Integrated Environmental Assessment and Management — Volume 00, Number 00—pp. 1–16
Received: 30 July 2023 | Revised: 24 November 2023 | Accepted: 10 January 2024

Workshop Synthesis

Implementation of the CREED approach for environmental assessments

Carolina Di Paolo,[1] Irene Bramke,[2] Jenny Stauber,[3,4] Caroline Whalley,[5] Ryan Otter,[6] Yves Verhaegen,[7] Lisa H. Nowell,[8] and Adam C. Ryan[9]

[1]*Dow Benelux B.V., Toxicology and Environmental Research and Consulting, Terneuzen, The Netherlands*
[2]*AstraZeneca BV, Den Haag, The Netherlands*
[3]*CSIRO Environment, Sydney, New South Wales, Australia*
[4]*La Trobe University, Wodonga, Victoria, Australia*
[5]*European Environment Agency, Copenhagen, Denmark*
[6]*Data Science Institute, Middle Tennessee State University, Murfreesboro, Tennessee, USA*
[7]*Concawe, Brussels, Belgium*
[8]*U.S. Geological Survey (Emeritus), Sacramento, California, USA*
[9]*International Zinc Association, Durham, North Carolina, USA*

EDITOR'S NOTE:

This article is part of the special series "Criteria for Reporting and Evaluating Exposure Datasets (CREED)." The papers, developed through a SETAC Technical Workshop, present a tested methodology to evaluate the reliability and relevance of measured chemical data for use in estimating exposure in environmental assessments. Specifically, these papers deliver a method to consistently and transparently evaluate whether a chemical dataset is fit for a specific assessment purpose, as not all measured chemical data are fit for all purposes; to identify limitations of the dataset that may qualify or constrain the use of the data; and to provide guidance to data generators on critical study characteristics that should be reported to ensure that their data are useful to the widest possible range of assessment types.

Abstract

Environmental exposure data are a key component of chemical and ecological assessments, supporting and guiding environmental management decisions and regulations. Measures taken to protect the environment based on exposure data can have social and economic implications. Flawed information may lead to measures being taken in the wrong place or to important action not being taken. Although the advantages of harmonizing evaluation methods have been demonstrated for hazard information, no comparable approach is established for exposure data evaluation. The goal of Criteria for Reporting and Evaluating Exposure Datasets (CREED) is to improve the transparency and consistency with which exposure data are evaluated regarding usability in environmental assessments. Here, we describe the synthesis of the CREED process, and propose methods and tools to summarize and interpret the outcomes of the data usability evaluation in support of decision-making and communication. The CREED outcome includes a summary that reports any key gaps or shortcomings in the reliability (data quality) and relevance (fitness for purpose) of the data being considered. The approach has been implemented in a workbook template (provided as Supporting Information), for assessors to readily follow the workflow and create a report card for any given dataset. The report card communicates the outcome of the CREED evaluation and summarizes important dataset attributes, providing a concise reference pertaining to the dataset usability for a specified purpose and documenting data limitations that may restrict data use or increase environmental assessment uncertainty. The application of CREED is demonstrated through three case studies, which also were used during beta testing of the methodology. As experience with the CREED approach application develops, further improvements may be identified and incorporated into the framework. Such development is to be encouraged in the interest of better science and decision-making, and to make environmental monitoring and assessment more cost-effective. *Integr Environ Assess Manag* 2024;00:1–16. © 2024 SETAC

KEYWORDS: CREED; Environmental assessment; Exposure data usability; Reliability and relevance evaluation; Reporting and evaluating criteria

This article contains online-only Supporting Information.
Address correspondence to cdipaolo@dow.com
Published on wileyonlinelibrary.com/journal/ieam.

INTRODUCTION

measuredEnvironmental exposure data are a key component of chemical and ecological assessments, supporting and guiding management and regulations. Datasets of

 DOI: 10.1002/ieam.4909

concentrations of chemicals in environmental matrices are routinely used (together with ecotoxicity data) for different risk assessment purposes. Practitioners, however, are often faced with uncertainties arising from dataset quality issues and/or data reporting gaps. Furthermore, the evaluation of the quality and relevance of hazard and exposure data is often hampered by the subjectivity inherent in expert judgment. Aiming for best practice, frameworks to determine the suitability of ecotoxicity data for specific purposes have been developed (European Commission, 2018; Kase et al., 2016; Klimisch et al., 1997; Moermond et al., 2017; Warne et al., 2018). One example is Criteria for Reporting and Evaluating Ecotoxicity Data (CRED), which supports evaluations of ecotoxicity data for reliability, relevance, and reporting (Moermond et al., 2016).

As reviewed in the accompanying article by Merrington et al. (forthcoming), previous initiatives have addressed issues that arise when chemical monitoring data are used in exposure assessment (e.g., OECD, 2013). Nonetheless, an approach for the harmonized assessment and reporting of exposure data for environmental purposes is lacking. Regulatory guidance for determining the suitability of exposure datasets for specific objectives is limited across jurisdictions, with considerable gaps and minimal instructions on how to process and evaluate exposure datasets. For example, risk assessment guidance from Europe and the United States (ECHA European Chemicals Agency, 2016; USEPA, 2015) typically mentions "quality of dataset" criteria, albeit at a relatively high level, but rarely includes criteria by which to determine the relevance of data for a given purpose. Consequently, the evaluation of exposure datasets is likely to be subject to inconsistencies of professional judgment, particularly when faced with ambiguous or incomplete information. This may include difficulties in dealing with nondetects (also called censored data), inadequate spatial and temporal coverage, a lack of analytical quality control information (as can be encountered with ecotoxicity datasets), and so forth. In addition, there is minimal guidance on how uncertainties can be effectively quantified and communicated, leading to uncertainty in whether the dataset is usable for a given assessment purpose.

To improve the transparency of, and the confidence in, environmental assessments that use chemical monitoring data to represent exposure, it is important to evaluate the underlying monitoring datasets for both their reliability and their relevance for the specific assessment purpose. Reliability refers to the inherent quality of a given dataset, based on sample collection methods, chemical analysis methods, and data processing and statistics. Relevance refers to the degree of suitability or appropriateness of a dataset to address a specific purpose or to answer the questions that have been defined by the assessor. There would be a clear practical benefit to assessors if systematic and transparent criteria were available to use in evaluating monitoring datasets for both reliability and relevance, including guidance on how to combine datasets from different sources, consider data representativity, and express uncertainty.

Criteria for Reporting and Evaluating Exposure Datasets (CREED) was initiated as a Society of Environmental Toxicology and Chemistry (SETAC)-supported activity to develop a framework and criteria for assessing the reliability, relevance, and usability of measured environmental exposure (monitoring) data, aiming to improve the transparency and consistency with which exposure data are evaluated for use in environmental assessments. The goal is to provide a framework through which expert judgment is guided and documented, making exposure data-use decisions transparent and systematic, and facilitating consistency from assessor to assessor. Following an initial stakeholder analysis, the approach was built, beta-tested, and refined to be applicable across jurisdictions and for a variety of assessor-defined purposes and is intended to provide a harmonized set of "best practices" for risk assessors. It can also be used by those generating data as a guide to the parameters that are important to collect and report, and by database owners as a guide for which data fields are considered important for use in environmental assessments, ultimately supporting the development of reporting standards for environmental monitoring data.

Not all measured environmental data are fit for all potential assessment purposes (such as trends assessment or compliance assessment). How monitoring data are treated and processed can have a considerable bearing upon how they can be used. The CREED approach provides best practice for evaluating monitoring data for reliability and relevance (toward a specified assessment purpose) to promote robust and consistent data use by and between environmental assessors (Merrington et al., forthcoming). The components of this framework are presented in the four papers in the CREED series, which, respectively, review the need for CREED, document existing guidance relating to environmental exposure data, and describe the development of CREED (Merrington et al., forthcoming), present reliability (Hladik et al., forthcoming) and relevance criteria (Peters et al., forthcoming) for evaluating and reporting environmental concentration data, and describe standardized approaches for scoring and summarizing data to facilitate "fit for purpose" data usage (this article).

Here, we describe the CREED framework and its implementation, and provide tools designed to summarize the outcomes of a CREED evaluation and interpret the overall usability of exposure data in a transparent way. We present a procedure for scoring datasets at two usability levels (Silver and Gold), each considering the specific assessment purpose, and provide a tool (presented as a workbook template in Supporting Information) that automates scoring of a dataset based on reliability and relevance criteria ratings as entered by the assessor. We also present a purpose-specific report card, which summarizes the CREED evaluation results in support of decision-making. Three case studies, representing different datasets and assessment purposes, were used in a beta test to demonstrate and refine the application of the CREED framework, and scoring worksheets are provided for the three case studies to illustrate CREED

implementation. It is hoped that the CREED framework will be improved in the future, incorporating the practical experience and knowledge gained through its application by experts in multiple exposure data evaluations.

METHODS

The CREED framework was developed by a diverse group of government, industry, and academic scientists and/or analysts over a two-year period (2021–2023), with much of the breakthrough work occurring during and after a SETAC technical workshop held in Copenhagen, Denmark, in May 2022 (Merrington et al., forthcoming). Development of the CREED framework, with intended applicability across jurisdictions and for a diverse array of assessment types, was necessarily an iterative process. The development of CREED entailed consideration of the relevant professional groups through a stakeholder analysis, and selection of effective summary and visualization tools for representing and communicating assessment findings. A first version of the framework was beta-tested by experts invited through an open-access SETAC webinar (https://www.setac.org/discover-events/webinars.html) and refined afterward. The approach was then implemented into a workbook template for assessors to easily follow the procedures and communicate findings, and demonstrated through its application to three case studies.

THE CREED APPROACH

The CREED framework includes

1. Purpose statement—the assessor defines the assessment purpose, clearly describing how the dataset is proposed to be used.
2. Gateway criteria—minimum requirements that a dataset or study must meet to warrant CREED evaluation. (If one or more of the gateway criteria are not met, the dataset or study should not be evaluated by the CREED approach unless the missing information can be provided.)
3. Reliability criteria—19 criteria across six classes: media, spatial, temporal, analytical, data handling and statistics, and supporting parameters.
4. Relevance criteria—11 criteria across the same six classes as above.
5. Data usability score and report card, which are both purpose-specific.

The general framework (Figure 1) describes how data usability is evaluated relative to a specified risk assessment purpose. Each component of the CREED workflow is described below, and in the final report card, which summarizes the outcomes of the CREED evaluation.

Assessment purpose

A clear definition of the specific assessment purpose for which the to-be-evaluated dataset is to be used is a prerequisite first step in the CREED procedure, because it is essential for the evaluation of data relevance (Peters et al., forthcoming). The CREED approach has been developed to allow the criteria for reliability and relevance to be applied generically (i.e., for any type of environmental assessment purpose), provided that the assessment purpose is sufficiently well-defined. This eliminates the need for diverging sets of criteria for different purposes. The assessment purpose is defined by the assessor at the outset and a clear description (i.e., the purpose statement) is included in the report card. Ideally, the purpose statement will specify the minimum information that is required for a relevance criterion to be "fully met" by the dataset and may also include the information required for a criterion to be "partly met," as appropriate. For additional guidance on how to construct an effective purpose statement, please refer to

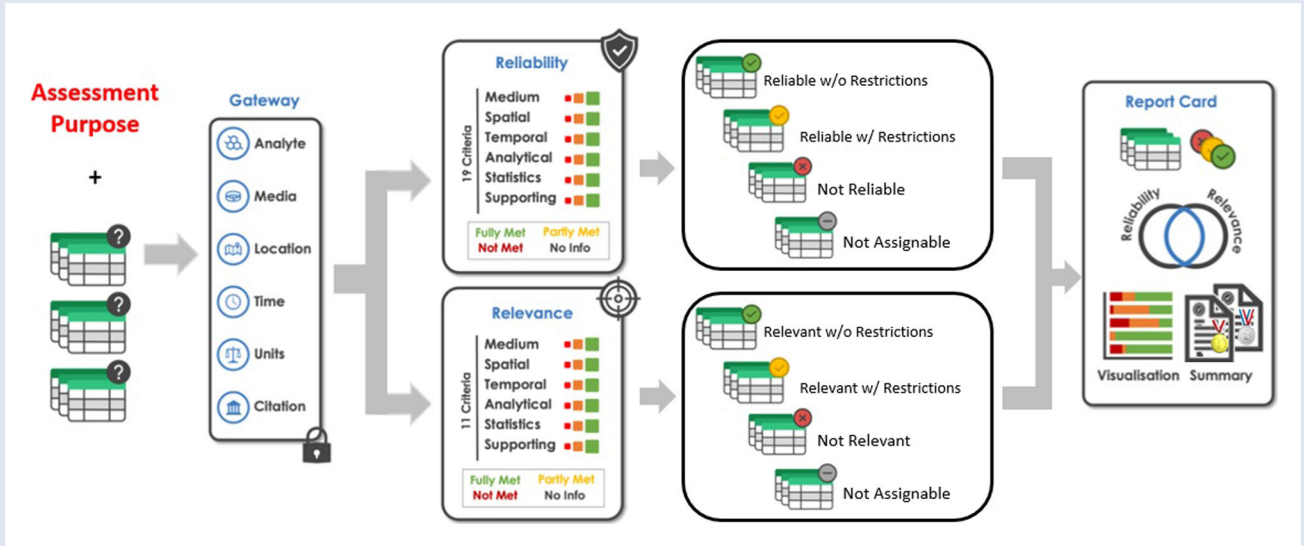

FIGURE 1 The CREED approach workflow

 DOI: 10.1002/ieam.4909

Peters et al. (forthcoming). If the assessor develops a well-defined purpose statement, then dataset evaluation should be straightforward for each criterion—resulting in systematic and consistent dataset evaluations, even by multiple assessors.

Gateway criteria

As an efficiency step, six gateway criteria (Figure 1; Table 1) were identified for determining whether there is sufficient information available in or about the dataset to even start a CREED evaluation (i.e., these define the minimum set of information required to assess the dataset's reliability, relevance, and usability). The gateway criteria address type of media and/or matrix (sampling medium), analyte identity, sampling site location, sampling date, units of measurement, and citation (source of data for traceability). If the dataset fails one or more gateway criteria, the dataset should not be evaluated in its current form under CREED. For example, if the gateway criterion on the analyte identity is not met (i.e., #2: Does the study specify which unique analyte is measured?), there is no point in doing a detailed evaluation for reliability (i.e., #6: Was/were the analyte(s) of interest suitably and definitively identified?) or relevance (i.e., #8: Was/were the reported analyte(s) appropriate for the given purpose?). Further details of gateway criteria are described by Hladik et al. (forthcoming). Failure at the gateway criteria step could prompt the assessor to search for the missing information. In any case, the gateway criteria pass/fail are communicated on the report card, in line with the goal of transparency for decision-making.

Detailed (reliability and relevance) criteria

Datasets that have passed the gateway criteria are evaluated for reliability and relevance in the context of the assessment purpose. There are 19 reliability criteria and 11 relevance criteria (Table 2), as detailed in the accompanying papers by Hladik et al. (forthcoming) and Peters et al. (forthcoming), respectively. Individual criteria are grouped into six classes of criteria, namely, medium (e.g., water, soil, etc.), spatial, temporal, analytical, data handling and statistics, and supporting parameters. Prospective users of the CREED procedure are encouraged to consult Hladik et al. (forthcoming) for details and examples pertaining to reliability criteria and Peters et al. (forthcoming) for details and examples on assessment purpose and relevance criteria.

For all reliability and relevance criteria, the assessor evaluates the dataset relative to the previously defined assessment purpose, assigning a rating to each criterion that reflects the extent to which the dataset satisfies that specific criterion. For each individual criterion, the dataset is rated with one of the five possible ratings (defined in Table 3): "fully met," "partly met," "not met/inappropriate," "not reported," and (only for circumstance-specific criteria, and only when the circumstances described by the criterion do not apply to the dataset) "not applicable." CREED calls for the assessor to record any data limitations (i.e., reasons for assigning a rating other than "fully met") identified during the evaluation process, for transparency.

Each individual reliability or relevance criteria is designated as either required or recommended (Table 2). This distinction underpins the two levels at which datasets are simultaneously scored under CREED, in that only required criteria are scored at the Silver level, whereas all criteria (required plus recommended) are scored at the Gold level.

There is a substantial overlap between CREED criteria (especially for reliability) and OECD's (2013) metadata requirements to support monitoring data. Although OECD (2013) asks some relevance-related questions (e.g., the objectives of the program, the proximity of sources, and discharge emissions), CREED differs by calling for the evaluation of 11 relevance criteria in relation to the specific purpose of the assessment. However, the combined suite of CREED required criteria have a similar function to OECD's (2013) minimum dataset for exposure assessment, and the suite of all CREED criteria (required plus recommended) are analogous in function to OECD's ideal dataset.

CREED two-level scoring approach

The rating of the individual reliability and relevance criteria by the assessor will automatically direct the assignment of the dataset into reliability and relevance categories, respectively—each at both Silver and Gold levels (Table 4). At the Silver level, only the required criteria determine the category and/or score assignment, whereas all criteria

TABLE 1 Gateway criteria

No.	Title	Gateway criterion
1	Sampling medium/matrix	Does the study specify which medium/matrix is sampled?
2	Analyte	Does the study specify which unique analyte is measured?
3	Spatial location	Does the study specify where samples were collected? At a minimum, the country is specified.
4	Year	Does the study indicate when samples were collected? At a minimum, the sampling year is reported.
5	Units	Does the study specify units of measurement?
6	Data source/citation	Does the study cite the source of data and/or is a suitable bibliographic reference available for the study?

TABLE 2 Detailed criteria for reliability (RB) and relevance (RV)

#	Theme	Class	Title	Silver (required)	Gold (recommended)	Criterion
RB1	Reliability	Media	Sample medium/matrix	S	G	Was the sampling medium/matrix reported in detail (for water: dissolved fraction or whole water; for sediment: sieved or whole; for soil, grain size; for biota, species, age, sex, tissue type), and was the matrix appropriate for the analyte of interest?
RB2	Reliability	Media	Collection method/sample type		G	Was the sample collection method reported? Examples include grab, depth- and width-integrated, discrete, composite, or time-integrated samples, or continuous monitoring.
RB3	Reliability	Media	Sample handling		G	Was information reported on sample handling (transport conditions, preservation, filtration, storage)? Was the type of container suitable for use with the analyte of interest (i.e., no loss or contamination)?
RB4	Reliability	Spatial	Site location	S	G	Were the site locations reported?
RB5	Reliability	Temporal	Date and time	S	G	Were the date and time of sample collection reported?
RB6	Reliability	Analytical	Analyte(s) measured	S	G	Was/were the analyte(s) of interest suitably and definitively identified?
RB7	Reliability	Analytical	Limit of detection and/or limit of quantification	S	G	Were limits of detection and/or quantification provided?[a]
RB8	Reliability	Analytical	Accreditation/quality management system	[S]	[G]	Were the laboratory and method accredited for all or almost all samples? Several national and international accreditation bodies are available (e.g., ISO, UKAS); was that laboratory and/or method certified to these standards? Was a quality system (such as, e.g., ISO 17025) adopted? *This is a shortcut criterion; if the answer is yes to any of these questions, then skip to question RB14).* *If no, then additional questions about analytical method are needed, so proceed to question RB9.*
RB9	Reliability	Analytical	Method (skip if method/lab are accredited)	S	G	Was the method sufficiently described or referenced, such that it can be reproduced if necessary? Was method validation included?
RB10	Reliability	Analytical	Lab blank contamination (skip if method/lab are accredited)		G	Was method blank contamination assessed with laboratory blanks?

(*Continued*)

DOI: 10.1002/ieam.4909

TABLE 2 (*Continued*)

#	Theme	Class	Title	Silver (required)	Gold (recommended)	Criterion
RB11	Reliability	Analytical	Recovery/accuracy (skip if method/lab are accredited)		G	Were method recovery/accuracy and/or uncertainty assessed by recovery of standard reference material (SRM) and/or were lab spike samples assessed?
RB12	Reliability	Analytical	Reproducibility/precision (skip if method/lab are accredited)		G	Were method reproducibility and/or uncertainty assessed with lab replicates and long-term control recoveries?
RB13	Reliability	Analytical	Field QC		G	Were quality control (QC) samples collected during field sampling (such as field blanks, spikes, replicates) to demonstrate the method performance for a given field study?
RB14	Reliability	Data handling and statistics	Calculations (consider only if the dataset contains calculated values)		(G)	If chemical concentrations were normalized or adjusted (e.g., to represent bioavailability or toxicity), then were the calculations explained and were they appropriate for the analyte and medium?
RB15	Reliability	Data handling and statistics	Significant figures (consider only if the dataset contains calculated values)		(G)	During calculations, were data reported to the appropriate number of significant figures or decimal places?
RB16	Reliability	Data handling and statistics	Outliers (consider only if the dataset mentions outliers)		(G)	For any outliers deleted from the dataset, was evidence provided that these outliers were due to an error in measurement or contamination?
RB17	Reliability	Data handling and statistics	Censored data (consider only if the dataset contains censored values [i.e., nondetects])	(S)	(G)	Were censored data reported correctly (e.g., as a numerical value plus a less-than sign or another indicator of a nondetect). If a substitution method was used for nondetects (e.g., censored data were replaced by zero, or by 1/2 or another fraction of the LOD/LOQ), then can the original censored data be restored by back-calculation using the reported LOD/LOQ?
RB18	Reliability	Data handling and statistics	Summary statistics procedures (*consider only if the dataset contains summary statistics*)		(G)	Were summary statistics calculated appropriately? If the dataset contained censored data, then were censored data included and were appropriate procedures used to determine summary statistics?
RB19	Reliability	Supporting parameters	Supporting data quality (*consider only if supporting parameters are required for the purpose*)		(G)	If any supporting parameters are required for the assessment purpose, then were the supporting parameter data provided, and were their methods and data quality addressed?

(*Continued*)

TABLE 2 (*Continued*)

#	Theme	Class	Title	Silver (required)	Gold (recommended)	Criterion
RV1	Relevance	Media	Sample medium/matrix	S	G	Was the sampling medium/matrix appropriate for the given purpose?
RV2	Relevance	Media	Collection method/sample type		G	Was the sample collection method (e.g., grab, depth- and width-integrated, discrete, composite, or time-integrated samples, or continuous monitoring) adequate for the given purpose?
RV3	Relevance	Spatial	Study area	S	G	Were the study area and number of locations sampled suitable for the given purpose?
RV4	Relevance	Spatial	Site type		G	Was the rationale for selection of sampling locations provided and is it suitable for the given purpose?
RV5	Relevance	Temporal	Sampling timespan	S	G	Were the samples collected over a time scale that was appropriate for the given purpose?
RV6	Relevance	Temporal	Sampling frequency	S	G	Over the timespan, was the sampling frequency appropriate for the given purpose?
RV7	Relevance	Temporal	Temporal conditions		G	Were conditions during sampling events documented and relevant for the given purpose (e.g., baseflow, storm events, planned/unplanned discharges, etc.)?
RV8	Relevance	Analytical	Analyte	S	G	Was/were the analyte(s) reported appropriate for the given purpose?
RV9	Relevance	Analytical	Sensitivity/LOD/LOQ	S	G	Was the method sensitive enough for the given purpose (i.e., were the LOD and/or LOQ below the benchmarks or metrics to which concentrations in the dataset will be compared)?
RV10	Relevance	Data handling and statistics	Summary statistics type (consider only if the data set contains summary statistics)		(G)	Were the summary statistics provided (e.g., median, geometric mean, arithmetic mean, percentiles) appropriate for the given purpose?
RV11	Relevance	Supporting parameters	Supporting parameters (consider only if supporting parameters are required for the purpose)		(G)	Were all supporting parameters provided that were needed to achieve the given purpose?

Note: Bracket around S or G indicates that RB8 is a shortcut criterion; if RB8 is fully met, RB9-RB12 can be skipped. Parentheses around S or G in Silver and Gold columns indicate that the criterion only applies under certain conditions, which are specified in the Title column.

Abbreviations: ISO, International Organization for Standardization; UKAS, United Kingdom Accreditation Service.

[a]The limit of detection (LOD, also called the Detection Limit, DL, or Method Detection Limit, MDL), is the minimum value that the method can determine is statistically different from blanks. The limit of quantification, LOQ, is the minimum value that the method can quantify with a defined uncertainty.

DOI: 10.1002/ieam.4909

TABLE 3 System of ratings for individual criteria

Rating	Definition
Fully met	All conditions of the criterion are satisfied by the study or dataset.
Partly met	Some of the conditions of the criterion are met for either part or all of the dataset, or all conditions are met by part of the dataset.
Not met/inappropriate	The data or approach were flawed or inappropriate for the analyte or assessment purpose.
Not reported	Insufficient information was provided to evaluate the criterion.
Not applicable	The circumstances required for the criterion do not apply to the dataset.

equired plus recommended) determine the category and/or score at the Gold level. Thus, the Silver level is less ambitious than the Gold level, which represents an ideal dataset where all criteria contribute to the category and/or score.

This two-level scoring system was developed recognizing that "perfect" (Gold standard) datasets are not common, while potentially usable datasets (i.e., those that meet basic criteria) are frequently encountered and should therefore be within the scope of CREED applicability for the framework to be of practical use. For example, some data characteristics are commonly not reported, such as sample handling method, field quality control procedures, or hydrologic conditions during sampling; because these are designated by CREED as recommended criteria (and not as required), datasets may still be usable at the Silver level even if these attributes are not provided. By documenting the limitations of a given dataset, the two-level scoring system allows the assessor to identify why a dataset might be less than ideal, but still usable for a given purpose.

Based on the ratings assigned by the assessor for the reliability and relevance criteria, the dataset is then automatically assigned (by predefined rules) to the appropriate overall reliability and relevance categories at both the Silver and Gold levels (Table 4). The overall categories available for reliability are "reliable without restrictions," "reliable with restrictions," "not reliable," or "not assignable." The categories available for relevance are "relevant without restrictions," "relevant with restrictions," "not relevant," or "not assignable."

The dataset reliability and relevance categories are then combined to determine the overall usability of the dataset for the given purpose. The usability categories are (i) "usable without restrictions"; (ii) "usable with restrictions"; and (iii) "not usable." The assigned categories of reliability, relevance, and usability are captured in the report card, along with the assessment purpose; these category assignments are always purpose-dependent. Importantly, the report card also lists the data limitations that the assessor specified during the reliability and relevance evaluations. If a dataset is scored as "not assignable" due to information having been not reported, the assessor might use the list of data limitations to identify additional supporting documentation, and attempt to fill the related information gap(s) to allow a more conclusive reevaluation of the dataset. In theory, an assessor might elect to use a dataset scored by CREED as "not usable," such as in the absence of other available datasets; in such a case, the data limitations (including missing information) that caused the "not usable" CREED score will be included in the CREED report card and can function as a list of unsupported assumptions, analogous to a warning label, regarding the use of that dataset for the assessment purpose.

Scoring workbook and report card

CREED has been implemented in a Microsoft Excel workbook template (File S1) so that assessors can readily follow the workflow (Figure 1). Copies of the template can be downloaded and used by assessors as a scoring tool to

TABLE 4 Overall reliability and relevance category definitions

Category	Description
Reliable without restrictions/relevant without restrictions	All reliability or relevance criteria (at the Gold level) or all required criteria (at the Silver level) for this study were met. The study is well designed and performed, and it does not contain flaws that affect the reliability of the study.
Reliable with restrictions/relevant with restrictions	Some required (at the Silver level) and/or recommended (at the Gold level) reliability or relevance information is only partly addressed. The study is generally well designed and performed, based on the information reported for this study. Some characteristics of the study will limit the applicability of the data.
Not reliable/not relevant	Not all reliability or relevance required criteria (at the Silver level) or all criteria (at the Gold level) criteria were met. The study has flaws in design or performance.
Not assignable	Information that is needed to evaluate the study is missing. This includes studies that do not give sufficient experimental details and that are only listed in abstracts or secondary literature (books, reviews, databases, etc.) or studies of which the documentation is not sufficient to evaluate reliability for one or more critical aspects.

evaluate their chosen datasets and generate a CREED report card for each dataset.

When using the workbook template to evaluate a dataset, the assessor should step through the tabs from left to right or click on the appropriate level of workflow described in the "CREED Workflow" tab. First, the assessor should enter an appropriate description of the assessment purpose on the "Purpose Statement" tab, and then enter dataset details on the "Dataset Details" tab. The purpose of entering the dataset details is simply to summarize basic characteristics of the dataset for inclusion in the report card. Defining the "Purpose Statement" is critical, as a clear statement will make the rating process simpler, more systematic, and (if multiple studies are being evaluated and/or multiple assessors are evaluating studies for the same assessment purpose) more consistent. Ideally, the purpose statement will specify both the optimum and minimum information thresholds required to meet the assessment purpose, as illustrated in the Supporting Information for the three case studies (Files S2–S5).

Next, the assessor should evaluate the gateway criteria on the "Gateway Criteria" tab, followed (if all gateway criteria are met) by both the reliability and relevance criteria on the "Reliability Criteria" and "Relevance Criteria" tabs, respectively. For any of the detailed reliability and relevance criteria that are less than "fully met," the CREED procedure calls for the assessor to specify in writing any data limitations for that criterion. Based on the criteria ratings and limitations entered in the worksheets by the assessor, the tool will automatically assign the dataset to reliability, relevance, and usability categories at both Gold and Silver levels and create a report card that summarizes the findings and data limitations.

The "Report" tab of the workbook provides a summary of the usability evaluation at the two scoring levels, which, as explained above, are intended to represent dataset quality relative to required criteria only (Silver) and to required plus recommended criteria (Gold). Therefore, a score at the Gold level is determined using all CREED criteria and a score at the Silver level is determined using only those criteria that datasets must meet for most assessment purposes.

Finally, the CREED workbook template offers the option to extract the full evaluation outcomes as a downloadable report card. This includes the dataset details; the assessment purpose statement and thresholds for meeting individual criteria; the gateway criteria pass/fail result; Silver- and Gold-level scores for data reliability, relevance, and usability (relative to the assessment purpose); and any dataset limitations.

CREED refinement and application—Beta test and case studies

Using an early version of CREED, two datasets were evaluated in a beta test of the methodology. The two datasets were obtained from the French "EauFrance" database (source: Naïades database website, url: http://www.naiades.eaufrance.fr/france-entiere#/), based on surface freshwater data in the Occitanie region of France during 2017–2021. These two datasets are provided here in Table S1. The objective of the beta testing was threefold: to evaluate the consistency of scoring among a tripartite pool of beta testers, to receive technical feedback to inform refinement of the approach, and to receive feedback regarding the utility and feasibility of the approach.

The beta-test participants received the datasets and predefined assessment purposes for the case studies, as well as instructions in the form of a Standard Operating Procedure (SOP). They were asked to take a purpose survey (online survey containing the technical CREED evaluation), and a general user perception survey. A total of 90 participants were invited to participate in the beta test, from across a wide variety of sectors (academia—21%; consulting—31%; government—34%; industry—14%), geographic regions (Oceania—15%; Asia—9%; North America—26%; South America—5%; Africa—5%; Europe—40%), and scientific roles (data generators—45%; risk assessors—43%; database owners—3%; other—9%). Of these, 22 participants responded to the general beta-test survey, as detailed in the Results and discussion section.

Findings of the beta test led to minor technical refinements and shaped the CREED approach as it is presented in this and the accompanying papers. Lessons learned from the technical component of the beta test and the perception survey results of the beta tests are described here. Of particular note, the beta test revealed the importance of having a well-defined purpose statement to improve the consistency of scoring among assessors. The technical survey results are not presented in detail here because the purpose statements used in the beta test were simpler than those recommended in the current CREED approach papers, so the results are not directly applicable to the current CREED procedure.

Using the same two datasets considered for the beta test, three case studies are presented to demonstrate the application of CREED and showcase typical outcomes of the CREED approach. The actual scoring of the case study datasets was performed independently by at least two CREED panelists. One dataset, focused on atrazine, was used in two case studies, each one with a different assessment purpose: case study #1 aimed to evaluate atrazine occurrence in surface waters, and case study #2 aimed to determine whether atrazine should be prioritized for regular monitoring, based on comparison with the European Union (EU) Water Framework Directive's (WFD's) annual average environmental quality standard (AA-EQS). A second dataset, containing cadmium and hardness data, was used in case study #3, which aimed to determine whether cadmium was in compliance with the EU WFD's AA-EQS (EU, 2008). The detailed assessment purpose statements are presented in Peters et al. (forthcoming) and in the completed CREED scoring and workflow templates in Files S3–S5.

RESULTS AND DISCUSSION

The CREED framework was developed and successfully tested to improve the transparency of, and the confidence

 DOI: 10.1002/ieam.4909

in, environmental assessments that use monitoring data to estimate exposure. The gateway criteria establish minimum information requirements to be met before a dataset can be evaluated for reliability and relevance under CREED. CREED then guides and documents dataset evaluation in relation to a specified assessment purpose, and communicates the results in a systematic and transparent manner, ultimately providing a usability score for that dataset for that assessment purpose. Aspects of CREED's development and implementation, including lessons learned, are discussed below.

CREED initial stakeholder analysis

Prior to the development of CREED, the main stakeholder groups for each specific assessment purpose were identified and evaluated for their expected degree of interest in and impact on the implementation of CREED for exposure data assessment. Interest was defined as the importance that each stakeholder group gives to, or the extent to which they would be affected by, the CREED implementation, and impact was defined as each group's level of influence or power determining the likelihood of CREED being implemented.

Stakeholder analysis is a tool for practitioners and decision-makers to identify and evaluate the main groups impacted by, and(or) interested in, the specific assessment purpose and this can be used to direct the subsequent communication strategy (Figure 2). The use of stakeholder analysis for environmental decisions has extensively been reviewed and recommendations are available in the literature (Bendtsen et al., 2021; Reed, 2008).

The aim of the initial stakeholder analysis was to identify the main groups and their respective perceived needs and/or aims regarding exposure data use, to consider how to involve them during the framework development, and to define a communication strategy. For this, the workshop participants (i) identified the main stakeholder groups (i.e., data users, data generators, data owners, and the general community); (ii) rated each group as high-medium-or-low for their likely interest in CREED and their likely impact on CREED implementation; and (iii) planned their involvement during the CREED development and implementation, as described below. Stakeholders were grouped as follows:

- Data users were defined as the practitioners who retrieve and evaluate the available datasets and ultimately decide on their usability for different purposes. This group was considered as having high interest and in general also high potential impact and/or influence on the CREED implementation. This finding highlighted the importance of involving practitioners as active players in the development of CREED, as achieved by means of the beta test.
- Data generators included those involved with data production steps such as planning and performing sampling, chemical analysis, data handling, and data reporting. This group was mostly considered as having medium interest and medium impact, except for those involved with risk-based monitoring, who were rated similarly to data users.
- Data owners are responsible for maintaining and providing datasets but may not have risk assessment expertise. These include monitoring scheme owners, environmental regulators, database owners, policy owners, and journals. Data owners were identified as having high impact, owing to their setting requirements for data entry. On the other hand, they were considered as having low to medium interest in CREED implementation, requiring succinct communication tools such as the report card to address their specific needs.
- Community stakeholders include the general public, special interest groups, nongovernmental organizations, and so forth. who may be interested in the data quality and trends over time, for example, in a regional catchment report card (Gladstone Healthy Harbour Partnership Report Card, 2022). Community was considered to have medium interest but in general low impact on the CREED implementation.

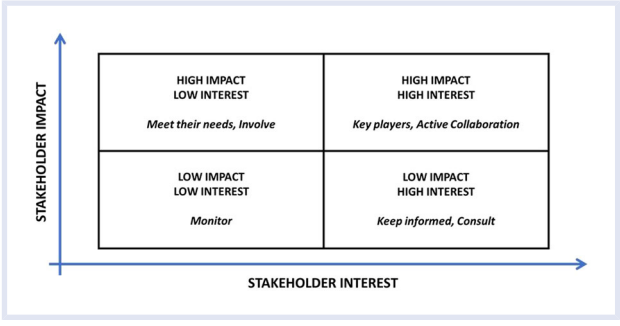

FIGURE 2 A stakeholder analysis was performed during the early stages of CREED development considering as variables the interest and impact of the main stakeholder groups in the implementation of the framework

Beta test—Perception survey and lessons learned

The perception survey response rate was ~25% (22 respondents out of 90 invited participants), with half of the participants indicating that they had over 10 years of experience in evaluating environmental exposure data. The participants included data generators (33%), risk assessors (50%), database owners (7%), and other groups (10%), which is well aligned with the "high interest/high impact" quadrant of the stakeholder analysis outlined in Figure 2.

The responses indicated that the majority of beta testers perceived CREED as both useful and feasible, as follows: CREED directly addresses a key knowledge gap (~80% agreement among respondents); the CREED gateway criteria were feasible to use immediately (~90%); and it was feasible to collect and report most of the information required to meet all the CREED criteria (70%).

Key results focused on potential for implementation of CREED in the beta testers' own area of responsibility within

a particular timeframe: 40% considered it possible to meet minimum requirements within one year, whereas 30% thought that it would take more than five years. The vast majority (95%) agreed that "Wide adoption of CREED and publications of CREED evaluations would help to highlight best practice in environmental monitoring," with 90% also agreeing that adoption of CREED would enable a common understanding of potential weaknesses in monitoring study design, and increase confidence in risk assessment positions that are based on monitoring data.

One of the most important lessons learned from the technical survey of the beta test was that the purpose statement should be defined in detail. It is difficult to judge the usability of a dataset if the purpose for which that dataset is to be used is vague or undefined. Without specificity in the purpose statement, assessors have to use their own best professional judgment to decide what is adequate to meet the criteria; this may result in inconsistencies among assessors who are evaluating the same dataset, or even by a single assessor evaluating multiple datasets. Therefore, for the case studies evaluated in the present CREED papers (which utilized the same datasets as in the beta test), the purpose statements were revised to include detailed thresholds for key criteria. To give one example, relevance criterion #3 asks if the study area and number of sampling locations are suitable for the given purpose. In the beta test, the purpose statement for case study #1 gave no specifics on the required number or density of sampling sites; on the basis of the dataset (Table S1) and a map of sampling sites provided to beta testers (File S2: Figure S1), 78% of beta testers selected "fully met" and 22% selected "partly met" because insufficient information was reported on sampling sites needed for the assessment. The purpose statement for case study #1 therefore has been improved for this paper and its companion papers (see File S3) and it now specifies that the site density should be >1 site per 100 km^2 for relevance criterion #3 to be "fully met" and >1 site per 1000 km^2 to be "partly met." Peters et al. (forthcoming) provide guidance on developing an effective purpose statement.

Another lesson learned from the beta test was that certain technical characteristics of the dataset may require explanation for assessors to be able to successfully evaluate a dataset for reliability; examples include censored data (an explanation of what this means and what techniques are appropriate for handling it) and the limits of quantification and detection (including where this information can be found or inferred within a dataset). The reliability discussion in Hladik et al. (forthcoming) includes technical guidance on these data characteristics, which should facilitate the application of the CREED reliability criteria.

Additional feedback from the beta testers also highlighted the need to clarify how to evaluate the dataset where only some of the data met specific criteria in full. For example, in case study #3, some samples in the cadmium dataset included data for the appropriate supporting parameter required as a toxicity-modifying factor (i.e., hardness), and some did not. Beta testers queried whether this meant that the dataset overall was usable or not. This prompted us to clarify possible reasons why individual criteria might be rated as "partly met."

CREED evaluation of case studies

Case studies #1 and #2 used the same atrazine dataset for two different assessment purposes and case study #3 used the cadmium dataset with its own assessment purpose. Both datasets met the gateway criteria and therefore progressed through the full CREED evaluation of the detailed reliability and relevance criteria. Since the reliability of the data does not typically vary with the assessment purpose, only a single reliability evaluation was performed for each of the two (atrazine and cadmium) datasets, independent of the assessment purpose (i.e., case studies #1 and #2 shared the same reliability evaluation). On the other hand, because the assessment of relevance and consequently also of usability are dependent on the specific assessment purpose, two separate relevance evaluations of the atrazine dataset were performed for case studies #1 and #2, and the cadmium dataset was evaluated once for relevance for case study #3.

The findings of the reliability evaluations for the case studies are described in Hladik et al. (forthcoming). To summarize, at the Silver level, the atrazine dataset (case studies #1 and #2) was scored as "reliable without restrictions" since all the required reliability criteria were "fully met." At the Gold level, the dataset was scored as "not assignable" because information was not reported on sample handling, storage, and transport; there also were data limitations due to incomplete reporting of information on sample collection methods. The cadmium dataset (case study #3) was scored as "reliable with restrictions" at the Silver level (because field quality control results were only generally reported) and as "not assignable" at the Gold level (due to incomplete information on sample collection method and missing information on sample handling, transport, and storage).

The detailed outcomes of the case study relevance evaluations are described in Peters et al. (forthcoming). For case study #1, on atrazine occurrence, the dataset was considered as "relevant with restrictions" at both Gold and Silver levels because the number of sampling locations partly met the sampling frequency requirement and because there was insufficient information provided on the rationale for site selection. For case study #2, on atrazine prioritization for monitoring, the dataset was considered as "relevant with restrictions" at both Gold and Silver levels, again because the number of sampling locations partly met the sampling frequency requirement. For case study #3, on cadmium compliance, the dataset was considered as "not assignable" at both Gold and Silver levels because compliance assessment requires that the number of sampling locations exceed 10% of the number of waterbodies in the region, and the dataset information did not provide the number of waterbodies in the region. Other limitations include insufficient information provided on the rationale for site selection

 DOI: 10.1002/ieam.4909

(Gold level) and only part of the dataset used an analytical method that was sensitive enough for the purpose (Gold and Silver levels).

After scoring a dataset for reliability and relevance, the usability score was determined based on the lower of the reliability or relevance scores within each of the Silver and Gold levels (Table 5; Files S3–S5). For example, in case studies #1 and #2 (atrazine occurrence evaluation and atrazine prioritization, respectively) at the Silver level, the relevance score dictated the usability score of "usable with restrictions" because it was lower (relevant with restrictions) than the reliability score (reliable without restrictions). For case studies #1 and #2 at the Gold level, the reliability score dictated the usability score of "not usable" because it was lower (not assignable) than the relevance score (relevant with restrictions). In terms of usability, if either the reliability or relevance score is not assignable, then it is impossible to know if the dataset is usable, so the conclusion is "not usable" because of missing information. For case study #3, the dataset was "not usable" at both the Silver and Gold levels.

Two-level scoring procedure

The CREED two-level (Gold and Silver) scoring procedure was designed to accommodate the reality that some information required to evaluate some criteria is commonly missing. Ideally, a dataset should at least "partly meet" all reliability and relevance criteria for the dataset to be usable. This ideal case is represented by the Gold level. However, for the CREED procedure to be useful for practitioners in the near term, a "passing" score must be achievable by a reasonable number of currently existing datasets. Thus, the Silver level represents a pragmatic compromise, in that it distinguishes between required and recommended criteria, and only the required criteria must be at least "partly met" for the dataset to be categorized as usable, with or without restrictions, at the Silver level. In practice, the Gold and Silver levels can be thought of as a mechanism for weighing certain criteria as more critical than others (e.g., required vs. recommended) for most purposes. The two-level scoring approach is consistent with the aim of CREED not to unduly penalize or discard data, but instead to highlight information gaps, enable transparency including possible bias in any assessments using the dataset being evaluated, and encourage data generators, users, and managers to fill the identified gaps in future.

While scores at the Silver and Gold levels are automatically provided after the assessor completes the relevance and reliability tabs in the CREED workbook, the scores should be considered together. The combination of Silver and Gold scores represents a continuum of information that can provide useful details to the assessor. For example, in the case studies #1 and #2 developed by CREED, the atrazine dataset was scored as not usable at the Gold level, but usable with restrictions at the Silver level, for both assessment purposes. Therefore, the atrazine dataset was considered as good enough for proceeding with both assessment purposes, even though the assessor and other stakeholders should be aware that data limitations exist and should be considered in related decision-making. Another example would be a scenario in which a dataset is assessed as "usable without restrictions" at the Silver level (because all required criteria are fully met), but "usable with restrictions" at the Gold level (because one or more recommended criteria are only partly met). In aggregating this dataset with other datasets, an assessor might decide to restrict the database to datasets that were usable at the Silver level, or at the Gold level—depending on many factors, including the assessment purpose, data availability, and the assessor's best professional judgment. Issues that arise when aggregating data are discussed in detail by the companion papers on reliability (Hladik et al., forthcoming) and relevance (Peters et al., forthcoming).

Data visualization tools and report card development

Visualization tools are routinely applied to facilitate the communication of data analysis findings to different audiences. CREED explored a number of visualization approaches that can provide varied levels of detail on the exposure data evaluation process. These might support communication strategies aligned with the specific assessment purpose and with the target stakeholder groups.

TABLE 5 Reliability, relevance, and usability scores for each of the three case studies to which the CREED approach was applied

Dataset	Reliability (RB) score	Purpose	Relevance (RV) score	Usability (U) score
Atrazine	*Silver level*: reliable without restrictions *Gold level*: not assignable	Case study #1: occurrence	*Silver level*: relevant with restrictions *Gold level*: relevant with restrictions	*Silver level*: usable with restrictions *Gold level*: not usable
		Case study #2: prioritization	*Silver level*: relevant with restrictions *Gold level*: relevant with restrictions	*Silver level*: usable with restrictions *Gold level*: not usable
Cadmium	*Silver level*: reliable with restrictions *Gold level*: not assignable	Case study #3: Compliance	*Silver level*: not assignable *Gold level*: not assignable	*Silver level*: not usable *Gold level*: not usable

CREED developed the report card template to summarize the assessment purpose, data usability, the dataset attributes, and data limitations. The report card aims to provide, in an accessible manner, the outcomes of the data usability evaluation and the key limitations of the datasets, both of which are important for decision-making, as they summarize the usability of the data in support of the specific assessment purpose. An example of such a report card summary is given for case study #1 in Figure 3, and the complete report card is provided in File S6. Scoring workbooks have been completed for case studies #1–3, including the resulting report cards (available as Files S3–S5).

We consider it important to provide a summary report that would give a short overview of the CREED evaluation outcome. However, we did not think it appropriate to reduce this outcome to a single score or number, so instead provided a text summary of the usability outcome in the "Report" worksheet of the scoring workbook (see examples for case studies in Files S3–S5). A more detailed understanding of the outcome can be found by studying the "Reliability Criteria" and "Relevance Criteria" worksheets in these scoring workbooks.

The role of best professional judgment in CREED

The aim of the CREED framework, in a complementary manner to CRED regarding aquatic ecotoxicity data, is to improve the consistency and transparency of exposure dataset evaluation. To achieve consistency in CREED evaluations among datasets and/or among assessors, we seek to minimize reliance on best professional judgment during the reliability and relevance criteria-rating step. As previously noted, this can be accomplished by setting specific thresholds as part of the purpose statement that define the ideal or optimal ("fully met") versus minimum acceptable ("partly met") thresholds required for the purpose. This does

(A)

DATASET DETAILS - Key Attributes of Dataset

Source (reference):	French National "EauFrance" Database (http://www.naiades.eaufrance.fr/france-entiere#/)		
Reported Analyte:	Atrazine	**Sample Medium/Matrix:**	Surface freshwater
Sampling Conditions:	variable depth (0.2-49 m)	**Study Area:**	Occitanie region, France
Number of Sites:	716 sites	**Site Density:**	more than 1 site per 1,000 km²
Number of Samples:	1-137, 93% of sites have > 4 samples per site	**Site Type(s):**	rivers, streams, lakes
Sampling Period:	2017 to 2021	**Sampling Frequency:**	1-18 samples per year
Sampling Method(s):	occurrence assessment		
Analytical Method(s):	Unfiltered. Either (1) extracted with SPE and analyzed by GC-MS/MS or HPLC-MS/MS; or (2) liquid-liquid extracted and analyzed by GC-MS		
Limit of Quantification:	< 0.0017 - 0.013 µg/L		
Other Details:	Study duration: 1 sample to 4 years, depending on site		

PURPOSE STATEMENT - Required Attributes for Dataset

Assessment of whether atrazine is present at state-of-the-art quantifiable concentrations in surface freshwaters within the Occitanie region, France, between 2017 and 2021, with at least two years of monitoring (fully met) and a minimum of four samples per year at each sampling location (partly met). The dataset should ideally have (high risk) agricultural areas included and a sampling density of one site per 100 km2 (728 sites, based on the size of the study area). This assessment is aimed to determine if withdrawing atrazine from use in the European Union has resulted in a reduction in its occurrence across the Occitanie region.

No.	Purpose Summary Table (optional)	Fully Met (FM)	Partly Met (PM)
RV1	**The sampling medium/matrix was appropriate for the given purpose.**	Surface freshwater	Not applicable
RV2	**The sample collection method** (e.g., grab, depth- and width-integrated, discrete, composite, or time-integrated samples, or continuous monitoring) **was adequate for the given purpose.**	All methods accepted	Not applicable
RV3	**The study area and number of locations sampled was suitable for the given purpose.**	Occitanie: >1 site per 100	Occitanie: >1 site per 1000
RV4	**The rationale for selection of sampling locations was provided and it is suitable for the given purpose.**	At least 25% of the samples	Any surface freshwater
RV5	**The samples were collected over a time scale that was appropriate for the given purpose.**	At least 2 years of monitoring	Not applicable
RV6	**Over the timespan, the sampling frequency was appropriate for the given purpose.**	At least 12 samples a year	At least 4 samples a year
RV7	**Conditions during sampling events were documented and relevant for the given purpose** (e.g., baseflow, storm events, planned/unplanned discharges, etc.).	Any sampling conditions	Not applicable
RV8	**The analyte(s) reported was/were appropriate for the given purpose.**	Atrazine	Not applicable
RV9	**The method was sensitive enough for the given purpose** (i.e., the LOD and/or LOQ were below the benchmarks or metrics to which concentrations in the dataset will be compared).	At least 90% of the LOQs should be < 0.01 ug L-1	At least 90% of the LOQs should be < 0.6 µg L-1 (EU
RV10	**The summary statistics provided** (e.g., median, geometric mean, arithmetic mean, percentiles) **were appropriate for the given purpose.**	Not applicable	Not applicable
RV11	**All supporting parameters that were needed to achieve the given purpose were provided.**	Not applicable	Not applicable

GATEWAY CRITERIA

No.	Title	Criterion	Answer
1	Sampling Medium/ Matrix	Does the study specify which medium/matrix is sampled?	Yes
2	Analyte	Does the study specify which unique analyte is measured?	Yes
3	Spatial Location	Does the study specify where samples were collected? At a minimum, there is enough information for the given purpose (e.g., country).	Yes
4	Year	Does the study indicate when samples were collected? At a minimum, there is enough information for the given purpose (e.g., sampling year).	Yes
5	Units	Does the study specify units of measurement?	Yes
6	Data source/ Citation	Does the study cite the source of data and/or is a suitable bibliographic reference available for the study?	Yes

FIGURE 3 Report card summary for case study #1. Dataset details, purpose statement, and scored gateway criteria are shown in (A), and overall fit for purpose summary, scored reliability criteria, and scored relevance criteria are shown in (B). The complete report card (from "PRINT SUMMARY" in CREED scoring tool) is available in File S6

(B)

FIT FOR PURPOSE

Assessment Level	Data Reliability	Data Relevance	Data Usability	Interpretation	Limitations/Restrictions
Silver (required level)	Reliable Without Restrictions	Relevant With Restrictions	Usable With Restrictions	Meets minimum requirements with restrictions/limitations	Relevance: Site density below optimal (1 in 100 km2) but acceptable (>1 per 1000 km2); Only a subset of sites has 2 or more years of monitoring; Only a subset of sites has >12 samples per year; some have >4 samples per year;
Gold (recommended level)	Reliability Not Assignable	Relevant With Restrictions	Not Usable	Missing data, so does not currently meet minimum requirements	Reliability: Only general info. provided on sample type, without details or reference; No info. on sample handling/transport/storage; Relevance: Site density below optimal (1 in 100 km2) but acceptable (>1 per 1000 km2); Only a subset of sites has 2 or more years of monitoring; Only a subset of sites has >12 samples per year; some have >4 samples per year; Sites are all surface water, but it is unknown how many sites are in agricultural settings;

RELIABILITY CRITERIA

Class	No.	Title	Criterion	Conclusion	Status
Media	RB01	**Sample Medium/ Matrix**	**Was the sampling medium/matrix reported in detail** (for water: dissolved fraction or whole water; for sediment: sieved or whole; for soil, grain size; for biota, species, age, sex, tissue type)**, and was the matrix appropriate for the analyte of interest?**	Fully Met	Required
	RB02	**Collection Method/ Sample Type**	**Was the sample collection method reported?** Examples include grab, depth- and width-integrated, discrete, composite, or time-integrated samples, or continuous monitoring.	Partly Met	Recommended
	RB03	**Sample Handling**	**Was information reported on sample handling** (transport conditions, preservation, filtration, storage)**? Was the type of container suitable for use with the analyte of interest?** (i.e., no loss or contamination)	Not Reported	Recommended
Spatial	RB04	**Site Location**	**Were the site locations reported?**	Fully Met	Required
Temporal	RB05	**Date and Time**	**Were the date and time of sample collection reported?**	Fully Met	Required
Analytical	RB06	**Analyte(s) Measured**	**Was the analyte(s) of interest suitably and definitively identified?**	Fully Met	Required
	RB07	**LoD and/or LoQ**	**Were limits of detection and/or quantitation provided?**	Fully Met	Required
	RB08	**Accreditation/ Quality Management System**	**Were the laboratory and method accredited for all or almost all samples? Several national and international accreditation bodies are available** (e.g. ISO, UKAS)**; Was that laboratory and/or method certified to these standards? Was a quality system** (such as e.g. ISO 17025) **adopted?** *If these criteria are 'Fully Met', please proceed to No. RB13. If not, please proceed to additional questions Nos. RB09-RB12.*	Partly Met	
	RB09	*Method*	*Was the method sufficiently described or referenced, such that it can be reproduced if necessary? Was method validation included?*	Fully Met	Required
	RB10	*Lab Blank Contamination*	*Was method blank contamination assessed with laboratory blanks?*	Fully Met	Recommended
	RB11	*Recovery/ Accuracy*	*Were method recovery/accuracy and/or uncertainty assessed by recovery of standard reference material (SRM) and/or were lab spike samples assessed?*	Fully Met	Recommended
	RB12	*Reproducibility/ Precision*	*Were method reproducibility and/or uncertainty assessed with lab replicates and long-term control recoveries?*	Fully Met	Recommended
	RB13	**Field QC**	**Were quality control** (QC) **samples collected during field sampling** (such as field blanks, spikes, replicates) **to demonstrate the method performance for a given field study?**	Fully Met	Recommended
Data Handling & Statistics	RB14	**Calculations** (if dataset contains calculated values)	**If chemical concentrations were normalised or adjusted** (e.g., to represent bioavailability or toxicity)**, then were the calculations explained and were they appropriate?**	Not Applicable	Recommended
	RB15	**Significant Figures** (if dataset contains calculated	**During calculations, were data reported to the appropriate number of significant figures or decimal places?**	Not Applicable	Recommended
	RB16	**Outliers** (if dataset mentions outliers)	**For any outliers deleted from the data set, was evidence provided that these outliers were due to an error in measurement or contamination?**	Not Applicable	Recommended
	RB17	**Censored Data** (if dataset contains censored values)	**Were censored data reported correctly** (e.g., as a numerical value plus a less-than sign or another indicator of a nondetect)**? If a substitution method was used for nondetects** (e.g., censored data were replaced by zero, or by 1/2 or another fraction of the LOD/LOQ)**, then can the original censored data be restored by back-calculation using the reported LOD/LOQ?**	Fully Met	Required
	RB18	**Summary Statistics Procedures** (if dataset contains summary statistics)	**Were summary statistics calculated appropriately? If the dataset contained censored data, then were censored data included and were appropriate procedures used to determine summary statistics?**	Not Applicable	Recommended
Supporting Parameters	RB19	**Supporting Data Quality** (if supporting parameters are required for the purpose)	**If any supporting parameters are required for the assessment purpose, then were the supporting parameter data provided, and were their methods and data quality addressed?**	Not Applicable	Recommended

RELEVANCE CRITERIA

Class	No.	Title	Criterion	Conclusion	Status
Media	RV01	**Sample Medium/ Matrix**	**Was the sampling medium/matrix appropriate for the given purpose?**	Fully Met	Required
	RV02	**Collection Method/ Sample Type**	**Was the sample collection method** (e.g., grab, depth- and width-integrated, discrete, composite, or time-integrated samples, or continuous monitoring) **adequate for the given purpose?**	Fully Met	Recommended
Spatial	RV03	**Study Area**	**Were the study area and number of locations sampled suitable for the given purpose?**	Partly Met	Required
	RV04	**Site Type**	**Was the rationale for selection of sampling locations provided and was it suitable for the given purpose?**	Partly Met	Recommended
Temporal	RV05	**Sampling Timespan**	**Were the samples collected over a time scale that was appropriate for the given purpose?**	Partly Met	Required
	RV06	**Sampling Frequency**	**Over the timespan, was the sampling frequency appropriate for the given purpose?**	Partly Met	Required
	RV07	**Temporal Conditions**	**Were conditions during sampling events documented and relevant for the given purpose** (e.g., baseflow, storm events, planned/unplanned discharges, etc.)**?**	Fully Met	Recommended
Analytical	RV08	**Analyte**	**Was/were the reported analyte(s) appropriate for the given purpose?**	Fully Met	Required
	RV09	**Sensitivity/ LOD/ LOQ**	**Was the method sensitive enough for the given purpose** (i.e., were the LOD and/or LOQ below the benchmarks or metrics to which concentrations in the dataset will be compared)**?**	Fully Met	Required
Data Handling & Statistics	RV10	**Summary Statistics Type** (if dataset contains summary statistics)	**Were the summary statistics provided** (e.g., median, geometric mean, arithmetic mean, percentiles) **appropriate for the given purpose?**	Not Applicable	Recommended
Supporting Parameters	RV11	**Supporting Parameters** (if supporting parameters are required for the purpose)	**Were all supporting parameters provided that were needed to achieve the given purpose?**	Not Applicable	Required

FIGURE 3 Continued.

not eliminate best professional judgment from CREED, but rather moves it out of the criteria-rating step to the earlier step of purpose-statement definition. For example, in case study #1, the decision to require that each sampling location should have 4–12 samples per year and a minimum of two years of data (Table S1 in Peters et al., forthcoming) was made using best professional judgment regarding how much data would be needed to represent each site within the desired occurrence assessment. This made the criteria-rating step, in which a given dataset would be evaluated as to whether the number of samples and sampling years were sufficient, very straightforward.

CREED implementation

Transparency of decision-making is increasingly important as stakeholders (e.g., citizens, industry, and regulators) seek to understand how and why decisions are made, especially where that decision may impact human and environmental health (Bayer, 2023; PAN Europe, 2022). It is rare to have a full and exhaustive set of information available before a decision must be taken. It is possible to direct further investigations to improve the knowledge base and close significant information gaps, but only if the basis on which the original decision was made is transparent.

CREED aims to support and encourage the different groups of stakeholders toward closing the information gaps on exposure data in support of environmental assessments. The implementation of CREED can be achieved by the combined efforts and interest of data generators, data users, and data owners. For data generators, CREED provides guidance on which parameters and metadata should be reported for their monitoring studies, so that external data users will be able to use these data in exposure assessment. For data users, CREED provides guidance on which parameters are important and should be looked for, when evaluating whether existing data are reliable and relevant for their specific assessment purpose. For data owners such as database managers, CREED specifies the types of data fields that should be included (preferably on a mandatory basis) in the database, so that data generators have a place to adequately describe their studies and data users can find the supporting information that they need to adequately evaluate exposure datasets for use in environmental assessments.

Ultimately, CREED-based reporting formats for monitoring data could be developed, and might serve as valuable input toward the future development of OECD harmonized templates for monitoring data. For example, an assessor might require that all datasets used for a given assessment be usable at the Silver level. Similarly, database owners might require that all datasets be usable at the Silver level to be included in large databases (e.g., Information Platform for Chemical Monitoring [IPCHEM] [Comero et al., 2020]; NORMAN [Dulio et al., 2020]). However, as data reporting can be a burdensome process, particularly when preparing a large dataset derived from many contributors, database owners may need to find a balance between achieving minimal acceptable standards and making it feasible for their data providers to report.

It is anticipated that over time, the *de minimis* standard may be raised as good practice becomes standard. We expect that, as experience with the CREED approach develops, further improvements to CREED may be identified. Such development is to be encouraged, not only in the interests of better science and decision-making but also to make monitoring and assessment more cost-effective.

CONCLUSION

Measures taken to protect the environment based on exposure data can have social and economic implications: Flawed information may lead to suboptimal measures being taken or to important action not being taken. CREED provides an approach and tools to improve the transparency and consistency with which exposure data are evaluated prior to use in environmental assessments. The identification and communication of dataset limitations concerning specific purposes are considered as a core value of the methodology. In this respect, CREED serves also as a data gap analysis tool, where potential weaknesses of the dataset are flagged and may suggest possible strategies to overcome the resulting use limitations. By prioritizing consistency and transparency of dataset evaluation prior to use in environmental assessments, CREED not only supports informed decision-making but can also be very valuable in the planning of future data collection campaigns. In the interest of better science and decision-making, and of the cost-effectiveness of environmental monitoring and assessment, it is encouraged that future improvements be made to the CREED approach, based on early experience with its application in exposure assessment.

AUTHOR CONTRIBUTION

Carolina Di Paolo: Conceptualization; methodology; writing—original draft; writing—review and editing. **Irene Bramke**: Conceptualization; investigation; writing—original draft; writing—review and editing. **Jenny Stauber**: Conceptualization; visualization; writing—original draft; writing—review and editing. **Caroline Whalley**: Conceptualization; writing—original draft; writing—review and editing. **Ryan Otter**: Conceptualization; data curation; formal analysis; resources; software; writing—original draft; writing—review and editing. **Yves Verhaegen**: Conceptualization; methodology; software; writing—original draft; writing—review and editing. **Lisa H. Nowell**: Conceptualization; visualization; writing—review and editing. **Adam C. Ryan**: Conceptualization; formal analysis; methodology; visualization; writing—original draft; writing—review and editing.

ACKNOWLEDGMENT

The authors thank Zhanyun Wang for technical contributions during development of the CREED framework and beta testing. Additionally, we thank the IEAM editors, Kimberly Taylor (U.S. Geological Survey), and two anonymous reviewers for their thoughtful and helpful reviews of the manuscript. We also thank the 22 participants of the beta testing for their feedback on the CREED approach. There was no funding for this article, although the SETAC Technical Workshop was sponsored by Concawe, GSK, Metals Environmental Research Association (MERA), SETAC, Syngenta, and Unilever.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DISCLAIMER

Caroline Whalley is currently employed at the European Environment Agency (EEA). The text in this article solely expresses the author's own views and not those of the EEA.

 DOI: 10.1002/ieam.4909

Jenny Stauber is employed at the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia. This paper has been reviewed and approved for publication by CSIRO. This paper has been peer-reviewed and approved for publication consistent with U.S. Geological Survey Fundamental Science Practices (https://pubs.usgs.gov/circ/1367/).

DATA AVAILABILITY STATEMENT

Data for case studies and tools presented in this manuscript are available in the Supporting Information. No additional data have been used in this article.

ORCID

Carolina Di Paolo http://orcid.org/0000-0002-4281-1034
Irene Bramke https://orcid.org/0009-0009-0807-0752
Jenny Stauber https://orcid.org/0000-0002-1231-3173
Ryan Otter https://orcid.org/0000-0002-1537-4027
Lisa H. Nowell https://orcid.org/0000-0001-5417-7264

SUPPORTING INFORMATION

Table S1. Case study dataset (Excel file), contains: Atrazine dataset from Occitanie, France (from the French National EauFrance database); Cadmium and hardness dataset from Occitanie, France (from the French National EauFrance database)

File S1. CREED workbook template.

File S2. Support and background information for the case studies.

File S3. Complete case study #1 scoring and workflow templates.

File S4. Complete case study #2 scoring and workflow templates.

File S5. Complete case study #3 scoring and workflow templates.

File S6. Report card for case study #1.

REFERENCES

Bayer. (2023). EU glyphosate renewal dossier submission, glyphosate renewal in the EU. https://www.bayer.com/en/agriculture/glyphosateeu

Bendtsen, E. B., Clausen, L. P. W., & Hansen, S. F. (2021). A review of the state-of-the-art for stakeholder analysis with regard to environmental management and regulation. *Journal of Environmental Management*, *279*, 111773. https://doi.org/10.1016/j.jenvman.2020.111773

Comero, S., Dalla Costa, S., Cusinato, A., Korytar, P., Kephalopoulos, S., Bopp, S., & Gawlik, B. M. (2020). A conceptual data quality framework for IPCHEM—The European Commission Information Platform for chemical monitoring. *TrAC: Trends in Analytical Chemistry*, *127*, 115879. https://doi.org/10.1016/j.trac.2020.115879

Dulio, V., Koschorreck, J., van Bavel, B., van den Brink, P., Hollender, J., Munthe, J., Schlabach, M., Aalizadeh, R., Agerstrand, M., Ahrens, L., Allan, I., Alygizakis, N., Barcelo', D., Bohlin-Nizzetto, P., Boutroup, S., Brack, W., Bressy, A., Christensen, J. H., Cirka, L., … Slobodnik, J. (2020). The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): Let's cooperate! *Environmental Science Europe*, *32*, 100. https://doi.org/10.1186/s12302-020-00375-w

ECHA European Chemicals Agency. (2016). *Guidance on information requirements and chemical safety assessment. Chapter R.16: Environmental exposure estimation, Version 3.0. ECHA-16-G-03-EN* (178 p.).

EU. (2008). *Directive 2008/105/EC of the European Parliament and Council on environmental quality standards in the field of water policy*. http://data.europa.eu/eli/dir/2008/105/2013-09-13

European Commission. (2018). *Common implementation strategy for the Water Framework Directive (2000/60/EC) Guidance Document No. 27: Technical guidance for deriving environmental quality standards*. Revised edition. European Communities.

Gladstone Healthy Harbour Partnership. (2022). *2022 Report card*. https://www.ghhp.org.au/2022-report-card#:~:text=Overall%2C%20the%20Gladstone%20Harbour%20Report,Harbour%2C%20good%20(B)

Hladik, M. L, Markus, A., Helsel, D, Nowell, L. H, Polesello, S., Rüdel, H., Szabo, D., & Wilson, I. (Forthcoming). Evaluating the reliability of environmental concentration data to characterize exposure in environmental risk assessments. *Integrated Environmental Assessment and Management*. In press. https://doi.org/10.1002/ieam.4893

Kase, R., Korkaric, M., Werner, I., & Ågerstrand, M. (2016). Criteria for reporting and evaluating ecotoxicity data (CRED): Comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. *Environmental Science Europe*, *28*, 7. https://doi.org/10.1186/s12302-016-0073-x

Klimisch, H. J., Andreae, M., & Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regulatory Toxicology and Pharmacology*, *25*(1), 1–5. https://doi.org/10.1006/rtph.1996.1076

Merrington, G., Nowell, L. H., & Peck, C. (Forthcoming). Using environmental concentration exposure datasets in environmental assessments. The development of Criteria for Reporting and Evaluating Exposure Datasets (CREED). *Integrated Environmental Assessment and Management*. In press. https://doi.org/10.1002/ieam.4899

Moermond, C., Beasley, A., Breton, R., Junghans, M., Laskowski, R., Solomon, K., & Zahner, H. (2017). Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. *Integrated Environmental Assessment and Management*, *13*(4), 640–651. https://doi.org/10.1002/ieam.1870

Moermond, C. T., Kase, R., Korkaric, M., & Ågerstrand, M. (2016). CRED: Criteria for reporting and evaluating ecotoxicity data. *Environmental Toxicology and Chemistry*, *35*(5), 1297–1309. https://doi.org/10.1002/etc.3259

OECD. (2013). *OECD environment, health and safety publications series on testing and assessment no. 185*. Guidance document for exposure assessment based on environmental monitoring. ENV/JM/MONO(2013)7. (79 p.).

PAN Europe. (2022). *Prolonging EU glyphosate authorisation for an extra year: Very bad news for nature and health*. https://www.pan-europe.info/press-releases/2022/10/prolonging-eu-glyphosate-authorisation-extra-year-very-bad-news-nature-and

Peters, A., Beking, M., Oste, L., Hamer, M., Vuaille, J., Harford, A. J., Backhaus, T., Lofts, S., Svendsen, C., & Peck, C. (Forthcoming). Assessing the relevance of environmental exposure data sets. *Integrated Environmental Assessment and Management*. In press. https://doi.org/10.1002/ieam.4881

Reed, M. S. (2008). Stakeholder participation for environmental management: A literature review. *Biological Conservation*, *141*, 2417–2431. https://doi.org/10.1016/j.biocon.2008.07.014

USEPA. (2015). *ProUCL version 5.1 user's guide. Statistical software for environmental applications for data sets with and without nondetect observations*. https://www.epa.gov/sites/default/files/2016-05/documents/proucl_5.1_user-guide.pdf

Warne, M., Batley, G., van Dam, R. A., Chapman, J. C., Fox, D. R., Hickey, C. W., & Stauber, J. (2018). *Revised method for deriving Australian and New Zealand water quality guideline values for toxicants*. https://doi.org/10.13140/RG.2.2.36577.35686

