TASK ORDER 68HE0C18F0792 UNDER
CONTRACT EP-C-17-017

EXTERNAL PEER REVIEW OF SHORT-TERM CHRONIC TOXICITY OF ALUMINUM TO THE FATHEAD MINNOW, *PIMEPHALES PROMELAS*: EXPANSION OF THE EMPIRICAL DATABASE FOR BIOAVAILABILITY MODELING

FINAL PEER REVIEW SUMMARY REPORT

October 17, 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 *Short-term chronic toxicity of Aluminum to the fathead minnow, Pimephales promelas: 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 the individual reviewer comments in response to each charge question 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.

1.1 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 fish 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 peer review to evaluate these fish toxicity tests, which were unpublished at the time the review was conducted.

1.2 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 and 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 fish 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 replied that an adequate number of concentrations (five plus a control) were tested. Four of them noted that this procedure followed standard EPA guidelines. One reviewer noted there had been two challenges during testing, but concluded that these were appropriately addressed such that all test concentrations were sufficiently characterized to provide meaningful accurate test results.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Each of the tests were conducted with 5 concentration plus controls. This is generally considered acceptable for establishing concentration-response relationships provided adequate range finding is conducted.	
Reviewer 2	Five concentrations of Al and a control were used for each test. This is technically adequate for calculating LC/EC values. The design is in compliance with the USEPA guidelines for toxicology testing with aquatic organisms. Two out of the 7 tests got survival concentration-response relationship that allowed calculation of NOEC, LOEC, and LC values. All anticipated sublethal endpoints were calculated based on concentration-response relationships of the growth data.	
Reviewer 3	Response: It appears that an adequate number and range of concentrations were used in this project to allow full characterization of the concentration response and allow determination of a scientifically-defensible chronic effect concentration. Rationale: The goal of this research project was to evaluate the effects of multiple water quality variables on the concentration-dependent toxicity of aluminum (Al) to the standard test vertebrate *Pimephales promelas*. The study was designed to increase the range of water quality variables under which a reasonable prediction of fish toxicity could be made under a given range of water quality variables. The test followed standard USEPA methodology (US EPA 2002). The methods included in the EPA	

Reviewer	Comments	EPA Response to Comments
	manual are referenced in Table IA, 40 CFR Part 136 regulations and, therefore, constitute approved methods for acute toxicity tests of fish. 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 predictable negative impact mainly on growth, and to a much lesser extent, survival of *P. promelas* (Santore et al., 2018; DeForest et al., 2018; Gensemer et al., 2017) The standard EPA protocol calls for five test concentrations and a control and this was followed in the present study. The concentrations of Al used were based on historical response data with *P. promelas* in other reconstituted water (Page 2-3, paragraph 1). Five concentrations is the standard number of concentrations used by most toxicity testing laboratories, allowing 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 and assumed to be the degree of negative change from which an organism cannot recover).	
Reviewer 4	Yes. The test was conducted following standard US EPA chronic testing methodology according to US EPA (2002) with modifications for testing with Al. 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 (section 2.5), it is judged to be adequate for the test purpose. The range of concentrations was chosen on the basis of preliminary results and by putting nominal water quality characteristics into the bioavailability models to predict effects (section 2-5). This seems a reasonable approach. In 7 of 7 tests it was possible to estimate chronic effect concentrations (NOEC, LOEC, EC_{10}, EC_{20} and EC_{50}; Table 3-11) for growth. For 5 of the 7 tests, there was no dose-response for survival, and no EC_x values could be estimated (Table 3-11).	
Reviewer 5	The study was performed following the agreed to protocol. One challenge was in the middle of the testing program that laboratory was moved from one location to another. I believe that the PI and Study Coordinator adequately evaluated potential difference in the culturing and resulting testing by	

Reviewer	Comments	EPA Response to Comments
	additional quality control procedures that adequately assessed that there was no differences. Each test was performed with five treatments and a control, with four replicates in a random arrangement. This procedure follows standard EPA test procedures. Control survival was acceptable in all testing. One issue that occurred during testing was that the dissolved aluminum concentrations were considerably lower than the total aluminum concentration. I believe the study team adequately addressed this issue in the interpretation of the study 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 a sufficient number of replicates (four replicates with 10 fish per replicate) was used for each test concentration, and that this procedure followed standard EPA guidelines. One reviewer explained that about 30 replicates is the optimal number statistically speaking. However, because the test conditions are carefully controlled, the study supports rigorous statistical testing using four replicates.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Four replicates were tested per condition, with each replicate represented by 10 individuals. The standard deviations of the toxicity responses were relatively modest, suggesting that replication was adequate for these studies.	
Reviewer 2	Yes, 4 replicates per treatment with 10 fish per replicate were usually used for this type of test with fish.	
Reviewer 3	Response: Yes, the number of replicates (four per Al treatment concentration and four in the untreated control) was sufficient to allow acceptable statistical rigor for a *P. promelas* chronic toxicity evaluation under the stated test conditions. Rationale: Four replicate chambers (with 10 organisms in each chamber) of each toxicant concentration and the control are the numbers recommended by the US EPA (2002). This number of replicates is used by most toxicity testing laboratories, allowing comparison	

Reviewer	Comments	EPA Response to Comments
	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 Charge 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 four replicates. My only question with the statistics entails the statement: "If the data met the assumptions of normality and homogeneity, the NOEC and LOEC were estimated using an Analysis of Variance…" (Page 2-6, paragraph 1). It is unclear how the authors proceeded if the data did not meet parametric assumptions.	
Reviewer 4	There were four replicates per test concentration. According to US EPA (2002, section 12.10.2.1), this is the recommended number of replicates for this kind of test.	
Reviewer 5	The number of replicates (four) and test concentrations (minimally five plus a control) were standard with in ecotoxicity testing with *Pimephales promelas*. Testing was also performed in a randomized manner concerning treatment and replicate placement. These are acceptable.	

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

The reviewers generally felt that the source, maintenance, and/or husbandry of the test organisms were not described well enough. One reviewer said the description of the feeding rations was not adequate, though other aspects of the source, maintenance, and husbandry description were adequate. Another reviewer noted that environmental conditions and maintenance procedures such as temperature, photoperiod, food, feeding rates, biomass/water volume, and water change were not described. A third reviewer wrote that husbandry of the test organisms was adequately described, but that information on diet was insufficient and unclear. A fourth reviewer replied that the source was clearly described, but maintenance and husbandry were only partly described. The fifth reviewer thought that the description was adequately presented.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Generally, the source maintenance and husbandry of the test organisms was well described, however the description of the feeding rations was not adequate. Reporting a volume of a food suspension is meaningless unless we know the density of the food items in the volume of water provided to the test organisms.	
Reviewer 2	Yes, the organisms were originally from Aquatic Biosystems and cultured at OSU for more than 10 years. However, due to the laboratory move, adult broodstocks were cultured at two different locations at slightly different water quality. For example, pH of 7.8-8.0 compared to 6.6-6.8 and hardness of 100-120 mg/L as CaCO3 compared to 132 mg/L as CaCO3. 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.	
Reviewer 3	Response: No, the source, maintenance, and husbandry of the *P. promelas* test organisms were not adequately described. Rationale: In the report, section 2.3.2 SOURCE, the authors state that the <24 hour old larval fish 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 (Test System, #7) the authors state that the newly hatched larval fish were fed 0.15 mL of a Yeast/Trout Chow/Cereal leaves mixture (YTC) and algae suspension (*Pseudokirchneriella subcapitata,* 1:1), twice daily (a.m. and p.m.). I believe this is what is normally fed to *Ceriodaphnia dubia* during culture and testing, not *to P. promelas*. Later on in Appendix A (2.3 Test Diet), the authors state that brine shrimp (Artemia) nauplii <24 hours old were fed to the test fish. Which diet was actually fed to the test fish (I am guessing the latter)? Also, the above two diets were stated to have been fed to the *P. promelas* during testing but not explicitly stated in the report or appendices that the test organisms were cultured and maintained under the same food regimen. 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	

Reviewer	Comments	EPA Response to Comments
	information. Husbandry of the test organisms during culture and testing as described appeared to be adequate.	
Reviewer 4	The source of the fish was clearly described. Fish were obtained from in-house cultures, and their original source was from Aquatic Biosystems (Fort Collins, CO, USA; section 2.3.2). The maintenance and husbandry were partly described. The culture water was described in detail (section 2.3.2), however I could find no other details on the maintenance or husbandry of the test organisms. I also checked the OSU Protocol No. Al-PP-CSR7d-035, provided as Appendix A, but could not find details there either. As the species has been cultured in house for many generations, and the fish were determined to be in good health prior to testing (as described in section 2.3.3), it can probably be assumed that maintenance and husbandry conditions were 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 to ensure that the fathead minnow were health and consistent in their toxicological response.	

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

Three reviewers responded that acclimatizing the embryos for four days seems sufficient. One of them commended the researchers for their detailed acclimation protocol and the use of non-metal chelating buffers and carbon dioxide headspace procedures to control acclimation and testing pH.

A fourth reviewer noted that criteria for determining when an organism is acclimated are rarely defined, therefore it is difficult to know whether four days of acclimation were sufficient. Also, it is unclear whether conditions such as temperature, light regime, and food were similar between cultures and test conditions. The fifth reviewer was concerned that it was not feasible to appropriately acclimate the organisms because the tests were initiated with larvae less than 24 hours old; this reviewer suggested that it might have been better to both rear the parents and allow the eggs to develop and hatch in the appropriate waters.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Because the tests were initiated with larvae <24 hours old, it is not feasible to appropriately acclimate the animals. The fertilized eggs of the test animals were hatched at hardness of 100 mg/L and apparently transferred to higher hardness waters during egg development. My understanding from reading this	

Reviewer	Comments	EPA Response to Comments
	vague text is that the eggs were transferred from hardness 100 waters to higher hardnesses for the final 4 days of development, but this should be clarified. It might have been better to rear the parents in the appropriate waters and allow the eggs to develop and hatch in the appropriate control waters.	
Reviewer 2	The organism acclimation to different hardness was described. The acclimation period was 4 days, which seems to be fine. No acclimation to different pH was mentioned.	
Reviewer 3	Response: It appears that the *P. promelas* were appropriately acclimated for test conditions at the time during which the toxicity testing was performed. Rationale: The *P. promelas* 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 $CaCO_3$, alkalinity of 70 mg/L as CaCO3, and pH of 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. For the higher hardness tests (hardness of 250 and 400 mg/L $CaCO_3$), embryos were acclimated over four days from the above described moderately hard water starting immediately after hatching. This should be sufficient time for complete acclimation.	
Reviewer 4	For all of the tests, the larvae were hatched in moderately hard, reconstituted lab water. For 5 of the tests, the larvae were kept in this water until test initiation; for the 2 higher hardness tests, larvae were acclimated to the higher hardness test water (250 and 400 mg/L as $CaCO_3$) for 4 days after hatching (section 2.3.4). Since criteria for determining when an organism is actually acclimated are rarely defined, it is difficult to say whether 4 days was sufficient. There is no further mention of acclimation in the report and therefore assumed that other conditions of the test (e.g., temperature, light regime, food) were similar between cultures and test conditions.	

Reviewer	Comments	EPA Response to Comments
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. In addition, I appreciate the use of non-metal chelating buffers and the CO_2 headspace procedures to control acclimation and testing pH in this study. The researches should be commended on this practice.	

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

The reviewers generally responded that the test endpoints were sufficiently defined. Three reviewers replied that the data acceptability criteria were well defined and explained, but a fourth reviewer said that they were not—for example, the predetermined acceptable range for normality and homogeneity for these tests were not stated by the authors. Also, the authors did not explain how data were handled when they did not meet assumptions of normality and homogeneity; if all data met those assumptions, it should be stated in the report. The fifth reviewer did not comment on the data acceptability criteria.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Only mortality and dry mass are provided as endpoints, with the analysis focused primarily on weight.	
Reviewer 2	Test endpoints (NOEC, LOEC, LCs, and ECs) were described in the statistical analysis section. Acceptability criteria for control survival and growth were mentioned. The results met the acceptability criteria.	
Reviewer 3	Response: Test endpoints were sufficiently defined and explained. Data acceptability criteria were not well defined and explained. Rationale: Although rather brief, the author's state under section 2.10.2 BIOLOGICAL MONITORING p. 2-5 that observations of live and dead fish were conducted on a daily basis from initiation to termination, and dead fish were removed immediately. 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	

Reviewer	Comments	EPA Response to Comments
	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 ". Incidentally, I made the same statement on my review of the *Ceriodaphnia dubia* aluminum toxicity report). 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 4	Test endpoints were survival and growth. According to OSU Protocol No Al-PP-CSR7d-035, death was defined as the lack of movement in response to gentle prodding (Protocol section 4.6). Growth was estimated as mean dry biomass at the end of the test (i.e., total dry weight of surviving organisms divided by the original number of organisms at test initiation; section 3.3). Quality criteria for the test are explicitly defined in section 4.9 of the Protocol (Appendix A).	
Reviewer 5	The test endpoints and data acceptability criteria were well defined and explained in the text. The authors had issues with dissolved concentrations being considerably lower than total (and this did not always follow a dose response relationship). I believe the authors adequately addressed it in their report. Since they are using measured concentrations for the expression of toxicity, it is being adequately represented in the conclusions.	

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

Four reviewers replied that the test solutions were fully described. Two of these four also thought test concentrations were sufficiently verified. Of the other two reviewers, one commented that while the target test concentrations were extensively tested and verified *during* the study, the report does not state whether the analytical equipment was tested and calibrated *prior* to the study. The other reviewer commented that stock concentrations are usually verified prior to use, but this was not mentioned in the report.

The final reviewer responded that preparation of the exposure waters was a major issue because the test solutions were only allowed to equilibrate for three hours before initiating the tests, which likely resulted in highly dynamic exposure conditions as the aluminum precipitated during the fish exposures. This reviewer also found the difference between dissolved and total concentrations disturbing and noted that error estimates were not provided for the measured constituents. Another reviewer also noted the large variation between the dissolved concentrations and total concentrations.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Test concentrations were not verified prior to testing because the test solutions were only allowed to equilibrate for 3 hours before the initiation of the tests. All varication appears post-hoc. In one test (Al1222) measured initial concentrations are significantly higher than targeted nominal concentrations. Preparation of the exposure waters is a major issue with these tests. Waters were made, pH adjusted and allowed to equilibrate for only 3 hours. This resulted in highly dynamic exposure conditions as Al is likely precipitating during the fish exposures. If the goal was to evaluate the physical effects of Al precipitates on larval fish, this might be appropriate, but it unclear to me how this reflects bioavailability and traditional toxicity evaluation. The difference between dissolved and total concentrations (especially comparing the "new" and 'old" dissolved concentrations is disturbing). Table 3-1 shows no error estimates in any of the measured constituents, though error estimates are provided for pH, conductivity and DO in table 3-2.	
Reviewer 2	The preparation of the test solutions was clearly described. The measured total Al were closed to the nominal concentrations but large variation between the measured dissolved concentrations and total concentrations was reported. Usually, stock concentrations are verified prior to use. However, it was not mentioned in the report. The authors mentioned that the stock concentrations were likely higher than the target concentrations. This likely resulted in consistently higher measured total Al concentrations than the target nominal concentrations.	
Reviewer 3	<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 t*he study but there it was not explicitly stated that the analytical equipment was tested and calibrated *prior* to the study. <u>Rationale:</u> It appears that the analytical portion of this project was very carefully performed and documented. The report provides an	

Reviewer	Comments	EPA Response to Comments
	extensive description of the analytical methodology used, including composition of sampling containers, commercial source, preparation, and storage of test substance (p. 2-1), 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), 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 in section 2.10.3 under Dissolved Metals but defined as "<0.45 **µg/L**" in Section 2.2, last sentence) 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 321 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 4	Yes. Preparation of test solutions is described in detail in section 2.5, and both total and dissolved Al were measured at test initiation as described in section 2.10. 3. Data verifying target test concentrations are provided in Tables 3.3 – 3.8.	
Reviewer 5	The test solutions were well described and were sufficiently verified prior to testing.	

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

Four of the reviewers responded that the manipulated test water quality variables were sufficiently measured. One reviewer commented that hardness and alkalinity were only measured for control water of each test at test initiation and that the frequency of the DOC measurement was not reported. Another reviewer

commented that water hardness and DOC would not be expected to vary during the test, so measuring these variables just prior to the start is sufficient.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	Yes, pH, DOC and hardness were well monitored during the tests.	
Reviewer 2	The procedure for controlling the quality of the test water, such as pH was clearly described. It was conducted carefully. Concentrations of DO, pH, conductivity, and temperature were measured daily and therefore sufficient. The measured values were around the target values. However, hardness and alkalinity were measured only for control water of each test at test initiation. No description for the frequency of DOC measurement was reported. These water quality parameters are usually measured at least for 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 3	<u>Response:</u> From the report 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. <u>Rationale:</u> Under Section 2.10 TEST MONITORING, subsection 2.10.1 WATER QUALITY the authors indicate that pH, conductivity, and dissolved organic carbon (DOC) were measured in each concentration at test initiation, once daily, and at test termination using a HACH HQ3od pH meter. These variables were measured in both the replenishment water and in one test chamber just prior to replenishment. 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	

Reviewer	Comments	EPA Response to Comments
	results. Both water hardness and DOC would not be expected to vary greatly during a test exposure and thus measurement just prior to the start of a test would be sufficient. The mean and raw values for the data from these analyses are presented in Tables 3-1 and 3-2 in the report, and Appendices C and D, respectively.	
Reviewer 4	Yes. These were measured at test initiation, once daily and at test termination in both "new" and "old" water as described in section 2.10.1. Data verifying that water quality variables were sufficiently maintained are provided in Table 3-2.	
Reviewer 5	Water quality variables were adequately manipulated. I believe that the use of the buffers, as well as the CO_2 headspace technique, were warranted for keeping these tight conditions concerning the challenging pH parameters used in this testing program.	

2.8 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)?

Four reviewers responded that the frequency of the chemical concentrations measured in test solutions was sufficient. One of these reviewers noted that the frequency is standard protocol for water quality variables and that the measurement methods usually provide highly accurate and reproducible results. Two of these four reviewers commented that accuracy was also sufficient; the other two did not comment on this point.

Three reviewers pointed out that the measured dissolved concentrations were considerably lower than the total concentrations. One of these reviewers said the authors adequately addressed this issue in the study conclusions by using measured concentrations for the expression of toxicity. However, another of these reviewers commented that this weakens confidence in the study's metal analysis and biological results; this reviewer was not convinced that instrument performance uncertainty was the explanation. The third of these reviewers did not think that either frequency and accuracy of the chemical concentrations measured in test solutions was sufficient and said that use the nominal concentrations or the mean measured new concentrations as descriptors of toxicity was arbitrary because the differences between the total and dissolved aluminum concentrations were extreme.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	No. These tests were conducted under highly variable conditions. It is completely arbitrary to use the nominal concentrations or even the mean measured "new" concentrations as descriptors of toxicity because the differences between the total and dissolved Al concentrations were extreme.	

Reviewer	Comments	EPA Response to Comments
Reviewer 2	Concentrations of 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. However, the measured dissolved Al concentrations were largely deviated from the total concentrations. This weakens the confidence of metal analysis and biological results of the study. One of the explanations for the variation was the uncertainty in performance of the instrument at different times. This explanation doesn't sound convincing because the measured total Al concentrations seem to be fine for all treatments throughout the study.	
Reviewer 3	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 toxicity test exposure. They were also measured in the renewal water prior to renewing 80% of the water in the test and control chambers (Section 2.9 TEST INITIATION, SOLUTION RENEWAL, AND FEEDING). The instrumentation used for these measurements 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	Comments	EPA Response to Comments
Reviewer 4	Yes. Total Al was measured in each treatment in newly prepared waters ("new") at test initiation, twice during the tests, and from a composite of replicates at test termination ("old"). Dissolved Al (< 0.45 µm) was similarly measured at test initiation and termination, but only once during the tests. Detailed results of the metal analyses are provided in Appendix B.	
Reviewer 5	I believe that the frequency and accuracy of the chemical concentrations were sufficiently performed through the duration of the test. The authors had issues with dissolved concentrations being considerably lower than total (and they did not always follow a dose response relationship). I believe the authors adequately addressed it in their report. Since they are using measured concentrations for the expression of toxicity, it is being adequately represented in the conclusions of this study. (See next charge question for additional input to this charge question).	

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

Two reviewers responded that the anomalies were explained and justified; a third reviewer appeared to share this opinion. One of these reviewers said that all anomalous data occurred within the water quality and aluminum measurement results and were few in number, thus explained/justified without the need for additional data or testing. A fourth reviewer replied that there were no significant anomalies in the data. A fifth reviewer answered, "not really, except for the procedure for controlling the pH of the test waters."

Reviewer	Comments	EPA Response to Comments
Reviewer 1	There were no significant anomalies in the data.	
Reviewer 2	Not really, except for the procedure for controlling the pH of the test waters.	
Reviewer 3	Response: All anomalous data occurred within the water quality and Al measurement results. These were few in number. They were explained/justified without the need for additional data or testing. Rationale:	

Reviewer	Comments	EPA Response to Comments
	• The authors report that adult *P. promelas* broodstock were moved to a new laboratory location and reared for a period of three months in well water with a hardness of 132 mg/L as CaCO3 and pH of 6.6 – 6.8. Larval fish from this adult broodstock were used for tests Al 1218 PPC, Al 1222 PPC, and Al 1225 PPC (Section 2.3.2 SOURCE, paragraph 2, page 2.1). No differences were observed between offspring from broodstock cultured in the two laboratory waters following reference toxicity testing (Section 2.3.3 ORGANISM HEALTH, paragraph 1, page 2.2). • The authors noted high variability in measurements of dissolved organic carbon (Section 3.1 TEST CONDITIONS, paragraph 1, page 3.1). They attribute this variability to the need to use multiple batches of Suwannee River Natural Organic Matter (Suwannee NOM) which has historically been variable in DOC. They also acknowledge the possibility of observed differences being due to variability in analytical technique. However, they did not feel the observed differences were significant and reported the DOC as measured. • Some upward pH drift occurred in some studies over the course of the exposure (Section 3.1 TEST CONDITIONS, paragraph 1, page 3.1). This drift was minimized using a buffer to control the pH and in two cases slightly adjusting the CO_2 atmosphere within the test chambers. • In the same paragraph as above the authors reported that the observed range of conductivity values was wide, with values increasing as the Al exposures increased. They feel this may have been an artifact arising from the need for increased pH adjustments in the higher exposures, which required addition of HCl and/or NaOH to maintain target pH values. • Under section 3.2 DEFINITIVE TEST CONCENTRATIONS the authors noted that the total Al from post exposure solutions resulted in variability in recovery. They believe this was primarily due to the difficulty in removing a completely homogenized aliquot from the sample chambers. • In the same section as above, paragraph 2, the authors observed that a few of the dissolved Al measurements were unexpectedly elevated and did not correspond to other dissolved samples from the same concentration (shown in Tables 3-3 to 3-9 as bolded values with an asterisk *). They	

Reviewer	Comments	EPA Response to Comments
	felt that these elevated concentrations were associated with breaching of the filter (related to the fact that larger insoluble hydroxide precipitate can almost immediately clog the filter and additional pressure on the filter is necessary to obtain sufficient sample volume for analysis). To address this, pressure on the filter was kept at a minimum and new filters were used once excessive pressure was apparent. In certain cases, they felt this elevated Al may have been an artifact of the method and the large concentrations of precipitated Al in the solutions. • The authors also noted in the section above that certain dissolved Al measurements in the high DOC tests resulted in dissolved Al below detection. They felt this was due to Al binding with DOC from denser and larger particulates of insoluble hydroxides in the higher exposures (this was also observed in testing by Gensemer et al. (2018) and Cardwell et al. (2018)). Also observed in the current (and previous studies), dissolved Al concentrations did not monotonically increase as total Al increased and also could not be directly correlated with the toxic response in the organisms. To address this, total Al concentrations for determining biological effect concentrations were used in the analyses. • No anomalous effects were observed in the biological results.	
Reviewer 4	The total Al concentrations were generally close to, but a bit higher than nominal concentrations. There were some technical difficulties measuring dissolved Al that led to high variability in measured values. The authors explain this as due to problems with filtering the samples and the fact that the majority of solutions were well above the solubility limits of Al (section 3.2). For this reason, results are based on total Al, rather than dissolved Al, which makes sense. There was some degree of variability in the DOC concentrations which the authors explain in section 3.1.	
Reviewer 5	I believe that the anomalies observed during testing were well explained and the justification was sufficiently presented and plausible (page 3-4 and 3-5). The authors had issues with dissolved concentrations being considerably lower than total (and did not always follow a dose response relationship). I believe the authors adequately addressed it in their report. Since they are using measured concentrations for the expression	

Reviewer	Comments	EPA Response to Comments
	of toxicity, it is being adequately represented in the conclusions of the study.	

2.10 Do the reported test results meet or exceed the data acceptability criteria required for derivation of ambient water quality criteria for the protection of aquatic life?

Three reviewers replied that the test results met the data acceptability criteria and a fourth reviewer said the study covered a wide range of water quality parameters suitable for BLM development and calibration. One of these reviewers commented that this study appears to have been carefully planned and executed, with all tests meeting control acceptability criteria; the study extends the range and thus applicability of the previously derived models; the study results are directly applicable to the EPA-developed Water Quality Criteria; and these new data will be useful for updating the BLM and MLR models. Two of these reviewers noted, respectively, that the control acceptability criteria are consistent with EPA 2002 guidance, and that the test and resulting data meet the minimal requirements for the National Guidelines.

The fifth reviewer failed to understand the rationale for conducting these kinds of experiments because two types of concurrent exposures (dissolved aluminum exposures and precipitates) are occurring.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	I fail to understand the rationale for conducting experiments in this manner. Two types of exposures are occurring concurrently. Dissolved Al exposures and precipitates on the fish are both occurring because the solubility limit of Al is often exceeded. What are these exposures attempting to simulate in nature? Perhaps a mixing zone of some sort where a waste stream is hitting a receiving water and precipitating Al on the resident fauna?	
Reviewer 2	This study covered a wide range of water quality parameters that are suitable for BLM development and calibration. The growth data demonstrated concentration-response relationships that are useful for calculating effect concentrations based on total concentrations but not based on dissolved concentrations.	
Reviewer 3	<u>Response:</u> The reported test results appear to 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, with all tests meeting control acceptability criteria (minimum of	

Reviewer	Comments	EPA Response to Comments
	80% survival and an average dry weight of surviving fish in control chambers of > 0.25 mg; USEPA 2002). The present results appear consistent with previous work in that they can be used to validate the current Al bioavailability models (both Biotic Ligand Model and Multiple Linear Regression models). The present study extends the range and thus applicability of the previously derived models, with the effective range of pH increasing from 8.0 to 8.2, of hardness from 127 to 422 422 mg/L of $CaCO_3$, and dissolved organic carbon from 5.0 to 11.58 mg/L. I agree with the author's prediction that these new data will be useful for updating the BLM and MLR models. 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.	
Reviewer 4	Yes. In all tests, control acceptability criteria (minimum of 80% survival and an average dry weight of surviving fish in control chambers of > 0.25 mg; dissolved oxygen concentration > 60 percent saturation) were met. In addition, temperature, dissolved oxygen and concentration of the test substance were satisfactorily maintained, based on time-weighted averages, over the test period. These criteria are defined in OSU Protocol No Al-PP-CSR7d-035 (Appendix A) and are consistent with US EPA (2002) guidance.	
Reviewer 5	I believe that these test results will strengthen the aluminum ambient water quality criteria. The tests and resulting data met the minimal requirements for the National Guidelines (Stephen et al., 1985).	

2.11 Is there any reason to be concerned with the use of the test results for criteria derivation?

Three reviewers stated they do not have any concerns with using the results for criteria derivation. One of these reviewers noted that this study along with a similar study on *C. dubia* increased the understanding of bioavailability and toxicity of aluminum to aquatic organisms.

Two reviewers expressed concerns. One was concerned about the large variation of the measured dissolved aluminum concentrations, because dissolved metal concentrations are usually used for evaluating metal bioavailability with the BLM approach. Given that, this reviewer does not know how the BLM can be applied to predict the bioavailability of aluminum in this report. The fifth reviewer was concerned that there is no discussion of the mechanistic model for how toxicity is occurring in *P. promelas* (e.g., whether due to

precipitates of dissolved concentrations and speciation). Also, the testing scenario employed makes it largely impossible to evaluate what the organisms are exposed to, particularly with respect to dissolved concentrations, which makes it arbitrary what the descriptor of toxicity is.

Reviewer	Comments	EPA Response to Comments
Reviewer 1	There is no articulation of a mechanistic model of how toxicity is occurring in these animals. If toxicity is assumed to be a physical issue due to precipitates, this should be articulated. If toxicity is assumed to be due to standard BLM – gill binding/uptake related phenomena, then the dissolved concentrations and speciation become much more important. In the testing scenario employed here, it is largely impossible to evaluate what the organisms are exposed to – particularly with respect to dissolved concentrations. It then becomes arbitrary what the descriptor of toxicity is – nominal, measured new, dissolved new, dissolved old? If solid phases are contributing to toxicity, then by logical extension, the concentrations of elements in the diets of animals should also be considered in criteria development.	
Reviewer 2	The concern is the large variation of the measured dissolved Al concentrations, especially at high concentrations. Dissolved metal concentrations are usually used for evaluating metal bioavailability, especially using the BLM approach. Given that said, I don't know how the BLM can be applied to predict the bioavailability of Al in this report.	
Reviewer 3	<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 and thus increase the accuracy of toxicity predictions based on ambient water quality values. 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, in concert with a very similar study evaluating the toxic effect of aluminum on *Ceriodaphnia dubia*, has achieved these objectives and has	

Reviewer	Comments	EPA Response to Comments
	increased the applicable range of previous predictive models used to derive an Al WQC. The actual numerical values of this expansion are listed above in Charge Question #10. 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 – 12) and thus may be confidently used in the water quality criterion derivation process.	
Reviewer 4	No, I do not believe so. Overall, the test protocol has been thoroughly described, is consistent with standard US EPA guidance for chronic testing, acceptability criteria have been met, and results have been documented in detail, analyzed appropriately, and interpreted reasonably.	
Reviewer 5	I have no concerns concerning the use of the test results in the criteria derivation process.	

3.0 ADDITIONAL COMMENTS PROVIDED

Reviewer	Comments	EPA Response to Comments
Reviewer 3	The toxicity of Al markedly increases as ambient pH increases from 4.0 to 4.5 due to the change in the predominant Al speciation from the free ion form Al^{3}+ to an increased hydroxy complexing form (Schofield and Trojnar 1980). The authors never mention this in their report, probably because they never tested pH below 6.0. Nevertheless, to put this project in proper prospective of evaluating Al toxicity, I believe the above toxicity phenomenon should be mentioned in the Introduction or the Discussion and Conclusion.	
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 review:** • Second paragraph, last sentence. The definition of dissolved needs to be cleaned up. As it is written, the initial part of the sentence is referring to Al concentration and the definition	

Reviewer	Comments	EPA Response to Comments
	of dissolved is referring to the filter pore size. Suggest the following sentence. *"All concentrations are expressed in micrograms Al per litter (µg/L Al) either as total or dissolved (defined as filtrate passing through a 0.45 µm filter)."* Note: it was presented correctly on p 2-5 under Section 2.10.3 Analytical Sampling. • While I could follow the description of Section 3.2 Definitive Test Concentrations section, it is very complex and is not easy to comprehend. I believe that this section would benefit from a Figure that provides a summary of the issues encountered and how they were addressed. This would assist the reader in clearly following the issues.	

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 3	Literature Cited: Cardwell AS, WJ Adams, RW Gensemer, E Nordheim, RC Santore, AC Ryan, WA Stubblefield. 2018. 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. DeForest DK, KV Brix, LM Tear, WJ Adams. 2018. Multiple Linear Regression models for predicting aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37:80-90. 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. Gensemer, R, J Gondek, P Rodriquez, JJ Arbildua, WA Stubblefield, AS Cardwell, RC Santore, A Ryan, WJ Adams, E Nordheim. 2017. Evaluating the effects of pH, hardness, and dissolved organic carbon on the toxicity of aluminum	

Reviewer	Comments	EPA Response to Comments
	to freshwater aquatic organisms under circumneutral conditions. Environ. Toxicol. Chem. 37(1):49-60. Santore R, AC Ryan, F Kroglund, PH Rodriguez, WA Stubblefield, AS Cardwell, WJ Adams, E Nordheim. 2018. Development and application of a biotic ligand model for predicting the chronic toxicity of dissolved and precipitated aluminum to aquatic organisms. Environ. Toxicol. Chem. 37:7079. Schofield, CL and JR Trojnar. 1980. Aluminum toxicity to brook trout (Salvelinus fontinalis) in acidified waters. In: Toribara T.Y., Miller M.W., Morrow P.E. (eds) Polluted Rain. Environmental Science Research. Springer, Boston, MA.	

APPENDIX A

CHARGE TO REVIEWERS

Technical Charge to External Peer Reviewers

Contract No. EP-C-17-017
Task Order 68HE0C18F0792 (ERG Task 14)
August 2018

External Peer Review of Fish 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 fish 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 test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?
5. Were test endpoints and data acceptability criteria well defined and explained?
6. Was preparation of test solutions fully described and target test concentrations verified prior to testing?
7. Were manipulated test water quality variables (e.g., pH, DOC, water hardness) measured with sufficient frequency and accuracy to represent intended levels?
8. 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)?
9. Were any anomalies in the test explained or justified with additional information or testing?
10. Do the reported test results meet or exceed the data acceptability criteria required for derivation of ambient water quality criteria for the protection of aquatic life?
11. Is there any reason to be concerned with the use of the test results for criteria derivation?

APPENDIX B

INDIVIDUAL REVIEWER COMMENTS

COMMENTS SUBMITTED BY REVIEWER 1

External Peer Review of Short-term Chronic Toxicity of Aluminum to the Fathead Minnow, *Pimephales promelas*: 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)?**

Each of the tests were conducted with 5 concentration plus controls. This is generally considered acceptable for establishing concentration-response relationships provided adequate range finding is conducted.

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?**

Four replicates were tested per condition, with each replicate represented by 10 individuals. The standard deviations of the toxicity responses were relatively modest, suggesting that replication was adequate for these studies.

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

Generally, the source maintenance and husbandry of the test organisms was well described, however the description of the feeding rations was not adequate. Reporting a volume of a food suspension is meaningless unless we know the density of the food items in the volume of water provided to the test organisms.

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

Because the tests were initiated with larvae <24 hours old, it is not feasible to appropriately acclimate the animals. The fertilized eggs of the test animals were hatched at hardness of 100 mg/L and apparently transferred to higher hardness waters during egg development. My understanding from reading this vague text is that the eggs were transferred from hardness 100 waters to higher hardnesses for the final 4 days of development, but this should be clarified. It might have been better to rear the parents in the appropriate waters and allow the eggs to develop and hatch in the appropriate control waters.

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

Only mortality and dry mass are provided as endpoints, with the analysis focused primarily on weight.

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

Test concentrations were not verified prior to testing because the test solutions were only allowed to equilibrate for 3 hours before the initiation of the tests. All varication appears post-hoc. In one test (Al1222) measured initial concentrations are significantly higher than targeted nominal concentrations. Preparation of the exposure waters is a major issue with these tests. Waters were made, pH adjusted and allowed to equilibrate for only 3 hours. This resulted in highly dynamic exposure conditions as Al is likely precipitating

during the fish exposures. If the goal was to evaluate the physical effects of Al precipitates on larval fish, this might be appropriate, but it unclear to me how this reflects bioavailability and traditional toxicity evaluation. The difference between dissolved and total concentrations (especially comparing the "new" and 'old" dissolved concentrations is disturbing.

Table 3-1 shows no error estimates in any of the measured constituents, though error estimates are provided for pH, conductivity and DO in table 3-2.

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

Yes, pH, DOC and hardness were well monitored during the tests.

8. 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)?

No. These tests were conducted under highly variable conditions. It is completely arbitrary to use the nominal concentrations or even the mean measured "new" concentrations as descriptors of toxicity because the differences between the total and dissolved Al concentrations were extreme.

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

There were no significant anomalies in the data.

10. Do the reported test results meet or exceed the data acceptability criteria required for derivation of ambient water quality criteria for the protection of aquatic life?

I fail to understand the rationale for conducting experiments in this manner. Two types of exposures are occurring concurrently. Dissolved Al exposures and precipitates on the fish are both occurring because the solubility limit of Al is often exceeded. What are these exposures attempting to simulate in nature? Perhaps a mixing zone of some sort where a waste stream is hitting a receiving water and precipitating Al on the resident fauna?

11. Is there any reason to be concerned with the use of the test results for criteria derivation?

There is no articulation of a mechanistic model of how toxicity is occurring in these animals. If toxicity is assumed to be a physical issue due to precipitates, this should be articulated. If toxicity is assumed to be due to standard BLM – gill binding/uptake related phenomena, then the dissolved concentrations and speciation become much more important. In the testing scenario employed here, it is largely impossible to evaluate what the organisms are exposed to – particularly with respect to dissolved concentrations. It then becomes arbitrary what the descriptor of toxicity is – nominal, measured new, dissolved new, dissolved old?

If solid phases are contributing to toxicity, then by logical extension, the concentrations of elements in the diets of animals should also be considered in criteria development.

COMMENTS SUBMITTED BY REVIEWER 2

External Peer Review of Short-term Chronic Toxicity of Aluminum to the Fathead Minnow, *Pimephales promelas*: 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)?**

Five concentrations of Al and a control were used for each test. This is technically adequate for calculating LC/EC values. The design is incompliance with the USEPA guidelines for toxicology testing with aquatic organisms. Two out of the 7 tests got survival concentration-response relationship that allowed calculation of NOEC, LOEC, and LC values. All anticipated sublethal endpoints were calculated based on concentration-response relationships of the growth data.

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, 4 replicates per treatment with 10 fish per replicate were usually used for this type of test with fish.

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

Yes, the organisms were originally from Aquatic Biosystems and cultured at OSU for more than 10 years. However, due to the laboratory move, adult broodstocks were cultured at two different locations at slightly different water quality. For example, pH of 7.8-8.0 compared to 6.6-6.8 and hardness of 100-120 mg/L as CaCO3 compared to 132 mg/L as CaCO3. 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 test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?**

The organism acclimation to different hardness was described. The acclimation period was 4 days, which seems to be fine. No acclimation to different pH was mentioned.

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

Test endpoints (NOEC, LOEC, LCs, and ECs) were described in the statistical analysis section. Acceptability criteria for control survival and growth were mentioned. The results met the acceptability criteria.

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

The preparation of the test solutions was clearly described. The measured total Al were closed to the nominal concentrations but large variation between the measured dissolved concentrations and total concentrations was reported. Usually, stock concentrations are verified prior to use. However, it was not mentioned in the report. The authors mentioned that the stock concentrations were likely higher than the

target concentrations. This likely resulted in consistently higher measured total Al concentrations than the target nominal concentrations.

7. 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 the quality of the test water, such as pH was clearly described. It was conducted carefully. Concentrations of DO, pH, conductivity, and temperature were measured daily and therefore sufficient. The measured values were around the target values. However, hardness and alkalinity were measured only for control water of each test at test initiation. No description for the frequency of DOC measurement was reported. These water quality parameters are usually measured at least for 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.

8. 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)?

Concentrations of 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. However, the measured dissolved Al concentrations were largely deviated from the total concentrations. This weakens the confidence of metal analysis and biological results of the study. One of the explanations for the variation was the uncertainty in performance of the instrument at different times. This explanation doesn't sound convincing because the measured total Al concentrations seem to be fine for all treatments throughout the study.

9. 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.

10. Do the reported test results meet or exceed the data acceptability criteria required for 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. The growth data demonstrated concentration-response relationships that are useful for calculating effect concentrations based on total concentrations but not based on dissolved concentrations.

11. Is there any reason to be concerned with the use of the test results for criteria derivation?

The concern is the large variation of the measured dissolved Al concentrations, especially at high concentrations. Dissolved metal concentrations are usually used for evaluating metal bioavailability, especially using the BLM approach. Given that said, I don't know how the BLM can be applied to predict the bioavailability of Al in this report.

COMMENTS SUBMITTED BY REVIEWER 3

External Peer Review of Short-term Chronic Toxicity of Aluminum to the Fathead Minnow, *Pimephales promelas*: 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:

It appears that an adequate number and range of concentrations were used in this project to allow full characterization of the concentration response and allow determination of a scientifically-defensible chronic effect concentration.

Rationale:

The goal of this research project was to evaluate the effects of multiple water quality variables on the concentration-dependent toxicity of aluminum (Al) to the standard test vertebrate *Pimephales promelas.* The study was designed to increase the range of water quality variables under which a reasonable prediction of fish toxicity could be made under a given range of water quality variables. The test followed standard USEPA methodology (US EPA 2002). The methods included in the EPA manual are referenced in Table IA, 40 CFR Part 136 regulations and, therefore, constitute approved methods for acute toxicity tests of fish. 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 predictable negative impact mainly on growth, and to a much lesser extent, survival of *P. promelas* (Santore et al., 2018; DeForest et al., 2018; Gensemer et al., 2017) The standard EPA protocol calls for five test concentrations and a control and this was followed in the present study. The concentrations of Al used were based on historical response data with *P. promelas* in other reconstituted water (Page 2-3, paragraph 1). Five concentrations is the standard number of concentrations used by most toxicity testing laboratories, allowing 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 and assumed to be the degree of negative change from which an organism cannot recover).

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 (four per Al treatment concentration and four in the untreated control) was sufficient to allow acceptable statistical rigor for a *P. promelas* chronic toxicity evaluation under the stated test conditions.

Rationale:

Four replicate chambers (with 10 organisms in each chamber) of each toxicant concentration and the control are the numbers 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 Charge 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 four replicates.

My only question with the statistics entails the statement: "If the data met the assumptions of normality and homogeneity, the NOEC and LOEC were estimated using an Analysis of Variance…" (Page 2-6, paragraph 1). It is unclear how the authors proceeded if the data did not meet parametric assumptions.

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

Response:

No, the source, maintenance, and husbandry of the *P. promelas* test organisms were *not* adequately described.

Rationale:

In the report, section 2.3.2 SOURCE, the authors state that the <24 hour old larval fish 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 (Test System, #7) the authors state that the newly hatched larval fish were fed 0.15 mL of a Yeast/Trout Chow/Cereal leaves mixture (YTC) and algae suspension (*Pseudokirchneriella subcapitata*, 1:1), twice daily (a.m. and p.m.). I believe this is what is normally fed to *Ceriodaphnia dubia* during culture and testing, not to *P. promelas*. Later on in Appendix A (2.3 Test Diet), the authors state that brine shrimp (*Artemia*) nauplii <24 hours old were fed to the test fish. Which diet was actually fed to the test fish (I am guessing the latter)?

Also, the above two diets were stated to have been fed to the *P. promelas* during testing but not explicitly stated in the report or appendices that the test organisms were cultured and maintained under the same food regimen. 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 test organisms appropriately acclimated for the type of test and test water conditions to represent their chronic sensitivity under those conditions?

Response:

It appears that the *P. promelas* were appropriately acclimated for test conditions at the time during which the toxicity testing was performed.

Rationale:

The *P. promelas* 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 $CaCO_3$, alkalinity of 70 mg/L as $CaCO_3$, and pH of 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. For the higher hardness tests (hardness of 250 and 400 mg/L $CaCO_3$), embryos were acclimated over four days from the above described moderately hard water starting immediately after hatching. This should be sufficient time for complete acclimation.

5. 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 author's state under section 2.10.2 BIOLOGICAL MONITORING p. 2-5 that observations of live and dead fish were conducted on a daily basis from initiation to termination, and dead fish were removed immediately.

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 "$\alpha = 0.05$ ". Incidentally, I made the same statement on my review of the *Ceriodaphnia dubia* aluminum toxicity report). 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.

6. 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 t*he study but there it was not explicitly stated that the analytical equipment was tested and calibrated *prior* to the study.

Rationale:

It appears that the analytical portion of this project was very carefully performed and documented. 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. 2-1), 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), 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 in section 2.10.3 under Dissolved Metals but defined as "<0.45 **µg/L**" in Section 2.2, last sentence) 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 321 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.

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

Response:

From the report 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, conductivity, and dissolved organic carbon (DOC) were measured in each concentration at test initiation, once daily, and at test termination using a HACH HQ3od pH meter. These variables were measured in both

the replenishment water and in one test chamber just prior to replenishment. 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 just prior to the start of a test would be sufficient. The mean and raw values for the data from these analyses are presented in Tables 3-1 and 3-2 in the report, and Appendices C and D, respectively.

8. 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 toxicity test exposure. They were also measured in the renewal water prior to renewing 80% of the water in the test and control chambers (Section 2.9 TEST INITIATION, SOLUTION RENEWAL, AND FEEDING). The instrumentation used for these measurements 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.

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

Response:

All anomalous data occurred within the water quality and Al measurement results. These were few in number. They were explained/justified without the need for additional data or testing.

Rationale:

- The authors report that adult *P. promelas* broodstock were moved to a new laboratory location and reared for a period of three months in well water with a hardness of 132 mg/L as $CaCO_3$ and pH of 6.6 – 6.8. Larval fish from this adult broodstock were used for tests Al 1218 PPC, Al 1222 PPC, and Al 1225 PPC (Section 2.3.2 SOURCE, paragraph 2, page 2.1). No differences were observed between offspring from broodstock cultured in the two laboratory waters following reference toxicity testing (Section 2.3.3 ORGANISM HEALTH, paragraph 1, page 2.2).

- The authors noted high variability in measurements of dissolved organic carbon (Section 3.1 TEST CONDITIONS, paragraph 1, page 3.1). They attribute this variability to the need to use multiple batches of Suwannee River Natural Organic Matter (Suwannee NOM) which has historically been variable in DOC. They also acknowledge the possibility of observed differences being due to variability in analytical technique. However, they did not feel the observed differences were significant and reported the DOC as measured.

- Some upward pH drift occurred in some studies over the course of the exposure (Section 3.1 TEST CONDITIONS, paragraph 1, page 3.1). This drift was minimized using a buffer to control the pH and in two cases slightly adjusting the CO_2 atmosphere within the test chambers.

- In the same paragraph as above the authors reported that the observed range of conductivity values was wide, with values increasing as the Al exposures increased. They feel this may have been an artifact arising from the need for increased pH adjustments in the higher exposures, which required addition of HCl and/or NaOH to maintain target pH values.

- Under section 3.2 DEFINITIVE TEST CONCENTRATIONS the authors noted that the total Al from post exposure solutions resulted in variability in recovery. They believe this was primarily due to the difficulty in removing a completely homogenized aliquot from the sample chambers.

- In the same section as above, paragraph 2, the authors observed that a few of the dissolved Al measurements were unexpectedly elevated and did not correspond to other dissolved samples from the same concentration (shown in Tables 3-3 to 3-9 as bolded values with an asterisk *). They felt that these elevated concentrations were associated with breaching of the filter (related to the fact that larger insoluble hydroxide precipitate can almost immediately clog the filter and additional pressure on the filter is necessary to obtain sufficient sample volume for analysis). To address this, pressure on the filter was kept at a minimum and new filters were used once excessive pressure was apparent. In certain cases, they felt this elevated Al may have been an artifact of the method and the large concentrations of precipitated Al in the solutions.

- The authors also noted in the section above that certain dissolved Al measurements in the high DOC tests resulted in dissolved Al below detection. They felt this was due to Al binding with DOC from denser and larger particulates of insoluble hydroxides in the higher exposures (this was also observed in testing by Gensemer et al. (2018) and Cardwell et al. (2018)). Also observed in the current (and previous studies), dissolved Al concentrations did not monotonically increase as total Al increased and also could

not be directly correlated with the toxic response in the organisms. To address this, total Al concentrations for determining biological effect concentrations were used in the analyses.

- No anomalous effects were observed in the biological results.

10. Do the reported test results meet or exceed the data acceptability criteria required for derivation of ambient water quality criteria for the protection of aquatic life?

Response:

The reported test results appear to 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, with all tests meeting control acceptability criteria (minimum of 80% survival and an average dry weight of surviving fish in control chambers of > 0.25 mg; USEPA 2002). The present results appear consistent with previous work in that they can be used to validate the current Al bioavailability models (both Biotic Ligand Model and Multiple Linear Regression models). The present study extends the range and thus applicability of the previously derived models, with the effective range of pH increasing from 8.0 to 8.2, of hardness from 127 to 422 422 mg/L of $CaCO_3$, and dissolved organic carbon from 5.0 to 11.58 mg/L. I agree with the author's prediction that these new data will be useful for updating the BLM and MLR models.

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.

11. Is there any reason to be concerned with the use of the test results for criteria derivation?

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 and thus increase the accuracy of toxicity predictions based on ambient water quality values. 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, in concert with a very similar study evaluating the toxic effect of aluminum on *Ceriodaphnia dubia*, has achieved these objectives and has increased the applicable range of previous predictive models used to derive an Al WQC. The actual numerical values of this expansion are listed above in Charge Question #10. 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 – 12) and thus may be confidently used in the water quality criterion derivation process.

General comments:

The toxicity of Al markedly increases as ambient pH increases from 4.0 to 4.5 due to the change in the predominant Al speciation from the free ion form Al^{3+} to an increased hydroxy complexing form (Schofield and Trojnar 1980). The authors never mention this in their report, probably because they never tested pH below 6.0. Nevertheless, to put this project in proper prospective of evaluating Al toxicity, I believe the above toxicity phenomenon should be mentioned in the Introduction or the Discussion and Conclusion.

Literature cited:

Cardwell AS, WJ Adams, RW Gensemer, E Nordheim, RC Santore, AC Ryan, WA Stubblefield. 2018. 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.

DeForest DK, KV Brix, LM Tear, WJ Adams. 2018. Multiple Linear Regression models for predicting aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37:80-90.

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.

Gensemer, R, J Gondek, P Rodriquez, JJ Arbildua, WA Stubblefield, AS Cardwell, RC Santore, A Ryan, WJ Adams, E Nordheim. 2017. 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(1):49-60.

Santore R, AC Ryan, F Kroglund, PH Rodriguez, WA Stubblefield, AS Cardwell, WJ Adams, E Nordheim. 2018. Development and application of a biotic ligand model for predicting the chronic toxicity of dissolved and precipitated aluminum to aquatic organisms. Environ. Toxicol. Chem. 37:7079.

Schofield, CL and JR Trojnar. 1980. Aluminum toxicity to brook trout (Salvelinus fontinalis) in acidified waters. In: Toribara T.Y., Miller M.W., Morrow P.E. (eds) Polluted Rain. Environmental Science Research. Springer, Boston, MA.

COMMENTS SUBMITTED BY REVIEWER 4

External Peer Review of Short-term Chronic Toxicity of Aluminum to the Fathead Minnow, *Pimephales promelas*: 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) with modifications for testing with Al. 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 (section 2.5), it is judged to be adequate for the test purpose. The range of concentrations was chosen on the basis of preliminary results and by putting nominal water quality characteristics into the bioavailability models to predict effects (section 2-5). This seems a reasonable approach. In 7 of 7 tests it was possible to estimate chronic effect concentrations (NOEC, LOEC, EC_{10}, EC_{20} and EC_{50}; Table 3-11) for growth. For 5 of the 7 tests, there was no dose-response for survival, and no EC_x values could be estimated (Table 3-11).

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?

There were four replicates per test concentration. According to US EPA (2002, section 12.10.2.1), this is the recommended number of replicates for this kind of test.

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

The source of the fish was clearly described. Fish were obtained from in-house cultures, and their original source was from Aquatic Biosystems (Fort Collins, CO, USA; section 2.3.2). The maintenance and husbandry were partly described. The culture water was described in detail (section 2.3.2), however I could find no other details on the maintenance or husbandry of the test organisms. I also checked the OSU Protocol No. Al-PP-CSR7d-035, provided as Appendix A, but could not find details there either. As the species has been cultured in house for many generations, and the fish were determined to be in good health prior to testing (as described in section 2.3.3), it can probably be assumed that maintenance and husbandry conditions were adequate.

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

For all of the tests, the larvae were hatched in moderately hard, reconstituted lab water. For 5 of the tests, the larvae were kept in this water until test initiation; for the 2 higher hardness tests, larvae were acclimated to the higher hardness test water (250 and 400 mg/L as $CaCO_3$) for 4 days after hatching (section 2.3.4). Since criteria for determining when an organism is actually acclimated are rarely defined, it is difficult to say whether 4 days was sufficient. There is no further mention of acclimation in the report and therefore assumed that other conditions of the test (e.g., temperature, light regime, food) were similar between cultures and test conditions.

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

Test endpoints were survival and growth. According to OSU Protocol No Al-PP-CSR7d-035, death was defined as the lack of movement in response to gentle prodding (Protocol section 4.6). Growth was estimated as mean dry biomass at the end of the test (i.e., total dry weight of surviving organisms divided by the original number of organisms at test initiation; section 3.3). Quality criteria for the test are explicitly defined in section 4.9 of the Protocol (Appendix A).

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

Yes. Preparation of test solutions is described in detail in section 2.5, and both total and dissolved Al were measured at test initiation as described in section 2.10. 3. Data verifying target test concentrations are provided in Tables 3.3 – 3.8.

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

Yes. These were measured at test initiation, once daily and at test termination in both "new" and "old" water as described in section 2.10.1. Data verifying that water quality variables were sufficiently maintained are provided in Table 3-2.

8. 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. Total Al was measured in each treatment in newly prepared waters ("new") at test initiation, twice during the tests, and from a composite of replicates at test termination ("old"). Dissolved Al (< 0.45 µm) was similarly measured at test initiation and termination, but only once during the tests. Detailed results of the metal analyses are provided in Appendix B.

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

The total Al concentrations were generally close to, but a bit higher than nominal concentrations. There were some technical difficulties measuring dissolved Al that led to high variability in measured values. The authors explain this as due to problems with filtering the samples and the fact that the majority of solutions were well above the solubility limits of Al (section 3.2). For this reason, results are based on total Al, rather than dissolved Al, which makes sense. There was some degree of variability in the DOC concentrations which the authors explain in section 3.1.

10. Do the reported test results meet or exceed the data acceptability criteria required for derivation of ambient water quality criteria for the protection of aquatic life?

Yes. In all tests, control acceptability criteria (minimum of 80% survival and an average dry weight of surviving fish in control chambers of > 0.25 mg; dissolved oxygen concentration > 60 percent saturation) were met. In addition, temperature, dissolved oxygen and concentration of the test substance were

satisfactorily maintained, based on time-weighted averages, over the test period. These criteria are defined in OSU Protocol No Al-PP-CSR7d-035 (Appendix A) and are consistent with US EPA (2002) guidance.

11. Is there any reason to be concerned with the use of the test results for criteria derivation?

No, I do not believe so. Overall, the test protocol has been thoroughly described, is consistent with standard US EPA guidance for chronic testing, acceptability criteria have been met, and results have been documented in detail, analyzed appropriately, and interpreted reasonably.

COMMENTS SUBMITTED BY REVIEWER 5

External Peer Review of Short-term Chronic Toxicity of Aluminum to the Fathead Minnow, *Pimephales promelas*: 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. One challenge was in the middle of the testing program that laboratory was moved from one location to another. I believe that the PI and Study Coordinator adequately evaluated potential difference in the culturing and resulting testing by additional quality control procedures that adequately assessed that there was no differences. Each test was performed with five treatments and a control, with four replicates in a random arrangement. This procedure follows standard EPA test procedures. Control survival was acceptable in all testing. One issue that occurred during testing was that the dissolved aluminum concentrations were considerably lower than the total aluminum concentration. I believe the study team adequately addressed this issue in the interpretation of the study 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 (four) and test concentrations (minimally five plus a control) were standard with in ecotoxicity testing with *Pimephales promelas*. Testing was also performed in a randomized manner concerning treatment and replicate placement. 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 to ensure that the fathead minnow were health and consistent in their toxicological response.

4. **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. In addition, I appreciate the use of non-metal chelating buffers and the CO_2 headspace procedures to control acclimation and testing pH in this study. The researches should be commended on this practice.

5. 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. The authors had issues with dissolved concentrations being considerably lower than total (and this did not always follow a dose response relationship). I believe the authors adequately addressed it in their report. Since they are using measured concentrations for the expression of toxicity, it is being adequately represented in the conclusions.

6. 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.

7. 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 the CO_2 headspace technique, were warranted for keeping these tight conditions concerning the challenging pH parameters used in this testing program.

8. 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. The authors had issues with dissolved concentrations being considerably lower than total (and they did not always follow a dose response relationship). I believe the authors adequately addressed it in their report. Since they are using measured concentrations for the expression of toxicity, it is being adequately represented in the conclusions of this study.

(See next charge question for additional input to this charge question.)

9. 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 and 3-5). The authors had issues with dissolved concentrations being considerably lower than total (and did not always follow a dose response relationship). I believe the authors adequately addressed it in their report. Since they are using measured concentrations for the expression of toxicity, it is being adequately represented in the conclusions of the study.

10. Do the reported test results meet or exceed expectations for the data acceptability required for the derivation of ambient water quality criteria for the protection of aquatic life?

I believe that these test results will strengthen the aluminum ambient water quality criteria. The tests and resulting data met the minimal requirements for the National Guidelines (Stephen et al., 1985).

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

I have no concerns concerning the use of the test results in the criteria derivation process.

Specific comments from review:

- Second paragraph, last sentence. The definition of dissolved needs to be cleaned up. As it is written, the initial part of the sentence is referring to Al concentration and the definition of dissolved is referring to the filter pore size. Suggest the following sentence. *"All concentrations are expressed in micrograms Al per litter (µg/L Al) either as total or dissolved (defined as filtrate passing through a 0.45 µm filter)."* Note: it was presented correctly on p 2-5 under Section 2.10.3 Analytical Sampling.
- While I could follow the description of Section 3.2 Definitive Test Concentrations section, it is very complex and is not easy to comprehend. I believe that this section would benefit from a Figure that provides a summary of the issues encountered and how they were addressed. This would assist the reader in clearly following the issues.