849
Northern Hallands coastal water	1993-2014	22	2	497
Northern Roskilde Fjord	1989-2009	21	3	1113
Northern Sound coastal water	1997-2012	16	1	209
Northern Sound Denmark	1988-2004	17	4	707
NW Sealand coastal water	1989-2005	12	1	92
Odense Fjord inner	1985-2009	25	1	732
Odense Fjord outer	1985-2014	30	1	1355
Onsala coastal water	1986-2014	29	1	329
Randers Fjord inner	1991-2014	23	1	500
Randers Fjord outer	1989-2009	19	1	403
Sejerø Bay	1989-2010	21	2	418
Skälderviken	1994-2014	21	6	399
Skive Fjord	1985-2014	30	3	1963
Southern Roskilde Fjord	1985-2014	30	1	836
Southern Sound coastal water	1992-1998	6	2	61
Vejle Fjord inner	1985-2014	29	2	751
Vejle Fjord outer	1985-2000	11	2	237