

Methodology for Derivation of Pesticide Water Quality Criteria for the Protection of Aquatic Life in the Sacramento and San Joaquin River Basins

Phase I: Review of Existing Methodologies



Prepared for the Central Valley Regional Water Quality Control Board

Patti L. TenBrook, Ph.D
and
Ronald S. Tjeerdema, Ph.D.

Department of Environmental Toxicology
University of California, Davis

April 2006

**Methodology for Derivation of Pesticide Water Quality Criteria for the Protection
of Aquatic Life in the Sacramento and San Joaquin River Basins. Phase I: Review of
Existing Methodologies**

Final Report Prepared for the Central Valley Regional Water Quality Control Board

Patti L. TenBrook

and

Ronald S. Tjeerdema

Department of Environmental Toxicology

University of California, Davis

April 2006

Executive Summary

The goal of this project is to develop a methodology for derivation of pesticide water quality criteria for the protection of aquatic life in the Sacramento River and San Joaquin River basins. The project will be accomplished in three phases. This is a report of the results of Phase I, which is a comparison and evaluation of existing criteria derivation methodologies from around the world. Phase II will be development of the new criteria derivation methodology. Phase III will be to apply the new methodology to derive criteria for up to five pesticides including diazinon and chlorpyrifos, two organophosphate insecticides of particular concern in the Sacramento and San Joaquin River basins due to listings under 303(d) of the federal Clean Water Act.

The approach for Phase I was to conduct an extensive literature search to find 1) criteria derivation methodologies currently in use, or proposed for use, throughout the world; 2) original studies supporting the methodologies; 3) proposed modifications of existing methodologies; and 4) relevant research in ecotoxicology and risk assessment. Based on literature discussing recent scientific thinking on water quality criteria derivation, a list was developed containing components to consider in evaluation and development of a water quality criteria derivation methodology. These components are discussed with respect to how they are, or are not, addressed by existing criteria derivation methodologies. Included in the discussion are methodologies from: (listed alphabetically) Australia/New Zealand, Canada, Denmark, the European Union/European Commission (EU/EC), France, Germany, The Netherlands, the Organization for Economic Co-operation and Development (OECD), South Africa, Spain, the United Kingdom (UK), and the United States (US), including the Great Lakes Region, and a few individual states whose methodologies diverge somewhat from USEPA (1985) guidance. This project is focused on development of pesticide criteria and so the review of methodologies is likewise focused on pesticides.

This report includes a brief discussion of water quality policy as it pertains to criteria derivation. Different types, levels or tiers of criteria that may be developed to satisfy policy requirements are described. The question of what levels of ecosystem organization must be protected to meet water quality goals is addressed, as is the importance of having some level of confidence that derived criteria will achieve those goals.

Derivation of scientifically sound criteria depends on the use of adequate amounts of high quality ecotoxicity data from diverse taxonomic groups. Ecotoxicity and physical-chemical data issues are reviewed including data sources and literature searches, data quality, data quantity, data kinds (physical-chemical, quantitative structure activity relationships, acute vs. chronic, hypothesis tests vs. regression analyses, single-species vs. multispecies, traditional vs. non-traditional endpoints, quantitative species sensitivity relationships), and data reduction.

Various aspects of exposure that are related to toxicity are discussed including magnitude, duration and frequency considerations, multipathway exposure, and water quality characteristics that affect toxicity. Two basic criteria derivation methodologies are discussed and critiqued: assessment factor (AF) methods and species sensitivity distribution (SSD) methods. Application of these methods by existing criteria derivation guidelines is described. Criteria calculation issues that are addressed include derivation and justification of assessment factors, degree of aggregation of taxa, selection of an appropriate distribution and an appropriate percentile level in SSDs, and confidence limits. Other considerations in criteria derivation include mixtures and multiple stressors, bioaccumulation and secondary poisoning, threatened and endangered species, harmonization and cross-media coherence of criteria, utilization of data, and encouragement of data generation. A brief discussion of criteria derivation guideline formats is presented, followed by conclusion section.

Three possible outcomes of this project are: 1) make no change in criteria derivation methodology (i.e. continue using the USEPA 1985 guidance); 2) adopt one of the other existing methodologies, or; 3) develop an entirely new methodology. Based on this review, the third outcome is most likely. This review has revealed that no single existing methodology is ideal, but elements of several of them could be combined, along with some newer risk assessment tools, into a usable, flexible criteria derivation procedure that will produce protective criteria. Phase II of this project will involve further exploration of the various elements and models presented here to determine which are appropriate for the new methodology. Among the reviewed methodologies, those from Australia/New Zealand (ANZECC & ARMCANZ 2000), The Netherlands (RIVM 2001) and the Great Lakes (USEPA 2003a) are recommended for comparison to the new methodology in Phase III of this project.

List of Tables

Table 1. Components to be addressed by water quality criteria derivation methodology	3
Table 2. Overview of major methodologies	5
Table 3. Definitions of levels of organization	10
Table 4. Overview of similarities and differences between key elements of six major criteria derivation methodologies.	

List of Figures

Figure 1. Generic illustration of SSD technique	59
Figure 2. Comparison of log-normal, log-logistic and log-triangular distributions	60

List of Acronyms and Abbreviations

AA	Annual Average
ACE	Acute-to-Chronic Estimation
ACR	Acute to Chronic Ratio
AEV	Acute Effect Value
AF	Assessment Factor
ANZECC	Australia and New Zealand Environment and Conservation Council
ARMCANZ	Agriculture and Resource Management Council of Australia and New Zealand
ASTM	American Society for Testing and Materials
BAF	Bioaccumulation Factor
BC	British Columbia
BCF	Bioconcentration Factor
CAS	Chemical Abstract Service
CCC	Criterion Continuous Concentration
CCME	Canadian Council of Ministers of the Environment
CDFG	California Department of Fish and Game
CEV	Chronic Effect Value
CMC	Criterion Maximum Concentration
CVRWQCB	Central Valley Regional Water Quality Control Board
CSTE/EEC	Scientific Advisory Committee on Toxicity and Ecotoxicity of Chemicals/European Economic Community
CWA	Clean Water Act
DTA	Direct Toxicity Assessment
EC	European Commission
EC _x	Concentration that affects x% of exposed organisms
ECL	Environmental Concern Level
ERL	Environmental Risk Level
EINECS	European Inventory of Existing Commercial Substances
EqP	Equilibrium Partitioning
EQS	Environmental Quality Standard
EU	European Union
FACR	Final Acute to Chronic Ratio
FAV	Final Acute Value
FCV	Final Chronic Value
FPV	Final Plant Value
FRV	Final Residue Value
GMAV	Genus Mean Acute Value
HC _x	Hazardous Concentration potentially harmful to x% of species
ICE	Interspecies Correlation Estimation
IUPAC	International Union of Pure and Applied Chemistry
K _H	Henry's law constant
K _{ow}	Octanol-water partition coefficient
K _p	Solid-water partition coefficient
LC _x	Concentration lethal to x% of exposed organisms

LOEC	Lowest Observed Effect Concentration
LOEL	Lowest Observed Effect Level
MAC	Maximum Allowable Concentration
MATC	Maximum Acceptable Toxicant Concentration
MPC	Maximum Permissible Concentration
msPAF	Multispecies Potentially Affected Fraction
MTC	Maximum Tolerable Concentration
NC	Negligible Concentration
NCDENR	North Carolina Department of Environment and Natural Resources
NOEC	No Observed Effect Concentration
OECD	Organization for Economic Co-operation and Development
parNEC	Parametric No Effect Concentration
pK _a	Acid dissociation constant
PNEC	Probable No Effect Concentration
QSAR	Quantitative Structure Activity Relationship
QSSR	Quantitative Species Sensitivity Relationship
QT	Quality Target
RI	Reliability Index
RIVM	National Institute of Public Health and the Environment, Bilthoven, The Netherlands
RPF	Relative Potency Factor
RWQCB	Regional Water Quality Control Board
SACR	Secondary Acute to Chronic Ratio
SAV	Secondary Acute Value
SCC	Secondary Chronic Concentration
SCV	Secondary Chronic Value
SETAC	Society of Environmental Toxicology and Chemistry
SMC	Secondary Maximum Concentration
SMCV	Species Mean Chronic Value
SMAV	Species Mean Acute Value
SRC _{ECO}	Ecosystem Serious Risk Concentration
SSD	Species Sensitivity Distribution
SWRCB	State Water Resources Control Board
TES	Threatened and Endangered Species
TEF	Toxic Equivalency Factor
TGD	European Union's Technical Guidance Document on Risk Assessment
TMDL	Total Maximum Daily Load
TRG	Tissue Residue Guideline
TSD	Technical Support Document for Water Quality-based Toxics Control
TV	Trigger Value
UK	United Kingdom
US	United States
USEPA	United States Environmental Protection Agency
VROM	Ministry of Housing, Spatial Planning and Environment, The Hague, The Netherlands
WCS	Water-based Criteria Subcommittee

WER Water Effect Ratio
WFD Water Framework Directive

1.0 Introduction

The goal of this project is to develop a methodology for derivation of pesticide water quality criteria for the protection of aquatic life in the Sacramento and San Joaquin River basins. The surface waters of these basins receive pesticide inputs in runoff and drainage from agriculture, silviculture, and residential and industrial storm water (CVRWQCB 2004). The term pesticide is defined by the Central Valley Regional Water Quality Control Board (CVRWQCB 2004) as (1) any substance, or mixture of substances which is intended to be used for defoliating plants, regulating plant growth, or for preventing, destroying, repelling, or mitigating any pest, which may infest or be detrimental to vegetation, man, animals, or households, or be present in any agricultural or nonagricultural environment whatsoever, or (2) any spray adjuvant, or (3) any breakdown products of these materials that threaten beneficial uses.

The project will be accomplished in three phases. This is a report of the results of Phase I, which is a comparison and evaluation of existing criteria derivation methodologies from around the world. Phase II will be development of the criteria derivation methodology. The new methodology may simply be one of the existing methodologies, a combination of features from existing methodologies, or an entirely new methodology based on the latest available research in aquatic ecotoxicology and environmental risk assessment. Phase III will be to apply the new methodology to derive criteria for up to five pesticides including diazinon and chlorpyrifos, two organophosphate insecticides of particular concern in the Sacramento and San Joaquin River basins due to listings under 303(d) of the federal Clean Water Act (CWA; CVRWQCB 2002).

The mission of California's nine Regional Water Quality Control Boards (RWQCB) is "to develop and enforce water quality objectives and implementation plans which will best protect the beneficial uses of the State's waters, recognizing local differences in climate, topography, geology and hydrology" (California SWRCB 2005). Toward that mission, each RWQCB is responsible for development of a "basin plan" for its hydrologic area. The "Water Quality Control Plan (Basin Plan) for the Sacramento River and San Joaquin River Basins," (CVRWQCB 2004) contains the following language regarding toxic substances in general, and pesticides in particular:

"...waters shall be maintained free of toxic substances in concentrations that produce detrimental physiological responses in human, plant, animal, or aquatic life."

"No individual pesticide or combinations of pesticides shall be present in concentrations that adversely affect beneficial uses."

"Discharges shall not result in pesticide concentrations in bottom sediments or aquatic life that adversely affect beneficial uses."

"Pesticide concentrations shall not exceed the lowest levels technically and economically achievable."

Table III-2A of the basin plan lists specific pesticide objectives for diazinon of 0.080 µg/L as a 1-h average and 0.050 µg/L as a 4-d average. Neither objective is to be exceeded more than once every three years on average. These objectives are based on aquatic life criteria for diazinon, which were derived by the California Department of Fish and Game (CDFG; Siepmann & Finlayson 2000) following US Environmental Protection Agency guidance (USEPA 1985). No other specific pesticide objectives have been adopted although numeric criteria have been developed for chlorpyrifos (Siepmann & Finlayson 2000). The CVRWQCB would like to develop specific objectives for more pesticides to provide clear goals for permitting and Total Maximum Daily Load (TMDL) programs. This project will provide a methodology to derive numeric criteria which may be used as specific pesticide objectives for the Sacramento and San Joaquin River basins.

The approach for Phase I was to conduct an extensive literature search to find 1) criteria derivation methodologies currently in use, or proposed for use, throughout the world; 2) original studies supporting the methodologies; 3) proposed modifications of existing methodologies; and 4) relevant and recent research in ecotoxicology and risk assessment. Four documents were found that provide a good overview of the latest scientific thinking in the field of water quality criteria derivation. First is a book, “Reevaluation of the State of the Science for Water-Quality Criteria Development” (Reilly et al. 2003) which is a report of conclusions reached by participants in a Society of Environmental Toxicology and Chemistry (SETAC) Pellston workshop. Second is the “Draft Report on Summary of Proposed Revisions to the Aquatic Life Criteria Guidelines” (USEPA 2002a). Third is a report from the United Kingdom (UK) Environment Agency called “Derivation and Expression of Water Quality Standards, Opportunities and Constraints in Adopting Risk-Based Approaches in EQS Setting” (EQS: environmental quality standard; Whitehouse et al. 2004). Finally, is a report from the Fraunhofer-Institute Molecular Biology and Applied Ecology, prepared on behalf of the European Commission (EC), called “Towards the Derivation of Quality Standards for Priority Substances in the Context of the Water Framework Directive” (Lepper 2002). Information in these reports, as well as conversations with state and federal regulators (Karkoski pers. comm. 2005; Denton pers. comm. 2005), were used to construct Table 1, which is a list of components to consider in evaluation and development of a water quality criteria derivation methodology.

In this report, the components listed in Table 1 are discussed with respect to how they are, or are not, addressed by existing criteria derivation methodologies. Included in the discussion are methodologies from: (listed alphabetically) Australia/New Zealand, Canada, Denmark, the European Union/European Commission (EU/EC), France, Germany, The Netherlands, the Organization for Economic Co-operation and Development (OECD), South Africa, Spain, the United Kingdom (UK), and the United States (US), including the Great Lakes Region, and a few individual states whose methodologies diverge somewhat from USEPA (1985) guidance. In some cases original documents were not available in English, but other resources containing summaries of those documents were available and were used for this report. Existing methodologies are

Table 1. Components to be addressed by a water quality criteria derivation methodology.

Category	Component (listed alphabetically)	Reference
Criteria types/uses (Section 4.0)	One type/level of criterion vs. multiple types/levels Use in regulatory programs	1,3,4,5 1,2,3,4
Protection level (Section 5.0)	Economically, ecologically, recreationally important species Ecosystem function and structure Individuals vs. populations Justification of percentile levels (i.e. 10 th , 5 th , 1 st) Probability of over- or underprotection	6 1,3,4 3 2,3,4 1,2,3,4
Ecotoxicity and physical-chemical data (Section 6.0)	Data quality and quantity Acceptability criteria Minimum data set Minimum literature search Taxa number and diversity Ecological relevance Kinds of data Acute (LCx/ECx, NOEC) vs. chronic (EC _x , NOEC) Ecosystem, field, semi-field, laboratory Multispecies vs. single-species Traditional vs. non-traditional endpoints	2,3,4 1,2,3,4 6 1,2,3 3,4 1,2,3,4 1,2,3,4 1,4 1,2,3,4
Criteria calculation (Section 7.0)	Bioaccumulation/secondary poisoning Community/ecosystem/population/foodweb models Confidence limits for criteria/explicitly stated uncertainty Degree of aggregation of taxa Derivation and justification of assessment/uncertainty factors Encouragement data generation Environmental fate of chemicals Exposure considerations Bioavailability Short-term/acute (including pulse) and long-term/chronic Magnitude, duration and frequency Monitoring considerations Recovery from toxic events Harmonization/coherence across media Incorporation of physical-chemical data Kinetic-based modeling/time to event analysis Mixtures/multiple stressors Multipathway (e.g. dietary) exposure Plants and animals combined vs. separate Risk assessment approach Separate acute and chronic criteria vs. single criterion Site-specificity Small data sets Species Sensitivity Distributions (SSD) Toxicant mode of action Threatened and endangered species Wildlife Utilization of available data	1,2,4 1,3,4 1,2,3,4 2 2,3,4,5 3,4 1,4 1,4 1,2,3,4 1,2,3,4 4 1,2 1,4 1,2,4 1,2,3 1,2,4 1,2 1,2,4 1,2,3,4 1,2 1,2,3 1,2,5 1,2,3,4 1,4 1,2 1,4 3,4,5
Guideline format (Section 8.0)	Understandable, navigable, usable	2

1. Reiley et al. 2003

2. USEPA 2002a

3. Whitehouse et al. 2004

4. Lepper 2002

5. Pers. comm. (Karkoski 2005; Denton 2005)

6. Not part of discussions in references 1-4, but part of existing criteria derivation methodologies in the US (USEPA 1985), Australia/New Zealand (ANZECC & ARMCANZ 2000), and/or The Netherlands (RIVM 2001)

evaluated against recent research and reports on criteria derivation techniques. In addition to pesticides, most of the methodologies address toxicity due to metals and other inorganic chemicals (e.g. ammonia), and non-pesticide organic chemicals. This project is focused on development of pesticide criteria and so the review of methodologies is likewise focused on pesticides. Some of the latest recommendations for water quality criteria derivation methodologies are simply not technically feasible at this time due to lack of data or lack of agreement among experts on techniques. However, thorough discussions of feasibility of approaches are beyond the scope of Phase I of this project and will be reserved for Phase II (development of a methodology for the Sacramento and San Joaquin River basins).

2.0 Summary of major methodologies reviewed

Many existing methodologies are discussed in this report, but the focus is on a few that either are widely accepted and used (USEPA 1985, RIVM 2001--updated from VROM 1994), represent unique approaches (CCME 1999), or are newer methodologies that incorporate and improve upon the best features of prior methodologies (ANZECC & ARMCANZ 2000, USEPA 2003a). The European Union's Technical Guidance Document on Risk Assessment (TGD) is also a focus because it represents the latest European guidance on derivation of predicted no effect concentrations for risk assessments, and EU member nations are starting to use the TGD for derivation of water quality criteria (Traas, pers. comm.).

Table 2 lists the six major methodologies, the types of criteria that are derived from them, and how the criteria are used. The USEPA methodology (1985) utilizes a statistical extrapolation procedure to derive criteria (section 7.2.2), while the Canadian methodology (CCME 1999) utilizes an assessment factor approach (section 7.2.1). All of the others utilize a combination of these two basic criteria derivation methods.

3.0 Water quality policy

It is important to note that different countries of the world have different environmental policies, which are reflected in their water quality criteria derivation methodologies. The European Union's (EU) Water Framework Directive (WFD) is the policy document guiding water quality protection efforts for EU member states. The WFD is a policy that is intended to "...contribute to pursuit of the objectives of preserving, protecting and improving the quality of the environment, in prudent and rational utilization of natural resources, and to be based on the precautionary principle and on the principles that preventive action should be taken, environmental damage should, as a priority, be rectified at source and that the polluter should pay." The precautionary principle may be summed up as follows:

Table 2. Overview of major methodologies

Method Title	Source	Year	Country	Criterion	Criterion description
Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses	USEPA	1985	United States	CMC: criterion maximum concentration	Used for setting water quality standards, setting discharge limits, and other regulatory programs; for protection from short-term exposure
				CCC: criterion continuous concentration	Used for setting water quality standards, setting discharge limits, and other regulatory programs; for protection from long-term exposure
A protocol for the derivation of water quality guidelines for the protection of aquatic life	CCME	1999	Canada	Guidelines	Single maximum which is not to be exceeded
Australian and New Zealand guidelines for fresh and marine water quality.	ANZECC & ARMCANZ	2000	Australia/ New Zealand	HRTV: high reliability trigger value	Derived from ≥ 1 multispecies or ≥ 5 single-species chronic data; not a mandatory standard; exceedance triggers further investigation
				MRTV: medium reliability trigger value	Derived from ≥ 5 acute data; not a mandatory standard; exceedance triggers further investigation
				LRTV: low reliability trigger value	Derived from < 5 acute or chronic data; not used as a guideline value
Guidance document on deriving environmental risk limits in The Netherlands	RIVM	2001	The Netherlands	NC: negligible concentration	Used to set environmental quality standards (EQS); EQS may or may not be legally binding
				MPC: maximum permissible concentration	Used to set environmental quality standards (EQS); EQS may or may not be legally binding
				SRC _{ECO} : ecosystem serious risk concentration	Used to set environmental quality standards (EQS); EQS may or may not be legally binding
Water quality guidance for the Great Lakes system	USEPA	2003	United States	Tier I CMC	Adopted into water quality standards or used to implement narrative criteria; for protection from short-term exposure
				Tier I CCC	Adopted into water quality standards or used to implement narrative criteria; for protection from long-term exposure
				Tier II CMC	Used only for implementation of narrative criteria; for protection from short-term exposure
				Tier II CCC	Used only for implementation of narrative criteria; for protection from long-term exposure
Technical guidance document on risk assessment É. Part II. Environmental Risk Assessment.	ECB	2003	European Union	PNEC: predicted no effect concentration	Used in risk assessment

“In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.” (Rio Convention 1992).

Applegate (2000) affirms that, while containing many precautionary elements, US policy does not adhere to the precautionary principle, as many other factors (especially economics) drive US environmental policy. In addition, the USEPA has embraced the use of ecological risk assessment to assess potential chemical hazards. Chapman et al. (1998) note that the precautionary principle undermines the risk assessment approach by either defining infinitely small no-effect concentrations or infinitely large safety factors. Although subscribing to the precautionary principle, EU member countries, Canada, The Netherlands, South Africa, Denmark, and Australia/New Zealand have incorporated risk assessment techniques into their water quality criteria derivation methodologies (Lepper 2002, ECB 2003, CCME 1999, RIVM 2001, Roux et al. 1996, Samsøe-Petersen & Pedersen 1995, ANZECC & ARMCANZ 2000). Thus, although arising from different policy tenets, many of the water quality criteria derivation techniques used throughout the world are applicable under US and California policy.

4.0 Criteria types and uses

Three types of water quality criteria are described by the USEPA: numeric, narrative, and operational (USEPA 1985). This project is concerned with derivation of numeric criteria which the Central Valley Regional Water Quality Control Board can use in setting water quality objectives. This section describes many different types of numeric criteria that may be derived according to various methodologies, depending on how the values are to be used and how much data are available.

Throughout the literature numeric water quality criteria are referred to by many different terms. For example, there are trigger values (TVs; ANZECC & ARMCANZ 2000), guidelines (CCME 1999), criteria (USEPA 1985, Samsøe-Petersen & Pedersen 1995, Roux et al. 1996), quality standards, thresholds (Lepper 2002, Zabel & Cole 1999), environmental risk limits (ERLs; RIVM 2001), maximum tolerable concentrations (MTCs; OECD 1995), predicted no effect concentrations (PNECs; ECB 2003), water quality objectives (WQOs; Bro-Rasmussen et al. 1994) and quality targets (BMU 2001, Irmer 1995). The common thread in all of these is that the values derived are scientifically based numbers which are intended to protect aquatic life from adverse effects of pesticides, without consideration of defined water body uses, societal values, economics, or other non-scientific considerations. This corresponds to what the USEPA calls a numeric criterion and it is the derivation of this type of number that is the subject of this report.

4.1 Numeric criteria vs. advisory concentrations

In the US, numeric criteria are derived for compounds when adequate toxicity, bioaccumulation and/or field data are available (USEPA 1986). These criteria may be used for such things as developing water quality standards or setting effluent limitations (USEPA 1985). If adequate data are not available for criteria derivation then advisory concentrations are derived. Advisory concentrations are used to interpret ambient water quality data. For example, if the ambient concentration of a chemical is below the advisory concentration, then there is no further concern; if the concentration is above the advisory concentration, then more data are collected, preferably enough to allow calculation of a criterion (USEPA 1986).

The USEPA Great Lakes water quality guidance (USEPA 2003a) provides for derivation of Tier I and Tier II criteria. Tier I criteria, which are derived from complete data sets according to the USEPA (1985) methodology, may be adopted as numeric criteria, may be used to adopt water quality standards, or may be used to implement narrative criteria. Tier II criteria, which are similar to USEPA advisory concentrations, are derived from incomplete data sets in a methodology similar to USEPA (1986), and are used only for implementation of narrative criteria.

4.2 Numeric criteria of different types and levels

Many existing criteria derivation methodologies include procedures for derivation of more than one level or type of criterion for each toxicant (OECD 1995, ANZECC & ARMICANZ 2000; La Point et al. 2003; RIVM 2001, Lepper 2002, USEPA 2003a). Derivation of separate acute and chronic criteria, as is done in the USEPA (1985) and UK methodologies (Zabel and Cole 1999), is not what is meant by “different types and levels of criteria.” Rather, this refers to either the derivation of different levels of criteria to meet different regulatory goals, or to the use of ecological risk assessment techniques with increasing levels of technical sophistication, leading to criteria with site-specific application and greater certainty (La Point et al. 2003). The second of these is directly related to how much and what kind of data are available for criteria derivation.

4.2.1 Different criteria levels to meet different goals

Compartment-specific environmental risk limits (ERL) are derived in The Netherlands (RIVM 2001). The three levels of ERLs are the ecosystem serious risk concentration (SRC_{ECO}), the maximum permissible concentration (MPC) and the negligible concentration (NC). The NC (concentration causing negligible effects to ecosystems) is calculated as the MPC divided by a safety factor of 100 and represents a regulatory target value. The MPC is a concentration that should protect all species in ecosystems from adverse effects. If concentrations in ambient waters are above the MPC, discharges can be further regulated. Between the MPC and the NC, principles of ALARA (as low as reasonably achievable) are used to continue reducing levels toward the NC (Warmer & Van Dokkum 2002). The SRC_{ECO} is a concentration at which ecosystem functions will be seriously affected or are threatened to be negatively affected (assumed

to be when 50% of species and/or 50% of microbial and enzymatic processes are possibly affected; RIVM 2001). Waters that exceed the SRC_{ECO} require clean-up intervention efforts.

In the French methodology (Lepper 2002), four threshold levels, corresponding to biological quality suitability classes for water bodies, are calculated for each substance. Threshold level 1, indicating negligible risk for all species, is derived from either chronic or acute toxicity data, with safety factors applied. The level 2 threshold, indicating possible risk of adverse effects for the most sensitive species, is derived from the same data as level 1, but with smaller safety factors applied. Levels 3 and 4, indicating probable or significant risk of adverse ecosystem effects, respectively, are derived solely from acute data. Tentative standards may be set if a minimum data set is not available. Freshwater standards may be used as tentative marine standards if insufficient marine data are available and there is no reason to suspect greater sensitivity in the marine species. None of threshold values derived by the French methodology is enforceable; values serve as references for assessments and actions.

4.2.2 Criteria of increasing site-specificity and certainty

The OECD (1995) methodology recognizes three levels of aquatic effects assessment and derives maximum tolerable concentrations (MTCs) for each level. An initial, or primary, assessment is based on laboratory toxicity data from only one or two representatives of primary producers, primary consumers and predators. An intermediate, or refined, assessment is based on results of chronic or semi-chronic laboratory tests. Field or semi-field studies are used for comprehensive assessments. MTCs derived by the OECD (1995) methodology are used to set environmental quality objectives. However, MTCs have different levels of reliability depending on how they are derived. An MTC derived from quantitative structure activity relationships (QSARs) has lower status than one derived from acute toxicity tests; one derived from acute toxicity tests has lower status than one derived from chronic tests; an MTC derived from a reliable, representative field test has the highest status. Lower status MTCs are used for setting priorities, rather than for setting objectives.

In Australia/New Zealand trigger values (TVs) of low, medium and high reliability are derived (ANZECC & ARMCANZ 2000). The reliability rating is dependent on how much data supports the value. Only medium and high reliability values are used as final guideline TVs. Low reliability values, which are similar to USEPA advisory concentrations, are interim figures, which, if exceeded, indicate the need for further data collection. High and medium reliability TVs are not pass/fail levels. If exceeded, a TV is re-evaluated and refined in a site-specific assessment. Further regulatory action usually occurs only if the site-specific value is exceeded (although risk managers have the option of using the more conservative, national TVs as enforceable values).

By whatever name, all of the numbers discussed (including those not currently used in setting water quality standards or objectives) represent efforts to estimate

concentrations of chemicals that, if exceeded, might lead to loss of designated uses of water bodies. When data are limited, numeric criteria of low site-specificity and high uncertainty can be derived. As more data become available, criteria can be refined for better site-specificity and greater certainty (Di Toro 2003; La Point et al. 2003).

5.0 Protection and confidence

Aquatic life water quality criteria are intended to protect aquatic life from exposure to toxic substances. But what really is the goal? Is it overall ecosystem protection, or protection of each individual in the ecosystem? And how certain can we be of meeting that goal? This section discusses how aquatic life protection goals are stated in various derivation methodologies, and how those goals have to be approached given the need to extrapolate ecosystem effects from single-species toxicity data. There is also a discussion of the importance of being able to state, with a quantified level of certainty, that criteria are achieving the intended level of protection.

5.1 Level of organization to protect

It is necessary to decide what level of organization (defined in Table 3) is to be protected by water quality criteria. Several derivation methodologies seek to protect individuals or species, expecting that by doing so, they will protect ecosystems. Canada's guiding principles for the development of freshwater aquatic life guidelines state that guidelines will consider all components of the ecosystem, and will be "set at such values as to protect all forms of aquatic life and all aspects of the aquatic life cycle" (CCME 1999). Similarly, the UK derives aquatic life EQSs for the protection of all aquatic species. The Netherlands has the goal of protecting all species in ecosystems from adverse effects (RIVM 2001).

Most of the reviewed methodologies specifically seek to protect aquatic ecosystems. Water quality criteria in South Africa are to "allow for the sustainable functioning of healthy and balanced aquatic ecosystems." This is achieved by developing criteria that are protective of representative, key species from a variety of trophic groups (Roux et al. 1996). France derives threshold levels that will maintain a water's suitability to support its biological function and other uses (Lepper 2002). The USEPA criteria are intended to protect "aquatic organisms and their uses" without specifically aiming to protect ecosystems. However, the methodology states that ecosystems can tolerate some stress and it is not necessary to protect all species at all times (USEPA 1985). Arguing that this feature of the USEPA (1985) methodology did not meet the needs of California regulators, Lillebo et al. (1988) developed a criteria derivation methodology specifically for use in California that was designed to ensure full protection of aquatic biological resources. In Australia/New Zealand the goal is "to maintain and enhance the 'ecological integrity' of freshwater and marine ecosystems, including biological diversity, relative abundance and ecological processes" (ANZECC & ARMCANZ 2000). German quality targets are designed to "maintain or restore a self-reproducing and self-regulating biocenosis of plants, animals, and microorganisms that is typical of the location

concerned and is as natural as possible” (Irmer et al. 1995). OECD guidelines provide methods for derivation

Table 3. Definitions of levels of organization.

Level	Definition	Reference
Individual	A single organism	Webster's New Collegiate Dictionary 1976
Species	A taxonomic grouping of morphologically similar individuals that actually or potentially interbreed	Curtis & Barnes 1981
Population	A group of individuals of one species that occupy a given area at the same time	Curtis & Barnes 1981
Community	All of the organisms inhabiting a common environment and interacting with one another	Curtis & Barnes 1981
Ecosystem	All organisms in a community plus the associated abiotic environmental factors with which they interact	Curtis & Barnes 1981
Ecosystem structure	The spatial and temporal relationships of biotic and abiotic components that support energy flow and biogeochemical processes in an ecosystem	Curtis & Barnes 1981
Ecosystem function	The processes by which energy flows and materials are cycled through an ecosystem	Curtis & Barnes 1981
Ecosystem engineer	Organisms that directly or indirectly modulate the availability of resources to other species	Lawton 1994
Keystone species	Species whose removal from a community would precipitate a further reduction in species diversity or produce other significant changes in community structure and dynamics	Daily et al. 1993

of criteria “where no adverse effects on the aquatic ecosystem are expected” (OECD 1995). Denmark derives water quality criteria that are defined as ecotoxicological no-effect-concentrations (Samsoe-Petersen & Pedersen 1995). The PNECs derived by the EU risk assessment methodology (ECB 2003) are intended to ensure “overall environmental protection,” while the Scientific Advisory Committee on Toxicity and Ecotoxicity of Chemicals (CSTE/EEC) states that WQOs should permit all stages in the life of aquatic organisms to be successfully completed, should not produce conditions that cause organisms to avoid habitat where they would normally be present, should not result in bioaccumulation, and should not alter ecosystem function (Bro-Rasmussen et al. 1994; originally in CSTE/EEC 1987). The state of North Carolina seeks to ensure aquatic life propagation and maintenance of biological integrity (North Carolina Department of Environment and Natural Resources 2003). As discussed previously, the mandate of the Central Valley Regional Water Quality Control Board is to maintain waters free of “toxic substances in concentrations that produce detrimental physiological responses in human, plant, animal, or aquatic life” (CVRWQCB 2004).

5.2 Portion of species to protect

In spite of somewhat differing goals, all methodologies are forced to rely a great deal on single-species toxicity data to derive criteria. As pointed out in ECB (2003), two important assumptions are critical to these methodologies, which seek ecosystem protection by extrapolation from single-species laboratory ecotoxicity tests: 1) ecosystem sensitivity depends on the most sensitive species; and 2) protecting ecosystem structure protects community function. This approach is common throughout the world due to the relative availability of data from single-species toxicity tests compared to multispecies or ecosystem data.

A corollary assumption is that ecosystems can sustain some level of damage (to individuals or populations, for example) from toxicants or other stressors and subsequently recover with no lasting harm. This assumption is not completely supported in the literature. As discussed by Spromberg & Birge (2005a, 2005b), whether or not population-level effects occur due to toxicant effects on physiological responses of individuals depends very much on life-history characteristics of the species in question. On one hand, Zabel & Cole (1999) point out that, in the case of algae, if a sensitive species were eliminated from an ecosystem, the photosynthetic function could be quickly replaced by another, less sensitive species. Ecosystem structure will have changed, but function is maintained. On the other hand, Daily et al. (1993) note that the disappearance of a single species could lead to the unraveling of community structure due to complex interactions among species. Lawton (1994) explores the importance of “ecosystem engineers” and states that loss of keystone species, whether they are engineers or important trophic links, may cause dramatic and sudden ecosystem changes.

It would seem, then, that ecosystems might not be protected if water quality criteria are derived by a method that does not have the goal of protecting of 100% of species. However, there is no way to ensure that level of protection because it is not possible to know the entire composition of an ecosystem. Even if that were possible it would not be possible to determine the sensitivity of all the component species. This document presents and evaluates alternative methods for estimating ecosystem no-effect concentrations by extrapolating from available toxicity data, the bulk of which is from single-species laboratory studies. To determine whether numbers derived from these methods are adequately protective (i.e., meet policy goals) requires validation of those numbers through field or semi-field studies.

5.3 Probability of over- or underprotection

To give environmental managers some knowledge of how likely it is that a criterion will provide the intended level of protection, criteria are best expressed with associated confidence limits. Criteria that overprotect lead to unnecessary expenditures, while criteria that underprotect may lead to ecosystem damage. Many criteria methodologies (Canada, France, Germany, UK) involve compilation of data, and then selection of the single most sensitive datum (often multiplied by an extrapolation factor) to represent the criterion (CCME 1999, Lepper 2002, Zabel & Cole 1999). Criteria

derived this way do not have confidence limits associated with them. They may be protective, but there is no way to know to what degree they may over- or underprotect. Criteria derived by the USEPA (1985) methodology also do not have associated confidence limits, in spite of using a species sensitivity distribution (SSD) methodology. Australia/New Zealand (ANZECC & ARMCANZ 2000), The Netherlands (RIVM 2001) and OECD (1995) use SSD techniques that derive criteria at specified confidence levels. For example, for a criterion derived at a 50% confidence level the true no-effect level may be either above or below the derived criterion with equal probability. If derived at a 95% confidence level, there is only a 5% chance that the true no-effect level lies below the derived criterion. This kind of information can provide environmental managers with some sense as to the reliability of criteria. Details regarding how confidence limits are generated are discussed further in Section 7.2.2.3.

6.0 Ecotoxicity and physical-chemical data

At the core of all criteria derivation methodologies lie ecotoxicological effects data. Good criteria must be based on good quality data of adequate taxonomic diversity. Physical-chemical data are important for proper interpretation of toxicity test data, for estimation of bioavailability, and for estimation of toxicity for some classes of chemicals. Thus criteria derivation methodologies must include clear guidance regarding how much of what kinds of data are required for calculation of criteria. A big challenge, which will be discussed further in section 6.3, is finding way to derive criteria from very small data sets. Ideally, it would be possible to derive scientifically sound criteria based on the minimum data sets typically required for pesticide registration procedures. The focus of this section is on what quality and quantity of data are required by existing methodologies.

6.1 Data sources and literature search

Whatever the derivation methodology, the most reliable, most certain criteria are derived from the largest and best quality data sets. It is very helpful for a criteria derivation methodology to include some guidance on where and how to find data. To avoid any perceptions that, say, a regulator has selected only data from very sensitive species, or that a chemical producer has selected only data from very tolerant species, there should also be explicit guidance regarding what constitutes a minimal literature search.

Of the reviewed methodologies, the Dutch methodology provides the most detailed information regarding sources of ecotoxicological and physical-chemical data (RIVM 2001). For plant protection products and biocides, data from registration application packets are used, as well as other relevant data. For other substances data are drawn from public literature. A list of data sources is given, which includes on-line databases (e.g. Current Contents, Biosis, Chemical Abstracts, Toxline), internal databases, handbooks (Mackay et al. 1992, 1993, 1995, 1997, 1999), libraries and even confidential data (note that USEPA 1985, expressly excludes the use of confidential or privileged data). Data used to derive MPCs must be from original sources (as opposed to review

articles, for example). Literature search efforts must be described and should go back to at least 1970. If four or more acceptable chronic studies are available, just a short overview of acute toxicity is acceptable. However, if there are fewer than four chronic data, then all acute data are evaluated. Both freshwater and marine data are collected; if statistical comparison indicates that they are not different, then data are combined.

In the Danish methodology (Samsøe-Petersen & Pedersen 1995) data are collected from handbooks, databases and searches of the open literature. Handbooks include ECETOC (1993), GESAMP (1989), Howard (1990), Howard (1991), KemI (1989), MITI (1992), Nikunen et al. (1990), Roth (1993) and Verschueren (1983). Databases include AQUIRE (1981-present), BIODEG (1992), and LOGKOW (1994). Biodegradability data are estimated using the BIODEG Probability Program (1992) when measured data are not available. Literature searches go back to 1985 and are conducted using BIOSIS. Details of a BIOSIS search profile are given in Annex 2 of Samsøe-Petersen & Pedersen (1995).

For derivation of criteria in Australia/New Zealand (ANZECC & ARMCANZ 2000) data are collected from international criteria documents, the USEPA AQUIRE database, papers from the open literature with acute and chronic toxicity data from field, semi-field and laboratory data, an internal database, and review papers on ecotoxicology. Physical-chemical data are drawn from electronic databases (such as HSDB, available via Toxnet at <http://toxnet.nlm.nih.gov/>) and from Verschueren (1983; most recent version 2001 CD-ROM) and Hansch et al. (1995). Spanish guidelines (Lepper 1999) specify that published data from all kinds of sources may be used to derive criteria. Principal data sources are on-line databases (e.g. AQUIRE, POLTOX, MEDLINE and others) and published water quality objectives.

In the UK, data for EQS derivation is taken from published literature, commercial databases, and unpublished sources (such as manufacturer data; Zabel & Cole 1999). The Canadian (CCME 1999) guidelines indicate what kinds of data should be sought, but do not specify data sources. OECD (1995), German (BMU 2001, Irmer et al. 1995), USEPA (1985), EU (ECB 2003, Bro-Rasmussen 1994), France (Lepper 2002) and South African (Roux et al. 1996) guidelines contain no specifics regarding where to find data or what constitutes an adequate literature search.

Without specific requirements regarding data sources and literature searches, data sets used in criteria derivation could be unnecessarily biased (unnecessary because acceptable data may be overlooked). To ensure inclusion of all relevant data, specific guidance should be given in the derivation methodology.

6.2 Data quality

To minimize uncertainty in water quality criteria, only data that meet stated quality standards should be used in criteria derivation. Toxicity and physical-chemical data should be from studies conducted according to accepted protocols that are appropriate for the chemical and organism being tested. All of the reviewed criteria

derivation methodologies have specific data quality requirements for physical-chemical data as well as for ecotoxicity data. In terms of quality, some, such as France, Germany and Spain, simply state that tests have to have been conducted according to accepted, standardized protocols or according to principles of good laboratory practice (Lepper 2002, BMU 2001, Irmer et al. 1995). Others list very specific data requirements, which are described in the following sections.

6.2.1 Physical-chemical data quality

Only a few of the guidelines state specific data quality parameters for some kinds of physical-chemical data. The Dutch methodology (RIVM 2001) requires that solid-water partition coefficients (K_p) be determined in batch experiments as in Bockting et al. (1993; for organic chemicals). Tests conducted according to OECD guidelines are also acceptable. The Netherlands guidance also points out that water solubility should be determined at an appropriate temperature, usually at 25° C which matches standard laboratory toxicity test temperatures. Since other physical-chemical parameters, such as vapor pressure, Henry's constant (K_H), octanol-water partition coefficient (K_{ow}), and solid-water partition coefficient (K_p) are also temperature-dependent, the temperature at which they were measured should also be noted and values should be adjusted if necessary (Schwarzenbach et al. 1993).

The OECD (1995) guidelines specify that K_{ow} values may be calculated using the ClogP3 algorithm of Hansch and Leo (1979), or may be taken from the THOR/Starlist database. Both the ClogP3 algorithm and the THOR/Starlist database are now accessible through the Bio-Loom program (Biobyt at www.biobtye.com). For highly hydrophobic compounds ($\log K_{ow} > 5$) the OECD (1995) methodology requires that the K_{ow} be determined by either the slow stirring or generator column method. The guidelines recommend expert evaluation of K_{ow} values, as there are many compounds for which reliable values cannot be determined. If measured data are not available, OECD (1995) allows that water solubility may be determined by appropriate QSARs that relate K_{ow} to solubility.

The USEPA (1985) has specific criteria for acceptance of bioconcentration factors (BCF). To be used in determination of final residue values (FRVs), BCFs must be from flow-through tests, must be based on measured concentrations of test substance in both tissue and test solution, and must be from tests that were long enough for the system to reach steady-state. For lipophilic materials, the percent lipid in the tissue must be reported. If a BCF was determined in an exposure that caused adverse effects in the test organism, it should not be used. If reported on a dry weight basis, BCF values must be converted to a wet weight basis. Finally, if more than one acceptable BCF is available, the geometric mean of available values is used, provided they are from exposures of the same length.

Any physical-chemical data used in derivation of water quality criteria should be evaluated to ensure that they were determined by appropriate methods. Generally, data from current, standard methods (e.g., ASTM, OECD) applied and performed correctly for

the chemical of interest, will be acceptable. Non-standard methods may also be appropriate, but only if valid reasons are given for deviation from standard methods. In regards to pesticides, which vary widely in characteristics such as hydrophobicity, water solubility, and ionizability, it is particularly important to verify that reported partition coefficients were determined correctly.

6.2.2 Ecotoxicity data quality

The EU Technical Guidance Document on Risk Assessment (TGD; ECB 2003) defines data quality in terms of reliability and relevance. Reliability is the inherent quality of a test relating to test methodology and the way that the performance and results of the test are described. Relevance refers to the extent to which a test is appropriate for a particular hazard or risk assessment. Reliable data are from studies for which test reports describe the test in detail and indicate that tests were conducted according to generally accepted standards. Relevance is judged by whether a study included appropriate endpoints, was conducted under relevant conditions, and if the substance tested was representative of the substance being assessed. EU criteria derivation guidance, as described by Bro-Rasmussen (1994), is very general with regard to data quality and primarily requires that data include details of tests used.

The UK, The Netherlands, Canada and Australia/New Zealand evaluate data and assign ratings depending on its reliability and/or relevance. In the UK, primary data are those classified as reliable and relevant and secondary data are those for which inadequate details are available. The evaluation is based purely on expert assessment of experimental procedures, test species, endpoints, and whether or not a dose-response relationship has been established. Primary data are used in derivation of EQSs; secondary data are used only as supporting information (Zabel & Cole 1999).

The Dutch methodology uses a reliability index (RI) to evaluate data (RIVM 2001). Reliable data (RI = 1) are from studies conducted and reported in accordance with internationally accepted test guidelines or Mensink et al. (1995). Less reliable data (RI = 2) are those from studies in less accord with accepted guidance or Mensink et al. (1995), and data deemed not reliable (RI = 3) are from studies not at all in accord with accepted guidance or Mensink et al. (1995). Data rated 1 or 2 are used in derivation of ERLs; data rated 3 are included in the final report, but are not used in criteria derivation.

Part of data quality is ensuring that it comes from properly conducted, well-documented studies. In The Netherlands (RIVM 2001), data must come from referenced studies that must include specific organism identification, information regarding purity of the test substance, details of the test, and clearly stated results. For systematic evaluation, data are subdivided by type (freshwater, marine, acute, chronic) and put into data tables. Table headings include: species (including scientific name), species properties (e.g. age, weight, lifestage), analysis of test compound (measured or not, Y or N), test type (flow-through, static-renewal, static), substance purity, test water, pH, water properties (e.g. hardness, salinity), exposure time, test criterion (e.g. LC₅₀ or NOEC, where LC₅₀ is the

concentration that is lethal to 50% of organisms, and NOEC is the no observed effect concentration), ecotoxicological endpoint (growth, reproduction, mortality, immobilization, morphological effects, histopathological effects), LC₅₀ values, NOEC values, notes, and reference information.

For data to be usable in criteria derivation in The Netherlands, specific toxicity test acceptability requirements must be met (RIVM 2001). These include: the purity of the test substance must be at least 80%, studies may not use animals collected from polluted sites, concentration of test substance may not exceed 10x the water solubility, no more than 1 ml/L of carrier solvent can have been used, and recovery of the substance needs to be 80% or more. Also, for compounds with short half-lives, the renewal frequency in a static-renewal test becomes important. In the Dutch methodology, if the $t_{1/2}$ is shorter than the renewal interval, the data are not used.

The Australia/New Zealand guidelines follow the standard operating procedures for the AQUIRE system (AQUIRE 1994) to rate toxicity studies according to how well they are documented (ANZECC & ARMCANZ 2000). In this rating system, weighted scores are applied to eighteen characteristics relating to test methodology. The two most heavily weighted characteristics are exposure duration and end-point; if those are not both recorded the study will not receive a strong rating. Other characteristics, which each receive very little weight, include control type, organism characteristics, chemical analysis method, exposure type, test location, chemical grade, test media, hardness/salinity, alkalinity, dissolved oxygen, temperature, pH, trend of effect, effect percent, statistical significance and significance level. Based on scores in these categories, data are rated as either C (complete), M (moderate), or I (incomplete). Only data rated C or M are used to derive guideline values. In addition, the Australia/New Zealand guidelines allow for use of data that has already been accepted and used in Dutch and Danish water quality documents. Clear direction on how to deal with outlying data is also given in the Australia/New Zealand guidelines, although the curve-fitting technique used in this methodology (described in section 7.2.2.1) minimizes the need to remove outliers. The Danish methodology also assesses data quality according to the AQUIRE system (Samsoe-Petersen & Pedersen 1995).

In addition to the Australia/New Zealand data quality guidelines already discussed, the ANZECC & ARCANZ (2000) guidelines provide specific toxicity test validity criteria. These include: test solutions should cover a geometrically-increasing series; a control and solvent control should be included; control mortalities should be less than 10% (or some other level, determined by the specific test method); other adverse effects in controls should be less than 20%; water quality parameters should be measured and should be within specified limits; a least significant difference for hypothesis tests should be calculated and reported; test organisms should be allowed sufficient time for acclimation to test water; loading of animals in test containers should be appropriate; measured test concentrations should not vary greatly from nominal concentrations; animals should be randomly assigned to test vessels and test vessels should be randomly placed in test chamber or room; any requirements for things such as timing of hatch, or timing and number of young produced should be met; source and health of test organisms

and stock cultures should be traceable; feeding and no-feeding requirements must be met; reference toxicant test results for test organisms should be available.

By the Canadian methodology (CCME 1999) each study is evaluated to ensure acceptable laboratory practices were used. Studies are classified as either primary, secondary, or unacceptable, with only primary and secondary data being used to derive guideline values. Primary data must be from toxicity tests conducted according to currently acceptable laboratory practices, but more novel approaches may be acceptable on a case-by-case basis. Also, for primary data, test concentrations must be measured at the beginning and end of the exposure period, and static tests are unacceptable unless test concentrations and environmental conditions were maintained throughout the test. Studies should have endpoints from partial or full life-cycle tests and should include determination of effects on embryonic development, hatching, germination, survival, growth and reproduction. Appropriate controls must be included and measurements of abiotic variables (e.g. temperature, pH, etc.) should be reported. Secondary data may come from tests conducted from a wider range of methodologies, and may include static tests, and test with endpoints such as pathological, behavioral or physiological effects. Nominal test concentrations are acceptable for secondary data, and, as for primary data, relevant abiotic variables and control responses should be reported.

Data used in derivation of criteria by the USEPA (1985) must be available in a publication or must be in the form of a typed, dated, and signed document (manuscript, memo, letter, etc.). Reports must include enough supporting information to indicate acceptable test procedures and reliable results. The USEPA also provides very specific data quality guidance (USEPA 1985). Tests are to be rejected if there was not a control treatment, if too many control organisms died or were stressed or diseased, or if improper dilution water was used. Tests using formulated mixtures or emulsifiable concentrates are not acceptable, but tests with technical grade materials are acceptable. For highly volatile or degradable materials, or for measuring chronic toxicity, tests should be flow-through with frequent measurements of test solution concentrations. Chronic test exposures should be life-cycle, partial life-cycle or early life-stage. Data are rejected if they are from tests with brine shrimp, species that do not have reproducing populations in North America, or organisms previously exposed to contaminants.

As in the USEPA methodology, South Africa (Roux et al. 1996) rejects data from tests in which there was no control treatment, too many control organisms died (> 10%), improper dilution water was used, organisms were previously exposed to contaminants, or where there was insufficient agreement of toxicity data within and between species. Data from tests with formulated mixtures are also rejected.

The OECD (1995) guidelines prefer toxicity data from tests conducted according to standardized methods. The guidelines also specify that it is important to consider water solubility, K_{ow} and bioaccumulation potential of a substance in assessing acceptability of acute toxicity data. If the solubility is below the LC_{50} , or if the test duration was too short given the K_{ow} and/or BCF (generally for $\log K_{ow} > 5$) then acute tests are not acceptable and only chronic data may be used. Further toxicity test acceptability requirements are

not given in OECD (1995), but OECD test guidelines include specific validity standards (e.g. OECD 1992).

Detailed data quality requirements must be part of a criteria derivation methodology. Specifics must ensure quality, but should not be so stringent that excessive data are rejected. The Netherlands, USEPA, Canada, Australia/New Zealand and OECD provide good guidance, and elements of these should be considered for inclusion in the new methodology.

6.3 Data quantity—ecotoxicity

Small ecotoxicity data sets are a common problem faced by regulators wanting to develop water quality criteria. Large data sets, representing numerous taxa in acute and chronic exposures, exist for very few chemicals. Basic data sets required for pesticide registration, typically containing only acute data for a few species, are available for many chemicals, but for some new chemicals, no ecotoxicity data are available. This section explores what kinds of water quality criteria can be derived by various methodologies from data sets of all sizes.

The quantity of ecotoxicological effects data required for criteria derivation varies quite a lot around the world and depends on what derivation methodology is used, what type of criterion is being developed (i.e., values to be used in standard setting vs. advisory values), and what level of uncertainty is acceptable in the criterion. Criteria are derived by extrapolating from effects seen in whatever data are available to real-world situations. The two basic methods for doing these extrapolations are application of assessment factors (AFs; discussed in section 7.2.1) and statistical extrapolation of species sensitivity distributions (SSDs; discussed in section 7.2.2). There is not much debate about appropriate levels of data for the AF method. Factors are applied according to how much of what kinds of data are available, and many methodologies allow for derivation of a numerical guideline value (as opposed to an enforceable criterion) for a contaminant based on as little as one datum that may be an estimated toxicity value rather than a measured one. On the other hand, for statistical extrapolation methods, there is little agreement among current methodologies, proposed methodologies or in the literature regarding how much data is needed to produce criteria with a reasonable level of uncertainty.

According to the Australian/New Zealand methodology (ANZECC & ARMICANZ 2000) high reliability TVs can be determined either directly from at least three multispecies chronic NOEC values, or from statistical extrapolation using at least five single-species chronic NOEC values (from five different species). A moderate reliability TV can be derived from at least five single-species acute toxicity values, and low reliability TV can be derived from a single acute or chronic toxicity datum.

The Dutch methodology (RIVM 2001) requires at least four chronic NOEC values of species of different taxa for a refined effects assessment, but for a preliminary effect assessment an ERL may be derived from a single LC₅₀ or QSAR estimate (see

section 6.4.2 regarding QSARs). Toxicity data estimated by QSARs may also be used in statistical extrapolation models.

The OECD (1995) guidelines present several methods for criteria derivation and each has its own data requirements. For statistical extrapolations by the methods of Aldenberg & Slob (1993) or Wagner and Løkke (1991), at least five chronic NOECs are required. To derive a final chronic value (FCV) by the USEPA methodology (USEPA 1985) requires chronic NOEC values for at least eight animal families including Salmonidae, a second family in the class Osteichthyes, a family in the phylum Chordata, a planktonic crustacean, a benthic crustacean, and insect, a family in a phylum other than Arthropoda or Chordata, and a family in any order of insects or any phylum not yet represented. Unlike the USEPA (1985) method, the OECD does not allow for derivation of a chronic criterion by application of an acute-to-chronic ratio (ACR) to a final acute value (FAV). For OECD (1995) assessment factor methods an environmental concern level (ECL) can be determined from a single LC₅₀ value. If no toxicity data are available, QSARs may be used for some classes of chemicals to estimate toxicity and values thus derived may be used in derivation of maximum tolerable concentrations (QSARs are discussed in detail in section 6.4.2).

For derivation of a FAV the USEPA (1985) requires acute toxicity data for species resident in North America from at least eight different families, as described above for the OECD methodology. The California Department of Fish and Game has derived criteria for carbaryl and methomyl using the USEPA (1985) SSD method when fewer than the eight required families are represented in the data set, using professional judgment to determine that species in the missing categories were relatively insensitive and their addition would not lower the criteria (Siepmann & Jones 1998; Monconi & Beckman 1996). A FCV may be calculated (acc. to USEPA 1985) in the same manner as the FAV if chronic data are available for at least 8 different families. Alternatively, a FCV may be derived by application of an ACR to a FAV if ACRs are available for aquatic species in at least three families provided that, of the three species, at least one is a fish, at least one is an invertebrate, and at least one is an acutely sensitive freshwater species. The USEPA methodology also requires data from at least one toxicity test with an alga or vascular plant and at least one acceptable bioconcentration factor (BCF). The South African methodology (Roux et al. 1996) has essentially the same data quantity requirements as the USEPA, with the exception that the data must be from species that are either indigenous to southern Africa, or are of local commercial or recreational importance.

The state of North Carolina follows USEPA (1985) FAV derivation procedures to determine acceptable acute toxicity levels, but also provides a means for derivation of an acceptable level of acute or chronic toxicity based on the lowest available LC₅₀ value (implying that a single value may be used; North Carolina Department of Environment and Natural Resources 2003). The Water Quality Guidance for the Great Lakes (USEPA 2003a) allows for derivation of Tier II criteria based on applying an assessment factor to the lowest genus mean acute value (GMAV) in the database. Although not explicitly stated, it appears that a Tier II criterion could be based on a single datum by this method.

The Canadian methodology (CCME 1999) requires at least three studies on at least three fish species resident in North America, including at least one cold-water species and one warm-water species. At least two of the fish studies must be chronic studies. The Canadian guidelines also require at least two chronic studies on at least two invertebrate species from different classes, at least one of which has to be a planktonic species resident in North America. At least one study of a freshwater vascular plant or algal species resident in North America is also required, unless a chemical is known to be highly phytotoxic, in which case at least four acute and/or chronic studies of nontarget plants or algae are required.

For effects assessment according to the EU TGD on risk assessment (ECB 2003) an assessment factor method can be used to derive a predicted no effect concentration (PNEC) from either one LC/EC₅₀ from each of three trophic levels (fish, crustacean, alga), or from one or more chronic NOECs. For statistical extrapolation by the species sensitivity distribution method (SSD; details in section 7.2.2), the TGD requires at least 10 chronic NOECs from eight taxonomic groups including two families of fish, a crustacean, an insect, a family in a phylum other than Arthropoda or Chordata, a family in any order of insect or any phylum not already represented, an alga and a higher plant.

In France data from three trophic levels (algae/plants, invertebrates, fish) are required for derivation of threshold values. If data from only two trophic levels are available, provisional thresholds are derived. If there are no data from particularly sensitive species, or if there are data for fewer than two trophic levels, then no criteria are derived (Lepper 2002).

German methodology requires chronic toxicity data from four trophic levels (bacteria/reducers, green algae/primary producers, small crustaceans/primary consumers, fish/secondary consumers) to derive criteria. If chronic NOECs are available for at least two trophic levels, acute data may be used to fill trophic level gaps, but must be multiplied by an acute-to-chronic extrapolation factor (0.1) and the result is a tentative criterion. If chronic data from at least two trophic levels are not available, no criterion can be derived (Lepper 2002, BMU 2001, Irmer et al. 1995).

In Spain, aquatic life criteria are derived from acute or chronic data for at least three species, which must include algae, invertebrates, and fish (Lepper 2002), while the UK requires acute or chronic data for algae or macrophytes, arthropods, non-arthropod invertebrates and fish (Zabel & Cole 1999). Neither of these methodologies indicates precisely how much data of each kind is required.

By several current derivation methodologies, water quality guideline values can be derived by the application of assessment factors even if there is no measured toxicity data (based on QSARs). For development of full, enforceable criteria that can be used directly in setting water quality standards, a large, diverse ecotoxicity database is required. Canadian guidelines (CCME 1999) require at least 6 data; others do not specify a number, but leave much to professional judgment. In all cases, as the number and

diversity of data increase, assessment factors decrease, thus reducing the uncertainty-driven conservatism in criteria values.

For statistical extrapolations by parametric techniques, data requirements range from $n = 4-10$. In discussing the use of statistical extrapolations for very small data sets Aldenberg & Luttik (2002) note that samples sizes as small as $n = 2$ can be used, although the values derived from samples as small as $n = 2-3$ are not of much practical use due to their very high level of uncertainty. In an analysis of the influence of data quantity and quality, and model choice on results of SSDs, Wheeler et al. (2002) found that a minimum of $n = 10$ was required to obtain a reliable estimate of a particular endpoint (e.g., an HC_5 ; hazardous concentration potentially harmful to 5% of species). Okkerman et al. (1991) conclude that, while seven data would be ideal, five data are adequate for the SSD procedure described by Van Straalen & Denneman (1989). According to Aldenberg & Slob (1993) the risk of under-protection of a 50% confidence limit estimate of the HC_5 (based on a log-logistic distribution) decreases considerably as sample size is increased from 2 to 5, but less so as it is increased from 5-10 and from 10-20.

Jagoë & Newman (1997) proposed using bootstrapping techniques with SSDs to avoid the issue of fitting available data to a particular distribution (discussed in section 7.1.2.1). Later, Newman et al. (2000) found that the minimum sample sizes required for a bootstrapping method ranged from 15 to 55. In a similar analysis, Newman et al. (2002) found that 40-60 samples were required to derive an HC_5 with an acceptable level of precision. Van Der Hoeven (2001) described a non-parametric SSD method that requires a minimum of 19 samples with as many as 59 required to derive an one-sided 95% confidence limit HC_5 estimate. Considering the general lack of ecotoxicity data it is understandable that none of the current criteria derivation methodologies utilizes a bootstrapping approach for SSD extrapolations. Grist et al. (2002) argue that a drawback of the bootstrap technique is that there is no legitimate way to determine a minimum sample size.

Based on this discussion, a sample size of 5 is the minimum needed for parametric statistical extrapolation procedures. For smaller data sets, only assessment factor derivation methods are appropriate. Minimal data sets available for derivation of criteria in California will be those required for registration under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) and those required by the California Department of Pesticide Regulation (DPR). According to 40 CFR Part 158.490 (1993), the minimum data required by FIFRA is an LC_{50} for a fish and an LC_{50} for a freshwater invertebrate. All other kinds of aquatic toxicity data are only conditionally required depending on planned pesticide usage, potential for transport to water, whether any acute LC/EC_{50} values were < 1 mg/L, whether estimated environmental concentrations are > 0.01 times any LC/EC_{50} , or if data indicate reproductive toxicity, persistence, or bioaccumulative potential. It is possible that for many new chemicals, only the two acute toxicity data will be available. The DPR has tiered data requirements (California DPR 2005a). The minimum data set includes LC_{50} s for one warm water and one cold water fish and for a freshwater invertebrate. Further testing is required for the same reasons discussed for FIFRA. Again, it is possible that no more than the minimum data will be

available for criteria derivation for new pesticides. An assessment factor criteria derivation method will be needed for these very small data sets.

6.4 Kinds of data

6.4.1 Physical-chemical data

Physical-chemical data are not used directly in the derivation of water quality criteria. However, they are valuable for assessment of toxicity test data (for example comparing test concentrations to solubility), for translation of criteria based on total concentration in water to dissolved concentration in water or to concentration in suspended matter, for assessment of factors that might affect toxicity (such as the effect of pH on the relative concentrations of ionized and unionized forms of chemicals), for estimation of physical-chemical parameters for which no measured values are available, for prediction of bioaccumulation or secondary poisoning potential (section 7.3.2), and especially for estimation of toxicity where data are lacking. For the purposes of this discussion, bioconcentration factors (BCF) and bioaccumulation factors (BAF) are included in the group of physical-chemical parameters, although it is recognized that they could as well be discussed as toxicological data.

The Netherlands methodology (RIVM 2001) requires collection of specific physical-chemical data. For each substance the following information is required: IUPAC name, CAS number, EINECS number (European Inventory of Existing Commercial Substances), structural formula (including diagram), empirical formula, molar mass, octanol-water partition coefficient (K_{ow}), water solubility, melting point, vapor pressure, Henry's law constant (K_H), acid dissociation constant(s) (pK_a), solid-water partition coefficients (K_p) and degradation information (i.e., hydrolysis, photolysis, biodegradation). The methodology includes procedures for calculation of a dimensionless K_H if measured constants are not available.

Physical-chemical data and environmental fate information are used in the Dutch methodology (RIVM1999) in a number of ways. For example, if a substance has a $t_{1/2}$ of less than 4 h, then the criterion is derived for stable degradation products, rather than for the parent compound. Also, if data are lacking for a particular environmental compartment, partitioning data can be used to estimate concentrations given a measured concentration in another compartment. Water solubility data are used to judge the reliability of aquatic toxicity studies (section 6.2.2), but may also be used together with vapor pressure and molecular weight data to calculate a Henry's Law constant. Suspended matter-water partition coefficients are used to calculate total toxicant concentrations in water based on the dissolved concentration (section 7.1.3), and octanol-water partition coefficients are used to estimate aquatic toxicity using QSARs, for estimation of BCF values, to determine potential risk of secondary poisoning, and for estimation of organic carbon-water partition coefficients. Finally, partitioning constants are used for harmonization procedures (Section 7.3.4).

The OECD (1995) methodology recommends that the following information be obtained for each compound: chemical structure, molecular weight, melting point, water solubility, K_{ow} , sediment-water partition coefficient (K_{sw}), and pK_a . Octanol-water partition coefficients may be used to estimate water solubility, or to derive QSAR estimates of toxicity. Van Leeuwen et al. (1992) showed that by using QSAR estimates, it is possible to develop a relationship between K_{ow} and the hazardous concentration for inert chemicals, and thus it is possible to derive MTCs and their associated confidence limits directly from K_{ow} values.

In the Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000), K_{ow} and BCF values are used to estimate bioaccumulative potential. The BCF also may be used in calculating water concentrations that will be protective of fish-eating predators from bioaccumulative chemicals. For derivation of low reliability target values for narcotic chemicals (when little to no toxicity data are available) the Australia/New Zealand guidelines utilize K_{ow} values to derive QSAR estimates of toxicity. Beyond K_{ow} and BCF values, the Australia/New Zealand guidelines provide no specific requirements for collection and reporting of physical-chemical data.

As discussed in section 7.1.3, the German derivation methodology utilizes the suspended particulate matter-water partition coefficient to express quality targets in terms of toxicant concentration in suspended particulate matter for compounds with partition coefficients greater than 1000 l/kg. Also, for protection of fisheries in Germany, BCF values are used to derive water quality targets based on maximum permissible pesticide residue values for fish.

The USEPA (1985) guidelines only explicitly require collection of bioaccumulation data, and then only if data are available indicating that residues are of toxicological concern. Other physical-chemical data, such as volatility, solubility and degradability, are required for evaluation of toxicity data. Bioaccumulation data (BCFs and BAFs) are used to derive the final residue value (FRV).

For development of a full guideline, Canada (CCME 1999) requires collection of environmental fate data. Specifically, information must be available on the mobility of the substance and where it is most likely to end up, on abiotic and biotic transformations that occur during transport and after deposition, on the final chemical form of the substance, and on the persistence of the substance in water, sediment and biota.

The Danish methodology (Samsøe-Petersen & Pedersen 1995) does not clearly specify what kinds of physical-chemical data must be collected, but criteria derivation documents indicate consideration of a wide range of data including CAS number, empirical formula, molecular weight, water solubility, K_H , BCF, and K_{ow} , as well as biodegradability data. Bioaccumulation data are used in deciding on the size of the assessment factor to be applied (see section 7.2.1 for discussion of assessment factors). Biodegradation data are used to determine whether criteria ought to be derived for the parent chemical or for a stable, toxic metabolite. If little is known about degradation products of a substance, then assessment factors will reflect this uncertainty.

According to EU guidance (Bro-Rasmussen et al. 1994) physical-chemical data requirements are very general, stating simply that “a summary of the main chemical and physico-chemical characteristics” must be collected. For criteria derivation, bioaccumulative potential and persistence can affect the size of the applied assessment factor. Also, K_{ow} values may be used to derive QSAR estimates of toxicity when toxicity data are lacking. For assessment of secondary poisoning potential, the EU risk assessment TGD (ECB 2003), utilizes K_{ow} values, adsorption data, hydrolysis and other degradation data, and molecular weight.

Spanish guidelines (Lepper 2002) require collection of physical-chemical data that may have some bearing on the toxicity of the substance. These include speciation, toxicokinetic properties, and relationships between toxicity and water quality parameters. The UK (Zabel & Cole 1999) and South African (Roux et al. 1996) guidelines indicate no specific uses for physical-chemical data in criteria derivation.

Physical-chemical data are used by various methodologies to improve interpretation of ecotoxicity data and to determine whether water quality criteria are set at levels that could potentially harm non-aquatic species (including humans). Without a good set of physical-chemical data it would not be possible to adequately assess potential effects of chemicals. Explicit details regarding the collection of physical-chemical data are an important part of a criteria derivation methodology.

6.4.2 Quantitative Structure Activity Relationships (QSARs)

QSARs are mathematical relationships between a chemical's structure and its toxicity. According to Jaworska et al. (2003) QSARs are simplified mathematic representations of complex chemical-biological interactions. They are most commonly developed by regression analysis, neural nets or classification methods (Jaworska et al. 2003). QSARs are used by several existing criteria derivation methodologies to fill in data gaps. That is, if little to no toxicity data are available for criteria derivation, toxicity for some kinds of compounds for some species can be estimated using QSARs.

The most commonly used chemical structural feature used in QSARs is the K_{ow} . QSARs are developed for classes of chemicals, such as inert, less inert, reactive and specifically acting chemicals (Verhaar et al. 1992). These classes were later described by Vaal et al. (1997b) as non-polar narcotics, polar narcotics, reactive compounds and specifically acting compounds. For fathead minnows Russom et al. (1997) further separated the specifically acting compounds into oxidative phosphorylation uncouplers, acetylcholinesterase inhibitors, respiratory inhibitors, electrophiles/proelectrophiles and central nervous system seizure agents. QSARs with good predictive power can be developed for narcotic chemicals from K_{ow} data alone, but for chemicals with a specific mode of toxic action, more physical-chemical data are needed, such as reactivity or pK_a , and the predictive models become more complex (Auer et al. 1990). Ramos et al. (1998) suggest that models based on real phospholipid membrane/water partitioning, rather than K_{ow} s, would more accurately predict toxicity of polar and non-polar narcotics. The recent

“Workshop on Regulatory Use of (Q)SARs for Human Health and Environmental Endpoints,” (summarized in Jaworska et al. 2003) produced a series of papers that provide guidance on assessing reliability, uncertainty and applicability of QSARs (Eriksson et al. 2003) and a review of the use of QSARs in international decision-making frameworks for prediction of ecologic effects and environmental fate of chemicals (Cronin et al. 2003).

When insufficient data are available, several water quality criteria derivation methodologies allow for the use of QSARs to estimate aquatic toxicity (discussed below). When assessing hazards of chemicals for which little to no ecotoxicity data are available, the USEPA Office of Pollution Prevention and Toxics (OPPT) uses QSARs under the Toxic Substances Control Act (TSCA) to estimate toxicity (Nabholz 1991). Toxicity values calculated from QSARs are used in statistical extrapolation or assessment factor methods to derive criteria. Neither the USEPA national nor the newer Great Lakes criteria derivation methodologies use QSARs in criteria derivation (USEPA 1985, 2003a).

Although recognizing that QSARs exist for many modes of toxic action, the Dutch guidelines allow the use of QSARs only for substances that have a non-specific mode of action (i.e., those acting by narcosis; RIVM 2001). The guidelines provide 19 QSARs for aquatic species representing 9 different taxa. NOECs estimated from QSARs may be used as input into extrapolation models for derivation of ERLs. In the UK (Zabel & Cole 1999) QSARs or other models may be used to predict toxicity in absence of other data, but such data are not used to derive EQSs (used only for support).

The OECD (1995) guidelines offer two QSAR approaches. First, is that of the USEPA Office of Pollution Prevention and Toxics (OPPT), which is based on the classification of chemicals by their structure without consideration of mode of toxic action. Specifics of this approach are described by Nabholz (2003). Second, is a method that classifies chemicals first by mode of action and then by chemical structure. The second approach is similar to that used in the Dutch methodology (RIVM 2001), except that the OECD provides QSARs for four classes of toxic mode of action as defined by Verhaar et al. (1992; inert/baseline, less inert, reactive and specifically acting chemicals). By the OECD (1995) methodology, if no toxicity data are available, QSARs may be used to derive MTCs. For inert chemicals, QSARs may be used to estimate toxicity for fish, *Daphnia*, and algae. For less inert chemicals, estimates may be made for fish. Due to lack of thorough evaluation, QSARs for reactive and specifically acting chemicals are not used to derive OECD MTCs (OECD 1995). If some toxicity data are available, then QSAR estimates of toxicity for inert chemicals are compared to measured values. If the values agree within a factor of 5, then the QSAR values may be used to extend the database for MTC derivation. MTCs derived solely from QSAR data are used only for priority setting purposes; they are not used to set environmental quality standards.

When toxicity data are lacking, the use of QSARs offers a way to estimate toxicity and fill data gaps for polar and non-polar narcotic chemicals. However, existing criteria derivation methodologies do not endorse the use of QSARs to estimate toxicity

for chemicals with specific modes of action. Pesticides of greatest concern in the Sacramento and San Joaquin Rivers include specifically-acting organophosphates and pyrethroids. Given the current state of the science, QSARs will not be useful in predicting toxicity for these kinds of chemicals.

6.4.3 Ecotoxicity data

Many kinds of ecotoxicity data exist in the literature. Results of short-term acute tests are available, as are results of long-term chronic tests, and sensitive life-stage tests, which may be used as predictors of chronic toxicity (USEPA 2002b). Some studies assess lethality, while others assess sub-lethal endpoints, including inhibition of growth or reproduction. Still others look at effects of toxicants on biochemical endpoints, such as inhibition of acetylcholinesterase or up-regulation of glutathione S-transferases. Some tests are performed on one species, while others utilize microcosms or mesocosms to study several species at the same time. Some tests are conducted under highly controlled laboratory conditions and some are conducted in field or semi-field settings. Each of these kinds of studies generates a value, or series of values, in the form of a lethal concentration that kills 50% of exposed organisms (LC_{50}), an effect concentration that adversely affects some portion (x) of exposed organisms (EC_x) or a no observed effect concentration (NOEC), which is the highest concentration of toxicant that causes a response that is not different from the control treatment. Results may also be reported as a lowest observed effect concentration (LOEC), the lowest concentration of toxicant that causes a response that *is* different from the control, or a maximum allowable toxicant concentration (MATC), which is the geometric mean of the NOEC and LOEC (USEPA 1987). This section is a discussion of the many different approaches to definition and usage of different kinds of data among existing criteria derivation methodologies. Further details about exactly which data are used in criteria calculation are included in the appropriate subsections of Section 7.0.

6.4.3.1 Acute vs. chronic

Water quality criteria need to be protective of aquatic life under conditions of long-term, continuous exposure, as well as under conditions of short-term, transient exposure. Long-term exposures are generally considered chronic exposures, while short-term exposures are considered acute. However, an acute exposure for an organism with a relatively long life-span would represent a chronic exposure for an organism with a relatively short life-span. Toxicity test data are often defined as either acute or chronic, but what those terms mean with respect to exposure duration varies with species. Thus it is important to have clear guidance regarding what kind of toxicity test data should be considered to represent acute versus chronic exposures, and what kind of criteria may be derived with acute versus chronic toxicity data.

The Netherlands guidelines give the very general definition that an acute exposure represents a relatively short period, while a chronic exposure represents enough time for a complete or partial life-cycle. Whether an exposure is acute or chronic depends on the physiology and life-cycle characteristics of the species (RIVM 2001). To make the

distinction clear, then the Dutch guidelines give more detailed definitions. Acute tests generally last less than 4 d and the results are reported as an LC₅₀ or EC₅₀. Chronic tests generally last more than 4 d and results are reported as a NOEC. However, for single-celled organisms (e.g. algae or bacteria), chronic NOECs may be obtained in less than 4 d. And to be very specific, the following guidance is offered: for algae, bacteria or protozoa, tests of 3-4 days are defined as chronic; for Crustacea and Insecta, tests of 48 or 96 h are acute; and for Pisces, Mollusca, and Amphibia tests of 96 h are acute, while early life-stage tests and 28-d growth tests are chronic (RIVM 2001). Only chronic NOECs are used for refined effect assessments; acute data are used with application of assessment factors in preliminary effect assessments.

Chronic toxicity data are preferred by OECD (1995) guidelines with either NOECs or MATCs being acceptable. However, acute data are also used, but with appropriate application of assessment factors (i.e., ACRs). The guidelines caution that substances with low water solubility or $\log K_{ow} > 5$, a 96-h acute exposure in water may not be long enough to see effects and it may only be possible to use chronic data for such substances. While not explicitly stated, then, it appears that this methodology considers exposures longer than 96 h to be chronic. By this methodology, NOECs may be estimated by conversion of LOECs (e.g., $\text{NOEC} = \text{LOEC}/2$), but only if the LOEC corresponds to a concentration causing > 20% effect.

The Australia/New Zealand guidelines (ANZECC & ARMCANZ 200) contain the very general description that acute tests are shorter than chronic tests, but go on to indicate that in actually applying the methodology, data from tests longer than 96 h were considered to be chronic, except for tests with single-celled organisms (for which 96-h tests are considered chronic). Chronic data are used to derive high reliability target values, while acute data are used to derive moderate reliability target values. NOEC and LC₅₀ data are both used in statistical extrapolations, but the resulting hazardous concentration determined with LC₅₀s is multiplied by an ACR.

The USEPA (1985) methodology utilizes acute LC₅₀ or EC₅₀ data to derive the Final Acute Value (FAV). The EC₅₀ data in this case are based on the percentage killed plus the percentage immobilized. EC₅₀ data relating to less severe effects are not used in calculation of the FAV. Acute toxicity data are described as those from 48-h tests with daphnids and other cladocerans, from 96-h tests with embryos and larvae of various shellfish species, or from 96-h tests with older life stages of shellfish species. Tests with single-celled organisms of any duration are not considered acute tests. The Great Lakes guidance (USEPA 2003a) expands on that a bit and states that any test that takes into account the number of young produced (e.g., protozoan tests) are not considered acute even if the duration is less than 96 h. Chronic tests in the USEPA guidance (USEPA 1985) are described as life-cycle tests (ranging from just over 7 d for mysids to 15 months for salmonids), partial life-cycle tests (all major life stages exposed in less than 15 months; specifically for fish that require more than a year to reach sexual maturity), or early life-stage test (ranging from 28 to 60 d; also specifically for fish). Chronic data (expressed as an MATC or a value determined by regression) are used to derive a Final Chronic Value (FCV), but the FCV may also be calculated by applying an ACR to the

FAV. The South African methodology (Roux et al. 1996) generally follows the USEPA (1985) methodology in terms of the use of LC/EC₅₀ and MATC data, but does not contain explicit descriptions of acute vs. chronic data.

The German guidelines (Irmer et al. 1995) use NOECs from studies of long-term toxicant exposure. If no chronic data are available, acute data may be multiplied by a factor and used instead. No guidance is given on how to classify tests as either acute or chronic.

For derivation of full guidelines in Canada (CCME 1999), at least two of three fish toxicity data must be from full or partial life-cycle (chronic) studies, and both invertebrate data must be from full or partial life-cycle studies. Rather than using NOEC values, as in most methodologies, the Canadian methodology uses the lowest observable effect level (LOEL; equivalent to LOEC) to derive guidelines. Acute-to-chronic ratios may be used to convert acute data to chronic. For most substances, a plant study of unspecified duration is required. However, for highly phytotoxic substances, four acute and/or chronic studies are required (with no definition given for acute vs. chronic for plants).

The UK guidelines (Zabel & Cole 1999) use both acute and chronic data for derivation of annual average (AA) concentrations, but use only acute data for derivation of maximum allowable concentrations (MAC). Chronic data may be in the form of chronic or sub-chronic NOECs, MATCs, or chronic EC₅₀s. No guidance is given on how to distinguish between acute and chronic data for non-plants, but the guidelines specify that algal growth tests lasting 48-72 h represent chronic exposures, and should not be used to derive an MAC. However, tests measuring algicidal effects in a 48-72-h exposure would be appropriate for derivation of an MAC. If, however, algae are the most sensitive of species tested for a substance, then a growth inhibition EC₅₀ may be used to derive an MAC.

Both acute LC₅₀ and chronic NOEC data may be used according to the EU methodology (Bro-Rasmussen 1994), but no definition of acute or chronic is given. The EU risk assessment TGD (ECB 2003) avoids the use of the terms acute and chronic and, instead, refers to short-term and long-term tests. Short-term results are in the form of LC/EC₅₀s and long-term results are in the form of NOECs, which may be estimated from LOECs, EC₁₀s or MATCs. The only guidance given regarding what duration constitutes a short-term vs. a long-term exposure is for algae studies, which are considered short-term if less than 72 h and long-term if 72 h or longer.

The Danish, French and German guidelines (Samsoe-Petersen & Pedersen 1995, Lepper 2002) all utilize both LC₅₀ and NOEC data to derive criteria, but none of them specifically define acute vs. chronic tests.

To ensure consistency in how toxicity data are used to derive criteria, the terms “acute” and “chronic” must be defined in the derivation guidelines. Once defined, the choice to use either acute or chronic data depends on what kind of criterion is being

calculated and what kinds of data are available. Acute criteria should be derived from acute data, and chronic criteria should be derived from chronic data, but when chronic toxicity data are lacking acute data may be used to derive chronic criteria.

6.4.3.2 Hypothesis tests vs. regression analysis

As discussed in other parts of this report, current criteria derivation methodologies use toxicity data that have been summarized in the form of a NOEC, LC₅₀, EC₅₀ or some other effect level (i.e., EC₅, EC₁₀, etc., or, more generally, EC_x). Which of these values is the best to use for derivation of protective criteria? Following is a discussion of toxicity data analysis methods, which particularly focuses on problems with using NOEC values and the challenges in using EC_x values.

Ecotoxicity test data are usually analyzed by one of two methods. Hypothesis tests, which are typically used for life-cycle, partial life-cycle, and early life-stage tests, compare treatment groups to a control group to determine which of the treatment groups is significantly different from the control (Stephan & Rogers 1985). A no observed effect concentration (NOEC) or no observed effect level (NOEL) and a lowest observed effect concentration or level (LOEC or LOEL, respectively) may be derived from this type of analysis. Some methodologies use the geometric mean of the NOEC and LOEC to calculate a maximum acceptable toxicant concentration (MATC). The other widely used method for analysis of ecotoxicity data is regression analysis, which is most commonly used for acute toxicity tests, but can as easily be applied to chronic tests. In regression analysis an equation is derived that describes the relationship between concentration and effects (Stephan & Rogers 1985). Thus it is possible to make point estimates of toxicant concentrations that will cause a given level of effect (EC_x), or to predict effects for a given level of toxicant.

Many problems with hypothesis testing are described in the literature. They are summed up succinctly by Stephan & Rogers (1985) who point out seven computational and five conceptual problems with hypothesis testing, and then discuss why regression analysis is a better alternative. The computational points are briefly described here; for the conceptual points and further details, the reader is referred to Stephan & Rogers (1985).

1) Hypothesis testing can only provide quantitative information about toxicant concentrations actually tested. The estimated effect values (i.e., NOEC and LOEC) have to be one of the tested concentrations with the true NOEC lying somewhere between the NOEC and LOEC. For regulatory purposes, such as deriving water quality criteria, a single number is needed, so regulators choose to use one or the other of the NOEC or LOEC, or they use an arithmetic or geometric mean of the NOEC and LOEC. As the authors point out, hypothesis tests provide no basis for such interpolations. In contrast, regression analysis determines a relationship between concentration and effect, and so provides a means to interpolate for estimation of effects at untested concentrations.

2) Hypothesis tests are sensitive to how carefully a test was conducted (i.e., a well-conducted test typically produces low variability within treatments) and how many replicates were used. In other words, the minimum detectable significant difference between treatments decreases with increased replication and with decreasing variability between replicates. In regression analysis, the point estimate is not affected by the number of replicates or the reproducibility among replicates; only the size of the confidence limits is affected.

3) In hypothesis testing, the selection of α (type I error rate), which is usually arbitrarily chosen at 0.05, can completely change the resulting NOEC value. With regression analysis, the confidence limits will change according to α , but the point estimate will not change.

4) The effect value obtained from a hypothesis test is completely dependent on what toxicant concentrations were actually tested. Regression analysis allows for estimation of a concentration that falls between those actually tested. Consequently, regression analysis provides a way to predict an effect level for a given concentration, which cannot be done with the results of hypothesis tests.

5) Changes in statistical procedure (such as data transformations) can have large effects on results of hypothesis tests due to the discontinuous nature of the data. For example, if the results of a hypothesis test are changed by a data transformation, the change in the resulting effect level will likely be at least a factor of two, which is the reciprocal of the typical dilution factor used in toxicity tests. However, in a regression analysis, the concentration-response curve is assumed to be a smooth continuous function and results are affected very little by small changes in statistical procedures.

6) Hypothesis testing does not properly interpret data inversions. That is, if a particular toxicant concentration caused a significant effect, but a higher concentration in the same test did not, then interpretation of hypothesis test results is difficult. The same kind of result analyzed by regression would just widen the confidence limits of the point estimate.

7) Hypothesis tests require averaging of experimental units across replicates. For example, if measured concentrations for a particular treatment vary, then the concentrations must be averaged before the hypothesis test can be conducted. With regression, each experimental unit can be treated independently. If concentrations vary within intended replicates, the results can be used without averaging.

The most important conceptual point made by Stephan & Rogers (1985) is that hypothesis tests give results that are statistically significant, but have nothing to do with the biological significance of effects. Hypothesis tests are typically performed with the Type I error rate (α) defined, but without proper definition of an acceptable Type II error rate (β) and without specifying an acceptable minimum significant difference. Thus, there is no linkage of the statistics to biology. Bruce et al. (1992) observe another shortcoming

of hypothesis testing, namely, that when results are reported just as a NOEC value, information on the concentration-response curve and variability in the data is lost.

Hoekstra & Van Ewijk (1993) give examples of how NOEL values (that is, no *observable-in-this-particular-test* effect level values) are often misinterpreted as no effect levels. They cite a study by Murray et al. (1979) in which thymus gland weight was potentially reduced by as much as 25% at the NOEL, with the uncertainty due to variability in the weight of the exposed thymus glands. A study by Speijers et al. (1986) resulted in a NOEL that could potentially cause a 73% reduction in response compared to control. Mount et al. (2003), likewise, note that tests with low variability may produce a LOEC representing responses 2-3% different from the control, while a test with high variability may produce a LOEC representing responses 40+% different from the control. Stephan & Rogers (1985) found that adverse effects ranging from 10-50% different from controls have been reported as “no statistically significant effect concentrations.” Suter et al. (1987) found effect levels at the MATC in fish tests ranging from 12% for hatching to 42% for fecundity. In a more recent short communication, Crane & Newman (2000) summarized findings of studies showing that the level of effect corresponding to reported MATCs for fish averaged 28% with a range of 0.1-84%, and that power analysis of hypothesis tests for standard *Daphnia magna* and *Ceriodaphnia dubia* tests revealed that these tests are able to detect effects ranging from 25-100%. Clearly, in spite of its name, the NOEC is not a no effect level, and for derivation of protective water quality criteria it would be unacceptable to use NOEC data corresponding to such potentially high effects.

Given the apparent agreement among toxicologists that regression analysis provides better effect level estimates than hypothesis tests (Stephan & Rogers 1985, Bruce et al. 1992, Grothe et al. 1996, Moore & Caux 1997), we are faced with the problem of having a large, otherwise usable, historical chronic toxicity data set in which results are reported as NOECs derived from hypothesis tests. In some cases (i.e., if enough raw data are included in the study report) data could be re-analyzed to determine point estimates. However, that still leaves the problem of deciding what effect level best represents a no effect level. The USEPA (1991) suggests that a NOEC (for all kinds of tests and all species) is approximately equivalent to an IC₂₅ (inhibition concentration; concentration causing 25% inhibition compared to the control), while Bruce et al. (1992) chose an EC₂₀ as a level of population effect that probably would not lead to adverse effects at the community level. However, Bruce et al. (1985) state that the decision as to what is a safe level should be based on biological criteria established with consideration for the species, the measured endpoint, test design, compound degradability, and the slope of the concentration-response curve. Results of a 1994 workshop in The Netherlands indicated a preference among participants (including regulators, industry, contract laboratories, statisticians and risk assessors) for use of an EC₅ or EC₁₀ to represent a no effect level (Van Der Hoeven et al. 1997). This was determined via a questionnaire with responses ranging from EC₁ to EC₂₅. Reasons given for choosing the EC₅ and EC₁₀ were admittedly completely non-scientific: the effect level should be small because an (almost) no effect level is intended; the effect level should not be too small because of problems with accuracy and model dependence; and the effect level should be a round number. Participants felt that the effect level should depend on ecological

consequences, but that would require species-dependent values when, politically, a single effect value for all species is preferable.

Other, novel ways of analyzing toxicity data have been proposed. These include the use of parametric threshold models to derive a parametric no effect concentration (parNEC; Van Der Hoeven et al. 1997; Bedaux & Kooijman 1994, Cox 1987), models based on dynamic energy budget (DEB) theory (Kooijman 1993, Kooijman et al. 1996, Kooijman & Bedaux 1996a, 1996b, Péry et al. 2002), the use of life table evaluation techniques (Daniels & Allan 1981, Gentile et al. 1982), case-based reasoning models (Van Den Brink et al. 2002), and the use of a double bootstrap procedure to estimate demographic toxicity (e.g., toxicant effect on population growth rate; Grist et al. 2003). These models are not well developed and the results they produce have not been thoroughly compared to existing data analysis methods.

A sound approach, then, seems to be the one proposed by participants in the 1994 workshop in The Netherlands (Van Der Hoeven et al. 1997). There was overwhelming support for replacing the NOEC by a more appropriate measure. However, they recognized the need for a transition period and concluded that NOEC data may be used as a summary statistic in ecotoxicity testing if the following are reported: a) the minimum significant difference; b) the actual observed difference from control; c) the statistical test used; and d) the test concentrations. Of the alternative NOEC replacements considered at the workshop, there was really no preference for either the EC_x or parNEC approach because both have merit, and further research is needed before a choice can be made. However, according to workshop participants, if the EC_x approach is used, then the x value should be 5 or 10%.

Statistical regression methods are commonly used and widely accepted for analysis of acute toxicity data. For analysis of chronic data hypothesis tests have been more widely used, but they have fallen out of favor due primarily to dependence on experimental design and unrestrained type II error rates. Regression methods are currently preferred for analysis of chronic data. The problem is that regression methods yield EC₅, EC₁₀, or other EC_x values and science does not offer a way to decide which of those values best represents a true no-effect level. Policy decisions will have to be made in order to decide what kind of chronic data are acceptable for use in criteria derivations.

6.4.3.3 Single-species (laboratory) vs. multispecies (laboratory/field/semi-field) data

The introduction of the USEPA (1985) methodology notes that it would be ideal if we could determine no-effect concentrations for water bodies by adding various concentrations of a chemical of concern to a several clean water bodies and determining the highest concentration that causes no effect. Clearly this is not an option, so we must rely on toxicity studies of smaller scale, ranging from single- and multispecies laboratory tests to multispecies field or semi-field (microcosm or mesocosm) tests. As models of environmental exposure, the order of preference is field tests, followed by mesocosm/microcosm tests, multispecies laboratory tests, and single-species laboratory tests. However, the most abundant, reliable and easily interpretable toxicity data are from

single-species laboratory tests. All of the other types of studies are criticized for lack of standardization, lack of replication and difficulty of interpretation. This section discusses these different kinds of toxicity data, their limitations, and how they may be used in criteria derivation

Multispecies data is problematic for use in criteria derivation due its paucity and variability. On the other hand there is much debate in the literature about whether or not single-species toxicity tests are good predictors of ecosystem effects. Shulz and Liess (2001) saw significant differences in fenvalerate effects on caddisflies due to both inter- and intraspecific interactions. However, as was discussed previously, single-species toxicity tests can be successfully used in various extrapolation procedures to determine concentrations that are protective of ecosystems (Maltby et al. 2005, Hose & Van Den Brink 2004, Okkerman et al. 1993, Versteeg et al. 1999, Emans et al. 1993, USEPA 1991).

In a review of the use of multispecies, model ecosystem tests for predicting effects of chemicals in the environment, Crane (1997) concludes that more information is needed on repeatability, reproducibility and predictive ability before such tests can be used confidently for prediction of environmental effects. Kraufvelin (1999) studied Baltic Sea hard bottom littoral mesocosms and concluded that repeatability, reproducibility and ecological realism of these mesocosms were poor enough to preclude the use of such data in predictive risk assessment, or for extrapolation to natural ecosystems. Sanderson (2002) reviewed the replicability of micro- and mesocosms and found that coefficients of variation (CV) averaged 45%, with a large, outdoor mesocosms averaging 51%. Also, 88% of biotic variables measured had no statistically significant results even with a replication level (n = 3-4) that should have yielded at most 75% insignificant results.

Another problem with field or semi-field studies is that they often have little to no replication due to unmanageable logistics. One of the reviewed methodologies (OECD 1995) cites a SETAC-Europe document (1992) that asserts that unreplicated experiments may be acceptable for responses that occur in a short period of time. However, Hanson et al. (2003) found that to detect a $\geq 25\%$ change from control in microcosm exposures of *Myriophyllum* spp. to haloacetic acids would require anywhere from 2-21 replicates depending on what endpoint is measured.

The question of how well single-species toxicity tests predict field effects has been addressed by many researchers. As discussed previously, water quality criteria derived from single-species tests are protective of ecosystems in many cases. Also, Borthwick et al. (1985) showed that laboratory-derived NOECs were predictive of field effects of fenthion on pink shrimp. Likewise, Crane et al. (1999) found that freshwater amphipod response to pirimiphos methyl was the same whether exposed in 250-mL laboratory beakers or 50,000-L pond mesocosms. A caveat to that study, though, is that the amphipods were caged in the mesocosm study and so did not experience full effects of the mesocosm environment. In a review of field validation of predictions based on laboratory-derived NOECs, Persoone & Janssen (1994) conclude that, in general, NOECs

derived from single-species laboratory studies relate well to single- and multispecies NOECs derived from field studies.

Field or semi-field data are used in the Dutch methodology for comparison with ERLs derived from single-species data (RIVM 2001). They are not used as input for ERL derivation. Nonetheless, to be usable, the data must meet specific requirements. Studies must show a distinct concentration-effect relationship, derive a reliable multi-species NOEC, include several taxonomic groups, must include at least two test concentrations and a control, must include two replicates per concentration. In addition, the concentration of compound should be measured several times throughout the study and physical-chemical parameters should be monitored (pH, temperature, hardness, TSS, etc.). Test endpoints should include biomass and population density as well as species diversity and species richness.

Although field and semi-field data are not used in criteria derivation, the OECD (1995) guidelines offer criteria for assessing the acceptability of ecosystem studies, as such studies are useful for assessing effects of chemicals under field conditions. To be acceptable, test results must include a NOEC for key components of the ecosystem with a concentration-response relationship. To avoid over-prediction of toxicity, the test should include an ecosystem recovery component. Test systems should include a range of taxonomic groups, preferably including fish, and must have properly simulated ecosystem properties such as nutrient cycling and trophic structure. Physical and chemical parameters (pH, dissolved oxygen, hardness, temperature) must be monitored throughout the test. Biological response measurements should include individual level parameters (survival, growth, reproduction, bioaccumulation) as well as population (age/size structure, production, recover rates) and community level (species composition, relative abundance) measurements. Tests must be conducted at time and space scales that are appropriate to the physical-chemical characteristics of the toxicant and life history of organisms. Ecosystem studies must include a control and 2-3 test concentrations and should be duplicated.

Due to these problems, and given the relative cost-effectiveness, reproducibility and reliability of single-species toxicity tests, most methodologies do not utilize multispecies data for criteria derivation. However Australia/New Zealand, Germany, the UK, and the EU (in risk assessment TGD) do have provisions for using field or microcosm data to derive criteria as long as it meets acceptability criteria (AZNCECC & ARMCANZ 2000, Irmer et al. 1995, Zabel & Cole 1999, ECB 2003). In practice, very few criteria are derived from field data. Methodologies that do not use field or semi-field data directly, do use them as a comparison to criteria derived from single-species data (RIVM 2001, OECD 1995). In some cases a final criterion may be adjusted if strong multispecies evidence indicates that the single-species criterion is over- or underprotective (USEPA 1985, USEPA 2003a, Zabel & Cole 1999, RIVM 2001).

6.4.3.4 Traditional vs. non-traditional endpoints

Survival, growth and reproduction are traditional measurement endpoints in ecotoxicity tests. Because these effects can be readily linked to population-level effects, they are favored for use in deriving water quality criteria that are to be protective of ecosystems. Non-traditional endpoints, such as endocrine disruption, enzyme induction, enzyme inhibition, behavioral effects, histological effects, stress protein induction, changes in RNA or DNA levels, mutagenicity, and carcinogenicity, are often more sensitive than traditional endpoints, but have had very few links established between effects seen at the individual level and effects at the population, community or ecosystem level. For this reason, they are rarely used for derivation of water quality criteria.

In the USEPA (1985) methodology, non-traditional endpoints fall into the category of “other data,” and are rarely used in criteria derivation. The recent “Ambient Aquatic Life Water Quality Criteria for Tributyltin (TBT) –Final” (USEPA 2003b) utilizes data from studies of imposex in the dogwhelk, *Nucella lapillus*, to set the final chronic criterion.

In calculating aquatic ERLs, The Netherlands includes only data from endpoints that affect species at the population level, such as survival, growth and reproduction (RIVM 2001). However, a broad range of effects is included in the category of “reproductive effects,” such as histopathological effects on reproductive organs, spermatogenesis, fertility, pregnancy rate, number of eggs produced, egg fertility, and hatchability (as described in Slooff 1992). Endpoints used for calculation of criteria based on secondary poisoning include fertility, pregnancy rate, number of live fetuses, pup mortality, eggshell thinning, egg production, egg fertility, hatchability and chick survival (as described in Romijn et al. 1993). Other endpoints, such as immobility or endocrine disruption, may be used only as evidence in support of derived ERLs if the endpoints are relevant to the species or are specific to a known toxicant mode of action. Carcinogenicity and mutagenicity endpoints are not used as studies of these endpoints are difficult to evaluate and population level consequences due to these effects are unknown (RIVM 2001).

The OECD (1995) methodology prefers use of traditional endpoints, such as survival, growth and reproduction, but biochemical endpoints may be considered as well, although clear guidance is not given. German policy makers have recognized the potential for adverse effects posed by endocrine disruptors, but have not yet incorporated these concerns into water quality targets due to lack of data on the presence of endocrine disruptors in water bodies and on concentration-dependent effects (BMU 2001). Similarly, the Danish methodology (Samsoe-Petersen & Pedersen 1995) excludes data for endpoints such as enzyme activity, or hemoglobin or hormone concentrations because such effects do not translate easily into population level effects. The Canadian methodology (CCME 1999) accepts tests with endpoints of pathological, behavioral and physiological effects as secondary data (used for derivation of interim guideline values). In South Africa, only lethality is accepted as an acute endpoint, which reflects an unequivocally, irreversible effect (Roux et al. 1996), but for chronic data, any adverse effect is accepted as an endpoint because that is consistent with the precautionary approach.

In a review article on establishing causality between exposure to endocrine disruptors and effects, Segner (2005) discusses three cases in which population-level effects in wildlife could be linked to environmental substances with endocrine activity: reductions in dogwhelk (*Nucella lapillus*) populations due to imposex caused by exposure to tributyltin; reduction in predatory bird populations due to egg-shell thinning caused by exposure to DDE; and decline in Atlantic salmon populations due to effects of 4-nonylphenol on the ability of smolts to osmoregulate. However, only in the case of tributyltin is there a strong case for endocrine disruption as the mechanism of the observed toxic effects. As Matthiessen (2000) notes, the TBT case does not mean that we should jump to conclusions about endocrine effects in individuals translating into population effects for other chemicals—the data just do not exist to draw that conclusion. In a project that incorporated laboratory, semi-field and field studies to determine relationships among biomarker effects, behavioral effects, reproductive effects and biomonitoring, Triebskorn et al. (2003) found that they could extrapolate from biochemical effects to population level effects, but they could not statistically link population and community level responses to chemical, limnological and geomorphological field data. The causal relationship between environmental conditions and effects could not be established. In addition, several of the enzyme induction responses were not useful indicators of exposure or effects (Triebskorn et al. 2003).

In a thorough review of fish bioaccumulation and biomarkers, Van Der Oost et al. (2003) state that it is hard to predict to what extent biochemical alterations in a population may influence the health of the population or ecosystem. They go on to discuss several cases in which fish diseases have been linked to pollutant exposure, and exposure was linked to biomarker response, but they conclude the discussion with further comments about the difficulty of correlating biomarker responses to higher level responses.

In a study of the effects of atrazine and its degradation products on routine swimming, antipredator responses, resting respiration, and growth in red drum larvae, Del Carmen Alvarez & Fuiman (2005) saw significant effects on swimming behaviors and growth. They also found higher rates of predator-prey interactions. However, the only quantitative prediction they could make about population effects was due to reduced growth rates. The authors postulated that increased metabolic rates, due to higher swimming rates, might lead to starvation, but no quantitative link was established.

Others have also noted problems with trying to understand the significance (to populations, communities or ecosystems) of biomarkers observed in individuals. Olsen et al. (2001) found natural variability as high as 2-fold in acetylcholinesterase and glutathione S-transferase levels in *Chironomus riparius* Meigen larvae exposed at thirteen uncontaminated sites. Such variability in the absence of toxicants suggests that it would be very difficult to discern toxicant effects by monitoring activity levels of these enzymes. In a later study, Crane et al. (2002) determined that acetylcholinesterase inhibition in *C. riparius* is a good predictor of demographically important effects (e.g., reproduction), caused by exposure to the insecticide pirimiphos methyl. In the same

study, Crane et al. (2002) found that pirimiphos methyl had no effect on glutathione S-transferase activity. Callaghan et al. (2002) likewise found that acetylcholinesterase activity in *C. riparius* was a robust and specific biomarker for exposure to organophosphate pesticides and was unaffected by temperature variation, while glutathione S-transferase activity was neither robust nor specific, with induction occurring at low temperature as well as with pesticide exposure. Enzyme induction was not linked to demographic effects in the study by Callaghan et al. (2002). In a field study of the effects of bleached kraft mill effluents on fish, Kleopfer-Sams & Owens (1993) found that induction of P450 enzymes was a good biomarker for exposure, but provided no predictive power for individual health or population level effects.

De Coen & Janssen (2003) have proposed a model for predicting population-level effects based on biomarker responses in *Daphnia magna*. By measuring digestive and metabolic enzyme activities, cellular energy allocation, DNA damage, and antioxidative stress activity, they used a multivariate partial least squares model to predict time to death, mean brood size, mean total young per female, intrinsic rate of natural increase, net reproductive rate, and growth. They found that the energy-based biomarker measurements combined with measurements of DNA integrity produced good predictions of population-level effects. Maboeta et al. (2003) found a link between a biomarker and population effects in earthworms. Teh et al. (2005) found that Sacramento splittail suffered reduced survival and growth, as well as cellular stress, after a three-month recovery period following a 96-h exposure to runoff from orchards treated with diazinon and esfenvalerate. Although no significant mortality occurred during the 96-h exposure period, histopathological abnormalities were observed after a one-week recovery period in clean water. While it appears that the histopathology may have predicted the population level effects in this case, no mechanistic link was made and it is possible that the reduced growth was due to other factors.

Studies showing a predictive relationship between biochemical, behavioral, or other non-traditional endpoints and population, community or ecosystem level effects are rare. Much more research is needed before non-traditional toxicity test endpoints will be generally useful as general predictors of ecosystem no-effect levels.

6.4.3.5 Data estimated from interspecies relationships

One criticism of using single-species toxicity data for derivation of water quality criteria is that such tests are performed on a very limited number of species. For the majority of species there are no toxicity data, which can be of particular concern in the case of threatened or endangered species that are at risk for chemical exposure. This section presents some tools that have potential for predicting toxicity to untested species based on toxicity to tested species.

The concept of quantitative species sensitivity relationships (QSSRs) was developed by Vaal et al. (1997a). They looked for patterns in sensitivity variation among 26 aquatic species for 21 toxicants. While species could be qualitatively grouped according to sensitivities (e.g., vertebrates were different from invertebrates), no

quantitative predictive model could be derived. The authors noted that to further develop QSSRs their findings need to be interpreted in terms of toxicokinetics, modes of action, and relevant species characteristics. In another study, Vaal et al. (1997b) found that acute lethality of non-polar and polar narcotics is highly predictable for a broad range of aquatic species. Reactive and specifically acting chemicals tend to be much more toxic with very high variation in sensitivity between species, and their toxicity is not predictable with current information and models.

Along the same lines, the USEPA has developed interspecies correlation estimation software (ICE v 1.0), which can be used to estimate acute toxicity of all kinds of compounds to aquatic species, genera and families having little to no data, based on species that do have adequate data sets (USEPA 2003d). Toxicity estimates made by interspecies correlations work well within taxonomic families, but less well as taxonomic distance increases. The ICE models generate estimated toxicity values with confidence limits to quantify uncertainty.

QSSRs, as described by Vaal et al. (1997a,b), are not yet developed well enough to be generally useful for estimating toxicity to untested species. The EPA ICE model offers a promising technique for generating toxicity estimates for untested species, including threatened or endangered species. Estimates from ICE could be used to supplement data sets so that important untested species may be included in criteria derivations. Estimates might also be used to evaluate whether criteria derived with tested species would be protective of untested species of particular concern.

6.5 Data reduction

Data that are to be used in criteria calculation procedures often require preliminary treatment. For example, if there are multiple data for a particular combination of species, substance and endpoint, then some method has to be specified for reducing those data into a single point for that species/substance/endpoint combination. Most methodologies utilize the geometric mean to represent the best estimate of the central tendency of toxicity data, but whether to use the geometric mean or the arithmetic mean for environmental chemical data is somewhat controversial. Parkhurst (1998) argues that, for environmental chemical concentrations, the arithmetic mean is superior to the geometric mean because it is unbiased, easier to calculate, scientifically more meaningful, and more protective of public health (due to the low bias of the geometric mean). He acknowledges a few cases in which a geometric mean is preferable. One of those cases that is pertinent to criteria derivation is that of averaging ratios, such as bioconcentration factors. Even for log-normally distributed data, Parkhurst states that the arithmetic mean is preferable since it is unbiased and makes more scientific sense. He gives the example of two data sets, A = (10, 90) and B = (40, 50). The arithmetic mean of A is larger, but the mean of the logarithms of B is larger. In such a case, according to Parkhurst, a statistical comparison based on the log-transformed data may be irrelevant or misleading.

The USEPA (1985) argues that for log-normally distributed data, the geometric mean is preferred over the arithmetic mean. Parkhurst's argument regarding the low bias of geometric means not being protective does not hold for toxicity data (as opposed to environmental concentration data), since lower values are more protective. Ease of calculation is not a legitimate argument against the geometric mean, which is easily calculated with hand calculators or computers. Returning to Parkhurst's example of sets A and B, the possibility of being misled by the geometric mean in this case is not of any more or less concern than being misled by the arithmetic mean because the differences in neither the raw data nor the log-transformed data are significant due to high variability. For data that are so widely variable, especially in very small sets, it may be that no central tendency will adequately describe the data. Thus, the geometric mean is a reasonable approach for reducing toxicity data from multiple tests to a single number for criteria derivation.

The USEPA (1985) specifies that species mean acute values (SMAVs) are to be calculated as the geometric mean of available species values, while the genus mean acute values (GMAVs) are calculated as the geometric mean of all SMAVs for a given genus. If data are available for life stages that are at least a factor of two more resistant than another life stage for the same species, then the data for the more resistant life stage is not used to calculate the SMAV because the goal is to protect all life stages. Likewise, if acute values for a species or genus differ by more than a factor of ten, then some or all of the values should be excluded (guidance on how to choose what to keep or exclude is not given). The SMAV may be calculated from the result of one or more flow-through test in which toxicant concentrations were measured, but if no such data are available, then data from static or static-renewal tests with nominal toxicant concentrations are used. The same procedure applies to chronic data. Chronic values used in calculations are either the geometric means of NOEC and LOEC values (i.e., MATC values), or a value derived by regression analysis (with no indication of what effect level should be used). The South African guidelines follow the USEPA procedure for data reduction (Roux et al. 1996), but only specify the use of chronic MATCs for criteria derivation.

The Dutch methodology offers very clear instructions regarding preliminary data processing (RIVM 2001). For a given substance, if several data are available for one species for the same endpoint, then the geometric mean is calculated. If data are available for one species, but for multiple endpoints, then the data for the most sensitive endpoint is used. If data are available for multiple life stages of one species, then data from the most sensitive life stage is used. All acceptable chronic toxicity data are converted into NOEC values as follows (RIVM 2001):

- The highest reported concentration not statistically different from the control ($p < 0.05$) is the NOEC
- The highest concentration showing 10% effect or less is considered the NOEC if statistical evaluation is not possible
- A reported LOEC is converted to a NOEC by use of factors (factors may be adjusted if justified by data)
 - $\text{NOEC} = \text{LOEC}/2$ for cases where: $10\% \text{ effect} < \text{LOEC} < 20\% \text{ effect}$

- NOEC = EC₁₀ for cases where: LOEC ≥ 20% effect and dose-response relationship is available
- NOEC = LOEC/3 for cases where: 20% < LOEC < 50% effect
- NOEC = LOEC/10 for cases where: 50 ≤ LOEC ≤ 80% effect
- NOEC is reported as ≥ [highest observed no effect concentration] if none of the treatment groups was significantly different from the control; these values are not used in statistical extrapolation methods.
- “Toxic Threshold” values, as defined by Bringmann & Kühn (1977) are regarded as NOECs
- For a maximum acceptable toxicant concentration (MATC) expressed as a range of values, the NOEC is the lower value; for MATC expressed as a single value, the NOEC = MATC/2.
- NOEC values expressed as total concentrations in water are converted to dissolved concentrations if the K_p and concentration of particulate matter are known.

Further data processing required by the Dutch methodology (RIVM 2001) is that if toxicity data for a particular toxicant appear to be bimodally distributed, then statistical analysis must be performed to determine if the apparent differences are significant. This would apply to differences between, for example, freshwater and saltwater species. If the differences are not significant, then the data are combined for criteria derivation. If the differences are significant, then separate criteria must be developed.

The EU risk assessment TGD (ECB 2003) also provides instructions for how to derive LC/EC₅₀ or NOEC values from studies in which those values are not reported:

- If raw data are available, the values can be calculated. The LC/EC₅₀ should be calculated by probit or other regression technique; the NOEC may be calculated by either hypothesis test or regression (the TGD does not claim a preference for one over the other due to the continuing controversy, previously discussed);
- Results presented as LC/EC₁₀₋₄₉ can be used as LC₅₀s, but if results are presented as LC/EC_{>50} they cannot be used;
- A LOEC representing an effect >10 and less than 20% may be converted to a NOEC: NOEC = LOEC/2.
- A LOEC representing an effect > 20% is not used; an EC10 is calculated from the data and is regarded as the NOEC
- If the percent effect of a LOEC is unknown, then a NOEC cannot be estimated;
- A NOEC may be estimated as: MATC/√2;
- An EC₁₀ from a long-term test can be considered a NOEC.

The EU TGD also describes data reduction procedures for cases where there are multiple data for one species (ECB 2003). Data are selected according to which ones reflect realistic European environmental parameters. Also, the database is evaluated to be sure that information will not be lost in averaging procedures (for example, due to very sensitive endpoints). After these initial screening steps, data for the most sensitive endpoints are selected. Multiple values for the same endpoint for the same species are researched to try to determine why they are different. If data are determined to be

equivalent, then the geometric mean of values is used. If reasons for differences are found, then data may be grouped according to appropriate factors (e.g., pH ranges). The effects of all of these possible data exclusions on the final effects assessment must be explored and explained.

By the OECD (1995) methodology, if several toxicity data are available for the same species measuring the same endpoint, then the geometric mean of the values is used. If data are available for the same species, but for different endpoints (i.e., survival vs. growth vs. reproduction) then only the lowest value is used. The OECD (1995) guidelines require the use of only chronic NOEC or MATC data for statistical extrapolation procedures. For tests in which only a chronic LOEC is reported, the NOEC may be calculated as $NOEC = LOEC/2$. This conversion can only be done if the LOEC corresponds to an effect less than 20%. If the measured effect is greater than 20%, then further toxicity testing is required at lower concentrations. The factor of 2 is representative of the typical interval between test concentrations, thus if the actual interval is known to be different then it should be used instead of 2.

For derivation of high reliability TVs the Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000) require that if several NOECs are available for different endpoints for the same species for a particular substance, then the lowest NOEC (i.e., the most sensitive endpoint) is used for criteria derivation. Also, if several NOECs are available for the same effect for the same species for a substance, then the geometric mean of values is used. Acute data, which are used to derive moderate reliability TVs, are reduced the same way.

Another type of data reduction that may be necessary is removal of outliers. While the USEPA (1985) has some vague advice regarding data that should be excluded, the Australia/New Zealand guidelines provide clear instructions on how to deal with outlying data (ANZECC & ARMCANZ 2000). First, for excessively high or low data points, original papers are consulted to try to determine an explanation for the variation (e.g., differences between nominal and measured concentrations, water quality factors, errors). Also, if data are bimodally distributed, only the lower of the two groups is retained. In this methodology, the use of curve-fitting statistical extrapolation models, reduces the need to remove outliers. Data are excluded if they are from unpublished studies or derived from studies with excessively wide concentrations ranges.

Because toxicity data for a given chemical may be available in different forms (i.e., NOEC, LOEC, LC/EC_x), for different exposure durations, and for different endpoints, it is necessary to provide some guidance for selecting or standardizing values for use in criteria derivation. Instructions should also be provided for how to reduce multiple data for a given chemical/species combination to a single value, as well as for how to manage bimodal distributions and outliers.

7.0 Criteria Calculation

This section is a discussion of how criteria values are calculated by different methodologies. Exposure factors that affect toxicity are discussed because they may influence how criteria are derived or expressed. Assessment factor and statistical extrapolation methods are described and evaluated and details of criteria calculations are given. Finally, other considerations in criteria derivation are discussed, including chemical mixtures and multiple stressors, bioaccumulation and secondary poisoning, threatened and endangered species, harmonization of criteria across environmental compartments, and utilization of data.

7.1 Exposure considerations

Deriving water quality criteria typically involves just the effects assessment portion of an ecological risk assessment, without an exposure assessment. However, it is possible to incorporate aspects of exposure into the effects assessment, and that is the nature of the discussion here. That is, it is a discussion of exposure factors that affect toxicity, not of how to estimate environmental exposures.

7.1.1 Magnitude, duration, frequency

Water quality criteria that are adequately protective of aquatic life must be defined such that protection is provided against exposures of varying magnitudes, durations and frequencies. A criterion designed to protect against ongoing, chronic toxicant exposure that is stated in terms of magnitude only will be overprotective in cases of brief, mild excursions above the criterion, but will be underprotective in cases of brief, large excursions. This is an important consideration in the Sacramento and San Joaquin River basins where short-term toxicant pulses, coincident with pesticide use or storm events, occur regularly (Bailey et al. 2000, Dileanis et al. 2002, Dileanis et al. 2003, Domagalski 2000, Dubrovsky et al. 1998, Kratzer et al. 2002, Kuivila et al. 1999, Werner et al. 2000). Many single-species studies have shown that pulse exposures to toxicants can cause significant effects in aquatic organisms that could lead to population-level effects (Shulz & Liess 2000, Brown et al. 2002, Ingersoll & Winner 1982, Forbes & Cold 2005, Cold & Forbes 2004, Hodson et al. 1983). However, in mesocosm studies or in population-level analysis of single-species tests of pulsed pesticide exposures, no long-term effects were found (Heckman & Friburg 2005, Reynaldi & Liess 2004, Pusey et al. 1994). In all of the latter studies, some period of recovery was required (reported as rapid or 2-3 wk). Presumably if a community were to receive another pulse exposure before full recovery, effects of the new pulse would compound those of the first. Thus, it is important to have water quality criteria that are defined in terms of magnitude, duration and frequency in such a way that monitoring programs can be readily designed to determine exceedances.

Two basic approaches to addressing exposure are found in existing criteria derivation methodologies. First is to incorporate some combination of magnitude, duration and frequency in each criterion statement (USEPA 1985, Zabel & Cole 1999, Roux et al. 1996). Second is to derive the magnitude only and leave duration and frequency determinations to site-specific management decisions (ANZECC &

ARMCANZ 2000, CCME 1999, Lepper 2002, BMU 2001, Irmer et al. 1995, OECD 1995, RIVM 2001, ECB 2003, Bro-Rasmussen et al. 1994, Samsøe-Petersen & Pedersen 1995).

The USEPA (1985) criteria are expressed in terms of magnitude, duration and frequency, with separate acute and chronic criteria. Magnitude is determined by analysis of effects data, but duration and frequency are the same for all toxicants. The allowable exposure durations, expressed as an averaging period of 4 d for chronic toxicity and 1 h for acute toxicity, are meant to restrict concentration fluctuations above the criteria in receiving waters. As mentioned previously, a number of studies have shown that pulses of high exposure can cause greater effects in single-species toxicity tests than the same average constant concentration. It follows that minimizing the length of the averaging period will minimize concentration fluctuation during the period. At the time the USEPA criteria guidelines were developed there were not many studies to support the notion that observed chronic toxicity was likely due to toxicant effects on a sensitive life stage over a relatively short period, but recent USEPA toxicity test guidance (USEPA 2002b) indicates that chronic toxicity may be estimated by sensitive life-stage tests lasting 4-7 days, in lieu of full life-cycle tests. Thus, for chronic toxicity, the 4-d averaging period seems reasonable. Four days is long enough to observe the equivalent of chronic toxicity, but minimizes opportunities for concentration fluctuations.

The 1-h period for acute toxicity seems somewhat arbitrary and is based on, 1) the fact that it is a shorter time period than a typical acute test; and 2) a non-referenced comment that “high concentrations of some materials can cause death in one to three hours” (USEPA 1985). The Technical Support Document for Water Quality-based Toxics Control (TSD; USEPA 1991) indicates that the 1-h period is derived from data on ammonia toxicity, which implies that it is a very conservative number for toxicants that are not as fast-acting as ammonia. While the supporting data seem lacking to support the 1-h averaging period, the importance of exposure duration to toxicity is well documented (Newman & Crane 2002). Alternative methods of addressing exposure duration in criteria derivation will be discussed later.

Finally, the frequency of one exceedance every 3 years is intended to allow ecosystem recovery (USEPA 1985). Again, this number seems arbitrary in that it is based on studies (not clearly referenced) that show that ecosystems take from 6 weeks to 10 years to recover (USEPA 1985). However, the TSD (USEPA 1991) indicates that although data were lacking in 1985 to relate criteria excursions to ecological effects, the criteria are designed such that a single marginal exceedance should cause little to no ecological effect. It goes on to argue that if marginal excursions are rare, then high-stress events that require recovery time, would be extremely rare, and the 3-year interval should be very protective. In the end, the 3-year period seems to have little scientific basis.

The UK methodology states criteria in terms of magnitude and duration, but not frequency. An annual average (AA) concentration is intended to protect ecosystems against long-term exposure, while a maximum allowable concentration (MAC) is meant to protect against transient concentrations that may cause acute toxicity (Zabel & Cole

1999). Although the AA and MAC are intended to protect against different exposure durations in general, they do not include specific statements regarding duration (such as the 1-h and 4-d averaging periods stated in USEPA criteria). Defined this way, determinations of whether the AA and MAC are being met or not depend on monitoring program design.

In Australia/New Zealand, Canada, EU member nations, and the state of North Carolina criteria are expressed in terms of magnitude only, and are designed to protect against long-term exposure (ANZECC & ARMCANZ 2000, CCME 1999, BMU 2001, Lepper 2002, Irmer et al. 1995, ECB 2003, Bro-Rasmussen et al. 1994, North Carolina Department of Environment and Natural Resources). In these cases, the values derived are intended to be used by water quality managers to develop enforceable standards (which take into account factors such as use designations and economic considerations), or to trigger further data collection. Thus, allowable frequency and duration of exceedances are part of the management process, rather than the criteria derivation process.

For countries that follow EU guidance, the pesticide criteria are values not to be exceeded by the 90th percentile of the levels monitored in water (Lepper 2002), thus duration and frequency of exceedances depend entirely on monitoring design. According to Lepper (2002), an EU Expert Advisory Forum, convened in 2001 and 2002, discussed the merits of various ways of analyzing monitoring data to determine compliance with the goals of the WFD. Possibilities included annual arithmetic mean, geometric mean, median, 90th percentile, and a maximum, never to be exceeded. The Forum concluded that, from a scientific standpoint, either the arithmetic mean or the 90th percentile would be the best way to determine a reference condition, but the choice between those two was a political question. In his report, Lepper (2002) proposes that the EU should consider the use of a maximum acceptable concentration value in addition to quality standards designed to assess annual reference conditions. The maximum acceptable concentration would be a concentration not to be exceeded any time and is intended to protect against episodic exposure events.

Within Canada, British Columbia (BC) has its own criteria derivation methodology (Government of British Columbia 1995), which closely resembles that of the CCME (1999) except that the BC guidelines recommend derivation of separate acute and chronic criteria for substances that are known to be acutely toxic. Similarly, South Africa utilizes a modified USEPA (1985) methodology in which final criteria are stated as either acute effect values (AEV) or chronic effect values (CEV; Roux et al. 1996). Thus, both BC and South Africa criteria address the role of exposure duration in toxicity, although they do not address frequency.

Analysis of ecotoxicity data by time to event (TTE) methods allows simultaneous consideration of exposure magnitude and duration in effects predictions (Newman & Crane 2002). Among other things, TTE models may be used for 1) estimates of effects over any time period, rather than just at the end of an arbitrary test period; 2) extrapolation from acute to chronic exposures; 3) analysis of time-varying exposure (e.g.

pulse exposures); and 4) determination of changes in relative risk through time (Crane et al. 2002).

Among current criteria derivation methodologies, only the Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000) allow the use of the TTE methods of Mayer et al. (1994) and Sun et al. (1995) to calculate chronic toxicity values from acute toxicity data. A computer program called ACE (acute to chronic estimation; v. 2.0) is available from USEPA to do these calculations (USEPA 2003c). Unfortunately, as noted in the Australia/New Zealand guidelines, it is almost impossible to obtain the raw data required to use these models.

Time-to-event methods are part of current discussions in the US and the UK in regards to possible revisions to derivation methodologies. The Water-based Criteria Subcommittee (WCS) of the USEPA is planning to propose that kinetic-based modeling be incorporated into revised guidelines (USEPA 2005). Although the exact model has not been determined, the workgroup is looking toward a model (or models) that will describe the time-course of toxicity and will include a toxicant accumulation component. In considering improvements to the UK methodology, Whitehouse et al. (2004) recommend use of survival time modeling, accelerated life testing, and theoretically-derived functions that may be used to account for the time-dependence of toxicity (as described by Dixon & Newman 1991, Newman & Aplin 1992, Newman & McCloskey 1996, Sun et al. 1995). These methods may be used to determine the risk of death within a given time interval depending on toxicant concentration. Whitehouse et al. (2004) determined that the two-step linear regression method of Mayer et al. (2002) is a relatively easy way to generate LC_0 values (basically chronic toxicity values derived from LC_{50} data), which may then be used in construction of species sensitivity distributions for determination of hazardous concentrations (discussed in section 7.2.2). As pointed out by Whitehouse et al. (2004) the data required for time-to-event analysis (i.e. survival at 0, 24, 48h, etc.) is usually collected during standard ecotoxicity tests, but is often not reported (and not obtainable). Thus, to use this type of analysis would not require entirely new test procedures, but simply new reporting procedures.

The USEPA WCS is considering the use of population models, which will provide a way for criteria to reflect population recovery after toxic events (USEPA 2005). Such a model was used by the USEPA in derivation of dissolved oxygen criteria for the Cape Cod to Cape Hatteras region (USEPA 2000). The WCS notes that population models are quite complex and their application could be prohibitively resource intensive. Nonetheless, they really offer the best way to determine the significance of effects on survival, growth and reproduction that are typically measured in laboratory toxicity tests. Ultimately, further literature searches for ecosystem recovery studies may be the most practical way to determine appropriate exceedance frequencies.

In whatever format criteria are stated, monitoring programs have to be designed to determine compliance. For criteria that are expressed as a single number (ANZECC & ARMCANZ 2000, OECD 1995, CCME 1999, RIVM 2001, Samsøe-Petersen & Pedersen 1995, Bro-Rasmussen 1994, Irmer et al. 1995, Lepper 2000) the risk manager has to

determine how often and with what frequency a criterion can be exceeded, and then design a monitoring program that will assess compliance. For criteria that include duration and frequency components (USEPA 1985, Roux et al. 1996, Zabel & Cole 1999), the risk manager has only to design the monitoring program.

Exclusion of duration and frequency components from criteria statements leaves those two factors solely to policy-based decisions. It would be better if these components could be science-based. The USEPA (1985) format of expressing criteria is a step toward that, but the duration and frequency values used in the acute and chronic criteria statements have little scientific basis. It is possible that a review of more recent literature could strengthen those values. To give risk managers more science-based information would require the use of time-to-event models to determine the duration component, and population models and/or good ecosystem recovery studies to determine the frequency component.

7.1.2 Multipathway exposure

Aquatic life is exposed to contaminants by two routes: water and food. Water quality criteria derived from single-species laboratory studies are based on water-only exposures, which may considerably underestimate the actual dose an animal receives in the environment from the combination of water and contaminated food sources (Benson et al. 2003). An extreme example is demonstrated in a study of effects of selenium on fish (Lemly 1985). Loss of diversity and reproductive failure occurred in fish communities exposed to selenium at concentrations 10-35 times lower than concentrations causing adverse effects in laboratory studies. Benson et al. (2003) note that the extent of dose underestimation caused by ignoring food exposure has not been well studied because the significance of the food pathway has only recently gained acceptance.

Studies of hydrophobic organic chemicals also point to the importance of dietary exposure. A model comparing food and water exposures of PCBs to lake trout in Lake Michigan determined that 99% of body burdens came from food exposure. Threespine sticklebacks accumulated significantly more hexachlorobenzene when feeding on contaminated *Tubifex tubifex* compared to water-only exposures (Egeler et al. 2001). Other studies have shown that the significance of dietary uptake varies, but it is not completely clear what factors determine whether food exposure will be important. Relationships between log K_{ow} values and dietary uptake of hydrophobic chemicals in fish have been reported, but the relationship is not consistent (Gobas et al. 1988, Qiao et al. 2000). Gobas et al. (1988) found an inverse relationship between log K_{ow} and dietary uptake efficiency, with efficiency decreasing for log K_{ow} values > 7 . On the other hand, Qiao et al. (2000) found that, for chemicals with log K_{ow} values of 5 or less, 98% of fish body burden was accounted for by gill uptake, whereas for a chemical with a log K_{ow} value of 7.5, 85% of body burden was from dietary uptake. The Qiao et al. (2000) model determined that food:water concentration ratios were important predictors of the relative uptake from the two routes. For ratios $> 10^7$ uptake was predicted to be 100% from diet; at about 10^5 uptake was equally from diet and water; and for ratios $< 10^3$ uptake was 100% from water. For modeling the relationship between log K_{ow} and exposure route,

environmentally relevant food:water concentration ratios (ranging from 191 to $10^{5.9}$) were used. Fisk et al. (1998) found a significant curvilinear relationship between $\log K_{ow}$ and dietary uptake efficiency, with efficiency increasing for $\log K_{ow}$ values between 5 and 7, and then dropping off for values above 7. Other studies have found no clear relationship between $\log K_{ow}$ and uptake efficiency. One study suggests a link between level of chlorination of dioxins (Loonen et al. 1991), and another suggests an activated transport mechanism for hydrophobic organic chemicals, with uptake efficiency dependent on molecular weight (Burreau et al. 1997).

While dietary uptake has been shown to be an important exposure route for many hydrophobic organic compounds, the theoretical basis for differences in dietary uptake efficiency is not clearly established. For narcotic chemicals (those exhibiting a non-specific mode of action), Traas et al. (2004) have developed a food web model for calculation of environmental quality criteria based on internal effect concentrations. This type of model is not based on exposure, but on concentrations of contaminants already in organisms. Thus, all exposure routes are incorporated. However, the model does not work for chemicals with specific toxic modes of action, a characteristic of most newer pesticides.

Until food web or other models are further developed to incorporate multipathway exposures into criteria derivation, the best approach seems to be to continue with water-only assessments. If studies show these criteria to be underprotective, and if the substance has a $\log K_{ow}$ between 5 and 7, then dietary uptake studies specific to the compound and species affected should be done to determine if exposure has been significantly underestimated. Many modern pesticides tend to be less hydrophobic (even water soluble) chemicals making the dietary exposure route less important.

7.1.3 Water quality characteristics

Further considerations in deriving water quality criteria are whether criteria should be expressed in terms of total chemical or bioavailable chemical, and whether criteria should be adjusted for other factors (e.g., pH, temperature, interactions with other substances) that are known to affect toxicity of some substances. Criteria based on any kind of toxicity test data are based on bioavailable chemicals and thus incorporate bioavailability. However, the issue is that laboratory tests are performed in clean water under controlled water quality conditions. Such tests do not reflect the effects that water quality parameters of natural waters may have on toxicity. The issue of bioavailability is discussed in virtually all of the existing methodologies, and is addressed with respect to metals in most. However, only a few quantitatively address bioavailability, or the potential effects of pH and temperature, on toxicity of organic chemicals.

In the UK, environmental quality standards (EQS) may be expressed either as total or dissolved concentrations, but which way it is done and how it is done is based on expert judgment on a case-by-case basis. The Canadian protocol provides no specific methodology to account for water quality factors that may affect toxicity (CCME 1999). Pawlisz et al. (1998) derived Canadian water quality guidelines for the pyrethroid

deltamethrin, and although they acknowledged that deltamethrin toxicity to insects is temperature-dependent, they did not address that issue in deriving the criteria. The European Commission's Technical Guidance Document on Risk Assessment (ECB 2003) discusses effects of pH on bioavailability and toxicity of ionizable organic chemicals in Appendix XI. It indicates that toxicity tests ought to be conducted at pH levels both above and below the pK_a for the test substance. However, since this is rarely done (because toxicity tests must be conducted in narrowly prescribed pH ranges) effects of pH on toxicity can only be qualitatively discussed as part of a risk assessment.

The Australia/New Zealand guidelines acknowledge that suspended solids, dissolved organic matter and total organic carbon levels in water may affect bioavailability, and thus toxicity, of organic compounds. However, the guideline authors did not feel that such solid/toxicant interactions are understood well enough to allow specific quantitative guidance for national criteria setting. Guidance is given for case-by-case, site-specific evaluation of bioavailability. If quantitative relationships exist between toxicity and some parameter affecting that toxicity, such as pH or temperature, then factors may be applied to calculate a site-specific target value. Lacking such a generally applicable quantitative relationship, the use of direct toxicity assessment (DTA) using local waters and local conditions is recommended (ANZECC & ARMCANZ 2000).

A few methodologies offer very specific guidance on how to express criteria as either total or dissolved concentrations. In Germany, if a substance has a suspended particulate matter-water partition coefficient greater than 1000 l/kg, the target is expressed as the level in suspended particulate matter, and is calculated as follows (Irmer 1995, adapted from LAWA 1997):

$$QT_{SPM}(\mu g/kg) = QT_{water}(\mu g/l) \cdot \frac{K}{10^{-6} \cdot K \cdot 25(mg/l) + 1} \quad (1)$$

Where QT_{SPM} = Quality Target in suspended particulate matter ($\mu g/kg$);

QT_{water} = Quality Target in water, total ($\mu g/l$)

K = partition coefficient (l/kg)

25 = Default concentration of suspended particulate matter (mg/l)

$10e-6$ = conversion factor (kg/mg)

In The Netherlands ERLs (MPC and NC) for water are reported for dissolved and total concentrations based on a standard amount of suspended matter (30 mg/l). The total concentration is calculated as follows (RIVM 2001):

$$MPC_{water_total} = MPC_{water_dissolved}(1 + K_{ppm} \times 0.001 \times 0.03) \quad (2)$$

Where: 0.001 = conversion constant (g/kg)

0.03 = content of suspended matter (g/l)

K_{ppm} = partition coefficient for suspended matter/water

And: $K_{ppm} = K_{oc} \times f_{oc}$ (3)

Where: K_{ppm} = partition coefficient for standard suspended matter (l/kg)
 K_{oc} = organic carbon-normalized partition coefficient (l/kg)
 f_{oc} = fraction organic carbon (standardized at 11.72%)

And: $NC_{water_total} = \frac{MPC_{water_total}}{100}$ (4)

Where: 100 is a safety factor to account for mixture effects

Both the German and Dutch methods for determining solid/water partitioning depend on an assumption of a standard concentration of solids of some standard composition. Unfortunately, partition coefficients are highly dependent on the composition of the solids and on the nature of the contaminant (Schwarzenbach 1993). Solid/water partition coefficients can be underestimated if colloids are not completely removed from the solution phase (Wu and Laird 2004). Without partition coefficients specific to the sediments in a given sample, calculations of dissolved vs. bound pesticides could produce erroneous results. For example, Wu and Laird (2004) determined that partition coefficients for chlorpyrifos in aqueous mixtures of six different smectites ranged from 45-6,846 l/kg. Burgess et al. (2005) found partition coefficients for nonylphenol ranging from 21.3 for cellulose to 9,770 for humic acid, indicating that even normalizing to organic carbon may not produce generally applicable partition coefficients. It makes little sense to try to select a single value from such wide ranges to represent partitioning behavior for solids of all compositions.

Another feature of the Dutch methodology is the recommendation to normalize ERLs to a specific pH, or to base the ERLs on the relevant chemical species, for chemicals whose speciation, and thus bioavailability and/or toxicity, depend on pH (RIVM 2001). This sort of adjustment would apply to weak organic acids, such as phenols. Degradation of compounds and metabolite formation are also considered in this methodology. For compounds with half-lives ($t_{1/2}$) less than 4 h the MPC must be derived from the stable degradates or metabolites.

The USEPA (1985) provides detailed instructions for determination of acute and chronic criteria in cases where toxicity to two or more species is related to a water quality characteristic (hardness, pH, temperature, etc.). This method is regularly applied to metals criteria, but could also apply to pesticides whose speciation depends on pH or whose toxicity depends on temperature. The key is that a demonstrable quantitative relationship must exist between toxicity and the water quality parameter. Criteria are then expressed as mathematical formulae that describe that relationship. The USEPA “Guidelines for Deriving Numerical Aquatic Site-Specific Water Quality Criteria by Modifying National Criteria” describes the water effect ratio (WER) technique to account for differences in bioavailability due to chemical-physical characteristics of the site water (USEPA 1984a) as follows:

$$\text{Water_Effect_Ratio} = \frac{\text{Site_Water_LC}_{50}}{\text{Laboratory_Water_LC}_{50}} \quad (5)$$

The site-specific maximum concentration is then equal to national maximum concentration multiplied by the WER. A similar procedure may be used for chronic toxicity.

The site water used in WER determination is to be collected under typical conditions (i.e., not during floods or storms). However, since pesticide loadings to surface waters are typically due to storm or agricultural runoff, and since suspended solids are also higher than normal during runoff events, it would be best to have a way to express criteria in terms that reflect the covariance of pesticides and suspended solids at the time a sample is taken. The simplest method would be to derive criteria based on dissolved concentrations (as is typically done), and then to use solids data together with measurements of total concentrations and partition coefficients to determine compliance. This could be achieved using the following equation, which is given in RIVM (2001) for converting total concentrations to dissolved concentrations:

$$C_{dissolved} = \frac{C_{total}}{1 + (K \cdot S)} \quad (6)$$

Where: $C_{dissolved}$ = concentration of chemical in dissolved phase
 C_{total} = total concentration of chemical in water
 K = solid-water partition coefficient (l/kg); may be expressed as K_{oc}/f_{oc}
 S = concentration of sediment in water (kg/L)

The resulting dissolved concentration would then be compared to the water quality criterion to determine compliance.

In the Central Valley of California, suspended solids levels vary greatly. The US Geological Survey reports levels ranging from 1-330 mg/L in samples from various streams in the Sacramento River Basin and from 1-5280 mg/L in the San Joaquin River Basin (USGS 2005a,b). For pesticides with high sediment-water partition coefficients, bioavailability could vary considerably with solids levels and, ideally, this factor should be considered in derivation of water quality criteria.

In addition to consideration of bioavailability, effects of other water quality factors should be considered in deriving criteria. For organic chemicals, this applies primarily to pH and temperature. As described in USEPA (1985), if data are available to establish quantitative relationships between water quality characteristics and toxicity, then criteria should be expressed as equations reflecting that relationship.

7.2 Basic methodologies

Two basic criteria derivation methodologies are in use or proposed for use throughout the world. The aim of both methods is to extrapolate toxicity values from

available data to values that will be protective of the environment. The assessment factor (AF) method involves multiplying the lowest value of a set of toxicity data by a factor to arrive at a criterion. The statistical extrapolation method involves the use of one of several similar species sensitivity distribution (SSD) techniques to determine the criterion. These two methods are discussed in more detail in the following sections. Some countries use exclusively one or the other methods, while some use a combination of methods, depending on data availability. In a 1993 review of the statistical procedure of Van Straalen & Denneman (1989) and the assessment factor method used by USEPA (1984b; now in Nabholz 1991), Calabrese & Baldwin (1993) concluded that these two methods produced the same results and at that time there was really no strong argument for selecting one method over the other.

One of the big advantages of SSDs over the AF method is that it is possible to derive a criterion with a known level of confidence (section 7.1.2.3). This is not true of the SSD method utilized by the USEPA (1985), in which only four data points are ultimately used to derive each criterion. However, for SSD methods that calculate an HC₅ using all available data, confidence can be quantified.

7.2.1 Assessment factor (AF) method

7.2.1.1 Criteria derivation by the AF method

France, Germany, Spain, the UK, and Canada utilize only the AF method for derivation of water quality criteria (Lepper 2002, BMU 2001, Zabel & Cole 1999, CCME 1999). Others, including Australia/New Zealand, The Netherlands, USEPA, the EU, Denmark, and OECD utilize a combination of the SSD and AF methods (RIVM 2001, USEPA 1985, ECB 2003, Bro-Rasmussen et al. 1994, Samsøe-Petersen & Pedersen 1995, OECD 1995).

In France, Spain, Germany, and the UK, criteria are derived by multiplying (or dividing, depending on how the factor is expressed) the lowest toxicity value from a minimum data set by a factor. One criterion is derived that is supposed to protect against long-term exposures (Lepper 2002, Irmer et al. 1995, BMU 2001). In France, AFs of 1-1000 are applied to single-species toxicity data. For derivation of low level criteria, acute data may be used with an AF of 1, but high level criteria are derived by applying an AF of 10 to chronic NOEC data or 1000 to acute data (Lepper 2002). In Spain, data corresponding to the most sensitive organism are used in criteria derivation. LC/EC₅₀ values are multiplied by a safety factor of 0.01 and chronic NOEC values by a factor of 0.1. Further safety factors are applied to account for lack of relevant species, persistence or bioaccumulative potential and genotoxic potential (Lepper 2002).

In the UK the lowest relevant and reliable adverse effect concentration in the data set is multiplied by a safety factor. A maximum acceptable concentration (MAC) to protect from acute toxicity is derived from acute data, is derived by application of a factor of 2-10 to the lowest available acute toxicity value. An annual average (AA) concentration to protect from chronic toxicity may be derived from either acute or

chronic data, or from acceptable field data, with application of appropriate factors (from 1-100) to the lowest available toxicity value (Zabel & Cole 1999).

The Canadian methodology (CCME 1999) uses chronic LOEC values to derive criteria. If there is an adequate data set, then the lowest LOEC is divided by a factor of 10. If only acute data are available, then the lowest LC/EC₅₀ value is divided by an ACR, if one is available. The resulting estimated chronic value is then divided by 10 to derive the criterion. If no ACR is available, then the criterion is derived directly from the lowest LC/EC₅₀ by dividing it by either 20 (for non-persistent chemicals) or 100 (for persistent chemicals).

The Netherlands methodology utilizes the AF method for derivation of MPC and SRC_{ECO} values through a process called “Preliminary effect assessment.” This is not the preferred derivation method and is used only in cases where there are not at least four chronic toxicity data from four different taxonomic groups, or if only acute toxicity data are available. Assessment factors are applied according to how much of what kinds of data are available, and they range from 1 to 1000. Similarly, the OECD recommends use of an AF method for limited data sets (OECD 1995), with factors ranging from 1-1000 depending on available data. A factor of 10 is applied to the lowest NOEC or QSAR estimate of chronic toxicity from a data set that includes at least algae, crustaceans and fish. If only acute data or QSAR estimates of acute data are available, then a factor of 100 is applied if the data set includes algae, crustaceans and fish, but a factor of 1000 is applied if only one or two species are represented.

Although the USEPA (1985) does not derive criteria if adequate data sets are not available, the state of North Carolina and the Great Lakes region utilize the AF method to derive criteria when data are lacking. In addition, for derivation of criteria in California, Lillebo et al. (1988) developed an AF method which uses LOEC values. For pesticides, this method involves finding the geometric mean of the three lowest LOEC values from acceptable tests and multiplying by a factor of 0.1. This criterion is intended to protect all species in an ecosystem through long-term exposure. In the state of North Carolina, if adequate data are not available for derivation of a FAV by the USEPA (1985) methodology, then a factor of 3 is applied to the lowest available LC₅₀ value to determine an acceptable acute value. North Carolina sets aquatic life standards based on chronic toxicity. In the absence of a chronic value, a measured ACR may be applied to an acute value. If no ACR is available, then the acute value is divided by 100 (for t_{1/2} > 96 h) or 20 (for t_{1/2} < 96 h; North Carolina Department of Environment and Natural Resources 2003). For derivation of Tier I aquatic life values, the Great Lakes methodology (USEPA 2003a) follows the USEPA (1985) guidelines. However, when not enough data are available for derivation of Tier I values, Tier II values are derived using an AF method. Secondary acute values (SAVs) are derived by dividing the lowest available GMAV by a factor ranging from 4.3 (if seven GMAVs are available) to 21.9 (if only one GMAV is available). The secondary maximum concentration (SMC) is the SAV divided by 2. The secondary chronic value (SCV) is derived in one of three ways: 1) the FAV (from a Tier I procedure) is divided by a secondary acute-to-chronic ratio (SACR; derivation is described in the next section); 2) the SAV is divided by the final acute-to-chronic ratio

(FACR from Tier I); or 3) the SAV is divided by the SACR. The secondary chronic concentration (SCC) is equal to the lower of the SCV or the final plant value.

In practice, all of the current Australia/New Zealand TVs that were derived from single-species toxicity tests were calculated by the SSD method, but the ANZECC & ARMCANZ (2000) guidelines include an AF method for cases where data are lacking. Some of the TVs were derived by applying a factor of 10 to the lowest of at least three acceptable multiple species tests. To derive moderate reliability TVs when only acute data for more than five species are available, a factor of 10 is applied prior to applying the ACR. No justification for choosing a factor of 10 is given. Low reliability TVs are derived by application of factors ranging from 20-1000, with larger factors for smaller data sets containing more acute than chronic data.

The Danish methodology utilizes the EU assessment factors for its AF method (Samsøe-Petersen & Pedersen 1995), but for its SSD method (that of Wagner & Løkke 1991) it does not allow the application of default ACRs to derive NOECs.

The South African methodology (Roux et al. 1996) follows that of the USEPA (1985) very closely, except that the FAV is divided by one of several safety factors (rather than 2 in all cases) to derive the acute effect value (AEV). The FCV is calculated as in the USEPA (1985) guidance, but again, safety factors ranging from 1-1000 are applied to derive the chronic effect value (CEV). If no chronic data and no ACRs are available a CEV is derived by multiplying the FAV or final plant value (FPV) by 1000. The FPV is the lowest result from a 96-h algae test or from a chronic test with a vascular plant.

7.2.1.2 Derivation and justification of factors

Assessment factor (AF), safety factor, application factor, acute-to-chronic ratio (ACR), margin of safety: all are terms that refer to a value that is used as a multiplier for measured toxicity values to account for uncertainty in using that number to predict real-world events. Chapman et al. (1998) reviewed the use of safety factors in ecological risk assessment. They point out that despite a lack of supporting data, standardized factors of 0.1, 0.05 and 0.01 are used throughout the world in various regulatory programs (often expressed as the inverse, 10, 20, and 100, to be used as divisors). They also note that current applications of safety factors are based on policy rather than on empirical science and that they result in values that are protective, but not predictive.

According to Irmer et al. (1995) factors used in derivation of German water quality criteria were based on internationally accepted practices until 1992 when politics deemed that factors should be limited to a total value of 0.01. The factor of 0.1 is applied to extrapolate from lab, single-species tests to field conditions, and one single further factor of 0.1 may be applied for various uncertainties (e.g. if data are available showing that a species from a taxon not normally tested is more sensitive). An acute-to-chronic factor of 0.1 does not count in the limit and may be applied separately when chronic data are not available. Irmer et al. (1995) argue that limiting the total applied safety factor to

0.01 results in weak water quality targets as it is not uncommon to have acute-to-chronic ratios as high as 1000.

Factors used in the preliminary effect assessment in The Netherlands are derived from two sources. First is the Technical Guidance Document (TGD) for derivation of the Predicted No Effect Concentration (PNEC; ECB 1996 and updated ECB 2003), and second is a USEPA (1984) document cited by Van De Meent et al. (1990). A more recent version of the same USEPA procedure is now available (Nabholz 1991). The TGD (ECB 2002) factors are based on the uncertainty associated with intra- and inter-laboratory variation of toxicity data, intra- and inter-species variations, short-term to long-term toxicity extrapolation and laboratory to field extrapolation (which includes mixture effects). For each of these extrapolations a factor (in this case divisor) of between 1 and 10 is applied and if multiple extrapolations are required, then the factor can be as high as 1000. The USEPA document (Nabholz 1991) recommends factors ranging from 1 to 1000, which also depend on the amount and type of data available. They are also based on uncertainties due to extrapolations from acute to chronic, laboratory to field and from extremely small (i.e. $n = 1$) acute data sets. For both of these sets of factors, if data are available an acute-to-chronic factor may be calculated, rather than using a default value of 10.

In the UK factors are assessed to account for uncertainty arising from extrapolation from one species to another, short to long exposure times, acute to chronic effects, chronic to ecosystem effects, and effects in one ecosystem to those in another (Zabel & Cole 1999). The size of the factor depends on the size of the data set and whether data are available for what is expected to be the most sensitive group. An additional factor may be applied if substance is bioaccumulative (usually substances with $MW < 700$ and $BCF > 100$ or $\text{Log } K_{ow} > 3$). Factors range from 1-1000 and their application relies a lot on expert judgment.

According to the OECD (1995) guidelines, assessment factors are empirically derived; that is they have no theoretical basis, but are based on experience in effects assessment. In this methodology a factor of 10 is applied for each of three possible extrapolation steps: 1) laboratory-derived NOEC to field; 2) short to long exposure times; and 3) acute to chronic effects. Alternatively, the OECD (1995) allows use of assessment factors given in the EU risk assessment methodology (ECB 2003, discussed elsewhere).

Acute-to-chronic ratios (ACRs) are used in the USEPA methodology (1985) to derive chronic criteria when chronic data are lacking. ACRs are calculated from chronic data for which at least one corresponding acute value is available (from the same study, or from a different study using the same dilution water). Species mean ACRs are calculated as the geometric mean of all available ACRs for that species. To calculate the final ACR, one of four methods is used: 1) for materials for which the species ACRs covary with the SMAV, the ACR is calculated using only species whose SMAVs are close to the FAV ("close" is not defined); 2) if there is no covariance, and all of the ACRs for a set of species are within a factor of ten, then the final ACR is calculated as the geometric mean of all species ACRs, including both freshwater and saltwater species;

3) for acute tests with shellfish embryos and larvae a final ACR of 2.0 is used, which makes the FCV equal to the criterion maximum concentration (CMC); and 4) for species with mean ACRs less than 2.0, a final ACR of 2.0 is used due to possible acclimation of test species to the toxicant. If a final ACR cannot be determined by any of these methods, then it is likely that neither the final ACR nor the FCV can be calculated.

Factors used in deriving SAVs in the Great Lakes guidance range from 4.3 to 21.9, depending on how many of the minimum Tier I data requirements are met. For example, if seven data from different families are available, then the factor is 4.3, but if only one datum is available, then the factor is 21.9. According to Pepin (pers. comm. 2005) these factors are based on a USEPA study by Host et al. (1995), which presents several methods for deriving factors to use for data sets that are smaller than the minimum eight values. A secondary acute to chronic ratio (SACR) is derived by using any available measured ACRs plus enough assumed ACRs of 18 to give a total of 3 ACRs. For example, if no measured ACRs are available, then three assumed ACRs of 18 are used. If two measured values are available, then just one assumed value is used. The geometric mean of the three values is then used as the SACR, which is used to calculate the SCV, just as the ACR is used to calculate and FCV. For the AF method used by the state of North Carolina, a factor of 3 is applied to the lowest available LC₅₀ value. Factors ranging from 20-100 are used as default ACRs. No justification for these factors is given (North Carolina Department of Environment and Natural Resources). Lillebo et al. (1988) use an additive toxicity model to derive the suggested factor of 0.1 to apply the geometric mean of the three lowest LOECs among acceptable studies. As this value was derived from metals effects data, the applicability to pesticides may or may not be valid.

The factor of 10 used in deriving TVs from multi-species data in Australia/New Zealand is to account for variations in mesocosm types and for the fact that more sensitive species may not have been in the test systems (ANZECC & ARMICANZ 2000). No particular justification is given for factors used to derive moderate and low reliability TVs, however, they are similar to those provided in the OECD guidelines (OECD 1995) on which much of the Australia/New Zealand methodology is based. Acute-to-chronic conversions are accomplished in the Australian/New Zealand guidelines in one of three ways: a chemical-specific ACR is applied; an LC₀ is calculated according to Mayer et al. (1994) and Sun et al. (1995); or a default ACR of 10 or more is applied. The chemical-specific ACR is the ratio of an acute EC₅₀ to a chronic NOEC. If multiple ACRs are available, the geometric mean of all ACRs for all species is used for derivation of criteria by the SSD method, while the ACR for the most sensitive organism is used for the AF method.

The factors used in the EU guidance (Bro-Rasmussen et al. 1994) range from 10 to account for experimental variability, to 100 to account for lack of NOEC data, to 1000 to account for lack of NOEC and LC₅₀ data. Although not specifically stated, the discussion in Bro-Rasmussen et al. (1994) suggests that the final EU factor could be adjusted if judged necessary due to bioaccumulative potential, persistence, carcinogenicity, mutagenicity, or other data supporting further concern. The French and Spanish methodologies utilize the EU factors, although Spain includes the possibility of a

factor as high as 100,000 if only acute data are available for a compound that is lacking ecotoxicity data for relevant species, is persistent or bioaccumulative, and has genotoxic potential (Lepper 2002). The EU risk assessment TGD (ECB 2003) uses assessment factors ranging from 1-1000, with the choice of the appropriate size depending to a large extent on professional judgment. The factors are intended to account for variability of laboratory toxicity data, variability within and between species, short-term to long-term exposure extrapolation, and laboratory to field extrapolation (which includes effects of mixtures). The more toxicity data available for species of different trophic levels, different taxonomic groups, and different lifestyles, the smaller is the applied factor. If only one acute value is available from each of three trophic levels, a factor of 1000 is applied. If only a single chronic NOEC is available from either a fish or daphnid test, a factor of 100 is applied. For two long-term NOECs from two different trophic levels, the factor is 50. A factor of 10 is used if chronic NOECs are available from at least three species representing three different trophic levels. Factors of 1-5 are applied to results of SSD extrapolations. For field or model ecosystem data, the size of the factor is on a case-by-case basis.

Acute and chronic safety factors used in the South African methodology (Roux et al. 1996) are intended to compensate for missing information. They are applied to account for lack of enough data to assess inter- and intra-species variability, and lack of chronic data. Acute safety factors range from 1-100 and depend on completeness of the data set. For example, if the minimum data set is available, and includes data for more than one test in at least three taxonomic groups, then a factor of 1 is applied. At the other extreme, if as little as one acute result is available, a factor of 100 is applied to the FAV. The chronic safety factors range from 1-100 and also depend on completeness of the data set. For example, if the chronic data base contains ACRs or chronic exposure data for at least on species from three different taxa (including at least one fish), then the factor is 1. If only acute data are available, then a factor of 1000 is applied to the FAV or FPV to arrive at the CEV. In the South African methodology, ACRs are derived by dividing the geometric mean of available acute values by the geometric mean of chronic values, where acute and chronic values were obtained in the same test, or in tests run in similar dilution water. The same concerns discussed for the USEPA methodology regarding covariance of SMAVs and ACRs, and for intra-species ACR variability, apply in the South African methodology.

Canadian factors range from 10 to 100, but the total factor applied could be higher if, for example, a measured ACR is higher than 10. A factor of 10 is applied to chronic data to account for variability in species sensitivity, extrapolation from laboratory to field, and differences in test endpoints. Higher level factors are applied to acute data when no chronic data are available and are used to extrapolate from acute to chronic exposures, or to derive criteria directly, as described earlier.

There is no theoretical basis for any of the assessment factors used by the various criteria derivation methodologies. They are all empirically derived numbers. The origin of generic factors of 10 for each step of uncertainty is not clear; those methodologies that indicate a reason for the selection of a value of 10 simply state that it is widely accepted.

Measured ACRs seem to have a firmer basis in empirical evidence, but they are usually derived for a particular chemical and for a particular species and then are applied to other species or groups of species, leading to further uncertainty in final criteria values. Issues surrounding the use of generic factors in ecological effects assessment are discussed in section 7.2.1.4.

Different default ACRs are used in different methodologies when no measured ACR is available. The Great Lakes guidance uses a value of 18 (USEPA 2003a), Canada uses either 2 or 10 (CCME 1999), the OECD, the USEPA Office of Pollution Prevention and Toxics and Australia/New Zealand use 10 (OECD 1995, Nabholz 1991, ANZECC & ARMICANZ 2000). Kenaga (1982) reports that ACRs were < 25 for 86% of 84 chemicals tested. However, for pesticides 70% of ACRs were > 25, with the largest at 18,100 for propanil. The large percentage of chemicals with ACRs < 25 was due to the fact that 93% of industrial organic chemicals fell into that category. Based on Kenaga's results, the USEPA "Guidelines for Deriving Ambient Aquatic Advisory Concentrations" (USEPA 1986) use a default ACR of 25 for calculation of advisory values, but only for low molecular weight non-ionizable organic chemicals. There is no evidence that default ACR values are appropriate for pesticides in general.

7.2.1.3 Aggregation of taxa

All of the AF methodologies, with the exception of the Great Lakes tier II procedure (USEPA 2003a), consider data for aquatic animals and plants together in criteria derivation. The criterion is based on the most sensitive species, regardless of such factors as taxon or toxicant mode of action. However, separate freshwater and saltwater criteria are typically derived. The issue of whether taxa are pooled or not is of more concern in SSD extrapolation procedures, thus it is discussed in more depth in section 7.2.2.4.

7.2.1.4 Evaluation of assessment factors

Assessment factors are recognized as a conservative approach for dealing with uncertainty in assessing risks posed by chemicals (Chapman et al. 1998). Further, Chapman et al. (1998) note that application of empirically based factors to toxicity data does not quantify uncertainty, but does reduce the probability of underestimating risk. At the same time, the use of AFs also greatly increases the possibility of overestimating risk. Chapman et al. (1998) are very concerned that AFs are typically applied generically, when they should be derived and used based on factors such as the scale, frequency and severity of potential environmental insults, or the steepness of a toxicant's dose-response curve. In their conclusion, Chapman et al. (1998) suggest the following principles for the use of safety factors: 1) data supercede extrapolation; that is, if data are available, they should be used; 2) extrapolation requires context; use of safety factors should be based on existing scientific knowledge; 3) extrapolation is not fact; estimates of effect levels obtained using safety factors should only be used as screening values, not as threshold values; 4) extrapolation is uncertain; safety factors should encompass a range rather than being a single value; 5) all substances are not the same; safety factors should be scaled

relative to different substances, potential exposures and nature of effects; and 6) unnecessary overprotection is not useful; safety factors for individual extrapolation steps should not exceed 10, and may be much lower.

Specifically addressing ACRs, Chapman et al. (1998) cited studies showing that measured ACRs can vary from 1 to 20,000. In view of this, it is unreasonable to apply a generic factor (whether of 10 or some other magnitude) across species and across substances, as is often done in criteria derivation if no chronic data are available. The reality remains, though, that adequate chronic data are generally not available and some means of extrapolation is needed. If an ACR is developed according to the principles for the use of safety factors (described above), then it will be derived in the context of the best scientific understanding of the substance and of the species under consideration, and should be a better predictor of chronic toxicity than a generic factor would be.

One possibility for reducing the need to use ACRs is found in the work of Duboudin et al. (2004) who have proposed a novel way of directly using acute toxicity data to determine a chronic HC₅ value. By using an acute to chronic transformation procedure derived from comparisons of acute and chronic SSDs within species categories, an acute data set is transformed into a chronic data set, which is then used to determine the HC₅ value.

An alternative to the use of assessment factors is the use of statistical extrapolation methods to make ecosystem predictions based on single-species toxicity data. These methods, which are discussed in the next section, still rely on a fair amount of acute-to-chronic extrapolation.

7.2.2 Species sensitivity distribution (SSD) method

According to Posthuma et al. (2002a), a number of ecologists and ecotoxicologists independently developed methods for assessment of ecosystem effects based on the variance in response to toxicants among species. The USEPA (1978a,b, 1979, 1985) was the first to utilize this species sensitivity distribution (SSD) method to extrapolate from a limited set of available data to derive water quality criteria designed to protect some portion of species in an ecosystem (Posthuma et al. 2002). In Europe, Kooijman (1987) developed the concept of deriving a hazard concentration for sensitive species (HCS; Van Straalen & Van Leeuwen 2002). According to Van Straalen & Van Leeuwen (2002) Kooijman's idea was soon followed with refinements and modifications by Van Straalen & Denneman (1989), Wagner and Løkke (1991) and Aldenberg & Slob (1993) and Aldenberg & Jaworska (2000). The most recent version of these SSD techniques appears in the ANZECC & ARMCANZ (2000) criteria derivation methodology. The main difference among these methodologies is in selection of the shape of the distribution that is used for the extrapolations (discussed further in section 7.2.2.1). Other differences are the kinds and quantity of data used in the extrapolation procedures (discussed in sections 6.2.3 and 6.3), what level of confidence is associated with derived criteria (section 7.2.2.3), and how data are aggregated in constructing the distribution (7.2.2.4). One area in which there is much agreement among methodologies

is in the selection of the 5th percentile as the cutoff for prediction of no-effect concentrations (section 7.2.2.2). Figure 1 provides a very general illustration of the SSD technique and will be referred to throughout the following discussion.

7.2.2.1 Appropriate distribution

The first step in the SSD methodology is to plot data in a cumulative frequency distribution. One approach then is to assume that those data are a random sample of all species and that if all species were sampled, they could be described by some

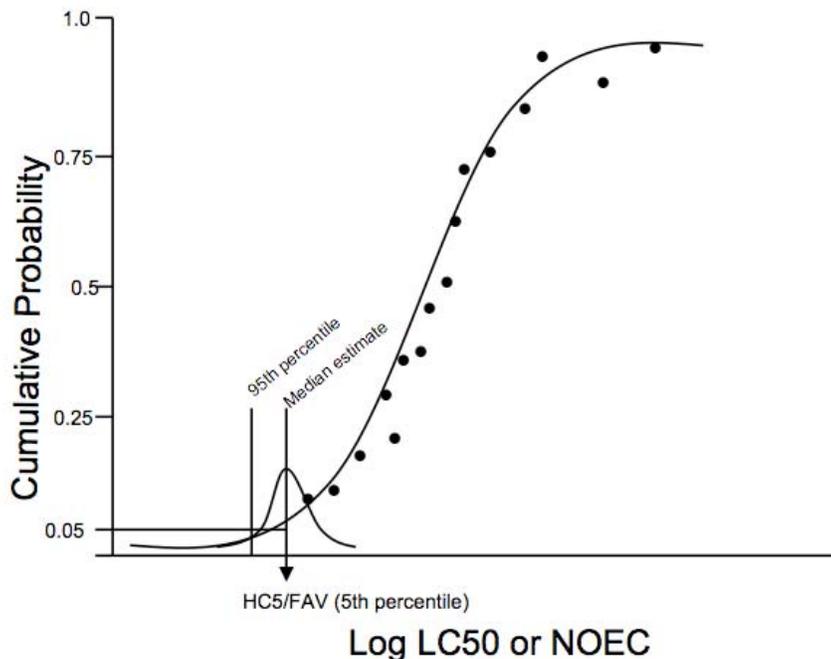


Figure 1. Generic illustration of SSD technique

distribution. The USEPA (1985) assumes a log-triangular distribution while The Netherlands methodology utilizes a log-normal distribution as described by Aldenberg & Jaworska 2000. The USEPA Office of Pesticide Programs utilizes a regression method based on a log-normal distribution, which can be used for either assessing risk (forward use) or for deriving environmental risk criteria (inverse use; Fisher & Burton 2003). Any of the SSD methods that utilize all available data (i.e., all except USEPA 1985) may be used either in the forward or inverse direction. The OECD (1995) methodology offers a choice of either the log-normal distribution method of Wagner & Løkke (1991), the log-logistic distribution method of Aldenberg & Slob (1993), or the triangular distribution of USEPA (1985) depending upon which distribution best fits the available data. Figure 2 shows the log-normal, log-logistic and log-triangular distributions.

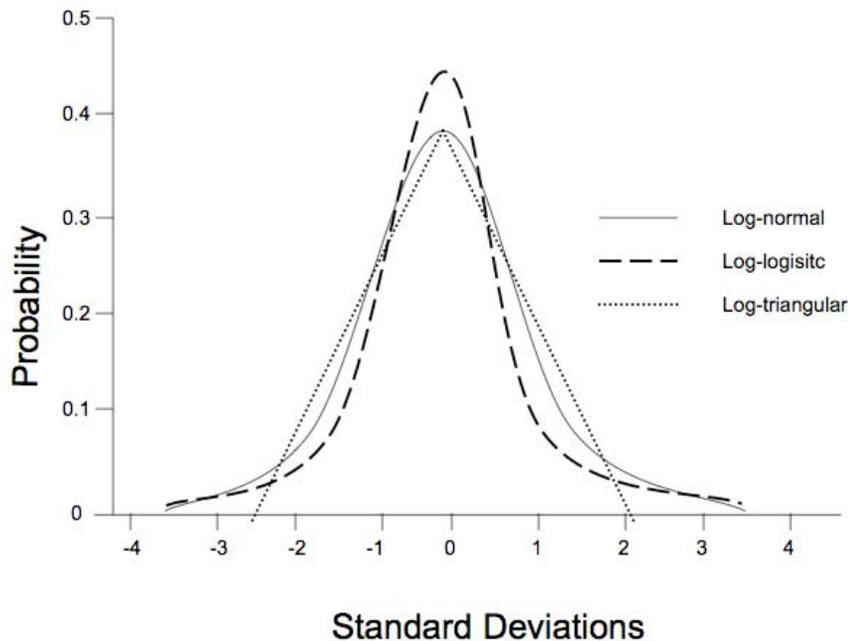


Figure 2. Comparison of log-normal, log-logistic and log-triangular distributions

The OECD (1995) observes two advantages to the USEPA (1985) method. First, because it uses a subset of some of the lowest available values, it is not affected by deviations of the highest values from the assumed distribution. Also, data reported as “greater than” may be used, which is not possible with other methods (Erickson & Stephan 1988). On the other hand, Okkerman et al. (1991) criticize the USEPA’s selection of the triangular distribution because it implies a toxicity threshold, it implies the possibility of a 100% protection level, and it only uses four (usually the lowest four) data to calculate a criterion. The authors of the Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000) did not adopt the USEPA (1985) SSD approach, noting that its data requirements are too stringent, that there is no biological basis for selecting a triangular distribution, that not all of the data are used, and that it assumes that a threshold toxicity value exists. In defense of the USEPA (1985) approach, Erickson and Stephan (1988) argue that, because the entire data set is used in setting percentile ranks and cumulative probabilities, calculation of the FAV using the four data points nearest the 5th percentile does not constitute “not using all the data. They interpret the use of those four data as a means of giving more weight to toxicity values nearest 5th percentile. This weighting leads to other problems, which are discussed in section 7.3.5.

To sidestep the issue of choosing an appropriate distribution, several researchers have suggested that it would be best to make no assumptions about distribution shapes, and to use non-parametric methods to estimate community or ecosystem effects based on

single-species toxicity tests (Jago & Newman 1997, Van Der Hoeven 2001, Grist et al. 2002). Grist et al. (2002) found significant differences between HC₅ values determined by parametric vs. non-parametric methods, with no pattern in which methods produced higher or lower estimates. Wheeler et al. (2002) suggest that to get the best HC₅ (or, generally, HC_p) estimate, data should be analyzed by four different SSD methods (two parametric and two non-parametric), selecting the one that gives the best fit. While bootstrapping techniques offer a solution to the distribution problem, they are very data intensive (as discussed earlier), and thus will not work for the many small data sets available for criteria derivation.

Arguments for one or the other distribution, or for making no distributional assumptions, are based on which distributions are easier to work with, or which ones better quell the criticism that SSDs are not valid because data usually do not fit the assumed distribution. In the end, all of the methods currently in use appear to derive protective criteria. In The Netherlands, the log-normal distribution was selected over a log-log distribution (Aldenberg & Slob 1993) because the distributions are not all that different, results obtained are not different, and the normal distribution provides powerful statistical tools (RIVM 2001). Likewise, the OECD (1995) concludes that the log-normal, log-logistic and triangular distribution methods give very similar results.

The ANZECC & ARMCANZ (2000) guidelines take the data-fitting idea a step further in a modification of the Dutch approach. By the Australia/New Zealand methodology data are fitted to one of a family of Burr distributions (Burr 1942, Shao 2000), and then the HC₅ extrapolation is done based on the best-fit distribution. This approach allows for derivation of high and moderate reliability TVs from data that would have failed assumptions of either log-normal or log-logistic distributions. Noting that the Dutch (MHSPE 1994 at the time, but now RIVM 2001) and Danish (Samsoe-Petersen & Pedersen 1995) SSD methodologies give very similar results, and differ only in the selection of either a log-logistic (Dutch; according to Aldenberg & Slob 1993) or log-normal (Danish; according to Wagner & Løkke 1991) distribution, the Australia/New Zealand guidelines chose to start with the Dutch approach because it had been more extensively evaluated and was easier to use. Advantages to the Dutch approach include that it uses the full range of available data, and a water manager can choose a level of protection and a level of uncertainty associated with a guideline value.

7.2.2.2 Percentile cutoff

To use an SSD method for criteria derivation requires selection of a percentile of the distribution as a cutoff point. This is often interpreted to mean that species lying above this point in the distribution will be protected as long as the concentration of chemical is below the concentration at the selected percentile, but species lying below the percentile would be harmed. The SSD methodologies reviewed here derive criteria using the 5th percentile of the distribution (Fig. 1). Some of the methodologies call this concentration an HC₅ (hazardous concentration affecting 5% of species). Van Straalen & Van Leeuwen (2002) note that it is not correct to interpret the HC₅ to mean that 5% of species will be harmed (as was argued, for example, by Lillebo et al. 1988, regarding the

USEPA 1985 methodology). Rather, the HC₅ approach is one method for derivation of a predicted no-effect concentration (PNEC), and although the choice of the 5th percentile is a purely a pragmatic one, it has been validated by field studies. Solomon et al. (2001), note that any percentile may be chosen as long as it can be validated against knowledge and understanding of ecosystem structure and function. Following is a discussion of percentile cutoff values used by some existing methodologies, why they were chosen, and whether they have been validated.

The USEPA rationale for choosing the 5th percentile is simply that criteria values derived using the 10th or 1st percentiles seemed too high and too low, respectively, and since the 5th falls between those, it was selected (Stephan 1985). Nonetheless, studies have shown good agreement between USEPA criteria and no-effect concentrations determined in experimental stream studies (USEPA 1991). The Dutch guidelines (RIVM 2001) use the 5th percentile for derivation of MPC values and the 50th percentile for calculation of the SRC_{ECO}. Specific reasons for these choices are not given, but the 5th percentile has been validated against field NOECs in studies by Emans et al. (1993) and Okkerman et al. (1993). The Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000) consider the question more rigorously, but still arrive at the 5th percentile level for the simple reasons that it works well in the Dutch guidelines (RIVM 2001) and it gives TVs that agree with NOEC values from multi-species tests. The reason for not regularly using a lower percentile is that the uncertainty is very high in the extreme tail of the distribution and the uncertainty can contribute more to the derived TV than the data. However, the Australia/New Zealand guidelines do use the 1st percentile as a default value for high conservation ecosystems, for bioaccumulative substances, and for cases where an important species is not protected at the 5th percentile level. To provide further information to water quality managers in Australia/New Zealand, other percentile levels are also calculated so that criteria are given based on the 1st, 5th, 10th and 20th percentiles.

Other researchers have also found good correlation between criteria derived from the 5th percentile of single-species SSDs and NOECs determined in multi-species tests (Maltby 2005, Hose & Van Den Brink 2004, Versteeg et al. 1999). On the other hand, Zischke et al. (1985) found that a laboratory-derived criterion concentration of pentachlorophenol was not protective of invertebrates and fish in outdoor experimental channels. Maltby et al. (2005) determined that concentrations of pesticides derived the 5th percentile of species sensitivity distributions with 95% confidence was protective of freshwater ecosystems, but concentrations derived with 50% confidence was not protective and required application of a safety factor.

The 5th percentile SSD cutoff, which has been validated against multi-species NOECs in several cases, is commonly used by current methodologies. It is a level that balances the desire to select a percentile near zero with the need to avoid utilizing the highly uncertain tails of the distributions.

7.2.2.3 Confidence limits

Once a percentile cutoff is chosen, it is necessary to decide what level of certainty is desired in the resulting concentration. The USEPA methodology (1985) does not provide a means to determine levels of confidence in the derived criteria. All other SSD methodologies result in a criterion derived from a specified percentile level and a specified level of confidence. Uncertainty in an extrapolated value is due to the risk that the extrapolated value is wrong (Aldenberg & Slob 1993). The distribution around the extrapolated value can be used to determine lower boundaries for the extrapolated value (Kooijman 1987, Van Straalen & Denneman 1989, Wagner & Løkke 1991, Aldenberg & Slob 1993). By evaluating this uncertainty, it is possible to state that the true HC₅ falls above (or below) the extrapolated value with, say, a 50%, 90%, 95% or other level of certainty. While all of these confidence levels may be calculated, the most statistically robust is the 50%, or median, estimate (ANZECC & ARMCANZ 2000, EVS Environmental Consultants 1999, Fox 1999). Again, the variability in the tails of the distribution tend to compound, rather than clarify, the uncertainties.

The Dutch methodology (RIVM 2001) utilizes the 50% confidence, or median, HC₅ estimate for derivation of MPCs, but they also report a 90% two-sided confidence interval. Likewise, the Dutch utilize the median estimate of the HC₅₀ for derivation of the SRC_{ECO}, but also report the 90% confidence interval. The Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000) follow the Dutch in using the median estimate of the HC₅ to derive the most probable estimate of the MTC. The Danish methodology, though, uses the lower 95th percentile estimate of the 5th percentile to derive criteria (although the Danish prefer to use an AF method; Samsøe-Petersen & Pedersen 1995). The EU risk assessment TGD utilizes the median PNEC estimate, but also considers the 95th percentile estimate in determining whether or not an assessment factor should be applied to the derived PNEC. The OECD guidance (OECD 1995) offers extrapolation factors to allow calculation of either median or 95th percentile HC₅ estimates, and leaves it to the user to choose which level to use. By way of example, Fig. 1 shows a median and lower 95th percentile estimates of the 5th percentile.

Maltby et al. (2005) investigated SSDs for pesticides and determined that the 95th percentile estimate of the SSD 5th percentile derived an HC₅ that was protective of ecosystems. The median 5th percentile level was protective in the case of a single pesticide application, but was not protective in the case of continuous or multiple applications. The authors suggest the use of a safety factor to address this. However, the SSDs in this study were constructed from acute toxicity data and it is not expected that an HC₅ derived from acute data would be protective in continuous exposure scenarios. Multiple exposures would be better addressed by consideration of a frequency component in the criterion statement.

7.2.2.4 Aggregation of taxa

As discussed earlier (7.2.2.1), one challenge in the use of SSDs is to fit the data to an appropriate distribution prior to extrapolation. One way to achieve a better fit is to

break data into groups rather than to pool it all together in one SSD. Data may be grouped according to toxic mode of action, habitat (e.g., freshwater vs. saltwater), reproductive strategy or life cycle (Solomon & Takacs 2002). Newman et al. (2000) found that cumulative frequency models that did not fit log-normal or log-logistic models had distinct shifts in slope corresponding to transitions among taxa in the ranked data set. When data are grouped according to taxa or toxic mode of action, more data sets fit the log-normal distribution (ECOFRAM 1999, Newman et al. 2002). Traas et al. (2002) also support the idea of separating data into sub-groups for different taxa, or according to toxic mode of action before constructing SSDs. In constructing SSDs for pesticides, Maltby et al. (2005) found that composition of taxonomic assemblages affected the hazard assessment, but groupings by habitat and geographic distribution had no effect.

The only criteria methodology that explicitly separates data into groups in constructing SSDs is the USEPA (1985), in which the SSD is constructed using animal data only. Plants are included in criteria derivation, but not directly. If a plant proves to be the most sensitive of species tested, then the final plant value (FPV) is the FCV. All other methodologies combine all aquatic data. The Netherlands methodology even includes NOECs derived from secondary poisoning analysis for birds and mammals (RIVM 2001). However, according to some of the guidelines, if statistical analysis shows that the data do not fit the assumed SSD distribution, or if data show a bimodal distribution, then data may be grouped to achieve a fit, with the most sensitive group used to derive the criterion, or with derivation of separate criteria (RIVM 2001, ECB 2003). In deriving target values by the Australia/New Zealand methodology (ANZECC & ARMCANZ 2000), which involves fitting data to one of several possible distributions, it was possible to use all data sets in their entirety (i.e., with all taxa combined).

The process of grouping and/or excluding data has been done in other studies. For example, in constructing an SSD for an ecological risk assessment of chlorpyrifos, Giesy et al. (1999) excluded data from rotifers, mollusks, and other insensitive organisms, although no statistical process was used to determine which data to exclude. Likewise, in a risk assessment of diazinon in the Sacramento and San Joaquin River basins, Novartis Crop Protection (1997) considered 10th percentile values for a combined fish and arthropod data set, as well as for separate fish and arthropod sets. The 10th percentile derived from the combined sets was 3,710 ng/L, while that for the fish alone was 79,900 ng/L and that for arthropods was 483 ng/L. Based on these numbers, combining the fish and arthropod data would lead to an underestimation of risk to arthropods, indicating that the data for the two groups should be analyzed separately.

When the goal of a water quality criterion is to protect all species in an ecosystem, it is important to include all species in the derivation procedure. However, it is reasonable, especially in construction of SSDs, to separate species into groups if a multimodal distribution is evident. At the same time, if there is no statistically significant difference between apparent groups (e.g., saltwater and freshwater, or plants and animals), then the data should be pooled for criteria derivation.

7.2.2.5 Criteria derivation procedures (by SSD method)

This section will simply present the nuts and bolts of each of the currently utilized SSD procedures. South Africa will not be discussed here because it utilizes the USEPA (1985) methodology. Likewise, the OECD methodology (OECD 1995) provides guidance for use of the USEPA (1985) SSD procedure, as well as those utilized (at the time) by The Netherlands (RIVM 2001) and Denmark (Samsøe-Petersen & Pedersen 1995), thus OECD SSD procedures will not be discussed further either.

7.2.2.5.1 USEPA (1985)

To calculate the FAV, the GMAVs are ordered from highest to lowest and assigned ranks from 1 to N. For each GMAV a cumulative probability (P) is calculated as $P = R/(N+1)$. The four GMAVs nearest to $P = 0.05$ are selected (for data sets with fewer than 59 GMAVs, this will always be the four lowest values in the set). Using the selected GMAVs and P values, the FAV is calculated as follows (see Erickson & Stephan 1988 for derivation):

$$s^2 = \frac{\sum((\ln GMAV)^2) - ((\sum(\ln GMAV))^2 / 4)}{\sum(P) - ((\sum(\sqrt{P}))^2 / 4)} \quad (7)$$

$$L = (\sum(\ln GMAV) - s(\sum(\sqrt{P}))) / 4 \quad (8)$$

$$A = s(\sqrt{0.05}) + L \quad (9)$$

$$FAV = e^A \quad (10)$$

Where: s^2 = variance of lowest four values in the data set
GMAV = genus mean acute value
P = percentile rank of datum
L is as defined in equation (8)
A is as defined in equation (9)
FAV = final acute value
e = base of the natural logarithm

The acute criterion, called the criterion maximum concentration (CMC), is equal to the FAV/2.

The FCV may be derived in the same manner if enough chronic data are available, however, the FCV is typically derived by application of an ACR to the FAV. The chronic criterion, called the criterion continuous concentration (CCC) is the lowest value among the FCV, the final plant value (FPV) or the final residue value (FRV; discussed further in section 7.2.3).

7.2.2.5.2 The Netherlands (RIVM 2001)

Environmental risk limits (ERLs) are derived using the SSD method of Aldenberg and Jaworska (2000). That is, HC_p values are calculated based on a log-normal SSD. The HC_5 and HC_{50} are calculated as:

$$\log HC_p = \bar{x} - k \cdot s \quad (11)$$

Where:

HC_p = Hazardous concentration for p% of species

\bar{x} = mean of log-transformed NOEC data

k = extrapolation constant depending on percentile, level of certainty and sample size (Table 1 in Aldenberg & Jaworska 2000)

s = standard deviation of log-transformed data

A computer program is available for making these calculations (RIVM 2004).

HC_5 values are used as maximum pollutant concentrations (MPC), which are used to derive environmental quality standards (EQS). A negligible concentration (NC), which serves as an EQS target value, is equal to the MPC/2. HC_{50} values are used as ecosystem serious risk concentrations (SRC_{ECO}), which are EQS intervention values (i.e., the ecosystem is seriously threatened because 50% of species are adversely affected).

7.2.2.5.3 Denmark (Samsoe-Petersen & Pedersen 1995)

The Danish methodology utilizes the SSD method of Wagner & Løkke (1991), which is essentially the same as that used in The Netherlands (RIVM 2001), but is stated in different terms and only calculates a lower one-sided confidence limit 5th percentile value. A value called a protection concentration (K_p) is calculated as follows:

$$K_p = \exp(\bar{x} - s \cdot k) \quad (12)$$

Where:

K_p = concentration protecting (100-p)% of species with a specified level of confidence

p = percentile cutoff level

\bar{x} = mean of log EC or log NOEC data

s = standard deviation

k = one-sided tolerance limit factor for a normal distribution depending on chosen confidence level (from Wagner & Løkke 1991).

The SSD method is only used in Denmark to estimate water quality criteria. An assessment factor method is preferred and given more weight in deriving criteria.

7.2.2.5.4 Australia/New Zealand (ANZECC & ARMCANZ 2000)

The Australian/New Zealand guidelines use the same method as the Dutch, but with a curve-fitting procedure that overcomes the problem of data that do not fit an assumed distribution. Using the program BurrliOZ v. 1.0.13 (CSIRO 2001; Campbell et al. 2000), data are first fitted to one of a family of Burr distributions (Burr 1942; the log-logistic distribution is in the Burr family). After an appropriate distribution is chosen, then the calculation of the median HC₅ value is the same as shown for the Dutch methodology (section 7.2.2.5.2), but utilizing extrapolation factors (k) derived for each of the distributions.

7.2.2.5.5 EU Risk Assessment Guidelines (ECB 2003)

Similar to the Australia/New Zealand approach, the EU TGD (ECB 2003) utilizes the Dutch SSD procedure, but with the provision that the distribution that best fits the data should be used. Either the Anderson-Darling or Kolmogorov-Smirnov test may be used to check goodness of fit. The PNEC is calculated as follows:

$$PNEC = \frac{5\%SSD(50\%c.i.)}{AF} \quad (13)$$

Where:

PNEC = predicted no effect concentration

5%SSD = concentration determined from species sensitivity distribution expected to protect 5% of species

50% c.i. = 50% confidence interval

AF = assessment factor of 1-5

7.2.2.6 Evaluation of SSD methodologies

When enough data are available, SSD methodologies provide a reasonable way to estimate ecosystem-level effects based on single-species data. A number of criticisms have been directed at SSDs and their use in setting regulatory limits. Most of the criticism stems from the underlying assumptions in SSD methodologies, some of which are discussed in the OECD (1995) and Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000). The most general of these assumptions are discussed here. First, is the assumption that the ecosystem is protected if 95% of species in the ecosystem are protected. This assumption may be particularly problematic if so-called keystone species are among the most sensitive to a toxicant. Any criterion derived by any method must be compared to data for species that are considered to be important for ecological, commercial or recreational reasons. If data indicate that important species will be harmed by the derived criterion, then an adjustment of the criterion is in order. The USEPA (1985) methodology stipulates that, if a species mean acute (chronic) value (SMAC or

SMCV, respectively) of a commercially or recreationally important species is lower than the calculated FAV (FCV), then the SMAC (SMCV) is used as the FAV (FCV).

Another assumption discussed by OECD (1995) and Australia/New Zealand (ANZECC & ARMCANZ 2000) is that the distribution of toxicity data is symmetrical. If insensitive species give very high toxicity values and sensitive species give very low values, then this bimodal distribution will likely have a very large standard deviation. The large standard deviation resulting from such data will lead to a very low estimate of the 5th percentile level. The best way to handle bimodal distributions seems to be to scrutinize data to see if outliers should be removed from the set (ANZECC & ARMCANZ 2000), or to split the data into two distributions and use the more sensitive data to derive criteria (RIVM 2001, ECB 2003). A third assumption (ANZECC & ARMCANZ 2000, OECD 1995) is that toxicity data represent independent, random samples from the distribution, which is not true as data tend to be from species that are easy to handle in the laboratory, or were selected for their sensitivity to a particular toxicant. For some species there are many data and for some there are none.

Posthuma et al. (2002a,b) point out a number of advantages, disadvantages, and ongoing issues in the use of SSD methods. Advantages include: 1) SSD methods are conceptually more transparent and scientifically more defensible than AF methods; 2) they are widely accepted by regulators and risk assessors; 3) they are understandable; 4) they allow risk managers to choose appropriate percentile levels and confidence levels; 5) they use commonly available ecotoxicity data; 6) they rely on relatively simple statistical methods; 7) they provide a way to assess mixtures; 8) they can be used to determine effects on species or on communities; and 8) they provide clear graphical summaries of assessment results. Disadvantages include: 1) SSD methods have not been proven to be more (or less) reliable than alternatives; 2) they require relatively large data sets; 3) they rely on statistics with no mechanistic components; 4) distributional assumptions may not be true; 5) multi-modal species distributions are problematic; 6) criteria based on lower confidence limits are overprotective; 7) test species are not a random sample; 8) there is no weighting of important species; 9) sensitive species may be over-represented; 10) important species may fall in the unprotected range; and 11) ecosystem functions are not represented. Ecological issues discussed by Posthuma et al. (2002a) include the problem of using data from a few species in laboratory conditions to represent responses of many species under field conditions. The authors note that laboratory data are often biased toward very sensitive or very tolerant species, and are from studies conducted in conditions that do not account for bioavailability and multiple routes of exposure. Statistical issues include choice of toxicological endpoint, data set distribution type, choice of percentile level to represent a no-effect, and methods of quantifying uncertainty.

In spite of violations of some of the assumptions, and in spite of the disadvantages, SSD methods have many advantages over AF methods in criteria derivation. Particularly important is the ability for risk managers to select appropriate percentile levels and confidence levels, which is not possible by the AF method. So far, criteria derived from SSDs have proven to be protective of ecosystems (section 5.1).

Further validation will come over time as the database of field studies expands (OECD 1995).

7.3 Other considerations in criteria derivation

7.3.1 Mixtures/multiple stressors

A recurring criticism of deriving water quality criteria from single-species, single-chemical laboratory toxicity tests is that such tests do not account for the multiple stressors facing organisms in the field. In the environment organisms must deal with chemical mixtures, physical stressors, and interactions with other organisms. Methods to incorporate the effects of temperature, pH, and other environmental factors into criteria derivation have been discussed (section 7.1.3). Species interactions can only be addressed in multi-species toxicity tests, which have been discussed (section 6.4.3.3). This section specifically addresses the effects of contaminant mixtures.

Results of stream monitoring in the US revealed that more than 50% of samples contained five or more pesticides (USGS 1998). The California Department of Pesticide Regulation reports that over 175 million pounds of hundreds of different pesticides were commercially applied in California in 2003 (CA DPR 2005b), thus it is likely that various mixtures will be present in surface waters due to transport processes such as drift and runoff. Studies of the effects of mixtures are few and represent an extremely small portion of the number of mixtures that could potentially occur in the environment. Water quality criteria derived from single-chemical exposures have proven to be protective of ecosystems, but the question is, if chemical A and B show additive (or synergistic or antagonistic) toxicity, then what level of each is acceptable in the environment. Lydy et al. (2004) discuss the challenges of regulating pesticide mixtures considering our limited knowledge of pesticide interactions. Alabaster & Lloyd (1982) report that joint toxicity of pesticide mixtures is more than additive in a high proportion of cases compared other kinds of toxicants. On the other hand Mount et al. (2003) point out that very few cases of extreme antagonism or synergism are observed in environmental mixtures and, therefore, an assumption of additivity is appropriate in most cases of chemical mixtures. Warne & Hawker (1995) proposed the funnel hypothesis, which states that deviations from additivity in mixtures decrease with increasing numbers of components in the mixture, thus for very complex mixtures, additivity models are likely to be valid.

Two models used to assess additive toxicity are the response addition model, in which the chemicals have different modes of action and do not interact with each other, and the concentration addition model, in which chemicals have the same mode of action, but do not interact with each other (Plackett & Hewlett 1952). According to Mount (2003), the response addition model is not widely accepted, as it is not a readily testable model, but the concentration addition model has been successfully tested for a number of modes of action and may be used to derive technically defensible criteria.

The concentration addition model is applied in the Water Quality Control Plan (Basin Plan) for the Sacramento River and San Joaquin River Basins (CVRWQCB 2004).

In the Basin Plan, in the cases where multiple chemicals with similar modes of action are present in a water body, water quality objectives are met if the following is true:

$$\sum_{i=1}^n \frac{C_i}{O_i} < 1.0 \quad (14)$$

where

C_i = concentration of toxicant i in water

O_i = water quality objective for toxicant i

In reviewing proposed Basin Plan amendments, Felsot (2005) noted that, in the case of diazinon and chlorpyrifos in particular, this additivity analysis is not appropriate because the denominator is based, not on actual toxicity values, but on an objective that includes a safety factor of 2. He proposes that a better way to determine compliance in the case of additive toxicity is to use the relative potency factor (RPF) approach, which is analogous to the toxic equivalency factor (TEF) approach used in assessing toxicity of dioxin and dioxin-like compounds. By the RPF approach, one chemical (usually the most toxic) is chosen to be the reference chemical and the potency of all other similarly-acting chemicals is expressed as a ratio of its toxicity to the toxicity of the reference. This ratio, the RPF, is multiplied by measured concentrations of each non-reference chemical to produce concentrations in terms of equivalents of the reference chemical. Compliance with the objective for the reference chemical is based on the sum of the measured reference chemical plus the concentrations of the equivalents.

The USEPA (1985) guidelines do not incorporate mixtures or multiple stressors into aquatic life criteria derivation. However, as discussed earlier, in the case of the Central Valley Regional Water Quality Control Board, regulators may use mixture models to assess compliance with objectives. Likewise, the Australia/New Zealand guidelines do not derive mixture criteria, but determine compliance using the following formula (ANZECC & ARMCANZ 2000):

$$TTM = \Sigma(C_i/WQG_i) \quad (15)$$

Where

TTM = total toxicity of the mixture

C_i = concentration of the i th component of the mixture

WQG $_i$ = water quality guideline for that component

If the TTM exceeds 1.0, then the water quality guideline has been exceeded.

For more complex mixtures (> 5 components), the Australia/New Zealand guidelines prefer the technique of direct toxicity assessment (DTA) of effluents and receiving waters (equivalent to whole effluent or ambient toxicity testing in the US). DTA is a good tool to determine if waters are able to support aquatic communities, but without follow-up toxicity identification evaluation, it does not provide information

regarding what chemical or chemicals in a mixture might be causing any observed toxicity. Again, this is a monitoring and compliance tool, not a way to address derivation of criteria for mixtures.

The Dutch, Danish, OECD, South African, Canadian, EU, Spanish and French methodologies do not directly address mixtures or multiple stressors (RIVM 2001, Samsøe-Petersen & Pedersen 1995, Roux et al. 1996, CCME 1999, Bro-Rasmussen et al. 1994, Lepper 2002). The EU risk assessment TGD and the German guidelines include mixture effects in lists of uncertainties to be addressed by assessment factors, but offers no further guidance on mixtures (ECB 2003, Irmer et al. 1995). In the UK combined EQSs may be derived for structurally similar substances with similar modes of action (Zabel & Cole 1999). Lepper (2002) proposes a method for derivation of quality standards that does not explicitly account for the toxicity of mixtures, but does utilize the AF method described in the EU risk assessment TGD (ECB 2003), which includes factors for mixture effects.

Another way to address mixtures is provided by Könemann (1981). He developed a Maximum Toxicity Index (MTI), which can be used to quantify toxicity of mixtures of two or more chemicals that have either simple similar or independent action. Könemann's model, and the others discussed to this point, only work to assess whether toxicity is additive or more-or-less than additive. Rider & LeBlanc (2005) have recently proposed a model that incorporates toxicokinetic chemical interaction as well as concentration addition and response addition. This model, called the integrated addition and interaction (IAI) model correctly predicted joint toxicity of 30 ternary mixtures containing known interacting chemicals. Models that assumed no interaction did not accurately predict joint toxicity.

Finally, SSDs offer a means of assessing mixture toxicity. Traas et al. (2002) discuss how to determine a multisubstance potentially affected fraction (msPAF; essentially the same as the HC_p) for cases of concentration addition (i.e., similar mode of action) and response addition (i.e., different modes of action). The calculations are somewhat complicated, and the reader is referred to Traas et al. (2002) for details. To calculate an overall msPAF, SSDs are first calculated individually for each chemical in the mixture. The concentration addition calculation method is then applied to groups with similar mode of action yielding a msPAF for each mode of action in the mixture. The individual PAF values that did not fit into any groups based on mode of action are aggregated with the msPAFs by the response addition calculation method to yield an overall msPAF for the mixture. It is current practice to apply the concentration addition calculation method to groups of chemicals with narcotic modes of action, as well as photosynthesis inhibitors and acetylcholinesterase inhibitors (Traas et al. 2002). Typically, this aggregation would be done within taxonomic groups, but Posthuma et al. (2002) carry the idea to the extreme and suggest that it is possible to aggregate across taxa to derive a msPAF for all species in an ecosystem. By this method, a msPAF value < 0.05 would indicate compliance with water quality criteria, while a msPAF > 0.05 would indicate non-compliance (assuming criteria were derived to protect 95% of species).

By all of the methods discussed, mixture toxicity is addressed at the compliance stage, not at the criteria derivation stage. While it would be ideal to actually derive criteria for mixtures, it would be an impossible task to try to develop criteria for all of the potential pesticide mixtures that could occur in a water body. To determine compliance based on criteria for individual chemicals, an appropriate model should be selected. If little is known about the actions and interactions of the chemicals in a mixture, then an additive assumption is reasonable and simple models may be used. However, if interactions are known to occur that lead to antagonistic or synergistic action, then a more complex model, such as that of Rider & LeBlanc (2005) should be used.

7.3.2 Bioaccumulation/secondary poisoning

Bioaccumulative chemicals pose risks that are not measured in standard laboratory toxicity tests. For chemicals that have bioaccumulative potential, many methodologies provide a way to incorporate bioaccumulation data into criteria derivation. This can be as simple as adjusting the size of the applied assessment factor (Zabel & Cole 1999, Samsøe-Petersen & Pedersen 1995, Bro-Rasmussen et al. 1994, Lepper 2002) or using a tissue residue level to determine a chronic criterion (USEPA 1985), but may involve converting food-based NOECs for fish-eating predators into water-based NOECs, which can be combined with other water-effects data in criteria derivation (RIVM 2001, OECD 1995). Others do not address bioaccumulation at all in aquatic life water quality criteria, but do so in other kinds of ecological effects assessments (CCME 1999, USEPA 2003a). For example, the Great Lakes guidance includes a procedure for derivation of water quality criteria for the protection of wildlife (USEPA 2003a). The South African methodology does not consider bioaccumulation at all (Roux et al. 1996).

The final residue value used in the USEPA (1985) methodology is intended to prevent exceedance of FDA action levels in recreationally or commercially important species, and to protect wildlife, including fishes and other animals that consume aquatic organisms in cases where adverse effects from this dietary exposure route have been demonstrated. The FRV is a water concentration derived by dividing a maximum permissible tissue concentration by a bioconcentration factor (BCF; uptake directly from water) or bioaccumulation factor (BAF; uptake from water and food). BAFs are preferred for the FRV calculation, but since BAFs are generally not available, BCFs are used. The maximum permissible tissue concentration may be an FDA action level for fish oil or the edible portion of fish or shellfish, or a maximum dietary intake that will not cause adverse effects on survival, growth, or reproduction. If multiple BCF values are available, the highest geometric mean species BCF is used. For protection of fish-eating wildlife, the BCF should be based on whole-body measurements, while for human health concerns it should be based on the edible portion of the fish (which could be the whole fish in some cases and for some cultures). The FRV is selected as the lowest of all residue values determined (for different species, including humans). If the FRV is the lowest of the FCV, FRV and final plant value (FPV), the chronic criterion is set as the FRV. The German approach is similar in that for protection of fisheries (human consumption) water quality targets may be based on dividing the allowable food residue by a bioconcentration factor (Irmer et al. 1995; BMU 2001).

Chemicals that do not pose a risk to primary producers or consumers, may pose risks to organisms, particularly terrestrial organisms, higher up the food chain if those chemicals have the potential to bioaccumulate. In The Netherlands methodology, this is addressed via consideration of secondary poisoning (RIVM 2001). For substances with $\text{Log } K_{ow} > 3$, molecular weight < 700 , low metabolism or excretion rate and/or other literature evidence of bioaccumulative potential, secondary poisoning must be considered in deriving criteria. Bioconcentration factors (BCFs), or bioaccumulation factors (BAFs; if available), are used to convert predator NOECs to water NOECs as follows:

$$NOEC_{water, fish-to-predator} = \frac{NOEC_{predator}}{BCF_{fish}} \times 0.32 \quad (16)$$

$$NOEC_{water, mussel-to-predator} = \frac{NOEC_{predator}}{BCF_{mussel}} \times 0.20 \quad (17)$$

where 0.32 and 0.20 are factors to correct for caloric content of food. These converted NOECs are combined with all other aquatic effects data from direct toxicity assessments to calculate an ecosystem MPC (MPC_{ECO}). MPCs are also calculated separately for predators and for the aquatic compartment and the independently derived values are reported for comparison.

The OECD (1995) also provides guidance for consideration of secondary poisoning. According to this methodology chemicals are likely to bioaccumulate if they have $K_{ow} > 3$, molecular weight < 1000 , molecular diameter $< 5.5 \text{ \AA}$, and molecular length $< 5.5 \text{ nm}$. Reactive and readily metabolized substances are not expected to bioaccumulate. The OECD requires that bioconcentration factors be expressed on a whole body fresh, or wet, weight basis and that they be lipid normalized. BCFs may be either measured experimentally, or may be estimated using the K_{ow} . The OECD determined that secondary poisoning risks to predatory fish is not a concern based on several modeling studies (OECD 1995 based on Barber et al. 1988, Gobas et al. 1988, Norstrom et al. 1976, Thomann & Connolly 1984). Therefore, only secondary poisoning in fish-eating mammals and birds is addressed. Unfortunately, the OECD guideline authors seem to have misinterpreted these studies. For example, Thomann & Connolly (1984) found that dietary uptake of PCBs accounted for 99% of body burden in adult trout. Also, Gobas et al. (1988) compiled data showing that average efficiency of absorption of hydrophobic organic chemicals from food for salmon and rainbow trout was 0.45 ± 0.06 for chemicals with $\text{log } K_{ow} < 7.0$, and 0.18 ± 0.04 for $\text{log } K_{ow} \geq 7.0$. Thus, dietary uptake may be less important for extremely hydrophobic chemicals, but it is still a measurable exposure route. Dietary exposure is discussed further in section 4.2. The OECD guidelines use toxicity data to derive a maximum concentration in food that will minimize risk for fish-eating wildlife (other than fish). This maximum concentration is divided by the BCF for fish to give the MTC for water (based on a method presented by Romijn et al. 1993). If the water MTC derived this way is lower than the MTC derived for protection of aquatic life, then secondary poisoning must be considered in setting criteria.

The EU risk assessment TGD (ECB 2003) describes potentially bioaccumulative chemicals as those that have a $\log K_{ow} > 3$, *or* are highly adsorptive, *or* belong to a class of chemicals known to be bioaccumulative, *or* have a structure that indicates bioaccumulative potential, *and* have no features that might mitigate bioaccumulative potential (e.g., short half-life). The TGD provides guidance for assessment of secondary poisoning, but from the angle of assessing risk to predators from dietary uptake, rather than water quality criteria setting. That is, BCFs and biomagnification factors (BMFs; relative concentration in predator compared to prey) are used to predict concentrations of contaminants in prey items based on concentrations measured in water using the following equation:

$$PEC_{oral,predator} = PEC_{water} \cdot BCF_{fish} \cdot BMF \quad (18)$$

where

$PEC_{oral,predator}$ = predicted environmental concentration in food
 PEC_{water} = predicted environmental concentration in water (from exposure assessment)
 BCF_{fish} = bioconcentration factor for fish on wet weight basis
 BMF = biomagnification factor in fish

By substituting NOEC values for the PEC values in equation 18, one could solve for $NOEC_{water}$ given a $NOEC_{oral,predator}$, BCF and BMF values:

$$NOEC_{water} = \frac{NOEC_{oral-predator}}{BCF_{fish} \cdot BMF} \quad (19)$$

Inclusion of the BMF value accounts for the fact the concentration of a contaminant in the food (fish) is due not only to direct uptake from water, but also to intake of contaminated organisms from lower trophic levels. For cases where measured BMF values are not available, default values based on several studies, are provided in the TGD. As noted, the USEPA (1985), OECD (1995) and RIVM (2001) guidelines discuss the use of BAFs (which include dietary uptake), but use BCFs (with no dietary component) in their calculations because they are more available. However, these methodologies make no attempt to correct the calculated water concentrations for dietary uptake, leading to water concentration values (e.g., FRVs in the case of USEPA) that are higher than they should be because the BMF term in the denominator reduces the $NOEC_{water}$.

The Canadian water quality criteria derivation protocol (CCME 1999) does not call for application of an additional factor for bioaccumulation, but to derive a full guideline for a bioaccumulative chemical requires that bioaccumulation data be reported; if such data are lacking, then only an interim guideline can be derived. Further, Canada has separate tissue residue guidelines (TRGs) for protection of fish-eating wildlife (CCME 1997), which are not translated into safe water levels, but are expressed in terms of safe levels in fish tissue. Due to lack of data on safe dietary levels of chemical for wildlife, many TRGs are based on human health studies.

Similar to the Canadian approach, the USEPA Great Lakes guidance (USEPA 2003a) does not incorporate bioaccumulation into aquatic life criteria. Rather, it provides for derivation of separate water quality criteria for the protection of wildlife and for protection of human health. The end result is separate water quality criteria for each of aquatic life, wildlife and human health. Presumably, the decision about which one(s) to apply to a given water body depends on beneficial use designations.

The authors of the Australia/New Zealand guidelines (ANZECC & ARMCANZ 2000) chose not to incorporate bioaccumulation into water quality criteria derivation guidelines for two major reasons. First, the link between concentrations of bioaccumulative chemicals in water and secondary poisoning is not strong. Second, there is insufficient international guidance for deriving bioaccumulation-based criteria. To address the uncertainty in the safe level of bioaccumulative chemicals, the Australia/New Zealand guidelines use the 1st percentile of the species distribution (rather than the 5th) to determine criteria for chemicals with log K_{ow} values between 3 and 7. Also, the guidelines allow for site-specific, case-by-case application of available methods for translating wildlife dietary levels into water quality criteria.

Given the potential for secondary poisoning effects in aquatic and terrestrial animals, as well as human health concerns (which affect commercially and recreationally important species), it is important to include a method of incorporating bioaccumulation into water quality criteria derivation. If a linkage can be made between dietary exposure and adverse effects (e.g., from wildlife studies, or FDA action levels) then those effects data should be used, along with BCFs, BAFs and BMFs to translate food-item concentration limits into water concentrations.

7.3.3 Threatened and endangered species

Due to their protected status, it is likely that very little toxicity test data will be available for threatened and endangered species (TES). However, it is important to ensure that these species are protected by water quality criteria. Setting national criteria for TES that have limited geographic range makes little sense, which explains why very few of the national criteria derivation methodologies specifically address TES. However, the goal of this project is to develop regional criteria, and so protection of TES in the Sacramento and San Joaquin River basins should be considered.

Among the few methodologies that do address TES, procedures are given for adjustment of criteria on a site-specific basis if data are available indicating that one or more TES may not be protected by the calculated criterion. For example, the Great Lakes guidance (USEPA 2003a) provides two methods for site-specific modifications to criteria for protection of TES (listed or proposed for listing): 1) if the SMAV for the TES or a surrogate species is lower than the FAV, then that SMAV may be used as the FAV; or 2) site-specific criteria may be derived using the criteria recalculation procedure described in the USEPA Water Quality Standards Handbook (USEPA 1994). The Australia/New Zealand guidelines suggest that TES may be protected through selection of surrogate

species (appropriate to a particular site) for inclusion in the data set used to derive TVs (ANZECC & ARMCANZ 2000).

The ICE and QSAR approaches (sections 6.4.3.5 and 6.4.2, respectively), provide the ability to quantitatively estimate toxicity for TES based on toxicity to surrogate species. While QSARs are limited to a few specifically-acting substances, ICE models can be applied to any substance. On the other hand, the ICE model has only been developed for acute toxicity, while QSARs exist for prediction of both acute and chronic toxicity. These two estimation techniques would probably best be used as a means to assess the potential for harm to TES by comparing estimated toxicity values to derived criteria.

7.3.4 Harmonization/coherence across media

The concept of harmonization is aptly described as follows (RIVM 2001): “The objective of the harmonization procedure is to compare the concentrations at steady-state in the receiving compartments...with the MPCs that have been derived for these compartments from the (eco)toxicological data. If this comparison indicates that maintaining the concentration in the primary compartment (the compartment of emission) at MPC level results in exceeding the MPC in any of the secondary compartments, the set of MPCs must be considered incoherent and has to be adjusted.” Briefly, according to the Dutch methodology (RIVM 2001) the scheme for harmonizing aquatic life water/sediment/soil ERLs is as follows: 1) if there is not sufficient direct toxicity data available for soil or sediment, then the ERL is derived from water data by the equilibrium partitioning (EqP) method, and this is the final, harmonized soil/sediment ERL; 2) if there is sufficient data to allow statistical extrapolation (by a refined effects assessment) of soil/sediment data, then the ERL is derived directly and no further harmonization is required; 3) if the soil/sediment ERL is determined by a preliminary effects assessment (by application of assessment factors to limited data sets), then this value is compared to the value determined by the EqP approach and the lower of the two values is taken as the harmonized ERL. This procedure is used with the caveat that there are uncertainties in both the ERLs and in the partition coefficients.

To harmonize ERLs with human health risk limits, The Netherlands methodology (RIVM 2001) uses a multimedia box model to estimate equilibrium concentrations in secondary (receiving) compartments given an ERL derived for a primary (emission) compartment. Utilizing Van De Meent’s (1993) SimpleBox environmental fate model, it is possible to determine if an ERL derived for a primary compartment has the potential to result in exceedance of an ERL or human health limit in another compartment. It is not clearly explained why this approach is not used to harmonize water, soil and sediment ERLs, but it seems that this type of model could be used to harmonize aquatic life criteria across environmental compartments.

In the German methodology the most sensitive asset (e.g., drinking water, aquatic life etc.) is taken as the basis for deriving the water quality objective (BMU 2001). For

example, if the drinking water target for a substance is 0.1 µg/L and the aquatic life target is 0.05 µ/L, then the aquatic life target becomes the objective.

Cross-media coherence of criteria is addressed by only the few methodologies mentioned here. Lack of attention to this issue is likely due to gaps in knowledge and lack of data for development of models to describe intermedia process. Benson et al. (2003) note that models have been successfully used for assessing possible conflicts between water and sediment criteria for some compounds, but fully integrated quantitative multimedia models are not available for making full intermedia assessments. While it may not be possible to derive fully integrated criteria, it is important to use what models are available to determine if exceedances of water quality criteria might adversely affect other environmental compartments.

7.3.5 Utilization of available data and encouragement of data generation

Many methodologies make very poor use of available data by using just the lowest (Lepper 2002, CCME 1999) or lowest few values in the data set (USEPA 1985, Roux et al. 1996). The SSD methodologies utilized in The Netherlands (RIVM 2001) and Australia/New Zealand (ANZECC & ARMCANZ 2000) make full use of data, including utilizing variability information to derive confidence limits for criteria. In particular, the Australia/New Zealand curve-fitting method reduces the need to remove outliers or truncate data sets showing some degree of multimodality.

A recurring theme throughout this review is that ecotoxicity data are generally too scarce to allow for derivation of criteria with a high level of certainty that they will neither over- nor underprotect aquatic ecosystems. Therefore, it would be beneficial if a criteria derivation methodology were designed to encourage data generation by all stakeholders. Okkerman et al. (1991) found that HC₅ values based on data for five species were lower than those based on nine species. This is because the uncertainty in the SSD method decreases with increasing sample size due to lower standard deviations and extrapolation factors.

Contrarily, for the USEPA (1985) method, which uses only the four values nearest the 5th percentile (the lowest four values in many cases) to calculate the FAV, additional data have different effects, depending upon whether the new data fall within the group of four nearest the 5th percentile. This is illustrated in a report prepared for the California State Water Resources Control Board by the Great Lakes Environmental Center (GLEC 2003). In appendix C of that report, the authors present results of various manipulations of a basic data set. First, with no change to the four values used in calculation of the FAV, simply increasing the number of samples (N), always increases the FAV as the variability in P values of the four data is reduced. Second, as the range of the four values increases (i.e., the variability of the four data increases), the FAV decreases because of the increased variability around the 5th percentile. The problem with the first of these kinds of data set manipulations is that, in an effort to derive higher criteria by the USEPA method, one could simply conduct more tests with insensitive species. Aside from causing the set to violate the log-triangular distribution assumption,

such data would drive the criterion upward in a predictable manner, based solely on N, because the new data would not be near the 5th percentile. With other SSD methodologies (i.e., those that do not ignore the upper part of the distribution) the best way to drive a criterion higher is to have a large, balanced data set, such that the variability in the whole set is reduced. By these other methods, if a data set were “padded” with extremely high or low values, outliers and bimodal distributions would be detected and the set would be modified to fix these problems prior to the SSD analysis (ANZECC & ARMCANZ 2000, RIVM 2001, ECD 2003). To encourage generation of balanced data sets, SSD methods that utilize all data (RIVM 2001, ANZECC & ARMCANZ 2000) are preferable.

Manufacturers and dischargers have little incentive to generate data if an AF method is used to derive criteria because new data showing low sensitivity are ignored, while new data showing high sensitivity will drive a standard lower (Whitehouse et al. 2004). This is because only the lowest data are used to set criteria by AF methods. However, AF methods do have some ability to encourage data generation because factors are smallest for the most complete data sets, and smaller factors yield higher criteria.

According to the Australia/New Zealand methodology, a high reliability TV can be based on results of three high quality field or mesocosm studies. There is no stipulation that such TVs will only be used if they are lower than those derived by extrapolation methods, thus multi-species research is encouraged (ANZECC & ARMCANZ 2000). On the other hand, if adequate field or mesocosm data are available that indicate a FCV should be lower than the one calculated by the USEPA (1985) methodology, then the FCV can be adjusted. This does not encourage generation of field or semi-field data by all stakeholders because the FCV can only be adjusted downward in this scenario.

8.0 Guideline Format

The greatest of criteria derivation methodologies will be of little use if it is not understandable, navigable and usable by environmental managers. Based on the reviewed guidelines, a well-formatted document will include a table of contents, a glossary, a list of acronyms, a flow chart or figure outlining the criteria derivation process, some introductory/background information, defense/explanation of selected approaches (including statements of assumptions and limitations), explicit guidance and instructions for each step of the process, details of calculations and statistical procedures, worked examples, references, and any data tables that may be needed for the process (e.g., taxonomic groups, extrapolation factors, assessment factors, etc.).

9.0 Conclusion

Table 4 is a summary of differences and similarities between key elements of the six methodologies identified in Table 2. This table highlights the strengths and weaknesses of each methodology in the areas of how data are used to derive criteria, how criteria are derived, and what other factors are considered in the final expression of criteria.

By existing methodologies, water quality criteria may be derived from single-species toxicity data by statistical extrapolation procedures (for adequate data sets) or by use of empirically-based assessment factors (for data sets of any size). Assessment factor methods are conservative and have a low probability of underestimating risk, with a concomitant high probability of overestimating risk. Extrapolation methods may also under- or overestimate risk, but in a quantifiable manner. In both methods, uncertainty is reduced with larger sample sizes. Methods are also available for criteria derivation using multispecies toxicity data, although this is rarely done due to lack of acceptable data.

Environmental toxicity of chemicals is affected by a number of factors. Given the current state of the science, some of these can be addressed in criteria derivation, and some cannot. Elements of magnitude, duration and frequency of exposure may be incorporated into criteria either through the use of time-to-event and population models, or by derivation of both acute and chronic criteria that are stated with duration and frequency components. Multipathway toxicant exposure is of concern for hydrophobic organic chemicals, but without food web models that work for chemicals with specific modes of action, it is not possible to incorporate a multipathway exposure component into criteria derivation. If data are available to establish quantitative relationships, criteria may be derived to reflect bioavailability and toxicity based on water quality characteristics.

Ecotoxicological effects and physical-chemical data are needed for criteria derivation. The quality and quantity of required data are clearly stated in existing methodologies, with very specific data quality requirements given in some cases. Lists of acceptable data sources, descriptions of adequate data searches, schemes for rating ecotoxicity data, specifications of kinds of data (e.g., acute vs. chronic), and instructions for data reduction are important and helpful features. Many methodologies present procedures for derivation of criteria from both large and small data sets. Very small data sets may be supplemented through the use of QSARs for some kinds of compounds, and through the use of models such as ICE (for prediction of toxicity to under-tested species) and ACE (for estimation of chronic toxicity from acute data).

Toxicity of mixtures is addressed by existing methodologies. In some cases, additional assessment factors are applied to criteria to account for exposure to mixtures, while in others, concentration addition models are used to assess compliance with criteria. Multiple stressors and bioaccumulation are also addressed in some methodologies by application of additional assessment factors. Methods are available for translating dietary exposure limits for humans or other fish-eating animals into water concentrations. For consideration of threatened and endangered species, a few options are available, which rely heavily on data from surrogate species to derive criteria. Utilizing partition coefficients, criteria may be harmonized across media to ensure that levels set to protect one compartment do not result in partitioning of substances to unacceptable levels in other compartments.

Table 4. Overview of similarities and differences between key elements of six major criteria derivation methodologies.

Method	Data used directly for derivation					SSD method ¹						AF method ²		Criteria Considerations											
	Sources	Evaluation criteria	QSARs allowed	Multispecies data	Endpoints not linked to SGR ³	Log Triangular	Log-normal	Burr family/best fit	Minimum number of data required	Minimum number of taxa required	Uncertainty quantified	All data used	Minimum number of data required	Minimum number of taxa required	Acute	Chronic	Magnitude	Duration	Frequency	Bioaccumulation	Additivity	Non-additivity	Bioavailability	Water quality	TES ⁴
USEPA (1985)		✓			R ⁵	✓			8	8					✓	✓	✓	✓	✓				✓	✓	
CCME (1999)		✓			S ⁶							6-9	5			✓	✓								
ANZEC/ARMCANZ (2000)		✓	✓	✓			✓		5	5	✓	✓	1	1		✓	✓			✓			✓	✓	✓
RIVM (2001)	✓	✓	✓				✓		4	4	✓	✓	1	1		✓	✓			✓				✓	
USEPA (2003)		✓				✓			8	8			1	1	✓	✓	✓	✓	✓				✓	✓	✓
ECB (2003)		✓		✓			✓		10	8	✓	✓	1	1		✓	✓			✓					

¹Species sensitivity distribution method

²Assessment factor method

³Survival/Growth/Reproduction

⁴Threatened and Endangered Species

⁵R = Rarely

⁶S = Secondary data only

Several existing methodologies prefer to derive criteria by statistical extrapolations that utilize entire data sets, as opposed to procedures that utilize only the lowest data point or points. Utilization of entire data sets allows derivation of confidence limits for criteria, and encourages data generation.

Three possible outcomes of this project are: 1) make no change in criteria derivation methodology (i.e. continue using the USEPA 1985 guidance); 2) adopt one of the other existing methodologies, or; 3) develop an entirely new methodology. Based on this review, the third outcome is most likely. Criteria derivation methodologies have improved over the past two decades as they incorporate more and more ecological risk assessment techniques. As such, no single existing methodology is ideal, but elements of several of them could be combined, along with some newer risk assessment tools, into a usable, flexible criteria derivation procedure that will produce protective criteria. Phase II of this project will involve further exploration of the various elements and models presented here to determine which are appropriate for the new methodology. Among the reviewed methodologies, those from Australia/New Zealand (ANZECC & ARMCANZ 2000), The Netherlands (RIVM 2001) and the Great Lakes (USEPA 2003a) are recommended for comparison to the new methodology in Phase III of this project. These three methodologies use widely accepted, scientifically defensible, approaches to criteria derivation. The Australia/New Zealand approach builds on that of The Netherlands, while the Great Lakes approach represents an updated version of USEPA (1985) guidelines.

10.0 References

Alabaster JS, Lloyd R. 1982. *Water Quality Criteria for Freshwater Fish*. Butterworth Scientific, Surrey, UK, pp. 253-314.

Aldenberg T, Jaworska JS. 2000. Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotox Environ Safe* 46: 1-18.

Aldenberg T, Luttik R. 2002. Extrapolation factors for tiny toxicity data sets from species sensitivity distributions with known standard deviation. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter II GW, Traas TP, eds. Lewis Publishers, CRC Press, New York, NY.

Aldenberg T, Slob W. 1993. Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotox Environ Safe* 25: 48-63.

ANZECC & ARMCANZ. 2000. Australian and New Zealand guidelines for fresh and marine water quality. Australian and New Zealand Environment and Conservation Council and Agriculture and Resource management Council of Australia and New Zealand, Canberra, Australia.

- Applegate JS. 2000. The precautionary preference: an American perspective on the precautionary principle. *Human Ecol Risk Assess* 6: 413-443.
- AQUIRE. 1981-present. AQUIRE database. US Environmental Protection Agency. Available through ECOTOX at <http://www.epa.gov/ecotox/>.
- AQUIRE (Aquatic Toxicity Information Retrieval Database). 1994. AQUIRE standard operating procedures. USEPA, Washington, DC.
- Auer CM, Nabholz JV, Baetcke KP. 1990. Mode of action and the assessment of chemical hazards in the presence of limited data: use of Structure-Activity Relationships (SAR) under TSCA, Section 5. *Environ Health Persp* 87: 183-197.
- Bailey HC, Deanovic L, Reyes E, Kimball T, Larson K, Cortright K, Connor V, Hinton DE. 2000. Diazinon and chlorpyrifos in urban waterways in Northern California, USA. *Environ Toxicol Chem* 19: 82-87.
- Barber MC, Suarez LA, Lassiter RR. 1988. Modeling bioconcentration of nonpolar organic pollutants by fish. *Environ Toxicol Chem* 7: 545-558.
- Bedaux JJM, Kooijman SALM. 1993. Statistical analysis of bioassays, based on hazard modeling. *Environ Ecol Stat* 1: 303-314.
- Benson WH, Allen HE, Connolly JP, Delos CG, Hall LW Jr, Luoma SN, Maschwitz D, Meyer JS, Nichols JW, Stubblefield WA. 2003. Exposure Analysis. In: *Reevaluation of the State of the Science for Water-Quality Criteria Development*, Reiley M, Stubblefield WA, Adams WJ, Di Toro DM, Hodson PV, Erickson RJ, Keating FJ Jr, eds. SETAC Press, Pensacola, FL.
- BIODEG. 1992. Biodegradation probability program (version 3.0). Now called BioWin. Available at http://www.syrres.com/esc/est_soft.htm.
- BMU. 2001. Environment Policy, Environment Resources Management in Germany, Part II, Quality of Inland Surface Waters. Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, Div. WAI 1(B), Postfach 12 06 29, Bonn Germany.
- Borthwick PW, Clark JR, Montgomery RM, Patrick JM Jr, Lores EM. 1985. Field confirmation of a laboratory-derived hazard assessment of the acute toxicity of fenthion to pink shrimp, *Penaeus duorarum*. In: *Aquatic Toxicology and Hazard Assessment: Eighth Symposium. ASTM STP 891*, Bahner RC, Hansen DJ, eds. American Society of Testing and Materials, Philadelphia, PA, pp. 177-189.
- Bringmann G, Kühn R. 1977. Befunde der Schädwirkung wassergefährdender Stoffe gegen *Daphnia magna* (Hazardous substances in water towards *Daphnia magna*). *Z Wasser Abwasserf* 10: 161-166.

Bro-Rasmussen F, Calow P, Canton JH, Chambers PL, Silva Fernandes A, Hoffmann L, Jouany J-M, Klein W, Persoone G, Scoullos M, Tarazona JV, Vighi M. 1994. EEC water quality objectives for chemicals dangerous to aquatic environments (List 1). *Rev Environ Contam Toxicol* 137: 83-110.

Brown MD, Carter J, Thomas D, Purdie DM, Kay BH. 2002. Pulse-exposure effects of selected insecticides to juvenile Australian crimson-spotted rainbowfish (*Melanotaenia duboulayi*). *J Econ Entomol* 95: 294-298.

Bruce RD, Versteeg DJ. 1992. A statistical procedure for modeling continuous toxicity data. *Environ Toxicol Chem* 11: 1485-1494.

Burgess RM, Pelletier MC, Gundersen JL, Perron MM, Ryba SA. 2005. Effects of different forms of organic carbon on the partitioning and bioavailability of nonylphenol. *Environ Toxicol Chem* 24: 1609-1617.

Burr IW. 1942. Cumulative frequency functions. *Ann Math Stat* 13: 215-232.

Burreau S, Axelman J, Broman D, Jakobsson E. 1997. Dietary uptake in pike (*Esox lucius*) of some polychlorinated biphenyls, polychlorinated naphthalenes and polybrominated diphenyl ethers administered in natural diet. *Environ Toxicol Chem* 16: 2508-2513.

California DPR. 2005a. Registration Desk Manual, Chapter 6, Data requirements for obtaining product registration and for label amendments. California Department of Pesticide Regulation, Sacramento, CA.

California DPR. 2005b. Pesticide Use Report, http://www.cdpr.ca.gov/docs/pur/pur03rep/03_pur.htm, California Department of Pesticide Regulation, Sacramento, CA.

California SWRCB. 2005. State Water Resources Control Board web site. <http://www.swrcb.ca.gov/about/mission.html>

Callaghan A, Fisher TC, Grosso A, Holloway GJ, Crane M. 2002. Effect of temperature and pirimiphos methyl on biochemical biomarkers in *Chironomus riparius* Meigen. *Ecotox Environ Safe* 52: 128-133.

Campbell E, Palmer MJ, Shao Q, Warne MStJ, Wilson D. 2000. BurrliOZ: A computer program for calculating toxicant trigger values for the ANZECC and ARMCANZ water quality guidelines. Perth WA.

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.

- CCME. 1997. Protocol for the derivation of Canadian tissue residue guidelines for the protection of wildlife that consume aquatic biota. Canadian Council of Ministers of the Environment, Ottawa.
- CEPA. 1999. Canadian Environmental Protection Act. CEPA Environmental Registry, http://www.ec.gc.ca/CEPARRegistry/the_act/.
- Chapman PM, Fairbrother A, Brown D. 1998. A critical evaluation of safety (uncertainty) factors for ecological risk assessment. *Environ Toxicol Chem* 17: 99-108.
- Cold A, Forbes VE. 2004. Consequences of a short pulse of pesticide exposure for survival and reproduction of *Gammarus pulex*. *Aquat Toxicol* 67: 287-299.
- Cox C. 1987. Threshold dose-response models in toxicology. *Biometrics* 43: 511-524.
- Crane M. 1997. Research needs for predictive multispecies tests in aquatic toxicology. *Hydrobiologia* 346: 149-155.
- Crane M, Attwood C, Sheahan D, Morris S. 1999. Toxicity and bioavailability of the organophosphorous insecticide pirimiphos methyl to the freshwater amphipod *Gammarus pulex* L. in laboratory and mesocosm systems. *Environ Toxicol Chem* 18: 1456-1461.
- Crane M, Chapman PF, Sparks T, Fenlon J, Newman MC. 2002. Can risk assessment be improved with time to event models? In: *Risk Assessment with Time to Event Models*, Crane M, Newman MC, Chapman PF, Fenlon J, eds. Lewis Publishers, Boca Raton, FL, pp. 153-166.
- Crane M, Newman MC. 2000. What level of effect is a no observed level? *Environ Toxicol Chem* 19: 516-519.
- Crane M, Sildanchandra W, Kheir R, Callaghan A. 2002. Relationship between biomarker activity and developmental endpoints in *Chironomus riparius* Meigen exposed to an organophosphate insecticide. *Ecotox Environ Safe* 53: 361-369.
- Cronin MTD, Walker JD, Jaworska JS, Comber MHI, Watts CD, Worth AP. 2003. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environ Health Perspectives* 111: 1376-1390.
- CSIRO. 2001. BurrliOZ v. 1.0.13. Commonwealth Scientific and Industrial Research Organization, Australia.
- CSTE/EEC. 1987. Internal report CSTE/87/101/XI from Directorate/EEC General for Environment, Nuclear Safety & Civil Protection, DG XI/A/2, Brussels.
- Curtis H, Barnes SN. 1981. *Invitation to Biology, Third Edition*. Worth Publishers, Inc. New York.

- CVRWQCB. 2004. The Water Quality Control Plan (Basin Plan) for the California Regional Water Quality Control Board Central Valley Region, fourth edition, the Sacramento River Basin and the San Joaquin River Basin.
- CVRWQCB. 2002. 2002 CWA Section 303(d) List of Water Quality Limited Segments. Central Valley Regional Water Quality Control Board web site. <http://www.swrcb.ca.gov/tmdl/docs/2002reg5303dlist.pdf>
- Daily GC, Ehrlich PR, Haddad NM. 1993. Double keystone bird in a keystone species complex. *Proc Natl Acad Sci USA* 90: 592-594.
- Daniels RE, Allan JD. 1981. Life table evaluation of chronic exposure to a pesticide. *Can J Fish Aquat Sci* 38: 485-494.
- De Coen WM, Janssen CR. 2003. A multivariate biomarker-based model predicting population-level responses of *Daphnia magna*. *Environ Toxicol Chem* 22: 2195-2201.
- Del Carmen Alvarez M, Fuiman LA. 2005. Environmental levels of atrazine and its degradation products impair survival skills and growth of red drum larvae. *Aquat Toxicol* 74: 229-241.
- Dileanis PD, Bennett KP, Domagalski JL. 2002. Occurrence and transport of diazinon in the Sacramento River, California, and selected tributaries during three winter storms, January-February, 2000. United States Geological Survey, Water-Resources Investigations Report 02-4101.
- Dileanis PD, Brown DL, Knifong DL, Saleh D. 2003. Occurrence and transport of diazinon in the Sacramento River and selected tributaries, California, during two winter storms, January-February, 2001. United States Geological Survey, Water-Resources Investigations Report 03-4111.
- Di Toro DM. 2003. Executive Summary. In: *Reevaluation of the State of the Science for Water-Quality Criteria Development*. SETAC Press, Pensacola, FL. pp. xxi-xxv.
- Domagalski J. 2000. Pesticides in surface water measured at select sites in the Sacramento River basin, California 1996-1998. United States Geological Survey, Water-Resources Investigations Report 00-4203.
- Duboudin C, Ciffroy P, Magaud H. 2004. Acute-to-chronic species sensitivity distribution extrapolation. *Environ Toxicol Chem* 23: 1774-1785.
- Dubrovsky NM, Kratzer CR, Brown LR, Gronberg JAM, Burow KR. 1998. Water quality in the San Joaquin-Tulare Basin, California 1992-95, United States Geological Survey Circular 1159, on line @ URL:<http://water.usgs.gov/pubs/circ1159>, updated April 20, 1998.

ECB. 2003. Technical guidance document on risk assessment in support of commission directive 93/67/EEC on risk assessment for new notified substances, commission regulation (EC) no. 1488/94 on risk assessment for existing substances, directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Part II. Environmental Risk Assessment. European Chemicals Bureau, European Commission Joint Research Center, European Communities.

ECETOC. 1993. Aquatic toxicity data evaluation. ECETOC Technical Report No. 56. ECETOC, Brussels.

ECOFRAM. 1999. Committee on FFRA risk assessment methods aquatic report. US Environmental Protection Agency, Washington, DC.

Egeler P, Meller M, Roembke J, Spoerlein P, Streit B, Nagel R. 2001. *Tubifex tubifex* as a link in food chain transfer of hexachlorobenzene from contaminated sediment to fish. *Hydrobiologia* 463: 171-184.

Emans HJB, Van Den Plassche EJ, Canton JH. 1993. Validation of some extrapolation methods used for effect assessment. *Environ Toxicol Chem* 12: 2139-2154.

Erickson RJ, Stephan CE. 1988. Calculation of the final acute value for water quality criteria for aquatic organisms. EPA/600/3-88-018. US Environmental Protection Agency, Washington, DC.

Ericksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspectives* 111: 1361-1375.

EVS. 1999. A critique of the ANZECC and ARMCANZ (1999) water quality guidelines. Prepared for: Minerals Council of Australia and Kwinana Industries Council. Final Report, October 1999, EVS, Vancouver, BC.

Felsot AS. 2005. A critical analysis of the draft report, "Amendments to the Water Quality Control Plan for the Sacramento River and San Joaquin River Basins for the Control of Diazinon and Chlorpyrifos Runoff into the Lower San Joaquin River" (Karkoski et al. 2004) and supporting documents. Prepared for the Central Valley Regional Water Quality Control Board, Sacramento, CA.

Fisher DJ, Burton DT. 2003. Comparison of two US Environmental Protection Agency species sensitivity distribution methods for calculation ecological risk criteria. *Hum Ecol Risk Assess* 9: 675-690.

Fisk AT, Norstrom RF, Cymbalisty CD, Muir DCG. 1998. Dietary accumulation and depuration of hydrophobic organochlorines: bioaccumulation parameters and their

relationship with the octanol/water partition coefficient. *Environ Toxicol Chem* 17: 951-961.

Forbes VE, Cold A. 2005. Effects of the pyrethroid esfenvalerate on life-cycle traits and population dynamics of *Chironomus riparius*—importance of exposure scenario. *Environ Toxicol Chem* 24: 78-86.

Fox DR. 1999. Setting water quality guidelines--A statistician's perspective. *SETAC News* 19(3): 17-18.

Gentile H, Gentile SM, Hairston HG Jr, Sullivan BK. 1982. The use of life-tables for evaluating the chronic toxicity of pollutants to *Mysidopsis bahia*. *Hydrobiologia* 93: 179-187.

GESAMP. 1989. The evaluation of the hazards of harmful substances carried by ships: Revision of GESAMP reports and studies No. 17. IMO Reports and studies No. 35. Group of Experts on the Scientific Aspects of Marine Protection (United Nations).

GLEC. 2003. Draft Compilation of existing guidance for the development of site-specific water quality objectives in the state of California. Great Lakes Environmental Center, Columbus, OH.

Giesy JP, Solomon KR, Coates JR, Dixon KR, Giddings JF, Kenaga EE. 1999. Chlorpyrifos: ecological risk assessment in North American aquatic environments. *Rev Environ Contam toxicol* 160: 1-129.

Gobas FAPC, Muir DCG, Mackay D. 1988. Dynamics of dietary bioaccumulations and faecal elimination of hydrophobic chemicals in fish. *Chemosphere* 17: 943-962.

Government of British Columbia. 1995. Derivation of water quality criteria to protect aquatic life in British Columbia. Government of British Columbia, Ministry of Land, Air and Water Protection, Water Quality Branch, <http://wlapwww.gov.bc.ca/wat/wq/BCguidelines/derive.html#can>.

Grist EPM, Crane M, Jones C, Whitehouse P. 2003. Estimation of demographic toxicity through the double bootstrap. *Wat Res* 37: 618-626.

Grist EPM, Leung KMY, Sheeler JR, Crane M. 2002. Better bootstrap estimation of hazardous concentration thresholds for aquatic assemblages. *Environ Toxicol Chem* 21: 1515-1524.

Grothe DR, Kickson KL, Reed-Judkins DK, eds. 1996. Whole effluent toxicity testing: An evaluation of methods and prediction of receiving system impacts. SETAC Press, Pensacola, FL.

- Hansch C, Leo A. 1979. Substituent constants for correlation analyses in chemistry and biology. John Wiley & Sons, New York, NY.
- Hansch C, Leo A, Hoekman D. 1995. Exploring QSAR. Hydrophobic, electronic, and steric constants. American Chemical Society, Washington, DC.
- Hanson ML, Sanderson H, Solomon KR. 2003. Variation, replication, and power analysis of *Myriophyllum* spp. microcosm toxicity data. *Environ Toxicol Chem* 22: 1318-1329.
- Heckman L-H, Friberg N. 2005. Macroinvertebrate community response to pulse exposures with the insecticide lambda-cyhalothrin using in-stream mesocosms. *Environ Toxicol Chem* 24: 582-590.
- Hodson PV, Blunt BR, Borgmann U, Minns CK, McGraw S. 1983. Effect of fluctuating lead exposures on lead accumulation by rainbow trout (*Salmo gairdneri*). *Environ Toxicol Chem* 2: 225-238.
- Hoekstra JA, Van Ewijk. 1993. Alternatives for the no-observed-effect level. *Environ Toxicol Chem* 12: 187-194.
- Hose GC, Van Den Brink PJ. 2004. Confirming the species-sensitivity distribution concept for endosulfan using laboratory, mesocosm, and field data. *Arch Environ Contam Toxicol* 47: 511-520.
- Host GE, Regal RR, Stephan CE. 1995. Analyses of acute and chronic data for aquatic life. US. Environmental Protection Agency, Washington, DC.
- Howard PH. 1990. Handbook of environmental fate and exposure data for organic chemicals. Vol. II: Solvents. ISBN 0-87371-204-8, Lewis Publishers, Chelsea, MI.
- Howard PH. 1991. Handbook of environmental fate and exposure data for organic chemicals. Vol. III: Pesticides. ISBN 0-87371-328-1, Lewis Publishers, Chelsea, MI.
- Ingersoll CG, Winner RW. 1982. Effect on *Daphnia pulex* (de geer) of daily pulse exposures to copper or cadmium. *Environ Toxicol Chem* 1: 321-327.
- Irmer U, Markard C, Blondzik K, Gottschalk C, Kussatz C, Rechenberg B, Schudoma D. 1995. Quality targets for concentrations of hazardous substances in surface waters in Germany. *Ecotox Environ Safe* 32: 233-243.
- Jago RH, Newman MC. 1997. Bootstrap estimation of community NOEC values. *Ecotoxicol* 6: 293-306.
- Jaworska JS, Comber M, Auer C, Van Leeuwen CJ. 2003. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ Health Perspectives* 111: 1358-1360.

- KemI. 1989. Environmentally hazardous substances. List of examples and scientific documentation. Kemikalieinspektionen 10/89, 303 pp. (in Swedish).
- Kenaga EE. 1982. Predictability of chronic toxicity from acute toxicity of chemicals in fish and aquatic invertebrates. *Environ Toxicol Chem* 1: 347-358.
- Kloepper-Sams PJ, Owens JW. 1993. Environmental biomarkers as indicators of chemical exposure. *J Haz Mat* 35: 283-294.
- Könemann H. 1981. Fish toxicity tests with mixtures of more than two chemicals: a proposal for a quantitative approach and experimental results. *Toxicology* 19: 229-238.
- Kooijman SALM. 1987. A safety factor for LC₅₀ values allowing for differences in sensitivity among species. *Wat Res* 21:269-276.
- Kooijman SALM, Bedaux JJM. 1996a. Analysis of toxicity tests on fish growth. *Wat Res* 30: 1633-1644.
- Kooijman SALM, Bedaux JJM. 1996b. Analysis of toxicity tests on *Daphnia* survival and reproduction. *Wat Res* 30: 1711-1723.
- Kooijman SALM. 1993. *Dynamic Energy Budgets in biological systems. Theory and applications in ecotoxicology*. Cambridge University Press, Cambridge, UK.
- Kooijman SALM, Hanstveit AO, Nyholm N. 1996. No-effect concentrations in algal growth inhibition tests. *Wat Res* 30: 1625-1632.
- Kratzer CR, Zamora C, Knifong DL. 2002. Diazinon and chlorpyrifos loads in the San Joaquin River Basin, California. January and February 2000. United States Geological Survey, Water-Resources Investigations Report 02-4103.
- Kraufvelin P. 1999. Baltic hard bottom mesocosms unplugged: replicability, repeatability and ecological realism examined by non-parametric multivariate techniques. *J Exp Mar Biol Ecol* 240: 229-258.
- Kuivila KM, Barnett HD, Edmunds JL. 1999. Herbicide concentrations in the Sacramento-San Joaquin Delta, California. United States Geological Survey, Water-Resources Investigations Report 99-4018 B.
- La Point TW, Belanger SE, Crommentuijn T, Goodrich-Mahoney J, Kent RA, Mount DI, Spry DJ, Vigerstad T, Di toro DM, Keating FJ Jr, Reiley MC. 2003. Problem Formulation. In: *Reevaluation of the State of the Science for Water-Quality Criteria Development*. SETAC Press, Pensacola, FL. pp. 1-14.

- LAWA. 1997. Zielvorgaben zum Schutz oberirdischer Binnengewässer. Band 1. Länderarbeitsgemeinschaft Wasser. Kulturbuchverlag Berlin GmbH, Berlin.
- Lawton JH. 1994. What do species do in ecosystems? *Oikos* 71: 367-374.
- Lemly AD. 1985. Toxicology of selenium in a freshwater reservoir: implications for environmental hazard evaluation and safety. *Ecotox Environ Safety* 10: 314-348.
- Lepper P. 2000. Towards the derivation of quality standards for priority substances in the context of the Water Framework Directive. Final Report of the Study Contract No. B4-3040/2000/30673/MAR/E1. Fraunhofer-Institute Molecular biology and Applied Ecology, Munich.
- Lillebo HP, Shaner S, Carlson D, Richard N. 1988. Water quality criteria for selenium and other trace elements for protection of aquatic life and its uses in the San Joaquin Valley. In: Technical Committee Report: Regulation of agricultural drainage to the San Joaquin River. Appendix D. California State Water Resources Control Board, Sacramento, CA.
- LOGKOW. 1994. LOGKOW octanol-water partition coefficient program. Now called KowWin. Available at http://www.syrres.com/esc/est_soft.htm. Syracuse Research Corporation, New York, NY.
- Loonen H, Parsons JR, Govres HAJ. 1991. Dietary accumulation of PCDDs and PCDFs in guppies. *Chemosphere* 23: 1349-1357.
- Lydy M, Belden J, Wheelock C, Hammock B, Denton D. 2004. Challenges in regulating pesticide mixtures. *Ecol Soc* 9: 1 [online] URL: <http://www.ecologyandsociety.org/vol9/iss6/art1/>
- Maboeta MS, Reinecke SA, Reinecke AJ. 2003. Linking lysosomal biomarker and population responses in a field population of *Aporrectodea caliginosa* (Oligochaeta) exposed to the fungicide copper oxychloride. *Ecotox Environ Safe* 56: 411-418.
- Mackay D, Shiu W-Y, Ma K-C. 1992. Illustrated handbook of physical-chemical properties and environmental fate for organic chemical. Volume I. Monoaromatic hydrocarbons, chlorobenzenes, and PCBs. Lewis Publishers, Boca Raton, USA.
- Mackay D, Shiu W-Y, Ma K-C. 1993. Illustrated handbook of physical-chemical properties and environmental fate for organic chemical. Volume III. Volatile organic chemicals. Lewis Publishers, Boca Raton, USA.
- Mackay D, Shiu W-Y, Ma K-C. 1995. Illustrated handbook of physical-chemical properties and environmental fate for organic chemical. Volume IV. Oxygen, nitrogen, and sulfur containing compounds. Lewis Publishers, Boca Raton, USA.

Mackay D, Shiu W-Y, Ma K-C. 1997. Illustrated handbook of physical-chemical properties and environmental fate for organic chemical. Volume V. Pesticide chemicals. Lewis Publishers, Boca Raton, USA.

Mackay D, Shiu W-Y, Ma K-C. 1999. Illustrated handbook of physical-chemical properties and environmental fate for organic chemical. CRC-LLC netbase, CD-rom version.

Maltby L, Blake N, Brock TCM, Van Den Brink PJ. 2005. Insecticide species sensitivity distributions: importance of test species and relevance to aquatic ecosystems. *Environ Toxicol Chem* 24: 379-388.

Matthiessen P. 2000. Is endocrine disruption a significant ecological issue? *Ecotoxicol* 9: 21-24.

Mayer FL, Ellersieck MR, Krause GF, Sun K, Lee G, Buckler DR. 2002. Time-concentration-effect models in predicting chronic toxicity from acute toxicity data. In: *Risk Assessment with Time to Event Models*, Crane M, Newman MC, Chapman PF, Fenlon J, eds. Lewis Publishers, Boca Raton, FL, pp. 39-67.

Mayer FL, Krause GF, Buckler DR, Ellersieck MR, Lee G. 1994. Predicting chronic lethality of chemicals to fishes from acute toxicity test data: concepts and linear regression analysis. *Environ Toxicol Chem* 13: 671-678.

Menconi M, Beckman J. 1996. Hazard assessment of the insecticide methomyl to aquatic organisms in the San Joaquin river system. Admin. Rep. 96-6. California Department of Fish and Game, Environ. Serv. Div., Rancho Cordova, CA.

Mensink BJWG, Montforts M, Wijkhuizen-Maslankiewicz L, Tibosch H, Linders JBHJ. 1995. Manual for summarizing and evaluating the environmental aspects of pesticides. Report no. 67101022. National Institute of Public Health and Environmental Protection (RIVM), Bilthoven, The Netherlands.

MITI. 1992. Biodegradation and bioaccumulation data on existing data based on the CSCL Japan. Japan chemical industry, Ecology-toxicology & information center. ISBN 4-89074-101-1.

Moore DRJ, Caux PY. 1997. Estimating low toxic effects. *Environ Toxicol Chem* 16: 794-801.

Mount DR, Ankley GT, Brix KV, Clements WH, Dixon DG, Fairbrother A, Hickey CW, Lanno RP, Lee CM, Munns WR, Ringer RK, Staveley JP, Wood CM, Erickson RJ, Hodson PV. 2003. Effects assessment: Introduction. In: *Reevaluation of the State of the Science for Water-Quality Criteria Development*, Reiley MC, Stubblefield WA, Adams WJ, Di Toro DM, Hodson PV, Erickson RJ, Keating FJ Jr, eds., SETAC Press, Pensacola, FL.

Murray FJ, Smith FA, Nitschke KD, Humiston CG, Kociba RJ, Schwetz BA. 1979. Three generation reproduction study of rats given 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) in the diet. *Toxicol Appl Pharmacol* 50: 241-252.

Nabholz JV. 1991. Environmental hazard and risk assessment under the United States Toxic Substances Control Act. *Sci Total Environ* 109/110: 649-665.

Nabholz JV. 2003. Toxicity assessment, risk assessment, and risk management of chemicals under TSCA in USA. Office of Pollution Prevention and Toxics, United States Environmental Protection Agency, Washington, DC.

Newman MC, Crane M. 2002. Introduction to time to event methods. In: *Risk Assessment with Time to Event Models*, Crane M, Newman MC, Chapman PF, Fenlon J, eds. Lewis Publishers, Boca Raton, FL, pp. 1-6.

Newman MC, Ownby DR, Mézin LCA, Powell DC, Christensen TRL, Lerberg SB, Anderson B-A. 2000. Applying species-sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient numbers of species. *Ecotoxicol Environ Chem* 19: 508-515.

Newman MC, Ownby DR, Mézin LCA, Powell DC, Christensen TRL, Lerberg SB, Anderson B-A, Padma TV. 2002. Species sensitivity distributions in ecological risk assessment: distributional assumptions, alternate bootstrap techniques, and estimation of adequate number of species. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter II GW, Traas TP, eds., Lewis Publishers, CRC Press, New York, NY.

Nikunen E, Leinonen R, Kultamaa A. 1990. Environmental properties of chemicals. Ministry of the Environment, Research report 91, Finland.

Norstrom RJ, McKinnon AE, De Freitas ASW. 1976. A bioenergetics-based model for pollutant accumulations by fish. Simulation of PCB and methylmercury in Ottawa River yellow perch *Perca flavescens*. *J Fish Res Board Can* 33: 248-267.

North Carolina Department of Environment and Natural Resources. 2003. "Redbook" Surface Waters and Wetlands Standards. North Carolina DENR, Division of Water Quality, NC Administrative Code 15A NCAC 02 B .0100 & .0200.

Novartis Crop Protection. 1997. An ecological risk assessment of diazinon in the Sacramento and San Joaquin River basins. Novartis Crop Protection, Environmental and Public Affairs Department Technical Report 11/97, Greensboro, NC.

OECD. 1981. *Guidelines for testing chemicals*. Organisation for Economic Co-operation and Development, Paris.

- OECD. 1992. Fish, early-life stage toxicity test. OECD guidelines for testing of chemicals. Organization for Economic Co-operation and Development, Paris.
- OECD. 1995. Guidance Document for Aquatic Effects Assessment. Organization for Economic Co-operation and Development, Paris.
- Okkerman PC, Van Den Plassche EJ, Emans HJB, Canton JH. 1993. Validation of some extrapolation methods with toxicity data derived from multiple species experiments. *Ecotox Environ Safe* 25: 341-359.
- Okkerman PC, Van Den Plassche EJ, Slooff W, Van Leeuwen CJ, Canton JH. 1991. Ecotoxicological effects assessment: a comparison of several extrapolation procedures. *Ecotox Environ Safe* 21: 182-193.
- Olsen T, Elerbeck L, Fisher T, Callaghan A, Crane M. 2001. Variability in acetylcholinesterase and glutathione S-transferase activities in *Chironomus riparius* Meigen deployed in situ at uncontaminated field sites. *Environ Toxicol Chem* 20: 1725-1732.
- Parkhurst DF. 1998. Arithmetic versus geometric means for environmental concentration data. *Environ Sci Technol* 32: 92A-98A.
- Pawlisz AV, Busnarda J, McLaughlin A, Caux P-Y, Kent RA. 1998. Canadian water quality guidelines for deltamethrin. *Environ Toxic Water* 13: 175-210.
- Persoone G, Janssen CR. 1994. Field validation of predictions based on laboratory toxicity tests. In: *Freshwater Field Tests for Hazard Assessment of Chemicals*. Hill IR, Heimbach F, Leeuwangh P, Matthiessen P, eds. CRC Press, Boca Raton, FL, pp. 379-397.
- Péry ARR, Flammarion P, Vollat B, Bedaux JJM, Kooijman SALM, Garric J. 2002. Using a biology-based model (DEBtox) to analyze bioassays in ecotoxicology: opportunities and recommendations. *Environ Toxicol Chem* 21: 459-465.
- Plackett RL, Hewlett PS. 1952. Quantal responses to mixtures of poisons. *J Royal Stat Soc B* 14: 141-163.
- Posthuma L, Traas TP, Suter III GW. 2002a. General introduction to species sensitivity distributions. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter III GW, Traas TP, eds., Lewis Publishers, CRC Press, Boca Raton, FL, pp. 3-10.
- Posthuma L, Traas TP, De Zwart D, Suter GW II. 2002b. Conceptual and technical outlook on species sensitivity distributions. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter III GW, Traas TP, eds., Lewis Publishers, CRC Press, Boca Raton, FL, pp. 475-508.

- Pusey BJ, Arthington AH, McClean J. 1994. The effects of a pulsed application of chlorpyrifos on macroinvertebrate communities in an outdoor artificial stream system. *Ecotox Environ Safe* 27: 221-250.
- Qiao P, Gobas FAPC, Farrell AP. 2000. Relative contributions of aqueous and dietary uptake of hydrophobic chemicals to the body burden in juvenile rainbow trout. *Environ Contam Toxicol* 39: 369-377.
- Ramos EU, Vaes WHJ, Verhaar HJM, Hermens JLM. 1998. Quantitative structure-activity relationships for the aquatic toxicity of polar and nonpolar narcotic pollutants. *J Chem Inf Comput Sci* 38: 845-852.
- Reiley MC, Stubblefield WA, Adams WJ, Di Toro DM, Hodson PV, Erickson RJ, Keating FJ Jr. 2003. *Reevaluation of the State of the Science for Water-Quality Criteria Development*. SETAC Press, Pensacola, FL.
- Reynaldi S, Liess M. 2005. Influence of duration of exposure to the pyrethroid fenvalerate on sublethal responses and recovery of *Daphnia magna* Straus. *Environ Toxicol Chem* 24: 1160-1164.
- Rider CV, LeBlanc GA. 2005. An integrated addition and interaction model for assessing toxicity of chemical mixtures. *Toxicol Sci* 87: 520-528.
- 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.
- RIVM. 2001. Guidance document on deriving environmental risk limits in The Netherlands. Report no. 601501 012. Traas TP, ed. National Institute of Public Health and the Environment, Bilthoven, The Netherlands.
- RIVM. 2004. ETX 2.0. Normal distribution based hazardous concentration and fraction affected. Designed by Van Vlaardingen P, Traas T, Aldenberg T, Wintersen A. National Institute of Public Health and the Environment, Bilthoven, The Netherlands.
- Romijn CAFM, Luttik R, Van De Meent D, Slooff W, Canton JH. 1993. Presentation of a general algorithm to include effect assessment on secondary poisoning in the derivation of environmental quality criteria. *Ecotox Environ Safe* 26: 61-85.
- Roth L. 1993. Wassergefährdende Stoffe. Ecomed verlag Gmbh, Landsberg/Lech.
- Roux DJ, Jooste SHJ, MacKay HM. 1996. Substance-specific water quality criteria for the protection of South African freshwater ecosystems: methods for derivation and initial results for some inorganic toxic substances. *S African J Sci* 92: 198-206.

- Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA. 1997. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ Toxicol Chem* 16: 948-967.
- Samsøe-Petersen L, Pedersen F (eds.). 1995. Water quality criteria for selected priority substances, Working Report, TI 44. Water Quality Institute, Danish Environmental Protection Agency, Copenhagen, Denmark.
- Sanderson H. 2002. Pesticide studies—replication of micro/mesocosm studies. *Environ Sci Pollut R* 6: 429-435.
- Schwarzenbach RP, Gschwend PM, Imboden DM. 1993. Environmental Organic Chemistry. John Wiley & Sons, Inc., NY, USA
- Schulz R, Liess M. 2001. Toxicity of fenvalerate to caddisfly larvae: chronic effects of 1- vs. 10-h pulse-exposure with constant exposures. *Chemosphere* 41: 1511-1517.
- Segner H. 2005. Developmental, reproductive, and demographic alterations in aquatic wildlife: establishing causality between exposure to endocrine-active compounds (EACs) and effects. *Acta hydrochim hydrobiol* 33: 17-26.
- SETAC-Europe. 1992. Guidance document on testing procedures for pesticides in freshwater mesocosms. From a meeting of experts on guidelines for static field mesocosm tests, held at Monks Wood Experimental Station, Abbots Ripton, Huntingdon, UK, 3-4 July 1991.
- Shao Q. 2000. Estimation for hazardous concentrations based on NOEC toxicity data: an alternative approach. *Environmetrics* 11: 583-595.
- Siepmann S, Jones MR. 1998. Hazard assessment of the insecticide carbaryl to aquatic organisms in the Sacramento-San Joaquin river system. Admin. Rep. 98-1. California Department of Fish and Game, Office of Spill Prevention and Response, Rancho Cordova, CA.
- Siepmann S, Finlayson B. 2000. Water quality criteria for diazinon and chlorpyrifos. California Department of Fish and Game.
- Slooff W. 1992. RIVM guidance document. Ecotoxicological effect assessment: deriving maximum tolerable concentrations (MTC) from single-species toxicity data. Report 719102 018, RIVM Bilthoven, The Netherlands.
- Solomon KR, Takacs P. 2002. Probabilistic risk assessment using species sensitivity distributions. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter GW II, Traas TP, eds. Lewis Publishers, New York, NY, pp. 285-314.

Speijers GJA, Franken MAM, Van Leeuwen FXR, Van Egmond HP, Boot R, Loeber JG. 1986. Subchronic oral toxicity study of patulin in the rat. Report 618314001. RIVM Bilthoven, The Netherlands.

Spromberg JA, Birge WJ. 2005a. Modeling the effects of chronic toxicity on fish populations: the influence of life-history strategies. *Environ Toxicol Chem* 24: 1532-1540.

Spromberg JA, Birge, WJ. 2005b. Population survivorship index for fish and amphibians: application to criterion development and risk assessment. *Environ Toxicol Chem* 24: 1541-1547.

Stephan CE. 1985. Are the "Guidelines for deriving numerical national water quality criteria for the protection of aquatic life and its uses" based on sound judgments? In: *Aquatic Toxicology and Hazard Assessment: Seventh Symposium, ASTM STP 854*, Cardwell RD, Purdy R, Bahner RC, eds., American Society for Testing and Materials, Philadelphia, PA, pp. 515-526.

Stephan CE, Rogers JW. 1985. Advantages of using regression analysis to calculate results of chronic toxicity tests. In: *Aquatic Toxicology and Hazard Assessment: Eighth Symposium. ASTM STP 891*, Bahner RC, Hansen DJ, eds., American Society for Testing and Materials, Philadelphia, PA, pp. 328-338.

Sun K, Krause GJ, Mayer FL Jr, Ellersieck MR, Basu AP. 1995. Predicting chronic lethality of chemicals to fishes from acute toxicity test data: theory of accelerated life testing. *Environ Toxicol Chem* 14: 1745-1752.

Suter GW II, Rosen AE, Linder E, Parkhurst DF. 1987. Endpoints for responses of fish to chronic toxic exposures. *Environ Toxicol Chem* 6: 793-809.

Teh SJ, Deng DF, Werner I, Teh FC, Hung SSO. 2005. Sublethal toxicity of orchard stormwater runoff in Sacramento splittail (*Pogonichthys macrolepidotus*) larvae. *Mar Environ Res* 59: 203-216.

Thomann RV, Connolly JP. 1984. Model of PCB in the Lake Michigan lake trout food chain. *Environ Sci Technol* 18: 65-71.

Traas TP, Van De Meent D, Posthuma L, Hamers T, Kater BJ, De Zwart D, Aldenberg T. 2002. The potentially affected fraction as a measure of ecological risk. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter GW II, Traas TP, eds. Lewis Publishers, New York, NY, pp. 315-344.

Traas TP, Van Wezel AP, Hermens JLM, Zorn M, Van Hattum AGM, Van Leeuwen CJ. 2004. Environmental quality criteria for organic chemicals predicted from internal effect concentrations and a food web model. *Environ Toxicol Chem* 23: 2518-2527.

Treibskorn R, Adam S, Behrens A, Beier S, Böhmer J, Braunbeck T, Casper H, Dietze U, Gernhöfer M, HOnnen W, Köhler H-R, Körner W, Konradt J, Lehmann R, Luckenbach T, Oberemm A, Schwaiger J, Segner H, Strmac M, Schüürmann G, Siligato S, Traunspurger W. Establishing causality between pollution an defects at different levels of biological organization: the VALIMAR project. *Hum Ecol Risk Assess* 9: 171-194.

USEPA. 1978a. Federal Register, 43: 21506-21518. May 18. US Environmental Protection Agency, Washington, DC.

USEPA. 1978b. Federal Register, 43: 29028. July 5. US Environmental Protection Agency, Washington, DC.

USEPA. 1984a. Guidelines for deriving numerical aquatic site-specific water quality criteria by modifying national criteria. EPA-600/3-84-099. US Environmental Protection Agency, Washington, DC.

USEPA. 1984b. Estimating “concern levels” for concentrations of chemical substances in the environment. Environmental Effects Branch, Health and Environmental Review Division (TS-796), Office of Toxic Substances, US Environmental Protection Agency, Washington, DC 20460-0001.

USEPA. 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. PB-85-227049. US Environmental Protection Agency, National Technical Information Service, Springfield, VA, USA.

USEPA. 1986. Guidelines for deriving ambient aquatic life advisory concentrations. EPA/822/R86/100. US Environmental Protection Agency, Washington, DC.

USEPA. 1987. 40 CFR Part 797. Environmental Effects Testing Guidelines. US Environmental Protection Agency, Washington, DC.

USEPA. 1991. Technical Support Document for Water Quality-based Toxics Control. EPA/505/2-90-001. US Environmental Protection Agency, Washington, DC.

USEPA. 1993. Federal Register, 40 CFR Part 158.490.

USEPA. 2000. Ambient Aquatic Life Water Quality Criteria for Dissolved Oxygen (Saltwater): Cape Cod to Cape Hatteras. EPA-822-R-00-012. US Environmental Protection Agency, Washington, DC.

USEPA. 2002a. Draft report on summary of proposed revisions to the aquatic life criteria guidelines. US Environmental Protection Agency, Washington, DC.

USEPA. 2002b. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organism, 4th edition. EPA-821-R-02-013. US Environmental Protection Agency, Washington, DC.

USEPA. 2003a. Water quality guidance for the Great Lakes system. Federal Register, 40 CFR Part 132. US Environmental Protection Agency, Washington, DC.

USEPA. 2003b. Ambient aquatic life water quality criteria for tributyltin (TBT) – Final. EPA 822-R-03-031. US Environmental Protection Agency, Washington, DC.

USEPA 2003c. Acute-to-chronic estimation (ACE v 2.0) with time-concentration-effect models, User manual and software. EPA/600/R-03/107. US Environmental Protection Agency, Washington, DC.

USEPA. 2003d. Interspecies correlation estimations (ICE) for acute toxicity to aquatic organisms and wildlife. II. User manual and software. EPA/600/R-03/106. US Environmental Protection Agency, Washington, DC.

USEPA. 2005. Science Advisory Board Consultation Document, Proposed Revisions to Aquatic Life Guidelines, Water-based Criteria, Water-based Criteria Subcommittee, US Environmental Protection Agency, Washington, DC.

USGS. 1998. Pesticides in surface and ground water of the United States: summary of the results of the National Water Quality Assessment Program (NAWQA). US Geological Survey, Washington, DC.

USGS. 2005a. Water Resource Data, California Water Year 2004, Volume 4, Northern Central Valley Basins and the Great Basin from Honey Lake Basin to Oregon State Line. US Geological Survey, Sacramento, CA.

USGS. 2005b. Water Resources Data, California Water Year 2004, Volume 3, Southern Central Valley Basins and the Great Basin from Walker River to Truckee River. US Geological Survey, Sacramento, CA.

Vaal M, Van Der Wal JT, Hermens J, Hoekstra J. 1997a. Pattern analysis of the variation in the sensitivity of aquatic species to toxicants. *Chemosphere* 35: 1291-1309.

Vaal M, Van Der Wal JT, Hoekstra J, Hermens J. 1997b. Variation in the sensitivity of aquatic species in relation to the classification of environmental pollutants. *Chemosphere* 35: 1311-1327.

Van De Meent D, Aldenberg T, Canton JH, Van Gesteel CAM, Slooff W. 1990. Desire for levels, background study for the policy document "Setting environmental quality standards for water and soil." National Institute of Public Health and the Environment, Bilthoven, The Netherlands.

Van Den Brink PJ, Roelsma J, Van Nes EH, Scheffer M, Brock TCM. 2002. PERPEST model, a case-based reasoning approach to predict ecological risks of pesticides. *Environ Toxicol Chem* 21: 2500-2506.

- Van Der Hoeven N. 2001. Estimating the 5-percentile of the species sensitivity distributions without any assumptions about the distribution. *Ecotoxicol* 10: 25-34.
- Van Der Hoeven N, Noppert F, Leopold A. 1997. How to measure no effect. Part I: Towards a new measure of chronic toxicity in ecotoxicology. Introduction and workshop results. *Environmetrics* 8: 241-248.
- Van Der Oost R, Beyer J, Vermeulen NPE. 2003. Fish bioaccumulation and biomarkers in environmental risk assessment: a review. *Environ Toxicol Pharm* 13: 57-149.
- Van Leeuwen CJ, Van Der Zandt PTJ, Aldenberg T, Verhaar HJM, Hermens JLM. 1992. Application of QSARs, extrapolation and equilibrium partitioning in aquatic effects assessment. I. Narcotic industrial pollutants. *Environ Toxicol Chem* 11: 267-282.
- Van Straalen NM, Denneman CAJ. 1989. Ecotoxicological evaluation of soil quality criteria. *Ecotox Environ Safe* 18: 241-251.
- Van Straalen NM, Van Leeuwen CJ. 2002. European history of species sensitivity distributions. In: *Species Sensitivity Distributions in Ecotoxicology*, Posthuma L, Suter II GW, Traas TP, eds., Lewis Publishers, CRC Press, Boca Raton, FL, pp. 19-34.
- Van Wijngaarden RPA, Brock TCM, Van Den Brink PJ. 2005. Threshold levels for effects of insecticides in freshwater ecosystems: a review. *Ecotoxicology* 14: 355-380.
- Verhaar HJM, Van Leeuwen CJ, Hermens JLM. 1992. Classifying environmental pollutants. 1: Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* 25: 471-491.
- Verscheuren K. 1983. *Handbook of environmental data on organic chemicals*, 2nd edition, Van Nostrand Reinhold Co., New York NY.
- Verscheuren K. 2001. *Handbook of environmental data on organic chemicals*, 4th edition, CD-ROM, Wiley Interscience, New York, NY.
- Versteeg DJ, Belanger SE, Carr GJ. 1999. Understanding single-species and model ecosystem sensitivity: data-based comparison. *Environ Toxicol Chem* 18: 1329-1346.
- VROM. 1994. Environmental quality objectives in The Netherlands. Ministry of Housing, Spatial Planning and Environment, The Hague, The Netherlands.
- Wagner C, Løkke H. 1991. Estimation of ecotoxicological protection levels from NOEC toxicity data. *Wat Res* 25: 1237-1242.
- Warmer H, Van Dokkum R. 2002. Water pollution control in the Netherlands, Policy and Practice. RIZA report 2002.009.

Warne MSJ, Hawker DW. 1995. The number of components in a mixture determines whether synergistic and antagonistic or additive toxicity predominate: the funnel hypothesis. *Ecotox Environ Safety* 31: 23-28.

Werner I, Deanovic LA, Connor V, de Vlaming V, Bailey HC, Hinton DE. 2000. Insecticide-caused toxicity to *Ceriodaphnia dubia* (cladocera) in the Sacramento-San Joaquin River Delta, California, USA. *Environ Toxicol Chem*, 19: 215-227.

Wheeler JR, Grist EPM, Leung KMY, Morritt D, Crane M. 2002. Species sensitivity distributions: data and model choices. *Marine Pollut Bull* 45: 192-202.

Whitehouse P, Crane M, Grist E, O'Hagan A, Sorokin N. 2004. Derivation and expression of water quality standards; opportunities and constraints in adopting risk-based approaches in EQS setting. R&D technical Report P2-157/TR. Environment Agency, Rio House, Almondsbury, Bristol.

Wu J, Laird DA. 2004. Interactions of chlorpyrifos with colloidal materials in aqueous systems. *J Environ Qual* 33: 1765-1770.

Zabel TF, Cole S. 1999. The derivation of environmental quality standards for the protection of aquatic life in the UK. *J CIWEM* 13: 436-440.

Zischke JA, Arthur JW, Hermanutz RO, Hedtke SF, Helgen JC. 1985. Effects of pentachlorophenol on invertebrates and fish in outdoor experimental channels. *Aquat Toxicol* 7: 37-58.