TASK ORDER 68HE0C18F0787 UNDER
CONTRACT EP-C-17-017

EXTERNAL PEER REVIEW OF CHRONIC TOXICITY OF ALUMINUM TO THE CLADOCERAN, *CERIODAPHNIA DUBIA*: EXPANSION OF THE EMPIRICAL DATABASE FOR BIOAVAILABILITY MODELING

FINAL PEER REVIEW SUMMARY REPORT

July 31, 2018

Submitted to:
U.S. Environmental Protection Agency
Office of Water, Office of Science and Technology
Health and Ecological Criteria Division
1200 Pennsylvania Avenue, NW
Washington, DC 20460
Attn: Diana Eignor
Eignor.Diana@epa.gov

Submitted by:
Eastern Research Group, Inc.
110 Hartwell Avenue
Lexington, MA 02421

ERG
www.erg.com

CONTENTS

1.0 INTRODUCTION

This report documents the results of an independent letter peer review of a toxicity report entitled *Chronic Toxicity of Aluminum to the Cladoceran, Ceriodaphnia dubia: Expansion of the empirical database for bioavailability modeling*, developed by Oregon State University. The peer review was organized for the U.S. Environmental Protection Agency (EPA), Office of Water (OW).

Eastern Research Group, Inc. (ERG), a contractor to EPA, organized this external peer review and developed this report. Section 2 presents, for each charge question, the individual reviewer comments and a summary of those comments. Section 3 provides additional reviewer comments or recommendations, and Section 4 presents new information (e.g., references) provided by reviewers. Appendix A provides EPA's charge to reviewers and Appendix B presents the complete set of comments submitted by each reviewer.

Background

EPA establishes national recommended Ambient Water Quality Criteria (AWQC) under the Clean Water Act (CWA). Section 304(a)(1) aquatic life criteria serve as recommendations to states and tribes by defining ambient water concentrations that will protect against unacceptable adverse ecological effects to aquatic life from exposure to pollutants in water. Aquatic life criteria address the CWA goals of providing for protection and propagation of fish and shellfish. Once EPA publishes final §304(a) recommended water quality criteria, states and authorized tribes may adopt these criteria into their water quality standards to protect designated uses of water bodies. As required by the CWA, EPA periodically reviews and revises §304(a) AWQC to ensure they are consistent with the latest scientific information. In support of this mission, EPA is working to update water quality criteria to protect aquatic life from aluminum in freshwater environments.

Oregon State University conducted invertebrate toxicity tests for aluminum that may be relevant to development of the model used to determine aquatic life criteria for aluminum. EPA charged ERG with organizing an independent focused, objective evaluation of these invertebrate toxicity tests, which were unpublished at the time the review was conducted.

Peer Reviewers

ERG identified, screened, and selected the following five experts who met technical selection criteria provided by EPA and had no conflict of interest in performing this review:

- **David Buchwalter, Ph.D.:** Associate Professor, Department of Biological Sciences, North Carolina State University.
- **Valery E. Forbes, Ph.D.:** Dean of the College of Biological Sciences, University of Minnesota.
- **William L. Goodfellow, M.S.:** Principal Scientist & Practice Director, Exponent.
- **Richard S. Grippo, Ph.D.:** Emeritus Professor of Environmental Biology, Arkansas State University.
- **Tham C. Hoang, Ph.D.:** Assistant Professor, Loyola University.

ERG provided reviewers with instructions, the invertebrate toxicity report, and the charge to reviewers (Appendix A of this report) prepared by EPA. Reviewers worked individually to develop written comments in response to the charge questions. After receiving reviewer comments, ERG summarized reviewers' responses to each charge question, noting areas of agreement and disagreement, where relevant (see Section 2).

2.0 SUMMARY OF REVIEWER COMMENTS ORGANIZED BY CHARGE QUESTION

This section summarizes reviewer comments by charge question. Each summary is followed by a table presenting individual reviewer responses to that charge question (see Appendix B for the complete set of reviewer comments).

2.1 Were an adequate number of concentrations tested to fully-characterize concentration-response and determine an accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?

All five reviewers responded that an adequate number of concentrations were tested. Four specifically noted that the tests followed standard EPA methodology, which calls for five concentrations and a control. One reviewer said that, except for one test, all could be used to estimate reproductive effects. Another reviewer pointed out that lethal effects could not be calculated. A third reviewer said that a regression model can be used to develop a chronic effect concentration. A fourth reviewer suggested adding a section in the report for protocol deviations and analytical issues.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. The test was conducted following standard US EPA chronic testing methodology according to US EPA (2002). This reference is not provided in the reference list (it should be), but presumably refers to EPA-821-R-02-013. According to this guidance, a minimum of 5 test concentrations and a control should be used in a definitive test. As each test in this study included 5 exposure concentrations and a dilution water control (p. 2-2), it is judged to be adequate for the test purpose. The range of concentrations chosen was also deemed adequate to achieve estimates of the desired effect levels for reproduction (10, 20, and 50% effect; Table 3-13). With the exception of one test in which effects on survival occurred, all test concentrations could be used to estimate reproductive effects.	
Reviewer 2	A total of nine different tests were conducted under different pH, hardness and DOC conditions. Five total Al concentrations plus controls were generally used in the various tests. This number of concentrations is generally considered adequate.	
Reviewer 3	Yes, 5 concentrations of Al and a negative control were used for each test. This design appeared to follow the EPA guidelines for toxicology testing with freshwater organisms. The concentrations used were low that did not result in complete mortality at the highest concentration of each test. Therefore, lethal effect concentrations (LCs) could not be calculated.	

Reviewer 4	Response: In my opinion, an adequate number of concentrations were tested to allow full characterization of the concentration response and allow determination of a scientifically-defensible chronic effect concentration. Rationale: This research project evaluated the effects of multiple water quality variables on the toxicity of Aluminum (Al) to the cladoceran Ceriodaphnia dubia. The goal of the study was to increase the range of water quality variables under which a reasonable prediction of invertebrate toxicity could be performed under a given set of water quality variables. The test followed standard USEPA methodology (US EPA 2002). The methods included in this manual are referenced in Table IA, 40 CFR Part 136 regulations and, therefore, constitute approved methods for acute toxicity tests. These methods were used in the present study with modifications to address different water types and pH levels. For example, concentrations were based on previous studies shown to cause a negative impact on C. dubia survival and reproduction. The standard EPA protocol calls for five test concentrations and a control and this was mostly followed in the present study. For one test (Test #: Al 1185 CDC; p. 12, Appendices (page 1, Appendix B) six concentrations of Al were used, plus a treatment labeled "non pH"). This was apparently a confirmatory test for comparison to results obtained at the Chilean Mining and Metallurgy Research Center (CIMM; Santiago, Chile) and Universidad Adolfo Ibañez (UAI; Santiago, Chile) and reported in Gensemer et al. (2018) as indicated on p. 29, paragraph 3. Five concentrations is the number usually followed by most toxicity testing laboratories including those administered by the US EPA (such as the EPA facility in Cincinnati, OH with which I am familiar). This allows the present study to be compared to the results of other laboratories and have such results be incorporated into the statistical model developed by the authors. This regression model can be used to develop a scientifically defensible chronic effect concentration such as the EC20 (dose which causes a 20% change from control response of the test organisms).	
Reviewer 5	The study was performed following the agreed to protocol. However, one study used a 45% bisection of the test concentrations rather than the protocol specified 50% bisection. While I do not believe that this is a fatal flaw in the analysis, I believe that it does warrant a section in the report for protocol deviations (rather than as only noted in Section 2.5 [page 2-2]).	

	This would also provide an opportunity to offer the analytical issues (as identified in Section 3.2 [page 3-4]). I also believe the authors should assess whether the analytical anomalies bias the results high, low, or neutral. This is very helpful in the use of these results. In my overall opinion, all test concentrations were sufficiently characterized to provide a meaningful and accurate description of the test results and the chronic toxicity of aluminum.	

2.2 Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?

All five reviewers responded that each test concentration had 10 replicates, which is a sufficient number. Two specifically noted that this is consistent with standard EPA methodology. One also noted that 10 replicates of each concentration allows for comparison of the results with previous and likely future results from other laboratories. One reviewer said that the report did not clearly indicate the number of organisms used per replicate chamber, but another stated that each contained one cladoceran. One reviewer noted that conditions were carefully controlled to reduce variability in organism response.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. There were 10 replicate chambers for each exposure concentration and control, each containing one cladoceran. This is consistent with US EPA guidance (EPA-821-R-02-013).	
Reviewer 2	Yes. Ten replicates per treatment is adequate.	
Reviewer 3	Yes, 10 replicates per treatment were usually used for this type of test. The report (section 2.9) did not clearly say the number of organisms used per replicate chamber.	
Reviewer 4	Response: Yes, the number of replicates (10 per Al treatment concentration and 10 in the non-treated control) was sufficient to allow sufficient statistical rigor for a C. dubia chronic toxicity evaluation under the stated test conditions. Rationale: Ten replicates of each toxicant concentration and the control is the number recommended by the US EPA (2002). This number of replicates is used by most toxicity testing laboratories, allowing comparison of the results of the present study with previous (and likely future) results from other laboratories.	

Reviewer	Comments	EPA Response to Comments
	Statistical dogma suggests that ≈30 replicates is the optimal number when evaluating biological data. However, in this (and most other toxicity testing laboratories) the test conditions were carefully controlled, using 1) moderately hard diluent water prepared in-house (please see question 7 below), 2) environmental chambers controlled for pH and light regimen, and 3) neonates that were all less than 24 hours old. All of these conditions will serve to reduce variability in organism response to exposure, which will support rigorous statistical testing using 10 replicates.	
Reviewer 5	The number of replicates (10) and test concentrations (minimally 5 plus a control) were standard with in ecotoxicity testing with *Ceriodaphnia dubia*. These are acceptable.	

2.3 Was the source, maintenance, and husbandry of test organisms well described?

Three reviewers replied that the source of the organisms was well described, but they did not think that the maintenance of the test organisms was. Two of the three did not think husbandry of the test organisms was well described, while the third said it appeared to be adequate. The fourth reviewer did think that the description of the test animals was adequately presented. The fifth reviewer said the section was too brief and lacked details of animal performance for the reference toxicant tests.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Partially. The source of the organisms was well described. They were obtained from in-house cultures that had been maintained for over 10 years and originally obtained from Aquatic BioSystems (Fort Collins, CO, USA) (p. 2-1). Maintenance and husbandry of the test organisms were not described in the report, although the authors did indicate that they conducted monthly tests with a reference toxicant (NaCl) to confirm that the organisms were in good condition (p. 2-1).	
Reviewer 2	Not particularly. This section was remarkably brief and lacking details of animal performance for the reference toxicant tests. The reporting of volumes of algal suspensions used for feeding are not useful unless cell densities are reported.	
Reviewer 3	Organisms were originally from Aquatic Biosystems and cultured at OSU for more than 10 years. Organisms were cultured in moderately hard water. Other environmental conditions and maintenance procedures were not described, such as	

Reviewer	Comments	EPA Response to Comments
	temperature, photoperiod (light:dark hours), food, feeding rates, biomass/water volume, water change, etc.	
Reviewer 4	Response: No, an adequate description of the source, maintenance, and husbandry of the C. daphnia test organism was not provided. Rationale: In the report, section 2.3.2 SOURCE, the authors state that the <24 hour old neonates were obtained from in-house cultures which have been maintained successfully at the Aquatic Toxicology laboratory at Oregon State University (Corvallis) for >10 years. In Appendix A, section 2.2 and 2.3, feeding diet and feeding regimen during toxicity testing were described. However, nowhere that I could find in the report was it explicitly stated that the test organisms were cultured and maintained under these same conditions. I believe this is an oversight in reporting, not a failure of procedure, and this oversight can be readily remedied by the authors by providing the missing information. Husbandry of the test organisms during culture and testing as described appeared to be adequate.	
Reviewer 5	The description of the test animals was adequately presented in the report. Reference toxicant testing was regularly performed as part of the quality assurance program.	

2.4 Were the control's survival rates acceptable?

All five reviewers responded that the control survival rates were acceptable. Two specifically said that the control treatments met EPA criteria (>80% survival and >60% surviving females having 15 or more neonates). One reviewer pointed out that the test with the poor control reproductive output should not be used.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. The authors report that in all tests, control acceptability criteria (> 80 % survival and > 60% surviving females having 15 or more neonates) were met (p. 3-14). These fulfill the criteria for test acceptability outlined in EPA-821-R-02-013.	
Reviewer 2	The average number of neonates/female in controls ranged from 22 to 37 with 42.5 reported from a "concurrent control".	

Reviewer	Comments	EPA Response to Comments
	The test with the poor control reproductive output (Al1199 CDC) should not be used.	
Reviewer 3	The survival of the control organisms of each test was 100%. This meets the test acceptability criteria of the test method (80-100%).	
Reviewer 4	Response: Yes, it appears that the survival rate of C. dubia used in the control (no aluminum) treatments met the accepted survival rate for this type of toxicity testing. Rationale: The standard methodology as developed by the US EPA (1982) calls for at least 80% survival of the control test organisms for the test to be considered valid. On p. 29, paragraph 2, the authors state that, in all tests, control acceptability criteria (> 80 % survival and > 60% surviving females having 15 or more neonates) were met. Table 3-12 (p. 30 of report) and Appendix D Raw Data both indicate that control survival was uniformly 100%, clearly meeting the EPA (2002) control standard for test acceptability.	
Reviewer 5	Control survival rates were acceptable.	

2.5 Were test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?

All five reviewers responded that the test organisms were appropriately acclimated to different hardness levels. One reviewer was impressed with the acclimation process and said that the researchers should be commended. Two reviewers pointed out that the report did not indicate whether the test organisms were acclimated for different pH or dissolved organic carbon (DOC).

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes, as far as hardness is concerned. Organisms cultured under standard conditions (100 mg/L as CaCO3) were used in the moderately hard water tests (120 mg/L as $CaCO_3$). Organisms were acclimated to the soft (60 mg/L as $CaCO_3$) and hard water (250 and 400 mg/L as $CaCO_3$) conditions for multiple generations (i.e., over two months), and survival and reproduction were reported to be excellent (p. 2-2). As far as indicated in the	

Reviewer	Comments	EPA Response to Comments
	report, there was no acclimation for different pH (tested range: 6.3 – 8.8; standard culture at 7.8-8.0) or DOC (tested range: 1-14 mg/L; standard culture unknown) conditions.	
Reviewer 2	The report only mentions acclimation of cultures to different hardness levels, but not pH and DOC or buffers.	
Reviewer 3	Yes, the acclimation of the organisms to the hardness of test waters (250 and 400 mg/L as CaCO3) for multiple generations and over more than 2 months should be adequate.	
Reviewer 4	Response: It would appear that the C. dubia used in these toxicity tests were appropriately acclimated for the stated test type and described test water conditions at the time the chronic toxicity testing was performed Rationale: The C. dubia used for the present study were reported (Section 2.3.4 ACCLIMATION p. 2-2;) as being cultured at the Ohio State University AquaTox laboratory, in a "moderately hard" reconstituted water that was prepared as detailed in standard USEPA methods (USEPA 2002). This diluent was reported to have a measured hardness of 100 mg/L as CaCO3 and pH of 7.8 – 8.0, p. 2-2). All acclimated cultures for all of the toxicity tests were successfully maintained in their respective laboratory water for multiple generations (2+ months). Organism survival and reproduction were reported as excellent and organism health was maintained over the period of acclimation. Note: In section 2.3.4, ACCLIMATION is erroneously labeled, in section 2.3.2 SOURCE, as section 2.4.3).	
Reviewer 5	I was quite impressed with the acclimation process used in this study. In many instances, researchers do not go to the length of details used for the acclimation protocol performed in this study. The researches should be commended on this practice.	

2.6 Were test endpoints and data acceptability criteria well defined and explained?

Four reviewers responded that the test endpoints were well defined and explained. The fifth reviewer did not comment on the test endpoints. Two reviewers believed the data acceptability criteria to be well defined and explained; whereas, two did not. The fifth reviewer did not comment on data acceptability. One reviewer said

that the software packages used to assess data have built-in tests for homogeneity of variance; however, the reviewer recommended explicitly discussing control performance. Another reviewer recommended that there be a separate section in the report to define the measured endpoints of the test. A third reviewer said that it would be useful to know the conditions in which the organisms were observed and determined to be alive or dead. A fourth reviewer suggested further evaluating one of the treatments that seemed to have inappropriate results.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. Test endpoints included NOEC and LOEC for survival and reproduction (if data met assumptions of normality and homogeneity), as well as effect concentrations (i.e., LC10/LC20/ LC50 for survival and ECx10/EC20/EC50 for reproduction). The authors mentioned that any concentrations for which significant survival effects occurred were not included in the analysis of reproductive effects. Acceptability criteria for temperature (25 +/- 2°C) and dissolved oxygen (>60%) were indicated (p. 3-1) and met. The authors documented the range of measured pH and DOC measurements (p. 3-1), but did not indicate what was considered an acceptable range (Note: there are no acceptability criteria defined in EPA guidance EPA-821-R-02-013 for these parameters). The authors report that Al concentrations among all quality control samples were within acceptability criteria of 85-115%, whereas the standard addition recoveries were within acceptability criteria of 116-102% with a few exceptions (n=7) (p. 3-4).	
Reviewer 2	Data acceptability criteria were not explicitly discussed but the software packages used to assess data have built in tests for homogeneity of variance, etc. Control performance should be explicitly discussed however.	
Reviewer 3	Determination of NOEC, LOEC, LCs, and ECs were described in the statistical analysis section. However, a separate section to define the measured endpoints of the test is recommended.	
Reviewer 4	<u>Response:</u> Test endpoints were sufficiently defined and explained. Data acceptability criteria were not well defined and explained. <u>Rationale</u> Although rather brief, the authors state under section 2.10.2 BIOLOGICAL MONITORING p. 2-5 that observations of live and dead organisms were conducted on a daily basis from initiation to termination, and that the numbers of young were counted	

Reviewer	Comments	EPA Response to Comments
	daily. This is sufficient to understand the test endpoints used, but it would be useful to know under what conditions the organisms were observed (light table? microscope? visual inspection only? time of day?) and how the test organisms were determined to be either dead or alive. Data acceptability criteria for this project were not offered. Most uses of data acceptance criteria involve some type of comparison among the data groups to determine if variability falls within a predetermined acceptable range but the predetermined acceptable range for normality and homogeneity for these tests were not stated by the authors. The only data acceptability evaluation offered was that if the data met the assumptions of normality and homogeneity, the NOEC and LOEC were estimated using an analysis of variance to compare (p. 2-6, the authors use "p = 0.05 "as the threshold for accepting a significant effect but the correct variable here would be "α = 0.05 "). There was no explanation offered on how the data were handled when the data did not meet assumptions of normality and homogeneity. If all data met those assumptions it should be stated in the report.	
Reviewer 5	The test endpoints and data acceptability criteria were well defined and explained in the text. I would like the authors to further evauluate the pH 6.3, hardness 60, DOC 2 treatment as to the appropriateness of the results. The 529 Al treatment had slightly better reproduction average than the next lower concentration (264.5 Al treatment). While I know that this sometimes happens, the control through the 529 Al treatment (represents 5 of the treatments) ranged in reproduction from 32.6 to 26.0 neonates (Table 3-12, page 3-15). This represents a wide range of treatment concentrations, with minimal change in neonate average production. I couldn't further evaluate whether there was something in this test that might explain this effect? All other tests looked adequate and were well defined and explained.	

2.7 Was preparation of test solutions fully described and target test concentrations verified prior to testing?

All five reviewers replied that the test solutions were fully described, and the target test concentrations were verified. One reviewer pointed out that one test (Al1185) did not have a day 3 sample reported. A second reviewer noted that verification of stock concentrations was not mentioned in the report. A third reviewer explained that analytical samples from each treatment were collected for analysis at test initiation, during the test, and at test termination; and that the authors discussed variabilities. A fourth reviewer commented that

test concentrations were extensively tested and verified during the study, but the report did not indicate whether this occurred prior to the study.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. Preparation of the test solutions is described in detail at the top of p. 2-3. Analytical samples from each treatment were collected for total Al and dissolved Al (<45 μm) analysis from newly prepared waters (after the 3-hr equilibrium period) at test initiation, during the tests, and from a composite of replicates at test termination (p. 2-5). Total Al concentrations prior to addition to test chambers were between 93 and 115% of nominal spiked concentrations, with four measurements outside of this range (with measurements of 75, 117, 120, and 130% of nominal). Total Al concentrations in test solutions measured in the replicate chambers at the end of the tests were more variable and the authors explained that it was more difficult to obtain homogeneous samples from the chambers and that these measurements were therefore less reliable (p. 3-4). In addition, dissolved Al concentrations were found to be highly variable, ranging from 0.1 to 111% of total Al. The authors explained that this was expected because the majority of solutions were well above solubility limits. There was some variability in the background levels of Al in the control water, presumably due to differences in natural organic matter.	
Reviewer 2	Test solutions that were aged 3 hours were taken on day 0 for both total and dissolved Al concentrations. All tests except Al 1185 CDC also had test solutions measured on days 3 and 6. The Al1185 tests did not have a day 3 sample reported.	
Reviewer 3	Yes, the preparation of the test solutions was fully described. The measured total Al were closed to the nominal concentrations. Usually stock concentrations are verified prior to use. However, it was not mentioned in the report.	
Reviewer 4	<u>Response:</u> Yes, the methods of test solution preparation were fully described. The target test concentrations (both of the treatment chemical, aluminum, and the evaluated water quality variables) appears to have been extensively tested and verified during the study but there is no indication that this occurred prior to the study. <u>Rationale:</u>	

Reviewer	Comments	EPA Response to Comments
	It appears that great attention was paid to chemical analyses in this project. The report provides an extensive description of the analytical methodology used, including composition of sampling containers, commercial source, preparation, and storage of test substance (p. 1-2), preparation and distribution of text concentrations (p. 2-1), method of pH control (p. 2-3), timing of collection, treatment and holding time of samples after collection, calibration of analytical instrumentation, use of blanks (p. 2-5), chain of custody documentation for samples analyzed, and data handling and storage of results. Analytical samples for each treatment were obtained from the newly prepared and equilibrated (3 hrs) test concentration prior to the start of the test but there is no indication that concentrations were verified before testing. Samples were taken for chemical analysis just prior to introduction of test organisms to the test chambers. According to Section 2.11 ANALYTICAL CONFIRMATION samples were analyzed for total and dissolved (defined as sample water that has passed through a 0.45 µM filter) using a Spectro Arcos ICP-OE according to US EPA Method 200.7. with quality control samples and spiked samples to determine % recovery. Appendix A (Protocol) indicates that this was a standard procedure for metal analysis to determine Al concentrations using an Inductively Coupled Plasma with either Optical Emission Spectrometry or Mass Spectrometry (p.7). The raw data for these analyses are provided in APPENDIX B – Metals Analytical Data and comprise the majority of the 405 pages of the appendices. Spiked samples were used to determine accuracy of analyses by calculating metal recovery and were shown to be within acceptable analytical limits.	
Reviewer 5	The test solutions were well described and were sufficiently verified prior to testing.	

2.8 Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?

All five reviewers responded that the test water quality variables (temperature, pH, conductivity, and dissolved oxygen) were measured with sufficient frequency and accuracy. However, one reviewer noted that the details for DOC could not be located. Another reviewer commented that the measurements for hardness and alkalinity were weak because they were only measured in the control water at test initiation. In contrast, a third reviewer noted that measurement of hardness and DOC only at the beginning is sufficient, because these variables are not expected to vary greatly.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. Temperature, pH, conductivity, and dissolved oxygen (DO) were measured in each concentration at test initiation, once daily, and at test termination. Hardness, alkalinity, ammonia, and total residual chlorine (TRC) were measured in the control water of each test at test initiation (p. 2-4). Other parameters (i.e., Calcium, magnesium, sodium, potassium, chloride, sulfate, cations, anions, and DOC) were measured by outside labs using accepted methods, but it is not entirely clear from the report how often these measurements were done.	
Reviewer 2	Temperature, pH, conductivity and DO were measured daily. Details of the frequency of verification for DOC concentrations were not found.	
Reviewer 3	The procedure for controlling test water quality, such as pH was clearly described. It was conducted carefully. Measurement of pH, DO, conductivity, and temperature were sufficient. The measured values represent the target values. However, hardness and alkalinity were measured only in the control water of each test at test initiation. This is weak rather than sufficient. These parameters are usually measured at least in control, the lowest and highest treatment concentrations at test initiation and termination to make sure the addition of toxicant into the test treatments does not change the water quality of the test water.	
Reviewer 4	Response: Yes, it appears that the manipulated test water quality variables (pH, hardness, and DOC; incorrectly called parameters in the report) were measured with sufficient frequency and accuracy to represent intended levels and allow incorporation into an updated predictive model of aluminum toxicity under varying water quality conditions. Rationale: Under Section 2.10 TEST MONITORING, subsection 2.10.1 WATER QUALITY the authors indicate that pH, hardness, and dissolved organic carbon (DOC) were measured during toxicity testing. pH was measured in each concentration at test initiation, once daily, and at test termination using a HACH HQ3od pH meter. Water hardness was measured in the control water of each test at test initiation using a colorimetric titration method following Standard Methods 2340B/C (APHA 2012). DOC was measured by an outside laboratory (Oregon State University Cooperative Chemical Analytical Laboratory (Corvallis, OR, USA)	

Reviewer	Comments	EPA Response to Comments
	using a Shimadzu TOC-VCNS total organic carbon analyzer (Shimadzu Scientific Instruments, Columbia, Maryland) following a Combustion method ((Standard Methods 5310B APHA 2012). All of the analytical instrumentation used are of sufficient quality to provide accurate, reproducible data results. Both water hardness and DOC would not be expected to vary greatly during a test exposure and thus measurement only at the beginning of the test would be sufficient. The mean and raw values for the data from these analyses are presented in Tables 3-1 and 3-1 in the report, and the Appendices C and D, respectively.	
Reviewer 5	Water quality variables were adequately manipulated. I believe that the use of the buffers as well as CO2 headspace was warranted for keeping these tight conditions with regards to the challenging pH parameter.	

2.9 Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?

All five reviewers thought that the frequency and accuracy of the chemical concentration measurements were generally sufficient. One reviewer responded that the measured concentrations of total aluminum were close to the nominal concentrations. However, two reviewers pointed out that dissolved aluminum concentrations were very inconsistent, which weakens confidence in the study. A fourth reviewer said that the measurement methods used by the researchers usually provide highly accurate and reproducible results in determining intended exposure levels.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. Al concentrations were measured at test initiation and once during each test, and from a composite of replicates at test termination. Samples were analyzed for total and dissolved (< 45 μm) Al using standard US EPA methods. Blanks and quality control samples were also run (p. 2-5).	
Reviewer 2	Generally, yes for total Al concentrations. Test Al1199 CDC reported considerable variation in total Al concentrations among days for a given nominal concentration. Dissolved Al concentrations were all over the map and incredibly inconsistent.	
Reviewer 3	Total and dissolved Al were measured in new and old waters at test initiation and termination and during the test period. This is sufficient. In addition, the measured concentrations of total Al	

Reviewer	Comments	EPA Response to Comments
	were closed to the nominal concentrations, presenting an accuracy of preparation and measurement of the test solutions. However, the measured dissolved Al concentrations were far away from the total concentrations. This weakens the confidence of this study.	
Reviewer 4	Response The frequency and accuracy of chemical concentrations of the non-manipulated water quality variables measured in test solutions appeared to be sufficient to represent intended exposure levels throughout the duration of the tests. Rationale Temperature, conductivity, and dissolved oxygen (DO) were measured in each concentration at test initiation, once daily from one of the test chambers at each concentration of aluminum, and at test termination. This frequency is standard protocol for water quality variables that may exhibit some variation in concentration over the duration of a test exposure. They were also measured in the renewal water prior to changing out the adult daphnids. These were reported to be calibrated prior to starting a measurement in Appendix A Protocol following Oregon State University Aquatic Toxicology Laboratory Standard Operating Procedures. These were measured using calibrated digital instrumentation as described in Section 2.4 DILUTION WATERS and reported in Table 2-1. Alkalinity, ammonia, and total residual chlorine (TRC), were measured in the control water of each test at test initiation using digital meters. Temperature was measured with a standard laboratory thermometer. Test solution pH was measured using a HACH (Loveland, CO, USA) HQ30d pH meter. These methods of measurement usually provide highly accurate and reproducible results sufficient to ensure determination of intended exposure levels.	
Reviewer 5	I believe that the frequency and accuracy of the chemical concentrations were sufficiently performed through the duration of the test. (see next charge question for additional input to this charge question).	

2.10 Were any anomalies in the test explained or justified with additional information or testing?

Three reviewers thought that the anomalies were explained or justified. One of these reviewers commented the total and dissolved aluminum measurement anomalies were explained and addressed. Another of these reviewers said the few anomalous data (i.e., DOC, pH, conductivity, and variability in total and dissolved aluminum recovery) were explained and justified without the need for additional data or testing. The third reviewer said that the anomalies could be classified as deviations from the protocol and suggested adding a section in the report to assess whether the anomalies potentially bias the results. Two reviewers did not think the anomalies were explained or justified.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes. The only anomalies were variability in the total Al concentrations measured in the chambers at the end of the test and in dissolved Al measurements. The authors explained these results (see answer to question 7). There was one test in which significant effects on reproduction occurred, and the authors addressed this by omitting the affected test concentrations from the reproductive effects analysis.	
Reviewer 2	No. Anomalies (see control reproduction in Al1199 CDC) were not explained or justified with additional testing.	
Reviewer 3	Not really, except for the procedure for controlling the pH of the test waters.	
Reviewer 4	<u>Response</u> The relatively few anomalous data were explained/justified without the need for additional data or testing. <u>Rationale</u> In Section 3. RESULTS AND CONCLUSIONS, subsection 3.1 TEST CONDITIONS the authors observed some variability in measured DOC. This has been observed in their testing laboratory previously and they believe it is due to using multiple batches of Suwanee Natural Organic Matter (NOM) which shows some variation in % DOC among batches. They also acknowledge that observed differences may be due to variability in analytical measurements. Because the DOC concentrations are reported as measured and not nominal, they should be acceptable for this project's goals of incorporation and expansion into the previously established predictive model. pH was maintained within 0.2 SU of the target pH in freshly prepared ("new") solutions after the equilibrium period. However, in some studies, an increase in pH occurred in the "old" waters	

Reviewer	Comments	EPA Response to Comments
	(pH up to 0.3 – 0.4 SU above the "new" waters) between each 24-hr water renewal. Both the use of the buffer to control pH, and also slightly adjusting the CO_2 atmosphere, limited observed pH drift within limits that allowed incorporation of mean pH values into the predictive model. Mean conductivity values remained consistent over the 24-hr period between water renewals. But in certain cases the range in conductivity was wide, primarily in the higher DOC tests (Table 3-2, p. 3-2). This is likely due to the higher DOC and cannot be eliminated as a (slightly) confounding factor. The authors also speculate that some increase in conductivity in the "old" water may be due to addition of food to the test chambers. The authors observed some variability in total Al recovery from "old" solutions and suggest this was primarily due to the difficulty in removing the entire homogenized aliquot because it has been altered during final enumeration of neonates by removing the organisms during counting (to prevent double counting). They believe this may have resulted in the accidental removal of precipitates from the non-homogeneous solution, potentially resulting in a misrepresentation of the entire fraction in the test chamber. Therefore, they feel that the "new" solutions are the most appropriate measurements for average exposure determination of Al. When comparing total Al to dissolved Al in the same sample, dissolved Al was much more variable than total Al, ranging from 0.1 to 111% of total Al. The author's expected this as the majority of solutions were well above solubility limits. The observed trend in dissolved concentrations was that higher percentages of dissolved/total were apparent in the lower exposure concentrations and percentages decreased as total Al increased. A few dissolved Al measurements were elevated and unexpected (and did not correspond to total dissolved Al samples from the identical concentration). The authors feel this is most likely associated with breaching of the 0.45 µM filter by insoluble Al clogging the filter and requiring additional pressure on the filter to obtain sufficient sample volume. The authors addressed this by keeping pressure on the filter at a minimum. Because (unlike most metals) the dissolved/free ion species of Al has relatively less effect on toxicity than the Al hydroxide species at circumneutral pH (6–8), and Al concentration–toxicity relationships correspond to total Al (Cardwell et al., 2017), total Al was incorporated into the predictive model .	

Reviewer	Comments	EPA Response to Comments
Reviewer 5	I believe that the anomalies observed during testing were well explained and the justification was sufficiently presented and plausible (page 3-4). However, these anomalies can be classified as deviations from protocol. I think this report would benefits from a section in the report presenting these identified anomalies and also the researchers should attempt to assess whether these anomalies potentially bias the results high, low, or neutral. I think that this section will help strength the report and further demonstrate a transparent process.	

2.11 Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?

One reviewer responded that the reported test results are directly applicable to the derivation of ambient water quality criteria (WQC) because the value is derived using a Multiple Linear Regression (MLR) model based on pH, DOC, and hardness, which are precisely the variables evaluated by this study. Therefore, the data can be used to refine the model. Another reviewer believes that the test results will strengthen the WQC and will be useful to the application of the Biotic Ligand Model (BLM) and MLR model but does not think the report presents the details needed to make this assessment. A third reviewer commented that the study covered a wide range of water quality parameters that are suitable for BLM development and calibration; and that reproductive results are useful for effect concentration determinations for total aluminum, but not dissolved aluminum. A fourth reviewer responded "as far as I can tell" because the authors followed standard EPA guidance and Good Laboratory Practice. The fifth reviewer wrote that it is impossible to answer this question without seeing the entire package of how water chemistry parameters are going to be used to model concentrations and link them to toxicity. This reviewer said it would be difficult to convince people that the dissolved concentrations can be predictive of toxicity without an understanding of how aluminum precipitates are toxic to daphnids.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	As far as I can tell. The authors followed standard US EPA guidance for conducting chronic toxicity tests with *Ceriodaphnia dubia* with some modifications to account for specific water types and to achieve effective pH control. The general US EPA criteria for test design and test acceptability were met, and the authors applied principles consistent with Good Laboratory Practice (GLP). Although documentation on culture maintenance and husbandry were not included in the report, the fact that the laboratory has been culturing this species successfully for over a decade and that control organisms showed acceptable performance, give little cause for concern related to maintenance and husbandry.	

Reviewer	Comments	EPA Response to Comments
Reviewer 2	Without seeing the entire package of how water chemistry parameters are going to be used to model both dissolved and particulate/precipitate concentrations and link these to toxicity, it is impossible to answer this question. The use of total recoverable Al as a descriptor for toxicity seems to run counter to BLM principles. Without direct evidence and mechanistic understanding of how Al precipitates are toxic to daphnids, it is going to be very difficult to convince people that the dissolved concentrations reported in these tests can be predictive of toxicity.	
Reviewer 3	This study covered a wide range of water quality parameters that are suitable for BLM development and calibration. Reproductive results showed concentration-response relationships that are useful for determination of effect concentrations based on total concentration basis but not for dissolved concentration basis.	
Reviewer 4	<u>Response</u> The reported test results do meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life. <u>Rationale</u> This study appears to have been carefully planned and executed and seems to compare well with the results of other similar studies and laboratories. For instance, the authors compared their (EC10/EC20 with 95% confidence interval results with Gensemer et al. (2018) using a one-sample paired-comparison t-test and found that the values were not statistically different between laboratories. The authors also endeavored to make the study results appropriate for inclusion in previously developed models. For example, the Biotic Ligand Model (BLM) uses Ca and Mg (in mg/L) as input variables to calculate hardness values and the multiple linear regression (MLR) for the Al toxicity prediction model on which the Water Quality Criterion is based uses hardness (as mg/L CaCO3). The calculated hardness values in Table 3-1 were used in the MLR analysis to maintain consistency between model input values derived from other studies. The results of this study are directly applicable to the EPA-developed WQC because that value is derived using an MLR model based on a site's pH, DOC, and hardness (EPA 2017). These water quality variables are precisely those evaluated by	

Reviewer	Comments	EPA Response to Comments
	manipulation in this study and thus the datasets can be included as part of the model refinement effort.	
Reviewer 5	I believe that these test results will strengthen the aluminum water quality criteria, however, I am not sure the results were meant to meet all of this charge question the way it was described. I am confident that these results will be very useful to the application of the BLM model and MLR model, however, the results presented in the report do not provide the details to make this assessment.	

2.12 Is there any reason to be concerned with the use of the test results in the criteria derivation process?

Two reviewers did not have concerns with using the test results in the WQC derivation process. One said that the main goal of the study was to increase the understanding of bioavailability and toxicity of aluminum to aquatic organisms, and that the objectives were met. A third reviewer said that three of the tests had very steep concentration-response relationships and were flagged as being useful for exploratory analysis only. This reviewer suggested running the models with and without these data to judge whether they can be used for model refinements. A fourth reviewer replied that the BLM cannot be applied to the dissolved aluminum concentration data from the study, because they were substantially off from the total concentrations. The fifth reviewer responded that this study was not able to help researchers' understanding of the effects of dissolved aluminum on *C. dubia*. This reviewer feels there is still considerable uncertainty with both dissolved and particulate aluminum forms.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Three of the tests had very steep concentration-response relationships and were flagged by the TRAP model as being useful for exploratory analysis only due to an inadequate number of partial effects. It is difficult to judge what the effect of including these test results in the Biotic Ligand Model and Multiple Linear Regression Model would be. Certainly the models could be run with and without these data and a judgement made as to whether their precision was sufficient for inclusion in the model refinements.	
Reviewer 2	The complexity of Al chemistry makes this very challenging. We do not appear to be closer to understanding the effects of dissolved Al and its speciation on C. dubia as a result of these studies, because the dissolved concentrations are not tractable due to precipitation issues.	

Reviewer	Comments	EPA Response to Comments
	The uncertainty of the kinetics of precipitate formation and the effects of those precipitates on different forms of aquatic life bring a large amount of uncertainty into the equation. How does a 3 hour equilibration period in the laboratory (with high buffer concentrations) translate to animal exposures in nature? It is interesting that EPA is willing to consider Al solid phases in toxicity characterization, but generally refuses to consider the effects of dietary exposures of metals – which are known to cause deleterious effects in aquatic life. Thus, there appears to be considerable uncertainty with respect to both dissolved and particulate Al forms. It would appear that both dissolved criteria based on BLB type principles and particulate criteria would be needed – or that a considerably large uncertainty factor would be applied to a total Al measurement.	
Reviewer 3	The concern about this study is the measured dissolved Al concentrations. Dissolved Al concentrations were totally off the total concentrations, especially at high concentrations. A few examples are the measured dissolved concentrations were below the detection limit or 7 or 45 ug/L at the total Al concentrations of 5000 and 10000ug/L (Table 3-6, Test Al 1205CDC), or 80-217 ug/L at the total Al concentrations of 300-12000ug/L (Table 3-8, Test Al 1198CDC). Dissolved metal concentration has been using for evaluating metal bioavailability, especially using the BLM approach. Given that said, I don't know how the BLM can be applied to the dissolved concentration data set in this report.	
Reviewer 4	<u>Response</u> I do not believe there is any significant reason to be concerned with using the test results from this report in the water quality criterion derivation process. <u>Rationale</u> The main goal of this project was to increase understanding of the bioavailability and toxicity of Al to aquatic organisms. To reach this goal, the main objectives of this project were 1) to quantify the effects of water quality on Al toxicity and 2) to use the results to develop a bioavailability-based model to predict Al toxicity across a wider range of certain water quality variables (specifically pH, hardness, and dissolved organic carbon). I believe this study has achieved these objectives and has increased the applicable range of previous predictive models	

Reviewer	Comments	EPA Response to Comments
	used to derive an Al WQC. The expansion included increasing pH from 8.10 up to 8.70, hardness (as $CaCO_3$) up to 428 mg/L from 123 mg/L, and dissolved organic carbon from 4.0 mg/L up to 12.30 mg/L. Comparison of the current model predicted effect concentrations with observed effect concentrations, for water types outside the previous range of model development, suggests very good predictive capabilities of this new model (Table 3 – 13) and thus may be confidently used in the water quality criterion derivation process. In terms of future Al toxicity testing with the goal of developing a new WQC, I would like to see the following suggestions to be considered: 1) Al toxicity tests performed with sodium aluminum sulfate (probably as $NaAl(SO_4)_2 \cdot 12H_2O$. This would help address the massive problem with sulfuric acid-derived acid mine drainage (AMD), of which elevated Al is often a constituent. There are more than 500,000 abandoned and inactive mines in 32 states and AMD has degraded more than 8,000 miles of streams in Appalachia alone. 2) I would have preferred to see pH controlled in a flow-thru set-up, perhaps using a digital controller (Grippo 1997) rather than by buffers, which introduce a possibly confounding effect on the results. A flow-through protocol has not yet been developed for fecundity of Ceriodaphnia dubia but development of such a protocol would significantly increase environmental realism.	
Reviewer 5	I have not concerns with regards to the use of the test results in the criteria derivation process.	

3.0 ADDITIONAL COMMENTS PROVIDED

Reviewer	Comments	EPA Response to Comments
Reviewer 4	<u>Suggestions to authors</u> - Authors frequently use the phrase “In order to”. Reducing this phrase to simply “To” will convey the same meaning with fewer words, enhancing the goal of preparing scientific prose that exhibits clarity and brevity.	

Reviewer	Comments	EPA Response to Comments
	- In Part 3.3 BIOLOGICAL RESULTS, paragraph 3 the authors state "The results were quite comparable to those reported in Gensemer et al. (2018) (EC10/EC20 with 95% confidence intervals: 504.4 (226 – 1126) µg/L total Al and 631.3 (362 - 1101) µg/L total Al, respectively). A one sample t-test was performed and the values were not statistically different between laboratories. Because the comparison was between two independent populations of test results (ration of EC_{10}/EC_{20} a two – sample t-test may have been more appropriate. - Table 3-12. Some of the data are set off by both asterisks and bold-type. In the text it is stated that this indicates significant differences. I suggest including an explanation of what the bold-face and asterisks denote in the table heading, rather than the text, so the reader does not have to go searching in the text to determine the meaning of these highlighted results.	
Reviewer 5	**General Comments:** I found this report to be well written and supported using the information in the appendices. I support the use of these results for the derivation of the aluminum ambient water quality criteria. **Specific comments from reviewer:** • While the *Ceriodaphnia* tests followed the protocols as presented in Appendix A, the test as described by US EPA is a 3-brood test. However as specified in the protocol, the tests were carried out with 7-days of exposure (and potentially extended another day if 3-broods did not occur) rather than as a 3-brood test. Thus, the average neonates were considerably higher than normal 3-brood tests. I think that this should be mentioned in the results. Also, some of the variability during testing might also be explained because the protocol did not specify that the neonates are <24 hours old (from an 8-hour window). While the researchers followed the protocol, these two issues are outside of the US EPA methods that were reported in the Methods and Materials section (page 2-1). • What was the normality of the dilute NaOH and HCl? (Section 2.5, page 2-3) • Section 2.8 it should be pH rather than all capital letters (page 2-3).	

Reviewer	Comments	EPA Response to Comments
	• Good spike response, however, I think the dissolved Al observation needs its own paragraph. It is buried in the middle of the second paragraph on page 3-4. • The report states that there was no protocol deviations and amendments, however, there were several deviations that were noted in the text (i.e., 45% bisections rather than 50% bisections). This section needs revised as well as I recommend, as stated above, the researchers should assess whether the deviations bias the results potentially high, low, or neutral.	

4.0 NEW INFORMATION PROVIDED BY REVIEWERS

This section presents all new information that reviewers provided in addition to or within their specific responses (presented in Section 2, above) to the charge questions.

Reviewer	Comments	EPA Response to Comments
Reviewer 4	References cited American Public Health Association (APHA). 2012. Standard Methods for the Examination of Water and Wastewater, 22nd edition. Washington, D.C. Cardwell AS, WJ Adams, RW Gensemer, E Nordheim, RC Santore, AC Ryan, WA Stubblefield. 2017. Chronic toxicity of aluminum, at a pH of 6, to freshwater organisms: empirical data for the development of international regulatory standards/criteria. Environ. Toxicol. Chem. 37:36-48. EPA 2002. Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms Fifth Edition October 2002. U.S. Environmental Protection Agency Office of Water (4303T) 1200 Pennsylvania Avenue, NW Washington, DC 20460. EPA 2017. United States Environmental Protection Agency. Fact Sheet: Draft Aquatic Life Ambient Water Quality Criteria for Aluminum in Freshwaters Office of Water EPA 820-F-17-002 July 2017. Gensemer, R, J Gondek, P Rodriquez, JJ Arbildua, WA Stubblefield, AS Cardwell, RC Santore, A Ryan, WJ Adams, E Nordheim. 2018. Evaluating the effects of pH, hardness, and dissolved organic carbon on the toxicity of aluminum to freshwater aquatic organisms under circumneutral conditions. Environ Toxicol Chem. 37:49-60. Grippo, R.S. 1997. A gravity-based system for controlling pH in flow-through aquatic toxicity experiments. Environmental Technology. 18:763-768.	

APPENDIX A

CHARGE TO REVIEWERS

Technical Charge to External Peer Reviewers

Contract No. EP-C-17-017
Task Order 68HE0C18F0787
July 2018

External Peer Review of Invertebrate Toxicity Tests for Aluminum

BACKGROUND

The U.S. Environmental Protective Agency (EPA) Office of Water is charged with protecting ecological integrity and human health from adverse anthropogenic, water-mediated effects, under the purview of the Clean Water Act (CWA). In concurrence with this mission, EPA is working to update water quality criteria to protect aquatic life from the presence of aluminum in freshwater environments. Invertebrate toxicity tests for aluminum have been conducted and are yet unpublished. EPA is seeking a focused, objective evaluation of these invertebrate toxicity tests that may be used in the development of the model used to determine aquatic life criteria for aluminum.

CHARGE QUESTIONS

1. Were an adequate number of concentrations tested to fully-characterize concentration-response and determine an accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?
2. Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?
3. Was the source, maintenance, and husbandry of test organisms well described?
4. Were the control's survival rates acceptable?
5. Were test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?
6. Were test endpoints and data acceptability criteria well defined and explained?
7. Was preparation of test solutions fully described and target test concentrations verified prior to testing?
8. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?
9. Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?
10. Were any anomalies in the test explained or justified with additional information or testing?
11. Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?
12. Is there any reason to be concerned with the use of the test results in the criteria derivation process?

APPENDIX B

INDIVIDUAL REVIEWER COMMENTS

COMMENTS SUBMITTED BY

REVIEWER 1

External Peer Review of Chronic Toxicity of Aluminum to the Cladoceran, *Ceriodaphnia dubia*: Expansion of the Empirical Database for Bioavailability Modeling

1. **Were an adequate number of concentrations tested to fully-characterize concentration-response and determine an accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?**

Yes. The test was conducted following standard US EPA chronic testing methodology according to US EPA (2002). This reference is not provided in the reference list (it should be), but presumably refers to EPA-821-R-02-013. According to this guidance, a minimum of 5 test concentrations and a control should be used in a definitive test. As each test in this study included 5 exposure concentrations and a dilution water control (p. 2-2), it is judged to be adequate for the test purpose. The range of concentrations chosen was also deemed adequate to achieve estimates of the desired effect levels for reproduction (10, 20, and 50% effect; Table 3-13). With the exception of one test in which effects on survival occurred, all test concentrations could be used to estimate reproductive effects.

2. **Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?**

Yes. There were 10 replicate chambers for each exposure concentration and control, each containing one cladoceran. This is consistent with US EPA guidance (EPA-821-R-02-013).

3. **Was the source, maintenance, and husbandry of test organisms well described?**

Partially. The source of the organisms was well described. They were obtained from in-house cultures that had been maintained for over 10 years and originally obtained from Aquatic BioSystems (Fort Collins, CO, USA) (p. 2-1). Maintenance and husbandry of the test organisms were not described in the report, although the authors did indicate that they conducted monthly tests with a reference toxicant (NaCl) to confirm that the organisms were in good condition (p. 2-1).

4. **Were the control's survival rates acceptable?**

Yes. The authors report that in all tests, control acceptability criteria (> 80 % survival and > 60% surviving females having 15 or more neonates) were met (p. 3-14). These fulfill the criteria for test acceptability outlined in EPA-821-R-02-013.

5. **Were test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?**

Yes, as far as hardness is concerned. Organisms cultured under standard conditions (100 mg/L as CaCO3) were used in the moderately hard water tests (120 mg/L as $CaCO_3$). Organisms were acclimated to the soft (60 mg/L as $CaCO_3$) and hard water (250 and 400 mg/L as $CaCO_3$) conditions for multiple generations (i.e., over two months), and survival and reproduction were reported to be excellent (p. 2-2). As far as indicated in the report, there was no acclimation for different pH (tested range: 6.3 – 8.8; standard culture at 7.8-8.0) or DOC (tested range: 1-14 mg/L; standard culture unknown) conditions.

6. Were test endpoints and data acceptability criteria well defined and explained?

Yes. Test endpoints included NOEC and LOEC for survival and reproduction (if data met assumptions of normality and homogeneity), as well as effect concentrations (i.e., LC10/LC20/LC50 for survival and ECx10/EC20/EC50 for reproduction). The authors mentioned that any concentrations for which significant survival effects occurred were not included in the analysis of reproductive effects. Acceptability criteria for temperature (25 +/- 2°C) and dissolved oxygen (>60%) were indicated (p. 3-1) and met. The authors documented the range of measured pH and DOC measurements (p. 3-1), but did not indicate what was considered an acceptable range (Note: there are no acceptability criteria defined in EPA guidance EPA-821-R-02-013 for these parameters). The authors report that Al concentrations among all quality control samples were within acceptability criteria of 85-115%, whereas the standard addition recoveries were within acceptability criteria of 116-102% with a few exceptions (n=7) (p. 3-4).

7. Was preparation of test solutions fully described and target test concentrations verified prior to testing?

Yes. Preparation of the test solutions is described in detail at the top of p. 2-3. Analytical samples from each treatment were collected for total Al and dissolved Al (<45 μm) analysis from newly prepared waters (after the 3-hr equilibrium period) at test initiation, during the tests, and from a composite of replicates at test termination (p. 2-5). Total Al concentrations prior to addition to test chambers were between 93 and 115% of nominal spiked concentrations, with four measurements outside of this range (with measurements of 75, 117, 120, and 130% of nominal). Total Al concentrations in test solutions measured in the replicate chambers at the end of the tests were more variable and the authors explained that it was more difficult to obtain homogeneous samples from the chambers and that these measurements were therefore less reliable (p. 3-4). In addition, dissolved Al concentrations were found to be highly variable, ranging from 0.1 to 111% of total Al. The authors explained that this was expected because the majority of solutions were well above solubility limits. There was some variability in the background levels of Al in the control water, presumably due to differences in natural organic matter.

8. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?

Yes. Temperature, pH, conductivity, and dissolved oxygen (DO) were measured in each concentration at test initiation, once daily, and at test termination. Hardness, alkalinity, ammonia, and total residual chlorine (TRC) were measured in the control water of each test at test initiation (p. 2-4). Other parameters (i.e., Calcium, magnesium, sodium, potassium, chloride, sulfate, cations, anions, and DOC) were measured by outside labs using accepted methods, but it is not entirely clear from the report how often these measurements were done.

9. Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?

Yes. Al concentrations were measured at test initiation and once during each test, and from a composite of replicates at test termination. Samples were analyzed for total and dissolved (< 45 μm) Al using standard US EPA methods. Blanks and quality control samples were also run (p. 2-5).

10. Were any anomalies in the test explained or justified with additional information or testing?

Yes. The only anomalies were variability in the total Al concentrations measured in the chambers at the end of the test and in dissolved Al measurements. The authors explained these results (see answer to question 7). There was one test in which significant effects on reproduction occurred, and the authors addressed this by omitting the affected test concentrations from the reproductive effects analysis.

11. Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?

As far as I can tell. The authors followed standard US EPA guidance for conducting chronic toxicity tests with *Ceriodaphnia dubia* with some modifications to account for specific water types and to achieve effective pH control. The general US EPA criteria for test design and test acceptability were met, and the authors applied principles consistent with Good Laboratory Practice (GLP). Although documentation on culture maintenance and husbandry were not included in the report, the fact that the laboratory has been culturing this species successfully for over a decade and that control organisms showed acceptable performance, give little cause for concern related to maintenance and husbandry.

12. Is there any reason to be concerned with the use of the test results in the criteria derivation process?

Three of the tests had very steep concentration-response relationships and were flagged by the TRAP model as being useful for exploratory analysis only due to an inadequate number of partial effects. It is difficult to judge what the effect of including these test results in the Biotic Ligand Model and Multiple Linear Regression Model would be. Certainly the models could be run with and without these data and a judgement made as to whether their precision was sufficient for inclusion in the model refinements.

COMMENTS SUBMITTED BY

REVIEWER 2

External Peer Review of Chronic Toxicity of Aluminum to the Cladoceran, *Ceriodaphnia dubia*: Expansion of the Empirical Database for Bioavailability Modeling

1. **Were an adequate number of concentrations tested to fully-characterize concentration-response and determine an accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?**

A total of nine different tests were conducted under different pH, hardness and DOC conditions. Five total Al concentrations plus controls were generally used in the various tests. This number of concentrations is generally considered adequate.

2. **Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?**

Yes. Ten replicates per treatment is adequate.

3. **Was the source, maintenance, and husbandry of test organisms well described?**

Not particularly. This section was remarkably brief and lacking details of animal performance for the reference toxicant tests. The reporting of volumes of algal suspensions used for feeding are not useful unless cell densities are reported.

4. **Were the control's survival rates acceptable?**

The average number of neonates/female in controls ranged from 22 to 37 with 42.5 reported from a "concurrent control". The test with the poor control reproductive output (Al1199 CDC) should not be used.

5. **Were test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?**

The report only mentions acclimation of cultures to different hardness levels, but not pH and DOC or buffers.

6. **Were test endpoints and data acceptability criteria well defined and explained?**

Data acceptability criteria were not explicitly discussed but the software packages used to assess data have built in tests for homogeneity of variance, etc. Control performance should be explicitly discussed however.

7. **Was preparation of test solutions fully described and target test concentrations verified prior to testing?**

Test solutions that were aged 3 hours were taken on day 0 for both total and dissolved Al concentrations. All tests except Al 1185 CDC also had test solutions measured on days 3 and 6. The Al1185 tests did not have a day 3 sample reported.

8. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?

Temperature, pH, conductivity and DO were measured daily. Details of the frequency of verification for DOC concentrations were not found.

9. Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?

Generally, yes for total Al concentrations. Test Al1199 CDC reported considerable variation in total Al concentrations among days for a given nominal concentration.

Dissolved Al concentrations were all over the map and incredibly inconsistent.

10. Were any anomalies in the test explained or justified with additional information or testing?

No. Anomalies (see control reproduction in Al1199 CDC) were not explained or justified with additional testing.

11. Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?

Without seeing the entire package of how water chemistry parameters are going to be used to model both dissolved and particulate/precipitate concentrations and link these to toxicity, it is impossible to answer this question. The use of total recoverable Al as a descriptor for toxicity seems to run counter to BLM principles. Without direct evidence and mechanistic understanding of how Al precipitates are toxic to daphnids, it is going to be very difficult to convince people that the dissolved concentrations reported in these tests can be predictive of toxicity.

12. Is there any reason to be concerned with the use of the test results in the criteria derivation process?

The complexity of Al chemistry makes this very challenging. We do not appear to be closer to understanding the effects of dissolved Al and its speciation on C. dubia as a result of these studies, because the dissolved concentrations are not tractable due to precipitation issues.

The uncertainty of the kinetics of precipitate formation and the effects of those precipitates on different forms of aquatic life bring a large amount of uncertainty into the equation. How does a 3 hour equilibration period in the laboratory (with high buffer concentrations) translate to animal exposures in nature? It is interesting that EPA is willing to consider Al solid phases in toxicity characterization, but generally refuses to consider the effects of dietary exposures of metals – which are known to cause deleterious effects in aquatic life.

Thus, there appears to be considerable uncertainty with respect to both dissolved and particulate Al forms. It would appear that both dissolved criteria based on BLB type principles and particulate criteria would be needed – or that a considerably large uncertainty factor would be applied to a total Al measurement.

COMMENTS SUBMITTED BY

REVIEWER 3

External Peer Review of Chronic Toxicity of Aluminum to the Cladoceran, *Ceriodaphnia dubia*: Expansion of the Empirical Database for Bioavailability Modeling

1. **Were an adequate number of concentrations tested to fully-characterize concentration-response and determine an accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?**

Yes, 5 concentrations of Al and a negative control were used for each test. This design appeared to follow the EPA guidelines for toxicology testing with freshwater organisms. The concentrations used were low that did not result in complete mortality at the highest concentration of each test. Therefore, lethal effect concentrations (LCs) could not be calculated.

2. **Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?**

Yes, 10 replicates per treatment were usually used for this type of test. The report (section 2.9) did not clearly say the number of organisms used per replicate chamber.

3. **Was the source, maintenance, and husbandry of test organisms well described?**

Organisms were originally from Aquatic Biosystems and cultured at OSU for more than 10 years. Organisms were cultured in moderately hard water. Other environmental conditions and maintenance procedures were not described, such as temperature, photoperiod (light:dark hours), food, feeding rates, biomass/water volume, water change, etc.

4. **Were the control's survival rates acceptable?**

The survival of the control organisms of each test was 100%. This meets the test acceptability criteria of the test method (80-100%).

5. **Were test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?**

Yes, the acclimation of the organisms to the hardness of test waters (250 and 400 mg/L as CaCO3) for multiple generations and over more than 2 months should be adequate.

6. **Were test endpoints and data acceptability criteria well defined and explained?**

Determination of NOEC, LOEC, LCs, and ECs were described in the statistical analysis section. However, a separate section to define the measured endpoints of the test is recommended.

7. **Was preparation of test solutions fully described and target test concentrations verified prior to testing?**

Yes, the preparation of the test solutions was fully described. The measured total Al were closed to the nominal concentrations. Usually stock concentrations are verified prior to use. However, it was not mentioned in the report.

8. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?

The procedure for controlling test water quality, such as pH was clearly described. It was conducted carefully. Measurement of pH, DO, conductivity, and temperature were sufficient. The measured values represent the target values. However, hardness and alkalinity were measured only in the control water of each test at test initiation. This is weak rather than sufficient. These parameters are usually measured at least in control, the lowest and highest treatment concentrations at test initiation and termination to make sure the addition of toxicant into the test treatments does not change the water quality of the test water.

9. Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?

Total and dissolved Al were measured in new and old waters at test initiation and termination and during the test period. This is sufficient. In addition, the measured concentrations of total Al were closed to the nominal concentrations, presenting an accuracy of preparation and measurement of the test solutions. However, the measured dissolved Al concentrations were far away from the total concentrations. This weakens the confidence of this study.

10. Were any anomalies in the test explained or justified with additional information or testing?

Not really, except for the procedure for controlling the pH of the test waters.

11. Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?

This study covered a wide range of water quality parameters that are suitable for BLM development and calibration. Reproductive results showed concentration-response relationships that are useful for determination of effect concentrations based on total concentration basis but not for dissolved concentration basis.

12. Is there any reason to be concerned with the use of the test results in the criteria derivation process?

The concern about this study is the measured dissolved Al concentrations. Dissolved Al concentrations were totally off the total concentrations, especially at high concentrations. A few examples are the measured dissolved concentrations were below the detection limit or 7 or 45 ug/L at the total Al concentrations of 5000 and 10000ug/L (Table 3-6, Test Al 1205CDC), or 80-217 ug/L at the total Al concentrations of 300-12000ug/L (Table 3-8, Test Al 1198CDC). Dissolved metal concentration has been using for evaluating metal bioavailability, especially using the BLM approach. Given that said, I don't know how the BLM can be applied to the dissolved concentration data set in this report.

COMMENTS SUBMITTED BY

REVIEWER 4

External Peer Review of Chronic Toxicity of Aluminum to the Cladoceran, *Ceriodaphnia dubia*: Expansion of the Empirical Database for Bioavailability Modeling

1. Were an adequate number of concentrations tested to fully-characterize concentration-response and determine an accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?

Response:

In my opinion, an adequate number of concentrations were tested to allow full characterization of the concentration response and allow determination of a scientifically-defensible chronic effect concentration.

Rationale:

This research project evaluated the effects of multiple water quality variables on the toxicity of Aluminum (Al) to the cladoceran Ceriodaphnia dubia. The goal of the study was to increase the range of water quality variables under which a reasonable prediction of invertebrate toxicity could be performed under a given set of water quality variables. The test followed standard USEPA methodology (US EPA 2002). The methods included in this manual are referenced in Table IA, 40 CFR Part 136 regulations and, therefore, constitute approved methods for acute toxicity tests. These methods were used in the present study with modifications to address different water types and pH levels. For example, concentrations were based on previous studies shown to cause a negative impact on C. dubia survival and reproduction. The standard EPA protocol calls for five test concentrations and a control and this was mostly followed in the present study. For one test (Test #: Al 1185 CDC; p. 12, Appendices (page 1, Appendix B) six concentrations of Al were used, plus a treatment labeled “non pH”). This was apparently a confirmatory test for comparison to results obtained at the Chilean Mining and Metallurgy Research Center (CIMM; Santiago, Chile) and Universidad Adolfo Ibañez (UAI; Santiago, Chile) and reported in Gensemer et al. (2018) as indicated on p. 29, paragraph 3. Five concentrations is the number usually followed by most toxicity testing laboratories including those administered by the US EPA (such as the EPA facility in Cincinnati, OH with which I am familiar). This allows the present study to be compared to the results of other laboratories and have such results be incorporated into the statistical model developed by the authors. This regression model can be used to develop a scientifically defensible chronic effect concentration such as the EC20 (dose which causes a 20% change from control response of the test organisms).

2. Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?

Response:

Yes, the number of replicates (10 per Al treatment concentration and 10 in the non-treated control) was sufficient to allow sufficient statistical rigor for a C. dubia chronic toxicity evaluation under the stated test conditions.

Rationale:

Ten replicates of each toxicant concentration and the control is the number recommended by the US EPA (2002). This number of replicates is used by most toxicity testing laboratories, allowing comparison of the

results of the present study with previous (and likely future) results from other laboratories. Statistical dogma suggests that ≈30 replicates is the optimal number when evaluating biological data. However, in this (and most other toxicity testing laboratories) the test conditions were carefully controlled, using 1) moderately hard diluent water prepared in-house (please see question 7 below), 2) environmental chambers controlled for pH and light regimen, and 3) neonates that were all less than 24 hours old. All of these conditions will serve to reduce variability in organism response to exposure, which will support rigorous statistical testing using 10 replicates.

3. Was the source, maintenance, and husbandry of test organisms well described?

Response:

No, an adequate description of the source, maintenance, and husbandry of the C. daphnia test organism was not provided.

Rationale:

In the report, section 2.3.2 SOURCE, the authors state that the <24 hour old neonates were obtained from in-house cultures which have been maintained successfully at the Aquatic Toxicology laboratory at Oregon State University (Corvallis) for >10 years. In Appendix A, section 2.2 and 2.3, feeding diet and feeding regimen during toxicity testing were described. However, nowhere that I could find in the report was it explicitly stated that the test organisms were cultured and maintained under these same conditions. I believe this is an oversight in reporting, not a failure of procedure, and this oversight can be readily remedied by the authors by providing the missing information. Husbandry of the test organisms during culture and testing as described appeared to be adequate.

4. Were the control's survival rates acceptable?

Response:

Yes, it appears that the survival rate of C. dubia used in the control (no aluminum) treatments met the accepted survival rate for this type of toxicity testing.

Rationale:

The standard methodology as developed by the US EPA (1982) calls for at least 80% survival of the control test organisms for the test to be considered valid. On p. 29, paragraph 2, the authors state that, in all tests, control acceptability criteria (> 80 % survival and > 60% surviving females having 15 or more neonates) were met. Table 3-12 (p. 30 of report) and Appendix D Raw Data both indicate that control survival was uniformly 100%, clearly meeting the EPA (2002) control standard for test acceptability.

5. Were test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?

Response:

It would appear that the C. dubia used in these toxicity tests were appropriately acclimated for the stated test type and described test water conditions at the time the chronic toxicity testing was performed

Rationale:

The C. dubia used for the present study were reported (Section 2.3.4 ACCLIMATION p. 2-2;) as being cultured at the Ohio State University AquaTox laboratory, in a "moderately hard" reconstituted water that was prepared as detailed in standard USEPA methods (USEPA 2002). This diluent was reported to have a measured hardness of 100 mg/L as CaCO3 and pH of 7.8 – 8.0, p. 2-2). All acclimated cultures for all of the toxicity tests were successfully maintained in their respective laboratory water for multiple generations (2+ months). Organism survival and reproduction were reported as excellent and organism health was maintained over the period of acclimation.

Note: In section 2.3.4, ACCLIMATION is erroneously labeled, in section 2.3.2 SOURCE, as section 2.4.3).

6. Were test endpoints and data acceptability criteria well defined and explained?

Response:

Test endpoints were sufficiently defined and explained. Data acceptability criteria were not well defined and explained.

Rationale

Although rather brief, the authors state under section 2.10.2 BIOLOGICAL MONITORING p. 2-5 that observations of live and dead organisms were conducted on a daily basis from initiation to termination, and that the numbers of young were counted daily. This is sufficient to understand the test endpoints used, but it would be useful to know under what conditions the organisms were observed (light table? microscope? visual inspection only? time of day?) and how the test organisms were determined to be either dead or alive.

Data acceptability criteria for this project were not offered. Most uses of data acceptance criteria involve some type of comparison among the data groups to determine if variability falls within a predetermined acceptable range but the predetermined acceptable range for normality and homogeneity for these tests were not stated by the authors. The only data acceptability evaluation offered was that if the data met the assumptions of normality and homogeneity, the NOEC and LOEC were estimated using an analysis of variance to compare (p. 2-6, the authors use "p = 0.05 "as the threshold for accepting a significant effect but the correct variable here would be "$\alpha = 0.05$ "). There was no explanation offered on how the data were handled when the data did not meet assumptions of normality and homogeneity. If all data met those assumptions it should be stated in the report.

7. Was preparation of test solutions fully described and target test concentrations verified prior to testing?

Response:

Yes, the methods of test solution preparation were fully described. The target test concentrations (both of the treatment chemical, aluminum, and the evaluated water quality variables) appears to have been extensively tested and verified during the study but there is no indication that this occurred prior to the study.

Rationale:

It appears that great attention was paid to chemical analyses in this project. The report provides an extensive description of the analytical methodology used, including composition of sampling containers, commercial source, preparation, and storage of test substance (p. 1-2), preparation and distribution of text concentrations (p. 2-1), method of pH control (p. 2-3), timing of collection, treatment and holding time of samples after collection, calibration of analytical instrumentation, use of blanks (p. 2-5), chain of custody documentation for samples analyzed, and data handling and storage of results. Analytical samples for each treatment were obtained from the newly prepared and equilibrated (3 hrs) test concentration prior to the start of the test but there is no indication that concentrations were verified before testing. Samples were taken for chemical analysis just prior to introduction of test organisms to the test chambers. According to Section 2.11 ANALYTICAL CONFIRMATION samples were analyzed for total and dissolved (defined as sample water that has passed through a 0.45 µM filter) using a Spectro Arcos ICP-OE according to US EPA Method 200.7. with quality control samples and spiked samples to determine % recovery. Appendix A (Protocol) indicates that this was a standard procedure for metal analysis to determine Al concentrations using an Inductively Coupled Plasma with either Optical Emission Spectrometry or Mass Spectrometry (p.7). The raw data for these analyses are provided in APPENDIX B – Metals Analytical Data and comprise the majority of the 405 pages of the appendices. Spiked samples were used to determine accuracy of analyses by calculating metal recovery and were shown to be within acceptable analytical limits.

8. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?

Response:

Yes, it appears that the manipulated test water quality variables (pH, hardness, and DOC; incorrectly called parameters in the report) were measured with sufficient frequency and accuracy to represent intended levels and allow incorporation into an updated predictive model of aluminum toxicity under varying water quality conditions.

Rationale:

Under Section 2.10 TEST MONITORING, subsection 2.10.1 WATER QUALITY the authors indicate that pH, hardness, and dissolved organic carbon (DOC) were measured during toxicity testing. pH was measured in each concentration at test initiation, once daily, and at test termination using a HACH HQ3od pH meter. Water hardness was measured in the control water of each test at test initiation using a colorimetric titration method following Standard Methods 2340B/C (APHA 2012). DOC was measured by an outside laboratory (Oregon State University Cooperative Chemical Analytical Laboratory (Corvallis, OR, USA) using a Shimadzu TOC-VCNS total organic carbon analyzer (Shimadzu Scientific Instruments, Columbia, Maryland) following a Combustion method ((Standard Methods 5310B APHA 2012). All of the analytical instrumentation used are of sufficient quality to provide accurate, reproducible data results. Both water hardness and DOC would not be expected to vary greatly during a test exposure and thus measurement only at the beginning of the test would be sufficient. The mean and raw values for the data from these analyses are presented in Tables 3-1 and 3-1 in the report, and the Appendices C and D, respectively.

9. Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?

Response

The frequency and accuracy of chemical concentrations of the non-manipulated water quality variables measured in test solutions appeared to be sufficient to represent intended exposure levels throughout the duration of the tests.

Rationale

Temperature, conductivity, and dissolved oxygen (DO) were measured in each concentration at test initiation, once daily from one of the test chambers at each concentration of aluminum, and at test termination. This frequency is standard protocol for water quality variables that may exhibit some variation in concentration over the duration of a test exposure. They were also measured in the renewal water prior to changing out the adult daphnids. These were reported to be calibrated prior to starting a measurement in Appendix A Protocol following Oregon State University Aquatic Toxicology Laboratory Standard Operating Procedures. These were measured using calibrated digital instrumentation as described in Section 2.4 DILUTION WATERS and reported in Table 2-1. Alkalinity, ammonia, and total residual chlorine (TRC), were measured in the control water of each test at test initiation using digital meters. Temperature was measured with a standard laboratory thermometer. Test solution pH was measured using a HACH (Loveland, CO, USA) HQ30d pH meter. These methods of measurement usually provide highly accurate and reproducible results sufficient to ensure determination of intended exposure levels.

10. Were any anomalies in the test explained or justified with additional information or testing?

Response

The relatively few anomalous data were explained/justified without the need for additional data or testing.

Rationale

In Section 3. RESULTS AND CONCLUSIONS, subsection 3.1 TEST CONDITIONS the authors observed some variability in measured DOC. This has been observed in their testing laboratory previously and they believe it is due to using multiple batches of Suwanee Natural Organic Matter (NOM) which shows some variation in % DOC among batches. They also acknowledge that observed differences may be due to variability in analytical measurements. Because the DOC concentrations are reported as measured and not nominal, they should be acceptable for this project's goals of incorporation and expansion into the previously established predictive model.

pH was maintained within 0.2 SU of the target pH in freshly prepared ("new") solutions after the equilibrium period. However, in some studies, an increase in pH occurred in the "old" waters (pH up to 0.3 – 0.4 SU above the "new" waters) between each 24-hr water renewal. Both the use of the buffer to control pH, and also slightly adjusting the CO_2 atmosphere, limited observed pH drift within limits that allowed incorporation of mean pH values into the predictive model.

Mean conductivity values remained consistent over the 24-hr period between water renewals. But in certain cases the range in conductivity was wide, primarily in the higher DOC tests (Table 3-2, p. 3-2). This is likely due to the higher DOC and cannot be eliminated as a (slightly) confounding factor. The authors also speculate that some increase in conductivity in the "old" water may be due to addition of food to the test chambers.

The authors observed some variability in total Al recovery from "old" solutions and suggest this was primarily due to the difficulty in removing the entire homogenized aliquot because it has been altered during final enumeration of neonates by removing the organisms during counting (to prevent double counting). They believe this may have resulted in the accidental removal of precipitates from the non-homogeneous solution, potentially resulting in a misrepresentation of the entire fraction in the test chamber. Therefore, they feel that the "new" solutions are the most appropriate measurements for average exposure determination of Al.

When comparing total Al to dissolved Al in the same sample, dissolved Al was much more variable than total Al, ranging from 0.1 to 111% of total Al. The author's expected this as the majority of solutions were well above solubility limits. The observed trend in dissolved concentrations was that higher percentages of dissolved/total were apparent in the lower exposure concentrations and percentages decreased as total Al increased. A few dissolved Al measurements were elevated and unexpected (and did not correspond to total dissolved Al samples from the identical concentration). The authors feel this is most likely associated with breaching of the 0.45 µM filter by insoluble Al clogging the filter and requiring additional pressure on the filter to obtain sufficient sample volume. The authors addressed this by keeping pressure on the filter at a minimum. Because (unlike most metals) the dissolved/free ion species of Al has relatively less effect on toxicity than the Al hydroxide species at circumneutral pH (6–8) , and Al concentration–toxicity relationships correspond to total Al (Cardwell et al., 2017), total Al was incorporated into the predictive model .

11. Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?

Response

The reported test results do meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life.

Rationale

This study appears to have been carefully planned and executed and seems to compare well with the results of other similar studies and laboratories. For instance, the authors compared their (EC_{10}/EC_{20} with 95% confidence interval results with Gensemer et al. (2018) using a one-sample paired-comparison t-test and found that the values were not statistically different between laboratories. The authors also endeavored to make the study results appropriate for inclusion in previously developed models. For example, the Biotic Ligand Model (BLM) uses Ca and Mg (in mg/L) as input variables to calculate hardness values and the multiple linear regression (MLR) for the Al toxicity prediction model on which the Water Quality Criterion is based uses hardness (as mg/L CaCO3). The calculated hardness values in Table 3-1 were used in the MLR analysis to maintain consistency between model input values derived from other studies.

The results of this study are directly applicable to the EPA-developed WQC because that value is derived using an MLR model based on a site's pH, DOC, and hardness (EPA 2017). These water quality variables are precisely those evaluated by manipulation in this study and thus the datasets can be included as part of the model refinement effort.

12. Is there any reason to be concerned with the use of the test results in the criteria derivation process?

Response

I do not believe there is any significant reason to be concerned with using the test results from this report in the water quality criterion derivation process.

Rationale

The main goal of this project was to increase understanding of the bioavailability and toxicity of Al to aquatic organisms. To reach this goal, the main objectives of this project were 1) to quantify the effects of water quality on Al toxicity and 2) to use the results to develop a bioavailability-based model to predict Al toxicity across a wider range of certain water quality variables (specifically pH, hardness, and dissolved organic carbon). I believe this study has achieved these objectives and has increased the applicable range of previous predictive models used to derive an Al WQC. The expansion included increasing pH from 8.10 up to 8.70, hardness (as $CaCO_{3)}$ up to 428 mg/L from 123 mg/L, and dissolved organic carbon from 4.0 mg/L up to 12.30 mg/L. Comparison of the current model predicted effect concentrations with observed effect concentrations, for water types outside the previous range of model development, suggests very good predictive capabilities of this new model (Table 3 – 13) and thus may be confidently used in the water quality criterion derivation process.

In terms of future Al toxicity testing with the goal of developing a new WQC, I would like to see the following suggestions to be considered:

1) Al toxicity tests performed with sodium aluminum sulfate (probably as $NaAl(SO_4)_2 \cdot 12H_2O$. This would help address the massive problem with sulfuric acid-derived acid mine drainage (AMD), of which elevated Al is often a constituent. There are more than 500,000 abandoned and inactive mines in 32 states and AMD has degraded more than 8,000 miles of streams in Appalachia alone.

2) I would have preferred to see pH controlled in a flow-thru set-up, perhaps using a digital controller (Grippo 1997) rather than by buffers, which introduce a possibly confounding effect on the results. A flow-through protocol has not yet been developed for fecundity of Ceriodaphnia dubia but development of such a protocol would significantly increase environmental realism.

Suggestions to authors

- Authors frequently use the phrase "In order to". Reducing this phrase to simply "To" will convey the same meaning with fewer words, enhancing the goal of preparing scientific prose that exhibits clarity and brevity.

- In Part 3.3 BIOLOGICAL RESULTS, paragraph 3 the authors state "The results were quite comparable to those reported in Gensemer et al. (2018) (EC10/EC20 with 95% confidence intervals: 504.4 (226 –

1126) µg/L total Al and 631.3 (362 -1101) µg/L total Al, respectively). A one sample t-test was performed and the values were not statistically different between laboratories. Because the comparison was between two independent populations of test results (ration of EC_{10}/EC_{20} a two – sample t-test may have been more appropriate.

- Table 3-12. Some of the data are set off by both asterisks and bold-type. In the text it is stated that this indicates significant differences. I suggest including an explanation of what the bold-face and asterisks denote in the table heading, rather than the text, so the reader does not have to go searching in the text to determine the meaning of these highlighted results.

References cited

American Public Health Association (APHA). 2012. Standard Methods for the Examination of Water and Wastewater, 22nd edition. Washington, D.C.

Cardwell AS, WJ Adams, RW Gensemer, E Nordheim, RC Santore, AC Ryan, WA Stubblefield. 2017. Chronic toxicity of aluminum, at a pH of 6, to freshwater organisms: empirical data for the development of international regulatory standards/criteria. Environ. Toxicol. Chem. 37:36-48.

EPA 2002. Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms Fifth Edition October 2002. U.S. Environmental Protection Agency Office of Water (4303T) 1200 Pennsylvania Avenue, NW Washington, DC 20460.

EPA 2017. United States Environmental Protection Agency. Fact Sheet: Draft Aquatic Life Ambient Water Quality Criteria for Aluminum in Freshwaters Office of Water EPA 820-F-17-002 July 2017.

Gensemer, R, J Gondek, P Rodriquez, JJ Arbildua, WA Stubblefield, AS Cardwell, RC Santore, A Ryan, WJ Adams, E Nordheim. 2018. Evaluating the effects of pH, hardness, and dissolved organic carbon on the toxicity of aluminum to freshwater aquatic organisms under circumneutral conditions. Environ Toxicol Chem. 37:49-60.

Grippo, R.S. 1997. A gravity-based system for controlling pH in flow-through aquatic toxicity experiments. Environmental Technology. 18:763-768.

COMMENTS SUBMITTED BY

REVIEWER 5

External Peer Review of Chronic Toxicity of Aluminum to the Cladoceran, *Ceriodaphnia dubia*: Expansion of the Empirical Database for Bioavailability Modeling

General Comments:

I found this report to be well written and supported using the information in the appendices. I support the use of these results for the derivation of the aluminum ambient water quality criteria.

Review Charge Questions:

1. **Were an adequate number of concentrations tested to fully-characterize concentration-response and determine and accurate and scientifically-defensible chronic effect concentration (e.g., EC20)?**

The study was performed following the agreed to protocol. However, one study used a 45% bisection of the test concentrations rather than the protocol specified 50% bisection. While I do not believe that this is a fatal flaw in the analysis, I believe that it does warrant a section in the report for protocol deviations (rather than as only noted in Section 2.5 [page 2-2]). This would also provide an opportunity to offer the analytical issues (as identified in Section 3.2 [page 3-4]). I also believe the authors should assess whether the analytical anomalies bias the results high, low, or neutral. This is very helpful in the use of these results.

In my overall opinion, all test concentrations were sufficiently characterized to provide a meaningful and accurate description of the test results and the chronic toxicity of aluminum.

2. **Was there a sufficient number of replicates for each test concentration and control to pass statistical rigor for the type of test and test conditions?**

The number of replicates (10) and test concentrations (minimally 5 plus a control) were standard with in ecotoxicity testing with *Ceriodaphnia dubia*. These are acceptable.

3. **Was the source, maintenance, and husbandry of test organisms well described?**

The description of the test animals was adequately presented in the report. Reference toxicant testing was regularly performed as part of the quality assurance program.

4. **Were the control's survival rates acceptable?**

Control survival rates were acceptable.

5. **Were test organisms appropriately acclimated for the type of test water conditions to represent their chronic sensitivity under those conditions?**

I was quite impressed with the acclimation process used in this study. In many instances, researchers do not go to the length of details used for the acclimation protocol performed in this study. The researches should be commended on this practice.

6. Were test endpoints and data acceptability criteria well defined and explained?

The test endpoints and data acceptability criteria were well defined and explained in the text. I would like the authors to further evauluate the pH 6.3, hardness 60, DOC 2 treatment as to the appropriateness of the results. The 529 Al treatment had slightly better reproduction average than the next lower concentration (264.5 Al treatment). While I know that this sometimes happens, the control through the 529 Al treatment (represents 5 of the treatments) ranged in reproduction from 32.6 to 26.0 neonates (Table 3-12, page 3-15). This represents a wide range of treatment concentrations, with minimal change in neonate average production. I couldn't further evaluate whether there was something in this test that might explain this effect? All other tests looked adequate and were well defined and explained.

7. Was preparation of test solutions fully described and target test concentrations verified prior to testing?

The test solutions were well described and were sufficiently verified prior to testing.

8. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?

Water quality variables were adequately manipulated. I believe that the use of the buffers as well as CO_2 headspace was warranted for keeping these tight conditions with regards to the challenging pH parameter.

9. Was the frequency and accuracy of chemical concentrations measured in test solutions sufficient to represent intended exposure levels throughout the duration of the test(s)?

I believe that the frequency and accuracy of the chemical concentrations were sufficiently performed through the duration of the test. (see next charge question for additional input to this charge question).

10. Were any anomalies in the test explained or justified with additional information or testing?

I believe that the anomalies observed during testing were well explained and the justification was sufficiently presented and plausible (page 3-4). However, these anomalies can be classified as deviations from protocol. I think this report would benefits from a section in the report presenting these identified anomalies and also the researchers should attempt to assess whether these anomalies potentially bias the results high, low, or neutral. I think that this section will help strength the report and further demonstrate a transparent process.

11. Do the reported test results meet or exceed expectations for use in model development for the derivation of ambient water quality criteria for the protection of aquatic life?

I believe that these test results will strengthen the aluminum water quality criteria, however, I am not sure the results were meant to meet all of this charge question the way it was described. I am confident that these results will be very useful to the application of the BLM model and MLR model, however, the results presented in the report do not provide the details to make this assessment.

12. Is there any reason to be concerned with the use of the test results in the criteria derivation process?

I have not concerns with regards to the use of the test results in the criteria derivation process.

Specific comments from reviewer:

- While the *Ceriodaphnia* tests followed the protocols as presented in Appendix A, the test as described by US EPA is a 3-brood test. However as specified in the protocol, the tests were carried out with 7-days of exposure (and potentially extended another day if 3-broods did not occur) rather than as a 3-brood test. Thus, the average neonates were considerably higher than normal 3-brood tests. I think that this should be mentioned in the results. Also, some of the variability during testing might also be explained because the protocol did not specify that the neonates are <24 hours old (from an 8-hour window). While the researchers followed the protocol, these two issues are outside of the US EPA methods that were reported in the Methods and Materials section (page 2-1).
- What was the normality of the dilute NaOH and HCl? (Section 2.5, page 2-3)
- Section 2.8 it should be pH rather than all capital letters (page 2-3).
- Good spike response, however, I think the dissolved Al observation needs its own paragraph. It is buried in the middle of the second paragraph on page 3-4.
- The report states that there was no protocol deviations and amendments, however, there were several deviations that were noted in the text (i.e., 45% bisections rather than 50% bisections). This section needs revised as well as I recommend, as stated above, the researchers should assess whether the deviations bias the results potentially high, low, or neutral.