Long-term environmental monitoring for assessment of change: measurement inconsistencies over time and potential solutions

Kari E. Ellingsen[1*], Nigel G. Yoccoz[1,2], Torkild Tveraa[1], Judi E. Hewitt[3], Simon F. Thrush[4]

[1] Norwegian Institute for Nature Research (NINA), Fram Centre, P.O. Box 6606 Langnes, 9296 Tromsø, Norway

[2] Department of Arctic and Marine Biology, UiT The Arctic University of Norway, 9037 Tromsø, Norway

[3] National Institute of Water and Atmospheric Research, Hamilton, New Zealand

[4] Institute of Marine Sciences, University of Auckland, Auckland, New Zealand

[*] **Corresponding author**: E-mail: kari.ellingsen@nina.no, Phone: +47 41223760, ORCID: 0000-0002-2321-8278

Acknowledgments: KEE was supported by the Norwegian Oil and Gas Association (project no. 20-2013), the Norwegian Environment Agency (project no. 1204110 and 4013045), the Norwegian Research Council (project no. 212135) and Norwegian Institute for Nature Research (NINA). NGY and TT were supported by NINA. We thank one anonymous referee for useful comments on this article.

Abstract

The importance of long-term environmental monitoring and research for detecting and understanding changes in ecosystems and human impacts on natural systems is widely acknowledged. Over the last decades a number of critical components for successful long-term monitoring have been identified. One basic component is quality assurance/quality control protocols to ensure consistency and comparability of data. In Norway, the authorities require environmental monitoring of the impacts of the offshore petroleum industry on the Norwegian continental shelf, and in 1996 a large-scale regional environmental monitoring program was established. As a case study, we used a sub-set of data from this monitoring to explore concepts regarding best practices for long-term environmental monitoring. Specifically, we examined data from physical and chemical sediment samples and benthic macro-invertebrate assemblages from 11 stations from six sampling occasions during the period 1996-2011. Despite the established quality assessment and quality control protocols for this monitoring program, we identified several data challenges, such as, missing values and outliers, discrepancies in variable and station names, changes in procedures without calibration, and different taxonomic resolution. Furthermore, we show that the use of different laboratories over time makes it difficult to draw conclusions with regard to some of the observed changes. We offer recommendations to facilitate comparison of data over time. We also present a new procedure to handle different taxonomic resolution so valuable historical data is not discarded. These topics have a broader relevance and application than for our case study.

Keywords: data comparability; long-term monitoring; macrobenthos; oil and gas industry; taxonomic resolution

Introduction

There is a widespread recognition of the importance of long-term environmental monitoring and research for evaluating ecological responses to disturbance, and documenting and providing baselines against which change or extremes can be evaluated (Lindenmayer and Likens 2010). However, there are many challenges to be addressed in order to successfully assess changes and their causes (Lindenmayer and Likens 2010; Hughes 2014; Lindenmayer et al. 2015). During the last decades, characteristics of effective ecological monitoring have been summarized, and so have the reasons why monitoring may fail (Yoccoz et al. 2001; Lindenmayer and Likens 2010; Lindenmayer et al. 2015). Different designs fit different purposes, and it is important to address the question of "why monitor?" and clearly specify the objectives of a proposed monitoring programme. It is also important to evaluate what type, magnitude and causes of effect can or need to be detected with different designs. While it is crucial that environmental monitoring of human impacts on natural systems can detect temporal changes with sufficient reliability and sensitivity, estimating unbiasedly environmental changes and the impact of drivers is challenging (Yoccoz et al. 2001; Desaules 2012). These are challenges that call for targeted monitoring programmes designed to address specific questions (Nichols and Williams 2006).

Data comparability is a key requirement for any long-term monitoring program (Cao and Hawkins 2011). Using soil monitoring as an example, Desaules (2012) stated that measurement instability occurred along the whole measurement chain, from sampling to the expression of results. National monitoring programs in the US, member states of the European Union (EU) and the states of Australia, among others, are facing the challenge of comparability (Hughes and Peck 2008; Buss et al. 2015). With regard to the implementation of two large-extent US surveys, Hughes and Peck (2008) stated that consistent methods and levels of effort at all sampling points are necessary to distinguish differences in status/trend in ecological condition from differences in protocol.

The petroleum activity on the Norwegian continental shelf started in the late 1960's in the south, i.e. the Ekofisk fields (Gray et al. 1990), and since then exploration and extraction have been moving gradually northwards. Environmental monitoring in the Norwegian sector began already in 1973 at Ekofisk, and a thorough analysis of the Ekofisk and Eldfisk fields was done by Gray et al. (1990). A more comprehensive analysis of fields along the shelf was later done by Olsgard and Gray (1995). In 1996, a large-scale regional offshore environmental monitoring program, hereafter called the Regional Monitoring, was established for the Norwegian continental shelf (Gray et al. 1999; Bakke et al. 2011; 2013), as required by the Norwegian authorities (Iversen et al. 2015). At the same time, the focus on quality assurance/quality control protocols increased. In total, the Regional Monitoring covers about 1000 stations on the Norwegian continental shelf (Norwegian Oil and Gas 2013). The Regional Monitoring covers all oil fields on the Norwegian continental shelf, and the purpose of the monitoring is to provide an overview of the environmental status and trends in relation to the petroleum activities. The shelf has been divided into regions, from Region I in the south (North Sea) to Region XI in the north (Barents Sea, see Iversen et al. 2015), although the history and level of petroleum activity varies a lot among the regions. The requirement from the authorities is that the monitoring should be repeated every third year in regions with petroleum activity. In addition to all the field specific stations and reference stations, a minimum of 10 so-called regional stations were established in each region with petroleum activity. Some of these regional stations are in fact reference stations for specific fields, whereas other regional stations are not linked to any fields. The intention was that the regional stations should not be impacted by the oil and gas industry. The purpose of the establishment of these regional stations was to provide data for long-term changes within a given region such as those due to climate change (Gray et al. 1999). Indeed, a critical evaluation of the environmental status at such regional stations can be crucial to understand human impacts vs natural variation in this marine system over space and time.

The availability of adequate and sustained funding is one of the critical elements for maintaining effective monitoring programs (Hewitt and Thrush 2007; Lindenmayer and Likens 2010; Mieszkowska et al. 2014). Financial constraints often result in a trade-off between sampling in space

or time (Hewitt and Thrush 2007). Indeed, financial costs can have a major impact on deciding an acceptable level of uncertainty and can be the limiting factor in choosing an analytical strategy (Bennett et al. 2014). The Regional Monitoring is an example of a long-term monitoring program where reliable funding has been secured because the oil and gas industry has a financial commitment.

Each year the Norwegian Oil and Gas Association has contracted out projects to consulting companies to conduct the Regional Monitoring in some selected regions (Iversen et al. 2015). This procedure is based on current Norwegian legislation. The consequence is that different companies have conducted sampling, data processing, laboratory analyses, and written the survey reports for different years. Although the Norwegian Environment Agency has provided guidelines for fieldwork and data processing procedures and reporting (Iversen et al. 2011; 2015), this process requires successful inter-laboratory comparisons in practice. If this is not the situation, this procedure is in contrast to the current international recommendations of data integrity, specifically linked to the critical component of stability and competence of staff for long-term monitoring (e.g., Lindenmayer and Likens 2010).

According to the guidelines for the Regional Monitoring, the results should be comparable over time and among different regions (Iversen et al. 2015). Data from 11 regional stations on the southern part of the Norwegian continental shelf (Region I, i.e. the Ekofisk region) have given us the opportunity to examine temporal (and spatial) patterns of benthic communities and environmental variables from six sampling occasions over a rather extensive area, and explore best practices for long-term environmental monitoring. We have identified a number of challenges with regard to measurement inconsistencies over time, including outliers and missing values, discrepancies in variable and stations names, changes in procedures without calibration, and different taxonomic resolution. Accordingly, the Regional Monitoring is facing challenges in terms of data comparability (see e.g. Hughes and Peck 2008; Buss et al. 2015). Here we provide recommendations linked to the specific challenges identified to facilitate comparison of data over time. Although we have only focused on a case study using regional stations from one particular region, this issue is also relevant for the field specific stations in this region, the other regions along the Norwegian continental shelf, and with regard to a comparison

of data among the different regions on the shelf. Clearly, these topics also have a broader relevance and application than for the Norwegian Regional Monitoring.

The Regional Monitoring Protocol

The Norwegian continental shelf has been divided into 11 regions (Iversen et al. 2015). Within each region, a sampling design is arranged with a set of field stations located at different distances (250 m, 500 m, 1000 m, 2000 m etc.) and in different directions (depending of the dominating direction of currents) from each oil field, and with accompanying reference stations. In addition, the set of regional stations are evenly spatially distributed within a given region (see Fig. 1 for Region I). The regional stations are sampled during the same monitoring surveys as for the field stations and reference stations. Over time there has been some changes in the number (and location) of regional stations within regions. One reason for this may be that new oil or gas fields are established on the shelf, and some regional stations are located too close to the new fields and they may now be regarded as impacted by the petroleum industry. Another reason may be that fields are closed, and if a regional station is the same as a reference station linked to a field, the monitoring at this station may end.

The guidelines for the Regional Monitoring, including quality assurance/quality control protocols, are regularly updated (Iversen et al. 2011; 2015). These guidelines have improved, and this may potentially have improved the data quality over time. Consulting companies conducting the fieldwork and data processing write reports that are evaluated by the Norwegian Environment Agency and an independent expert group, appointed by the agency, and finally the reports are approved by the Environment Agency (Iversen et al. 2015). The data are collected in the Environmental Monitoring Database (hereafter called the MOD-database), owned by the Norwegian Oil and Gas Association. Although the survey reports are evaluated, the actual datasets that are included in the MOD-database are not evaluated by any independent expert group. The Norwegian Environment Agency has an open data policy and all data are open to public scrutiny (Olsgard and Gray 1995), which also makes the data potentially available to the scientific community. The amount of data collected in the MOD-

database may be considered as large at an international scale, for example, data from Regions I-IV and VI-VII have been collected every third year, starting in 1996, 1997 or 1998 depending on the region.

In the Regional Monitoring, biological, physical and chemical samples are taken from the bottom sediments with a 0.1 m^2 van Veen grab. At each station, five replicates for analyses of macrobenthos are taken. Biological samples are sieved on a 1 mm round-hole diameter sieve, and retained fauna are fixed in formalin for later identification. Three additional grabs are taken at each station for analyses of sediment variables. Sub-samples are taken from the upper 5 cm of the sediment for analyses of physical sediment characteristics, and from the upper 1 cm for chemical analyses. Sample station positioning employs a differential GPS (global positioning system, accuracy of <10 m) with the vessel held in position with a dynamic positioning system (DP).

At the time when the Regional Monitoring was established, sectorial based monitoring was commonly employed. However, during the last decades there has been a gradual change towards a more ecosystem-based monitoring. Marine biodiversity faces unprecedented threats from multiple pressures arising from human activities operating at global, regional and local scales (e.g., Mieszkowska et al. 2014; Thrush et al. 2015). On the Norwegian continental shelf, for example, the oil and gas industry is unlikely to be the only driver of change in a changing world and in a multi-use ecosystem. With regard to the seafloor, fishing activity is one example of human disturbance (Frid et al. 2000; Thrush and Dayton 2002; Kaiser et al. 2006). Other potential drivers of changes are climate change including ocean acidification (Gattuso et al. 2015). Yet, none of these potential multiple stressor effects have been used to inform either the monitoring design of the oil and gas industry on the Norwegian continental shelf, nor its analysis.

Case study: Regional stations from Region I

We used data on soft-sediment macrobenthos and chemical and physical characteristics of the sediment from regional stations collected at the southern part of the Norwegian continental shelf

(Region I, i.e. the Ekofisk-region; Iversen et al. 2011; 2015, Fig. 1). The sampling frame spanned approximately 130 km in a south-north direction and approximately 70 km from east to west (56°02' to 57°08' N, 2°30' to 3°49' E). The sampling was conducted in May-June, starting in 1996 with a repetition every third year, i.e. data from six sampling occasions (1996-2011) was available for our case study. Water depth (m) at the regional stations was similar (ranging from 65 to 72 m), and the sediment was dominated by fine sand (Table 1). Different consulting companies conducted the fieldwork at the different sampling occasions (Online Resource 1). The faunal identification for each monitoring year was performed by one of two consulting companies (hereafter called laboratory A or B, Online Resource 1), but sometimes with additional assistance by other national or international experts (for further information see survey reports: Cochrane et al. 2009; Jensen et al. 2000; Mannvik et al. 1997; 2012; Nøland et al. 2003; 2006). We used a version of the MOD-database from March 2013. In Region I there have been some changes in the number of regional stations over time. We selected 11 regional stations for our case study based on the criteria that the regional stations should have been investigated at least five times. Based primarily on number of sampling occasions, we selected data on total organic matter (TOM), sediment median grain size, sorting, skewness, kurtosis, gravel, total sand, fine sand, coarse sand, silt-clay (i.e. pelite; fraction of sediment < 0.0063 mm), total hydrocarbons (THC), polycyclic aromatic hydrocarbons (PAH), and the metals barium (Ba), cadmium (Cd), chromium (Cr), copper (Cu), iron (Fe), mercury (Hg), lead (Pb) and zinc (Zn) for our case study.

Data challenges and recommendations

The processing of the data extracted from the MOD-database has been time-consuming. It was also difficult to find all relevant data in the MOD-database, because some of the regional stations and some of the variables had been given different names in different years both in the MOD-database and in the survey reports. For a number of species, different synonyms are given for different years. For the physical and chemical characteristics of the sediments, a number of variables have only been measured (or included in the MOD-database) sporadically, and it is not obvious why some of these

variables have been measured at all. In our analyses we have used the values given in the database if there were discrepancies between the MOD-database and the survey reports.

Outliers, errors and missing values

Despite the current quality assessment and quality control protocols required by the Norwegian Environment Agency (Iversen et al. 2015), we identified several data challenges when examining data from the 11 regional stations in our case study. This shows that theory and practice is not always working hand in hand. First, we plotted all the selected environmental variables to identify outliers, erroneous values or missing values. One example of an outlier is that the TOM value for one replicate was about 8 times lower than for the other replicates (station 8 in 1996: 0.11 for replicate number 3 versus 0.83-0.93 for the other four replicates). Such outliers (in historical data) may be deleted prior to further analyses. However, this might be a data-entry error, and inspection of raw data sheets may confirm this. Furthermore, values of Cr in the sediment samples from 2011 are anomalous for some stations (stations 6, 9, 11, and 12, Fig. 2a). Ideally, such samples should have been rerun in the laboratory immediately. When this has not been done, these values can be corrected prior to further analyses, for example, by replacing them by a prediction from either a station-specific linear model having year as a predictor to take into account the trends observed in each station, and excluding the anomalous 2011 observations (as shown in Fig. 2b), or alternatively by a linear model for all stations but including a year by station interaction. Such imputation of missing data can be taken into account in latter analyses, as long as it is known which value is a direct measurement and which one is a derived value (Little and Rubin 1987). Note that we have only Cr data from 4 sampling occasions prior to 2011, so this correction should be checked when more data are available from future monitoring. For Zn, the values in 2011 are anomalous at several stations (Fig. 3), and we have therefore not included Zn in other analyses. Likewise, values of Cd in the sediment from 2005 for stations 4 and 5 were significantly higher than in all other samples both in space and time (Fig. 4). We have not included Cd in other analyses because there are few sampling occasions at several stations (only 3 or 4 years if the year 1996 is not included).

In future we recommend that values are plotted immediately after laboratory analyses in order to identify erroneous and missing values and rerun analyses when necessary. For environmental variables, the replicates could be immediately plotted against the previous results (from other years and other stations from the same year). If they do not show the same trend or fall outside a given threshold (e.g. 2 SD) of the previous data, another subsample from each replicate could be done, and if only one replicate lies outside, that replicate could be repeated. All data should be archived irrespective of their anomalousness, as they might represent warning signals of change and not just measurement errors.

For one station (station 9) from 1996 there was no faunal data given in the MOD-database. Likewise, some sediment characteristics (e.g., fine sand, coarse sand, gravel, sediment median grain size, sorting, skewness, and kurtosis) and some chemical variables (Cr, Cd, Cu, Fe, and Hg) were not measured on all sampling occasions. One specific example of missing data is that median grain size was only given for one station (station 9) from 1999 in the MOD-database. Raw data sheets may confirm if there were no more samples analysed in 1999. Cr was not measured in 1996, but it can be included in temporal analyses excluding this first year. For the years 1996, 2008 and 2011 values of "fine sand" are given in the database, whereas for the other years values of "total sand" is given. Although most of the "total sand" is "fine sand" this is obviously problematic for temporal analyses. In general, a relationship between variables might be used to make predictions about missing values. However, lack of strong relationships did not allow for any precise estimation of missing values in our case study.

Grooming the data will always be necessary, and some level of errors or problems with the data is expected with regard to the size of this database (see also Hughes and Peck 2008). In the future, we suggest that all data should be tested and corrected for errors before they are included in the MOD-database. It is much easier to resolve a problem in long-term data when it is identified in a timely fashion and while observers and methods are still available for examination and discussion (Lindenmayer and Likens 2010). With data included in a database, it is difficult to remember what

happened in the laboratory many years ago. When errors are identified, this should immediately be reported to the organisation responsible for updating and maintaining the database. It is important to immediately contact the laboratory responsible for the analyses and try to find out what might be the reason for errors. Our findings highlight that the actual *use* of long-term data sets is the primary way that errors, artefacts or other problems are uncovered (Lindenmayer and Likens 2010). Otherwise, errors and inconsistencies may discourage thorough reanalysis.

The use of different laboratories over time

The sediment median grain size values from 1996 are clearly at a higher level than for the rest of the sampling period (Fig. 5). This greatly influences the estimated trends; a linear mixed model with station as a random factor and year as a fixed effect would result in a decrease per year of -0.062 (s.e. = 0.0046), but removing 1996 more than halves the decrease to -0.028 (s.e. = 0.0031), with non-overlapping 95% confidence intervals. One particular laboratory was responsible for the analyses of physical properties of the sediment in 1996, while different laboratories performed the analyses for other years (Online Resource 1). As 1996 is the first year in this regional environmental monitoring program, we have not included median grain size values from 1996 in further analyses of temporal patterns.

We used data on TOM, fraction of silt and clay in the sediment, THC, PAH, Ba, Pb, and Cr for further analyses of temporal patterns. We used Principal Component Analyses (PCA), using standardized variables (subtracting the mean and dividing by the standard deviation). Replicates were first averaged by station and year, and then standardized. The first PCA axis is a linear combination of the variables maximizing the sum of squared correlations with all environmental variables. We therefore checked if the different variables were approximately linearly correlated. To analyze change over time, we used linear mixed models, with station and station by year as nested random factors. This allows for dependency among observations made on the same station in a given year and of repeated observations of the same station in different years, as well as estimating variance components

300 (repeated measurements of the same station in a given year, and stations over time). The analyses were
301 implemented in R, version 3.1.3 (R Core Team 2015). We used the libraries lme4 for linear mixed
302 models, and ade4 for PCA.
303
304 For a number of individual environmental variables (Ba, PAH, THC) where we had not already
305 identified any serious issues with the data (see above), the inclusion of the year 1996 in temporal
306 analyses turned out to be important for whether there was a trend in the data or not (Table 2).
307 Moreover, multivariate analyses based on a combined set of variables (TOM, silt-clay content, THC,
308 PAH, Ba and Pb) also revealed that 1996 was clearly separated from all following years of sampling
309 (Fig. 6), and this year had higher mean standardized values and lower variability than the following
310 years for the variables included in the analyses (Fig. 6). Importantly, in 1996, different laboratories
311 were used for sampling and analyses of both sediment organic chemistry, metals and physical
312 properties than for the next years (Online Resource 1). Note that if 1996 is omitted from the time
313 series, there is no clear trend in the data (Fig. 6). Although using a single laboratory to test all samples
314 may ensure consistency in the quality of results, it does not guarantee adequate quality. In this case, as
315 when several laboratories are participating in testing, it is imperative that inter-laboratory comparisons
316 are conducted (see Arrouays et al. 2012; Ross et al. 2015). Each year, some samples can be distributed
317 among the different laboratories that are involved in the monitoring over time. The results can then be
318 compared as a blind test each year. As an example, Ross et al. (2015) collected and distributed forest
319 soil samples to 15 laboratories in the eastern United States and Canada, and tested for variability
320 among laboratories for a number of soil properties. They recommended the continuation of reference
321 soil exchange programs to quantify the uncertainty associated with these analyses. Alternatively,
322 periodic analysis of reference samples is a widespread quality assurance procedure (Desaules 2012).
323 Replicate samples from the first sampling occasion or baseline study (site reference samples) can be
324 reanalysed simultaneously with the corresponding samples of each following sampling occasion. The
325 results can then be corrected based on the reanalysed first campaign samples, which correspond to a
326 site-specific control. This procedure would also enable different staff/laboratories to participate in the

long-term monitoring. However, it is important to evaluate issues of long-term storage effects on measurements, and if this could complicate the analyses (see Ross et al. 2015).

Changes in protocols without calibration

Over time, there has been changes in the protocols of the Regional Monitoring. Because all the datasets from the different years are collected in the MOD-database, and the changes in methods are not flagged in the database, users may download data without knowledge about these changes. No calibration has been done when protocols have been changed, i.e., running different protocols in parallel, at least this information is not given in the database. The monitoring of TOM in the sediment is one example of changes in procedures. From 2005, only one value was given for TOM in the MOD-database, because there was a change in procedure and sub-samples from three replicated grabs were pooled prior to analyses instead of analysing the three replicates separately (Nøland et al. 2006). Another example involves PAHs, which represent a group of compounds, and for the laboratory analyses of PAHs in the sediment not every year has a description (either in the survey reports written by the consulting companies or in the MOD-database) of exactly which compounds have been analyzed and included in the term 'PAH' in the MOD-database. It is therefore unclear if these values are comparable among years. Finally, only one replicate was analysed for metals in the sediments in 2011 compared to three or more replicates in previous years (Mannvik et al. 2012). This is particularly problematic because values of some of the metals in the sediment samples from 2011, such as Cr and Zn, are anomalous for several stations (see above, Fig. 2a, Fig. 3). It is not possible to decide if these changes are caused by changes in the monitoring program or other factors. Importantly, modifying methods can affect the ability to track changes in condition over repeated surveys (e.g. Hughes and Peck 2008). If the protocols need to be changed, we recommend calibrating the new methods with the previous methods, documenting the change and any effects on the measured variable (Lindenmayer and Likens 2010; Lindenmayer et al. 2015), and giving information on this in the database.

The need for a modernisation of the MOD-database

In our case study, we have only focused on 11 stations from the MOD-database, however, our findings are relevant for the entire Regional Monitoring. Importantly, a modification of the MOD-database could hinder a number of errors and problems with the data, by for example not accepting errors in species names, confusion with regard to (different) names of a given station or variable used for different years, different level of information included in the database (e.g. “total sand” or “fine sand”), and not accepting values larger or smaller than a given level. It could also take into account changes in nomenclature over time. Currently, it takes more than 1.5 years after the data are collected until they are included in the MOD-database (Iversen et al. 2015). Ideally, the data should be secured in the MOD-database much more quickly. In our opinion, the MOD-database needs to be modernised and restructured in order to secure the data in a better way and make the data more accessible for users (including those processing the samples). Furthermore, we recommend that all analyses of data from the Regional Monitoring, including the survey reports written by the consulting companies, should be based on data downloaded from the (modified) MOD-database, and not from different databases owned and controlled by the different consulting companies.

The importance of effective communication of knowledge can never be over emphasised. When the knowledge is easy to access and openly available to all, including the management, industry, and the public, this process will be more efficient. Interactive visualisation of data, analyses and results at websites, can efficiently provide such information on environmental status and trends, with only minimal effort of the user. We note that the construction of such interactive websites for long-term monitoring data is currently rapidly increasing (Loraine et al. 2015), and we propose that the data challenges outlined here could have been easily detected at an earlier stage if interactive tools for analyses had been made available.

Uneven taxonomic resolution: potential consequences and recommendations

382 According to the current guidelines for the Regional Monitoring, the aim is that the taxa should be
383 identified to the species level, as far as possible (Iversen et al. 2015). All taxa collected during
384 sampling are included in the MOD-database, irrespective of the level of identification. For our case
385 study, 25% (98 of 388 taxa) of the data downloaded from the MOD-database were identified to a
386 coarser taxonomic level than species level, either at one or more sampling occasions. Over time,
387 different laboratories have been responsible for the species identification (Online Resource 1).
388 Consequently, taxa with "uncertain" classification (i.e. not identified to the species level) may vary
389 among years (and regions) if the competence of staff/laboratories differ. A taxon in a species list
390 extracted from the MOD-database can therefore appear as two different taxonomic units, and this
391 complicates the comparison of faunal patterns at stations over time (and space).
392
393 A new procedure to handle different taxonomic resolution
394
395 The large amount of historical data (since 1996) from the Regional Monitoring are highly valuable.
396 However, when we are using data from the MOD-database for scientific purpose, our aim is that
397 unidentified taxa should only be included in data analyses if they cannot be mistaken for other
398 identified species (or taxa in general). In earlier publications using data from the MOD-database, we
399 have subjectively processed the data prior to data analyses by either deleting or pooling taxa with
400 uncertain classifications (e.g., Ellingsen 2001; 2002; Ellingsen and Gray 2002). However, this
401 procedure is time consuming and, in addition, it is not easily repeatable for other users. Here we
402 present a new procedure, described and implemented in an R script, in order to transparently adjust the
403 faunal data with uncertain taxonomic classifications prior to data analyses. The procedure can be
404 illustrated using the two taxonomic levels genus and species. A genus can appear either as such (e.g.,
405 *Sphaerodorum* spp) or as a species (e.g., *Sphaerodorum gracilis*). If there is only one species in a
406 genus, the genus can unambiguously be affected to the species, and either can be used. If there are
407 more than one species, we can consider two alternative solutions: alternative 1) "lumping" all species
408 in the genus and consider the genus as the only taxa, or alternative 2) removing observations appearing
409 only as genus and keep the species observations ("splitting"). The problem is that there are situations

where it makes sense to use the first alternative (e.g. first years appear only as genus, later years only as many different species), whereas the second alternative may appear best in some cases (few observations appear as genus and most as species). However, the first alternative may appear best if most observations appear as genus and few as species, since removing the genus would remove much information. We used the two approaches (lumping/splitting) in our case study to assess the sensitivity of analyses to the two choices, but the most important is to make the choices transparent (and easily modified if judged necessary) to assess the robustness of results.

Patterns of faunal composition – with and without correction of uncertain classifications

Prior to the data analyses, taxonomic groups not properly sampled by the methods used (Nematoda, Foraminifera), colonial groups (Porifera, Hydrozoa, Bryozoa), pelagic crustaceans (Calanoida, Mysidacea, Hyperiidae, Euphausiacea), and juveniles were excluded from the species list. Data analyses were done on species abundance data at the replicate level. For our case study, the total number of taxa in the unadjusted data set (i.e. directly downloaded from the MOD-database, but excluding the taxonomic groups mentioned above) from all 11 regional stations and all six sampling years was 388 taxa. The total number of taxa after adjusting for taxonomic classification uncertainties was 294 (modification alternative 1, i.e. based on lumping) or 314 (modification alternative 2, i.e. based on splitting).

We wanted to explore whether there were any differences in patterns of faunal composition based on the unadjusted data vs the lumping/splitting data. In order to examine faunal composition over time (and space) we used two approaches: Nonmetric Multidimensional Scaling (NMDS) using Bray-Curtis distance to measure dissimilarity among samples (which corresponds to one station in a given year), and Canonical Correspondence Analysis (CCA). We used both these analytical approaches because both are commonly used when analyzing faunal composition, yet they are based on somewhat different methodology. The analyses were implemented in R, version 3.1.3 (R Core Team 2015). We used the libraries vegan for NMDS and ade4 for ordination analyses.

Using the unadjusted data set in the NMDS analysis clearly showed a separation between the years 1996, 1999-2002-2005, and 2008-2011 (Fig. 7). It is important to note that laboratory A was responsible for sampling and taxa identification in 1996 and 2008-2011, whereas laboratory B was responsible for the other three years (Online Resource 1). Because of this, it is difficult to ascertain if, for example, there is a real difference between 1996 and the following years, or if this is a laboratory effect. A CCA analysis based on the unadjusted faunal data also showed that the year 1996 was clearly separated from all other years (Fig. 8), and this was even clearer than for the NMDS analysis (Fig. 7). Within any laboratory, taxonomic competence can change over time, and among laboratories such differences can be even larger. For the Regional Monitoring, this means that for one year a particular taxon is identified to the species level, whereas for another year the same taxon may be identified to a coarser taxonomic level, meaning that the taxon may be the same but it appears as different taxa for the two years if data from these two years are combined. Some of the consultancy firms that perform the Regional Monitoring have already speculated on potential effects of different practices. In a report focusing on the period 1996-2006, Renaud et al. (2008) found strong inter-annual differences in community structure, and suggested that this was likely due to changes in industry practices, as well as natural variability in recruitment and mortality. Here we have illustrated the potential consequences of comparing data among years (and laboratories) without considering the issue of stability and taxonomic competence.

After applying our suggested new procedure for correcting the uncertain classifications, i.e. our alternative 1 "lumping data" to 294 taxa, we see that the difference between 1999-2002-2005 and the other years is not as large as for the unadjusted data (NMDS-plot, Fig. 7). Using our alternative 2 "splitting data", we also found that the difference between 1999-2002-2005 and the other years is not as large (NMDS, Online Resource 2). Likewise, using the adjusted "lumped" data set in the CCA analysis showed that the faunal composition among years appear to be more similar than using the unadjusted faunal data, although the first axis is still related to the separation between 1996 and all other years (Fig. 8, for modification alternative 2 see Online Resource 2). This means that without

466 adjusting the data prior to the data analyses the years appear as more dissimilar to each other than what
467 they likely are. Yet, the differences were still substantial after adjustment.
468
469 For the shorter period from 1999 to 2005, when one laboratory was responsible for all three surveys,
470 there was a change in faunal composition from 1999 to 2002 to 2005 (Fig. 7). However, with only
471 three sampling occasions our ability to ascertain if this is a temporal trend is of course limited. Short
472 time series often constrain inferences about change because trend detection is limited by the number of
473 data points and temporal extent affecting whether a cyclic pattern is identified as a monotonic trend.
474 This requires a decision about how to treat trend detection involving consideration of the magnitude of
475 change vs statistical significance and the strong examination of patterns in residuals.
476
477 While we have suggested a procedure that takes into account the problem of uncertain species
478 identification in terms of data analysis, other options include taxonomists going back to archived
479 voucher specimens or image library. One alternative approach could also be to compare data at a
480 higher taxonomic level than at the species level (see e.g. Terlizzi et al. 2009; Fontaine et al. 2015), or
481 only focus on particular species or groups of species. However, we might expect the full community
482 identification to be more sensitive to identify community changes over space and time, although this
483 require further examination.
484
485 We have shown that the current procedures and, in particular, the use of different laboratories over
486 time strongly limit the utility of the historical data (since 1996) from the Norwegian continental shelf,
487 despite that the goal in the Regional Monitoring is to detect long-term trends and also provide an
488 estimate of the background conditions. Lindenmayer and Likens (2010) emphasize that several
489 seemingly small factors can contribute enormously to the success of long-term monitoring, sometimes
490 out of proportion to expectations. Stability and competence of staff is one such critical component, and
491 indeed, consistency increases comparability of data (Hughes and Peck 2008). Accordingly,
492 management of monitoring programmes that allow different companies to be responsible for

493 implementation over space or time, without extensive inter-calibration is not in accordance with the
494 current international recommendations (Arrouays et al. 2012).
495
496 **Summary and recommendations**
497
498 We have used a sub-set of data from a large-scale offshore regional environmental monitoring
499 program in Norway to explore concepts regarding best practices for long-term environmental
500 monitoring. The purpose of the Norwegian monitoring is to provide an overview of the environmental
501 status and trends in relation to the petroleum activities on the Norwegian continental shelf. Although,
502 there has been a focus on quality assurance/quality control protocols, we have identified a number of
503 challenges with regard to measurement inconsistencies, including discrepancies in variable and station
504 names, changes in procedures without proper calibration, different taxonomic resolution and changes
505 to nomenclature as well as missing values and outliers. Currently, it is difficult to decide if some of the
506 observed temporal changes at the stations in our case study are caused by natural variation, human
507 pressure, or simply by changes in the monitoring program over time, such as e.g. changes in
508 laboratory. We provide recommendations linked to these challenges to facilitate comparison of data
509 over time. We also present a new procedure, described and implemented in an R script, in order to
510 transparently adjust faunal data with uncertain taxonomic classifications prior to data analyses. Tightly
511 connected to these issues, we suggest that the data are carefully secured in a modernised database,
512 including that the data are constantly updated, regularly scrutinised for errors and rigorously reviewed,
513 and to make the data more accessible for users in Norway and elsewhere (see e.g. Fölster et al. 2014).
514 At present, a large part of the text in the database is given in Norwegian, which limit the availability
515 for international users. The construction of interactive websites for long-term monitoring data is
516 currently rapidly increasing, and we propose that the data challenges outlined here could have been
517 easily detected at an earlier stage if interactive tools for analyses were made available. Importantly,
518 frequent examination and use of data also result in important discoveries and stimulate new research
519 and management questions (Lindenmayer and Likens 2010). The Norwegian monitoring is an example
520 of a long-term monitoring program where reliable funding has been secured because the oil and gas

industry has a financial commitment. What is important is that the available resources are utilised in a way that warrant both spatial and temporal comparisons. Furthermore, during the last decades it has been a gradual change worldwide from a sectorial based monitoring (that is, with a focus on oil and gas industry only) towards a more ecosystem-based monitoring. We advocate for revision and updating of the Norwegian regional environmental monitoring program where potential multiple stressor effects (e.g. bottom fishing, climate change) are used to inform the monitoring of the oil and gas industry on the Norwegian continental shelf. This process would require a tight collaboration between authorities, different industries and scientists.

References

Arrouays, D., Marchant, B. P., Saby, N. P. A., Meersmans, J., Orton, T. G., Martin, M. P., Bellamy, P. H., Lark, R. M., & Kibblewhite, M. (2012). Generic issues on broad-scale soil monitoring schemes: a review. *Pedosphere*, 22, 456-469.

Bakke, T., Green, A. M. V., & Iversen, P. E. (2011). Offshore environmental monitoring in Norway – Regulations, results and developments. In K. Lee & J. Neff (Eds.), *Produced Water* (pp. 481-491). Springer, NY (Chapter 25).

Bakke, T., Klungsøyr, J., & Sanni, S. (2013). Environmental impacts of produced water and drilling waste discharges from the Norwegian offshore petroleum industry. *Marine Environmental Research*, 92, 154-169.

Bennett, J. R., Sisson, D. S., Smol, J. P., Cumming, B. F., Possingham, H. P., & Buckley, Y. M. (2014). Optimizing taxonomic resolution and sampling effort to design cost-effective ecological models for environmental assessment. *Journal of Applied Ecology*, 51, 1722-1732.

548 Buss, D. F., Carlisle, D. M., Chon, T.-S., Culp, J., Harding, J. S., Keizer-Vlek, H. E., et al., (2015).
549 Stream biomonitoring using macroinvertebrates around the globe: a comparison of large-scale
550 programs. *Environmental Monitoring and Assessment*, 187, 4132, doi: 10.1007/s10661-014-4132-8.
551
552 Cao, Y, & Hawkins, C. P. (2011). The comparability of bioassessments: a review of conceptual and
553 methodological issues. *Journal of the North American Benthological Society*, 30(3), 680-701.
554
555 Cochrane, S., Palerud, R., Wasbotten, I. H., Larsen, L. H., & Mannvik, H. P. (2009). Offshore
556 sediment survey of Region I, 2008. Akvaplan-niva report no. 4215 - 02. Akvaplan-niva, Tromsø,
557 Norway. 314 pp.
558
559 Desaules, A. (2012). Measurement instability and temporal bias in chemical soil monitoring: sources
560 and control measures. *Environmental Monitoring and Assessment*, 184, 487-502.
561
562 Ellingsen, K. E. (2001). Biodiversity of a continental shelf soft-sediment macrobenthos community.
563 *Marine Ecology Progress Series*, 218, 1-15.
564
565 Ellingsen, K. E. (2002). Continental shelf soft-sediment benthic biodiversity in relation to
566 environmental variability. *Marine Ecology Progress Series*, 232, 15-27.
567
568 Ellingsen, K. E., & Gray, J. S. (2002). Spatial patterns of benthic diversity: is there a latitudinal
569 gradient along the Norwegian continental shelf? *Journal of Animal Ecology*, 71, 373-389.
570
571 Fontaine, A., Devillers, R., Peres-Neto, P. R., & Johnson, L. E. (2015). Delineating marine ecological
572 units: a novel approach for deciding which taxonomic group to use and which taxonomic resolution to
573 choose. *Diversity and Distributions*, 21, 1167-1180.
574

575 Frid, C. L. J., Harwood, K. G., Hall, S. J., & Hall J. A. (2000). Long-term changes in the benthic
576 communities on North Sea fishing grounds. *ICES Journal of Marine Science*, 57, 1303-1309.
577
578 Fölster, J., Johnson, R. K., Futter, M. N., & Wilander, A. (2014). The Swedish monitoring of surface
579 waters: 50 years of adaptive monitoring. *Ambio*, 43, 3-18.
580
581 Gattuso, J.-P., Magnan, A., Billé, R., Cheung, W. W. L., Howes, E. L., Joos, F., et al., (2015).
582 Contrasting futures for ocean and society from different anthropogenic CO2 emissions scenarios.
583 *Science*, 349.
584
585 Gray, J. S., Clarke, K. R., Warwick, R. M., & Hobbs, G. (1990). Detection of initial effects of
586 pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. *Marine*
587 *Ecology Progress Series*, 66, 285-299.
588
589 Gray, J. S., Bakke, T., Beck, H. J., & Nilssen, I. (1999). Managing the environment effects of the
590 Norwegian oil and gas industry: from conflict to consensus. *Marine Pollution Bulletin*, 38(7), 525-
591 530.
592
593 Hewitt, J. E., & Thrush, S. F. (2007). Effective long-term ecological monitoring using spatially and
594 temporally nested sampling. *Environmental Monitoring and Assessment*, 133, 295-307.
595
596 Hughes, B. (2014). Monitoring: Garbage In Yields Garbage Out. *Fisheries*, 39(6), 243-243, doi:
597 10.1080/03632415.2014.915813.
598
599 Hughes, R. M., & Peck, D. V. (2008). Acquiring data for large aquatic resource surveys: the art of
600 compromise among science, logistics, and reality. *Journal of the North American Benthological*
601 *Society*, 27(4), 837-859.
602

603 Iversen, P. E., Green, A. M. V., Lind, M. J., Petersen, M. R. H., Bakke, T., Lichtenthaler, R., et al.,
604 (2011). Guidelines for offshore environmental monitoring: The petroleum sector on the Norwegian
605 Continental Shelf. Climate and Pollution Agency. TA number 2849/2011. 49 pp.
606
607 Iversen, P. E. Lind, M. J., Ersvik, M., Rønning, I., Skaare, B. B., Green, A. M. V., et al., (2015).
608 Guidelines for environmental monitoring of petroleum activities on the Norwegian continental shelf.
609 The Norwegian Environment. Agency M-number M-300/2015. 60 pp. (In Norwegian)
610
611 Jensen, T., Gjøs, N., Nøland, S.-A., Oreld, F., Møskeland, T., Bakke, S. M., et al., (2000).
612 Environmental Monitoring 1999, Region I – Ekofisk. Technical Report. Report no. 2000-3238. Det
613 Norske Veritas & Sintef Applied Chemistry, Norway. 294 pp.
614
615 Kaiser, M. J., Clarke, K. R., Hinz, H., Austen, M. C. V., Somerfield, P. J., & Karakassis, I. (2006).
616 Global analysis of response and recovery of benthic biota to fishing. *Marine Ecology Progress Series*,
617 311, 1-14.
618
619 Lindenmayer, D. B., Burns, E. L., Tennant, P., Dickman, C. R., Green, P.T., Keith, D. A., et al.,
620 (2015). Contemplating the future: Acting now on long-term monitoring to answer 2050's questions.
621 *Austral Ecology*, 40, 213-224.
622
623 Lindenmayer, D. B., & Likens, G. E. (2010). *Effective ecological monitoring*. CSIRO Publishing,
624 London, 170 pp.
625
626 Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley and Sons,
627 New York.
628

629 Loraine, A. E., Blakley, I. C., Jagadeesan, S., Harper, J., Miller, G., & Firon, N. (2015). Analysis and
630 visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome
631 Browser. *Methods in molecular biology (Clifton, N. J.)*, 1284, 481-501.
632
633 Mannvik, H. P., Pearson, T., Pettersen, A., & Lie Gabrielsen, K. (1997). Environmental monitoring
634 survey Region I 1996. Main Report. Akvaplan-niva report no. 411.96.996-1. Akvaplan-Niva, Tromsø,
635 Norway. 246 pp.
636
637 Mannvik, H. P., Wasbotten, I. H., Cochrane, S., & Moldes-Anaya, A. (2012). Miljøundersøkelse
638 Region I, 2011. Akvaplan-niva report no. 5339.02. Akvaplan-niva, Tromsø, Norway. 196 pp. (In
639 Norwegian).
640
641 Mieszkowska, N., Sugden, H., Firth, L. B., & Hawkins, S. J. (2014). The role of sustained
642 observations in tracking impacts of environmental change on marine biodiversity and ecosystems.
643 *Philosophical Transactions of the Royal Society A*, 372, 20130339.
644
645 Nichols, J. D., & Williams, B. K. (2006). Monitoring for conservation. *Trends in Ecology &*
646 *Evolution*, 21, 668-673.
647
648 Norwegian Oil and Gas (2013). Environmental Report 2013. The Norwegian Oil and Gas Association.
649 http://www.norskoljeoggass.no/en/Publica/Environmentalreports/Environmental-report-2013/.
650
651 Nøland, S. A., Gjøs, N., Bakke, S. M., & Oreld F. (2003). Environmental Monitoring 2002, Region I –
652 Ekofisk. Main report. Technical Report. Report no. 2003-0338. Det Norske Veritas/Sintef, Norway.
653 316 pp.
654
655 Nøland, S. A., Bakke, S. M., Rustad, I., & Brinchmann, K. M. (2006). Environmental Monitoring
656 Region I, 2005. Main Report. Report no. 2006-0187. Det Norske Veritas, Norway. 344 pp.

Olsgard, F., & Gray, J.S. (1995). A comprehensive analysis of the effects of offshore oil and gas exploration and production on the benthic communities of the Norwegian continental shelf. *Marine Ecology Progress Series*, 122, 277-306.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Renaud, P. E., Jensen, T., Wassbotten, I., Mannvik, H. P., & Botnen, H., (2008). Offshore sediment monitoring on the Norwegian shelf. A regional approach 1996-2006. Akvaplan-niva report no 3487 – 003. Akvaplan-Niva, Tromsø, Norway. 95 pp.

Ross, D.S., Bailey, S.W., Briggs, R.D., Curry, J., Fernandez, I.J., Fredriksen, G., et al., (2015). Inter-laboratory variation in the chemical analysis of acidic forest soil reference samples from eastern North America. *Ecosphere*, 6(5), 73. http://dx.doi.org/10.1890/ES14-00209.1.

Terlizzi, A., Anderson, M. J., Bevilacqua, S., Fraschetti, S., Wlodarska-Kowalcuk, M., & Ellingsen, K. E. (2009). Beta diversity and taxonomic sufficiency: Do higher-level taxa reflect heterogeneity in species composition? *Diversity and Distributions*, 15, 450-458.

Thrush, S. F., & Dayton, P. K. (2002). Disturbance to marine benthic habitats by trawling and dredging: implications for marine biodiversity. *Annual Review of Ecology and Systematics*, 33, 449-473.

Thrush, S. F., Ellingsen, K. E., & Davis, K. (2015). Implications of fisheries impacts to seafloor biodiversity and Ecosystem-Based Management. *ICES Journal of Marine Science*, 73 (Supplement 1), i44-i50, doi:10.1093/icesjms/fsv114.

685 Yoccoz, N. G., Nichols, J. D., & Boulinier, T. (2001). Monitoring of biological diversity in space and
686 time. *TRENDS in Ecology & Evolution,* 16, 446-453.
687

688 **Figure legends**

689

690 **Fig. 1.** Overview and location of 11 regional stations (red circles) and the petroleum installations
691 (stars) in Region I (Ekofisk) on the southern part of the Norwegian continental shelf, in the North Sea.

692

693 **Fig. 2.** (a) Chromium (Cr) in sediment samples before correction. Note that values in 2011 are
694 anomalous for stations 9, 11, 12 and 6. (b) Cr, where anomalous values in 2011 were replaced by a
695 prediction from a station-specific linear model having year as a predictor to take into account the
696 trends observed in each station, and excluding the anomalous 2011 observations.

697

698 **Fig. 3**. Zinc (Zn) in sediment samples. Note that there is a trend in the data prior to 2011 at several
699 stations, but that the values in 2011 are lower. There is only one replicate in 2011.

700

701 **Fig. 4.** Cadmium (Cd) in sediment samples. Note that two stations (station 4 and 5) have values that
702 are too high in 2005.

703

704 **Fig. 5.** Median grain size (Md) of sediment samples from all regional stations from all years of
705 sampling. Note that the values from 1996 are clearly higher than for the rest of the period.

706

707 **Fig. 6.** Principal component analysis (PCA) of the environmental variables total organic matter
708 (TOM), silt-clay content (pelite; fraction of sediment < 0.063 mm), total hydrocarbons (THC),
709 polycyclic aromatic hydrocarbons (PAH), barium (Ba), lead (Pb) and chromium (Cr) for the regional
710 stations. Top row: without Cr, including 1996. Left: Ellipses summarizing the distribution of
711 individual samples in different years (1996-2011), 1996 appearing as an outlier; Right: Correlation

712 circle of environmental variables with the two first PCA axes; the first PCA axis is positively
713 correlated with all variables, the second axis is positively correlated with TOM and silt-clay content
714 and negatively with THC. Bottom left: distribution of individual samples for a PCA with Cr but
715 excluding 1996. Bottom right: Distribution of averaged standardized values by years, showing the
716 high values for 1996 and the smaller variability compared to other years.

717

718 **Fig. 7.** Non-metric multidimensional scaling (NMDS) based on Bray-Curtis distance (using square
719 root transformed abundance data) and unadjusted faunal data, i.e. 388 taxa (d = 0.5) (left), and
720 adjusted faunal data with regard to uncertain taxonomic classifications, i.e., 294 taxa (modification
721 alternative 1, based on "lumping") (d = 0.2) (right). For the procedure on adjusting the faunal data see
722 text 'A new procedure to handle different taxonomic resolution'.

723

724 **Fig. 8**. Canonical Correspondence Analysis (CCA) of faunal data, left on unadjusted data, i.e. 388
725 taxa, right on adjusted data, i.e., 294 taxa (modification alternative 1, based on "lumping"). Top is for
726 axes 1 and 2, bottom is for axes 1 and 3. Only year as a categorical variable was used as a predictor.
727 For the procedure on adjusting the faunal data see text 'A new procedure to handle different
728 taxonomic resolution'.

729

730

Table 1. Regional stations in the study area (Region I), with information on water depth (m), geographical position and sediment characteristics. Latitude and longitude are in decimal degrees. Sediment variables (range at station over time): THC: total hydrocarbons (mg/kg, 0-1 cm); Ba: barium (mg/kg, 0-1 cm); TOM: total organic matter (%); and sand (%). For the years 1996, 2008 and 2011 values of "fine sand" is given in the MOD-database; whereas for the years 1999, 2002 and 2005 values of "total sand" is given. The variable "sand" is a mixture of these, and is therefore not used in further analyses. For TOM, one outlier is excluded (for station 8 in 1996, replicate number 3).

Regional station	Depth	Latitude (°N)	Longitude (°E)	THC	Ba	Sand	TOM
1	71	57.15	2.77	1.20-5.17	18.0-101.0	82.12-96.53	0.71-1.09
2	65	56.92	3.33	1.20-4.77	11.0-31.0	86.16-97.70	0.81-1.18
3	67	56.55	3.46	2.60-9.55	26.0-94.0	87.83-97.40	0.80-0.99
4	68	56.25	3.83	2.40-7.12	32.0-121.0	83.39-97.06	0.82-1.05
5	69	57.00	2.50	2.31-7.16	19.0-76.2	78.99-95.61	0.72-1.20
6	70	56.75	2.67	1.93-8.25	26.0-42.6	84.72-96.35	0.82-1.17
7	72	56.50	2.75	2.50-5.57	22.0-88.0	84.45-96.70	0.77-1.23
8	71	56.04	3.46	1.60-5.82	20.0-77.0	87.10-99.00	0.68-1.03
9	66	57.12	3.18	1.37-2.95	5.0-17.3	85.00-98.64	0.54-0.79
11	67	56.24	3.16	1.30-9.54	21.0-98.4	86.02-98.20	0.70-0.92
12	65	56.96	2.99	0.60-9.52	18.0-105.0	84.46-99.70	0.66-1.13

Table 2. Estimates of linear yearly trends of environmental variables (with se in parenthesis) obtained with a linear mixed model with station and station by year random effects. Statistically linear trends (at the 0.05 level) are indicated in bold/italics. TOM: total organic matter (%); Ba: barium; PAH: polycyclic aromatic hydrocarbons; THC: total hydrocarbons.

Variables	**With 1996**	**Without 1996**
TOM	***-0.0069 (0.0022)***	***-0.0090 (0.0029)***
Ba	***-0.67 (0.32)***	0.43 (0.35)
PAH (x1000)	***-0.72 (0.29)***	0.0 (0.25)
THC	***-0.098 (0.026)***	-0.032 (0.032)

Data accessibility: Faunal data and a description of the procedure of lumping/splitting taxa in the species list prior to data analyses (including the R script) will be made available from the Dryad Digital Repository after publication.

Supplementary Information

Additional Supporting Information may be found in the online version of this article.

Online Resource 1.

Table S1. Consulting companies responsible for fieldwork, identification of taxa, and laboratory analyses.

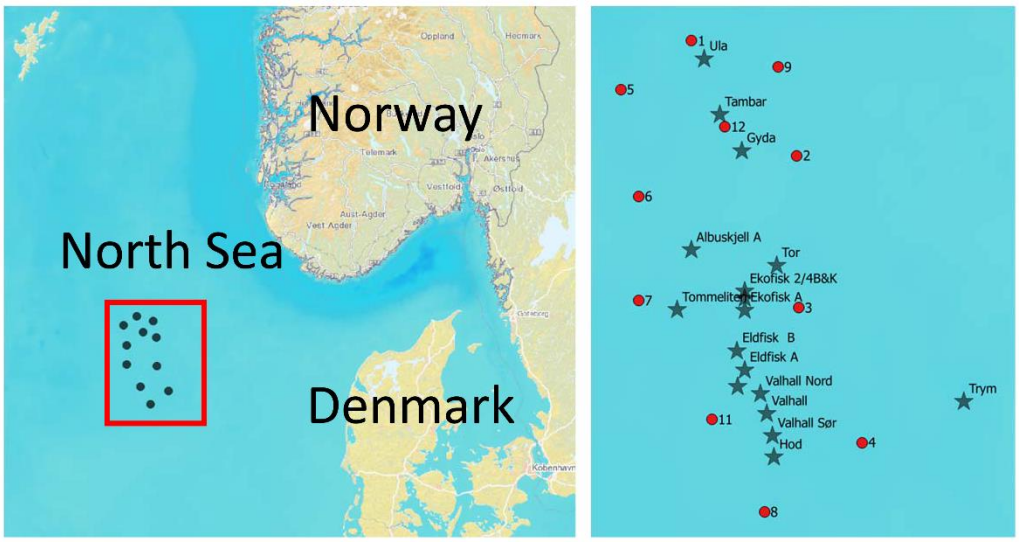
Online Resource 2.

Fig. S1. Non-metric multidimensional scaling (NMDS) using faunal data based on the splitting procedure.

Fig. S2. Canonical Correspondence Analysis (CCA) using faunal data based on the splitting procedure.

766 **Figure 1.**

767



768

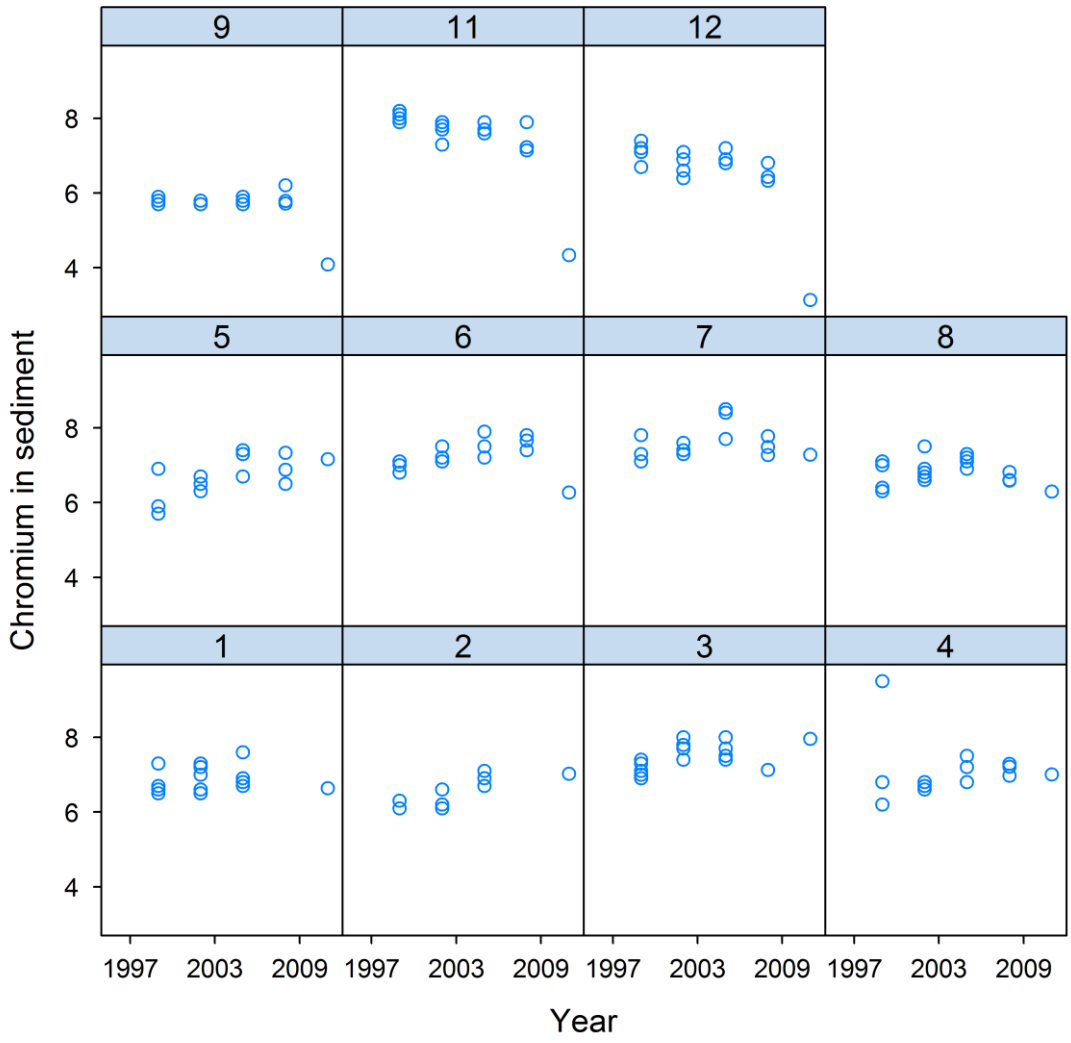
769 **Fig. 1.** Overview and location of 11 regional stations (red circles) and the petroleum installations

770 (stars) in Region I (Ekofisk) on the southern part of the Norwegian continental shelf, in the North Sea.

771

772 **Figure 2a.**

773



774

775

776

Figure 2b.

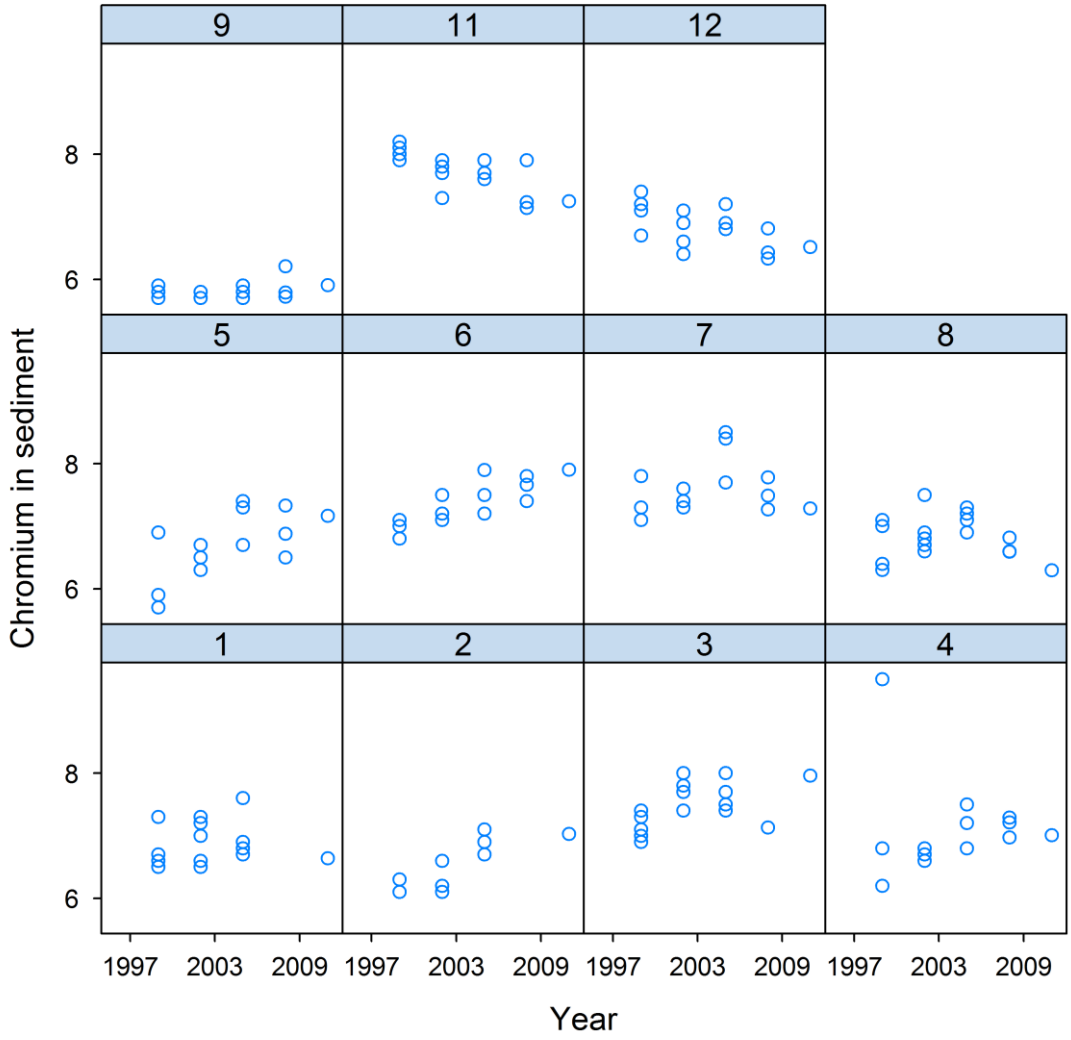
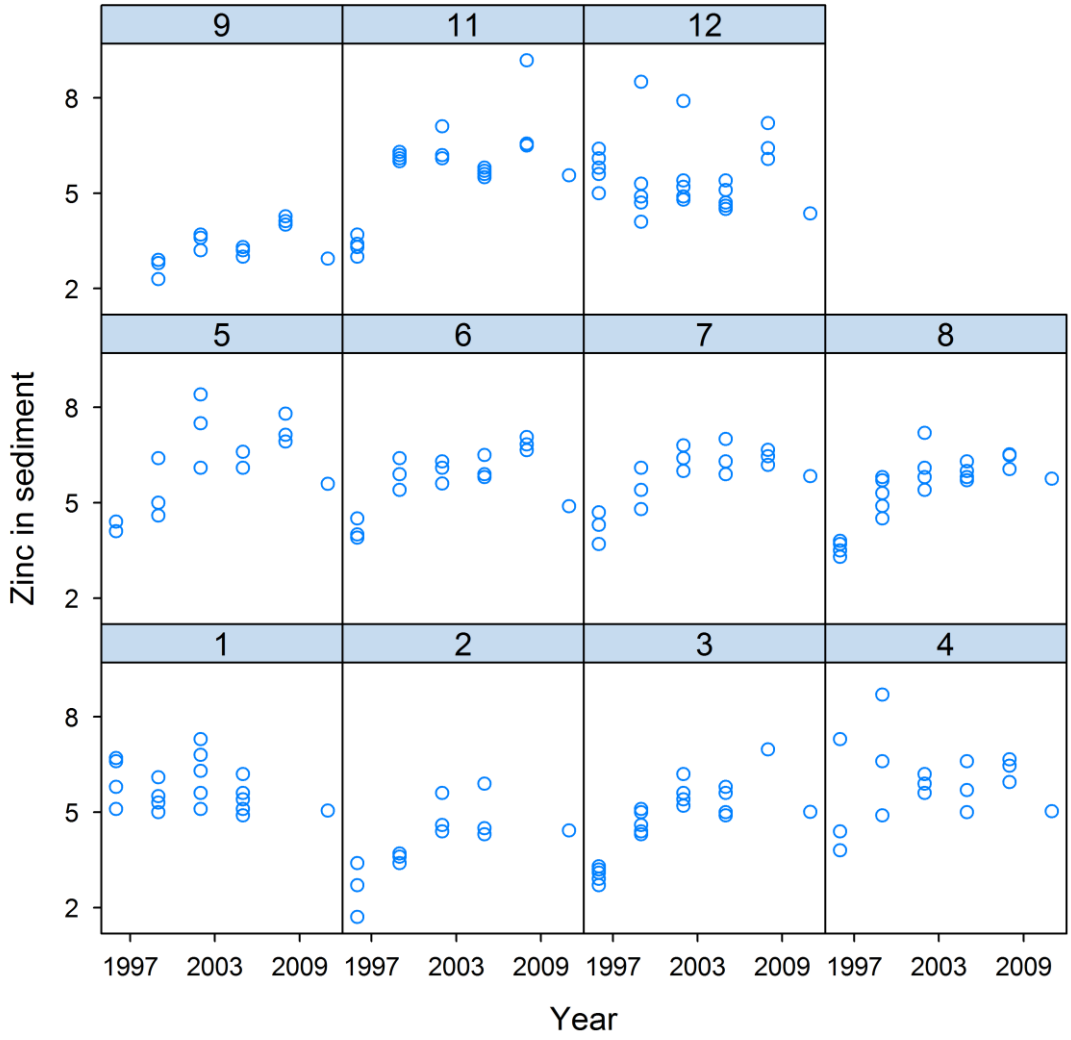


Fig. 2. (a) Chromium (Cr) in sediment samples before correction. Note that values in 2011 are anomalous for stations 9, 11, 12 and 6. (b) Cr, where anomalous values in 2011 were replaced by a prediction from a station-specific linear model having year as a predictor to take into account the trends observed in each station, and excluding the anomalous 2011 observations.

786 **Figure 3.**



787

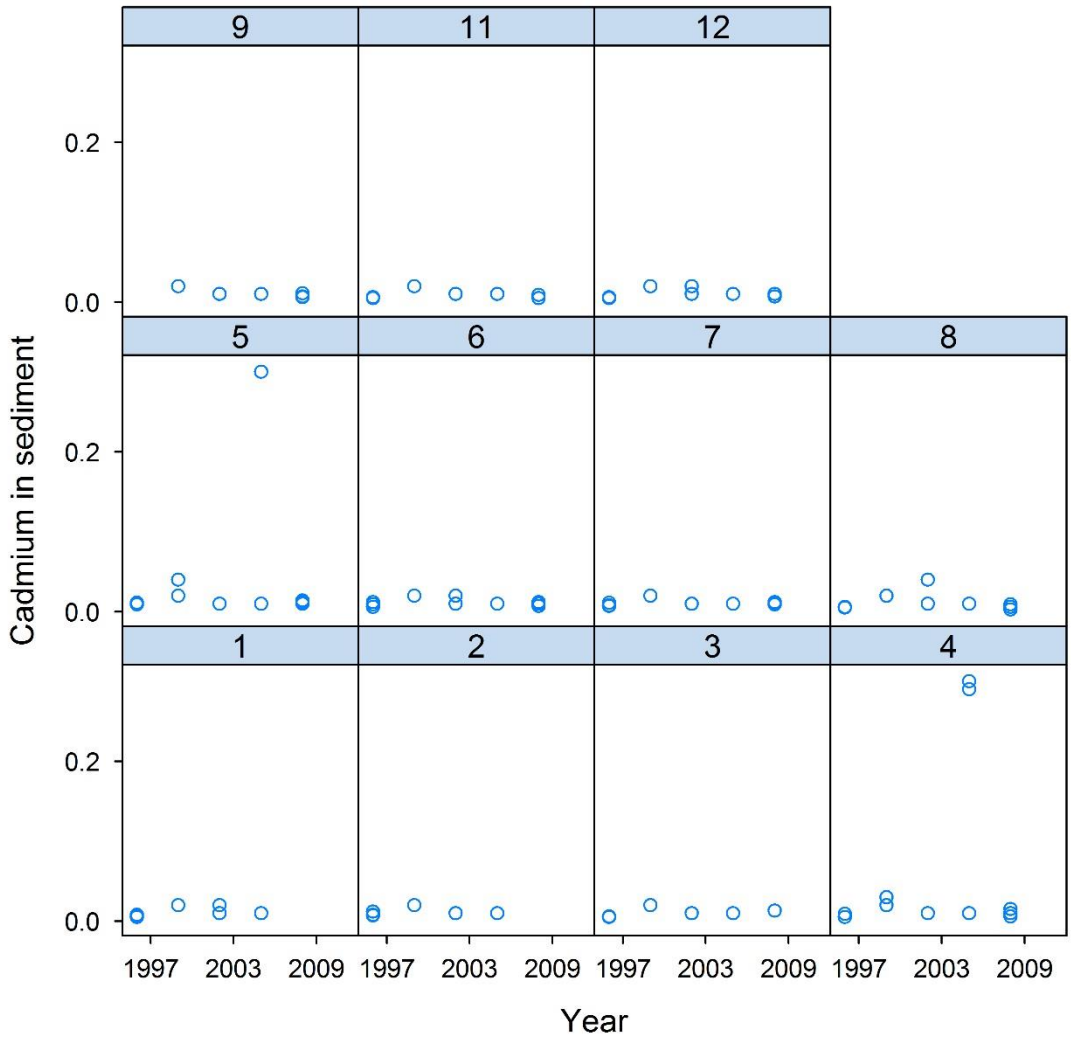
788 **Fig. 3.** Zinc (Zn) in sediment samples. Note that there is a trend in the data prior to 2011 at several
789 stations, but that the values in 2011 are lower. There is only one replicate in 2011.

790

791

792 **Figure 4.**

793



794

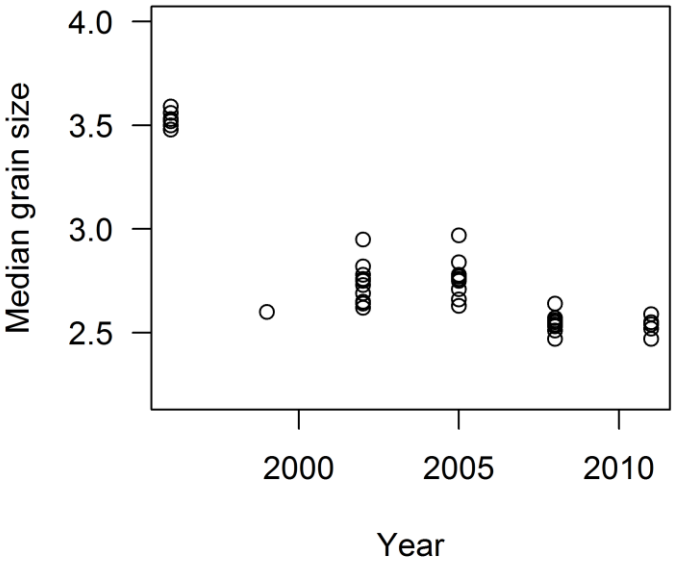
795

796 **Fig. 4.** Cadmium (Cd) in sediment samples. Note that two stations (station 4 and 5) have values that

797 are too high in 2005.

798

799 **Figure 5.**



800

801 **Fig. 5.** Median grain size (Md) of sediment samples from all regional stations from all years of
802 sampling. Note that the values from 1996 are clearly higher than for the rest of the period.

803

Figure 6.

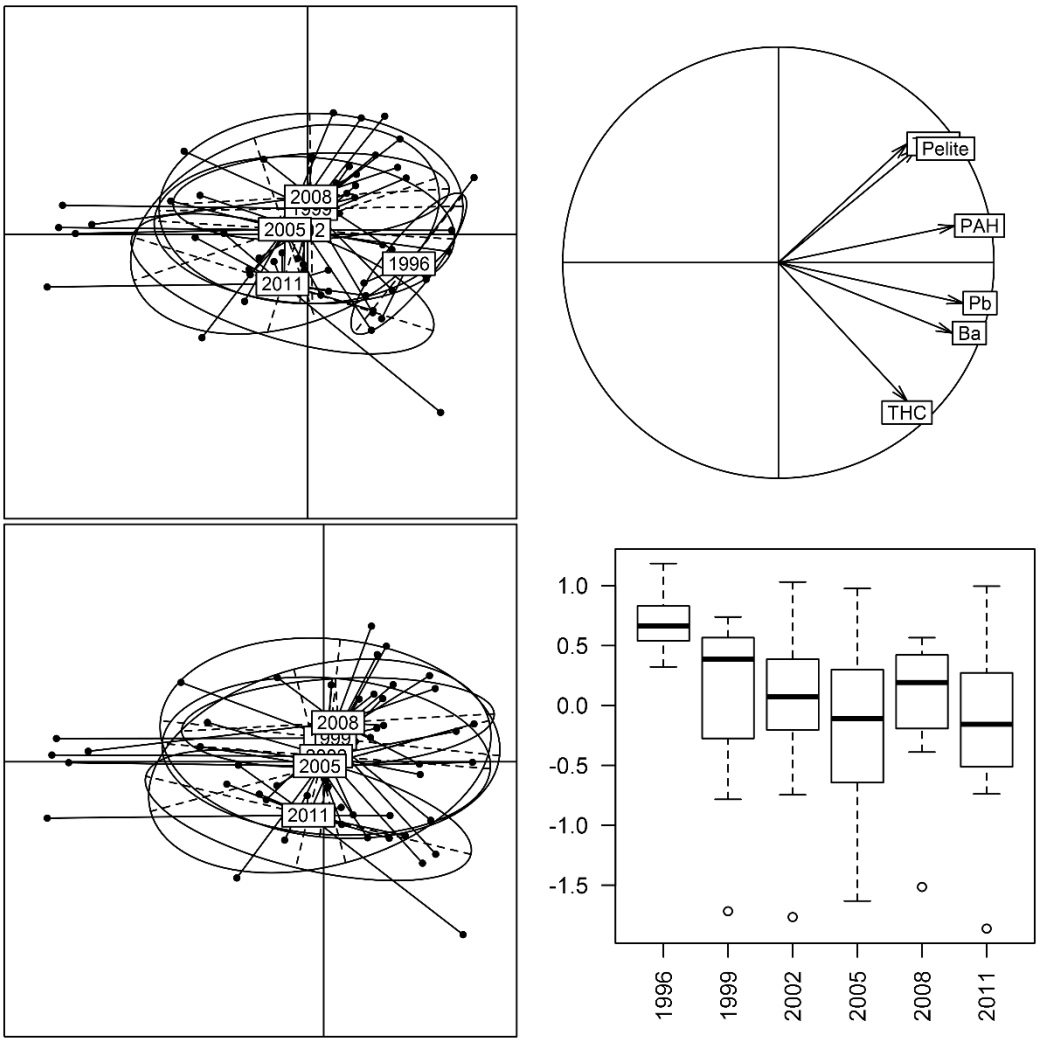
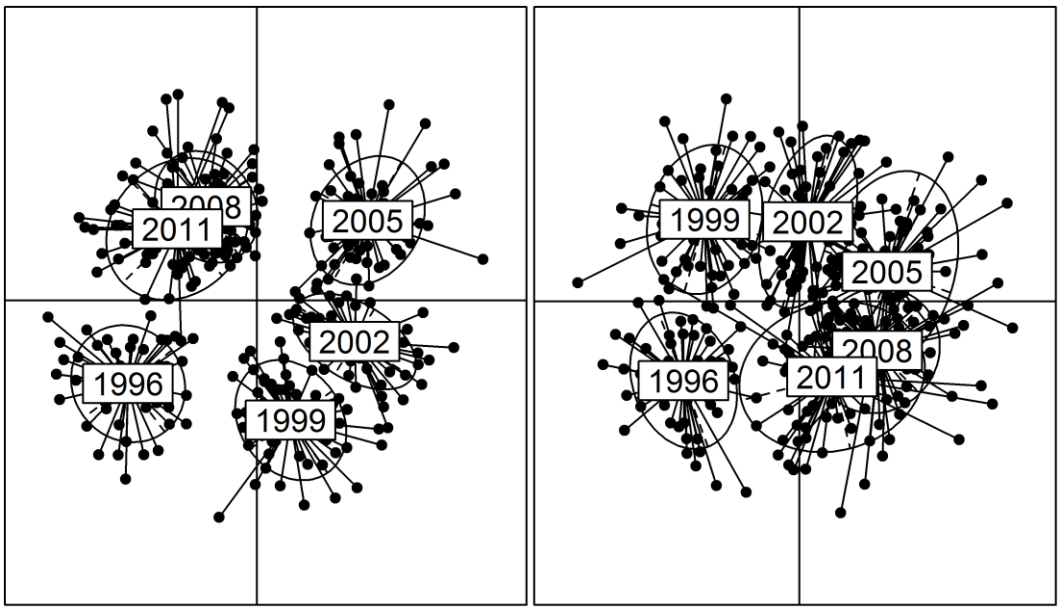


Figure 6. Principal component analysis (PCA) of the environmental variables total organic matter (TOM), silt-clay content (pelite; fraction of sediment < 0.063 mm), total hydrocarbons (THC), polycyclic aromatic hydrocarbons (PAH), barium (Ba), lead (Pb) and chromium (Cr) for the regional stations. Top row: without Cr, including 1996. Left: Ellipses summarizing the distribution of individual samples in different years (1996-2011), 1996 appearing as an outlier; Right: Correlation circle of environmental variables with the two first PCA axes; the first PCA axis is positively correlated with all variables, the second axis is positively correlated with TOM and silt-clay content and negatively with THC. Bottom left: distribution of individual samples for a PCA with Cr but excluding 1996. Bottom right: Distribution of averaged standardized values by years, showing the high values for 1996 and the smaller variability compared to other years.

820 **Figure 7.**

821



822

823

824 **Fig. 7.** Non-metric multidimensional scaling (NMDS) based on Bray-Curtis distance (using square
825 root transformed abundance data) and unadjusted faunal data, i.e. 388 taxa (d = 0.5) (left), and
826 adjusted faunal data with regard to uncertain taxonomic classifications, i.e., 294 taxa (modification
827 alternative 1, based on "lumping") (d = 0.2) (right). For the procedure on adjusting the faunal data see
828 text 'A new procedure to handle different taxonomic resolution'.

829

Figure 8.

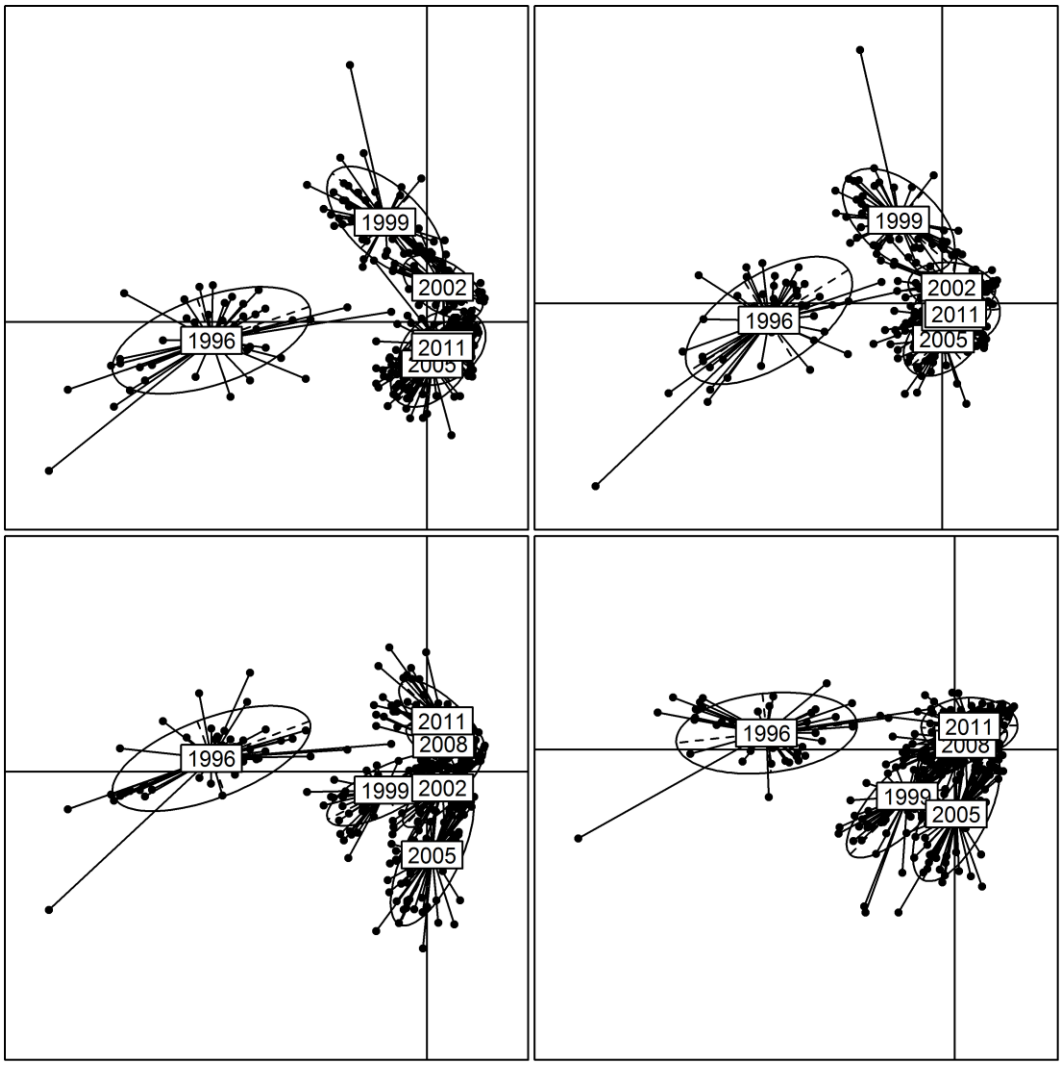


Fig. 8. Canonical Correspondence Analysis (CCA) of faunal data, left on unadjusted data, i.e. 388 taxa, right on adjusted data, i.e., 294 taxa (modification alternative 1, based on "lumping"). Top is for axes 1 and 2, bottom is for axes 1 and 3. Only year as a categorical variable was used as a predictor. For the procedure on adjusting the faunal data see text 'A new procedure to handle different taxonomic resolution'.

Supplementary information. Table.

Article title: Long-term environmental monitoring for assessment of change: measurement inconsistencies over time and potential solutions

Journal name: Environmental Monitoring and Assessment

Author names: Kari E. Ellingsen[1*], Nigel G. Yoccoz[1,2], Torkild Tveraa[1], Judi E. Hewitt[3] and Simon F. Thrush[4]

[1] Norwegian Institute for Nature Research (NINA), Fram Centre, P.O. Box 6606 Langnes, 9296 Tromsø, Norway

[2] Department of Arctic and Marine Biology, UiT The Arctic University of Norway, 9037 Tromsø, Norway

[3] National Institute of Water and Atmospheric Research, NZ

[4] Institute of Marine Sciences, University of Auckland, NZ

[*] Corresponding author: E-mail: kari.ellingsen@nina.no

Online Resource 1.

Table S1. Consulting companies (identified by letters) responsible for fieldwork during each sampling occasion, identification of taxa, and analyses of organic chemistry, metals and physical properties of the sediment samples. For further information, see survey reports written by consulting companies (Cochrane et al. 2009; Jensen et al. 2000; Mannvik et al. 1997; Mannvik et al. 2012; Nøland et al. 2003; Nøland et al.2006).

Year	Field work	Taxa identification	Analyses of sediment in laboratory		
		Main responsible	Organic chemistry	Metals	Physical properties
1996	A, C	A	C	G	F
1999	B, D	B	D	D	D
2002	B, D	B	D	D	D
2005	B, E	B	E	E	E
2008	A, C	A	C	G	C
2011	A, C	A	C	H	C

References

Cochrane, S., Palerud, R., Wasbotten, I. H., Larsen, L. H., & Mannvik, H. P. (2009). Offshore sediment survey of Region I, 2008. Akvaplan-niva report no. 4215 - 02. Akvaplan-niva, Tromsø, Norway. 314 pp.

Jensen, T., Gjøs, N., Nøland, S.-A., Oreld, F., Møskeland, T., Bakke, S. M., et al., (2000). Environmental Monitoring 1999, Region I – Ekofisk. Technical Report. Report no. 2000-3238. Det Norske Veritas & Sintef Applied Chemistry, Norway. 294 pp.

Mannvik, H. P., Pearson, T., Pettersen, A., & Lie Gabrielsen, K. (1997). Environmental monitoring survey Region I 1996. Main Report. Akvaplan-niva report no. 411.96.996-1. Akvaplan-Niva, Tromsø, Norway. 246 pp.

Mannvik, H. P., Wasbotten, I. H., Cochrane, S., & Moldes-Anaya, A. (2012). Miljøundersøkelse Region I, 2011. Akvaplan-niva report no. 5339.02. Akvaplan-niva, Tromsø, Norway. 196 pp. (In Norwegian).

888 Nøland, S. A., Gjøs, N., Bakke, S. M., & Oreld F. (2003). Environmental Monitoring 2002, Region I –
889 Ekofisk. Main report. Technical Report. Report no. 2003-0338. Det Norske Veritas/Sintef, Norway.
890 316 pp.
891
892 Nøland, S. A., Bakke, S. M., Rustad, I., & Brinchmann, K. M. (2006). Environmental Monitoring
893 Region I, 2005. Main Report. Report no. 2006-0187. Det Norske Veritas, Norway. 344 pp.
894
895

Supplementary information. Figures.

Article title: Long-term environmental monitoring for assessment of change: measurement inconsistencies over time and potential solutions

Journal name: Environmental Monitoring and Assessment

Author names: Kari E. Ellingsen[1*], Nigel G. Yoccoz[1,2], Torkild Tveraa[1], Judi E. Hewitt[3] and Simon F. Thrush[4]

[1] Norwegian Institute for Nature Research (NINA), Fram Centre, P.O. Box 6606 Langnes, 9296 Tromsø, Norway

[2] Department of Arctic and Marine Biology, UiT The Arctic University of Norway, 9037 Tromsø, Norway

[3] National Institute of Water and Atmospheric Research, NZ

[4] Institute of Marine Sciences, University of Auckland, NZ

[*] Corresponding author: E-mail: kari.ellingsen@nina.no

Online Resource 2.

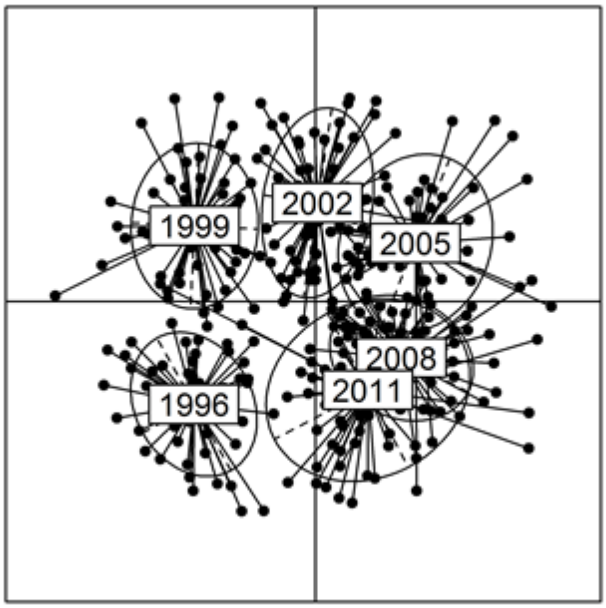


Fig. S1. Non-metric multidimensional scaling (NMDS) based on Bray-Curtis distance (using square root transformed abundance data) and adjusted faunal data with regard to uncertain taxonomic classifications, i.e., using 314 taxa (d = 0.2) (i.e. modification alternative 2, based on splitting; for the procedure on adjusting the faunal data see text 'A new procedure to handle different taxonomic resolution').

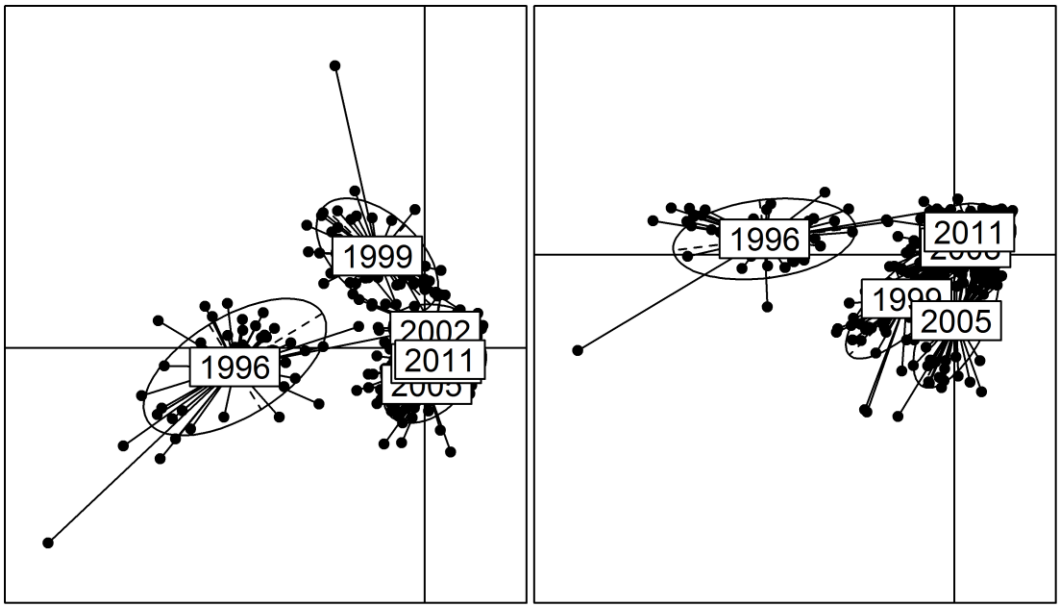


Fig. S2. Canonical Correspondence Analysis (CCA) of faunal data, using adjusted data, i.e., 314 taxa (i.e. modification alternative 2, based on splitting). Left is for axes 1 and 2, right is for axes 1 and 3. Only year as a categorical variable was used as a predictor. For the procedure on adjusting the faunal data see text 'A new procedure to handle different taxonomic resolution'.