Environmental Toxicology and Chemistry—Volume 40, Number 12—pp. 3441–3464, 2021
Received: 20 July 2021 | Revised: 4 October 2021 | Accepted: 8 October 2021

Hazard/Risk Assessment

Coral Ecotoxicological Data Evaluation for the Environmental Safety Assessment of Ultraviolet Filters

Emily E. Burns* and Iain A. Davies

Personal Care Products Council, Washington, DC, USA

Abstract: There is growing interest in the environmental safety of ultraviolet (UV) filters found in cosmetic and personal care products (CPCPs). The CPCP industry is assessing appropriate environmental risk assessment (ERA) methods to conduct robust environmental safety assessments for these ingredients. Relevant and reliable data are needed for ERA, particularly when the assessment is supporting regulatory decision-making. In the present study, we apply a data evaluation approach to incorporate nonstandard toxicity data into the ERA process through an expanded range of reliability scores over commonly used approaches (e.g., Klimisch scores). The method employs an upfront screening followed by a data quality assessment based largely on the Criteria for Reporting and Evaluating Ecotoxicity Data (CRED) approach. The method was applied in a coral case study in which UV filter toxicity data was evaluated to identify data points potentially suitable for higher tier and/or regulatory ERA. This is an optimal case study because there are no standard coral toxicity test methods, and UV filter bans are being enacted based on findings reported in the current peer-reviewed data set. Eight studies comprising nine assays were identified; four of the assays did not pass the initial screening assessment. None of the remaining five assays received a high enough reliability score (R_n) to be considered of decision-making quality (i.e., R1 or R2). Four assays were suitable for a preliminary ERA (i.e., R3 or R4), and one assay was not reliable (i.e., R6). These results highlight a need for higher quality coral toxicity studies, potentially through the development of standard test protocols, to generate reliable toxicity endpoints. These data can then be used for ERA to inform environmental protection and sustainability decision-making. *Environ Toxicol Chem* 2021;40:3441–3464. © 2021 Personal Care Products Council. *Environmental Toxicology and Chemistry* published by Wiley Periodicals LLC on behalf of SETAC.

Keywords: Data reliability; Ecotoxicology; Environmental risk assessment; Cosmetic and personal care products; Coral

INTRODUCTION

In recent years, there has been a growing interest in the environmental safety of cosmetic and personal care product (CPCP) ingredients in academic, public, and regulatory spheres. In particular, CPCP ingredients including microplastics (Burns & Boxall, 2018), parabens (Yamamoto et al., 2011), and most recently ultraviolet (UV) filters (Mitchelmore et al., 2021) are receiving attention. The CPCPs can enter the aquatic environment through their intended use and subsequent wash-off, either directly (e.g., swimming) or indirectly through down-the-drain release to wastewater (Burns et al., 2021). Therefore, the CPCP industry is developing product stewardship programs to assess the environmental safety of ingredients. The use of rigorous and standardized environmental risk assessment (ERA) procedures has become increasingly important as ingredient bans based on limited scientific evidence have been enacted, such as sunscreen ingredient bans in Palau (Bill SB 10-135; Remengesau, 2018) and Hawaii (Bill SB 2571; State of Hawaii Senate, 2018). These bans were not based on the results of comprehensive ERAs, and highlight the need for suitable ERA approaches that are protective of ecologically important organisms such as corals (Mitchelmore et al., 2021).

The CPCP industry aims to develop a systematic risk-based prioritization approach that begins with lower tier screening-level assessments, to identify which ingredients are the highest priority to assess using higher tier ERA methods. These assessments will focus on down-the-drain freshwater exposure scenarios on which most of the ecotoxicological hazard and

This article includes online-only Supporting Information.

* Address correspondence to burnse@personalcarecouncil.org
Published online 10 November 2021 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5229

exposure knowledge is based; however, in certain scenarios, for example, products used at the beach and coastal locations, the ERA exposure scenario may need to be extended to include direct wash-off during recreation (Burns et al., 2021). In these special circumstances, it is particularly important to consider relevant marine toxicological data (e.g., cnidarians, mollusks, echinoderms; European Chemicals Agency [ECHA], 2008). These methods will be used to derive risk thresholds in freshwater and marine environments and, if exceeded, trigger risk management and mitigation activities to reduce exposure to an environmentally safe level, similar to current practice by the US Environmental Protection Agency (USEPA) or the European Union under the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation.

The CPCP ERA, in an effort to reduce duplication of environmental data, will consider both peer-reviewed and standardized data required by regulatory authorities, which is aligned with efforts to include all relevant information within ERA frameworks in the United States (USEPA, 2011) and Europe (ECHA, 2008). Ecotoxicity data published in the peer-reviewed literature often capture endpoints, species, or taxa outside standardized testing protocols that can provide useful and ecologically sensitive information that could otherwise be missed (Ågerstrand et al., 2017a). Including nonstandard data from peer-reviewed literature also maximizes the utility of often publicly funded research, which is also aligned with industry's ethical commitment to the three "R's" of animal testing (reduction, refinement, and replacement) and potentially enhances the credibility of ERA among the public, retailers, regulators, and policy-makers (Mebane et al., 2019). It is important that all ecotoxicological data, peer-reviewed or otherwise, be subject to an evaluation of reliability and relevance to determine the adequacy of a study for regulatory, decision-making, or higher tier risk assessment purposes (Kase et al., 2016).

Reliability can be described as the inherent quality of a study, determined through a combined assessment of test design, reporting, performance, and analysis with sufficient information provided to demonstrate the reproducibility and accuracy of the results and independently repeat the test (Hartmann et al., 2017; Klimisch et al., 1997; Moermond et al., 2017). Relevance can be defined as the suitability of the data for a particular hazard identification or risk characterization. This includes the exposure concentration (e.g., below solubility), endpoint (e.g., individual or population-level), species, life stage, and exposure route (ECHA, 2008; Klimisch et al., 1997; Rudén et al., 2017). For example, a reliable study may not always be relevant; it depends on the goal of the assessment (e.g., a marine sediment endpoint may not be suitable for terrestrial risk assessment). To ensure consistency and transparency in this decision-making process, systematic reporting and documentation of the reliability and relevance assessments are needed (Martin et al., 2019). Hartmann et al. (2017) identified a key issue with peer-reviewed studies: a trade-off is often made whereby relevance is favored over reliability, and, although the study may be scientifically valid, the regulatory ERA validity is not met. Study evaluations are therefore critical because the use of low-quality (unreliable) or irrelevant data could lead to overestimates or, more concerningly, underestimates of risk, both of which could be costly through unnecessary mitigation or an overlooked hazard (Harris et al., 2014). An ERA is an inherently uncertain process (Institute of Medicine, 2013; National Research Council, 2009), and it is therefore essential to limit further uncertainties by using high-quality data.

Ecotoxicological data reliability

A systematic approach to evaluate the quality of ecotoxicological data was first proposed by Klimisch et al. (1997), with the goal of harmonizing data evaluation processes worldwide, ultimately for the ERA process, but also to improve the overall quality of the science. Klimisch et al. (1997) created four data reliability categories to classify studies based on how they were conducted and reported: *reliable* (1); *reliable with restrictions* (2); *unreliable* (3); and *unassignable* (4). Categories 1 and 2 were deemed suitable for risk assessment; however, Category 1 data are always preferred when multiple data points exist for a similar endpoint. In addition, Category 3 data can also be useful as supporting information, particularly when the results are similar to those reported in higher quality studies. There have been several criticisms of the Klimisch method (Kase et al., 2016; Moermond et al., 2016), which led to the development of new approaches that built on the foundation provided by Klimisch (as reviewed by Moermond et al., 2017). These reliability tools fall broadly into three categories: pass/fail, numerical score, or categorization. These approaches assess the following study attributes in various levels of detail: test setup, test compound, test organism, test design and conditions, results, and statistics. The USEPA Office of Pesticide Programs (OPP) applies a pass/fail approach, whereby all criteria ($n = 28$) need to be met for a study to be included in the ERA (USEPA, 2011). Numerical scoring is a less rigid approach; a score is assigned based on criteria met, which dictates the Klimisch category it falls within (Breton et al., 2009). Alternatively, Moermond et al. (2016) developed Criteria for Reporting and Evaluating Ecotoxicity Data (CRED), a categorization method for which objective criteria are combined with expert judgment (Moermond et al., 2017). Each criterion is evaluated as fully, partially, or not fulfilled, and a final reliability score is awarded based on expert judgment. The CRED method was also one of the key approaches reviewed in the development of methods for systematic review in USEPA Toxic Substances Control Act risk evaluations (USEPA, 2018).

Expert judgment within a risk assessment is unavoidable, particularly when nonstandard species and endpoints (e.g., corals) are assessed. The key is that such expert judgment should be consistently and transparently applied (Ingre-Khans et al., 2019) to convey decision-making and facilitate necessary scientific scrutiny. Inevitably, different aspects of study quality will be prioritized based on an assessor's expertise (Hartmann et al., 2017). The method therefore needs to be structured with well-defined criteria that can be as consistently and

transparently applied as possible, reducing bias or perceived bias.

Calls for CPCP bans have been posited in the peer-reviewed literature based on the outcomes of individual hazard studies using nonstandard species and endpoints (McCoshum et al., 2016; Zhong et al., 2019) without consideration of their results in the context of risk, other existing data, and standard ERA frameworks, and regardless of their quality or relevance for this regulatory purpose. Furthermore, consideration of data quality and standard ERA frameworks was also absent from recent regulatory bans on the use of certain UV filters, for example, benzophenone-3 (BP3) and ethylhexyl methoxycinnamate (EHMC) in Hawaii. These actions highlight the need for CPCP ingredients to follow a transparent and credible ERA process that includes an assessment of data reliability when higher tier ERA and/or regulatory decision-making are involved. To address this gap, we developed an ecotoxicological data evaluation method based on a combination of established approaches from the peer-reviewed literature and regulatory agencies to determine the appropriate use of nonstandard toxicity data in ERA. Based on the reliability score, the data broadly fall into three categories: suitable only for preliminary ERA, potentially suitable for use in a higher tier ERA to inform decision-making alongside or in lieu of appropriate regulatory data (e.g., REACH data), or discarded because the data are either unreliable or not relevant. We bring together the strengths of existing reliability assessment methodologies into an approach that is streamlined and well suited to addressing the unique challenges posed by nonstandard tests and organisms. An extended reliability scoring system is proposed that offers increased flexibility over the four Klimisch categories, but is also compatible with Klimisch through the application of expert judgment. We apply the evaluation and scoring methods to a UV filter and coral and ecotoxicological case study and discuss the results in the context of both relevance and reliability for ERA.

MATERIALS AND METHODS

Reliability assessment methodology

A minireview of existing ecotoxicity data reliability assessment methodologies was conducted, and a variety of methods were identified that are reported in the peer-reviewed literature and currently used by regulators in the United States, Canada, and Europe. The strengths of these approaches were brought together to build a method that incorporates credibility, consistency, and transparency within a streamlined framework (see Figure 1).

To streamline the assessment process, two relevance questions and three key reliability questions applicable to both standard and nonstandard studies are proposed as screening questions (Table 1). A study is not subject to the data quality evaluation if the second relevance question (i.e., RQ2) or two or more reliability screening questions are failed (i.e., RQ3–RQ5; Table 2). An added benefit of the upfront relevance screening is that it requires the user to state a clear problem formulation and define the scope of endpoints/species/matrices for consideration in the assessment, thereby focusing the process on

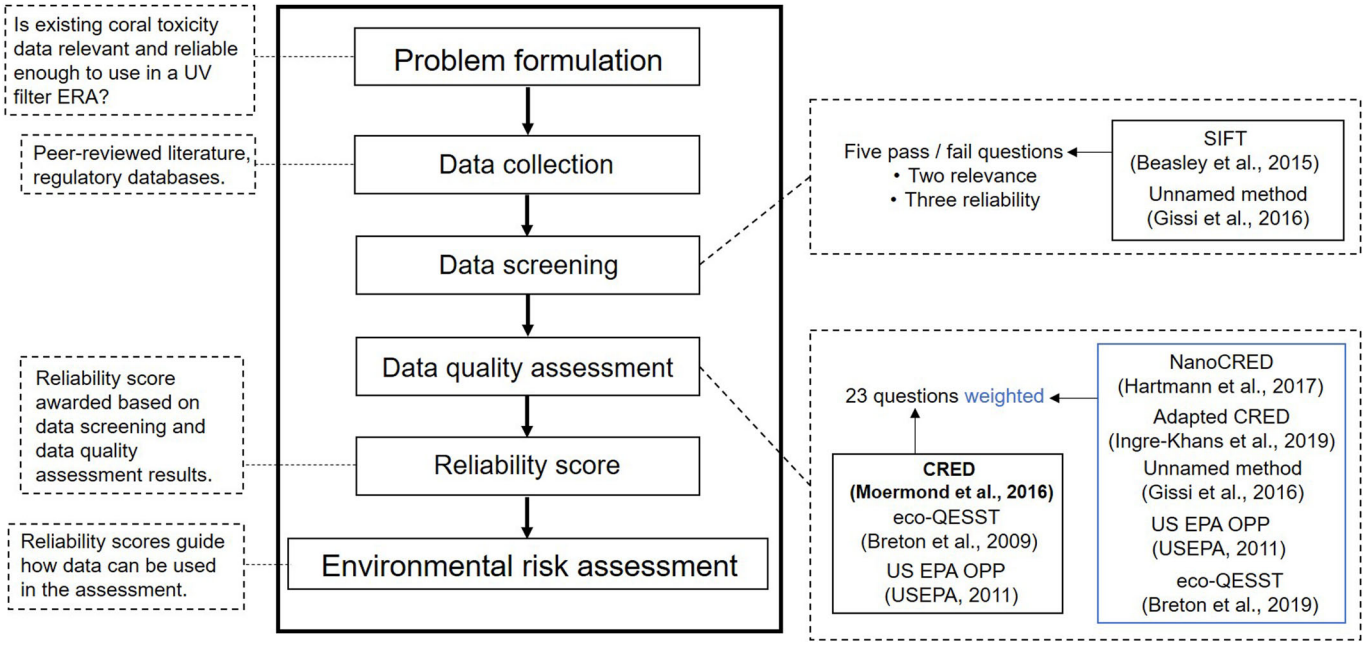

FIGURE 1: A simplified roadmap of how the data reliability assessment precedes the environmental risk assessment (ERA) process. The data reliability assessment consists of a data screening step followed by a 23-question data quality assessment; then the results from both steps determine the reliability score. Studies that fail the screening assessment are not subject to the data quality assessment and do not receive a reliability score. The existing data reliability methods that informed both steps are included in the adjacent boxes. The Criteria for Reporting and Evaluating Ecotoxicity Data (CRED) method (Moermond et al., 2016) is bolded because it largely informed the data quality assessment questions. SIFT = Stepwise Information-Filtering Tool; eco-QESST = Ecotoxicological Quality Evaluation System and Scoring Tool; USEPA OPP = US Environmental Protection Agency Office of Pesticide Programs.

TABLE 1: The ecotoxicological data assessment applied in the case study[a]

		Score	Comment[b]
Screening questions			
RQ1	Is the endpoint ecologically relevant?	Pass/fail	Standard ecologically relevant endpoints pertain to mortality, reproduction, or growth, but nonstandard endpoints with demonstrated ecological relevance (e.g., physiological, behavioral, biochemical) can be included (see Warne et al., 2018). Ecologically relevant endpoints should be outlined prior to assessment based on species (e.g., coral) relevant to the problem formulation.[c]
RQ2	Is the test organism relevant to the compartment, test compound, and/or assessment?	Pass/fail	The relevance of a particular study will be dependent on the problem formulation that defines the scope of the assessment.[d] See text for problem formulation, briefly, intact coral at any life stage originating from any region. Exposures are limited to the water phase and relevant test compounds are organic ultraviolet (UV) filters approved for use in the United States. UV filter mixtures are beyond the scope of the assessment.
RQ3	Was a negative control and solvent control (if necessary) at least duplicated?	Pass/fail	A negative control must be performed and if a solvent was used, a solvent control must also be included.
RQ4	Are ≥4 treatment concentrations included (including control) or experiment specifically designed as limit test?	Pass/fail	Most OECD tests require at least five test concentrations. A limit test is an exception and can pass with fewer treatment concentrations if no effects are observed. Expert judgment is needed to determine whether these tests are designed appropriately based on their results and endpoint(s) reported.
RQ5	Are endpoints based on measured concentrations if they deviate by ≥20% of the nominal concentration? If only nominal endpoints are presented, is any analytical verification undertaken?	Pass/fail	If analytical measurements are insufficient or absent to determine the average exposure concentration over the test, this criterion cannot be passed. An exception is when the test chemical has been demonstrated to be stable over the period of the test; however, expert judgment and careful consideration of test duration and other physicochemical properties are required (e.g., solubility, log K_{OW}).
Data quality assessment			
1	Is the biological endpoint stated and defined?	10, 5, 0	The studied endpoints and how they are quantified should be clearly stated and ideally their relevance validated (e.g., the method to quantify mortality, growth, or bleaching). Reduce to five points if endpoint descriptions are incomplete.
2	Are relevant validity criteria stated and met?	10, 5, 3, 0	For guideline or modified guideline studies, all validity criteria stated in the guideline must be met unless acknowledged in the modification. Reduce to five points if inferred, but data not presented or explicitly stated. For nonstandard studies, if validity criteria not mentioned, reduce score to five and apply general criteria or criteria used in a similar test if possible.[e] If some criteria met, but not all can be evaluated, score three points. If not possible to evaluate any criteria, no points awarded.
3	Is the test system defined and appropriate (flow-through, semistatic, or static conditions)?	5, 3, 2, 0	Flow-through conditions (award five points), semistatic conditions (award three points if renewal rate ≤48 h, two points if >48 h), static-conditions (award two points). The flow rate must also be stated. Consider stability of test substance and test organism requirements when scoring. With the use of expert judgment, five points can be awarded to a static test when it is the most appropriate option, as is the case for short-term coral reproductive assays (e.g., coral 5-h fertilization studies, see Gissi et al., 2017).
4	Is the test substance concentration maintained at ±20% throughout the exposure?	5, 3, 0	Award three points if analyte losses are acknowledged and strategies to maintain test substance concentration are applied. If analyte not monitored or only measured at the beginning of the test, no points awarded unless substance is demonstrated to be stable (award full points).
5	Is the test system appropriate for the test organism?	5, 3, 0	Consult guideline if appropriate and determine if test system is recommended for the test organism; reduce score if test system is permitted but not preferred. [f] Broadly, the score cannot exceed three points if adult coral is exposed in test medium that does not have flow (water movement) suitable for the test species (see Mitchelmore et al., 2021). Expert judgment is required to determine whether water movement is required based on the test species, life stage, and test length.

(*Continued*)

TABLE 1: (*Continued*)

		Score	Comment[b]
6	Is the biological effect level stated?	5, 3, 2, 0	In general, effect levels derived from regressions are awarded five points (e.g., ECx, LCx, no-effect concentration [NEC]). Three points are awarded if the maximum acceptable toxicant concentration (MATC) is reported (or can be calculated) and two points are awarded if only a NOEC or LOEC is reported. However, there are exceptions and expert judgment should be used to determine whether the test design is robust in terms of the effect level stated. For example, a NOEC could be awarded five points when a limit test is conducted or in a chronic test when the variability of control data indicates an EC10 cannot be reliably estimated. No points are awarded if biological effect level is not stated.
Is a parallel reference toxicant study conducted? Evaluate based on scenario:			
7a	If required in relevant guideline.	5, 0	Award five points if reference toxicant is included and effect was in the validity range stated in the guideline; award no points if outside validity range or not included. Award full points if reference toxicant studies are conducted periodically in accordance with the appropriate guideline. Award full points if reference toxicant not required by guideline and not included.
7b	If studying a wild organism.	5, 0	Award five points if a reference toxicant informed by a relevant guideline is included. If wild organism is nonstandard, evaluate under the next Scenario 7c.
7c	If studying a nonstandard organism/nonstandard endpoint.	5, 0	Award five points if reasoning is provided for reference toxicant selection (e.g., informed by relevant guideline or literature) and dose–response relationship is observed for the endpoint.
8	Is the test substance specifically identified (e.g., CAS no.) and source reported?	4, 2, 0	Award two points if only source is provided. Full points can be awarded if other details to clearly identify the test substance are provided.
9	Is the test substance purity reported?	4, 2, 0	If purity not reported, but analytical verification of test solutions undertaken, a score of two can be awarded.
10	Is the experiment appropriately replicated and not pseudoreplicated?	4, 2, 0	Consult appropriate guideline if applicable. If not applicable, expert judgment should be used determine whether the study design (replication) is suitable for the statistical model used. Reduce score to two if endpoints can be recalculated avoiding pseudoreplication.
11	Is a significant concentration–response relationship demonstrated?	4, 2, 0	Concentration–response should be demonstrated by at least five test concentrations (including control); reduce score if fewer concentrations. Reduce score if endpoints are reported but concentration-response relationship not demonstrated in figures/tables. If experiment is designed as a limit test and no effects are observed, full points can be awarded.
12	Was a suitable statistical method/model described to determine the toxicity?	3, 0	Statistical methods need to be fully reported and appropriate. For example, if hormeses occurs, is a suitable statistical model applied?
13	Was the significance level listed for NOEC/LOEC/MATC as ≤0.05 and for the NEC/LCx/ECx was an estimate of variability reported?	3, 0	No points awarded if not reported or if a statistical endpoint is not reported.
14	Was the exposure duration stated? Is the duration appropriate considering the species, life stage, endpoint, and reported effect concentration (acute or chronic)?	3, 1, 0	Consult relevant guideline for appropriate exposure duration. If nonstandard study, apply expert judgment. Reduce score if exposure duration reported, but not recommended/inappropriate. For coral, the duration for acute and chronic endpoints by life stage reported by Warne et al. (2018) were applied. No points awarded if duration not stated. If a recovery period is also relevant to the study but duration is not reported, no points awarded.
15	Is a suitable test concentration separation factor used?	3, 1, 0	Scaling factor of ≤3.2 is ideal, and 10 is considered the maximum. Close attention should be paid to spacing when deriving a NOEC. Award three points if <10, two points if 10, and no points if >10.
16	Do the test concentrations adequately bracket the biological endpoint?	3, 1,0	No points awarded if effect values are calculated by extrapolation rather than interpolation or if a statistical endpoint is not reported. If a LOEC is reported at the lowest test concentration, no points awarded. Reduce score if NOEC is observed at the highest test concentration, unless it is above test substance solubility (analytical confirmation required).
17	Are organisms appropriately acclimatized to test conditions?	3, 0	Award no points if acclimation period is not mentioned. For adult coral, sufficient healing time prior to toxicity testing should also be included.

(*Continued*)

TABLE 1: (*Continued*)

		Score	Comment[b]
18	Are organisms well described? (e.g., length, mass, age, strain, sex, etc.).	3, 2, 0	The descriptive parameters will change based on species. For coral in particular, a good description would include Latin name, origin, size, and whether individuals are from the same or genetically different colonies (reduce to two points if not reported).
19	Are the test vessels appropriate for the test substance?	3, 0	Assess based on the physicochemical properties of the test substance. For example, glass is preferred for many organic compounds, but this may not be the case for inorganics/metals. Test vessel choice can be justified by demonstrating sorption does not occur.
20	Are analytical methods described and appropriate QA/QC reported?	3, 1, 0	Reward one point if the method is reported without quality control/assurance criteria. These include limits of detection and quantification, recovery, method precision, and blank reporting.
Test medium parameters[g]			
21a	Dissolved oxygen	2, 1, 0	
21b	Temperature	2, 1, 0	
21c	pH	2, 1, 0	
21d	Salinity/conductivity	2, 1, 0	
21f	Species specific—include if specific parameters needed.	2, 1, 0	Include this criterion if there is a specific test medium parameter that needs to be measured to ensure test quality (e.g., iodine or nitrate).
22	If used, is solvent in the appropriate range?	2, 0	In general, solvent should not exceed 0.1 ml/L in accordance with OECD guidelines (OECD, 2019a, 2019b).
23	Is the solvent suitable for the test species?	2, 1, 0	Consult relevant guideline. It is an OECD acceptable solvent? Use of dimethyl sulfoxide (DMSO) reduces score if nonstandard species.

[a]The assessment consists of five screening questions covering critical relevance and reliability criteria followed by a 23-question data quality assessment. A study is not subject to the 23-question data quality assessment if relevance question RQ2 is failed or if two or more reliability screening questions are failed (i.e., RQ3–RQ5). A final reliability score is assigned based on the result from the screening assessment and the 23-question data quality assessment (see Table 2). Weighting can be adapted to suit the needs of the particular assessment and specific criteria can be excluded if not relevant. The comment column provides basic guidance for evaluation, but note this is not comprehensive and expert judgment should be exercised.

[b]During an evaluation, a comment can be included to convey the assessor's reasoning for giving a particular score to enhance transparency.

[c]When a guideline study (or modified or similar) is used, the required endpoints of that guideline must be reported. If different endpoints are reported but pertain to mortality, reproduction, or growth, the endpoint is considered relevant for the screening evaluation, but would not be preferred over standard endpoints if available. Deviation from the endpoints listed is permitted if the authors include a reference toxicant, evidence of repeatability, and a correlation to an established mortality, reproduction, or growth endpoint.

[d]The problem formulation could vary depending on the nature for the assessment, for example, a persistence, bioaccumulation, and toxicity (PBT) assessment, environmental compartmental risk assessment, or species-specific risk assessment that could have differing data requirements in different jurisdictions. A clear statement of the types of studies that are in scope prior to identifying studies to be assessed is needed. For more information and guidance on the development of problem formulations for environmental risk assessment please see USEPA (1998).

[e]Expert judgment is needed when no validity criteria are proposed. Consult a similar guideline when possible. When not possible, check for anomalies in controls (e.g., mortality, growth, effect) and between controls (e.g., solvent and negative).

[f]For example, corals often need flowing water to be maintained in a healthy condition so a static exposure would be inappropriate. Similarly, for the fish early-life stage toxicity test No. 210 (OECD, 2013), flow-through conditions are preferred, but in certain cases semistatic conditions can be acceptable.

[g]Award full points if measured and maintained throughout the test and reported. Reduce score to one if parameter is reported and inferred as maintained. Reduce score to one if only measured in dilution water or only at start of the test. Award score of zero if parameter range not appropriate for test species. For guideline studies check acceptable ranges of test medium parameters.

OECD = Organisation for Economic Co-operation and Development; MATC = maximum acceptable toxicant concentration; NEC = no-effect concentration; LC = lethal concentration; NOEC = no-observed-effect concentration; LOEC = lowest-observed-effect concentration; QA/QC = quality assurance/quality control.

relevant studies prior to conducting a lengthy reliability assessment. Furthermore, the three reliability questions were designed to cover key study design elements that, if two or more are failed, strongly indicate the study is unlikely to be reliable.

Studies that pass the screening assessment are then subject to a 23-question data quality assessment (Table 1). The assessment questions are largely based on the CRED approach, but aspects from Ecotoxicological Quality Evaluation System and Scoring Tool (Eco-QESST; Breton et al., 2009), the USEPA OPP (USEPA, 2011), and a method developed for marine tropical species (Gissi et al., 2016) also informed the data quality assessment (Figure 1). The CRED method permits the user to weigh criteria to help this process because not all criteria will impact study reliability equally. For example, a missing control is of greater concern for reliability than if a guideline is followed (Moermond et al., 2016). In the proposed tool, a numerical score is awarded for each question that is weighted based on the importance of each criterion to overall study reliability. Others have assigned weightings to CRED criteria (see Hartmann et al., 2017; Ingre-Khans et al., 2019), and these approaches, along with weighing recommendations of others (see Breton et al., 2009; USEPA, 2011), were used to derive the value of each data quality question (Figure 1 and Table 1).

The weightings can be altered, nonrelevant questions may be removed, or question guidance could be refined to address species-specific or test compound–specific considerations. For example, when a carrier solvent is not used, the related questions (22 and 23) are removed. This tool is still not "plug and play," as noted by Hartmann et al. (2017), and expert judgment will be required within the assessment. The key is that the expert judgment should be transparent, facilitated by including a comment explaining the score and thereby helping other assessors or readers to determine whether they agree with the outcome (Moermond et al., 2017).

Study reliability classification system

The study reliability classification and how it generally compares with Klimisch data reliability categories are shown in Figure 2. In the proposed method, a greater number of categories are included, described in Table 2. This provides the assessor with more options for studies that, according to Klimisch, are deemed *reliable with restrictions*. For example, studies that have one major issue are more easily differentiated from a study that contains several minor issues. These categories can be roughly translated to Klimisch scoring; however, exercising expert judgment is suggested if this is necessary. The Klimisch approach also has global regulatory implications, and it is our intent for the proposed method to be standalone from these processes. Our proposed method for awarding a final reliability score based on the screening and data quality assessment is presented in Table 2. The thresholds for each reliability category are flexible and can be altered if the assessor determines it would be more appropriate. For example, passing thresholds for categories R1 and R2 could be increased for assessing standard test method data (for example Organisation for Economic Co-operation and Development [OECD] guidelines [2004; 2012; 2013; 2019a]) when guidelines require many of the criteria covered in the tool to be met for a test to be deemed valid. Alternatively, an assessor could exercise expert judgment and over-ride the scoring system if needed; however, this decision should be justified.

For some contaminants it is expected that there will be limited data, which will have varying levels of reliability. Regardless, there is still a need to conduct the evaluation with available data. This could result in the use of data that may not meet the Klimisch et al. (1997) reliability standards. Alternatively, the proposed scoring method could provide reliability context for use in a preliminary or screening-level ERA with the goal of prioritizing critical data gaps that should be filled prior to conducting a higher tier ERA, or, for example, the method could be used to provide additional lines of evidence (e.g., R3 and R4 scoring studies; Table 2). This issue is not considered with current reliability methodologies. For example, Markovic et al. (2018) set out to generate a species sensitivity distribution (SSD) for nanoparticles and determined that the NanoCRED data reliability method proposed by Hartmann et al. (2017), would rule out all or much of the existing data on the grounds of insufficient quality. A method leading to a lower exclusion rate, ToxRTool (Schneider et al., 2009), was therefore applied to conduct a preliminary ERA. As an alternative solution to a similar problem, Gissi et al. (2016) presented SSDs that were based on varying levels of data quality. The data reliability categories presented could help accommodate this type of situation. When only R3 or lower data are available, a preliminary ERA can be conducted to provide recommendations for refinements through the collection of higher tier data or by filling key data gaps. However, in terms of regulatory ERA and decision-making, R1 and R2 studies from this method would be potentially suitable for this purpose, and lower scoring studies would be limited to supporting evidence.

The result is a systematic approach to evaluate the reliability of primarily nonstandard toxicity studies that are relevant to a particular problem formulation, as outlined in Figure 1. This is achieved by first developing the problem formulation, which includes relevant species, endpoints, and chemicals, identifying potentially suitable studies, and conducting a five-question screening assessment to identify relevant studies and studies likely to be reliable (e.g., score above R6), followed by a 23-question data quality assessment (Table 1), with the result informing the final reliability score (Table 2).

UV filter and coral toxicity case study

The goal of the case study was to evaluate the relevance and reliability of published coral toxicity studies for application to UV filter ERAs using the data quality evaluation process described. Following the process outlined in Figure 1, a problem formulation is first required. For this case study, the problem formulation is whether existing coral toxicity data are

relevant and reliable for use in a higher tier UV filter ERA in the United States with decision-making implications. The scope of the assessment is therefore limited to intact corals (i.e., whole-organism studies) of all life stages (e.g., larval or adult). Because coral are colonial organisms, coral fragments (nubbins) are appropriate for adult whole organisms. Coral species can originate from any region due to the current paucity in data. In the future, when more data are available, regional refinement of the scope to species relevant to the United States and its overseas territories (e.g., Indo-Pacific or Caribbean species) could be considered. Because coral are a colonial species, assays should be designed to consider normal variations in coral by covering differences within and between colonies of the same species (Shafir et al., 2003, 2007). Studies that include only genetically identical individuals (from the same colony) are therefore of reduced relevance. Ecologically relevant endpoints include growth (adults or coral recruits), mortality (e.g., sloughing of tissue to the point of skeletal exposure), reproduction (e.g., larval production, larval settlement, larval metamorphosis, and fertilization of gametes), and bleaching (Mitchelmore et al., 2021). Bleaching (expulsion of symbiotic algae) is a stress response that can lead to a coral's reduced ability to survive, grow, or reproduce and is thus of ecological relevance (Anthony et al., 2009; Douglas, 2003; Hughes et al., 2017, 2019). Bleaching can be quantified in numerous ways including algal cell counts, chlorophyll a content, or visually by examination of coral pigment, for example, the Coral Watch coral health chart (Summer et al., 2019). Nonecologically relevant endpoints are sublethal responses for which a clear link to an ecologically relevant effect has yet to be demonstrated (e.g., morphological changes, behavioral responses, and impacts on the photosynthetic abilities of the symbiont algae). These endpoints are of reduced relevance and are only suitable for a preliminary ERA (in the absence of ecologically relevant data) or as additional lines of evidence (see *Discussion, Screening assessment*).

Relevant study compounds include organic UV filters approved for use in the United States (see Mitchelmore et al., 2021). Exposures are limited to the water phase, and specific considerations for coral are covered in question five of the data quality assessment (Table 1). To achieve the highest score for question five, the exposure needs to occur in flowing or agitated water (see *Discussion, test species*). Question 21f is not included for this case study. Full screening and data quality assessments with comments for each study are given in the Supporting Information.

RESULTS

Description of the case studies included

In total, eight studies that investigated the ecotoxicological effects of UV filters on coral were identified for assessment in the case study (Tables 3 and 4). A summary of the physicochemical properties of the UV filters studied is provided in Table 3, and a brief summary of the ecotoxicological investigations is presented in Table 4. The scope was limited to coral studies due to the hypothesis that this taxon is uniquely sensitive to UV filter exposure and is therefore important to consider within ERAs (Danovaro et al., 2008; Downs et al., 2016). On the other hand, a recent review has challenged this hypothesis, but more work is required to validate and standardize coral testing prior to drawing conclusions on the relative sensitivity of corals in comparison with other standard test species in terms of UV filter exposure (Pawlowski et al., 2021).

In all studies, exposed corals were hard coral (reef-building), with the exception of *Xenia* sp.; a soft coral studied by McCoshum et al. (2016). All coral species studied maintain a symbiotic relationship with algae (dinoflagellates). The studies included cover both acute tests ranging from 24 h to 14 days (Danovaro et al., 2008; Downs et al., 2016; He et al., 2019a, 2019b; McCoshum et al., 2016; Stien et al., 2019) and chronic tests ranging from 35 to 41 days (Fel et al., 2019; Wijgerde et al., 2020). Researchers studied adult coral fragments (nubbins) and/or larvae (planula) collected from wild-caught or laboratory-cultured organisms (Table 4). The endpoints studied were varied, including mortality, deformity, larval settlement, bleaching, algal density, polyp retraction, growth reduction, photosynthetic efficiency, and metabolomic changes. In addition, one study reported identical endpoints under dark and light conditions to demonstrate the potential phototoxicity of the UV filter (Downs et al., 2016). The He et al. (2019b) study was split into two evaluations because the larvae test system and design was significantly different from those of adults. The in vitro cell line (calcioblast) toxicity data reported by Downs et al. (2016) were not included in the case study because they are beyond the scope of assessment (i.e., not whole organism). The validity of cell lines as a surrogate for whole-coral toxicity is uncertain, as discussed in detail by Mitchelmore et al. (2021). All but one study, that of Danovaro et al. (2008), was published within the past 5 years, indicating that this is a growing research field still in the early stages of development.

Screening assessment

The results of the screening assessment are summarized in Table 5. Two of the studies failed RQ1 because only nonecologically relevant endpoints for ERA were reported (Fel et al., 2019; Stien et al., 2019). Other studies did report nonecologically relevant endpoints, but passed RQ1 because an ecologically relevant endpoint was also included. By failing RQ1, a study cannot receive a reliability score of higher than R3 (Table 2). This is because ecologically relevant endpoints are needed for regulatory or higher tier ERA. Seven of the eight studies were determined to be relevant for the data quality assessment by passing RQ2 (Table 5). McCoshum et al. (2016) failed RQ2 because a sunscreen formulation was studied, and the UV filters within the sunscreen formulation were not quantitatively characterized or tested individually. Thus it cannot be determined whether any effect observed is the result of a single UV filter, a mixture of UV filters, or another ingredient in the formulation. Single-component toxicity data are prioritized over mixture toxicity as the toxicity of a mixture should ideally be calculated from the toxicity of individual

TABLE 2: Descriptions of the reliability scores awarded to studies based on their results from the screening and data quality assessment (see Table 1)

Reliability score	Screening evaluation	Data quality score	Description	ERA Interpretation
R1	Pass RQ1–RQ5	≥80%	This study is well designed and of high quality. No significant issues identified that reduce the reliability.	Potentially suitable for regulatory decision-making/higher tier ERA.
R2	Pass RQ1–RQ5	≥70%–79%	A well-designed and executed study with minor limitations that somewhat reduce the reliability of the results.	Potentially suitable for regulatory decision-making/higher-tier ERA (secondary to R1 studies, if available).
R3	Pass RQ1–RQ5 Fail RQ1; pass RQ2–RQ5 Fail 1 of RQ3–RQ5[c]	≥60%–69% ≥60% ≥70%	The study design and/or execution contained many minor limitations or a major limitation that significantly reduces the reliability of the results.	Preliminary assessment only. Can serve as additional line of evidence, with limitations stated. Useful for prioritizing higher quality studies.
R4	Pass RQ1–RQ5 Fail RQ1; pass RQ2–RQ5 Fail 1 of RQ3–RQ5[c]	≥50%–59% ≥50%–59% ≥60%–69%	The study contains many limitations to the point where the results should be interpreted with caution.	Apply expert judgment to determine whether useful for preliminary assessment, but clearly state limitations. [b] Can serve as additional line of evidence.
R5	Pass RQ1–RQ5 Fail RQ1; pass RQ2–RQ5 Fail 1 of RQ1–RQ5[c]	<50% <50% ≥50%–59%	The study has major design flaws and/or is poorly executed and cannot be considered reliable.	Study not useful for preliminary assessment. Can be supporting evidence if result similar to higher scoring study.
R6	Fail 1 of RQ1–RQ5[c]	<50%	The study design and/or execution is unsuitable for ERA and the results are highly unreliable.	Disregard, study not reliable or useful for ERA (even as supporting evidence).
NA1	Fail RQ2	N/A	Study does not pass relevance screening. Data quality score not evaluated.	Disregard, study not useful for problem formulation.
NA2	Fail ≥2 of RQ3–RQ5	N/A	Study does not pass reliability screening. Data quality score not evaluated.	Disregard, study not reliable or useful for ERA.

[a]The scoring is meant to serve as a guide to help derive a transparent and consistent reliability score, but expert judgment and context should also be considered when awarding the final reliability score.
[b]These studies can be used at the preliminary assessment stage, but priority should be given to replacing with higher quality data.
[c]In this case RQ1 can either be passed or failed, but RQ2 must be passed. If RQ2 failed, award reliability score of NA1.
ERA = environmental risk assessment.

components (OECD, 2019b). Furthermore, McCoshum et al. (2016) only included nubbins (adults) from a single colony in their test design.

The three reliability screening questions were designed to address three key areas of a study: adequate controls (RQ3), a suitable number of concentrations to observe a dose–response depending on test design (RQ4), and analytical verification of the test chemical concentration (RQ5). The study of Fel et al. (2019) was the only one to pass all three reliability screening criteria (i.e., RQ3–RQ5). Downs et al. (2016) and He et al. (2019a, 2019b) each failed one reliability screening question, indicating that the highest reliability score achievable for these studies is R3 (Table 2). Wijgerde et al. (2020), McCoshum et al. (2016), and Danovaro et al. (2008) failed two of the reliability screening questions and therefore do not pass the screening assessment.

Wijgerde et al. (2020) and Stien et al. (2019) both failed RQ3 because they did not include a negative control. Adequate controls are essential to conducting a reliable ecotoxicity study (Harris et al., 2014). This is important because, without a negative control, there is increased potential for false negatives (type II errors; Weyman et al., 2012). Wijgerde et al. (2020) provide a potential example of this: 33% mortality of *Acropora tenuis* was observed in the solvent control. Without a negative control, it cannot be determined whether effects occurred due to test conditions or possibly the solvent chosen. Meanwhile, Danovaro et al. (2008) included both a solvent and negative control, but only for one test species in one of the two in situ test locations.

Two studies failed RQ4, because only a single test concentration was included (McCoshum et al., 2016; Wijgerde et al., 2020). Neither study was designed as a limit test (i.e., tested near solubility), and in both cases an effect was observed in the single concentration studied, preventing the calculation of a no-observable-effect concentration (NOEC). Reichelt-Brushett and Harrison (2005) had previously noted this issue with coral research in which observed ecotoxicological values could not be used for decision-making because only two study concentrations were included. Certain situations, such as a limit test (either acute or chronic) could include fewer treatments without observation of a dose–response, but still pass RQ4. This is an aspect in which expert judgment is critical because both the test design and the results in treatments and controls (e.g., significant effect, no effect, variability in control) need to be considered. For example, when one is calculating a NOEC, increasing the number of replicates at the expense of treatments is suitable to achieve sufficient statistical power (ECHA, 2008). On the other hand, if an effect is observed in the lowest treatment, a NOEC cannot be derived. The number of test treatments should be at least five according to OECD guidance (see OECD, 2004, 2012, 2013, 2019a). The reason for this is to both sufficiently bracket the endpoint and observe a significant dose–response relationship. To achieve this with fewer than five treatments is challenging, even with a range-finding test. When all these factors are considered together, they show why a higher number of treatments are favored.

The reliability screening question failed most often was RQ5, conducting analytical verification of test concentration and basing endpoints on measured concentrations, if appropriate. Four studies conducted no analytical monitoring (Danovaro et al., 2008; Downs et al., 2016; McCoshum et al., 2016; Stien et al., 2019), whereas a further two did conduct monitoring but inappropriately reported endpoints based on

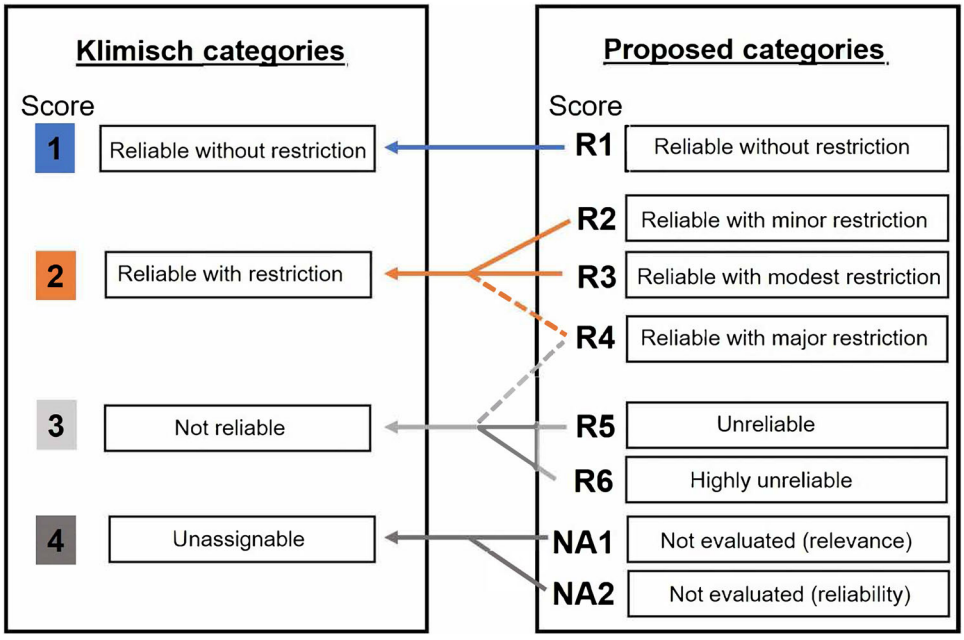

FIGURE 2: The data reliability scores used in the present study (right) and how they roughly compare with Klimisch categories (left; Klimisch et al., 1997). In the case of R4, depending on the nature of criteria that are considered unreliable, a comparable Klimisch category score of 2 or 3 could be appropriate. The proposed categories in our study are standalone and not intended to be interpreted in terms of Klimisch categories. The scoring approach we provide is meant to serve as a guideline; expert judgment and context should always be considered when a final score is awarded.

nominal exposures (He et al., 2019a, 2019b). Without measuring the concentration of UV filters in the test system, there is no way to determine what the coral was actually exposed to, which can lead to under- or overestimates of toxicity (Harris et al., 2014). Turner and Renegar (2017) observed similar issues in a review of coral toxicity studies with petroleum hydrocarbons and stated that the usefulness of a toxicity study is to determine a threshold concentration that can be compared with concentrations observed in the environment to inform chemical management. Without an actual measure of exposure, this cannot be achieved. Overall, the screening evaluation presented in Table 5 indicates that only four of the eight studies would be eligible to go through the subsequent data quality assessment and receive a final reliability score.

Data quality assessment

All studies were run through the data quality assessment regardless of their screening score as a proof of concept (Table 1), and the results are presented in Table 6. The larval settlement assay conducted by He et al. (2019b) received the highest data quality score, 72%. This was followed by the Fel et al. (2019) long-term study on adults (68%). The two He et al. (2019a, 2019b) adult assays were similar, 64% and 62%, with the difference between these scores explained by the concentration of solvent used. He et al. (2019b) exceeded the recommended maximum concentration of 0.1 ml/L (OECD, 2019b). In contrast to the larval assay conducted by the same authors (He et al., 2019b), the He et al. (2019a, 2019b) adult studies received lower scores because dose–response relationships were not observed. This also prevented the calculation of appropriate endpoints for ERA (e.g., median effect or lethal concentrations [EC(LC)50]), suggesting that the dosing range for certain UV filters (e.g., BP3, BP4) needed to be refined (He et al., 2019b). Conversely, the reported test concentration far exceeded solubility for EHMC, octocrylene, and benzophenone-8 (BP8), which suggests an inappropriate test design (e.g., a limit test may have been a more suitable option). Furthermore, the He et al. (2019a, 2019b) coral adult studies were also pseudoreplicated, because individual experimental units (coral nubbins) were exposed in the same treatment bottles. When these factors were taken together, the result was an approximately 10% lower data quality assessment score for the adult assays in comparison with the larvae settlement assay (Table 6; see the Supporting Information, Tables S1–S8 for detailed assessments of each study).

Generally speaking, the studies that failed the screening assessment also had the lowest data quality scores. Wijgerde et al. (2020) is the exception, scoring higher than Downs et al. (2016), yet failing the reliability screening. This is largely due to the comprehensive characterization and suitability of the test system for coral and appropriate characterization of the test chemical (BP3) reported by Wijgerde et al. (2020). The major issues with the Wijgerde et al. (2020) study (lack of controls and a single test concentration) were addressed in the screening assessment and do not contribute to the data quality score reported in Table 6. McCoshum et al. (2016) had the lowest score (30%), receiving points for only 8 of the 21 questions evaluated in our study.

Data reliability scores

The final data reliability scores reported in Table 6 consider the screening results (Table 5) and the data quality scores, also presented in Table 6. This score will inform how the data can be used for ERA, whether they are of potential regulatory standard (e.g., used for decision-making), suitable for preliminary ERA, suitable only as supporting evidence, or not suitable due to lack of relevance and/or reliability (see Table 2). No study was classified as R1 or R2, indicating that none of the studies were of potentially suitable quality for a regulatory ERA, beyond supporting evidence. Four studies were determined to be suitable for preliminary or screening-level ERA by scoring either R3 or R4. The larval settlement assay reported by He et al. (2019b) received the highest reliability score, R3, which is indicative of a study that is well designed but does have flaws that lower the reliability. These flaws were largely the reporting of endpoints based on nominal rather than measured concentrations, significant losses of the test chemical from the test system, and suitability of the test system for the coral. Fel et al. (2019) also scored R3; this was the only study to receive points for including a reference toxicant and to pass the three reliability screening questions (RQ3–RQ5). The adult assays conducted in the He et al. (2019a, 2019b) studies scored R4 due to the failure of a screening question and the less than 70% data quality scores achieved. The final study to receive a reliability score was that of Downs et al. (2016), which had the lowest score achievable (R6). Less than 50% was awarded in the data quality assessment, and a screening question was not passed (see Tables 5 and 6). The score indicates that this study is unreliable and not useful for preliminary ERA, even as supporting evidence. The remainder of the studies resulted in scores of NA1 and/or NA2 because they failed the screening assessment. Their low data quality scores (e.g., 30%, 33%, 38%, and 54%) support the conclusion from the screening assessment, that conducting lengthy data quality assessments on these studies is unnecessary because achieving reliability scores of R4 or higher is unlikely.

DISCUSSION

Screening assessment

Screening assessments successfully delineated between studies that were likely to receive a reliability score of R4 or higher. Based on these results, we conclude that the elements evaluated in the screening approach suitably streamline the reliability assessment process by focusing lengthy evaluations on studies that are likely to result in a score of R4 or higher, rather than R5 or R6 which are not overly useful for ERA. Interestingly, the study that failed the relevance screening assessment (score of NA1), McCoshum et al. (2016), also failed the reliability screening assessment (score of NA2). The value of

TABLE 3: Summary of the organic ultraviolet (UV) filters authorized for use as sunscreen ingredients in the United States[a] that have been evaluated in the peer-reviewed studies included in the case study

			Physicochemical properties[c]		
INCI name (INN)[b]	CAS no.	Abbreviation	Log K_{OW}	Solubility (μg/L)	Associated coral toxicity studies
Butyl methoxydibenzoylmethane (avobenzone)	70356-09-1	AVO	6.1	27	Fel et al. (2019) Danovaro et al. (2008) McCoshum et al. (2016)
Homosalate (homosalate)	118-56-9	HMS	6.34	400	Danovaro et al. (2008) McCoshum et al. (2016)
Ethylhexyl methoxycinnamate (octinoxate)	83834-59-7; 5466-77-3	EHMC	6	51	Danovaro et al. (2008) He et al. (2019a)
Ethylhexyl salicylate (octisalate)	118-60-5	EHS	6.36	500	Danovaro et al. (2008); McCoshum et al. (2016)
Octocrylene (octocrylene)	6197-30-4	OC	6.1	40	Danovaro et al. (2008); Fel et al. (2019); He et al. (2019a); McCoshum et al. (2016); Stein et al. (2019)
Benzophenone-3 (oxybenzone)	131-57-5	BP3	3.45	6000	Downs et al. (2016); He et al. (2019b); McCoshum et al. (2016); Wijgerde et al. (2020)
Benzophenone-4 (sulisobenzone)	4065-45-6	BP4	0.52	3.0×10^8	He et al. (2019b)
Benzophenone-8 (dioxybenzone)	131-53-3	BP8	2.33	13	He et al. (2019b)

[a]See Mitchelmore et al. (2021).
[b]UV filters are identified by their International Nomenclature of Cosmetic Ingredients (INCI) name and their international nonproprietary name (INN).
[c]Experimental physicochemical properties were obtained from publicly available Registration, Evaluation, Authorization and Restriction of Chemicals technical registration dossiers maintained by the European Chemicals Agency (2021).

the upfront assessment of both reliability and relevance, similar to the Stepwise Information-Filtering Tool (SIFT; Beasley et al., 2015), is a key streamlining mechanism to identify appropriate studies for the problem formulation prior to conducting time-consuming data quality assessments.

The relevance assessment (RQ1) also found that non-ecologically relevant endpoints for ERA were commonly reported (see Table 7). These endpoints are sublethal indicators of stress (Nordborg et al., 2020), and a reproducible relationship to population-level ecological effects has yet to be demonstrated (Warne et al., 2018). Such endpoints are of reduced relevance for ERA, but are still useful to monitor because they can provide insights into the toxic mode of action. For example, photosynthetic efficiency (quantum yield), an endpoint reported by Fel et al. (2019), quantifies the impact on the photosynthesizing ability of symbiont algae. Significant reductions in quantum yield have been observed, particularly when coral are exposed to photosystem II inhibitors such as diuron (Jones & Kerswell, 2003). However, this response is variable among coral species and is not clearly correlated with bleaching or other ecologically relevant effects, despite being a precursor to these effects in some cases (Negri et al., 2005).

Downs et al. (2016) reported morphological changes in planulae (deformity) in response to BP3 exposure. Deformity has been shown to be a precursor for mortality in coral larvae (Epstein et al., 2000) and an indicator of sublethal toxicity in other cnidarians (Echols et al., 2016), but in terms of ecological relevance it is not favored over endpoints that directly relate to population-level effects such as mortality, larval settlement, or metamorphosis (Nordborg et al., 2021). Polyp retraction is a behavioral response, and the data do suggest that this is a sensitive response; for example, He et al. (2019a) observed more polyp retraction at lower UV filter concentrations than any other response (Table 7). Renegar and Turner (2021) also observed that polyp retraction progressed to tissue attenuation and eventually mortality in coral exposed to polycyclic aromatic hydrocarbons. These data indicate there should be more investigation into a potential link to ecologically relevant population effects; however, a direct relationship needs to be established before this response could be used in higher tier ERA or for decision-making as a validated ecologically relevant endpoint. Other researchers have also found that toxicological thresholds for coral are difficult to compare between studies due to variability in the methods used and the endpoints reported (Negri et al., 2018; Nordborg et al., 2018). Therefore, it would be useful to establish standardized endpoints that are comparable and reproducible for ERA (Gissi et al., 2017; Nordborg et al., 2021).

Data quality assessment

The discussion of data quality assessment is organized by study aspect including test setup and system, test species, test substance, results, and statistics, rather than numerical score. Questions refer to the data quality assessment presented in Table 1 and results from the assessment presented in Table 6.

Test setup. There is no standardized test for coral (e.g., OECD, International Organization for Standardization, or

TABLE 4: Brief description of coral toxicity studies examining the effects of ultraviolet (UV) filters evaluated as part of the case study

Study	Species	Duration	Study compounds	Dosing	Endpoints	Comment
Danovaro et al. (2008)[a]	*Acropora* sp., *A. pulchra*, *Stylophora pistillata* (adult)	Not reported	EHMC, BP3, AVO, OC, EHS, sunscreen formulation	10–100 µL/L	Bleaching rate, bleaching initiation, algal density	Exposed wild adult coral in bags of filtered seawater in situ. Proposed that UV filters promoted viral infection, possibly playing an important role in coral bleaching
Downs et al. (2016)	*S. pistillata* (planulae); coral cells multiple species[b]	8–24 h	BP3	22.8–228,000 µg/L	Mortality, deformity, DNA damage, chlorophyll fluorescence, coral cell mortality	Reported that BP3 induced ossification of planula (encasing planula entirely in own skeleton). Also reported BP3 was genotoxic and reduced chlorophyll fluorescence (NOECs) and derived an EC50 and LC50 for planulae deformity and mortality. Applied correction factor to coral cell toxicity data to represent planulae mortality
McCoshum et al. (2016)	*Xenia* spp. (adult)	72 h exposure, 28-day recovery	Sunscreen formulation (BP3, HMS, OC, EHS, AVO)	0.26 ml/L	Growth	Soft coral species exposed to a sunscreen formulation containing multiple UV filters and inactive ingredients. Reduced growth was observed
Fel et al. (2019)[a]	*S. pistillata* (adult)	35 days	OC, AVO	10–5000 µg/L[c]	Photosynthetic efficiency	Tried to identify whether UV filters affected coral symbionts similarly to the pesticide diuron. No adverse effects were observed up to UV filter solubility
He et al. (2019a)	*Seriotopora caliendrum*, *Pocillopora damicornis* (adult)	7 days	EHMC, OC, co-exposure of EHMC and OC, sunscreen formulation	0.1–1000 µg/L	Mortality, bleaching, polyp retraction, algal density	*S. caliendrum* found to be more sensitive to EHMC than *P. damicornis*. Mortality increased in the sunscreen formulation exposures
He et al. (2019b)[a]	*S. caliendrum*, *P. damicornis* (adult and larvae)	14 days (larvae); 7 days (adult)	BP3, BP4, BP8	0.1–1000 µg/L[d]	Mortality, bleaching, polyp retraction, algal density, larval settlement	Adults were found to be more sensitive to benzophenones than larvae. An EC50 for larval settlement was able to be calculated for BP8. The remainder of endpoints reported were either LOECs or NOECs
Stien et al. (2019)	*P. damicornis* (adult)	7 days	OC	5–1000 µg/L	Polyp retraction, metabolomic changes	Identified OC transformation products in coral tissue that were lipophilic. The metabolomic profile indicated significant changes at 50 µg/L OC, whereas visually, coral polyps closed at 300 µg/L. The metabolic changes were hypothesized to be linked to mitochondrial dysfunction
Wijgerde et al. (2020)	*S. pistillata*, *Acropora tenuis*	42 days	BP3	1 µg/L	Mortality, growth, algal density, photosynthetic yield	Studied the effect of temperature and BP3. The effect of temperature was significant for *A. tenuis*. Only minimal effects from BP3 exposure alone were observed for both species

[a]Other compounds were included in the study but are not included in this summary because it is limited to ultraviolet filters authorized for use in the United States (see Mitchelmore et al. 2021).
[b]Coral cells were collected from *Stylophora pistillata*, *Pocillopora damicornis*, *Acropora cervicornis*, *Monstasteae annularis*, *Monstasteae cavernosa*, *Porites astreoides*, *Porites divaricata*.
[c]OC dosing range 100–5000 µg/L; AVO dosing range 10–5000 µg/L.
[d]BP8 dosing in the larvae settlement definitive test with *S. caliendrum* was 10–1000 µg/L.
AVO = butyl methoxydibenzoylmethane; BP3 = benzophenone-3; BP4 = benzophenone-4; BP8 = benzophenone-8; EHMC = ethylhexyl methoxycinnamate; EHS = ethylhexyl salicylate; HMS = homosalate; OC = octocrylene.

TABLE 5: Screening assessments for case study of coral ultraviolet filter toxicity tests[a]

Study	Screening assessment (Pass/fail)					Result
	RQ1[d]	RQ2[e]	RQ3[f]	RQ4[g]	RQ5[h]	
Danovaro et al. (2008)	Pass	Pass	Fail	Pass	Fail	NA2
Downs et al. (2016)	Pass	Pass	Pass	Pass	Fail	Fail 1 of RQ3–RQ5
Fel et al. (2019)	Fail	Pass	Pass	Pass	Pass	Fail RQ1; Pass RQ2–RQ5
He et al. (2019a)	Pass	Pass	Pass	Pass	Fail	Fail 1 of RQ3–RQ5
He et al. (2019b)[i]	Pass	Pass	Pass	Pass	Fail	Fail 1 of RQ3–RQ5
He et al. (2019b)[j]	Pass	Pass	Pass	Pass	Fail	Fail 1 of RQ3–RQ5
McCoshum et al. (2016)	Pass	Fail	Pass	Fail	Fail	NA1/NA2
Stein et al. (2019)	Fail	Pass	Fail	Pass	Fail	NA2
Wijgerde et al. (2020)	Pass	Pass	Fail	Fail	Pass	NA2

[a]The screening result is combined with the data quality assessment result (see Table 1) to derive a final reliability score (see Table 2). A result of NA1[b] or NA2[c] indicates the screening assessment is failed and the study is not subject to the data quality assessment and will not receive a reliability score.
[b]NA1 indicates the study failed RQ2 and will not be subject to the data quality assessment because it is not relevant.
[c]NA2 indicates the study failed two or more of RQ3–RQ5 and will not be subject to the data quality assessment because it is highly likely it is not reliable.
[d]Is the endpoint ecologically relevant?
[e]Is the test organism relevant to the compartment, test compound and/or assessment?
[f]Was a negative control and solvent control (if necessary) at least duplicated?
[g]Are ≥4 treatment concentrations included (including control) or specifically designed as a limit test?
[h]Are endpoints based on measured concentrations if they deviate by ≥20% of the nominal concentration? If only nominal endpoints are presented, is any analytical verification undertaken?
[i]Two scores are awarded to He et al. (2019b); this score is for the adult assay.
[j]Two scores are awarded to He et al. (2019b); this score is for the larval settlement assay.

USEPA). However, significant efforts toward developing such protocols for larval fertilization, survival, and settlement assays (Gissi et al., 2017; Leigh-Smith et al., 2018; Negri et al., 2018; Reichelt-Brushett & Hudspith, 2016; Reichelt-Brushett & Harrison, 2000), as well as adult lethal and sublethal (coral condition) assays (Gissi et al., 2019; Hédouin et al., 2016; Renegar et al., 2017; Renegar & Turner, 2021; Shafir et al. 2003, 2007), can be found in the peer-reviewed literature. None of the studies could follow a guideline method, but existing efforts toward assay standardization in the peer-reviewed literature were largely not considered. An exception is the study of Downs et al. (2016), in which a modification of OECD test guideline 236 (2013; fish embryo acute test) was cited. A lack of test guidelines could explain the observed variability in test design, exposure conditions, endpoints evaluated, and level of experimental detail included in the studies. Expert judgment was required to establish appropriate basic validity criteria (Question 2). It was therefore necessary for the evaluation to become an iterative process to ensure consistency and reflect key knowledge gleaned throughout the process. In this way, it was possible to apply consistent expert judgment across the studies.

The fulfillment of test validity criteria (Question 2) is weighted heavily, similarly to the Eco-QESST method (Breton et al., 2009). Efforts to establish validity criteria for coral studies have been made (Summer et al., 2019), in particular, control responses for larval fertilization and larval settlement should exceed 80% and 70%, respectively (Gissi et al., 2017; Negri et al., 2018). However, these particular criteria were not relevant to the majority of studies assessed. This issue is further complicated by a lack of water quality requirements (e.g., suitable ranges of dissolved oxygen, temperature, salinity, nutrients, and elements) broadly applicable for hard coral (hermatypic scleractinian corals) to aid in the validity assessment of the exposure medium. Instead, three basic validity criteria applicable to any ecotoxicity study were assessed (Table 1), which eases the challenge for assessing validity in nonstandard tests and species (Moermond et al., 2016). The Wijgerde et al. (2020) study is a good example of why following these basic validity criteria is important. Substantial mortality was observed for *A. tenuis* in the solvent control (33%), which is beyond acceptable levels of control mortality in standard test guidelines (e.g., 10%–20% in OECD, 2004, 2013, 2012). In addition, Danovaro et al. (2008) observed 16% zooxanthellae release in the solvent control, which was higher than some experimental treatments with the test substance and the negative control, but statistical significance was not reported. McCoshum et al. (2016) did not report control mortality or growth in addition to initial polyp counts in treatment, which is important because growth (based on number of polyps) was the observed endpoint. Beyond the basic validity criteria assessed, it would be beneficial to further establish coral-specific criteria because validity criteria in an ecotoxicological assay is an important quality control measure. This is particularly true for nonstandard assays for which typical endpoint responses and exposure conditions are not well established. Rigorous reporting of control responses and inclusion of typical validity parameters (e.g., dissolved oxygen) would improve both the consistency of and confidence in future coral toxicity studies.

The test duration was highly variable among the studies, and it was not always clear whether the endpoint was acute or chronic (Question 16). For example, He et al. (2019b) calculated an EC50 for larval settlement from a 14-day test. This is a chronic reproductive endpoint, but an EC10 or no-effect concentration is not reported, which would be useful for ERA. Warne et al. (2018) reported that due to the variable life history

TABLE 6: Abbreviated data quality assessment for case study of ultraviolet filter toxicity tests in coral[a]

Data quality assessment result	Danovaro et al. (2008)	Downs et al. (2016)	Fel et al. (2019)	He et al. (2019a)	He et al. (2019b)[b]	He et al. (2019b)[c]	McCoshum et al. (2016)[d]	Stein et al. (2019)	Wijgerde et al. (2020)
1. Biological endpoint stated and defined?	10	10	10	10	10	10	10	5	10
2. Are relevant validity criteria stated and met?	3	3	5	5	5	5	0	0	0
3. Is the test system used defined (e.g., static conditions)?	2	3	2	2	2	2	2	3	5
4. Is the test substance concentration maintained ± 20%?	0	0	3	0	0	0	0	0	3
5. Is the test system appropriate for the test organism?	0	0	5	3	3	3	5	3	5
6. Biological effect stated?	0	5	3	2	2	5	0	0	0
7. Is a parallel reference toxicant study conducted?	0	0	5	0	0	0	0	0	0
8. Test substance identified and source reported?	4	0	2	4	4	4	4	4	4
9. Test substance purity reported?	0	0	2	4	4	4	0	4	2
10. Is the experiment appropriately replicated?	4	4	2	0	0	4	0	0	4
11. Significant dose–response relationship demonstrated?	0	2	0	0	0	4	0	0	0
12. Suitable statistical method/model used to determine toxicity?	0	3	3	3	3	0	0	0	0
13. Significance level/variability reported for statistical endpoint?	0	0	3	3	3	1	0	0	0
14. Is exposure duration stated and appropriate?	0	3	3	1	1	3	1	3	3
15. Is a suitable test concentration separation factor used?	3	1	1	1	1	3	0	3	0
16. Do test concentration adequately bracket the endpoint?	0	3	3	3	3	3	0	0	0
17. Are organisms appropriately acclimatized to test conditions?	0	0	3	3	3	2	3	3	3
18. Are organisms well described?	3	3	3	2	2	3	0	3	3
19. Test vessels appropriate for the test substance?	0	3	3	3	3	3	3	3	0
20. Are analytical methods described and QA/QC reported?	0	0	1	3	3	3	0	0	3
21. Test medium parameters (total out of 22a – f)	0	2	2	8	8	8	1	3	6
22. If used, is solvent in the appropriate range?	2	2	2	2	0	0	–	0	2
23. Is the solvent suitable for the test species?	2	1	2	2	2	2	–	1	1
Total	33%	48%	68%	64%	62%	72%	30%	38%	54%
Reliability score	NA2	R6	R3	R4	R4	R3	NA1/NA2	NA2	NA2

[a]The total possible data quality score is presented as a percentage out of a maximum score of 100. The reliability scores were awarded based on the scheme presented in Table 2. Full data quality assessments for each study can be found in the Supporting Information.
[b]Two scores are awarded to He et al. (2019b), this score is for the adult assay.
[c]Two scores are awarded to He et al. (2019b), this score is for the larval settlement assay.
[d]McCoshum et al. (2016) maximum score was out of 96 rather than 100 because questions 22 and 23 were not evaluated as solvent not used in the study.

strategies of invertebrate taxa, it was difficult to generally classify appropriate acute and chronic text durations. This could help explain why the durations of the coral acute and chronic tests were so variable among the studies, which was also identified by Mitchelmore et al. (2021). Warne et al. (2018) provide guidance on appropriate test durations of acute and chronic coral tests based on the life stage and endpoint, which we used to assess the studies (see Table 1, Question 14). Briefly, adult/juvenile chronic tests are 14 or more days, whereas acute tests are less than 14 days for all endpoints. Embryo/larvae acute test durations are less than 7 days, except for studying larval development/metamorphosis, which is less than 48 h. Conversely, embryo/larvae chronic test durations for larval development/metamorphosis assays are 48 h or more, whereas embryo fertilization studies can be very short, 1 h or more, despite being chronic assays.

Test species and test system. Determining the appropriateness of the test system (Question 5) was challenging to evaluate without known ranges of acceptability for exposure conditions for the species of coral studied. Basic water quality parameters (e.g., pH, salinity, dissolved oxygen, key nutrients) were largely unreported (Question 21), despite being required and even part of the validity criteria for most guideline studies (see OECD, 2004, 2012, 2013). The study of Wijgerde et al. (2020) is an exception: temperature, salinity, alkalinity, calcium, nitrate, and phosphate were measured and maintained. The authors stated that their laboratory conditions did not reflect an oligotrophic reef, but rather the ideal combination for keeping coral healthy for the duration of a laboratory exposure. Furthermore, large variability in the exposure medium was observed. Four studies used artificial seawater, and three studies used filtered seawater in their static exposures (Danovaro et al., 2008; He et al., 2019a, 2019b). Fel et al. (2019) used unfiltered seawater renewed weekly throughout their 35-day exposure, raising concerns over the variability in exposure conditions. Reporting of the average and range of water quality parameters during a test is not only ecotoxicological good practice, but can greatly aid in the interpretation of results when anomalies arise. For example, Hédouin et al. (2016) observed species-specific changes in coral toxicity due to small changes in temperature (~3 °C), highlighting the importance of recording and maintaining species-appropriate temperature conditions.

A range of light intensities was used across the studies evaluated. Three studies reported that the light intensity led to significant evaporation and that test vessels required reconstitution during the exposure (as noted in the concurrent planulae experiment conducted under the same conditions by Downs et al., 2014, 2016; He et al., 2019a, 2019b); light intensity was even lowered to reduce evaporation (He et al., 2019b). He et al. (2019b) posited that differences in exposure conditions, in particular light intensity, could explain the difference in BP3 toxicity they observed in comparison with Downs et al. (2016). Indeed, Downs et al. (2016) did observe differences in toxicity under light and dark conditions (see Table 6), which suggests that for UV filters in particular, light conditions could be an important factor for toxicity. To ensure adequate light for the test species and to identify potential relationships between light conditions and toxicity, we suggest better characterization, or even standardization of an acceptable range, of light exposure in terms of both spectrum and quantity in coral studies.

A final observation on the appropriateness of the test systems for adult coral was the presence or absence of flowing water (water movement), a point that was also raised in the recent Mitchelmore et al. (2021) review. Evidence compiled by Turner and Renegar (2017) demonstrated that corals have increased sensitivity to static conditions compared with flow-through conditions and suggested that differences in toxicity could be related to depletion of oxygen and accumulation of waste products. Indeed, many coral toxicity studies on metals and other contaminants include detailed descriptions of flowing (including flow-through) or agitated exposure systems (Gissi et al., 2019; Leigh-Smith et al., 2018; Negri et al., 2011, 2018; Renegar & Turner, 2021; Summer et al., 2019); such studies have suggested that stagnant exposure conditions are not ideal for maintaining healthy adult coral. Employment of a flow-through system would also mitigate the evaporation issues observed in the static and semistatic exposures (Downs et al., 2016; He et al., 2019a, 2019b) and permit the use of appropriate lighting. Evaporation impacts the test substance concentration and the stability of water quality parameters, and efforts should therefore be made to avoid it. Looking forward to future studies, it would be useful to identify a selection of representative and sensitive test species for each region inhabited by coral (e.g., Pacific, Meso-American, Indian Ocean, etc.) for which acceptable ranges for these parameters can be derived (e.g., light intensity and period, pH, dissolved oxygen, temperature, key nutrients) and to determine how these parameters can be maintained without degrading the quality of the test system (e.g., reducing evaporation).

The test acclimation period varied among the studies and between larvae and adults (Question 17). In two studies, wild-caught organisms were immediately exposed to the test substance, indicating that no acclimation period was undertaken (Danovaro et al., 2008; Downs et al., 2016). Coral collected from the wild were also used in three further studies (Fel et al., 2019; He et al., 2019a, 2019b), but these corals were acclimated or spawned in the laboratory setting for a suitable period of time before ecotoxicity testing. When adult corals are tested, it is necessary to cut fragments from the mother colony, and these fragments need time to heal and acclimatize prior to testing (Negri et al., 2011; Shafir et al., 2003). The length of the "healing period" varied by study (3–4 weeks for hard coral, 96 h for soft coral). There is currently no acclimation protocol for adult coral, but Gissi et al. (2019) suggested monitoring coral condition by observing the color and presence of skeletal growth at the base of the coral nubbins. Danovaro et al. (2008) included no acclimation period after cutting adult corals. On the other hand, much shorter acclimation periods are appropriate for larvae, due to the shorter time frame between key life history events (e.g., fertilization or settlement). For example, when studying the effect of manganese, Summer et al. (2019)

TABLE 7: Summary of the coral toxicity endpoints assessed in the case study presented by descending reliability score (see Table 5)[a]

Score		UV filter	Life stage	Species	Endpoint	Result (µg/L)	Major issues	Reference
Suitable for preliminary ERA (R3 and R4)	R3	AVO	Adult	SP	*NOEC (photosynthetic yield)*	*1000 (87)*[b]	Semistatic renewal, no statistically significant dose–response relationship, poorly defined test medium, photosynthetic yield not ecologically relevant	Fel et al. (2019)
		OC	Adult	SP	*NOEC (photosynthetic yield)*	*1000 (519)*[b]		
		BP-8	Larvae	SC	*EC50 (settlement)*	*530.1*	Static exposure, only nominal endpoints reported, significant analyte losses (all concentrations <LOD at end of test), no reference toxicant	He et al. (2019b)
	R4	EHMC	Adult	SC	LOEC (polyp retraction)	10	Static exposure, only nominal endpoints reported, significant analyte losses, no significant dose–response relationship, no reference toxicant, pseudoreplication, and issues with test medium evaporation	He et al. (2019a)
					LOEC (bleaching, mortality)	*1000*[c]		
					NOEC (AD)	*≥1000*[c]		
				PD	*LOEC (polyp retraction)*	*1000*[c]		
					NOEC (AD, bleaching, mortality)	*≥1000*[c]		
		OC	Adult	SC, PD	*LOEC (polyp retraction)*	*1000*		
				SC, PD	*NOEC (AD, bleaching, mortality)*	*≥1000*		
		BP-3	Adult	SC	LOEC (polyp retraction)	10	Static exposure, only nominal endpoints reported, significant analyte losses, no significant dose–response relationship, no reference toxicant, pseudoreplication, and issues with test medium evaporation	He et al. (2019b)
					LOEC (bleaching)	1000		
					NOEC (AD, mortality)	≥1000		
				PD	NOEC (AD, bleaching, mortality, PR)	≥1000		
			Larvae	SC	LOEC (bleaching, mortality)	1000		
				PD	NOEC (bleaching, mortality)	≥1000		
		BP-4	Adult	SC, PD	NOEC (AD, bleaching, mortality, PR)	≥1000		
			Larvae	SC, PD	NOEC (bleaching, mortality)	≥1000		
		BP-8	Adult	SC	LOEC (polyp retraction)	10		
					LOEC (AD, bleaching, mortality)	*100*[c]		
				PD	*LOEC (AD, bleaching, mortality, PR)*	*1000*		
			Larvae	SC	*LOEC (bleaching)*	*1000*[c]		
					LOEC (mortality)	*500*[c]		
				PD	*NOEC (bleaching, mortality)*	*≥1000*[c]		
Unreliable	R6	BP-3	Larvae	SP	LOEC (chlorophyll fluorescence reduction, light)[d]	2.28	Test compound and purity cannot be confirmed, no analytical monitoring, no reference toxicant, wild organisms exposed in artificial sea water without acclimation, inappropriate and poorly documented test system, relevant validity criteria not reported	Downs et al. (2016)
					NOEC (chlorophyll fluorescence reduction, dark)[d]	22.8		
					NOEC (DNA damage, light and dark conditions)	22.8		
					EC50 (deformity, light conditions)	49		

(*Continued*)

TABLE 7: (*Continued*)

Score		UV filter	Life stage	Species	Endpoint	Result (µg/L)	Major issues	Reference
					EC50 (deformity, dark conditions)	137		
					LC50 (mortality, light conditions)	139		
					LC50 (mortality, dark conditions)	799		
Not scored, failed screening	NA2	BP-3	Adult	SP, AT	LOEC (photosynthetic yield)	1 (0.06)[b]	Missing controls, inappropriate study design (single test concentration and not limit test), no reference toxicant, basic validity criteria cannot be evaluated	Wijgerde et al. (2020)
				SP, AT	NOEC (mortality)	≥1 (0.06)[b]		
				SP	NOEC (growth, AD)	≥1 (0.06)[b]		
		OC	Adult	PD	*NOEC (metabolic profile)*	*5*	Missing controls, no analytical monitoring, no significant dose–response reported, no statistical endpoint reported, no reference toxicant, basic validity criteria cannot be evaluated	Stien et al. (2019)
					LOEC (polyp retraction)	*300*		
		AVO EHMC EHS OC BP-3	Adult	*Acropora* sp.	Not calculable (bleaching, AD)	—	Missing controls, no analytical monitoring, no significant dose–response reported, treatment data not fully reported, no statistical endpoints derived	Danovaro et al. (2008)
		EHMC BP-3	Adult	AP	Not calculable (bleaching, AD)	—		
	NA1/ NA2	Mixture	Adult	*Xenia* sp.	LOEC (growth)	0.26 ml/L[e]	Inappropriate study design (single test concentration and not limit test), no analytical monitoring, no reference toxicant, basic validity criteria cannot be evaluated. Tested sunscreen formulation containing multiple UV filters without analytically quantifying them	McCoshum et al. (2016)

[a]Endpoints that are underlined are not ecologically relevant. Ecologically relevant endpoints pertain to mortality, growth, reproduction (e.g., fertilization, larval settlement), and bleaching. Non-ecologically relevant endpoints are considered to be behavioral (e.g., poylp retraction), photosynthesis-related (e.g., symbiont photosynthetic yield or respiration), biomarkers, gene expression, genotoxicity (e.g., DNA damage), cell line responses, and tissue swelling. Note that with further research and a clear demonstration of a direct link to an ecologically relevant endpoint, the ecologically relevant status could be updated. Endpoints reported in italics indicate that the concentration is above the solubility of the test compound (see Table 3). All effect concentrations are nominal unless otherwise stated. Note that no study derived endpoints potentially suitable for regulatory or higher tier ERA (i.e., a score of R1 or R2).

[b]Mean measured concentration.

[c]Analytical data from the study indicated the treatment concentration was below solubility at the end of the test due to analyte losses, but a mean measured exposure concentration was not reported to confirm the exposure concentration throughout the test.

[d]Downs et al. (2016) quantified chlorophyll fluorescence as an indication of bleaching; however, the quantification method was reported to be a gross estimation of bleaching because it was not compatible with the geometry of coral larvae (Downs et al., 2014) and is therefore considered nonrelevant.

[e]Reported the amount of sunscreen formula in the test treatment, but concentrations of individual UV filters in the formula not reported.

AD = algal density; AP = *Acropora pulchra*; AT = *Acropora tenuis*; AVO = butyl methoxydibenzoylmethane; BP3 = benzophenone-3; BP4 = benzophenone-4; BP8 = benzophenone-8; EC = effect concentration; EC50 = median effect concentration; EHMC = ethylhexyl methoxycinnamate; EHS = ethylhexyl salicylate; ERA = environmental risk assessment; LC50 = median lethal effect concentration; LOD = limit of detection; LOEC = lowest observable effect concentration; NOEC = no observable effect concentration; OC = octocrylene; PD = *Pocillopora damicornis*; PR = polyp retraction; SC = *Seriatopora caliendrum*, SP = *Stylophora pistillata*, UV = ultraviolet.

used a 15-min acclimation period for coral larvae. However, the parent colonies were maintained in a laboratory holding tank until larvae had spawned, rather than being directly collected from the wild and exposed in artificial seawater as Downs et al. (2016) reported.

Test substance. Clear identification and concentration/purity of the test substance are key criteria to fulfill, but three studies either did not receive points for this question or required expert judgment to evaluate it (Question 8). In the case of Downs et al. (2016), a Chemical Abstracts Service number is provided, but it describes benzophenone-2 rather than BP3. In the case of Fel et al. (2019), chemical structures are provided, but the source of the test substance is not. In five studies, the test compound purity was not reported (Question 9), although this was mitigated by analytical verification of test concentrations in two of the five studies (Fel et al., 2019; Wijgerde et al., 2020).

Many UV filters are highly hydrophobic and poorly soluble and can be classified as difficult to test compounds (OECD, 2019b). As a result, carrier solvents were used in all but one study to help disperse the UV filters into solution. Even so, Fel et al. (2019) commented that a "nonsolubilized" fraction of organic UV filters was still present despite using solvent and that the organic UV filters were difficult to analytically quantify. Two studies used excessive solvent in their tests (more than 0.1 ml/L; Question 22), and dimethyl sulfoxide (DMSO) was used as a solvent in three studies (Question 23; Downs et al., 2016; Stien et al., 2019; Wijgerde et al., 2020). The solvent DMSO is not considered acceptable (USEPA, 2011), because it can increase uptake of the test substance across cell membranes (Kais et al., 2013). These phenomena cannot be captured with a solvent control and therefore resulted in score reduction due to possible over- or underestimation of effects, particularly in chronic studies (Turner et al., 2012).

In addition to solubility challenges, UV filters are unstable in aqueous test systems (Mitchelmore et al., 2021). Question 4 was designed to identify when measured exposure concentrations were variable throughout the duration of the test, which provides an indication as to whether the test setup was appropriate for the test chemical, whereas RQ5 in the screening assessment determined whether any analytical monitoring was conducted and whether endpoints were inappropriately reported only as nominal (see Table 1). This is important because endpoints estimated from nominal values are of reduced reliability: it cannot be confirmed that test organisms were exposed to the correct test chemical and concentration (Harris & Sumpter, 2015). There can be exceptions to this requirement that call for expert judgment to evaluate (see Table 1); however, generally a lack of any analytical verification or inappropriate use of nominal concentrations (e.g., measured concentrations deviated by greater than 20% of the nominal concentration) will reduce reliability. For example, if RQ5 is failed, the study cannot exceed a reliability score of R3, making it potentially only suitable for preliminary or screening-level ERA.

Of the eight studies, only four included analytical monitoring even though many UV filters fell under the description of difficult to work with substances (OECD, 2019b). This is evidenced by the monitoring data from all four studies demonstrating significant losses of the test compound from the test system, with losses greatly exceeding 20% of the nominal concentration, indicating that endpoints should be based on measured concentrations rather than nominal. Wijgerde et al. (2020) reported a mean measured concentration of 0.06 μg/L, 6% of the nominal dose of BP3 (1 μg/L). He et al. (2019b) reported only 2% of the nominal dose of BP3 remaining at the end of the 7-day test. More concerningly, all test concentrations of BP8 in the definitive larval settlement assay were below the limit of detection on Day 14 (10–1000 μg/L nominal). He et al. (2019a) also observed significant losses for EHMC and octocrylene, for which only 2% and 24%–61%, respectively, remained in the test system after the 7-day exposure. Fel et al. (2019) observed significant losses of avobenzone and octocrylene, with only 8% and 52%, respectively, remaining from their initial 1000 μg/L nominal exposure. These significant loses highlight why, particularly for difficult to test compounds such as several UV filters, conducting analytical verification and reporting endpoints based on measured values are critical for accurate endpoint determination (Moermond et al., 2016). A flow-through test design with appropriate test vessels (e.g., glass) and adequate setup including test chemical equilibration and stability assessment would be an ideal strategy to mitigate organic UV filter concentration stability issues. Providing water flow would also address the adult coral husbandry concerns associated with exposure medium stability outlined earlier in the *Test species and test system* section.

Results and statistics. A dose–response relationship was observed in only two studies (Question 11). For an effect to be reliably demonstrated, a dose–response relationship must be shown (Molander et al., 2015; Sumpter et al., 2016). There are, however, exceptions, as noted for RQ4 (Table 1), such as limit tests. He et al. (2019b) derived an EC50 for larval settlement by including an initial range-finding test and as a result were able to fit a suitable model to the data and not extrapolate to determine the EC50. The other endpoints studied all resulted in either no effect or a lowest-observable-effect concentration (LOEC). Several LOECs corresponded to a 100% effect (e.g., mortality) at the highest concentration tested, whereas no effect was observed in the next highest test concentration. These results indicate that the dosing needs to be adjusted to observe a dose–response relationship (i.e., by conducting a definitive test). Conducting a range-finding study prior to the definitive test is useful for this purpose. Downs et al. (2016) was the only other study that derived statistical endpoints based on a dose–response curve.

Another consideration was appropriate replication (Question 10). Without a test guideline, this criterion was assessed through expert judgment. The minimum number of replicates used was assessed as three. Pseudoreplication was observed in four studies (He et al., 2019a, 2019b; McCoshum et al., 2016; Stien et al., 2019), a problem commonly observed in ecotoxicology that can affect the inferential statistics applied to the data (Krull et al., 2013), resulting in lowered reliability.

Finally, although not directly evaluated through the data quality assessment, the reporting of raw data was rarely complete. Limited additional data could be found in the supporting information, but it was often incomplete or consisted only of derived values (e.g., averages) rather than raw data and would not meet the ecotoxicological reporting standards suggested by Mebane et al. (2019).

Data reliability scores

When the final reliability scores are translated into Klimisch categories, the advantage of our proposed extended scoring system emerges. An R4 score in our approach would most often translate to an R3 (unreliable) in the Klimisch method (Figure 2). However, a paucity of suitable coral data for risk assessment is evident from our evaluation. We also identified significant differences in study data quality that might be missed using a Klimisch approach. For example, the R4 score for the He et al. (2019a, 2019b) studies with adults was compared with the He et al. (2019b) larval study, which received a higher reliability score of R3. Data from R3 studies are preferred to data from R4 studies (Table 2). Studies scoring R3 and R4 are of appropriate quality to use in a preliminary or screening-level ERA, for which they can be used in combination with their reliability score to conduct preliminary assessments and identify knowledge gaps, but will only provide supporting information in a regulatory or higher tier ERA. In addition, it was found that all the R3 and many of the R4 endpoints were above the solubility of the test compound (see Table 7). The remainder of the studies assessed are not suitable for higher tier or preliminary ERA because they are missing key aspects that resulted in a failing of the screening assessment (Danovaro et al., 2008; McCoshum et al., 2016; Stien et al., 2019; Wijgerde et al., 2020), or because they received a score of R6, indicating unreliability (Downs et al., 2016).

General discussion

Ecotoxicity testing on coral has been conducted for other chemicals, including diuron (Jones, 2005), hydrocarbons (Negri et al., 2016), and heavy metals (Reichelt-Brushett & Harrison, 2005). However, as an ecotoxicological test species generating endpoints for risk assessment, relatively little work has been conducted in comparison with other taxa (Reichelt-Brushett, 2012). One of the main goals of conducting a reliability assessment is to ensure that data used for regulatory and higher tier risk assessment are of a suitable quality and fit-for-purpose. More broadly, such an assessment provides a mechanism to assess and incorporate peer-reviewed data into a regulatory or higher tier ERA regardless of whether standard test species or protocols were followed. A challenge identified in our study is that it is more difficult to assess species from historically underrepresented taxa in ecotoxicology. This is because established animal husbandry, testing protocols, and validity criteria are not widely known. At the current time, this is laboratory-specific knowledge that, although useful for judging the quality of work, is traditionally omitted from publication (Mebane et al., 2019). Knowledge in this field is rapidly growing, and thus methods applied today may be deemed unsuitable in the future. To address this limitation, the reliability scores resulting from our case study are nonbinding; instead, they provide a guide to interpret test data for use in ERA. As we learn more about best practice for a particular species, test system design, or analytical verification, we propose revisiting the assessments and updating them as needed. More broadly, authors of nonstandard tests in particular should utilize the supporting information to justify and fully describe test design and setup to promote repetition or further development of the work (Harris et al., 2014). Hanson et al. (2017) provide an informative overview of what information should be reported to enhance the value of a study that aims to enhance environmental protection. On the other hand, many of the criteria assessed are general to any ecotoxicological study (e.g., appropriate controls, dose–response relationships, appropriate solvents), and similarly to the findings of Ågerstrand et al. (2011), it was these general criteria that reduced the reliability of the peer-reviewed studies.

The weighting of questions in the data quality assessment, which is intended to help simplify and guide the process, could also be a limitation if applied too strictly. Moreover, considering the context of criteria that are not met is also important when one is determining the overall reliability score. It should be acknowledged that this is a subjective determination guided by the data quality questions and weighting, but ultimately facilitated through expert judgment. Transparency is the key to overcoming this limitation; a clear description of why a particular question is scored can provide a platform of discussion for those who disagree with a particular decision (Kase et al., 2016).

The results from our assessment indicate that none of the current studies on the toxicity of UV filters to coral are of potential regulatory or decision-making quality, R1 or R2 (Table 7). Despite this lack of reliability, a common theme of perceived regulatory importance was identified in the authors' recommendations and conclusions. This point was also recently identified and explored in the broader context of ecotoxicology by Brain and Hanson (2021). The authors' perceptions of the impact and value of a study is important because it influences how the scientific community and the public perceive the policy/regulatory importance of the results. The recommendations from the studies analyzed as part of the coral toxicity case study fell into two broad categories: (1) based on their findings, more research should be conducted on the subject, and (2) their findings can/should support coral reef management. Interestingly, the studies that were perceived to be evidence for policy-making/regulation were those determined to be least reliable for risk assessment. For example, studies with generally higher reliability scores called for further investigation and more data to perform refined risk assessment (Fel et al., 2019; He et al., 2019a, 2019b). On the other hand, McCoshum et al. (2016) suggested that, based on their findings, sunscreen usage should be limited by coral reef managers, and beach-goers should limit their use when swimming near coral reefs. Meanwhile, studies have suggested that their

data be directly fed into coral reef management action plans (Downs et al., 2016) or used to inform regulators seeking to develop coral reef protection measures (Danovaro et al., 2008). This is problematic because it encourages perception-driven policy-making based on individual studies versus unbiased science-based policy-making based on ERA. In our view, science-based policy-making is an effective approach to chemical management. A major process that underpins this process is ERA, which, when based on reliable science, can help decision-makers and risk managers chose the appropriate course of action for environmental protection (USEPA, 1998, 2003).

It is possible that the authors are not aware of the ERA approach and why reliable data are needed to support decision-making (Ågerstrand et al., 2017b; Breton et al., 2009). Meanwhile, regardless of reliability, the authors' recommendations have been used to influence policy-makers' priorities and decisions. As a result, UV filter bans (e.g., Bill SB-2571; State of Hawaii Senate, 2018) aiming to protect coral reefs are based on the most unreliable coral studies to date (Danovaro et al., 2008; Downs et al., 2016). This phenomenon is not limited to UV filters. For example, the environmental impact of microplastic is an exponentially growing field of research that has been largely built on the perception that microplastics (in particular cosmetic microbeads) are harmful to the environment. However, in the context of an ERA, this has yet to be demonstrated (Besseling et al., 2019; Burton, 2017). Policy-making efforts globally to reduce the environmental burden of primary microplastics (i.e., microbeads) have been undertaken despite their almost inconsequential fraction of the plastics found in the environment (Burns & Boxall, 2018).

A similar discourse has recently been observed pertaining to the environmental safety of the pharmaceutical metformin. Niemuth et al. (2015) reported that environmentally relevant concentrations of the widely used active pharmaceutical ingredient could cause endocrine disruption in fish and should be added to the list of potential endocrine-disrupting compounds. In response, Sumpter et al. (2016) identified shortcomings in their methodology including one test concentration, the status of control fish, and the repeatability of their results, among other things. Klaper and Niemuth (2016) responded by stating their research was "hypothesis-generating," drawing on a point made by Collins and Tabak (2014), in that there is an "over interpretation of creative hypothesis-generating experiments, which are designed to uncover new avenues of inquiry rather than provide definitive proof for any single question." The scientific community would benefit greatly from authors' clearly stating their study is an exploratory investigation and as a result has limited regulatory impact (Burns & Davies, 2020).

Issues with data reliability are not limited to peer-reviewed data. During the development of the Ecological Threshold for Toxicological Concern (Eco-TTC) tool (Connors et al., 2019), the EnviroTox database was created by compiling data from various United States, European, and Japanese regulatory databases and peer-reviewed literature. Only 41% of the possible entries were included after being assessed for duplication, relevance, validity, and acceptability based on the SIFT methodology developed by Beasley et al. (2015). Poorly conducted studies can actually hinder or delay environmental protection efforts because they need to be repeated, undermining their intended purpose and utility (Hanson et al., 2017). Moreover, they can trigger costly investigations to refute a poorly executed study. Another reason why data reliability assessments are so valuable is that independent assessment determines whether the design and execution of a particular study could facilitate the generation of credible results. Although value judgments will still differ throughout scientific and public spheres, data reliability scores provide a mechanism for the discussion of environmental protection to be founded on credible data or a clear understanding of the limitations of unreliable data.

CONCLUSIONS

In the present study we assessed the reliability of currently available coral ecotoxicity studies for higher tier or regulatory ERA of UV filters. The reliability approach built on existing methods (e.g., CRED, SIFT) and expanded on the commonly used four-category reliability scoring system derived by Klimisch. Coral are a nonstandard test species, and UV filters are challenging test substances, making these studies an ideal assessment for the usability and suitability of the proposed data evaluation approach. The approach has been designed primarily to assess data pertaining to CPCP ingredients in nonstandard test species and methods, but can also be applied to other consumer chemicals. The results of the case study indicated that none of the coral studies published were reliable for higher tier risk assessment and decision-making purposes without other lines of evidence. Four of the studies were determined to be suitable for preliminary ERA purposes. If a preliminary ERA indicates unacceptable risk, the generation of higher quality data is needed prior to decision-making and concluding on risk. Poorly scoring studies were successfully identified in the screening process, providing evidence that the screening assessment prior to the intensive data quality assessment is broadly suitable. Although difficult to interpret from an ERA perspective, these studies have highlighted a significant need to generate higher quality coral toxicity data, potentially through the development of robust standard test protocols, to obtain reliable endpoints suitable for higher tier ERA and decision-making.

Supporting Information—The Supporting Information is available on the Wiley Online Library at https://10.1002/etc.5229.

Acknowledgments—The authors thank the Environmental Safety Committee (ESC) and the ESC chair, A. Carrao (Kao, USA), for their support of the present study. The authors also thank the reviewers for their helpful contributions to the manuscript. The present study was funded by the Personal Care Products Council (PCPC).

Disclaimer—E.E. Burns and I.A. Davies are employees of the Personal Care Products Council.

Author Contributions Statement—**Emily E. Burns:** Conceptualization; investigation; methodology; formal analysis; writing-original draft, review, editing. **Iain A. Davies:** Supervision; funding acquisition; writing–review, editing.

Data Availability Statement—Data, associated metadata, and calculation tools are available from the corresponding author (burnse@personalcarecouncil.org).

REFERENCES

Ågerstrand, M., Breitholtz, M., & Rudén, C. (2011). Comparison of four different methods for reliability evaluation of ecotoxicity data: A case study of non-standard test data used in environmental risk assessments of pharmaceutical substances. *Environmental Sciences Europe*, *23*(17), 1–15. https://doi.org/10.1186/2190-4715-23-17

Ågerstrand, M., Brenig, M., Führ, M., & Schenten, J. (2017a). Refining tools to bridge the gap between academia and chemical regulation: Perspectives for WikiREACH. *Environmental Sciences: Process and Impacts*, *19*, 1466–1473. https://doi.org/10.1039/c7em00422b

Ågerstrand, M., Sobek, A., Lilja, K., Linderoth, M., Wendt-Rasch, L., Wernersson, A. S., & Rudén, C. (2017b). An academic researcher's guide to increased impact on regulatory assessment of chemicals. *Environmental Sciences: Processes and Impacts*, *19*, 644–655. https://doi.org/10.1039/c7em00075h

Anthony, K. R. N., Hoogenboom, M. O., Maynard, J. A., Grottoli, A. G., & Middlebrook, R. (2009). Energetics approach to predicting mortality risk from environmental stress: A case study of coral bleaching. *Functional Ecology*, *23*, 539–550. https://doi.org/10.1111/j.1365-2435.2008.01531.x

Beasley, A., Belanger, S. E., & Otter, R. R. (2015). Stepwise Information-Filtering Tool (SIFT): A method for using risk assessment metadata in a nontraditional way. *Environmental Toxicology and Chemistry*, *34*(6), 1436–1442. https://doi.org/10.1002/etc.2955

Besseling, E., Redondo-Hasselerharm, P., Foekema, E. M., & Koelmans, A. A. (2019). Quantifying ecological risks of aquatic micro- and nanoplastic. *Critical Reviews in Environmental Science Technology*, *49*, 32–80. https://doi.org/10.1080/10643389.2018.1531688

Brain, R. A., & Hanson, M. L. (2021). The press sells newspapers, we should not sell ecotoxicology. *Environmental Toxicology and Chemistry*, *40*(5), 1239–1240. https://doi.org/10.1002/etc.5003

Breton, R. L., Gilron, G., Thompson, R., Rodney, S., & Teed, S. (2009). A new quality assurance system for the evaluation of ecotoxicity studies submitted under the new substances notification regulations in Canada. *Integrated Environmental Assessment and Management*, *5*(1), 127–137. https://doi.org/10.1897/IEAM_2008-026.1

Burns, E. E., & Boxall, A. B. A. (2018). Microplastics in the aquatic environment: Evidence for or against adverse impacts and major knowledge gaps. *Environmental Toxicology and Chemistry*, *37*(11), 2776–2796. https://doi.org/10.1002/etc.4268

Burns, E. E., Ciszar, S. A., Roush, K. S., & Davies, I. A. (2021). National scale down-the-drain environmental risk assessment of oxybenzone in the United States. *Integrated Environmental Assessment and Management*, *17*(5), 951–960. https://doi.org/10.1002/ieam.4430

Burns, E. E., & Davies, I. A. (2020). The toxicological effects of oxybenzone, an active ingredient in suncream personal care products, on prokaryotic alga *Arthrospira* sp. and eukaryotic alga *Chlorella* sp.: Methodological issues. *Aquatic Toxicology*, *226*, 105501. https://doi.org/10.1016/j.aquatox.2020.105501

Burton, G. A. (2017). Stressor exposures determine risk: So, why do fellow scientists continue to focus on superficial microplastics risk? *Environmental Science and Technology*, *51*, 13515–13516. https://doi.org/10.1021/acs.est.7b05463

Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature*, *505*, 612–613. https://doi.org/10.1038/505612a

Connors, K. A., Beasley, A., Barron, M. G., Belanger, S. E., Bonnell, M., Brill, J. L., de Zwart, D., Kienzler, A., Krailler, J., Otter, R., Phillips, J. L., & Embry, M. R. (2019). Creation of a curated aquatic toxicology database: EnviroTox. *Environmental Toxicology and Chemistry*, *38*(5), 1062–1073. https://doi.org/10.1002/etc.4382

Danovaro, R., Bongiorni, L., Corinaldesi, C., Giovannelli, D., Damiani, E., Astolfi, P., Greci, L., & Pusceddu, A. (2008). Sunscreens cause coral bleaching by promoting viral infections. *Environmental Health Perspectives*, *116*(4), 441–447. https://doi.org/10.1289/ehp.10966

Douglas, A. E. (2003). Coral bleaching—How and why? *Marine Pollution Bulletin*, *46*(4), 385–392. https://doi.org/10.1016/S0025-326X(03)00037-7

Downs, C. A., Kramarsky-Winter, E., Fauth, J. E., Segal, R., Bronstein, O., Jeger, R., Lichtenfeld, Y., Woodley, C. M., Pennington, P., Kushmaro, A., & Loya, Y. (2014). Toxicological effects of the sunscreen UV filter, benzophenone-2, on planulae and in vitro cells of the coral, *Stylophora pistillata*. *Ecotoxicology*, *23*(2), 175–191. https://doi.org/10.1007/s10646-013-1161-y

Downs, C. A., Kramarsky-Winter, E., Segal, R., Fauth, J., Knutson, S., Bronstein, O., Ciner, F. R., Jeger, R., Lichtenfeld, Y., Woodley, C. M., Pennington, P., Cadenas, K., Kushmaro, A., & Loya, Y. (2016). Toxicopathological effects of the sunscreen UV filter, oxybenzone (benzophenone-3), on coral planulae and cultured primary cells and its environmental contamination in Hawaii and the U.S. Virgin Islands. *Archives of Environmental Contamination and Toxicology*, *70*, 265–288. https://doi.org/10.1007/s00244-015-0227-7

Echols, B. S., Smith, A. J., Gardinali, P. R., & Rand, G. M. (2016). The use of ephyrae of a scyphozoan jellyfish, *Aurelia aurita*, in the aquatic toxicological assessment of Macondo oils from the Deepwater Horizon incident. *Chemosphere*, *144*, 1893–1900. https://doi.org/10.1016/j.chemosphere.2015.10.082

Epstein, N., Bak, R. P. M., & Rinkevich, B. (2000). Toxicity of third generation dispersants and dispersed Egyptian crude oil on Red Sea coral larvae. *Marine Pollution Bulletin*, *40*(6), 497–503. https://doi.org/10.1016/S0025-326X(99)00232-5

European Chemicals Agency. (2021). *Registered substances*. Retrieved September 2, 2021, from https://echa.europa.eu/information-on-chemicals/registered-substances

European Chemicals Agency. (2008). Chapter R.10: Characterisation of dose [concentration]-response for environment. In *Guidance on information requirements and chemical safety assessment*.

Fel, J. P., Lacherez, C., Bensetra, A., Mezzache, S., Béraud, E., Léonard, M., Allemand, D., & Ferrier-Pagès, C. (2019). Photochemical response of the scleractinian coral *Stylophora pistillata* to some sunscreen ingredients. *Coral Reefs*, *38*, 109–122. https://doi.org/10.1007/s00338-018-01759-4

Gissi, F., Reichelt-Brushett, A. J., Chariton, A. A., Stauber, J. L., Greenfield, P., Humphrey, C., Salmon, M., Stephenson, S. A., Cresswell, T., & Jolley, D. F. (2019). The effect of dissolved nickel and copper on the adult coral *Acropora muricata* and its microbiome. *Environmental Pollution*, *250*, 792–806. https://doi.org/10.1016/j.envpol.2019.04.030

Gissi, F., Stauber, J., Reichelt-Brushett, A., Harrison, P. L., & Jolley, D. F. (2017). Inhibition in fertilisation of coral gametes following exposure to nickel and copper. *Ecotoxicology and Environmental Safety*, *145*, 32–41. https://doi.org/10.1016/j.ecoenv.2017.07.009

Gissi, F., Stauber, J. L., Binet, M. T., Golding, L. A., Adams, M. S., Schlekat, C. E., Garman, E. R., & Jolley, D. F. (2016). A review of nickel toxicity to marine and estuarine tropical biota with particular reference to the South East Asian and Melanesian region. *Environmental Pollution*, *218*, 1308–1323. https://doi.org/10.1016/j.envpol.2016.08.089

Hanson, M. L., Wolff, B. A., Green, J. W., Kivi, M., Panter, G. H., Warne, M. S. J., Ågerstrand, M., & Sumpter, J. P. (2017). How we can make ecotoxicology more valuable to environmental protection. *Science of the Total Environment*, *578*, 228–235. https://doi.org/10.1016/j.scitotenv.2016.07.160

Harris, C. A., Scott, A. P., Johnson, A. C., Panter, G. H., Sheahan, D., Roberts, M., & Sumpter, J. P. (2014). Principles of sound ecotoxicology. *Environmental Science and Technology*, *48*, 3100–3111. https://doi.org/10.1021/es4047507

Harris, C. A., & Sumpter, J. P. (2015). Could the quality of published ecotoxicological research be better? *Environmental Science and Technology*, *49*, 9495–9496. https://doi.org/10.1021/acs.est.5b01465

Hartmann, N. B., Ågerstrand, M., Lützhøft, H. C. H., & Baun, A. (2017). NanoCRED: A transparent framework to assess the regulatory adequacy of ecotoxicity data for nanomaterials—Relevance and reliability revisited. *NanoImpact*, *6*, 81–89. https://doi.org/10.1016/j.impact.2017.03.004

He, T., Tsui, M. M. P., Tan, C. J., Ma, C. Y., Yiu, S. K. F., Wang, L. H., Chen, T. H., Fan, T. Y., Lam, P. K. S., & Murphy, M. B. (2019a). Toxicological effects of two organic ultraviolet filters and a related commercial

sunscreen product in adult corals. *Environmental Pollution*, *245*, 462–471. https://doi.org/10.1016/j.envpol.2018.11.029

He, T., Tsui, M. M. P., Tan, C. J., Ng, K. Y., Guo, F. W., Wang, L. H., Chen, T. H., Fan, T. Y., Lam, P. K. S., & Murphy, M. B. (2019b). Comparative toxicities of four benzophenone ultraviolet filters to two life stages of two coral species. *Science of the Total Environment*, *651*, 2391–2399. https://doi.org/10.1016/j.scitotenv.2018.10.148

Hédouin, L. S., Wolf, R. E., Phillips, J., & Gates, R. D. (2016). Improving the ecological relevance of toxicity tests on scleractinian corals: Influence of season, life stage, and seawater temperature. *Environmental Pollution*, *213*, 240–253. https://doi.org/10.1016/j.envpol.2016.01.086

Hughes, T. P., Kerry, J. T., Alvarez-Noriega, M., Alvarez-Romero, J. G., Anderson, K. D., Baird, A. H., Babcock, R. C., Beger, M., Bellwood, D. R., Berkelmans, R., Bridge, T. C., Butler, I. R., Byrne, M., Cantin, N. E., Comeau, S., Connolly, S. R., Cumming, G. S., Dalton, S. J., Diaz-Pulido, G., … Wilson, S. K. (2017). Global warming and recurrent mass bleaching of corals. *Nature*, *543*, 373–377. https://doi.org/10.1038/nature21707

Hughes, T. P., Kerry, J. T., Baird, A. H., Connolly, S. R., Chase, T. J., Dietzel, A., Hill, T., Hoey, A. S., Hoogenboom, M. O., Jacobson, M., Kerswell, A., Madin, J. S., Mieog, A., Paley, A. S., Pratchett, M. S., Torda, G., & Woods, R. M. (2019). Global warming impairs stock—Recruitment dynamics of corals. *Nature*, *568*, 387–390. https://doi.org/10.1038/s41586-019-1081-y

Ingre-Khans, E., Ågerstrand, M., Rudén, C., & Beronius, A. (2019). Improving structure and transparency in reliability evaluations of data under REACH: Suggestions for a systematic method. *Human and Ecological Risk Assessment*, *26*(1), 212–241. https://doi.org/10.1080/10807039.2018.1504275

Institute of Medicine. (2013). *Environmental decisions in the face of uncertainty*. The National Academies Press. https://doi.org/10.17226/12568

Jones, R. (2005). The ecotoxicological effects of Photosystem II herbicides on corals. *Marine Pollution Bulletin*, *51*, 495–506. https://doi.org/10.1016/j.marpolbul.2005.06.027

Jones, R. J., & Kerswell, A. P. (2003). Phytotoxicity of Photosystem II (PSII) herbicides to coral. *Marine Ecology Progress Series*, *261*, 149–159. https://doi.org/10.3354/meps261149

Kais, B., Schneider, K. E., Keiter, S., Henn, K., Ackermann, C., & Braunbeck, T. (2013). DMSO modifies the permeability of the zebrafish (*Danio rerio*) chorion—Implications for the fish embryo test (FET). *Aquatic Toxicology*, *140–141*, 229–238. https://doi.org/10.1016/j.aquatox.2013.05.022

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 Sciences Europe*, *28*, 7. https://doi.org/10.1186/s12302-016-0073-x

Klaper, R. D., & Niemuth, N. J. (2016). On the unexpected reproductive impacts of metformin: A need for support and new directions for the evaluation of the impacts of pharmaceuticals in the environment. *Chemosphere*, *165*, 570–574. https://doi.org/10.1016/j.chemosphere.2016.08.048

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.

Krull, M., Barros, F., & Newman, M. (2013). Pseudoreplication in ecotoxicology. *Integrated Environmental Assessment and Management*, *9*, 531–533. https://doi.org/10.1002/ieam.1440

Leigh-Smith, J., Reichelt-Brushett, A., & Rose, A. L. (2018). The characterization of iron (III) in seawater and related toxicity to early life stages of scleractinian corals. *Environmental Toxicology and Chemistry*, *37*(4), 1104–1114. https://doi.org/10.1002/etc.4043

Markovic, M., Kumar, A., Andjelkovic, I., Lath, S., Kirby, J. K., Losic, D., Batley, G. E., & McLaughlin, M. J. (2018). Ecotoxicology of manufactured graphene oxide nanomaterials and derivation of preliminary guideline values for freshwater environments. *Environmental Toxicology and Chemistry*, *37*(5), 1340–1348. https://doi.org/10.1002/etc.4074

Martin, O. V., Adams, J., Beasley, A., Belanger, S., Breton, R. L., Brock, T. C. M., Buonsante, V. A., Burgos, M. G., Green, J., Guiney, P. D., Hall, T., Hanson, M., Harris, M. J., Henry, T. R., Huggett, D., Junghans, M., Laskowski, R., Maack, G., Moermond, C. T. A., … Ågerstrand, M. (2019). Improving environmental risk assessments of chemicals: Steps towards evidence-based ecotoxicology. *Environment International*, *128*, 210–217. https://doi.org/10.1016/j.envint.2019.04.053

McCoshum, S. M., Schlarb, A. M., & Baum, K. A. (2016). Direct and indirect effects of sunscreen exposure for reef biota. *Hydrobiologia*, *776*, 139–146. https://doi.org/10.1007/s10750-016-2746-2

Mebane, C. A., Sumpter, J. P., Fairbrother, A., Augspurger, T. P., Canfield, T. J., Goodfellow, W. L., Guiney, P. D., LeHuray, A., Maltby, L., Mayfield, D. B., McLaughlin, M. J., Ortego, L. S., Schlekat, T., Scroggins, R. P., & Verslycke, T. A. (2019). Scientific integrity issues in environmental toxicology and chemistry: Improving research reproducibility, credibility, and transparency. *Integrated Environmental Assessment and Management*, *15*(3), 320–344. https://doi.org/10.1002/ieam.4119

Mitchelmore, C. L., Burns, E. E., Conway, A., Heyes, A., & Davies, I. A. (2021). A critical review of organic ultraviolet filter exposure, hazard, and risk to corals. *Environmental Toxicology and Chemistry*, *40*(4), 967–988. https://doi.org/10.1002/etc.4948

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

Molander, L., Ågerstrand, M., Beronius, A., Hanberg, A., & Rudén, C. (2015). Science in risk assessment and policy (SciRAP): An online resource for evaluating and reporting in vivo (eco)toxicity studies. *Human and Ecological Risk Assessment*, *21*(3), 753–762. https://doi.org/10.1080/10807039.2014.928104

Negri, A. P., Brinkman, D. L., Flores, F., Botte, E. S., Jones, R. J., & Webster, N. S. (2016). Acute ecotoxicology of natural oil and gas condensate to coral reef larvae. *Science Reports*, *6*, 21153. https://doi.org/10.1038/srep21153

Negri, A. P., Flores, F., Röthig, T., & Uthicke, S. (2011). Herbicides increase the vulnerability of corals to rising sea surface temperature. *Limnology and Oceanography*, *56*(2), 471–485. https://doi.org/10.4319/lo.2011.56.2.0471

Negri, A. P., Luter, H. M., Fisher, R., Brinkman, D. L., & Irving, P. (2018). Comparative toxicity of five dispersants to coral larvae. *Science Reports*, *8*, 3034. https://doi.org/10.1038/s41598-018-20709-2

Negri, A. P., Vollhardt, C., Humphrey, C., Heyward, A., Jones, R., Eaglesham, G., & Fabricius, K. (2005). Effects of the herbicide diuron on the early life history stages of coral. *Marine Pollution Bulletin*, *51*, 370–383. https://doi.org/10.1016/j.marpolbul.2004.10.053

Niemuth, N. J., Jordan, R., Crago, J., Blanksma, C., Johnson, R., & Klaper, R. D. (2015). Metformin exposure at environmentally relevant concentrations causes potential endocrine disruption in adult male fish. *Environmental Toxicology and Chemistry*, *34*(2), 291–296. https://doi.org/10.1002/etc.2793

Nordborg, F. M., Brinkman, D. L., Ricardo, G. F., Agustí, S., & Negri, A. P. (2021). Comparative sensitivity of the early life stages of a coral to heavy fuel oil and UV radiation. *Science of the Total Environment*, *781*, 146676. https://doi.org/10.1016/j.scitotenv.2021.146676

Nordborg, F. M., Flores, F., Brinkman, D. L., Agustí, S., & Negri, A. P. (2018). Phototoxic effects of two common marine fuels on the settlement success of the coral *Acropora tenuis*. *Science Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-26972-7

Nordborg, F. M., Jones, R. J., Oelgemöller, M., & Negri, A. P. (2020). The effects of ultraviolet radiation and climate on oil toxicity to coral reef organisms—A review. *Science of the Total Environment*, *720*, 137486. https://doi.org/10.1016/j.scitotenv.2020.137486

National Research Council. (2009). *Science and decisions: Advancing risk assessment*. The National Academies Press. https://doi.org/10.17226/12209

Organisation for Economic Co-operation and Development. (2004). *Test No. 202: Daphnia sp. Acute immobilisation test. In OECD guidelines for the testing of chemicals, Section 2*. https://doi.org/10.1787/9789264069947-en

Organisation for Economic Co-operation and Development. (2012). *Test No. 211: Daphnia magna reproduction test. In OECD guidelines for the testing of chemicals, Section 2*. https://doi.org/10.1787/9789264185203-en

Organisation for Economic Co-operation and Development. (2013). *Test No. 236: Fish embryo acute toxicity (FET) test. In OECD guidelines for the testing of chemicals, Section 2*.

Organisation for Economic Co-operation and Development. (2019a). *Test No. 203: Fish, acute toxicity test. In OECD guidelines for the testing of chemicals, Section 2*. https://doi.org/10.1787/9789264069961-en

Organisation for Economic Co-operation and Development. (2019b). *Guidance document on aquatic toxicity testing of difficult substances and mixtures. In OECD series on testing and assessment*. https://doi.org/10.1787/0ed2f88e-en

Pawlowski, S., Moeller, M., Miller, I. B., Kellermann, M. Y., Schupp, P. J., & Petersen-thiery, M. (2021). UV filter used in sunscreens—A lack in current coral protection? *Integrated Environmental Assessment and Management*, *17*(5), 926–939. https://doi.org/10.1002/ieam.4454

Reichelt-Brushett, A. (2012). Risk assessment and ecotoxicology limitations and recommendations for ocean disposal of mine waste in the Coral Triangle. *Oceanography*, *25*(4), 40–51. https://doi.org/10.5670/oceanog.2012.66

Reichelt-Brushett, A., & Hudspith, M. (2016). The effects of metals of emerging concern on the fertilization success of gametes of the tropical scleractinian coral *Platygyra daedalea*. *Chemosphere*, *150*, 398–406. https://doi.org/10.1016/j.chemosphere.2016.02.048

Reichelt-Brushett, A. J., & Harrison, P. L. (2000). The effect of copper on the settlement success of larvae from the scleractinian coral *Acropora tenuis*. *Marine Pollution Bulletin*, *41*(7–12), 385–391. https://doi.org/10.1016/S0025-326X(00)00131-4

Reichelt-Brushett, A. J., & Harrison, P. L. (2005). The effect of selected trace metals on the fertilization success of several scleractinian coral species. *Coral Reefs*, *24*, 524–534. https://doi.org/10.1007/s00338-005-0013-5

Renegar, D. A., & Turner, N. R. (2021). Species sensitivity assessment of five Atlantic scleractinian coral species to 1-methylnaphthalene. *Science Reports*, *11*, 529. https://doi.org/10.1038/s41598-020-80055-0

Renegar, D. A., Turner, N. R., Riegl, B. M., Dodge, R. E., Knap, A. H., & Schuler, P. A. (2017). Acute and subacute toxicity of the polycyclic aromatic hydrocarbon 1-methylnaphthalene to the shallow-water coral *Porites divaricata*: Application of a novel exposure protocol. *Environmental Toxicology and Chemistry*, *36*(1), 212–219. https://doi.org/10.1002/etc.3530

Remengesau, T. E. Jr. (2018). Subject: Signing statement SB no. 10–135, SD1, HD1 (the Responsible Tourism Education Act of 2018). Retrieved February 17, 2021, from: http://extwprlegs1.fao.org/docs/pdf/pau181409.pdf

Rudén, C., Adams, J., Ågerstrand, M., Brock, T. C., Poulsen, V., Schleka, C. E., Wheeler, J. R., & Henry, T. R. (2017). Assessing the relevance of ecotoxicological studies for regulatory decision making. *Integrated Environmental Assessment and Management*, *13*(4), 652–663. https://doi.org/10.1002/ieam.1846

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., Hartung, T., & Hoffmann, S. (2009). "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicology Letters*, *189*, 138–144. https://doi.org/10.1016/j.toxlet.2009.05.013

Shafir, S., Van Rijn, J., & Rinkevich, B. (2003). The use of coral nubbins in coral reef ecotoxicology testing. *Biomolecular Engineering*, *20*(4–6), 401–406. https://doi.org/10.1016/S1389-0344(03)00062-5

Shafir, S., Van Rijn, J., & Rinkevich, B. (2007). Short and long term toxicity of crude oil and oil dispersants to two representative coral species. *Environmental Science and Technology*, *41*(15), 5571–5574. https://doi.org/10.1021/es0704582

State of Hawaii Senate. (2018). Details of Bill S.B. No. 2571, S.D. 2, H.D. 2, C.D. 1. A bill for an act. Hawaii State Capitol, Honolulu, HI, USA. Retrieved February 17, 2021, from: https://www.capitol.hawaii.gov/session2018/bills/SB2571_CD1_.HTM

Stien, D., Clergeaud, F., Rodrigues, A. M. S., Lebaron, K., Pillot, R., Romans, P., Fagervold, S., & Lebaron, P. (2019). Metabolomics reveal that octocrylene accumulates in *Pocillopora damicornis* tissues as fatty acid conjugates and triggers coral cell mitochondrial dysfunction. *Analytical Chemistry*, *91*, 990–995. https://doi.org/10.1021/acs.analchem.8b04187

Summer, K., Reichelt-Brushett, A., & Howe, P. (2019). Toxicity of manganese to various life stages of selected marine cnidarian species. *Ecotoxicology and Environmental Safety*, *167*, 83–94. https://doi.org/10.1016/j.ecoenv.2018.09.116

Sumpter, J. P., Scott, A. P., & Katsiadaki, I. (2016). Comments on Niemuth, N.J. and Klaper, R.D. 2015. Emerging wastewater contaminant metformin causes intersex and reduced fecundity in fish. *Chemosphere* 135, 38–45. *Chemosphere*, *165*, 566–569. https://doi.org/10.1016/j.chemosphere.2016.08.049

Turner, C., Sawle, A., Fenske, M., & Cossins, A. (2012). Implications of the solvent vehicles dimethylformamide and dimethylsulfoxide for establishing transcriptomic endpoints in the zebrafish embryo toxicity test. *Environmental Toxicology and Chemistry*, *31*(3), 593–604. https://doi.org/10.1002/etc.1718

Turner, N. R., & Renegar, D. A. (2017). Petroleum hydrocarbon toxicity to corals: A review. *Marine Pollution Bulletin*, *119*(2), 1–16. https://doi.org/10.1016/j.marpolbul.2017.04.050

U.S. Environmental Protection Agency. (1998). *Guidelines for ecological risk assessment* (EPA/630/R-95/002F).

U.S. Environmental Protection Agency. (2003). *A summary of general assessment factors for evaluating the quality of scientific and technical information* (EPA 100/B-03/001). Science Policy Council.

U.S. Environmental Protection Agency. (2011). *Evaluation guidelines for ecological toxicity data in the open literature. Procedures for screening, viewing, and using published open literature toxicity data in ecological risk assessments*. Office of Pesticide Programs.

U.S. Environmental Protection Agency. (2018). *Application of systematic review in TSCA risk evaluations* (EPA Document# 740-P1-8001). Office of Chemical Safety and Pollution Prevention.

Warne, M., Batley, G., van Dam, R., Chapman, J., Fox, D., Hickey, C., & Stauber, J. L. (2018). *Revised method for deriving Australian and New Zealand water quality guideline values for toxicants*. Prepared for the Revision of the Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Australian and New Zealand Governments and Australian State and Territory Governments.

Weyman, G. S., Rufli, H., Weltje, L., Salinas, E. R., & Hamitou, M. (2012). Aquatic toxicity tests with substances that are poorly soluble in water and consequence for environmental risk assessment. *Environmental Toxicology and Chemistry*, *31*(7), 1662–1669. https://doi.org/10.1002/etc.1856

Wijgerde, T., van Ballegooijen, M., Nijland, R., van der Loos, L., Kwadijk, C., Osinga, R., Murk, A., & Slijkerman, D. (2020). Adding insult to injury: Effects of chronic oxybenzone exposure and elevated temperature on two reef-building corals. *Science of the Total Environment*, *733*, 139030. https://doi.org/10.1016/j.scitotenv.2020.139030

Yamamoto, H., Tamura, I., Hirata, Y., Kato, J., Kagota, K., Katsuki, S., Yamamoto, A., Kagami, Y., & Tatarazako, N. (2011). Aquatic toxicity and ecological risk assessment of seven parabens: Individual and additive approach. *Science of the Total Environment*, *410–411*, 102–111. https://doi.org/10.1016/j.scitotenv.2011.09.040

Zhong, X., Downs, C. A., Che, X., Zhang, Z., Li, Y., Liu, B., Li, Q., Li, Y., & Gao, H. (2019). The toxicological effects of oxybenzone, an active ingredient in suncream personal care products, on prokaryotic alga *Arthrospira* sp. and eukaryotic alga *Chlorella* sp. *Aquatic Toxicology*, *216*, 105295. https://doi.org/10.1016/j.aquatox.2019.105295

