Water-Quality Benchmarks for the Protection of Aquatic Life: A Review of Methodologies

Anson Main, PhD
March 23, 2021

Executive Summary

Degradation of surface water systems by chemicals such as pesticides is a major concern for government regulators, conservationists, and the general public. However, the criteria that are used to protect freshwater aquatic resources may differ across the state, national, and international level. Most organizations (e.g., government, NGOs) use available toxicity data to determine thresholds to protect aquatic systems and their resident organisms. These thresholds are considered *water-quality benchmarks,* which are simply defined as a value by which measured concentrations can be compared to help evaluate the potential effects of pesticides on water quality.

The purpose of this analysis memo is to broadly describe the methodologies behind development of key water-quality benchmarks (WQBs) from California Water Boards, the US EPA, and other countries. Data quality is of the utmost importance when determining WQBs. Many organizations specify a range of taxonomic diversity required for inclusion in developing benchmarks and further specify the types of studies required (i.e., acute or chronic), associated physical-chemical data, and test conditions. As a case study, the development of current benchmarks for imidacloprid is briefly discussed herein.

Derivation of aquatic benchmark values is typically accomplished by one of two approaches: an assessment factor (AF) or a species sensitivity distribution (SSD). The aim of both methods is to develop a regulatory value from available ecotoxicity data that will protect the majority of organisms within the aquatic environment. AFs are considered a conservative approach that divides the lowest collected toxicity data by a numeric safety factor to develop a criterion. By comparison, an SSD is a statistical distribution of toxicity values that are a response of a selected species (often specific required taxa) for a given pesticide. An SSD can be used to derive a protective value for aquatic systems expressed in terms of percentile of organisms protected (e.g., 5^{th}, 10^{th}). Although SSDs are often more scientifically defensible than AFs, one major drawback is that they require larger data sets. Depending on the pesticide of interest, these data may not exist.

Within the scientific literature, there is no consensus on whether one method is better than another for deriving a WQB. Instead, careful consideration must be made regarding the purpose of the derived value. This includes the inclusion (or exclusion) of specific taxa and the data sources available. Although requiring species from a range of taxonomic groups may appear more robust, it is possible that available toxicity tests are biased toward sensitive or insensitive species that are also not truly representative of local conditions. Similarly, a lack of data may require regulators to calculate conservative benchmarks until more data become available. Conversely, new data may lead to greatly reduced values as additional toxicity results become available. The derivation and development of WQBs is complex as they are dependent on the

availability of high quality data and use of defensible scientific criteria when including or excluding studies (i.e., quality control). These considerations should not, however, preclude regulators from developing WQBs or relying on them to make risk management decisions.

1. Introduction

The protection and conservation of aquatic life in freshwater surface water systems from pesticide contamination is one of the major priorities for numerous countries across the globe. However, the derivation and development of WQBs is complex as they are dependent on the availability of high quality data and use of defensible scientific criteria when including or excluding studies (i.e., quality control). The criteria that are used to evaluate a water body at risk may differ between organizations, government agencies, or countries. The level at which protection of aquatic resources occurs (i.e., how protective) may be the major difference between regulatory approaches to similar pesticide problems. In the United States, the US EPA has embraced the use of ecological risk assessment as a tool to assess potential chemical hazards on surface water (TenBrook *et al*., 2008). The Office of Pesticide Programs uses ecological risk assessments that compare exposure values to LC_{50}/EC_{50}s (e.g., Aquatic Life Benchmarks based on a single species), whereas a more protective approach (i.e., requiring specific taxa and test conditions) is taken by the US EPA Office of Water using their ambient water quality criteria (US EPA, 1985). This is in stark contrast to the European Union which uses the *precautionary principle* to protect aquatic resources. This principle dictates that lack of full scientific certainty shall not be used as a reason for postponing measures that prevent degradation of the environment (Rio Convention, 1992). Compared to ecological risk assessment, the precautionary principle often defines low, no-effect concentrations (i.e., predicted concentrations where no toxic effects are likely) or utilizes large safety factors specific to the chemical of interest (Chapman *et al*., 1998).

Even though there are many different approaches to protecting aquatic life, numerous countries, organizations, and agencies derive *water-quality benchmarks* (WQBs) for specific chemicals of interest (e.g., pesticides). A WQB is defined as "a threshold value against which measured concentrations can be compared to help assess the potential effects of contaminants on water quality" (USGS, 2019). These benchmarks are often described as numeric values that are derived from methodologies such as species sensitivity distributions (SSD) or assessment factors (AF) that will be discussed herein. However, depending on the specific country or organization, WQBs may be calculated differently or referred to by many different names including: criteria (US EPA); water quality objectives (California Water Boards); guidelines (Canada); environmental risk limits (Netherlands); and water quality targets (TenBrook *et al*., 2008). Benchmark is a generic term selected here to describe the various values. The major difference between specific terms often is related to whether it is legally enforceable (e.g., a "standard") or more simply advisory (e.g., a "guideline" (USGS, 2019)). Regardless of the nomenclature, WQBs are derived from scientifically based numbers intended to protect aquatic life from potential negative effects of pesticide use.

Therefore, the objective of this review is to broadly describe the derivation of – and methodology behind – key WQBs from California and around the globe.

2. Use of single species versus multi-species to derive numeric WQBs

A key consideration in determining WQBs is the level of biological organization that the chosen benchmark is intended to protect. This may include methodologies used to protect specific individual species, representative taxa, ecosystem function, or entire aquatic communities. Some countries develop guidelines to protect biological communities rather than individual taxa. For example, Canada's freshwater aquatic guidelines are set to protect all biotic components (e.g., algae, macrophytes, invertebrates) of the ecosystem, whereas the US EPA specifically protects "aquatic organisms and their uses". The term "uses" is not directly defined by the US EPA; instead, they indicate that monitoring programs should adequately monitor species of concern to the public such as fresh and salt water fish and macroinvertebrates (US EPA, 1985). The US EPA, however, acknowledge that ecosystems can tolerate some level of stress (TenBrook *et al*., 2008). Although many countries differ in their chosen WQBs, there is a greater reliance on single-species toxicity data to derive most protective criteria. There is often a paucity of available studies that have evaluated multispecies or ecosystem responses to pesticides. In response, most methodologies for deriving WQBs seek ecosystem protection by reliance on extrapolation from single-species laboratory tests. These methodologies further assume that: (1) ecosystem sensitivity depends on its most sensitive species and (2) protecting ecosystem structure will in turn protect community function (TenBrook *et al*., 2008). Similar to the US EPA's Aquatic Life Benchmarks, countries such as Canada, France, Germany, and the UK compile available toxicity data and then select the single most sensitive datum to derive their respective benchmarks. However, calculated benchmarks do not have confidence limits associated with them. Numerical national water quality criteria values derived by the US EPA's Office of Water also do not have confidence limits despite using an extrapolation method such as a SSD. Because many countries have specific requirements for deriving aquatic benchmarks, one of the main goals for developing the University of California-Davis methodology (UCDM) was to derive criteria for a range of compounds (e.g., pyrethroids) that do not meet the specific US EPA data requirements (Fojut *et al*., 2012). The UCDM strongly recommends the use of an SSD for calculating WQBs with the main differences from the US EPA as follows: (1) UCDM provides a thorough, transparent procedure for study evaluation; (2) a more advanced SSD that uses the geometric means of five species and a specified distribution (e.g., Burr Type III, Log-logistic); (3) alternate approaches if data requirements for the SSD or acute-to-chronic ratio cannot be met; and (4) considerations for mixture toxicity (TenBrook *et al*., 2010; Fojut *et al*., 2012).

3. Background on the State of California's WQO

California's present system of water quality control was established in 1969 under the Porter-Cologne Water Quality Control Act (P-CA). The P-CA [CWC, Section 13050 (h)] defines *water quality objectives* (WQO) as "the limits or levels of water quality constituents or characteristics which are established for the reasonable protection of beneficial uses of water or the prevention of nuisance within a specific area." Therefore, the State Water Resources Control Board (SWRCB) and nine Regional Water Quality Control Board's water quality control programs focus on the prevention of "pollution" defined by the P-CA as an alteration of the quality of the waters of the state by waste to a degree which unreasonably affects beneficial uses [CWC, Section 13050 (I)]. In California, WQO can be implemented in a Water Quality Control Plan

(CCRWQCB, 2017) that is implemented regionally via a Basin Plan and may be specified as either numeric or in narrative form. Numeric objectives establish enforceable receiving water concentrations; however, many water quality objectives are implemented in narrative form. This does not mean that narrative objectives are not enforceable, but rather that they describe a requirement or prohibit a condition considered harmful to a beneficial use. Numeric thresholds in California are derived for protection of human and aquatic organisms; therefore, they are highly varied. These thresholds may include: drinking water standards (i.e., maximum contaminant levels), criteria maximum concentrations, and criteria continuous concentrations that protect aquatic organisms from acute and chronic exposures to pollutants, respectively (SWRCB, 2016; CCRWQCB, 2017).

4. WQBs: The use of Assessment Factors and Species Sensitivity Distributions

There are two basic derivation methodologies used to determine WQBs throughout the world: the AF and the statistical extrapolation method (e.g., SSD technique). The aim of both methods is to extrapolate a reliable threshold value from available ecotoxicity data that will be protective of the aquatic environment. Put simply, the AF method divides the lowest value of the collected toxicity data by a factor to develop a criterion. Almost all of the available AF methodologies include data for aquatic plants and animals together when deriving benchmarks (TenBrook *et al*., 2008). AFs are considered to be a simple and conservative approach for dealing with uncertainty when determining potential risks of chemicals. However, the possibility of over- or underestimating risk is greatly increased when using AFs. There are concerns that if factors are applied generically instead of mathematically derived or included taxa are either too sensitive or insensitive, AFs may not be protective or unnecessarily protective. Importantly, AFs should be context specific and based on existing scientific knowledge (Chapman *et al*., 1998). Some current WQBs may be based on a single datum or may be an estimated toxicity value (e.g., based on a quantitative structure activity relationship (QSAR)) rather than an actual measured value. Many European Union countries (e.g., France, Germany, Spain, and the UK) and Canada support the precautionary principle by relying solely on the AF method for derivation of WQBs (Table 1). For example, the Canadian methodology uses the chronic lowest-observed-effect concentration (LOEC) values to determine WQBs. If adequate data exist, the lowest available LOEC value is divided by a safety factor of 10 to determine protection (CCME, 1999; TenBrook *et al*., 2008). In contrast, the US EPA, Denmark, the Organisation for Economic Co-operation and Development (OECD), and Australia use a combination of the SSD and AF methods.

Assessment factors (also known as “safety factors” or “uncertainty factors”) can vary widely from 1 to 1000 and are applied based on the amount of data available and the kinds of data available (Table 1). For example, the OECD recommends AFs if data are limited. The OECD divides the lowest no-observable-effect concentration (NOEC, chronic toxicity) by a factor of 10 if the data include algae, crustaceans, and fish. This factor is raised to 100 for acute data and 1000 if only one or two species are represented. Regulatory programs throughout the world have used standardized factors of 10, 20, and 100 – despite having supporting data - more as a policy decision to assure protection rather than basing these factors on empirical science (Chapman *et al*., 1998).

An SSD is a statistical distribution that describes the response of a selection of species to the toxic effects of a certain pesticide. In order to be representative of ecosystems, some countries have strict requirements for the minimum number (e.g., US EPA: up to eight taxa) or type of organisms to include such as aquatic insects, fish (warm and cold water), and plants. The assumption behind the use of an SSD is that sensitivities of a selection of species can be described by some distribution. Available ecotoxicological data are then seen as a sample from this distribution to estimate parameters of an SSD (Posthuma *et al*., 2001). The estimated points are visualized as a cumulative distribution function where effect concentrations are plotted as either acute or chronic toxicity tests (see Figure 1; Giddings *et al*., 2014). When deriving environmental water quality criteria, a cutoff percentage (*p*) of a hazardous concentration (HC) is then chosen to protect all species with LC/EC_{50}'s above the calculated "safe" concentration (HC_p). In the earliest methods, the 5^{th} percentile of a chronic toxicity distribution has been chosen as a concentration protective for the majority of species in a community (i.e., HC_5). This does not mean that 5% of the species will knowingly be harmed, but rather an HC_5 assists in deriving a predicted no-effect concentration (PNEC) for an ecosystem of interest. However, there are some differences among various SSD methods such as the chosen shape of the distribution that is used for extrapolations. In addition, data aggregation, the kinds or quantity of data used, and the level of confidence associated with the criteria may be different among SSDs (TenBrook *et al*., 2008). Depending on the needs or the research question of interest, some researchers will divide data into groups, exclude less-sensitive species or taxa, or develop SSDs for combined and separate data sets. Although SSDs are empirically derived, they do not account for the potential for ecological interactions, the habitat needs of taxa, the importance of functional groups, or account for keystone species (Newman *et al*., 2000).

5. Variation of WQBs among countries and organizations

5.1 The main criteria included in derivation

Where possible, the criteria included in developing WQBs should be based on a range of taxonomic diversity. Importantly, physical-chemical data are needed for proper interpretation of included toxicity test data, estimations of bioavailability, and for estimation of potential cumulative toxicity for multiple chemicals. Often there is no clear guidance regarding how many studies should be included or what kind of data (e.g., acute or chronic studies, physical-chemical data) are required for calculation of WQBs. Similarly, there can be differences among countries in what is considered to be an acute exposure versus a more chronic exposure in their individual guidelines. For example, the Netherlands define an acute exposure as simply lasting a short period while chronic exposure should continue through at least part of a life cycle. Australian guidelines generally describe acute tests as shorter than chronic tests, with tests longer than 96 hrs considered to be chronic. Both acute and chronic are terms that need to be clearly defined in any methodology and should only be included as such when determining either acute or chronic guidelines. However, often chronic toxicity data are lacking leaving acute data to be used to derive chronic water quality criteria (TenBrook *et al*., 2008) and data sources can be highly varied. For example, acute endpoints (LC_{50} and EC_{50}) are often used interchangeably, but the resultant concentrations for a given species may be over an order of magnitude in difference and

care must be made during the data selection process to insure consistency. Guidelines in Spain indicate published data from all sources may be used to derive WQBs. Canadian guidelines specify the kinds of data that should be sought, but not the sources of the listed data. The OECD, German, US EPA, EU, France, and South African guidelines contain no specific descriptions of what constitutes an adequate literature search or where to find data necessary data for inclusion (TenBrook *et al*., 2008).

5.2 Concerns of data quality

In order to minimize uncertainty behind water quality criteria, only data that meet pre-set standards should be included for consideration and assessment. Many countries specify that any toxicological tests need to be conducted in settings that adhere to good laboratory practices (GLP). In agreement, some countries specify the physical-chemical parameters that must be included when evaluating the quality of toxicological data. The Netherlands requires that water solubility should be determined at ~25°C and other temperature-dependent parameters such as Henry's constant, vapor pressure, and the octanol-water partition coefficient (K_{OW}) should be reported (Table 2). The US EPA is specific about inclusion of volatility, solubility, and degradability to evaluate toxicity data; Canada requires environmental fate data; and, the Danish methodology simply specifies that a wide-range of data should be considered. However, inclusion of specific physical-chemical parameters is highly variable as some countries (e.g., UK, South Africa) have guidelines that do not specify evaluation for physical-chemical data when deriving WQBs (Table 2).

Depending on the pesticide of interest, there may be an exhaustive amount of data to consider when determining WQBs. However, it is important to identify the quality of the toxicity studies that are being used to derive any benchmark. Data quality is an oft discussed concern with respect to developing water quality guidelines. For example, the EU specifically defines data quality through two terms: reliability and relevance. In essence, *reliability* relates to test methodology, the quality of the testing, and the way that both performance and results are described. The *relevance* refers to the appropriateness of a test for a particular hazard or risk assessment (TenBrook *et al*., 2008). Therefore, *reliable data* are made up of studies that clearly report testing methods used and that tests were conducted according to accepted standards or GLP. The UK, the Netherlands, Canada, and Australia/New Zealand evaluate ecotoxicity data by assigning ratings based on reliability and relevance; however, the UK considers only primary data (highly reliable and highly relevant) to be part of this classification. The guidelines in Australia and New Zealand develop weighted scores that are applied to 18 characteristics of test methodology with heavy weighting on exposure duration and endpoint. When determining national water quality criteria, the US EPA consider criteria data only if they are published or in the form of a typed, dated, and signed document with enough detail to illustrate accepted test procedures were used to obtain reliable results. There are many criteria by which the US EPA Office of Water (OW) may reject tests including no control treatment, improper dilution of water, and/or if too many organisms died during testing (TenBrook *et al*., 2008). In contrast, whereas the OW requires a specific number of taxa for consideration, the US EPA Office of Pesticide Programs (OPP) has a somewhat different assessment. The OPP does not specify the

number of studies or test organisms. Instead, OPP uses aquatic toxicity data that rely more heavily on the selection of endpoints from the most sensitive species tested in acceptable studies (US EPA, 2004).

The UCDM illustrates a detailed numeric rating system for single-species effects studies by assigning a relevance score and a reliability score. Scores are designed as relevant (R), less relevant (L), or not relevant (N) with only R and L scores evaluated for reliability (Fojut *et al*., 2008). Although data quality evaluation is often focused on existing single-species toxicity tests, it is important to recognize that laboratory data sets may be also biased toward tolerant or sensitive species. Field conditions are also considerably different than those conditions maintained in the laboratory environment (Posthuma *et al*., 2001). Compared to a laboratory, there are uncertainties with interpreting a clear cause and effect in a more representative system (e.g., field, mesocosm). This may be attributed to differences in water quality parameters, ecological interactions with other organisms (e.g., predator-prey), or the addition of unstudied compounds. Therefore, mesocosm studies appear to be the most contested type of data for inclusion in determining WQBs despite many regulators and scientists agreeing to their overall biological relevance. For example, Canada's Pest Management Regulatory Agency (PMRA) indicate the collective interpretation of imidacloprid mesocosm data was challenging due to deficiencies such as (1) an inadequate number of exposure concentrations; (2) short study duration; (3) application regimes that are not representative of most exposure scenarios; and, (4) a low abundance of sensitive invertebrate species prevented reliable statistical evaluation (PMRA, 2016).

5.3 Quantity and kinds of data that are required for deriving WQBs

Depending on the WQBs being developed, the methodology being used, and the requirements of the country or organization, the quantity of ecotoxicological effects data required for evaluation is often vastly different. Though the two basic methods for extrapolating from effects data are the application of AF and statistical extrapolation of SSDs, there is little guidance on what constitutes appropriate levels of data when using the AF method. In some cases, this may lead to use of the most sensitive datum to develop an aquatic toxicity threshold. Methods that utilize statistical extrapolation are often in disagreement over how many studies, specific species, or data points are needed to produce sound criteria. Australia and New Zealand consider at least five single-species chronic NOEC values (from five different species) to be the minimum criteria in developing high reliability trigger values (Table 3). In deriving a final chronic value, the US EPA OW requires chronic NOEC values for at least eight animal families including two fish species, two crustaceans, an insect, a member of the family chordata, and two other unrepresented families (Table 3). The US EPA OPP does not indicate a specific number of taxon to be included for determination of either acute or chronic values. The UCDM acute data set requires five representative taxa in order to use an SSD for calculation of acute WQB. These required taxa include a warm water fish, a species in the family Salmonidae, a planktonic crustacean, a benthic crustacean, and an insect (Fojut *et al*., 2012). The Canadian methodology for deriving acute WQBs further requires that at minimum of one study of a freshwater plant or algal species indigenous to North America be included. Regardless of the specific country's

criteria, as the number of data points increase (including the diversity of test species), AFs decrease and thereby reduce uncertainty and conservatism in derived criteria values. Additionally, a sample size of five is the minimum needed when employing the use of parametric statistical extrapolation techniques whereas only AF methods are appropriate for smaller data sets (TenBrook *et al*., 2008). Typically, WQBs are still derived from single-species toxicity tests. The US EPA considers these tests to be not only the most abundant, but to contain the most reliable and easily interpretable data as other studies (e.g., mesocosms, field studies) are often criticized for lack of interpretability, replication, and standardization.

6. Imidacloprid: A case study of developing WQBs around the globe

In more recent years, systemic insecticides such as the neonicotinoid imidacloprid have garnered increased attention and scrutiny due, in part, to their frequent detections in surface water systems. In response, several countries have reevaluated their own WQBs in an effort to protect freshwater resources. Importantly, the derivation of WQBs and methodology selected by many groups is highly variable. Even when similar methods are applied to derive WQBs, the data used or the approach to criteria assessment such as inclusion or exclusion of specific studies may be different. In 2017, the US EPA OPP re-calculated their imidacloprid acute and chronic benchmarks for the protection of aquatic life. The previous imidacloprid chronic benchmark was set at 1,050 ng/L (US EPA, 2014) and calculated based on a no-observed-adverse-effect level (NOAEL) for *Chironomus dilutus*; however, the test duration was not reported (US EPA, 2008). This OPP reevaluation process included studies from registrants and those available from the open literature that were classified as either qualitative or quantitative. However, no higher-tier ecological effects studies were part of this evaluation. Based on the most sensitive aquatic invertebrates, mayflies, the US EPA has calculated an acute value of 385 ng/L and a chronic value of 10 ng/L for the presence of imidacloprid in water (Table 4). These values were based on the mayfly EC_{50} of 770 ng ai/L (*C. dipterum*) divided by an AF of 2 and a chronic NOAEC of 10 ng/L for the mayfly, *C. horaria* (US EPA, 2017).

In contrast, the Central Coast Regional Water Quality Control Board (CCRWQCB) contracted the University of California - Davis to derive imidacloprid-specific aquatic life criteria for watersheds under the CCRWQCB. Of the original 41 studies evaluated, 14 acute studies yielding 32 toxicity values from 29 taxa met the criteria of "reliable and relevant" (Bower and Tjeerdema, 2018). However, as only four of the five required taxa requirements were met by the existing studies, UC Davis was unable to use an SSD to derive acute toxicity criterion. Instead, an AF of 7.5 (based on four species) was used to develop an acute value that was then divided by a factor of 2 equating to a final acute value of 70 ng/L (Table 4). Similarly, an acute-to-chronic ratio was used to calculate the chronic WQB as highly rated (both reliable and relevant) acute and chronic studies were only available for *Daphnia magna*. A final chronic value of 14 ng/L was calculated from representative studies. Using an AF to calculate a criterion involves a high degree of uncertainty as there is a potential for either under- or over-protection based on the representation of sensitive species in the available dataset (Bower and Tjeerdema, 2018).

Other governments or independent researchers have either used SSDs or a range of approaches to calculate their own imidacloprid WQB from registrant-generated studies and the open literature.

Additional studies with new taxa being evaluated are constantly becoming available so selection of a WQB is often a moving target as the science evolves. Recently, the governments of Canada, the Netherlands, and researchers from universities in Canada, Australia, and Germany have published their imidacloprid WQB. Canada's PMRA calculated an SSD for freshwater invertebrate acute and chronic endpoints resulting in values of 360 ng/L (HC_5 of LC_{50} values) and 41 ng/L (HC_5 of EC_{50} values; PMRA, 2016). These values were based on available acute and chronic toxicity endpoints for 32 (acute) and 10 (chronic) freshwater invertebrate species (PMRA, 2016). These new values are in contrast to previous imidacloprid guidelines developed by the Canadian Council of Ministers of the Environment (CCME; 2007). In 2007, an interim WQB was developed to protect freshwater organisms where an AF of 10 was used to calculate a single value of 230 ng/L (CCME, 2007). In contrast to other calculations, the Dutch government included mesocosm studies as the most ecologically relevant way of assessing exposure to determine acute values. They further justified constructing SSDs from more sensitive taxonomic groups alone as required data on macrophytes were missing; compared to insects, primary producers are not sensitive to insecticides (Smit *et al*., 2015). Importantly, these derived values still included an assessment factor of three applied to the final data indicating a combination of methods. University researchers such as Morrissey et al. (2015) have used SSDs to generate acute and chronic toxicity thresholds (i.e., WQB) for robust datasets that included 49 species of aquatic insects and crustaceans from 12 invertebrate orders. The authors calculated the lower confidence interval of HC_5 from SSDs using 137 acute (LC_{50}) and 36 chronic ($L[E]C_{50}$) toxicity tests where they used all neonicotinoid compounds weighted and standardized to imidacloprid. This resulted in an acute threshold of 200 ng/L and a chronic threshold of 35 ng/L (Morrissey *et al*., 2015).

In a partnership with environmental consultants (including Stantec, Intrinsik Environmental Sciences, and Stone Environmental), Bayer CropScience developed an independent assessment of existing acute and chronic studies of imidacloprid toxicity (see Whitfield-Aslund et al., 2017). Using an SSD derived from only acute studies that had an acceptable rating, the research team used a reduced dataset of nine studies including only the most sensitive endpoint for each included species (Whitfield-Aslund et al. 2017). If multiple studies reported suitable endpoints for the same species, geometric means replaced the specific species value as the lowest acceptable endpoint. *Daphnia magna*, being orders of magnitude less sensitive than the next highest endpoint, were removed from further consideration resulting in an acute HC_5 of 1,730 ng/L (Table 4). Two further chronic SSDs were calculated separately for both laboratory and higher-tier (mesocosm, semi-field, and field) studies resulting in chronic HC_5 values of 39 ng/L and 1,010 ng/L, respectively (Whitfield-Aslund *et al*., 2017). The authors noted that chronic imidacloprid exposure may be overestimated when the potential for recovery is not accounted for in any type of evaluation.

7. Discussion

7.1 Does the inclusion of certain taxa influence calculations of WQBs?

A major concern in developing WQBs is related to the representativeness of included test species. It is impossible to include all potential taxa to be truly representative of the majority of

aquatic ecosystems requiring protection from pesticides. Typically, test species are selected based on their management under laboratory conditions, sensitivity to toxicants, acceptable standardized testing procedures, and their ecological relevance. Numerous organizations or countries require specific representative taxa when developing WQBs. One criticism is that a truly representative sample from the aquatic environment would include at least 50% insects. However, test species are usually chosen to be representative of different taxonomic groups or trophic levels (Posthuma *et al*., 2001) despite the fact that many more "representative" species may not be sensitive to contaminants. As imidacloprid is an insecticide, much of the existing literature has focused on toxicity testing of invertebrate species, which are known to be more sensitive compared to vertebrate organisms. But, not all organisms will respond to a toxicant the same way. In their evaluation of ecological thresholds for neonicotinoids in surface waters, Morrissey et al. (2015) indicated that some standard invertebrate test species are insensitive to neonicotinoid insecticides. Despite *Daphnia magna* being the global industry standard for invertebrate toxicity testing, it is far less sensitive to neonicotinoids (orders of magnitude higher than the geometric mean for many other aquatic invertebrate species; Morrissey *et al*., 2015). The inclusion of numerous toxicity tests that evaluate less sensitive species may influence calculated values for WQBs. When calculating a more generic SSD, the European Water Framework Directive recommend that an AF of 10 is applied to account for the extrapolation from a 50% effect level to the no-effect level. It is unclear as to which AF should be used when a specific SSD is constructed for the most sensitive species groups excluding other required taxa (Smit *et al*., 2015). Additionally, the reliance on requiring specific taxa for derivation of WQBs may also eliminate certain types of statistical extrapolation such as SSDs. For instance, UCDM could not use the SSD method to calculate acute and chronic values for imidacloprid due to a lack of highly rated aquatic plant and animal toxicity data. Available imidacloprid acute and chronic datasets were further missing required values for warm water fish (Bower and Tjeerdema, 2018). Importantly, the goal of developing the UCDM was to create a method that yields statistically robust criteria, similar to the USEPA by allowing for more flexible calculation methods to accommodate pesticide datasets that are variable in both overall size as well as diversity (Fojut *et al*., 2012).

7.2 Is there a potential for bias based on the data or method that is selected?

The introduction of bias into any scientific evaluation is always a concern for researchers and regulatory agencies alike. This bias could be in the form of excluding key studies when determining WQBs, inadvertently missing existing studies or choosing to ignore research from specific sectors (e.g., industry, academia, government). To minimize bias in the data sets used for derivation of WQBs, data requirements should be specified for literature searches and data sources. This specification will ensure inclusion of relevant data (TenBrook *et al*., 2008). As many organizations and countries do not specify data requirements, it is challenging to understand what constitutes an adequate literature search and subsequent evaluation. Similarly, multiple data for a particular species should be reduced down to one data point (e.g., species mean acute value (SMAC)) for inclusion in extrapolation methods (i.e., SSD) to avoid bias based on over-representation of taxonomically similar species (TenBrook *et al*., 2010).

Additionally, there may potential of bias based on the specific method for determining the WQB. For example, selection of the most sensitive taxonomic group may inadvertently bias aquatic protection criteria to be overly conservative. In their periodic registration review of imidacloprid, the US EPA developed their aquatic life benchmark by selecting the most sensitive endpoint from an acute study to determine protection levels needed for aquatic ecosystems. Specifically, Ephemeroptera (mayflies) were selected as the most sensitive taxonomic group and the mayfly, *Caenis horaria*, was chosen as the most sensitive species from an acute study (Roessink *et al*., 2013) considered acceptable for quantitative use (US EPA, 2016). However, critics indicate that acute and chronic thresholds based on one species may be inappropriate. As the test species is not native to North America, the relevance of the chosen study is questioned. Guidelines that are based on the single most sensitive datum do not have confidence limits associated with them and although protective, it is unclear as to what degree they are over- or under-protective (TenBrook *et al*., 2008). Importantly, SSDs are not without their potential bias as they are reliant on the use of laboratory-derived data that may be biased toward studies of sensitive or tolerant species. In addition, there is a bias toward the use of mortality data despite sublethal effects also being important in determining loss of local populations (Newman *et al*., 2000).

7.3 AF vs. SSD - Is one method better?

Extrapolation methods (e.g., SSD) are generally considered more robust when compared to the use of AF (Table 5). In the Netherlands, preference is given to results from an SSD or from model ecosystem studies (i.e., mesocosms) since both of these attempt to assess ecosystem effects through a more robust approach (Smit *et al*., 2015). If large data sets or those based on model ecosystems are unavailable, AFs provide a method for determining WQB from limited available data. However, although AFs are conservative and have a low probability of underestimating risk, one drawback is that AFs may greatly increase the possibility of overestimating risk based on the data used and chosen factor (TenBrook *et al*., 2008). Many states and countries use a range of factors with some up to 1,000 for chronic values (TenBrook *et al*., 2010). Therefore chosen AFs should be based on scientific knowledge (e.g., context for extrapolation, data limitations) and most AF values should not exceed 10 in order to reduce the potential for overprotection (Chapman *et al*., 1998). It should be noted that there are potential consequences when relying on one data point to aid in decision making by regulators. As mentioned above, the selection of the most sensitive datum may have issues concerning regional relevance of the test organism or the quality of the scientific study itself. Selection of the single sensitive datum may, therefore, be less representative of an actual aquatic system regardless of the level of protection provided.

Although SSDs appear more transparent and are often more scientifically defensible than AFs, one drawback is that SSDs do require larger data sets (Table 5). The number of data points included in development of an SSD is critical as are the conclusions that are based on them. By using an entire data set, confidence limits can be calculated for derived criteria. This is not possible with less reliable methods such as the use of the most sensitive data point (TenBrook *et al*., 2008). Despite potential issues of reliability, the approach of using the most sensitive data point is currently used by the US EPA in calculating its Aquatic Life Benchmarks (ALB). By

using the most sensitive datum, ALB are easily calculated which can be readily used to interpret water monitoring data. However, SSDs also have disadvantages as their included test species are not considered randomly sampled and critics argue SSDs are not considered more reliable than alternatives (e.g., single datum ALB; Posthuma *et al*., 2001). In particular, SSDs may be constructed from data that are perceived to be biased. As acute data sets are typically used for extrapolation, available lab data may have used test organisms biased toward sensitive or insensitive species that are not truly representative of local field conditions or ecosystems. Extrapolation based on single-species toxicity data is unable to account for ecological interactions or higher ecosystem-level events (Posthuma *et al*., 2001). Similar to AFs, extrapolation methods may also suffer from over- or under-estimating risk, but this uncertainty is greatly reduced when larger data sets are used (TenBrook *et al*., 2008).

In order to determine true environmental exposure, field studies are typically the most useful, followed by mesocosm/microcosm experiments, multi-species laboratory tests, and single-species laboratory tests. Unfortunately, although field or semi-field experiments are considered the best determinant of environmental exposure, they are often criticized for their lack of replication, poor standardization, and the challenge of interpretation (TenBrook *et al*., 2008). Therefore, single-species toxicity tests have become more heavily relied upon for WQBs due to their abundance, reliability, and ease of interpretation.

7.4 The inclusion of the "safety factor"

Data uncertainty such as extrapolations from the laboratory to the field or the limited availability of toxicity studies often leads to inclusion of a safety factor when determining a WQB. This may occur with both AF (which is in and of itself a safety factor) and SSD approaches. Safety factors are designed to account for uncertainty from experimentally derived numbers that are used to predict a real-world outcome (TenBrook et al., 2008). Many groups would argue that to protect all species, it is necessary to apply a safety factor to account for the unknown relative sensitivity of various test species (Elmegaard and Jagers op Akkerhuis, 2000). Dividing by a factor (e.g., 10, 100) is therefore used to estimate a safe level for a pesticide in the environment. Although these approaches have no relevance to *actual* uncertainty, they greatly reduce the potential for underestimating risk (Chapman et al., 1998). However, they are not without their criticism as safety factors are almost always generalized even if they are based on scientific data. Though safety factors are used to be protective, overestimation of risk is a strong possibility. It is plausible that being overly conservative may indicate a toxicological concern that is not based on scientific data alone. For example, if data are available for several species, organizations such as the OECD or US EPA typically select the most sensitive species to determine safety factors that are then applicable to all species (Chapman et al., 1998). Additionally, depending on the pesticide of interest, the most sensitive species may be a moving target. As chronic datasets are often less available, many SSDs also use a safety factor to account for missing data or an inadequate representation by specific taxa (TenBrook et al., 2008; Smit et al., 2015). The size of the safety factor is typically proportional to the amount of data that is available with larger safety factors being applied when data are few (Elmegaard and Jagers op Akkerhuis, 2000). Importantly, selection of a safety factor is typically a policy decision rather than a science-based

decision due to data insufficiency when extrapolating from a known to unknown circumstance (e.g., varied exposure durations; Chapman et al., 1998). Whenever possible, appropriate data should be used over safety factors and extrapolation requires context since it is neither certain nor absolute (Chapman et al., 1998).

8. Conclusion

The development of WQBs is a useful tool for regulatory agencies when determining toxicity thresholds that could be used to prevent adverse impacts on aquatic environments. Currently, however, there is no "one size fits all" approach that could be universally adopted by government or regulatory organizations. Instead, there are many different methods behind WQB development including use of AF, SSD, or selecting the most sensitive species and endpoint to be protective of a greater number of species. Each has its own scientific merit and associated limitations. Importantly, regulatory decision-making scenarios are often influenced by a range of issues and may ultimately be a reflection of geographic scope, regional pesticide use, and local or national concerns (e.g., agricultural productivity, pest management). Rather than suggest that any one method for developing WQBs is superior to another, a calculated protective value should ultimately be tailored to the goals of the specific regulatory group. As many stakeholders are likely to be impacted by development and regulatory applications of WQBs, regulators should carefully weigh the factors that contribute to pesticide exceedances in aquatic ecosystems.

Table 1. The use of Assessment Factors (AF) in existing methodologies by country or organization. Data presented are adapted from TenBrook *et al*. (2010).

Guidelines	Only use AF	Based on lowest toxicity value	NOEC or LOEC?	Acute: Factor value*	Chronic: Factor value*	Default ACR
North America						
Canada[1]	X		LOEC	NA	10-100	10
USEPA			--	2	NA	2
Europe						
EU[1]			NOEC	NA	1-1000	10
France[1]	X	X	NOEC	NA	1-1000	
Germany[1]	X	X	NOEC	NA	10-1000	10
Netherlands[1]			NOEC	NA	1-1000	
Spain[1]	X	X	NOEC	NA	1-100	
UK	X	X	NOEC	2-10	1-100	
Africa						
South Africa			--	1-100	1-1000	
Australia						
Australia/NZ			NOEC	NA	10-1000	10+
Organizations						
OECD			NOEC	100-1000	10	

*Value ranges are listed - chosen factors are ultimately dependent on types of data available

[1] Factor is applied to LOEC/NOEC value as specified or to LC/EC_{50} values if other values are unavailable

-- indicates data is missing or was unclear in the literature

Table 2. Description of the physical-chemical data requirements of various water quality guidelines developed around the world. An X indicates that the information is required and/or specified as being of importance during an evaluation of toxicity data.

	Physical-chemical requirements												
Guidelines*	Molecular weight	Molar mass	K_{OW}	Solubility	Melting point	Vapor pressure	Henry's Law	pK_a	K_p	K_{SW}	Degradation Info	BCF	EnvFate
North America													
Canada													X
UC Davis	X			X	X	X	X	X				X	X
USEPA				X		X					X	X	
Europe[1]													
Denmark	X		X	X			X				X	X	
EU	X		X								X		
Germany												X	
Netherlands		X	X	X	X	X	X	X	X		X		
Australia													
Australia/NZ			X									X	
Organizations													
OECD	X		X	X	X			X		X			

*Physical-chemical data is not required of all guidelines during evaluation of toxicity data.
[1]France, Spain, and the UK do not specify physical-chemical data in criteria derivation OR require data that specifically influences toxicity not presented here.

Table 3. An outline of the number of studies required per specific category of taxa when evaluating toxicity data to determine water quality guidelines. Some guidelines specify a minimum number of chronic NOEC (No Observable Effect Concentration) data points that must be included in evaluations. The – indicates that no value has been specified. FCV = final chronic value based on criteria from the EPA Office of Water.

	Min. Chronic	No. Required	**Number of studies required per category of taxa**						
Guidelines*	NOEC	Different Taxa	Algae/Plants[a]	Bacteria[a]	Crustaceans[a]	Fish[a]	Insects[a]	Other Phylum[b]	Unspecified[b]
North America									
Canada[1]	4	5	1		1	3	1		
UC Davis	5	5			2	2	1		
USEPA[2]	8	8			2	2	2	2	
USEPA (FCV)	3	5	1			1	2	1	
Europe									
EU	10	8	2		1	2	2	1	
France	2	3	1			1	1		
Germany	2	4	1	1	1	1			
Netherlands	4	4							4
Spain	--	3	1			1	1		
UK	--	4	1		1	1	1		
Africa									
South Africa[3]	8	8			2	2	2	2	
Australia									
Australia/NZ	3	5							5
Organizations									
OECD	5	8			2	2	2	2	

*individual guidelines often vary substantially per country or organization.
[1]fish species must be residents of N. America; if the chemical is phytotoxic, at least four algae/plant studies are required.
[2]chronic NOEC values are required for at least eight animal families.
[3]species must be indigenous to S. Africa and/or be of commercial or cultural importance to the country.
[a]there are specific requirements for some categories depending on the guideline, but at least one study is required from this category.
[b]these categories typically specify a different family or taxa from others already represented OR there are no specific taxa required.

Table 4. A description of the key imidacloprid water quality guidelines calculated by government agencies, industry, and educational institutions. Information presented is divided into acute and chronic values with the derivation method indicated as either assessment factor (AF), species sensitivity distribution (SSD), or the most sensitive species endpoint. The number of species (No. Species) evaluated in each guideline is listed.

Authors	Year	AI grade	Acute (ng/L)	Taxon	Derivation method(s)	No. Species	Origin
Smit et al.	2015	Any	200	Invertebrate	AF, SSD + AF, mesocosm	31	Netherlands
Morrissey et al.	2015	Any	200	Invertebrate	SSD	42	Canada
PMRA Canada	2016	--	360	--	SSD	32	Canada
Whitfield-Aslund	2017	Any	1,730	Invertebrate	SSD	11	USA
US EPA	2017	Any	385	Single species	most sensitive species, AF	1	USA
Raby et al.	2018	AI	1,080	Invertebrate	SSD	42	Canada
UCD Criterion	2018	AI	70	All	AF	--	California

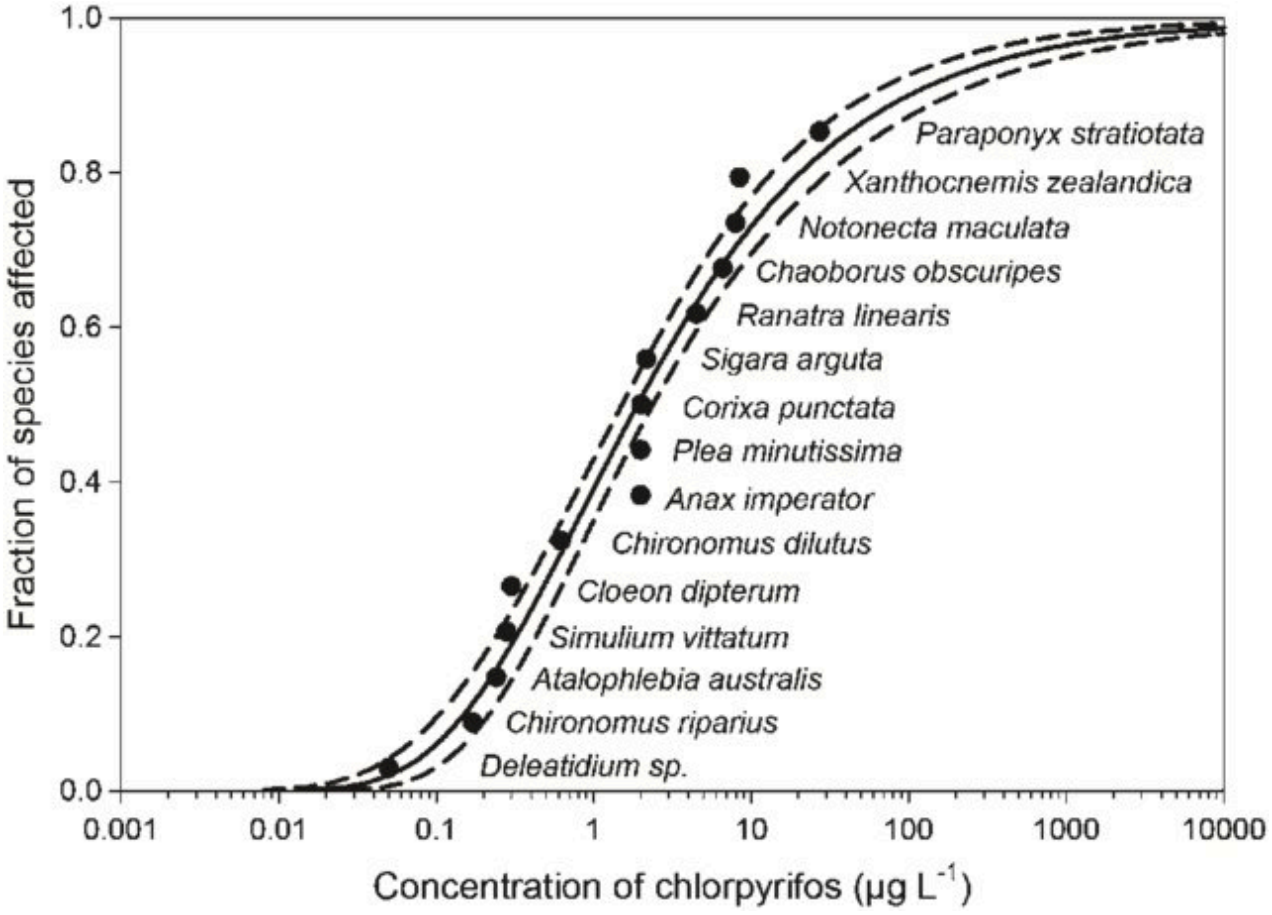
Authors	Year	AI grade	Chronic (ng/L)	Taxon	Derivation method(s)	No. Species	Origin
Smit et al.	2015	Any	8.3	Invertebrate	AF, SSD + AF	19	Netherlands
Morrissey et al.	2015	Any	35	Invertebrate	SSD, low CI	18	Canada
PMRA Canada	2016	--	40	--	SSD	10	Canada
Whitfield-Aslund	2017	Any	39	Invertebrate	SSD	11	USA
Whitfield-Aslund	2017	Any	1,010	Invertebrate	Taxon sensitivity distribution	15	USA
US EPA	2017	Any	10	Single species	most sensitive species, AF	1	USA
UCD Criterion	2018	AI	14	All	AF	1	California

-- indicates data is missing or was unclearly reported

Table 5. The major "pros" and "cons" of choosing an AF versus an SSD approach to developing WQBs.

AF	Pro	Con
	Can be calculated with one species	Test species are often limited
	Conservative estimate - likely protective of a range of species	Not always clear as to which AF value should be used
	Useful when large data sets are unavailable for a pesticide of interest	Selection of the most sensitive taxa may bias aquatic protection to be overly conservative
	Less prescriptive than SSDs	Often based on policy rather than science
		May be overly protective based on extrapolation of data or biased based on benchmark being derived from one value.
SSD	**Pro**	**Con**
	Derived from a range of taxa	Requires more data points than an AF
	Generally considered more robust than AF methods	Included species do not always reflect all aquatic environments
	Confidence intervals can be calculated	Test species are not randomly sampled
	Conclusions drawn are often more representative of an aquatic system	Reliant on lab-derived data that may be biased toward sensitive/tolerant species
		Results may be biased depending on chosen species to include in models
		May still require the addition of a safety factor to balance missing data

Figure 1. An example of a Species Sensitivity Distribution (SSD) evaluating the effect of the pesticide chlorpyrifos on aquatic insects. Extracted from Giddings et al. (2014).



References

Bower, J.C., and Tjeerdema, R.S. (2018). Water Quality Criteria Report for Imidacloprid. Phase III: Application of the pesticide water quality criteria methodology. Retrieved from: https://www.waterboards.ca.gov/centralcoast/water_issues/programs/tmdl/pesticide_criteria.html

Canadian Council of Ministers of the Environment (CCME; 1999). A protocol for the derivation of water quality guidelines for the protection of aquatic life. Canadian Environmental Quality Guidelines. Canadian Council of Ministers of the Environment, Ottawa.

Canadian Council of Ministers of the Environment (CCME; 2007). Canadian water quality guidelines for the protection of aquatic life: Imidacloprid. In: Canadian environmental quality guidelines, 1999, Canadian Council of Ministers of the Environment, Winnipeg.

Central Coast Regional Water Quality Control Board. (CCRWQCB; 2017). Water Quality Control Plan for the Central Coastal Basin, September 2017 Edition. California Environmental Protection Agency. Retrieved from: http://www.waterboards.ca.gov/centralcoast/publications_forms/publications/basin_plan/index.shtml

Chapman PM., Fairbrother A., and Brown D (1998). A critical evaluation of safety (uncertainty) factors for ecological risk assessment. *Environmental Toxicology and Chemistry*, 17, 99-108.

Elmegaard, N., and Jagers op Akkerhuis, G.J.A.M. (2000). Safety factors in pesticide risk assessment. Differences in species and acute-chronic relations. National Environmental Research Institute, Silkeborg, Denmark. 60 pp. – NERI Technical Report No. 325. Retrieved from: https://www.dmu.dk/1_viden/2_publikationer/3_fagrapporter/rapporter/fr325.pdf

Fojut, T. L., Palumbo, A. J., and Tjeerdema, R. S. (2012). Aquatic life water quality criteria derived via the UC Davis method: II. Pyrethroid insecticides. In *Aquatic life water quality criteria for selected pesticides* (pp. 51-103). Springer, Boston, MA.CWC (water quality)

Giddings, J., Williams, W.M., Solomon, K.R., and Giesy, J.P. (2014). Risks to Aquatic Organisms from Use of Chlorpyrifos in the United States. *Reviews of Environmental Contamination and Toxicology*, 231, 119-162.

Morrissey, C. A., Mineau, P., Devries, J. H., Sanchez-Bayo, F., Liess, M., Cavallaro, M. C., and Liber, K. (2015). Neonicotinoid contamination of global surface waters and associated risk to aquatic invertebrates: a review. *Environment international*, *74*, 291-303.

Newman, M.C., Ownby, D.R., Mézin, L.C.A., Powell, D.C., Christensen, T.R.L., Lerberg, S.B., and Anderson, B-A. (2000). Applying species-sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient number of species. *Environmental Toxicology and Chemistry*, 19, 508-515.

Pest Management Regulatory Agency. (PMRA; 2016) *Proposed Re-evaluation Decision – Imidacloprid* (Publication No. PRVD2016-20). Retrieved from: http://publications.gc.ca/site/archivee-archived.html?url=http://publications.gc.ca/collections/collection_2016/sc-hc/H113-9-2016-16-eng.pdf

Posthuma, L., Suter II, G. W., and Traas, T. P. (2001). *Species sensitivity distributions in ecotoxicology*. CRC press.

Rio Convention (1992) United Nations Conference on Environment and Development: Rio Declaration on Environment and Development, June 14, 1992. Reprinted in Intl. Legal Materials 31: 874–879.

Roessink, I., Merga, L. B., Zweers, H. J., and Van den Brink, P. J. (2013). The neonicotinoid imidacloprid shows high chronic toxicity to mayfly nymphs. *Environmental Toxicology and Chemistry*, *32*, 1096-1100.

Smit, C. E., Posthuma-Doodeman, C. J. A. M., Van Vlaardingen, P. D., and De Jong, F. M. W. (2015). Ecotoxicity of imidacloprid to aquatic organisms: derivation of water quality standards for peak and long-term exposure. *Human and Ecological Risk Assessment: An International Journal*, *21*, 1608-1630.

State Water Resources Control Board. 2016. A Compilation of Water Quality Goals, 17th Edition, January 2016. California Environmental Protection Agency. Retrieved from: https://www.waterboards.ca.gov/water_issues/programs/water_quality_goals/

TenBrook, P. L., Tjeerdema, R. S., Hann, P., and Karkoski, J. (2008). Methods for deriving pesticide aquatic life criteria. In *Reviews of Environmental Contamination and Toxicology Volume 199* (pp. 1-92). Springer, Boston, MA.

TenBrook, P. L., Palumbo, A. J., Fojut, T. L., Hann, P., Karkoski, J., and Tjeerdema, R. S. (2010). The University of California-Davis Methodology for deriving aquatic life pesticide water quality criteria. In *Reviews of Environmental Contamination and Toxicology Volume 209* (pp. 1-155). Springer, New York, NY.

US EPA. (1985). Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses. Retrieved from: https://www.epa.gov/sites/production/files/2016-02/documents/guidelines-water-quality-criteria.pdf

US EPA. (2004). Overview of the Ecological Risk Assessment Process in the Office of Pesticide Programs, U.S. Environmental Protection Agency: Endangered and Threatened Species Effects Determinations. Retrieved from: https://www.epa.gov/endangered-species/ecological-risk-assessment-process-under-endangered-species-act

US EPA. (2008). Problem Formulation for the Registration Review of Imidacloprid. Retrieved from: https://www.regulations.gov/document?D=EPA-HQ-OPP-2008-0844-0003

US EPA. (2014). OPP Pesticide Toxicity Database.

US EPA. (2016). Preliminary Aquatic Risk Assessment to Support the Registration Review of Imidacloprid. Retrieved from: https://www.regulations.gov/document?D=EPA-HQ-OPP-2008-0844-1086

USGS (2019). *Water-Quality Benchmarks for Contaminants*. Retrieved from: https://www.usgs.gov/mission-areas/water-resources/science/water-quality-benchmarks-contaminants?qt-science_center_objects=0#qt-science_center_objects

Whitfield-Aslund, M., Winchell, M., Bowers, L., McGee, S., Tang, J., Padilla, L., Greer, C., Knopper, L., and Moore, D. R. (2017). Ecological risk assessment for aquatic invertebrate communities exposed to imidacloprid as a result of labeled agricultural and nonagricultural uses in the United States. *Environmental Toxicology and Chemistry*, *36*, 1375-1388.