Received: 2 June 2021 | Revised: 20 July 2021 | Accepted: 29 July 2021

Critical Perspectives

The Weight-of-Evidence Approach and the Need for Greater International Acceptance of Its Use in Tackling Questions of Chemical Harm to the Environment

Andrew C. Johnson,[a,*] John P. Sumpter,[b] and Michael H. Depledge[c]

[a]UK Centre for Ecology and Hydrology, Wallingford, United Kingdom
[b]Institute of Environment, Health and Societies, College of Health and Life Sciences, Brunel University London, Uxbridge, United Kingdom
[c]European Centre for Environment and Human Health, University of Exeter Medical School, Knowledge Spa, Royal Cornwall Hospital, Truro, Cornwall, United Kingdom

Abstract: As we attempt to manage chemicals in the environment we need to be sure that our research efforts are being directed at the substances of greatest threat. All too often we focus on a chemical of concern and then cast around for evidence of its effects in an unstructured way. Risk assessment based on laboratory ecotoxicity studies, combined with field chemical measurements, can only take us so far. Uncertainty about the range and sufficiency of evidence required to take restorative action often puts policymakers in a difficult situation. We review this conundrum and reflect on how the "Hill criteria," used widely by epidemiologists, have been applied to a weight-of-evidence approach (a term sometimes used interchangeably with ecoepidemiology) to build a case for causation. While using a set of such criteria to address sites of local environmental distress has been embraced by the US Environmental Protection Agency, we urge a wider adoption of weight-of-evidence approaches by policymakers, regulators, and scientists worldwide. A simplified series of criteria is offered. Progress will require a sustained commitment to long-term wildlife and chemical monitoring over a sufficient geographic spread. Development of a comprehensive monitoring network, coupled with assembling evidence of harm in a structured manner, should be the foundation for protecting our ecosystems and human health. This will enable us to not only judge the success or failure of our efforts but also diagnose underlying causes. *Environ Toxicol Chem* 2021;40:2968–2977. © 2021 The Authors. *Environmental Toxicology and Chemistry* published by Wiley Periodicals LLC on behalf of SETAC.

Keywords: Weight of evidence; Chemicals; Environment; Populations; Risk

INTRODUCTION

Those working to protect the environment and humans from chemical pollution could consider themselves as having two main duties: 1) to ensure that the characteristics of new chemicals and likely exposure levels would not put biodiversity, ecosystem processes, and humans at risk (prospective risk assessment), and 2) to ensure that biodiversity, ecosystem processes, and human populations are not being damaged by the chemicals in current use (retrospective risk assessment or impact). Protecting the environment from exposure to unnecessary risk from new chemicals is a vital part of preserving living organisms. The ultimate aim for all those working in the field is that our competence and knowledge will lead eventually to a consensus to prevent harmful chemicals being marketed. Notwithstanding the often considerable amount of chemical safety and environmental fate data that industries have to provide for many jurisdictions for their products, fears about new types of effects and dangers from mixtures have led to a climate of perpetual uncertainty, if not fear, for the future of natural ecosystems (Bergstrom et al., 2021). Indeed, it is now considered by many that chemical pollution is one of the major drivers of biodiversity loss today (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, 2019), as well as damage to human health (Landrigan et al., 2018), and that more coherent approaches will be needed (Wang et al., 2021). Unfortunately, prospective analysis and risk assessment cannot guarantee identification of which chemicals might be harmful and which might not. Strategies such as quantitative structure–activity relationships and the ToxCast program in the United States (Dix et al., 2007) are attempting to detect chemicals with hazardous properties before they reach the

* Address correspondence to ajo@ceh.ac.uk
Published online 4 August 2021 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5184

environment, but to date the information that these approaches have provided has not been easy to translate into risk assessment (Villeneuve et al., 2019). We still have only a rudimentary understanding of how the combined effects of natural abiotic factors (e.g., temperature, salinity, oxygen levels) and biotic factors (nutritional state, reproductive condition) combine with chemical toxicity to reduce the Darwinian fitness of populations in nature. Furthermore, current regulatory practices are failing to prevent the presence of chemicals of unknown toxicity in ecosystems, as well as in humans, from being used and widely disseminated throughout the environment (Gold & Wagner, 2020; Washington et al., 2020).

Once chemicals are in routine use and likely to enter ecosystems, the classic and most common approach to assessing whether damage might be occurring is shown as "method A" in Figure 1. This illustrates the use of a combination of laboratory ecotoxicity studies with field measurements of the suspect chemical.

Yet questions persist over whether the observations from laboratory ecotoxicity studies, particularly regarding chronic toxicity and nonlethal endpoints, actually translate into damaging effects in populations in situ (Adams, 2003; Johnson & Sumpter, 2016; Munkittrick & McCarty, 1995; Suter et al., 1985). Less common are field tests, such as for pesticides in plots; but these suffer from the limited ability to assess impacts because of the short duration of the test (Joy et al., 2005). It may be that we underestimate impacts in the environment due to the different vulnerabilities of various species, of their life stages (juveniles, adults, etc.), and of the presence of multiple biotic and abiotic stressors (including mixtures of chemicals). We still face great difficulties in predicting the likely damaging consequences of repeated exposure to chemicals over the life course of each species. Alternatively, we may overestimate toxic effects because of compensating factors. The work of Fahlman et al. (2021) is a good example of where impacts demonstrated on fish in the laboratory (antidepressant effects on behavior) did not occur in the field because they were subsumed by wider environmental influences. Even when compensation does appear to occur, it may be difficult to interpret. For example, a population may persist because it becomes genetically resistant to the presence of a toxic chemical, but we may not wish to accept such a situation if the population passes its pollutant load on to predators or the resistant population is more susceptible to other toxic chemicals, rendering the ecosystem more vulnerable overall to future threats.

Unfortunately, there are also concerns that some of the research published about chemicals in the environment as used in "method A" could be misleading (Figure 1). These include growing concerns about a declining quality in published ecotoxicological work (Harris et al., 2014) and increasingly extravagant claims of significant damage (Hanson & Brain, 2020; Mebane et al., 2019). An example of some of these problems is the increased frequency of claims by scientists that the chemical exposure level used in their laboratory ecotoxicity tests was "environmentally relevant" when in fact it was not (Weltje & Sumpter, 2017). It is extremely unusual for the authors of papers reporting that a chemical causes adverse effects to a particular species in a laboratory study to replicate it or follow up that study by investigating whether those same effects are occurring in wild organisms exposed to the same chemical in the natural environment. Hence, it is often impossible to know if the chemical used in the laboratory studies is or is not a threat to free-living populations of that organism. Interpretation of the evidence provided by industry and academia can be contradictory and appear to reflect views belonging to two different ideological tribes, with the case of bisphenol A being an unfortunate example (Myers et al., 2009).

With the emergence of the precautionary principle, a philosophical underpinning exists for policymakers to restrict chemical use, even where the amount of evidence for harm is very limited (Gee, 2006; Wynne, 1992). Where the risks are high but uncertainty remains, applying the precautionary principle appears sensible. Nevertheless, it has been argued that the greater the social or economic impact of a restriction, the

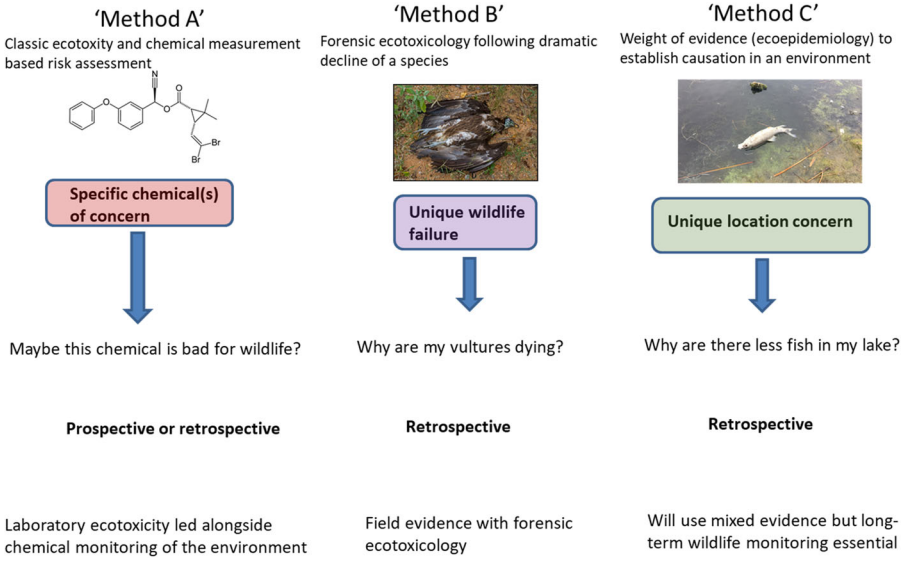

FIGURE 1: Examples of approaches used to establish whether or what chemical is causing harm or poses an unacceptable risk.

greater should be the expectation of coherence in the accumulated scientific evidence such as it is (Hill, 1965). The problem for society lies not with the precautionary principle but with the absence of an accepted minimum range and amount of evidence that is used to underpin such decision-making. Relying only on information from the route of "method A" (Figure 1) is a potentially weak foundation for calls to apply the precautionary principle.

The use of what might be called "forensic ecotoxicology," following field observations of an unexpected decline in a particular species, has had an excellent track record (described as "method B" in Figure 1) in establishing cause and effect. These include deductions on the significance of organochlorine pesticides, such as dichlorodiphenyltrichloroethane, as being responsible for egg shell thinning in predatory birds (Ratcliffe, 1970), the relationship between imposex in mollusks and tributyltin (Gibbs et al., 1991), then more recently between Asian vulture decline and diclofenac (Oaks et al., 2004). However, it is not clear if this approach alone has entirely explained the problems in eel populations (Geeraerts et al., 2011) or is suitable when multiple different groups of organisms are in decline.

Finally, we come to considering the formal structured approach to addressing an environmental problem, which has often been described as the "weight-of-evidence" approach (represented as "method C" in Figure 1). We have found it hard to differentiate the discussion and methods of ecoepidemiology from the weight-of-evidence approach. We use weight of evidence to refer to a structured approach for gathering multiple lines of evidence (Burton et al., 2002). In a review of the term, Weed (2005) recommended that authors always define the elements they are employing in a paper where the term weight of evidence is invoked. This confusion and past vagueness over terms have not helped the adoption of these approaches. Nevertheless, the inspiration for this approach of gathering multiple lines of evidence comes from epidemiology, most famously from the criteria listed in Hill (1965).

THE HILL CRITERIA AND THEIR APPLICATION

The field of chemical risks to the natural environment can benefit from reflecting on the work of Hill (1965) together with analysis of his "criteria," as reviewed by Susser (1991). This view led Fox (1991) to urge its adoption to complement the investigative approach reliant on field observations described as "ecoepidemiology" (Bro-Rasmussen & Lokke, 1984). It is worth pointing out that epidemiology deals with populations of one species—*Homo sapiens*. In applying it to multiple species in ecosystems, it has to be modified to consider ecological interactions and the incredibly diverse biology of species, so weight of evidence may be the more appropriate term. We are aware, however, that weight of evidence as a term is not ideal because it might be assumed to be referring to simply the amount of evidence supporting a hypothesis, which is fundamental to the philosophy of science (Popper, 1963). A better description might be "multiple lines of evidence," but it is perhaps too late to introduce a new term, so we will stick with weight of evidence.

Although Hill warned us against using his list as an infallible guide to causation, it is reasonable to assume that the more criteria which apply to the question at issue, the more confidence might be given as to whether causation exists. He offered the following criteria. 1) Strength: This where a distinct association can be distinguished. In epidemiology it might be described as where the incidence of a disease (harmful population effect) is clear. He gives the examples of the incidence of lung cancer in smokers being 10 times that of nonsmokers and that of John Snow's London residents having a 14 times greater death rate from cholera in 1855 because of a communal water pump drawing from sewage-contaminated water compared to their neighbors with a sewage-free water pump. 2) Consistency: Has the effect been observed to fit the same pattern by different people in different places, circumstances, and times? 3) Specificity: Here, damage is linked to the most exposed and vulnerable part of the population. Hill was considering unique exposures, such as for people working in a particular industry having adverse problems to a degree not seen in those in other environments. 4) Temporality: Here, harm is linked to a moment in time when exposure to a harmful agent began (and similarly would decline in time once the agent is withdrawn). 5) Biological gradient: This is the well-known exposure–response relationship where greater exposure might reasonably be expected to cause more damage. 6) Plausibility: This is where the relationship is biologically plausible. This is an interesting, somewhat subjective criterion, which, as Hill said, we cannot demand (because biological knowledge may be different tomorrow from what it is today). 7) Coherence: This is where different sources of evidence are brought together to complement the argument. Thus, linking the deaths from cholera to drinking water from a contaminated source is strengthened by also detecting *Vibrio* bacteria in the water. 8) Experiment: In this case Hill is asking does an intervention, such as reducing exposure through changing industrial practice, reduce the incidence of the linked disease? 9) Analogy: Hill suggests that we might be able to refer to similar agents (perhaps in molecular structure) which we know have caused serious problems in the past to warn us of danger.

We consider that an important part of the Hill criteria is the prominence, not to say dominance, of criteria that relate to evidence coming from the field rather than the laboratory. Some examples of the taking up of some, if not all, of these criteria by many proponents of ecoepidemiology/weight of evidence are shown in Table 1.

In a review of the utility of these different criteria for those involved in weight of evidence/ecoepidemiology–type studies by Collier (2003), it was found that some were less useful. Regarding specificity, few examples of highly stressor-specific symptoms could be found. Also, exposure–response relationships along a biological gradient were weaker in studies examining effects on communities, as opposed to studies of effects on individuals and populations.

TABLE 1: Examples of the Hill (1965) criteria picked out and used by a selection of authors

Strength of association	Consistency of observations	Specificity of association	Temporality (time order)	Biological gradient	Plausibility	Coherence	Experiment or intervention	Analogy
(Fox, 1991)	(Fox, 1991)	(Fox, 1991)	(Fox, 1991)	(Fox, 1991)	(Fox, 1991)			
(Gilbertson, 1997)	(Gilbertson, 1997)	(Gilbertson, 1997)	(Gilbertson, 1997)			(Gilbertson, 1997)		
(Ankley & Giesy, 1998)	(Cormier et al., 2003)	(Cormier et al., 2003)	(Ankley & Giesy, 1998) (Cormier et al., 2003)	(Cormier et al., 2003)	(Ankley & Giesy, 1998) (Cormier et al., 2003)	(Ankley & Giesy, 1998) (Cormier et al., 2003)	(Ankley & Giesy, 1998) (Cormier et al., 2003)	(Cormier et al., 2003)
(Collier, 2003)	(Collier, 2003)		(Collier, 2003)	(Collier, 2003)	(Collier, 2003)	(Collier, 2003)	(Collier, 2003)	
(Adams, 2003)	(Adams, 2003)	(Adams, 2003)	(Adams, 2003)	(Adams, 2003)	(Adams, 2003)	(Adams, 2003)	(Adams, 2003)	
(Moraes & Molander, 2004)	(Moraes & Molander, 2004)	(Moraes & Molander, 2004)	(Moraes & Molander, 2004)	(Moraes & Molander, 2004)	(Moraes & Molander, 2004)	(Moraes & Molander, 2004)		
(Suter et al., 2007)	(Suter et al., 2007)		(Suter et al., 2007)	(Suter et al., 2007)	(Suter et al., 2007)	(Suter et al., 2007)	(Suter et al., 2007)	(Suter et al., 2007)

Examples of where past and present environmental problems linked to chemicals might fit the Hill criteria are shown in Table 2.

The key references we used to help with our scoring were Matthiessen et al. (1995) for tributyltin and mollusks, Oaks et al. (2004) for diclofenac and vultures, Sumpter (2005) for estrogens and fish, Newton and Bogan (1974) for organochlorines and raptors, Woodcock et al. (2016) for bees and neonicotinoids, Yamamuro et al. (2019) for fish and neonicotinoids, Desforges et al. (2018) for polychlorinated biphenyls and killer whales, Geeraerts et al. (2011) for persistent organic pollutants and eels, and Johnson and Sumpter (2016) for pharmaceuticals and fish.

The review of significant chemical issues using the Hill criteria may not be a perfect fit with the topic of chemicals in the environment. The scoring provided in Table 2 could be seen as somewhat subjective, and others may have more or less confidence in the fit of these observations to the Hill criteria than we do ourselves. However, the exercise is illuminating in two ways: firstly, even chemical issues where we are confident of causation and where we have taken action do not score 9 out of 9; secondly, issues where we have not reached a settled conclusion would indeed score poorly despite many years of research. It is somewhat depressing that both Solomon et al. (2008) and Hayes et al. (2011) refer explicitly to the Hill criteria yet come to very different conclusions on the causality of atrazine and endocrine disruption in amphibians. The lesson here is that the criteria being referred to should be explained at the outset and the subsequent analysis as dispassionate and independent as possible. The assessment by Hayes et al. (2011) did not appear to recognize the prominence given to field evidence over laboratory data, which is at the heart of the Hill criteria.

WHERE THE WEIGHT-OF-EVIDENCE APPROACH HAS BEEN ADOPTED

Largely driven by scientists based in North America and Canada, a set of approaches has been recommended for environmental scientists to address causation. This drive is said to have come in part from requirements placed on regulators from the US Clean Water Act and notable pollution hot spot issues. Legal issues may also have helped propel the subject, particularly where human health impairment is believed to be due to chemical pollution (Susser, 1986). These have tended to assess whether and to what extent local chemical contamination has damaged wildlife in a specific lake or catchment (Bro-Rasmussen & Lokke, 1984; Kapo & Burton, 2006; Suter et al., 1999; Underwood, 1992). The US Environmental Protection Agency (USEPA) has made its own attempts to help guide investigators into whether any of a multitude of stressors is causing a local biological decline (Causal Analysis/Diagnosis Decision Information System [CADDIS], https://www.epa.gov/caddis). The stressor identification framework on which CADDIS is based was first published as a guidance document in 2000 (Cormier et al., 2000).

TABLE 2: How past and present chemical issues might score under the Bradford Hill criteria

Issue	1) Strength	2) Consistency	3) Specificity	4) Temporality	5) Biological gradient	6) Plausibility	7) Coherence	8) Experiment such as intervention	9) Analogy	Score
TBT and mollusks	Yes, impact worst in harbors and marinas compared to other locations	Yes, observed consistently in many places where compound used	Yes, imposex was unique to exposed population	Yes, impact linked to compounds arriving on market and recovery associated with withdrawal	Yes, strong relationship with exposure	Mechanism not fully understood	Yes, damage could be simulated in the lab	Yes, withdrawal led to recovery	Not in this case	7/9
Diclofenac and Asian vulture	Yes, impact overwhelming for carcass-feeding birds	Yes, observed consistently in many places where compound used	Yes, unique to exposed population	Yes, impact linked to compounds arriving on market	+/– Some knowledge but not so clear	Yes, as it turns out, potential for liver damage was known in humans	Yes, damage associated with diclofenac accumulation. Lab data on liver damage available	Too early to say if withdrawal leading to recovery	Not in this case	6/9
Estrogens and fish	Yes, impact on vtg and intersex was strong for exposed fish	Yes, observed consistently with respect to wastewater and EDC exposure.	Yes, strongest impacts linked to wastewater exposure	Not enough preceding records to be sure	Excellent confirming data for EDCs	Yes, fish share estrogen receptors with humans, so impacts could be expected	Yes, excellent link between lab and field-based information	Unclear, but reductions in NPE and EE2 may be having positive effects	Compounds with estrogen-like molecular structure tend to have these effects	6/9
Organochlorine insecticide and raptors	Yes, big impact on top predators	Yes, observed consistently in many places where compounds used	Yes, impact worst in top predator compared to other birds	Yes, impact linked to compounds arriving on market and recovery associated with withdrawal	+/– Some knowledge but not so clear	Not sure mechanism yet understood	+/– Range of experimental evidence of damage to different organisms even if mechanism not understood	Yes, withdrawal led to recovery	Not in this case but useful later	5/9
Neonicotinoids and bees	Pollinators preferring plants which contained neonicotinoids worst hit	Evidence of pollinator decline but not many studies clearly link to neonicotinoids	Yes, excellent UK evidence	Yes, excellent evidence of temporality	Yes, more impacts linked to more exposure	Probable that long-term damage is plausible and yet was not predicted	Yes, different streams of evidence have come together	Too early to say, perhaps	Arguably any nerve-damaging agent may be predicted to be a risk	5/9
Neonicotinoids and fish	Yes, drastic in exposed lake because effect indirect: food supply reduced	No, one study only so far	Effect was on the fish food source, not the fish directly	Yes, impact followed neonicotinoid arrival	Yes, as far as their food source, the invertebrates are concerned	Yes, but indirect via loss of their food source	Yes, lab and field evidence support neonicotinoid impact on food web	Too early to say, perhaps	Arguably any nerve-damaging agent may be predicted to be a risk	5/9

(*Continued*)

TABLE 2: (*Continued*)

Issue	1) Strength	2) Consistency	3) Specificity	4) Temporality	5) Biological gradient	6) Plausibility	7) Coherence	8) Experiment such as intervention	9) Analogy	Score
PCBs and killer whales	Yes, dramatic reproduction failures in killer whales	Yes, similar picture offered around the world	Yes, killer whales have the highest levels and their reproductive failure most evident	Possibly, although not enough historic records	Fairly good evidence on effect vs. exposure	Yes, damage to calf through release to milk and long lactation stage	Yes, good lab data on PCB complement field observations	Not possible, legacy POPs still with us	Arguable this is the case	5/9
POPs and eels	Certainly eels are in trouble, and link to POPs is not unlikely	Similar hypothesis of eel problems worldwide linked to POPs	Unique lifestyle and lipid content fit POP damage hypothesis	Unclear	Unclear but possible	Yes, uptake and damage of released POPs in eels is plausible mechanism	Yes, many lab studies on POP toxicity might support field observations	Not possible, legacy POPs still with us	Arguable this is the case	3/9
Pharmaceuticals and fish	Little dramatic evidence of harm	No consistent pattern	Unclear	Unclear, some nontoxic effects could be linked; however, many exposed populations improving over time	Yes, range of endpoints would have biological gradient	Yes, they are vertebrates like us	No strong coherence between lab and field	Not available	N/A	2/9

TBT = tributyltin; PCB = polychlorinated biphenyl; POP = persistent organic pollutant; vtg = vitellogenin; EDC = endocrine-disrupting compound; NPE = nonylphenol polyethoxylate; EE2 = 17α-ethynylestradiol; N/A = not available.

The CADDIS website was first published in 2005, with substantial revisions in 2007 and 2010. A broader approach has been to ascribe less than expected macroinvertebrate biodiversity statistically to a range of stressors, including the resident chemical mixture (largely reflecting the presence of traditional toxic chemicals such as metals; Posthuma et al., 2016).

While in North America, weight of evidence/ecoepidemiology has typically been used to guide studies into finding the causes of single or multiple wildlife problems at particular locations (Cormier et al., 2003), it can also be argued as being amenable to being used proactively to deduce the likelihood of a chemical(s) being responsible for serious harm to wildlife (leading to population-level consequences) or in effect to eliminate them as a cause. We may be mistaken, but we are not aware that a weight-of-evidence approach has been officially adopted by regulators and policymakers outside North America.

SIMPLIFYING AND RANKING CRITERIA FOR A WEIGHT-OF-EVIDENCE APPROACH TO CHEMICALS IN THE ENVIRONMENT

As shown in Table 1, a number of scientists have taken on the Hill criteria and assembled them in different ways to suit investigations of apparent problems in the natural environment. In Figure 2 we offer our own ranked sequence of criteria which could be used to eliminate a chemical, or mixture of chemicals, from a known source as being responsible for wildlife harm such as biodiversity loss.

We consider that field observations of declines in a wildlife population are the central plank in building a case for action. If at least 5 out of the 10 criteria in Figure 2 were satisfied, this would be grounds for concern and a reaction would be warranted. We acknowledge that looking for evidence of damage at the population level is very crude and can detect only the most drastic of impacts. However, if protecting the wildlife in our natural environments is our central concern, then this approach must be at the heart of protection and indeed restoration.

The need for long-term monitoring

It will not have escaped notice that to test such weight-of-evidence criteria it is necessary to be in possession of long-term wildlife monitoring data (at least annual and ideally married to sublethal biomarker measurements and chemical monitoring data) with an extensive geographic coverage. This is not a given, and, in fact, long-term monitoring programs appear to be in decline, if not under threat, everywhere. Unfortunately, different types of monitoring are often not integrated (e.g., wildlife numbers, physiological assessments based on biomarkers, behavioral changes, and chemical monitoring taking place at different times and places in the same ecosystem). We must accept that it is extremely important to maintain the resources for monitoring and assessing biodiversity over very long periods (decades), suitably designing studies to discern contaminant impacts (Jensen, 2019). A significant proportion of monitoring sites should correspond to human pressure locations such as those affected by agriculture, wastewater, and industrial discharges. European nations start from a position of some strength regarding the aquatic environment because of monitoring being reported on performance for the Water Framework Directive (Vaughan & Ormerod, 2012). Each state in the United States conducts biomonitoring surveys of its aquatic environments, usually including measurements of water chemistry and quality and wildlife surveys. The public can obtain these data from departments of environmental quality in each state. On the

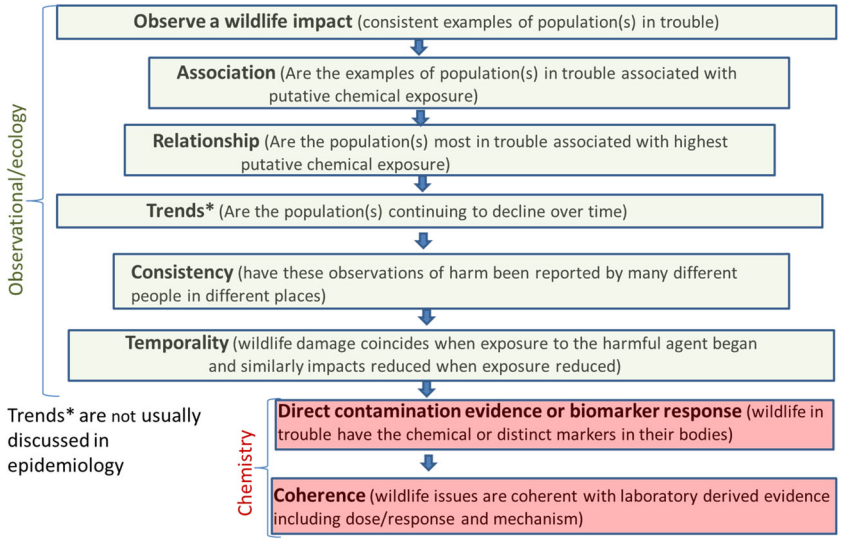

FIGURE 2: Proposed evidence sequence to follow to eliminate whether chemicals whose source is known are a cause of serious wildlife damage such as biodiversity loss.

terrestrial side, the Breeding Bird Surveys is a large-scale, long-term bird monitoring program (collaboration of the US Geological Survey and Environment Canada) where data collection follows a strict protocol (Belden et al., 2018). The situation for detection of harm within the terrestrial environment seems more ad hoc in Europe, and in the United Kingdom there is a tendency to rely much more on recording by nongovernmental organizations (e.g., the Royal Society for the Protection of Birds) and amateur naturalists, which, while often excellent, does seem a rather tenuous method for the state to ensure protection (Outhwaite et al., 2020). The United Kingdom at least does monitor both wildlife and chemicals around its coasts, although assessing changes in populations and their causes is difficult (Nicolaus et al., 2016).

Appropriateness and practicality

There are those who argue that "if you see a decline in wildlife, it is already too late." We have refuted this argument in previous work (Johnson & Sumpter, 2016), but suffice to say, most of the chemicals whose impact we still debate have been on the market for decades. Another criticism might be that this approach "sets the bar too high." Most would agree that national energies and resources should be focused especially on protecting the wildlife in greatest peril. Such efforts may be compromised if we dissipate our efforts and limited resources too widely and chase after issues which present negligible or nonexistent threats currently. We fear that continuously focusing our research funding into a succession of 3 year–funded projects centered on laboratory ecotoxicity tests of chemicals may overlook real areas of concern that would be discoverable by long-term monitoring efforts. The centrality that retrospective environmental monitoring could have in chemical safety has been highlighted (Milner & Boyd, 2017). We do not underestimate the difficulty in untangling cause and effect from within a natural environment with multiple stressors and compensating factors. Nor do we dismiss worrying trends of chemical use that are gathering pace over the coming years. The medical epidemiologist cannot and does not shrink from this challenge, and neither should we.

Some might argue that this approach takes us no further forward in assessing whether complex mixtures of chemicals are harming wildlife. However, if the source of the mixture is known or suspected, for instance, the mixtures of chemicals discharged in wastewater, then criteria 1 to 6 could help begin to provide answers. In this case, exposure can be inferred from the percentage of wastewater in the receiving waters (Jobling et al., 2006). The approach is a retrospective one. The problem has to occur and be identified before utilizing the approach we recommend. We have made the case elsewhere for the development of policies that might help to avoid or greatly reduce chemical threats in the environment in the first place (Collins et al., 2020).

In an ideal world, the next generation of chemicals would be much less toxic and persistent than their predecessors, and our prospective risk assessment would be much better than that achieved using current methods, so chemical problems in the environment would gradually become a thing of the past (Johnson et al., 2020). But a moral duty would remain to demonstrate, via monitoring, that we have succeeded, and the cause of any deterioration should not be put at the foot of chemicals without compiling evidence using a weight-of-evidence approach.

Back in 1997 it was argued passionately within governments and international bodies that we should adopt a formal structure for evidence gathering and analysis to support decision-making and restoration using approaches such as weight of evidence (Gilbertson, 1997), particularly where chemicals are accused of causing harm. The United States had very persuasive proponents, such as Cormier, Norton, and Suter, which helped in developing the CADDIS approach by the USEPA. In this modern era of chemical anxieties and deep concern over biodiversity loss, we would urge ministries of the environment and regulatory authorities in more nation-states to formally acknowledge and adopt these weight-of-evidence approaches, as offered in Figure 2.

CONCLUSIONS

The advantages for scientists, regulators, and policymakers in formally adopting a weight-of-evidence approach to challenges from chemicals in the environment include: a transparent mechanism behind decision-making, and the evidence largely comes from "neutral sources." This potentially takes a lot of the political heat out of the debate, and ensures that responses and efforts are proportionate and reflect what an informed society demands.

We believe that if 1) we monitored wildlife well enough, and 2) we used these criteria to score chemical or substance challenges, then our field and the appropriate policy responses would be on a much firmer footing than they are today. The weight-of-evidence approach reviews all relevant data, rather than focusing on a limited suite of test results, which are often of little relevance to real-world circumstances.

It is guaranteed that policymakers and regulators will face at some stage in the future (if they aren't already) unexpected chemical and substance effects and/or unanticipated declines in wildlife populations that will lead to a public clamor for action. When these moments arise, having a coherent and transparent set of criteria for assembling all the evidence available will help everyone. We recommend that those who advocate more precautionary approaches reflect on the weight-of-evidence approach. With contentious issues, it is not so much the quantity of evidence available, but rather whether multiple lines of different evidence exist which together can infer causation. This weight-of-evidence approach will appeal to all stakeholders, a factor which is critical in gaining consent prior to moving to control chemicals, where necessary, and hopefully restoring any damage that has been done.

Acknowledgment—The authors are grateful for funding from the Natural Environment Research Council (grant NE/S000100/1) for the ChemPop project, which has supported the present study, and for the meetings of the Defra Hazardous Substances

Advisory Committee, which stimulated the discussions that led to this publication.

Data Availability Statement—Data, associated metadata, and calculation tools are available from the corresponding author (ajo@ceh.ac.uk).

REFERENCES

Adams, S. M. (2003). Establishing causality between environmental stressors and effects on aquatic ecosystems. *Human and Ecological Risk Assessment*, *9*, 17–35.

Ankley, G. T., Giesy, J. P. 1998. *Endocrine disruptors in wildlife: A weight-of-evidence perspective*. SETAC, Pensicola, FL, USA.

Belden, J. B., McMurry, S. T., Maul, J. D., Brain, R. A., & Ghebremichael, L. T. (2018). Relative abundance trends of bird populations in high intensity croplands in the central United States. *Integrated Environmental Assessment and Management*, *14*, 692–702.

Bergstrom, D. M., Wienecke, B. C., van den Hoff, J., Hughes, L., Lindenmayer, D. B., Ainsworth, T. D., Baker, C. M., Bland, L., Bowman, D. M. J. S., Brooks, S. T., Josep G. Canadell, J. G., Constable, A. J., Dafforn, K. A., Depledge, M. H., Dickson, C. R., Duke, N. C., Helmstedt, K. J., Holz, A., & Johnson, C. R. (2021). Combating ecosystem collapse from the tropics to the Antarctic. *Global Change Biology*, *27*, 1692–1703.

Bro-Rasmussen, F., & Lokke, H. (1984). Ecoepidemiology—A casuistic discipline describing ecological disturbances and damages in relation to their specific causes: Exemplified by chlorinated phenols and chlorophenoxy acids. *Regulatory Toxicology and Pharmacology*, *4*, 391–399.

Burton, G. A., Chapman, P. M., & Smith, E. P. (2002). Weight-of-evidence approaches for assessing ecosystem impairment. *Human and Ecological Risk Assessment*, *8*, 1657–1673.

Collier, T. K. (2003). Forensic ecotoxicology: Establishing causality between contaminants and biological effects in field studies. *Human and Ecological Risk Assessment*, *9*, 259–266.

Collins, C., Depledge, M., Fraser, R., Johnson, A., Hutchison, G., Matthiessen, P., Murphy, R., Owens, S., & Sumpter, J. (2020). Key actions for a sustainable chemicals policy. *Environment International*, *137*, 4.

Cormier, S. M., Norton, S. B., & Suter, G. W. (2000). *Stressor identification guidance document* (EPA/822/B-00/025). US Environmental Protection Agency, Washington, DC. Retrieved 2021 August, 17, from https://cfpub.epa.gov/ncea/caddis/recordisplay.cfm?deid=20685

Cormier, S. M., Norton, S. B., & Suter, G. W. (2003). The US Environmental Protection Agency's stressor identification guidance: A process for determining the probable causes of biological impairments. *Human and Ecological Risk Assessment*, *9*, 1431–1443.

Desforges, J. P., Hall, A., McConnell, B., Rosing-Asvid, A., Barber, J. L., Brownlow, A., De Guise, S., Eulaers, I., Jepson, P. D., Letcher, R. J., Levin, M., Ross, P. S., Samarra, F., Vikingson, G., Sonne, C., & Dietz, R. (2018). Predicting global killer whale population collapse from PCB pollution. *Science*, *361*, 1373–1376.

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., & Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, *95*, 5–12.

Fahlman, J., Hellstrom, G., Jonsson, M., Fick, J. B., Rosvall, M., & Klaminder, J. (2021). Impacts of oxazepam on perch (*Perca fluviatilis*) behavior: Fish familiarized to lake conditions do not show predicted anti-anxiety response. *Environmental Science & Technology*, *55*, 3624–3633.

Fox, G. A. (1991). Practical causal inference for ecopidemiologists. *Journal of Toxicology and Environmental Health*, *33*, 359–373.

Gee, D. (2006). Late lessons from early warnings: Toward realism and precaution with endocrine-disrupting substances. *Environmental Health Perspectives*, *114*, 152–160.

Geeraerts, C., Focant, J. F., Eppe, G., De Pauw, E., & Belpaire, C. (2011). Reproduction of European eel jeopardised by high levels of dioxins and dioxin-like PCBs? *Science of the Total Environment*, *409*, 4039–4047.

Gibbs, P. E., Bryan, G. W., & Pascoe, P. L. (1991). TBT-induced imposex in the dogwhelk, *Nucella lapillus*—Geographical uniformity of the response and effects. *Marine Environmental Research*, *32*, 79–87.

Gilbertson, M. (1997). Advances in forensic toxicology for establishing causality between Great Lakes epizootics and specific persistent toxic chemicals. *Environmental Toxicology and Chemistry*, *16*, 1771–1778.

Gold, S. C., & Wagner, W. E. (2020). Filling gaps in science exposes gaps in chemical regulation. *Science*, *368*, 1066–1068.

Hanson, M., & Brain, R. (2020). Context and perspective in ecotoxicology. *Environmental Toxicology and Chemistry*, *39*, 1655.

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 & Technology*, *48*, 3100–3111.

Hayes, T. B., Anderson, L. L., Beasley, V. R., de Solla, S. R., Iguchi, T., Ingraham, H., Kestemont, P., Kniewald, J., Kniewald, Z., Langlois, V. S., Luque, E. H., McCoy, K. A., Muñoz-De-Toro, M., Oka, T., Oliveira, C. A., Orton, F., Ruby, S., Suzawa, M., Tavera-Mendoza, L. E., & Willingham, E. (2011). Demasculinization and feminization of male gonads by atrazine: Consistent effects across vertebrate classes. *Journal of Steroid Biochemistry and Molecular Biology*, *127*, 64–73.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300.

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES Secretariat, Bonn, Germany.

Jensen, O. P. (2019). Pesticide impacts through aquatic food webs. *Science*, *366*, 566–567.

Jobling, S., Williams, R., Johnson, A., Taylor, A., Gross-Sorokin, M., Nolan, M., Tyler, C. R., van Aerle, R., Santos, E., & Brighty, G. (2006). Predicted exposures to steroid estrogens in UK rivers correlate with widespread sexual disruption in wild fish populations. *Environmental Health Perspectives*, *114*, 32–39.

Johnson, A. C., Jin, X. W., Nakada, N., & Sumpter, J. P. (2020). Learning from the past and considering the future of chemicals in the environment. *Science*, *367*, 384–387.

Johnson, A. C., & Sumpter, J. P. (2016). Are we going about chemical risk assessment for the aquatic environment the wrong way? *Environmental Toxicology and Chemistry*, *35*, 1609–1616.

Joy, V. C., Pramanik, R., & Sarkar, K. (2005). Biomonitoring insecticide pollution using non-target soil microarthropods. *Journal of Environmental Biology*, *26*, 571–577.

Kapo, K. E., & Burton, G. A. (2006). A geographic information systems-based, weights-of-evidence approach for diagnosing aquatic ecosystem impairment. *Environmental Toxicology and Chemistry*, *25*, 2237–2249.

Landrigan, P. J., Fuller, R., Acosta, N. J. R., Adeyi, O., Arnold, R., Basu, N. N., Baldé, A. B., Bertollini, R., Bose-O'reilly, S., Boufford, J. I., Breysse, P. N., Chiles, T., Mahidol, C., Coll-Seck, A. M., Cropper, M. L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., & Zhong, M. (2018). The Lancet Commission on pollution and health. *Lancet*, *391*, 462–512.

Matthiessen, P., Waldock, R., Thain, J. E., Waite, M. E., & Scropehowe, S. (1995). Changes in periwinkle (*Littorina littorea*) populations following the ban on TBT-based antifoulings on small boats in the United Kingdom. *Ecotoxicology and Environmental Safety*, *30*, 180–194.

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., Orgego, 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*, 320–344.

Milner, A. M., & Boyd, I. L. (2017). Toward pesticidovigilance: Can lessons from pharmaceutical monitoring help to improve pesticide regulation? *Science*, *357*, 1232–1234.

Moraes, R., & Molander, S. A. (2004). Procedure for ecological tiered assessment of risks (PETAR). *Human and Ecological Risk Assessment*, *10*, 349–371.

Munkittrick, K. R., & McCarty, L. S. (1995). An integrated approach to aquatic ecosystem health: Top-down, bottom-up or middle-out? *Journal of Aquatic Ecosystem Health*, *4*, 77–90.

Myers, J. P., vom Saal, F. S., Akingbemi, B. T., Arizono, K., Belcher, S., Colborn, T., Chahoud, I., Crain, D. A., Farabollini, F., Guillette, L. J., Jr., Hassold, T., Ho, S.-m., Hunt, P. A., Iguchi, T., Jobling, S., Kanno, J., Laufer, H., Marcus, M., McLachlan, J. A., & Zoeller, R. T. (2009). Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: The case of bisphenol A. *Environmental Health Perspectives*, *117*, 309–315.

Newton, I., & Bogan, J. (1974). Organochlorine residues, eggshell thinning and hatching success in British sparrowhawks. *Nature*, *249*, 582–583.

Nicolaus, E. E. M., Wright, S. R., Bolam, T. P. C., Barber, J. L., Bignell, J. P., & Lyons, B. P. (2016). Spatial and temporal analysis of the risks posed by polychlorinated biphenyl and metal contaminants in dab (*Limanda limanda*) collected from waters around England and Wales. *Marine Pollution Bulletin*, *112*, 399–405.

Oaks, J. L., Gilbert, M., Virani, M. Z., Watson, R. T., Meteyer, C. U., Rideout, B. A., Shivaprasad, H. L., Ahmed, S., Chaudry, M. J. I., Arshad, M., Mahmood, S., Ali, A., & Khan, A. A. (2004). Diclofenac residues as the cause of vulture population decline in Pakistan. *Nature*, *427*, 630–633.

Outhwaite, C. L., Gregory, R. D., Chandler, R. E., Collen, B., & Isaac, N. J. B. (2020). Complex long-term biodiversity change among invertebrates, bryophytes and lichens. *Nature Ecology & Evolution*, *4*, 384–392.

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Posthuma, L., Dyer, S. D., de Zwart, D., Kapo, K., Holmes, C. M., & Burton, G. A. (2016). Eco-epidemiology of aquatic ecosystems: Separating chemicals from multiple stressors. *Science of the Total Environment*, *573*, 1303–1319.

Ratcliffe, D. A. (1970). Changes attributable to pesticides in egg breakage frequency and eggshell thickness in some British birds. *Journal of Applied Ecology*, *7*, 67–115.

Solomon, K. R., Carr, J. A., Du Preez, L. H., Giesy, J. P., Kendall, R. J., Smith, E. E., & Van Der Kraak, G. J. (2008). Effects of atrazine on fish, amphibians, and aquatic reptiles: A critical review. *Critical Reviews in Toxicology*, *38*, 721–772.

Sumpter, J. P. (2005). Endocrine disrupters in the aquatic environment: An overview. *Acta Hydrochimica et Hydrobiologica*, *33*, 9–16.

Susser, M. (1986). Rules of inference in epidemiology. *Regulatory Toxicology and Pharmacology*, *6*, 116–128.

Susser, M. (1991). What is a cause and how do we know one—A grammar for pragmatic epidemiology. *American Journal of Epidemiology*, *133*, 635–648.

Suter, G. W., Barnthouse, L. W., Breck, J. E., Gardner, R. H., & O'Neill, R. V. (1985). Extrapolating from the laboratory to the field: How uncertain are you? In R. D. Cardwell, R. Purdy, & R. C. Bahner (Eds.), *Aquatic toxicology and hazard assessment: Seventh symposium* (ASTM STP 854, pp. 400–413). American Society for Testing and Materials.

Suter, G. W., Barnthouse, L. W., Efroymson, R. A., & Jager, H. (1999). Ecological risk assessment in a large river-reservoir: 2. Fish community. *Environmental Toxicology and Chemistry*, *18*, 589–598.

Suter, G. W., Cormier, S. M., & Norton, S. B. (2007). Ecoepidemiology and causal analysis. In G. W. Suter (Ed.), *Ecological risk assessment*. CRC, Pensicola, FL, USA, pp 39–68.

Underwood, A. J. (1992). Beyond baci—The detection of environmental impacts on populations in the real, but variable, world. *Journal of Experimental Marine Biology and Ecology*, *161*, 145–178.

Vaughan, I. P., & Ormerod, S. J. (2012). Large-scale, long-term trends in British river macroinvertebrates. *Global Change Biology*, *18*, 2184–2194.

Villeneuve, D. L., Coady, K., Escher, B. I., Mihaich, E., Murphy, C. A., Schlekat, T., & Garcia-Reyero, N. (2019). High-throughput screening and environmental risk assessment: State of the science and emerging applications. *Environmental Toxicology and Chemistry*, *38*, 12–26.

Wang, Z. Y., Altenburger, R., Backhaus, T., Covaci, A., Diamond, M. L., Grimalt, J. O., Lohmann, R., Schaffer, A., Scheringer, M., Selin, H., Soehl, A., & Suzuki, N. (2021). We need a global science-policy body on chemicals and waste. *Science*, *371*, 774–776.

Washington, J. W., Rosal, C. G., McCord, J. P., Strynar, M. J., Lindstrom, A. B., Bergman, E. L., Goodrow, S. M., Tadesse, H. K., Pilant, A. N., Washington, B. J., Davis, M. J., Stuart, B. G., & Jenkins, T. M. (2020). Nontargeted mass spectral detection of chloroperfluoropolyether carboxylates in New Jersey soils. *Science*, *368*, 1103–1107.

Weed, D. L. (2005). Weight of evidence: A review of concept and methods. *Risk Analysis*, *25*, 1545–1557.

Weltje, L., & Sumpter, J. P. (2017). What makes a concentration environmentally relevant? Critique and a proposal. *Environmental Science & Technology*, *51*, 11520–11521.

Woodcock, B. A., Isaac, N. J. B., Bullock, J. M., Roy, D. B., Garthwaite, D. G., Crowe, A., & Pywell, R. F. (2016). Impacts of neonicotinoid use on long-term population changes in wild bees in England. *Nature Communications*, *7*, Article 12459.

Wynne, B. (1992). Uncertainty and environmental learning—Reconceiving science and policy in the preventive paradigm. *Global Environmental Change: Human and Policy Dimensions*, *2*, 111–127.

Yamamuro, M., Komuro, T., Kamiya, H., Kato, T., Hasegawa, H., & Kameda, Y. (2019). Neonicotinoids disrupt aquatic food webs and decrease fishery yields. *Science*, *366*, 620–623.

