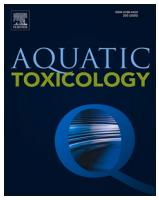
Aquatic Toxicology 286 (2025) 107480

Contents lists available at ScienceDirect

Aquatic Toxicology

journal homepage: www.elsevier.com/locate/aqtox



Compilation and standardization of oil toxicity data on early life stages of fish to support population-level oil spill impact modeling

Sami Vikkula [a,1,2,*], Samu Mäntyniemi [b], Sakari Kuikka [a]

[a] University of Helsinki, Fisheries and Environmental Management Group, Ecosystems and Environment Research Programme, Faculty of Biological and Environmental Sciences, P.O Box 65, Viikinkaari 1, P, Helsinki, FI-00014, Finland
[b] Natural Resources Institute Finland, Latokartanonkaari 9, Helsinki, FI-00790, Finland

ARTICLE INFO

Dataset link: Navigating Data Diversity Obstacles in Assessing Oil Spill Impacts on Fish Early Life Stages (Original data)

Keywords:
Oil spills
Fisheries
Impact assessment
Toxicity model
Data heterogeneity
Polycyclic aromatic hydrocarbons

ABSTRACT

The risk of oil spills has increased in recent years due to rising tanker traffic and the emergence of poorly maintained vessels. While the environmental impacts of oil spills are wide-ranging, their effects on fish populations remain contested, prompting the need for further research and the development of improved tools and methodologies for impact assessment. One complicating factor is the heterogeneity of existing laboratory exposure data, which hinders their usage in population-level oil spill impact assessment. This study addresses that gap by compiling a dataset from peer-reviewed laboratory exposure studies on the early life stages (ELS) of fish exposed to oil. Through a systematic literature review, we identified relevant studies and developed novel standardization methods to improve data comparability. These methods included converting diverse exposure metrics to a baseline polycyclic aromatic hydrocarbon (PAH) concentration metric, calculating the geometric mean and time-weighted average concentrations, and modeling missing control concentrations. The resulting dataset encompasses multiple fish species and oil types, as well as wide exposure time and PAH concentration ranges, in order to support impact assessments across diverse spill scenarios in fish population dynamics models. Although our methodology significantly increased the amount of usable data, the species and oil type coverage remained uneven, requiring model structures that accommodate information borrowing. We provide recommendations to improve future experimental reporting and suggest methodological extensions for the standardization methods. This study demonstrates how structured data compilation and standardization can cost-effectively expand the applicability of existing experimental data for oil spill impact assessment.

1. Introduction

The risk of oil spills in aquatic environments remains a critical concern as tanker traffic continues to increase. Since 2020, 37 spills, each involving at least seven metric tons of oil, have collectively released 38,000 t of oil into marine ecosystems. (ITOPF, 2025) The recent rise in the number of old, uninsured, and poorly maintained tankers has further heightened this risk (Caprile and Leclerc, 2024). The environmental consequences of oil spills depend on several factors, including the volume and type of oil released, environmental conditions, and the structure of the affected ecosystem (Barron et al., 2020). Due to the economic importance of fisheries, the effects on fish populations have received considerable attention in oil spill impact assessment (e.g., Carroll et al., 2018; Spromberg et al., 2024; Sumaila et al., 2012; Vikebø et al., 2025).

Early life stages (ELS) of fish are particularly vulnerable to oil due to their small size, high lipid content, and low metabolic rate of oil toxins (Petersen and Kristensen, 1998; Sørhus et al., 2021). The toxicity of oil has been strongly associated with polycyclic aromatic hydrocarbons (PAHs) (e.g., Adams et al., 2014; Hodson, 2017). In ELS, PAHs can interfere with critical developmental processes, causing a range of lethal and sublethal effects (Incardona et al., 2023, 2021; Sørhus et al., 2016). As a result, laboratory oil exposure studies on ELS have largely focused on the effects of PAHs (e.g., Bérubé et al., 2022; Bonatesta et al., 2022; Carls et al., 1999; Jones et al., 2020; Schiano Di Lombo et al., 2021; Simning et al., 2019; Wu et al., 2012).

* Corresponding author.
E-mail address: sami.vikkula@helsinki.fi (S. Vikkula).
[1] Web of Science: https://www.webofscience.com/wos/author/record/KRP-9761-2024
[2] University of Helsinki: https://researchportal.helsinki.fi/en/persons/sami-juhani-vikkula

https://doi.org/10.1016/j.aquatox.2025.107480
Received 11 March 2025; Received in revised form 1 July 2025; Accepted 1 July 2025
Available online 2 July 2025

Contradictory findings on the population-level impacts of oil spills via ELS mortality persist. While some studies link ELS exposure and associated cardiotoxic effects to population declines (Incardona et al., 2015, 2014, 2013), others attribute declines primarily to juvenile and adult mortality (Ainsworth et al., 2018) or report that oil-induced ELS mortality did not impair population reproductive capacity (Carroll et al., 2018). These contradictions highlight the need for methodologies that reduce uncertainty.

One approach to assessing population-level impacts through oil-induced ELS mortality is using a toxicity model to combine data from multiple laboratory exposure studies into a summary effect and use that in a population dynamics model to evaluate the impacts of different spill scenarios (Ainsworth et al., 2018; Carroll et al., 2022; Dornberger et al., 2016; Klok et al., 2014). This approach is often hindered by data heterogeneity (Hodson et al., 2019). A major source of heterogeneity is the use of varying exposure metrics, including nominal loading, total petroleum hydrocarbon (TPH) concentration, and a multitude of different PAH metrics (Bejarano et al., 2023; Redman and Parkerton, 2015).

Additional sources of heterogeneity include the use of different exposure regimes and exposure concentration measurements (Bejarano et al., 2014; Hodson et al., 2019). For example, spiked regime studies, where oil concentrations decline over time, frequently report exposure using initial concentrations, which may underestimate PAH toxicity and render such studies difficult to compare with continuous concentration regimes. Moreover, it has been suggested that toxicity models should rely exclusively on data from spiked regime studies with short exposure times, as these are considered most representative of typical oil spill conditions in the field (Bejarano et al., 2014). However, this approach fails to account for spills in confined environments or those caused by ruptured pipelines or oil well blowouts, where elevated PAH concentrations can persist for extended periods (Hodson et al., 2019).

The absence of standardized protocols for conducting and reporting oil exposure studies contributes to variability in the level of detail provided across studies (Adams et al., 2017). Studies have frequently failed to report background PAH concentrations in control treatments, and the reported exposure treatment concentrations include both oil and background PAHs (Bejarano et al., 2023). The sources of the water used in exposure studies vary from natural sources (for instance, seawater) to tap water, and background PAH concentrations are found in both (WHO, 2022). Fish population dynamics models incorporate parameters for the natural mortality of ELS which is mortality caused by typical environmental conditions, including background PAHs. Therefore, when integrating toxicity model estimates for oil spill impact assessment, it is essential that the estimates reflect only the effects of oil-sourced PAH compounds.

Thus, the objectives of this article are: (1) To compile a dataset from peer-reviewed laboratory exposure studies investigating ELS exposed to crude oil or oil products, with a specific focus on mortality responses to oil-sourced PAH compounds. To achieve this, we conduct a systematic review to identify data across multiple species and oil types, encompassing a range of exposure times and PAH concentrations, in order to support impact assessments across diverse spill scenarios in fish population dynamics models. (2) To introduce novel standardization methods to improve data comparability in the context of oil toxicity to ELS. These include the standardization of exposure metrics and concentration measurements, as well as the modeling of missing control PAH concentration measurements. Collectively, these methods serve as preliminary examples of how data availability can be expanded without the need for costly new laboratory exposure studies.

2. Materials and methods

2.1. Systematic review and data extraction

Based on the intended use of the dataset and the requirements of our standardization methods, we established a set of inclusion criteria for article selection. Details of these criteria are provided in Appendix A Section 1.1. In brief, only peer-reviewed articles reporting primary data were eligible for inclusion. We did not restrict our searches by publication date or geographic location. We conducted searches in two rounds. In the first round, we included articles on any species. In the second round, we focused on species for which the first round yielded the highest number of results. We included only those studies in which ELS were exposed to crude oil or refined oil products, excluding studies that involved exposure to individual compounds or compound groups. No restrictions were applied regarding exposure levels or exposure times.

We derived search term combinations (Table A1) based on the inclusion criteria and a couple of scoping searches. First round of searches was done in Helka (University of Helsinki Library, 2025), which is a centralized article database and search engine of University of Helsinki. It has access to most international journal databases such as Scopus, ProQuest, ScienceDirect, and Springer. After we had gone through the Helka results, we conducted complementary searches in Google scholar for the selected species. We also found possible articles for inclusion from the references of the search results. For summation of our results, we decided to group herring, salmon, and cod subspecies into overarching groups, but we also recorded the actual names of the species.

We applied our inclusion criteria on the search results from both rounds in three phases. First, we screened only abstracts and titles and included articles if the title and/or abstract indicated that the article is a laboratory exposure study of any fish life stage exposed to oil or oil compounds. Next, we read the full texts and excluded articles based on all other criteria except exposure metrics. We finally made those exclusions when we standardized exposure metrics (see Section 2.2). We imported all articles in Zotero (Takats et al., 2024).

We extracted data that reflected our inclusion criteria (Table A2). To ensure that exposure concentrations in the data reflected only oil-sourced PAHs, we calculated the additional PAH concentration-covariate in the dataset by subtracting the control treatment PAH concentration from the standardized exposure concentrations of the same experiment. Some of the extracted data was not part of our inclusion criteria but could prove useful for some models (Table A3). As this data was not part of our inclusion criteria, some of the articles had missing values. We extracted data from text, graphs, and tables in both the main texts and the supporting information. All extracted information was double-checked and recorded in our own templates (Appendix B) and supporting information (Table A5).

2.2. Standardization of exposure metrics

To improve data comparability and increase the volume of usable data, we decided to convert the original PAH concentration metrics of some of the articles to a common reference metric, hereafter referred to as the baseline. The purpose of this was to ensure that metrics across studies contain the same set of compounds, making them comparable. We selected the PAH metric used by Short et al. (1996), as described in Short and Harris (1996), as our baseline. This metric was most frequently used in our dataset (see Tables A4 and A5). Moreover, we considered it sufficiently comprehensive in terms of compound coverage, as it includes not only the 16 U.S. Environmental Protection Agency (EPA) priority PAHs but also a range of alkylated PAHs known to be relevant for toxicity assessment (Andersson and Achten, 2015).

If an article reported a PAH metric that had compounds not included in the baseline metric, we aimed to exclude those compounds from the reported concentrations. When the article provided a breakdown of individual PAH compound concentrations in the exposure solutions, we excluded the non-baseline compounds. Ideally, such breakdowns were provided for each individual experiment; however, if a breakdown was reported for only one experiment, we applied it to all experiments. In cases, where no breakdown was available, we evaluated whether the non-baseline compounds were likely to be present in notable

concentrations (Table A5). This assessment was based on the oil type, its degree of weathering, and the method used to dissolve PAHs into the exposure solution. To support this evaluation, we extracted information from the articles on dissolution methods and weathering procedures, and we consulted the American Petroleum Institute (API) gravity values of the oils (Table A5) and their corresponding weight classifications according to Wang et al. (2003).

The method used to dissolve oil into the exposure medium influences both the types and proportions of PAH compounds present in the solution. One approach, known as passive dosing, involves mixing oil with gravel and placing it in a flow-through chamber, where PAHs dissolve into water passing through the oiled gravel (Carls et al., 1999; Heintz et al., 1999). Another common method involves mixing oil and water using varying levels of mechanical energy, resulting in high-energy water-accommodated fractions (HEWAF) or low-energy water--accommodated fractions (LEWAF), or incorporating a chemical dispersant to create chemically enhanced water accommodated fractions (CEWAF) (Bérubé et al., 2022; Jones et al., 2020; Wu et al., 2012). The solubility of PAHs generally increases as molecular weight and the number of aromatic rings decrease (Adams et al., 2014). HEWAFs contain higher overall PAH concentrations, and a greater proportion of heavier PAHs compared to LEWAFs, and CEWAFs tend to reflect the PAH composition of whole oil (Gardiner et al., 2013).

Both oil type and the degree of weathering influence the composition of PAH compounds in oil (Wang et al., 2003; Wu et al., 2012). For example, heavy oils typically contain higher concentrations of PAHs overall and are enriched in heavier PAH compounds. In contrast, light oils are dominated by lighter PAHs. As oil undergoes weathering, the relative abundance of lighter compounds decreases. This is due to their higher volatility, which makes them more prone to both dissolution and evaporation.

When evaluating whether non-baseline compounds in a non-baseline PAH metric were likely to be present in notable concentrations, we first identified which non-baseline compounds the metric included. We then assessed whether the oil type, its degree of weathering, and dissolution method supported the presence of those compounds. For example, if the analytical method associated with a metric measured heavier non-baseline PAH compounds, and (1) the oil type in the article was heavy, (2) the dissolving method was HEWAF or CEWAF, and/or (3) the oil was weathered, we assumed that the solution likely contained notable concentrations of those compounds. Conversely, if (1) the oil type was light, (2) the dissolving method was LEWAF, and/or (3) the oil was not weathered, we assumed that the likelihood of these heavy compounds being present was low.

Similarly, for non-baseline compounds consisting of lighter PAHs, if (1) the oil type was light, (2) the dissolution method was LEWAF, and/or (3) the oil was not weathered, we assumed a higher probability of notable concentrations. On the other hand, if (1) the oil type was heavy, (2) the dissolution method was HEWAF or CEWAF, and/or (3) the oil was weathered, we considered the probability of significant concentrations of those light compounds to be low.

If we had reasons to believe that non-baseline compounds were present in notable concentrations, we attempted to identify another article that reported a breakdown of individual PAH compound concentrations for a solution prepared using the same oil type (similarly weathered) and a comparable dissolution method. In this context, we categorized oils dichotomously as either weathered or un-weathered. If no suitable breakdown was found, we excluded the article from the dataset.

We also removed non-baseline compounds from the reported background PAH concentrations in the control treatments. If an article provided a breakdown of compound concentrations for the controls, we eliminated the non-baseline compounds based on that and if not, we applied a correction using a proportion equal to half of the smallest proportion previously used to eliminate compounds from the exposure experiments of the same article.

Some articles employed PAH metrics that included significantly fewer compounds than the baseline. Due to resource limitations and this article being purposed to provide a preliminary exploration of this methodology, we did not attempt to estimate or add the missing baseline compounds. This task was left to the future applications of the model.

We also sought to convert articles that used TPH as an exposure metric. To do this, we identified articles that reported both TPH and PAH concentrations for a solution prepared with the same oil type (similarly weathered) and using similar dissolution method, and used these to calculate a PAH/TPH ratio. Additionally, the articles had to provide a sufficient breakdown of the measured PAH compounds so that we could exclude non-baseline compounds. If no suitable article was found, the article was excluded from the dataset.

One article in the dataset reported both TPH and PAH concentrations from a separate WAF sample prepared using one of the oils tested in their study (Table A5, Study ID 3.2.5). They fitted a regression model to predict TPH concentrations from PAH concentrations. We inverted their regression equation to predict PAH concentrations from the reported TPH values. The article also provided a breakdown of PAH compounds, which we used to remove non-baseline compounds. For the remaining experiments, involving other oil types, we used other articles to obtain PAH/TPH ratios.

2.3. Standardization of concentration measurements

The articles we had in our dataset differed in respect of exposure regimes and time points when mortalities were measured (Fig. 1). Exposure regimes were either spiked or continuous concentration regimes. Spiked regime can have one or multiple spikes. Multiple spikes means that the concentration is renewed once or more during the experiment. Upon renewal, the aim is to reset the concentration back to the initial concentration and the average concentration should be the same between renewals. Single spike concentrations are not renewed. Spiked regimes created with oil-coated gravel demonstrate concentrations that decline slower than spiked regimes created by dissolving PAH compounds in water and separating the WAF (Hodson et al., 2019). This is because new oil compounds dissolve from the gravel.

In some articles, dead individuals were counted and removed from the exposure chambers throughout the exposure time. In other cases, mortality was measured after exposure had ended and test subjects were removed to a dish containing clean control water. In these cases, dead individuals were counted either directly after the individuals had been

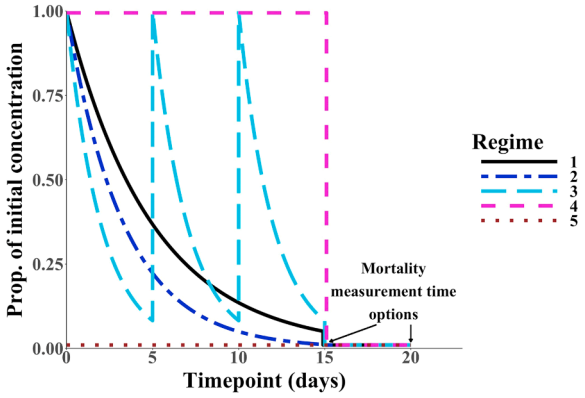


Fig. 1. Hypothetical examples of exposure regimes. Each has exposure time of 15 days. Mortality is measured either directly after exposure or after a 5-day purification period. Exposure regimes are: 1. Spiked regime with oil mixed into gravel and no renewal, 2. Spiked regime with WAF and no renewal, 3. Spiked regime with WAF and renewal every 5 days, 4. Continuous concentration with continuous flow of WAF, and 5. Control regime with no added oil compounds. Note that control concentration is also above 0, because of background PAH concentration in water.

placed in control water or after a predetermined purification period. These experiment characteristics and the original concentration measurements of the articles determined the need to standardize the reported measurements and what methods to use. If the exposure regime was continuous concentration, we assumed that the reported concentrations persisted throughout the exposure. If an article used a spiked exposure regime and reported exposure with initial concentrations, we calculated the geometric means of the exposure concentration measurements from the initial to the minimum:

$$\left(\prod_{i=1}^{n} C_i\right)^{\frac{1}{n}} \tag{1}$$

where C_i is concentration measurement i and n is the number of concentration measurements from the initial to the minimum. The use of geometric mean concentrations has been recommended for toxicity assays with declining exposure concentrations (OECD, 2019). If an article reported multiple concentration measurements from the start of the exposure to the end, we used those measurements to calculate the geometric means. If the article already reported a geometric mean, arithmetic mean, or a median concentration as measure of exposure, we accepted them as sufficient.

Most of the spiked regime articles reported only initial concentrations, which meant we had to model concentration measurements for the other time points before we could calculate the geometric means. As data for the model, we looked for laboratory exposure studies that had measured PAH concentrations at multiple time points throughout the exposure time (Appendix A Section 1.3, Table A6). We used a Bayesian exponential decay model to describe weathering of PAH compounds through time:

$$Y_i = Beta(a_i,\ b_i) \tag{2}$$

$$a_i = \theta\ \pi_i \tag{3}$$

$$b_i = \theta(1 - \pi_i) \tag{4}$$

$$\theta \sim U(1, 100) \tag{5}$$

$$\pi_i = exp(-\alpha_i\ t_i^{\nu}) \tag{6}$$

$$\nu \sim Beta(1, 1) \tag{7}$$

where response variable Y_i is the proportion of the initial exposure concentration on day t_i of observation i. The response variable follows a beta distribution with shape parameters a_i and b_i. The variance of the response variable is controlled by θ. The expected value of the response variable π_i is modelled with an exponential decay model (6) in which the percent decrease in proportion per day, $exp(-\alpha_i)$, slows down as the number of days increases controlled by parameter ν. The model structure was adapted from Vähätalo et al. (2010). The weathering rate α_i we modelled as a regression model on the log scale so that it only gets positive values:

$$\ln\alpha_i = \beta_0 + \beta_{j[i]} + c_{h[i]} \tag{8}$$

$$c_h \sim N\left(0,\ \sigma^2_{Oil\ type}\right) \tag{9}$$

$$\beta_0 \sim N(0,\ 100\ 000) \tag{10}$$

$$\beta_j \sim N(0,\ 100\ 000) \tag{11}$$

$$\sigma_{Oil\ type} \sim U(0.001,\ 10) \tag{12}$$

where β_0 is the intercept, β_j is the effect of experiment type j and $j[i]$ indexes the experiment type of observation i. Experiment type is either gravel or WAF, where WAF includes all dissolving methods other than gravel. The effect of oil type h on α is modelled as a random intercept c_h (9) and $h[i]$ indexes the oil type of observation i. Advantage of using a random intercept is that in addition to the effects of oils in the model data, we can estimate the effect of a common oil on α, which is any oil not included in the model dataset.

We ran the model with JAGS (Plummer, 2003) using R version 4.3.3 (R Core Team, 2024) and R package R2jags version 0.7–1 (Su and Yajima, 2021) with 3 parallel chains. It took 5000 burn-in iterations and 400 000 further iterations per chain until convergence was achieved. We saved every 400th iteration of the simulated parameter values. We used uninformative normal priors for the β coefficients (10, 11) and uniform priors for ν (7), $\sigma_{Oil\ type}$ (12), and θ (5). We assessed convergence with trace plots and Gelman-Rubin diagnostics (Gelman and Rubin, 1992) (Fig. A2, Table A7). We validated the model fit (Fig. 2) using Bayesian p-values (Gelman, 2003) (Fig. A3).

Next, we predicted the missing concentration measurements for the articles. First, we predicted the proportions of initial concentrations using all saved simulated parameter values and substituting t_i in (6) with the time points we wanted to predict concentrations for. These time points spanned from 0.1 to 83 days. We predicted proportions from the whole range of time points for all oil type and exposure regime-pairs (Fig. A4). When calculating geometric means for an article, we only used those mean predicted proportions that corresponded with the time points that were missing concentration measurements in the article, and that were predicted using parameters corresponding to the oil type and exposure regime in the article. If an article used an oil type that we did not have in the model data, we used proportions that were predicted for common oil. To predict the missing concentration measurements, we multiplied the predicted proportions with the reported initial concentrations. If a predicted concentration measurement was below the control concentration in the article, we marked it as equal to the control concentration. This was according to our assumption that in none of the articles would the concentration decrease below the background PAH concentration.

For articles with single spike concentration regimes, we calculated a geometric mean from the concentration measurements at the end of each exposure day. For articles with multiple spikes, we calculated a geometric mean for one exposure interval between renewals, assuming it to be the same for all intervals. In all articles with renewals, the time between renewals was consistent for each interval. If exposure time or interval was less than one day, we used the initial concentration and the concentration at the end of exposure or interval for the geometric mean. If exposure time or interval was more than one day but ended before the end of last day, we used initial concentration, concentrations at the end of each full day, and one more at the end of exposure or interval.

If mortality was measured directly after exposure, cumulatively during the exposure, or right after moving the test subjects to clean water, we did not do any further standardization. If the experiment used a purification period before mortality was measured, we calculated a time-weighted average:

$$\frac{C_{exp}\ a}{(a+b)} + \frac{C_{pur}\ b}{(a+b)} \tag{13}$$

where C_{exp} is the exposure concentration (geometric mean or other applicable measurement), C_{pur} is the concentration during the purification period (which we assume is the same as control concentration), a is the exposure time and b is the duration of the purification period. The use of time-weighted average concentrations has been recommended by Hodson et al. (2019). We applied this to all articles that applied a purification period no matter what exposure regime was used.

2.4. Modeling missing control concentrations

If an article fulfilled all other inclusion criteria but was missing control concentration measurements, we decided to model them. For

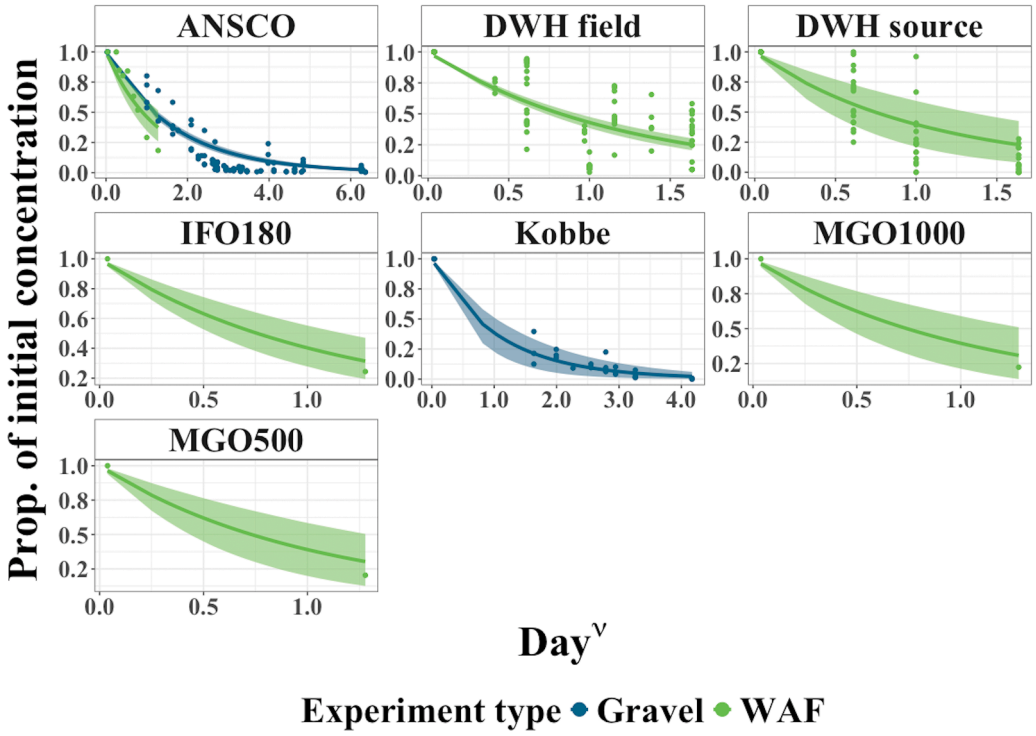


Fig. 2. Fitted values of the model for predicting missing exposure concentrations plotted over data points. Each panel is a different oil type. On the x-axis is exposure time transformed with parameter ν. The scales of the x-axes are different for each oil type. On the y-axis are the proportions of initial concentrations. Lines are means and ribbons are 95 % probability intervals.

this, we searched for articles that had reported control concentrations (Appendix A Section 1.4, Table A8). We decided to use an adaptation of the model described in Section 2.3, but instead of proportions declining as a function of time they decline as a function of the lowest exposure concentrations:

$$Y_i = Beta(a_i, b_i) \quad (14)$$

$$a_i = \theta_{j[i]}\ l_i\ \pi_i \quad (15)$$

$$b_i = \theta_{j[i]}\ l_i(1-\pi_i) \quad (16)$$

$$\theta_j \sim U(1,100) \quad (17)$$

$$\pi_i = exp(-\alpha_i\ l_i^{\nu_{j[i]}}) \quad (18)$$

$$\nu_j \sim Beta(1,1) \quad (19)$$

where response variable Y_i is the proportion of the lowest exposure concentration l_i of observation i. The response variable follows a beta distribution with shape parameters a_i and b_i. The variance of the response variable is controlled by θ_j, which varies by control type, and is weighted with the lowest exposure concentrations l_i. Control types are gravel in water or just water and $j[i]$ indexes the control type of observation i. The expected value of the response variable π_i is modelled with an exponential decay model in which the percent decrease in proportion per unit of lowest exposure concentration, $exp(-\alpha_i)$, slows down as the lowest exposure concentration increases controlled by parameter ν_j, which is different for both control types. The decline rate α_i we modelled as a regression model on the log scale:

$$\ln\alpha_i = \beta_0 + \beta_{j[i]} \quad (20)$$

$$\beta_0 \sim N(0,\ 100\ 000) \quad (21)$$

$$\beta_j \sim N(0,\ 100\ 000) \quad (22)$$

where β_0 is the intercept and β_j is the effect of control type. We ran the model using the same software as for the weathering model (see Section 2.3) with 3 parallel chains. It took 5000 burn-in iterations and 400 000 actual iterations per chain until the chains were properly converged. We saved every 400th iteration of the simulated parameter values. We used the same priors for parameters as in the weathering model. We assessed convergence with trace plots and Gelman-Rubin diagnostics (Fig. A6, Table A9). We validated the model fit (Fig. 3) with Bayesian p-values (Fig. A7).

We predicted missing control concentration measurements by first using the saved parameter values of appropriate control type specific β_j to calculate decline rates α. Then we substituted l_i in (18) with the lowest exposure concentrations of the articles and calculated expected proportions of lowest exposure concentrations π of which we took means (Fig. A8). Then we multiplied these mean proportions with the article-specific lowest exposure concentrations to get missing control concentrations.

3. Results

Our literature search in the Helka database yielded 960 search results (Fig. 4). After removing duplicates and excluding articles based on title and abstract, we had 165. These articles represented 46 species groups and had both lethal and sublethal endpoints. After reading the full texts, the number of remaining articles was 69, which represented 33 species groups. Only nine groups had three or more articles. We decided to focus on eight of those groups: cod, capelin, herring, mahi-mahi, rainbow trout, salmon, sheepshead minnow, and zebrafish. For these groups, we had 42 articles at this point and next, we searched for more results in Google Scholar which gave us 477 initial results in total. A majority of those were duplicates or articles that had already been reviewed during the Helka searches. After reading full texts, we got 21 additional articles for the selected groups. Most were for salmon and herring species. At this point, we had 63 articles from the two databases and had applied all

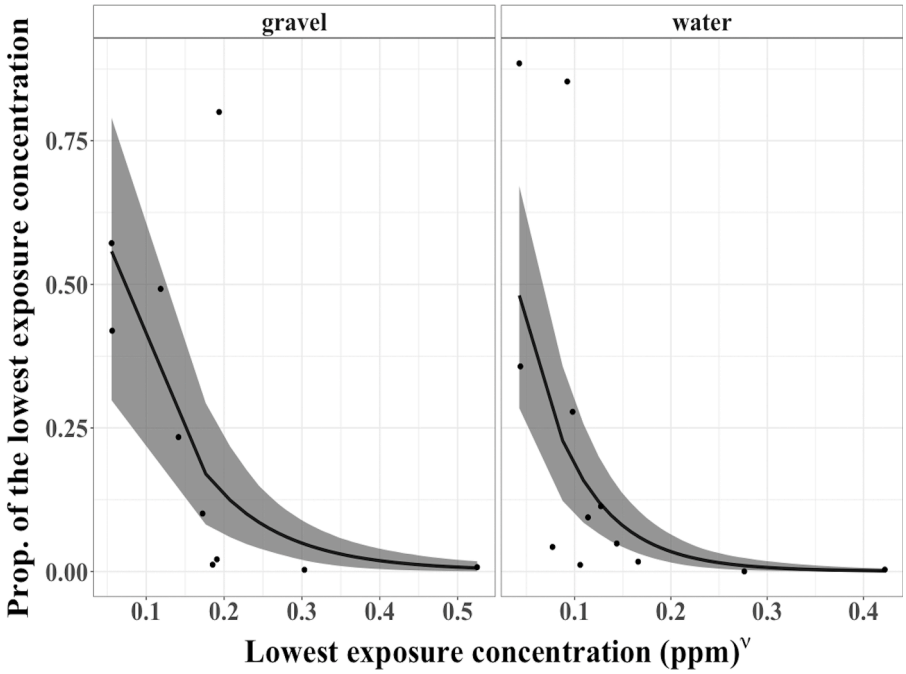


Fig. 3. Fitted values of the model for predicting missing control concentrations plotted over data points. On the x-axis are the lowest exposure PAH concentrations in parts per million (ppm) transformed with parameter ν. The scales of the x-axes are different for both control types. On the y-axis are the proportions of lowest exposure PAH concentrations. Lines are means and ribbons are 95 % probability intervals.

inclusion criteria except those related to exposure metrics. Next, we applied our standardization methods to the search results.

While standardizing exposure metrics, we had to exclude 28 articles, of which 14 had metrics with too few compounds compared to the baseline. The other 14 articles had either wrong exposure metrics, or we could not standardize them. After these exclusions, we no longer had articles for capelin which reduced the number of species groups to seven. No further exclusions were made to the dataset after this step. We were able to standardize the metrics of 20 articles in our dataset. Next, we proceeded to standardize the concentration measurements and model the missing control concentrations of the dataset articles.

Only five articles required no standardization of concentration measurements. Six articles reported suitable concentration measurements but included purification periods, which necessitated the calculation of time-weighted averages. Geometric mean concentrations were calculated for 24 articles, five of which involved purification periods. Two articles provided concentration measurements from multiple time points, but we still had to calculate the geometric means. We modeled missing concentration measurements for 22 of the dataset articles, many of which contained multiple experiments with varied exposure times and initial concentrations. In some cases, the exposure regimes varied in the same article, making individual experiments initially incomparable. We modeled missing control concentrations for 20 articles. Using the combination of our standardization methods we were able to significantly increase the amount of comparable data since our final dataset contained only two articles that required no standardization.

Our final dataset included data from 35 articles in total (Table A5; Appendix B), from which we extracted 649 mortality responses (408 for eggs and 241 for larvae) with varying exposure concentrations and exposure times. Of those responses, 548 were from oil exposures and 101 from controls. The distribution of responses by species groups was as follows (number of responses; percentage from total): cod (95; 14.6 %), herring (135; 20.8 %), mahi-mahi (125; 19.3 %), rainbow trout (137; 21.1 %), salmon (62; 9.6 %), sheepshead minnow (82; 12.6 %), and zebrafish (13; 2.0 %). Oil types with most responses were: ANSCO (132; 20.3 %), *Deepwater Horizon* (DWH) source oil (122; 18.8 %), Troll (64; 9.9 %), DWH field oil (42; 6.5 %), Clearwater McMurray dilbit (27; 4.2 %), Lloydminster conventional heavy crude oil (27; 4.2 %), HFO 6303 (24; 3.7 %), MESA (20; 3.1 %), HFO 7102 (15; 2.3 %), and Cosco Busan Bunker oil (12; 1.8 %). There were 11 more oil types with 10 or less responses each. The additional PAH concentration values in the dataset ranged from 2.02×10^{-7} to 1.85 ppm and the exposure time values from 0.1 to 120 days (Table A10). These ranges varied largely across different species and oil types.

4. Discussion

4.1. Interpretation of the results

The composition of species and oil types in the final dataset, along with the variability in exposure regimes, reflects the retrospective nature of oil spill impact assessment. Laboratory exposure studies are typically conducted in response to major spill events, such as the *Exxon Valdez* or the DWH blowout. For instance, Pacific herring and ANSCO are associated with the *Exxon Valdez* spill, while mahi-mahi and the DWH oil types are connected to the DWH blowout.

The variation in exposure regimes is a result of efforts to replicate the conditions of specific spills. For example, most *Exxon Valdez* studies used spiked gravel exposures, as the spilled oil contaminated gravel in herring spawning beaches, which served as a sustained source of PAH compounds (Short and Harris, 1996). In contrast, DWH blowout studies frequently employed continuous concentration regimes, since oil well blowouts can maintain elevated PAH concentrations over extended periods (Camilli et al., 2010). The differences in PAH concentration ranges observed between the dataset species and oil types can likewise be attributed to simulating specific spill conditions.

Differing exposure times between dataset species is largely explained by differences in the lengths of their egg and larva phases. For instance, egg exposures were often conducted from fertilization to hatching, which explains why, for example, eggs of salmon species had longer exposure times than others. Additionally, exposure times may have been influenced by the commonly held rationale that laboratory exposure studies should use relatively short durations (<96 h) (Bejarano et al., 2014). Such durations may not adequately capture the effects of spills

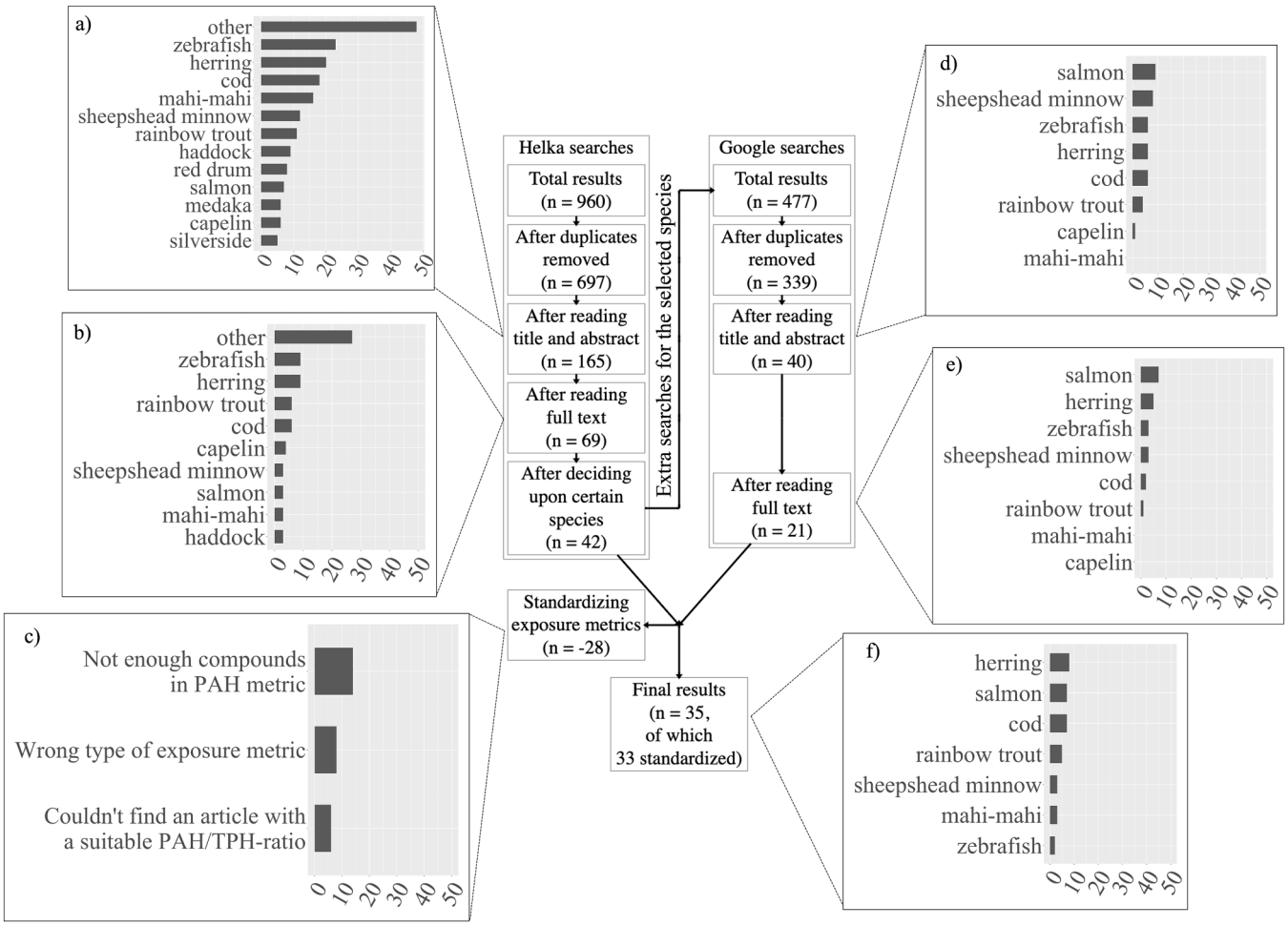


Fig. 4. Flowchart of the systematic review process. Bar plots a, b, d, e, and f show the number of articles per species in different stages of the process. Bar plot c shows the distribution of excluded articles per exclusion reason when we were standardizing exposure metrics.

where PAH concentrations remain elevated for extended periods (Hodson et al., 2019).

Our findings during the standardization of exposure metrics reflect the evolving understanding of how different PAH compounds contribute to oil toxicity (Bejarano et al., 2023; Hodson, 2017). For example, while the U.S. EPA 16 priority PAHs have frequently been used, they do not include alkylated PAHs, which make up 85–95 % of all PAH compounds in oil, or PAHs containing heteroatoms, which have been recognized as major contributors to toxicity. This shift in focus has driven demands for more comprehensive analytical methods. We observed that the use of TPH as a metric was more common in older studies, likely due to the limited availability or high cost of PAH analytical techniques at the time (Hodson et al., 2019). More recently, there has been some support for using TPH, based on the recognition of the toxic potential of oil constituents other than PAHs (Meador and Nahrgang, 2019). Given the ongoing development of knowledge and the retrospective nature of oil spill impact assessment, heterogeneity in study design and reporting is likely to persist. Consequently, the development of standardization methods, such as those presented here, is essential for enabling the use of diverse data sources in toxicity modeling.

4.2. Methodological contribution

Due to retrospective oil spill impact assessment, data is dominated by a few species and oil types, while data for others remain limited. This issue is well known and has led, for instance, to the use of surrogate species or the combining of data of different species (Carroll et al., 2018; Dornberger et al., 2016; Kalter and Passow, 2023; Klok et al., 2014; Olsen et al., 2013). The use of surrogate species is not ideal as biological traits can vary significantly across species (Carroll, 2022). Our standardization methods offer a means to mitigate this issue by increasing the amount of comparable data, thereby reducing the reliance on surrogate species or cross-species data aggregation.

Previous toxicity modeling studies have typically relied on data from similar exposure regimes (Barron et al., 2013; Bejarano et al., 2017; Bejarano and Barron, 2014). We suggest that by standardizing concentration measurements it is possible to use data from different regimes in the same toxicity model, even if spiked regimes report only initial concentrations. Consequently, our dataset enables toxicity modeling based on a broader range of data than would be possible if only studies using specific exposure regimes were included.

Another rationale for standardizing concentration measurements is that initial concentrations of spiked regimes tend to underestimate PAH toxicity (Hodson et al., 2019). Reporting initial concentrations has been common in spiked regime studies; in our dataset, 19 articles required standardization for this reason. The reliance on initial concentrations in such studies has also confounded the debate about oil toxicity to ELS. For instance, following the *Exxon Valdez* spill, several laboratory exposure studies reached divergent conclusions about the PAH concentration levels that could cause acute mortality in ELS (Brannon et al., 2006; Carls et al., 1999; Heintz et al., 1999; Neff et al., 2013; Page et al., 2012). This entire debate was centered on initial concentrations reported for spiked regimes.

Our dataset includes only standardized concentration measurements

for spiked regimes, which avoids underestimation of PAH toxicity (Fig. 5). Because PAH concentration measurement is costly and time-consuming, our standardization method may be particularly valuable for studies with limited resources, where conducting multiple measurements to compute geometric means is not feasible. Additionally, our modeling approach uses Bayesian methods, enabling estimates to be used as priors in subsequent analyses and updated as new data becomes available.

We also observed that many studies employed a purification period. Our method accounts for this by calculating a time-weighted average that incorporates both the exposure concentration and the concentration during the purification period. To our knowledge, this issue has not been explicitly addressed in previous toxicity modeling studies or standardization efforts.

The importance of reporting background pollutants has been previously emphasized (Bejarano et al., 2023). Despite this, we identified only one other approach considering missing control PAH concentrations (Gagliardi et al., 2016), where missing concentrations were estimated as 50 % of the lowest detection limit of the PAH analysis method used. Many of the articles that we reviewed, however, did not report detection limits. By contrast, our approach can model missing control concentrations in any study that provides the lowest reported exposure concentration and the type of water used in the exposure. This is an important methodological contribution since, according to our results, the lack of reporting control concentration measurements seems to be common and thus a significant hindrance to impact assessment in fish population dynamics models.

Approaches to handling variation in exposure metrics have differed across previous toxicity modeling studies. In some cases, the heterogeneity of metrics was acknowledged as a limitation, but the data were otherwise left unadjusted (Dornberger et al., 2016). In other instances, models focused exclusively on studies that used the same metric (Barron et al., 2013; Kalter and Passow, 2023). The approach most similar to ours was used by Bejarano and Barron (2014), who only included studies that reported PAH compound breakdowns, allowing them to include parent PAHs and their alkylated homologs in the analysis. While they also included TPH-based studies, these were analyzed in separate models from those based on PAHs.

Our method, on the other hand, allows for the conversion of TPH metrics and diverse PAH metrics into a common, comparable metric. It also enables the standardization of metrics from studies that did not report their own compound concentration breakdowns. This significantly increases the volume of data that can be incorporated into a single toxicity model. Given the absence of standardized reporting requirements, methods such as ours provide an important means of bypassing gaps in reporting compound concentration breakdowns.

4.3. *Implications for oil toxicity modeling and oil spill impact assessment*

Although our standardization methods increased the volume of usable data, the imbalance in data availability across species and oil types will likely result in greater uncertainty in toxicity model estimates for underrepresented groups. Reducing this uncertainty would require new exposure studies to generate additional data but this is resource-intensive. Given these constraints, modeling strategies should be developed to maximize the utility of existing data (De Laender et al., 2011; Hodson et al., 2019).

One such strategy could be using hierarchical toxicity models, where species and oil-type effects are treated as random effects. This modeling framework estimates all group-level effects within a single structure and enables information borrowing across groups (Gelman and Hill, 2007). For example, the effect of a given species on the PAH slope coefficient would be estimated as a weighted average of information on that species and all other species in the dataset. Similar approaches have been successfully applied in fisheries stock assessment (Kanaji et al., 2023; Prévost et al., 2003; Punt et al., 2011).

Information borrowing can reduce uncertainty where data is sparse. Furthermore, in a hierarchical model, species and oil types can learn from each other's PAH concentration and exposure time ranges, thereby broadening the range of scenarios that can be analyzed for each group.

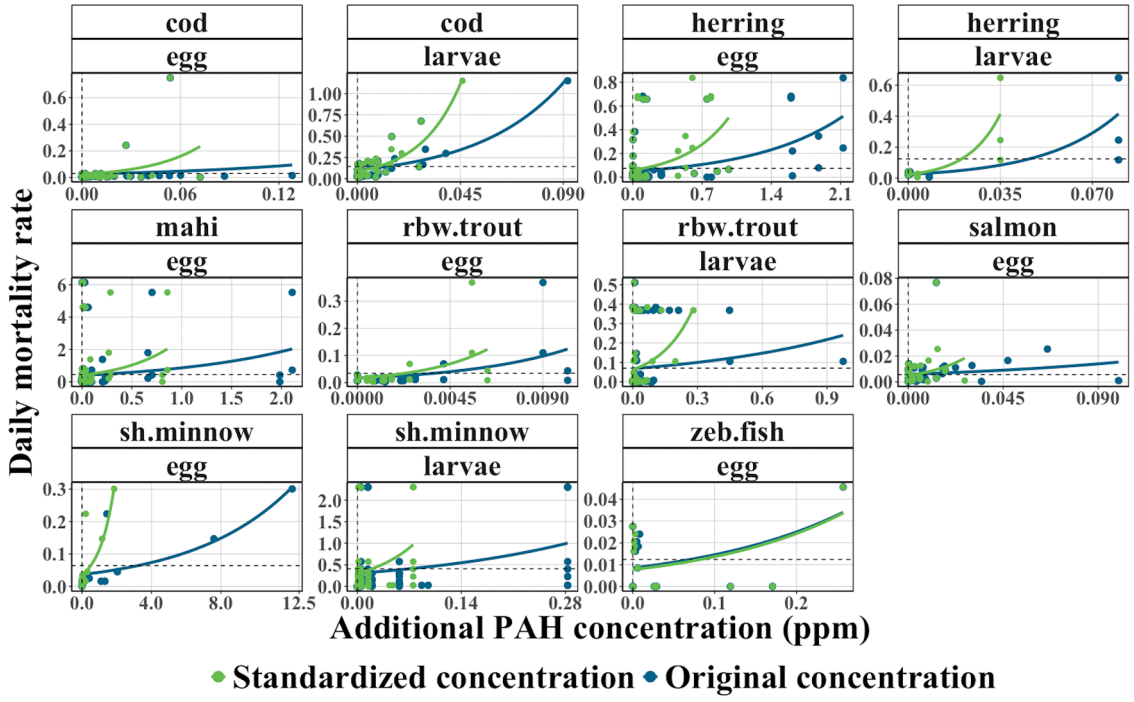


Fig. 5. Comparison of mortality data with standardized exposure concentration measurements and the originals. The scales of the x and y axes vary freely between panels. Daily mortality rate has been calculated as: −1 * log (1 - mortality)/exposure time. Additional PAH concentration in parts per million (ppm) is calculated as: exposure concentration – control concentration. The vertical dotted lines correspond to additional PAH concentration = 0. The horizontal dotted lines correspond to the average daily mortality rates in the controls of the species/life stage combinations. Smoothers were applied to highlight underlying trends in the data.

As our dataset compiles data from multiple species and oil types, it is well suited for hierarchical modeling. In addition, hierarchical models estimate a common population effect that represents the effect of a new group, which can serve as a proxy for species or oil types for which no data is available (Gelman and Hill, 2007).

As we did not limit our inclusion of articles to studies with short exposure times or relatively low PAH concentrations, the dataset is suitable for assessing impacts from oil spill scenarios involving sustained high concentrations (Hodson et al., 2019). This means our dataset supports the application of dose-response curve models, which allow impact assessments across a range of exposure concentrations and durations. This contrasts with species sensitivity distributions (SSDs), which are typically based on median lethal (LC50) or median effective (EC50) concentration data derived from short-duration (24–96 h) exposures (Barron et al., 2013; Bejarano and Farr, 2013).

Our standardization methods can also be adapted to other impact assessment contexts. For example, studies on the sublethal effects of oil on ELS are heterogeneous, and data is unevenly distributed among species (Dornberger et al., 2016; Hodson et al., 2019). Also, to predict oil spill impacts using toxicity model effect estimates based on standardized concentration measurements, it is necessary to calculate geometric mean and/or time-weighted average concentrations from post-spill field measurements, which requires numerous water samples. If frequent sampling is not feasible, the missing concentration measurements can be predicted using the weathering rate parameter estimates from our standardization model (6). In such cases, only a single initial concentration measurement at the onset of the spill is required.

The methods may also be adaptable to toxicity data for other pollutants that consist of complex chemical mixtures and/or are subject to weathering. For example, as interest in alternative maritime fuels continues to grow (WEF, 2024), new areas will increasingly be at risk of exposure. We anticipate that this will necessitate impact assessments in previously unstudied settings, however, these efforts may be limited by sparse data, the absence of standardized toxicity testing protocols, and high data heterogeneity. In such cases, adapting our standardization methods may offer a cost-effective strategy to increase the amount of comparable data.

4.4. Limitations

Limitations of this study relate to data availability, heterogeneity in original study designs, and structural constraints of the compiled dataset. Moreover, the standardization methods are preliminary and exploratory, and thus entail limitations stemming from assumptions made during data transformation and modeling, which introduce uncertainty into the estimates. The key limitations are summarized below:

- **Uncertainty related to variation in oil composition between batches:** Several parts of our methods are affected by the fact that oil composition can vary between batches, even for the same oil type (Bérubé et al., 2022). For example, we used only one article per oil to determine PAH/TPH ratios, which limits their generalizability. Also, some of the PAH breakdowns came from different batches than the target study. And finally, a PAH concentration in our dataset might not represent the same mixture of compounds as a field sample with the same concentration value.
- **Laboratory-based exposure conditions:** The data were generated under laboratory conditions designed to mimic environmental conditions, but these setups cannot capture the full complexity of real-world exposure scenarios.
- **Incomplete removal of non-baseline compounds in some cases:** In a few cases, we could not fully remove non-baseline PAH compounds because breakdowns were not comprehensive enough. They were likely only present at low concentrations, but they still might have contributed to toxicity.
- **Assumptions when deciding whether non-baseline compounds are likely present:** We made decisions about removing non-baseline compounds based on oil type, weathering, and how the oil was dissolved, however, some studies had contradicting details (see Table A5).
- **Baseline PAH metric does not necessarily cover all toxic oil compounds:** Our baseline PAH metric leaves out possible toxic compounds (Meador and Nahrgang, 2019).
- **Simplified weathering model:** The model does not account for variables like temperature. Also, we could not standardize exposure concentrations in the model since most articles did not provide breakdowns past the start of the experiment. This means the weathering rates in (6) might over- or under-predict depending on the compound compositions in the model data compared to baseline.
- **Assuming geometric mean and time-weighted average concentrations:** These quantities were used because we assume that they represent the effective concentrations that eggs and larvae are subject to in spiked exposures and exposures involving purification periods. Assuming other quantities would naturally lead to differing results. The use of geometric mean and time-weighted average is however, recommended in this context (Hodson et al., 2019; OECD, 2019).
- **Limitations in modeling missing control concentrations:** We grouped controls into just two types: gravel and water (which included seawater, tap water, etc.). The different control waters under the type “water” most likely contain different average levels and types of PAHs. Additionally, the model may over-predict if the model data happens to contain exceptionally high control concentrations compared to the lowest exposure concentrations, leading to low decline rates. Also, the link between control concentration and lowest exposure concentration is vague however, the available information in laboratory exposure studies does not offer a better alternative.
- **Uneven species and oil representation in the dataset:** Some species were tested with only a few oils, and some oils with only a few species (Table A5). Consequently, a toxicity model might struggle to separate the effect of species from the effect of oil type.
- **Possible collinearity between predictors:** Differences in PAH concentration ranges across oils, species, and life stages (especially in egg data—see Fig. A9) could cause collinearity between continuous and categorical variables. The same applies for exposure time (Fig. A10). Generalized variance inflation factor (GVIF) values (Fox and Monette, 1992) suggest that this is not a major issue except for exposure time in larvae data (Appendix A, Section 2).

4.5. Recommendations

This study highlights several areas where future research and data reporting practices could be improved to enhance the reliability and usability of oil toxicity datasets and standardization methods:

- **Studies should include their own PAH breakdowns:** Ideally, every exposure study would report PAH compound breakdowns. This recommendation has been made before by Bejarano et al. (2023). This would reduce the need to rely on data from other articles in standardization and ensure consistency in oil type, batch, weathering level, and dissolution method. Breakdowns should be reported for control treatments as well.
- **Exposure concentrations should be reported as time-weighted or geometric means:** Spiked exposure regime studies should report either time-weighted average concentrations (if a purification phase is used) or geometric means. This would make them directly comparable to continuous flow studies and avoid underestimating exposure effects. A similar recommendation has been made by Hodson et al. (2019).

- **Background PAH concentrations should always be reported:** This would help account for background contamination and improve the accuracy of toxicity models that support impact assessment in fish population dynamics models.
- **Improve the modeling of missing control concentrations:** Future versions of the model should include more independent variables, such as the type of water (for instance, seawater, tap water), whether the water was filtered, and the geographical location, which can influence background PAH levels (Zhang et al., 2021). Future studies should also investigate whether variation in control concentrations in (18) could be more accurately explained by variables other than the lowest exposure concentration.
- **Toxicity models should explore different covariate combinations:** To deal with the uneven distribution of data between species and oil types, future models should test different combinations of covariates. This could help avoid confusing the effect of oil type with that of species. Also, the possibility of collinearity between continuous covariates and categorical ones should be investigated. This applies especially to exposure time covariate in larvae data. By allowing information borrowing across groups, hierarchical models can address inconsistencies in PAH concentration and exposure time ranges across species and oil types.
- **Use transformations when dealing with skewed covariates:** Additional PAH concentrations, control PAH concentrations, and exposure time data in our dataset are skewed so transforming the data (for instance, using logarithms) may help improve model performance (Fig. A11).
- **Consider environmental and experimental variables when using the dataset in toxicity models**: Factors, such as salinity, oil weathering, and UV exposure, were reported in many of the included studies. These variables should be considered when applying the dataset in a toxicity model.

5. Conclusions

The primary objectives of this article were to: (1) compile a dataset from peer-reviewed laboratory exposure studies examining ELS mortality due to exposure to oil-sourced PAHs, and (2) develop novel standardization methods to improve data comparability and expand the volume of usable data in the context of oil toxicity to ELS. By conducting a systematic review across multiple species, oil types, and exposure conditions, we compiled a dataset for ELS oil toxicity modeling that supports the assessment of oil spill impacts on fish populations across a broad range of oil spill scenarios, thereby enhancing the capacity of decision-makers to plan and respond effectively to diverse spill events.

Through the development and application of our standardization methods, we were able to integrate heterogeneous experimental data into a unified framework, suitable for modeling population-level oil spill impacts. We demonstrated that substantial increases in comparable data can be achieved without the need for new laboratory exposure studies. This work also helped to consolidate and clarify the main sources of heterogeneity in ELS laboratory oil exposure studies. These include the use of diverse PAH and TPH exposure metrics, variations in exposure concentration measurements, differences in exposure regimes, incomplete reporting of background PAH concentrations, and inconsistent PAH compound coverage due to differing analytical techniques.

Despite the benefits, limitations remain. These include uncertainties introduced by variations in oil composition across batches, incomplete experimental reporting, and the uneven distribution of data across species and oil types. Nevertheless, the dataset and methods we developed here offer a solid basis for ELS oil toxicity modeling and can support advanced modeling strategies, such as hierarchical approaches, that mitigate the effects of data imbalance. Furthermore, our standardization approach demonstrates the potential for adaptation to a broader range of environmental impact assessments.

To improve future applications, we repeat the often-stated recommendation to report breakdowns of individual compound concentrations in laboratory exposure studies. Background PAH levels should also be reported and standardized concentration measurements be used. Our work highlights the continued need for method development to address persistent data heterogeneity in ELS oil exposure studies and supports a more integrated, resource-efficient approach toward oil spill impact assessment and environmental impact assessment in general.

Artwork

All figures and tables with prefix ‘A’ are found in Appendix A. We made the colors in all figures using color blind friendly palettes, if deemed necessary.

AI declaration

During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript only. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Sami Vikkula: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Samu Mäntyniemi:** Writing – review & editing, Validation, Supervision, Software, Methodology. **Sakari Kuikka:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding for this work came from two projects: 1. WISE project (https://wiseproject.fi/en/info/) funded by Research Council of Finland (Strategic Research Funding, grant number 336255, 2021) and 2. Gyroscope project (https://sites.utu.fi/gyroscope/) funded by Research Council of Finland (Targeted Academy Projects, grant number 353059, 2023).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.aquatox.2025.107480.

Data availability

Navigating Data Diversity Obstacles in Assessing Oil Spill Impacts on Fish Early Life Stages (Original data) (Mendeley Data)

References

Adams, J., Bornstein, J.M., Munno, K., Hollebone, B., King, T., Brown, R.S., Hodson, P.V., 2014. Identification of compounds in heavy fuel oil that are chronically toxic to rainbow trout embryos by effects-driven chemical fractionation. Environ. Toxicol. Chem. 33, 825–835. https://doi.org/10.1002/etc.2497.

Adams, J., Charbonneau, K., Tuori, D., Brown, R.S., Hodson, P.V., 2017. Review of methods for measuring the toxicity to aquatic organisms of the water accommodated

fraction (WAF) and chemically-enhanced water accommodated fraction (CEWAF) of petroleum (Canadian Science Advisory Secretariat research document No. 1919-5044; 2017/064). https://publications.gc.ca/site/eng/9.850393/publication.html.
Ainsworth, C.H., Paris, C.B., Perlin, N., Dornberger, L.N., Iii, W.F.P., Chancellor, E., Murawski, S., Hollander, D., Daly, K., Romero, I.C., Coleman, F., Perryman, H., 2018. Impacts of the Deepwater Horizon oil spill evaluated using an end-to-end ecosystem model. PLOS ONE 13, e0190840. https://doi.org/10.1371/journal.pone.0190840.
Andersson, J.T., Achten, C., 2015. Time to say goodbye to the 16 EPA PAHs? Toward an up-to-date use of PACs for environmental purposes. Polycycl. Aromat. Compd. 35, 330–354. https://doi.org/10.1080/10406638.2014.991042.
Barron, M.G., Hemmer, M.J., Jackson, C.R., 2013. Development of aquatic toxicity benchmarks for oil products using species sensitivity distributions. Integr. Environ. Assess. Manag. 9, 610–615. https://doi.org/10.1002/ieam.1420.
Barron, M.G., Vivian, D.N., Heintz, R.A., Yim, U.H., 2020. Long-term ecological impacts from oil spills: comparison of Exxon Valdez, Hebei Spirit, and Deepwater Horizon. Environ. Sci. Technol. 54, 6456–6467. https://doi.org/10.1021/acs.est.9b05020.
Bejarano, A.C., Adams, J.E., McDowell, J., Parkerton, T.F., Hanson, M.L., 2023. Recommendations for improving the reporting and communication of aquatic toxicity studies for oil spill planning, response, and environmental assessment. Aquat. Toxicol. 255, 106391. https://doi.org/10.1016/j.aquatox.2022.106391.
Bejarano, A.C., Barron, M.G., 2014. Development and practical application of petroleum and dispersant interspecies correlation models for aquatic species. Environ. Sci. Technol. 48, 4564–4572. https://doi.org/10.1021/es500649v.
Bejarano, A.C., Clark, J.R., Coelho, G.M., 2014. Issues and challenges with oil toxicity data and implications for their use in decision making: a quantitative review. Environ. Toxicol. Chem. 33, 732–742. https://doi.org/10.1002/etc.2501.
Bejarano, A.C., Farr, J.K., 2013. Development of short, acute exposure hazard estimates: a tool for assessing the effects of chemical spills in aquatic environments. Environ. Toxicol. Chem. 32, 1918–1927. https://doi.org/10.1002/etc.2255.
Bejarano, A.C., Gardiner, W.W., Barron, M.G., Word, J.Q., 2017. Relative sensitivity of Arctic species to physically and chemically dispersed oil determined from three hydrocarbon measures of aquatic toxicity. Mar. Pollut. Bull. 122, 316–322. https://doi.org/10.1016/j.marpolbul.2017.06.064.
Bérubé, R., Lefebvre-Raine, M., Gauthier, C., Bourdin, T., Bellot, P., Triffault-Bouchet, G., Langlois, V.S., Couture, P., 2022. Comparative toxicity of conventional and unconventional oils during rainbow trout (Oncorhynchus mykiss) embryonic development: from molecular to health consequences. Chemosphere Oxf. 288. https://doi.org/10.1016/j.chemosphere.2021.132521, 132521–132521.
Bonatesta, F., Khursigara, A.J., Ackerly, K.L., Esbaugh, A.J., Mager, E.M., 2022. Early life-stage Deepwater Horizon crude oil exposure induces latent osmoregulatory defects in larval red drum (Sciaenops ocellatus). Comp. Biochem. Physiol. Part C Toxicol. Pharmacol. 260, 109405. https://doi.org/10.1016/j.cbpc.2022.109405.
Brannon, E.L., Collins, K.M., Brown, J.S., Neff, J.M., Parker, K.R., Stubblefield, W.A., 2006. Toxicity of weathered Exxon Valdez crude oil to pink salmon embryos. Environ. Toxicol. Chem. 25, 962–972. https://doi.org/10.1897/05-129R1.1.
Camilli, R., Reddy, C.M., Yoerger, D.R., Van Mooy, B.A.S., Jakuba, M.V., Kinsey, J.C., McIntyre, C.P., Sylva, S.P., Maloney, J.V., 2010. Tracking hydrocarbon plume transport and biodegradation at Deepwater Horizon. Science 330, 201–204. https://doi.org/10.1126/science.1195223.
Caprile, A., Leclerc, G., 2024. Russia's "shadow fleet": bringing the threat to light (Briefing No. PE 766.242). EPRS (European Parliamentary Research Service). https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/766242/EPRS_BRI%282024%29766242_EN.pdf.
Carls, M.G., Rice, S.D., Hose, J.E., 1999. Sensitivity of fish embryos to weathered crude oil: part I. Low-level exposure during incubation causes malformations, genetic damage, and mortality in larval pacific herring (Clupea pallasi). Environ. Toxicol. Chem. 18, 481–493. https://doi.org/10.1002/etc.5620180317.
Carroll, J., Frøysa, H.G., Vikebø, F., Broch, O.J., Howell, D., Nepstad, R., Augustine, S., Skeie, G.M., Bockwoldt, M., 2022. An annual profile of the impacts of simulated oil spills on the Northeast Arctic cod and haddock fisheries. Mar. Pollut. Bull. 184, 114207. https://doi.org/10.1016/j.marpolbul.2022.114207.
Carroll, J., Vikebø, F., Howell, D., Broch, O.J., Nepstad, R., Augustine, S., Skeie, G.M., Bast, R., Juselius, J., 2018. Assessing impacts of simulated oil spills on the Northeast Arctic cod fishery. Mar. Pollut. Bull. 126, 63–73. https://doi.org/10.1016/j.marpolbul.2017.10.069.
De Laender, F., Olsen, G.H., Frost, T., Grøsvik, B.E., Grung, M., Hansen, B.H., Hendriks, A.J., Hjorth, M., Janssen, C.R., Klok, C., Nordtug, T., Smit, M., Carroll, J., Camus, L., 2011. Ecotoxicological mechanisms and models in an impact analysis tool for oil spills. J. Toxicol. Environ. Health A 74, 605–619. https://doi.org/10.1080/15287394.2011.550567.
Dornberger, L., Ainsworth, C., Gosnell, S., Coleman, F., 2016. Developing a polycyclic aromatic hydrocarbon exposure dose-response model for fish health and growth. Mar. Pollut. Bull. 109, 259–266. https://doi.org/10.1016/j.marpolbul.2016.05.072.
Fox, J., Monette, G., 1992. Generalized collinearity diagnostics. J. Am. Stat. Assoc. 87, 178–183. https://doi.org/10.2307/2290467.
Gagliardi, B.S., Pettigrove, V.J., Long, S.M., Hoffmann, A.A., 2016. A meta-analysis evaluating the relationship between aquatic contaminants and chironomid larval deformities in laboratory studies. Environ. Sci. Technol. 50, 12903–12911. https://doi.org/10.1021/acs.est.6b04020.
Gardiner, W.W., Word, J.Q., Word, J.D., Perkins, R.A., McFarlin, K.M., Hester, B.W., Word, L.S., Ray, C.M., 2013. The acute toxicity of chemically and physically dispersed crude oil to key arctic species under arctic conditions during the open water season. Environ. Toxicol. Chem. 32, 2284–2300. https://doi.org/10.1002/etc.2307.
Gelman, A., 2003. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing*. Int. Stat. Rev. 71, 369–382. https://doi.org/10.1111/j.1751-5823.2003.tb00203.x.
Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models, Analytical Methods For Social Research. University Press, Cambridge. https://www.cambridge.org/higheducation/books/data-analysis-using-regression-and-multilevel-hierarchical%20models/32A29531C7FD730C3A68951A17C9D983#overview.
Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–472. https://doi.org/10.1214/ss/1177011136.
Heintz, R.A., Short, J.W., Rice, S.D., 1999. Sensitivity of fish embryos to weathered crude oil: part II. Increased mortality of pink salmon (Oncorhynchus gorbuscha) embryos incubating downstream from weathered Exxon Valdez crude oil. Environ. Toxicol. Chem. 18, 494–503. https://doi.org/10.1002/etc.5620180318.
Hodson, P.V., 2017. The toxicity to fish embryos of PAH in crude and refined oils. Arch. Environ. Contam. Toxicol. 73, 12–18. https://doi.org/10.1007/s00244-016-0357-6.
Hodson, P.V., Adams, J., Brown, R.S., 2019. Oil toxicity test methods must be improved. Environ. Toxicol. Chem. 38, 302–311. https://doi.org/10.1002/etc.4303.
Incardona, J., Carls, M., Holland, L., Linbo, T., Baldwin, D., Myers, M., Peck, K., Tagal, M., Rice, S., Scholz, N., 2015. Very low embryonic crude oil exposures cause lasting cardiac defects in salmon and herring. Sci. Rep. 5, 13499. https://doi.org/10.1038/srep13499.
Incardona, J.P., Gardner, L.D., Linbo, T.L., Brown, T.L., Esbaugh, A.J., Mager, E.M., Stieglitz, J.D., French, B.L., Labenia, J.S., Laetz, C.A., Tagal, M., Sloan, C.A., Elizur, A., Benetti, D.D., Grosell, M., Block, B.A., Scholz, N.L., 2014. Deepwater Horizon crude oil impacts the developing hearts of large predatory pelagic fish. Proc. Natl. Acad. Sci. 111, E1510–E1518. https://doi.org/10.1073/pnas.1320950111.
Incardona, J.P., Linbo, T.L., Cameron, J.R., French, B.L., Bolton, J.L., Gregg, J.L., Donald, C.E., Hershberger, P.K., Scholz, N.L., 2023. Biological responses of Pacific herring embryos to crude oil are quantifiable at exposure levels below conventional limits of quantitation for PAHs in water and tissues. Environ. Sci. Technol. 57, 19214–19222. https://doi.org/10.1021/acs.est.3c04122.
Incardona, J.P., Linbo, T.L., French, B.L., Cameron, J., Peck, K.A., Laetz, C.A., Hicks, M.B., Hutchinson, G., Allan, S.E., Boyd, D.T., Ylitalo, G.M., Scholz, N.L., 2021. Low-level embryonic crude oil exposure disrupts ventricular ballooning and subsequent trabeculation in Pacific herring. Aquat. Toxicol. 235, 105810. https://doi.org/10.1016/j.aquatox.2021.105810.
Incardona, J.P., Swarts, T.L., Edmunds, R.C., Linbo, T.L., Aquilina-Beck, A., Sloan, C.A., Gardner, L.D., Block, B.A., Scholz, N.L., 2013. Exxon Valdez to Deepwater Horizon: comparable toxicity of both crude oils to fish early life stages. Aquat. Toxicol. 142–143, 303–316. https://doi.org/10.1016/j.aquatox.2013.08.011.
ITOPF, 2025. Oil tanker spill statistics 2024 - ITOPF. https://www.itopf.org/knowledge-resources/data-statistics/oil-tanker-spill-statistics-2024/.
Jones, E.R., Simning, D., Serafin, J., Sepúlveda, M.S., Griffitt, R.J., 2020. Acute exposure to oil induces age and species-specific transcriptional responses in embryo-larval estuarine fish. Environ. Pollut. 263, 114325. https://doi.org/10.1016/j.envpol.2020.114325.
Kalter, V., Passow, U., 2023. Quantitative review summarizing the effects of oil pollution on subarctic and arctic marine invertebrates. Environ. Pollut. 319, 120960. https://doi.org/10.1016/j.envpol.2022.120960.
Kanaji, Y., Sasaki, H., Hakamada, T., Okamura, H., 2023. Hierarchical modelling approach to estimate the abundance of data-limited cetacean species and its application to fishery-targeted and rarely seen delphinid species off Japan. ICES J. Mar. Sci. 80, 1643–1657. https://doi.org/10.1093/icesjms/fsad091.
Klok, C., Nordtug, T., Tamis, J.E., 2014. Estimating the impact of petroleum substances on survival in early life stages of cod (Gadus morhua) using the dynamic energy budget theory. Mar. Environ. Res. 101, 60–68. https://doi.org/10.1016/j.marenvres.2014.09.002.
Meador, J.P., Nahrgang, J., 2019. Characterizing crude oil toxicity to early-life stage fish based on a complex mixture: are we making unsupported assumptions? Environ. Sci. Technol. 53, 11080–11092. https://doi.org/10.1021/acs.est.9b02889.
Neff, J.M., Page, D.S., Landrum, P.F., Chapman, P.M., 2013. The importance of both potency and mechanism in dose–response analysis: an example from exposure of Pacific herring (Clupea pallasi) embryos to low concentrations of weathered crude oil. Mar. Pollut. Bull. 67, 7–15. https://doi.org/10.1016/j.marpolbul.2012.12.014.
OECD, 2019. Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures, OECD Series on Testing and Assessment. OECD Publishing, Paris.
Olsen, G.H., Klok, C., Hendriks, A.J., Geraudie, P., De Hoop, L., De Laender, F., Farmen, E., Grøsvik, B.E., Hansen, B.H., Hjorth, M., Jansen, C.R., Nordtug, T., Ravagnan, E., Viaene, K., Carroll, J., 2013. Toxicity data for modeling impacts of oil components in an Arctic ecosystem. Mar. Environ. Res. 90, 9–17. https://doi.org/10.1016/j.marenvres.2013.05.007.
Page, D.S., Chapman, P.M., Landrum, P.F., Neff, J., Elston, R., 2012. A perspective on the toxicity of low concentrations of petroleum-derived polycyclic aromatic hydrocarbons to early life stages of herring and salmon. Hum. Ecol. Risk Assess. 18, 229–260. https://doi.org/10.1080/10807039.2012.650569.
Petersen, G.I., Kristensen, P., 1998. Bioaccumulation of lipophilic substances in fish early life stages. Environ. Toxicol. Chem. 17, 1385–1395. https://doi.org/10.1002/etc.5620170724.
Plummer, M., 2003. JAGS: a program for analysis of bayesian graphical models using gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). Vienna, Austria, 124. https://mcmc-jags.sourceforge.io/.
Prévost, E., Parent, E., Crozier, W., Davidson, I., Dumas, J., Gudbergsson, G., Hindar, K., McGinnity, P., MacLean, J., Sættem, L.M., 2003. Setting biological reference points for Atlantic salmon stocks: transfer of information from data-rich to sparse-data

situations by bayesian hierarchical modelling. ICES J. Mar. Sci. 60, 1177–1193. https://doi.org/10.1016/j.icesjms.2003.08.001.
Punt, A.E., Smith, D.C., Smith, A.D.M., 2011. Among-stock comparisons for improving stock assessments of data-poor stocks: the "Robin Hood" approach. ICES J. Mar. Sci. 68, 972–981. https://doi.org/10.1093/icesjms/fsr039.
R Core Team, 2024. R: a language and environment for statistical computing. https://www.R-project.org/.
Redman, A.D., Parkerton, T.F., 2015. Guidance for improving comparability and relevance of oil toxicity tests. Mar. Pollut. Bull. 98, 156–170. https://doi.org/10.1016/j.marpolbul.2015.06.053.
Schiano Di Lombo, M., Weeks-Santos, S., Clérandeau, C., Triffault-Bouchet, G., Langlois Valérie, S., Couture, P., Cachot, J., 2021. Comparative developmental toxicity of conventional oils and diluted bitumen on early life stages of the rainbow trout (Oncorhynchus mykiss). Aquat. Toxicol. 239. https://doi.org/10.1016/j.aquatox.2021.105937, 105937–105937.
Short, J., Jackson, T.J., Larsen, M.L., Wade, T., 1996. Analytical methods used for the analysis of hydrocarbons in crude oil, tissues, sediments, and seawater collected for the natural resources damage assessment of the Exxon Valdez oil spill. In: Proceedings of the Exxon Valdez Oil Spill Symposium, American Fisheries Society Symposium. American Fisheries Society, pp. 140–148.
Short, J.W., Harris, P.M., 1996. Chemical sampling and analysis of petroleum hydrocarbons in near-surface seawater of prince William sound after the Exxon Valdez oil spill. In: Proceedings of the Exxon Valdez Oil Spill Symposium, American Fisheries Society Symposium. American Fisheries Society, pp. 17–28. https://www.researchgate.net/publication/312949659_Analytical_methods_used_for_the_analysis_of_hydrocarbons_in_crude_oil_tissues_sediments_and_seawater_collected_for_the_natural_resources_damage_assessment_of_the_Exxon_Valdez_oil_spill.
Simning, D., Sepulveda, M., De Guise, S., Bosker, T., Griffitt, R.J., 2019. The combined effects of salinity, hypoxia, and oil exposure on survival and gene expression in developing sheepshead minnows, Cyprinodon variegatus. Aquat. Toxicol. 214. https://doi.org/10.1016/j.aquatox.2019.105234, 105234–105234.
Sørhus, E., Donald, C.E., da Silva, D., Thorsen, A., Karlsen, Ø., Meier, S., 2021. Untangling mechanisms of crude oil toxicity: linking gene expression, morphology and PAHs at two developmental stages in a cold-water fish. Sci. Total Environ. 757, 143896. https://doi.org/10.1016/j.scitotenv.2020.143896.
Sørhus, E., Incardona, J.P., Karlsen, Ø., Linbo, T., Sørensen, L., Nordtug, T., van der Meeren, T., Thorsen, A., Thorbjørnsen, M., Jentoft, S., Edvardsen, R.B., Meier, S., 2016. Crude oil exposures reveal roles for intracellular calcium cycling in haddock craniofacial and cardiac development. Sci. Rep. 6, 31058. https://doi.org/10.1038/srep31058.
Spromberg, J.A., Allan, S.E., Scholz, N.L., 2024. Potential population-level impacts of future oil spills on Pacific herring stocks in Puget Sound. Hum. Ecol. Risk Assess. Int. J. 30, 138–163. https://doi.org/10.1080/10807039.2023.2301529.
Su, Y., Yajima, M., 2021. R2jags: using R to run "JAGS". https://CRAN.R-project.org/package=R2jags.
Sumaila, U.R., Cisneros-Montemayor, A.M., Dyck, A., Huang, L., Cheung, W., Jacquet, J., Kleisner, K., Lam, V., McCrea-Strub, A., Swartz, W., Watson, R., Zeller, D., Pauly, D., 2012. Impact of the Deepwater Horizon well blowout on the economics of US Gulf fisheries. Can. J. Fish. Aquat. Sci. 69, 499–510. https://doi.org/10.1139/f2011-171.
Takats, S., Stillman, D., Cheslack-Postava, F., Abaev, B., Bagdonas, M., Jellinek, A., Najdek, T., Petrov, D., Rentka, M., Vasilakis, M., Venčkauskas, A., Wang, X., Sha, Y., 2024. Zotero. https://www.zotero.org/.
University of Helsinki Library, 2025. Helka [WWW Document]. Helka. URL https://helka.helsinki.fi (accessed 7.13.22).
Vikebø, F.B., Nepstad, R., Matuszak, M., Rikardsen, E.S.U., Laurel, B.J., Meier, S., Eriksen, E., Röhrs, J., Christensen, K.H., Smieszek-Rice, M., Hoel, A.H., Huserbråten, M., 2025. Polar cod early life stage exposure to potential oil spills in the Arctic. Aquat. Toxicol. 281, 107293. https://doi.org/10.1016/j.aquatox.2025.107293.
Vähätalo, A.V., Aarnos, H., Mäntyniemi, S., 2010. Biodegradability continuum and biodegradation kinetics of natural organic matter described by the beta distribution. Biogeochemistry 100, 227–240. https://doi.org/10.1007/s10533-010-9419-4.
Wang, Z., Hollebone, B., Fingas, M., Fieldhouse, B., Sigouin, L., Landriault, M., Smith, P., Noonan, J., Thouin, G., Weaver, J., 2003. Characteristics of Spilled Oils, Fuels, and Petroleum Products: 1. Composition and Properties of Selected Oils (No. 600R03072). Environmental Protection Agency. https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1000AE6.txt.
WEF, 2024. Barriers to scaling zero-emission fuel supply in shipping [WWW Document]. World Economic Forum. URL https://www.weforum.org/publications/fuelling-the-future-of-shipping-key-barriers-to-scaling-zero-emission-fuel-supply/(accessed 10.17.24).
WHO, 2022. Guidelines For Drinking-Water quality: Fourth Edition Incorporating the First and Second Addenda. World Health Organization, Geneva. https://www.who.int/publications-detail-redirect/9789240045064.
Wu, D., Wang, Z., Hollebone, B., McIntosh, S., King, T., Hodson, P.V., 2012. Comparative toxicity of four chemically dispersed and undispersed crude oils to rainbow trout embryos. Environ. Toxicol. Chem. 31, 754–765. https://doi.org/10.1002/etc.1739.
Zhang, Xue, Zhang, Z.-F., Zhang, Xianming, Yang, P.-F., Li, Y.-F., Cai, M., Kallenborn, R., 2021. Dissolved polycyclic aromatic hydrocarbons from the Northwestern Pacific to the Southern Ocean: surface seawater distribution, source apportionment, and air-seawater exchange. Water Res. 207, 117780. https://doi.org/10.1016/j.watres.2021.117780.