PEER REVIEW REPORT

External Letter Peer Review for Aluminum Criteria Model

Peer Reviewers:

Tham C. Hoang, Ph.D.
Christopher Mebane, Ph.D.
Gregory Möeller, Ph.D.
Scott Smith, Ph.D.
John D. Stark, Ph.D.

Contract No. EP-C-17-023
Task Order 68HE0C18F0815

Prepared for:

U.S. Environmental Protection Agency
Office of Water, Office of Science and Technology
1200 Pennsylvania Ave., NW
Washington, DC 20460
Attn: Diana Eignor

Prepared by:

Versar, Inc.
6850 Versar Center
Springfield, VA 22151

November 2, 2018

TABLE OF CONTENTS

I. INTRODUCTION

The U.S. 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. The draft aluminum criteria model is being updated and there will be new multilinear equations included, a plant module will be included plus some other additions.

EPA is undertaking this task to obtain a focused, objective evaluation through external peer review of the aluminum model used to determine aquatic life criteria.

Versar selected the following five scientific experts to serve as peer reviewers:

Peer Reviewers:

Tham C. Hoang, Ph.D.
Loyola University Chicago
Chicago, IL 60660

Christopher Mebane, Ph.D.
U.S. Geological Survey
Idaho Water Science Center
Boise, ID

Gregory Möller, Ph.D.
University of Idaho-Washington State University School of Food Science
Environmental Sciences Program
Moscow, ID

Scott Smith, Ph.D.
Wilfrid Laurier University
Waterloo, ON,
Canada

John D. Stark, Ph.D.
Washington State University
Washington Stormwater Center
Puyallup Research and Extension Center
Puyallup, WA

II. CHARGE TO PEER REVIEWERS

1. Please review the DeForest *et al.* 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for *Ceriodaphnia dubia* and *Pimephales promelas*" dated 8/24/18.

- Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?
- Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.
- Does the pooled model behave similarly as the non-pooled models?

2. Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

- Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.
- Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.
- Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?
- Please provide suggestions of alternate approaches, if any.

3. Ease of Use:

- Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?
- Do you have any other suggestions to improve the ease of use?

III. PEER REVIEWER COMMENTS BY CHARGE QUESTIONS

Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	Prior to agreeing to conduct this review, I have been working on an NAS panel on an update of the 2015 EPA Multi-Sector General Stormwater Permit (MSGP). Because aluminum is a stormwater benchmark monitoring requirement for some of the sectors in this permit, I have familiarized myself with the original aquatic life criteria developed for aluminum (1988). I have also briefly looked over the 2017 draft document. I therefore appreciate the difficulty of working with metal toxicity and risk assessments for aquatic ecosystems. As pointed out in the Deforest memorandum and other papers (see the special edition of ET&C 37(1) 2018 for a number of papers dealing with aluminum toxicity), including the 2017 draft, the editorial by Adams et al. 2018 (ET&C 37(1) 34–35, aluminum toxicity is dependent upon water quality characteristics (pH, hardness, DOC), not unlike other metals, including copper and zinc. The Biotic Ligand model has been used in the past but it is difficult to use. I found that the multiple linear regression (MLR) model approach outlined in the Deforest memorandum is well-thought out. I am particularly impressed with the Calculator as it produce excellent results and is easy to use. The additional studies (new toxicity data since the original ALC in 1988) included in this document are of great value as they increased all of the R^2 values. The MLR model is a great improvement over past models because it incorporates pH, DOC, and hardness as these values relate to bioavailability and hence toxicity. The MLR can be used to normalize acute and chronic toxicity data to a set of predetermined water quality conditions. The MLR was also used to determine what water quality parameters are of value and which are not as important in terms of R^2. Furthermore, the authors determined that a pooled MLR model had higher adjusted and predicted R^2 values compared to the species specific models. This conclusion was justified by the results of the individual and pooled models. I agree that the results of these models indicate that the pooled model should be used in place of individual models.	
2	I have reviewed the documents provided by Versar that are presented in the below Table. An updated version of the Memorandum was provided on September 12. The Al criteria presented in these documents was developed based on multiple linear regression	

Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	model approach. Two MLR criteria models were developed. One is for individual species (non-pooled model) and the other is for a combination of 2 species of C. dubia and P. promelas (pooled model). The model development was clearly described in DeForest et al. 2018 paper. The Memorandum presented an update to the models of DeForest et al. 2018 at which, new data for C. dubia and P. promelas were used for calculation of the model coefficients (slopes). A pooled model that combined data for C. dubia and P. promelas was also presented in the Memorandum. The provided scenarios of data that had a pH range of 5-9, a DOC range of 0.5-10 mg/L, and a hardness range of 25-400 mg/L as CaCO3 were used to run the models and calculate the CMC and CCC values. A relative site-by-site comparison of the CMC and CCC values of the pooled and non-pooled models was conducted by calculating the ratio of the CMC and CCC values predicted by the pooled model to those predicted by the non-pooled model (Fig A and B). Below are some general comments for the model development and performance. Some of these comments will be further discussed and presented in the answers to the charge questions. • The MLR model approach is for sure easier to use than the Biotic Ligand Model approach. However, the BLM takes metal speciation and bioavailability into account and can be applied for various environmental conditions. The MLR is a statistical approach and its application is logically limited the range of environmental conditions that was used for model development. Most of the data used for the model development were coming from laboratory research that used formulated water which is cleaner and less extreme than field waters. Given the complicated chemistry of Al, especially in different pH conditions, I am not sure how well the MLR model prediction will represent the natural environment. • The current data (including the addition of the new data set) don't seem to be strong for a multiple regression analysis that get involved with at least 3 variables and interaction terms between them including a quadratic term, such as for pH (pH*pH). When such regression models are developed, data of factorial design experiments are more suitable for use. The limitation of data used for the model development might end up with a model that is less	

Table 1. General Impressions

REVIEWER NO.	REVIEWER COMMENT	EPA RESPONSE

representative and hence less accurate prediction, especially for cases that the data are outside or at the boundary of the current range and for other species rather than the two species used for the model calibration.

- There are advantages and disadvantages between the pooled and non-pooled models. The non-pooled model clearly distinguish the dependence of Al toxicity on water quality. For examples, quadric model for pH and P. subcapitata and C. dubia but linear for P. promelas. The pooled model combined C. dubia and P. promelas data and likely excluded the quadratic term. This might make the model be biased to P. promelas. Since data for other fish species are not sufficient and the dependence of Al toxicity on pH for other fish species is unknown, the current pooled model might not be representative. The conclusion of using the pooled model instead of non-pooled model for predicting Al criteria is less convincing. The pooled model predictions are much higher than the non-pooled model predictions for low and high pH cases. This doesn't sound that the pooled model criteria is protective although it is more convenient and preclude the need to recalculate genus species distribution.
- Given the MLR criteria- a statistical approach, 95% confidence intervals can be used instead of the acceptable prediction of 2-fold above and below the perfect prediction that has been used by the BLM approach.

File Name	Description
MLR Model_Pooled Slopes_Aluminum Criteria Calculator_8.29.18.xlsm	Pooled Slopes Aluminum Calculator
MLR Model_Individual_Slopes_Aluminum Criteria Calculator_8.29.18.xlsm	Individual Slopes Aluminum Calculator
Appendix A 9-5-18.xlsx	Appendix A file is to be used to check models for charge question #2
DeForest_et_al-2018-Environmental_Toxicology_and_Chemistry.pdf	DeForest et al. 2018 Paper
DeForest Aluminum MLR Models Update Memo (2018-08-24).pdf	DeForest Memo to EPA

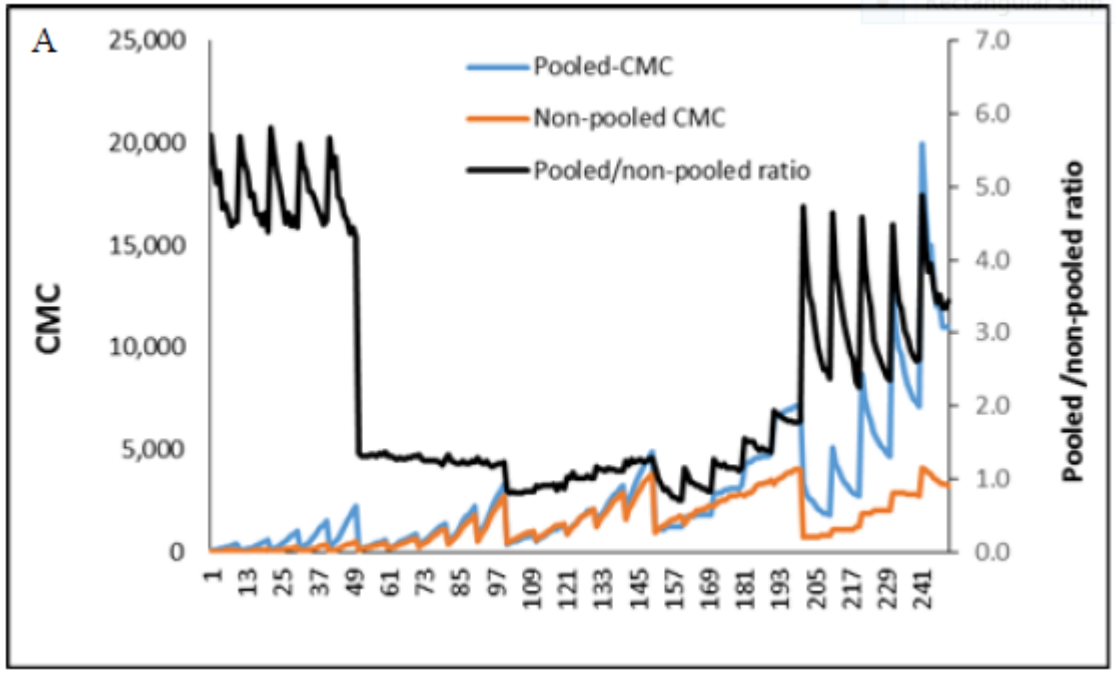
Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
		

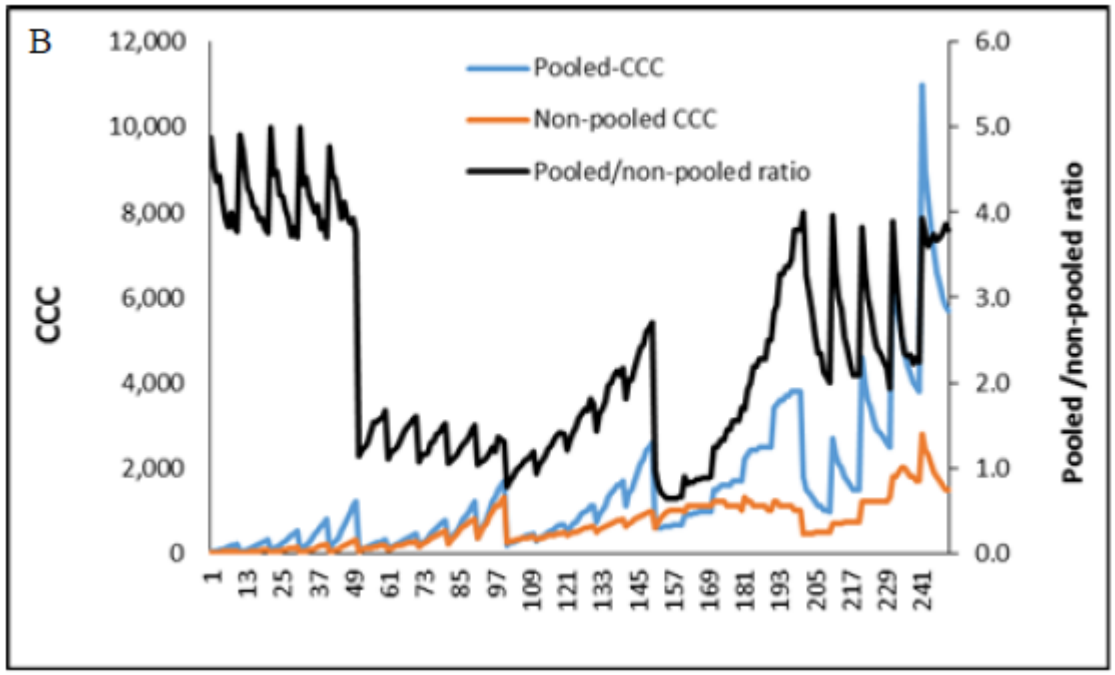
Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
		
3	It is clear that the scope of this review is to evaluate different possible aluminum criteria calculators (excel spreadsheets) all based on multiple linear regression (MLR). The primary purpose of this review is to evaluate and provide written comments on EPA's Aluminum Criteria Calculator/Model and answer three charge questions. The focus of the review is on two Excel spreadsheets with multiple tabs that contain the aluminum model. A user's guide is included in the Excel spreadsheets as a ReadMe tab. The starting place for this MLR process is the recent DeForest et al. (2017) paper along with more recent data and revised MLR models (memo from DeForest et al., 2018). From these MLR models, which predict ECx concentrations as a function of pH, hardness and DOC, spreadsheets were built to predict effect concentrations as a function of those 3 water chemistry variables and convert them to CCC and Criterion	

Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	Maximum Concentration (CMC) for use by stake holders. Spreadsheets were built using old and new data (the old data spreadsheet is already available online, the new spreadsheets are what are being evaluated here). The new data spreadsheets include either pooled or non-pooled versions. The initial impression of the proposed Criteria Calculator is that it was a good choice to use the familiar Excel software platform. Essentially all potential end-users (scientists, consultants, permit writers, …) will be familiar with Excel. This comfortable environment is a good choice for this tool. These models are designed for ease of use, using the common and familiar excel interface, and have been designed with the end user in mind. There is excellent transparency in how easy it is to find the underlying MLR equations within the spreadsheet, as well as seeing all the effects data that are used in the original MLR modelling. The information presented is accurate (the spreadsheets seem to apply the DeForest equations correctly) and for the most part presented clearly (see some exceptions below). In terms of soundness of conclusions, there were no conclusions to evaluate. Just the software tools.	
4	The use of multiple linear regression (MLRs) in metals criteria is an important step for translating the advances of biotic ligand modeling (BLMs) and related bioavailability research into functional criteria. Particularly with aluminum, they are a huge step forward from the old pH groups and can be both predictive of toxicity when exceeded, and protective of aquatic life uses when met. EPA has successfully used nonlinear regressions for many years with their ammonia criteria, and the educated public (i.e., dischargers, regulators) should have no problem working with these. The new toxicity dataset development and comprehensive data reduction and modeling are exemplary and hopefully harbingers for approaches with other outdated criteria. This review focused on comparing the performance of two MLR models. The outputs of the two models were often dissimilar, which was not expected. Comparisons with BLM outputs and other comparisons of MLR outputs with test calculations and natural	

Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	waters suggested that the individual or "non-pooled" MLR models has the better performance of the two. It was not clear that the pooled model would be as protective as intended by the guidelines for developing water quality criteria. Unfortunately, the severely compressed review schedule and my overlapping field work prevented a more in-depth review of the underlying math, and precluded taking time to ask the developers if I was interpreting and using the model correctly. Some of my criticisms could well be off the mark owing to the haste of this review. I did see the 12 September 2018 email that there was a correction to the memo and model, but with my overlapping field work and the long processing times to run the model, I did not have opportunity to go back and repeat my analyses before the 20 September 2018 deadline.	
5	The work is a very well-executed model development based on a highly-screened aquatic toxicity dataset that offers a significant advancement in environmental risk assessment of aluminum in freshwater. The authors of the DeForest et al. 2018 paper and the subsequent peer-reviewed citations represent experienced and qualified experts in the related fields. The enlarged dataset offered in the work of the OSU Aquatic Toxicology Lab has appropriately increased the value and usefulness of the MLR approach, and furthermore allows defendable pooled MLRs. The approach and dataset presented are peer-reviewed and represent our best available knowledge moving forward to update and improve the current three-decade-old approach to quantifying aluminum risk in aquatic ecosystems. The papers, data, and technical memorandum used in the supporting material present a convincing case for moving forward. Although the actual model spreadsheet would be improved with better notation and comments fields for novice users, and a much better effort at user guidance, the overall MLR model appears well developed. The model spreadsheet supporting documentation needs work before general distribution since the user base is less than familiar with this approach. The Readme appears written by experts for an audience of users with similar expertise and that is most often not the case at the state regulatory level, especially in smaller states. General	

Table 1. General Impressions		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	release of the criteria calculating model with its present level of documentation may lead to confusion and frustration with many users. The guidance for this review was somewhat challenging as well. For example the use of "Non-pooled" and "Individual" for the same thing was confusing. The models pre-loaded with scenarios was also somewhat mysterious at first, because I would assume you want the user base to fill in water quality scenarios of concern and run the model for specific results related to their management concerns. The Pooled Model does not appear to produce results consistent with the output of Non-pooled Model when comparing a side-by-side scenario data set. Hence, unless there is a reason for the rather large non-concordance of the two output sets, possibly due to user error, the Pooled Model would not be appropriate for use and appears to be generally overprotective.	

Table 2: Charge Question 1 ***Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.*** **Charge Question 1a.** ***Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?***		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	Yes. In fact, results of these MLR equations show that the addition of the new toxicity data improve the models.	
2	Yes, the MLR models developed by DeForest et al. 2018 are basically statistical models. Therefore, the models will be more confident if more data are used for model calibration. The Memorandum mentioned the improvement (higher R2 values) when new data set was included. In addition, the new data set covered a wider range of water quality parameters. Therefore, the updated models logically can be used to predict the toxicity of Al for a wider range of water quality, such as hardness, pH, and DOC.	
3	Yes it is appropriate to include the new toxicity data in the MLR equation. The original DeForest paper specifically mentions that data expanding the range of pH, DOC and hardness would be required to use the model for parameters outside the calibration range. A limitation of MLR models, because they are empirical, is that you cannot use them for waters outside the calibration range. Expanding the calibration range is exactly appropriate. Examination of Figures 1-4 in the DeForest memorandum clearly show that effect concentration predictions only negligibly change with this added data.	
4	Yes. The new toxicity data fills gaps in the tested water quality conditions that were lacking earlier.	

Table 2: Charge Question 1

Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.

Charge Question 1a.

Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?

REVIEWER NO.	REVIEWER COMMENT	EPA RESPONSE
5	The DeForest et al. 2018 ETC paper is the most comprehensive attempt at developing a model of the aquatic toxicity of aluminum in three decades. The paper develops a multiple linear regression model based on DOC, pH, and hardness conditions that are derived from a robust, screened aquatic toxicity data set. The regression analysis was on data from P. subcapitata, C. dubia, and P. promelas. The predictive MLR model demonstrated the ability to predict chronic toxicity with variable DOC, pH, and hardness conditions within a factor of two for 91% of the tests explored. There have been four citations of this paper in the very short period since its publication – achieving a highly cited notation. However, most of these have one of the authors as a co-author, and two contain the additional Al aquatic toxicity data of Gensemer et al. The additional co-authors on these papers as well as their publication in the leading journals in the field suggest the research is if the highest quality. The MLR approach thus demonstrates in this peer-reviewed paper, its viability for use in a regulatory science arena related to risk management of the freshwater aquatic toxicity of aluminum. It is appropriate and necessary to integrate the new toxicity data into the MLR equations. The OSU Aquatic Toxicology Lab data completes and enhances the MLR robustness specifically because of the targeted test quality and range of water quality conditions of the data set. The regulatory science community is fortunate that this data set became available during the review phase of the 2017 Draft Aquatic Life Criteria for Aluminum in Freshwater. As demonstrated	

Table 2: Charge Question 1 *Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.* **Charge Question 1a.** *Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?*		
REVIEWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	in the September 12, 2018, updated August 24, 2018, Memorandum, Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas, the integration of the new toxicity data expands the DOC, pH and hardness ranges where the MLR can be reliably used.	

Table 3: Charge Question 1b. *Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	All of the MLRs are appropriately parameterized. I would not add anything to the model inputs. However, it was interesting to me that the ln(DOC) x pH term was excluded in the *C. dubia* model but retained in the *P. promelas* model. As a modeler, I have encountered scenarios like this in the past. Sometimes, this is just a matter of inadequate data sets.	
2	The idea of combining fish and invertebrate data to develop a pooled model sounds reasonable because the model then can be used for predicting toxicity for both fish and invertebrate. However, it is not clear to me on how the sensitivity of each species was quantitatively taken into account. The Memorandum did mention that a species term and terms for each of the independent variables and their interactions were included in the pooled model but I don't see them in the results and conclusion. Equations 5 to 8 are separately for C. dubia and P. promelas. No slope for species term and intercept value was presented for the pooled models on page 6 of the Memorandum.	
3	The MLR method in the original DeForest paper is mathematically and scientifically sound. The parameters for both models were derived from this method so yes the parameters are sound. It is a limitation of empirical models that there is no theoretical basis for the values of the parameters so there is no theory to compare the values to. For this approach it is sufficient that the data points are described by the MLR parameters in a statistically best sense.	
4	It's hard to say with confidence. Certainly, in the DeForest and others' update memo, the pooled model performs very well fitting the Ceriodaphnia and fathead minnow data. However, in comparisons between the pooled model, the non-pooled model, and the aluminum BLM (Santore et al. 2018), the outputs were sometime quite different. Conceptually, these patterns should be similar between the models. They weren't. Unfortunately, in this type of	

Table 3: Charge Question 1b. ***Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.***		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	comparison, while the comparisons are reassuring when they are similar, when they are dissimilar it is not obvious why or which model is more believable. However, some aspects of the pooled MLR do seem amiss, with the flat response for hardness and a much greater magnitude of change for the DOC than for the individual slopes MLR or the BLM. Generally, the performance looks better for the non-pooled model, but that would have to be weighed against any advantage of reduced complexity and possibly better response from stakeholders for the pooled model.	
5	The pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized. The published DeForest et al. 2018 paper, and the subsequent works that cite this paper, develop a significant level of background in the peer-reviewed literature about the dominant water quality characteristics influencing aluminum aquatic toxicity. In the MLRs, ln(DOC), pH, and ln(Hard) are used in a common and defendable manner to define probability distributions in the scope of this risk assessment. The ground-truthing of the model with toxicity testing results suggests robustness. *"…the updated dataset supported development of a pooled MLR model that had comparably high adjusted and predicted R2 values compared to the species-specific MLR models. The pooled models also provided a similar level of accuracy in predicted EC10s and EC20s compared to the species-specific models*."	

Table 4: Charge Question 1c. *Does the pooled model behave similarly as the non-pooled models?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	Yes. The pooled model does behave similarly to the non-pooled models. In fact, the R^2 were somewhat higher of the pooled model compared to the individual models. A strong case is made by DeForest et al. 2018, for the use of the pooled model over the use of the individual models.	
2	The predictions of the two models for various scenarios showed a similar trend (Fig A and B) but relatively the predictions of the two models at low and high pH are about 5 time different as discussed above.	
3	Yes. There are three attached figures at the end of this document that demonstrate the same behavior of the pooled and non-pooled models (Figures 1 to 3). The individual (non-pooled) model and the pooled model both show protection (increasing EC20) as DOC increases and hardness increases for all 3 pHs plotted. *C. Dubia* was used as the example for these calculations. There are differences between the two models. The pooled model tends to show lower effect concentrations but the relative differences are never more than a factor of 2 and this only occurs at extremely low hardness values. The differences tend to be much smaller than that. More significantly it can be seen that by plotting the data used to calibrate the model (blue dots on Figures 1-3) the data and the model agree, although the pooled data does not agree as well as the individual data. This is to be expected because the pooled data has to satisfy more points simultaneously. The agreement between pooled and individual ECx predictions is also clearly shown by the four figures in the DeForest memo as mentioned in comment 1(a) above.	
4	Sometimes it is similar, but at other times the models are quite different. I looked at the patterns between the models in several ways – comparing to each other and the BLM (Figure 1), comparing their patterns in natural waters (Figure 2), comparing	

Table 4: Charge Question 1c. *Does the pooled model behave similarly as the non-pooled models?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	their performance with the test values provided here (Figure 3) and comparing back to the Ceriodaphnia toxicity data. *[Note: Figures are included in the section following Table 11 entitled "Five Figures and Reference from Reviewer 4's Review"]*	
5	No, see Question 2 results below. When the conditions of Appendix A are copied into fields C, D, and E the CMC and CCC results generated in columns H and I for the Non-Pooled and Pooled models are quite different. The model authors state in their technical memoranda: *"…the updated dataset supported development of a pooled MLR model that had comparably high adjusted and predicted R2 values compared to the species-specific MLR models. The pooled models also provided a similar level of accuracy in predicted EC10s and EC20s compared to the species-specific models*." *"The pooled aluminum MLR models provided a similar level of accuracy in EC10 and EC20 predictions for C. dubia and P. promelas as the species-specific MLR models. For C. dubia, the percentage of predicted EC10s and EC20s within a factor of two of observed was unchanged (94% and 97%, respectively) (Figure 3). For P. promelas, the percentage of predicted EC10s and EC20s within a factor of two of observed decreased from 94% to 90% for EC10s and from 97% to 94% for EC20s (Figure 4)."* *"Because the pooled MLR model performs well, there no longer appears to be any benefit in using species-specific MLR models for ambient water quality criteria development.(my emphasis) Use of the pooled model would preclude the need to recalculate the aluminum genus sensitivity*	

Table 4: Charge Question 1c. *Does the pooled model behave similarly as the non-pooled models?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**

	distribution for each water chemistry of interest. Instead, chronic aluminum criteria could be condensed to a single equation, such as the existing hardness-based criteria for several metals or the pooled MLR-based criteria for copper described in Brix et al. (2017). The slopes from the recommended pooled models are:
	
	• *Pooled slopes from EC10 model:*
	  ○ *ln(DOC) = 0.645*
	  ○ *pH = 1.995*
	  ○ *ln(Hard) = 2.255*
	  ○ *ln(Hard)×pH = -0.284*
	• *Pooled slopes from EC20 model:*
	  ○ *ln(DOC) = 0.592*
	  ○ *pH = 1.998*
	  ○ *ln(Hard) = 2.188*
	  ○ *ln(Hard)×pH = -0.268"*
	
	C. dubia
	$ln(EC10) = -8.618 + 0.645 \times ln[DOC] + 1.995 \times pH + 2.255 \times ln[Hard] - 0.284 \times ln[Hard] \times pH$ *(5)*
	
	$ln(EC20) = -8.555 + 0.592 \times ln[DOC] + 1.998 \times pH + 2.188 \times ln[Hard] - 0.268 \times ln[Hard] \times pH$ *(6)*
	
	P. promelas
	$ln(EC10) = -7.606 + 0.645 \times ln[DOC] + 1.995 \times pH + 2.255 \times ln[Hard] - 0.284 \times$	

Table 4: Charge Question 1c. *Does the pooled model behave similarly as the non-pooled models?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	ln[Hard] × pH *(7)* *ln(EC20) = -7.500 + 0.592 × ln[DOC] + 1.998 × pH + 2.188 × ln[Hard] – 0.268 × ln[Hard] × pH* *(8)* In these analyses, the authors appear to successfully defend use of a pooled MLR model in large part due to the expanded OSU data set made available in 2018. However, when same pH, DOC and Hardness field scenarios are loaded into the Non-pooled and Pooled models, the CMC and CCC results appear considerably different (see #2 below).	

Table 5: Charge Question 2 ***Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.*** **Charge Question 2a.** ***Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.***		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	I compared the resulted of the non-pooled to the pooled results and found that the pooled results were similar to the individual results. The Criterion Maximum Concentration (CMC) is the highest concentration of a chemical in water that aquatic organisms can be exposed to acutely without causing an adverse effect. The Criterion Continuous Concentration (CCC) is the highest concentration of a chemical in water that aquatic organisms can be exposed to indefinitely without resulting in an adverse effect. The CMC is usually higher than the CCC and this is exactly what the MLR models predict.	

Table 5: Charge Question 2 ***Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.*** **Charge Question 2a.** ***Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.***		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
2	The predicted CMC and CCC values by the pooled and non-pooled models were plotted in Fig. A and B above. The first 50 data points are for pH 5 scenarios. The last 50 data points are for pH 9 scenarios. The ratio of the pooled to non-pooled CMC and CCC values were also plotted. It can be seen that the model predictions are not the same across the pH values and more pH dependent. At pH 5 and 9, the predicted CMC and CCC values by the pooled model were approximately 5 times higher than those by the non-pooled model. Both models seem to give similar predicted CMC and CCC values at pH between 6 and 8 (ratio ~ 1). This pH range captures most pH data used to develop the models (few data points with pH between 5 and 6). Outside of this pH range, especially at pH 5 and 9, the predictions are likely extrapolated because no pH 5 and 9 was used for model calibration. Therefore, the predictions might not be confident at these pH conditions.	
3	Results of the side by side modelling are presented in the attached Figures 4 to 7. *[Note: Figures are included in the section following Table 11 entitled "Seven Figures from Reviewer 3's Review"]* Figure 4 demonstrates that the pooled spreadsheet often estimates higher CMC and CCC. It is unclear why Appendix A data were selected for this exercise though. Much of the pHs are outside the calibration range of the MLR. Unlike a mechanistic approach like a BLM, MLR cannot be extrapolated outside the calibration range. I am not clear on how this outside the range data was handled in the calculations. At one point in the instructions it just says it is flagged – but it was not when I ran the spreadsheet. It seems the flag might only work when DOC is too high? Later in the "read me" tab it says the excel model will default to the maximum recommended conditions when parameters are outside the range. I do not know if this was done, or exactly what this means. For parameters outside the range, are they just flagged? Or is the computational approach modified in some way. Some clarity is needed.	

Table 5: Charge Question 2 ***Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.*** **Charge Question 2a.** ***Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.***		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	In addition the documentation (read me) tab says that the range goes to pH of 9, but the DeForest memo states 8.1 is the calibration range. pH is of course on a log scale so 8 and 9 are an order of magnitude different. If we focus on the data that is within the calibration range of DeForest's proposed equations the pooled and individual results are very similar (Figure 4 and 5 below) and cluster around the one to one line. The tendency is that at low DOC the pooled results are lower and for high DOC the pooled results are higher.	
4	The combinations of pH, DOC, and hardness values provided in Appendix A is a similar type of evaluation as that I used with the BLM responses in Figure 1. In Figure 2, the best agreement is with the water quality conditions most commonly represented in the datasets and used to develop the models (pH 6-7 and pH 8 at low DOC), so agreement in this range is expected. The magnitude of difference between the models is substantial in some circumstances. For instance, with DOC the non-pooled model has toxicity sharply reduced (exponential increase in CCC) as DOC increases from 0.1 to about 2 mg/L, followed by a reduction in slope and slow increases. The non-pooled values steadily and steeply increase (Figure 1). The non-pooled CCC is about 500 µg/L by 2 mg/L DOC and only increases to 700 by 12 mg/L DOC. In contrast for the same values (2 and 12 mg/L DOC) the pooled model predicts much higher values, 900 and 2600 µg/L. The BLM predicts a linear reduction in toxicity (that is, a linear increase to the EC20 values) over this same range but the absolute values are much lower, about 70 to 250 µg/L for DOCs of 2 and 12 respectively (Figure 1). Granted it's not completely correct to compare CCC and Ceriodaphnia responses, but Ceriodaphnia are reasonably sensitive for the dataset (4th out 13 taxa) their EC20s should be slightly higher than the CCC for the same conditions. In figure 1, they generally were not higher.	

Table 5: Charge Question 2 *Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.* **Charge Question 2a.** *Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	[Note: Figures are included in the section following Table 11 entitled "Five Figures and Reference from Reviewer 4's Review"]	
5	The water conditions listed in Appendix A were pasted into columns C, D, and E of the Non-Pooled Model (individual slopes) and the Pooled Model (pooled slopes). The model calculated CCC and CMC were copied into a self-constructed Side-by-Side comparison spreadsheet for analysis and inspection. The data were plotted in a scatter graph for visual trend analysis and were further analyzed by fundamental statistical analyses. I did not attempt to quantify or analyze the difference any further. Upon generation of CCC and CMC values for the range of water conditions in Appendix A, there appears to be a significant positive bias for the pooled model result over the individual model result. The positive bias is generally smallest at higher water hardness levels, although more advanced multiparameter analyses may yield a different outcome.	

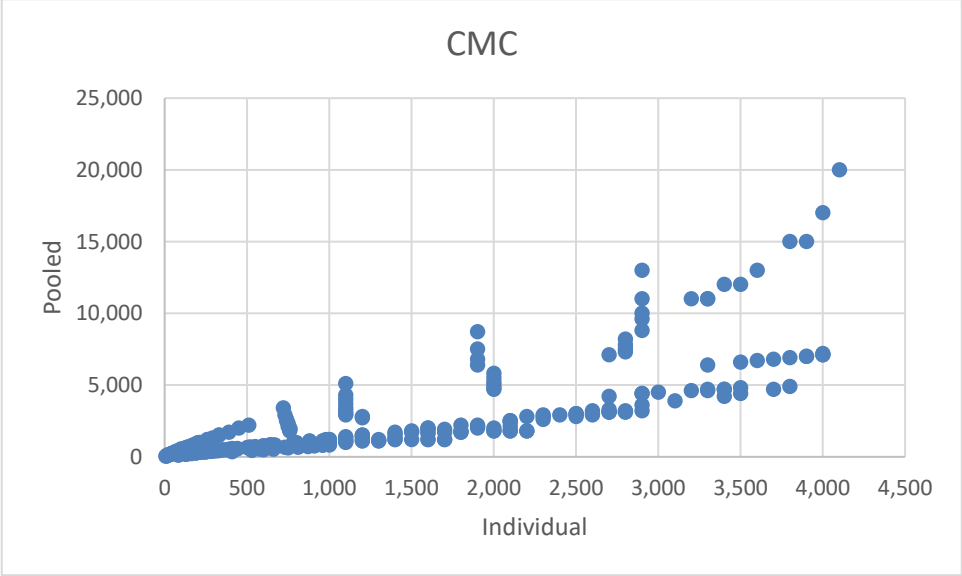
Table 5: Charge Question 2 *Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.* **Charge Question 2a.** *Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
		

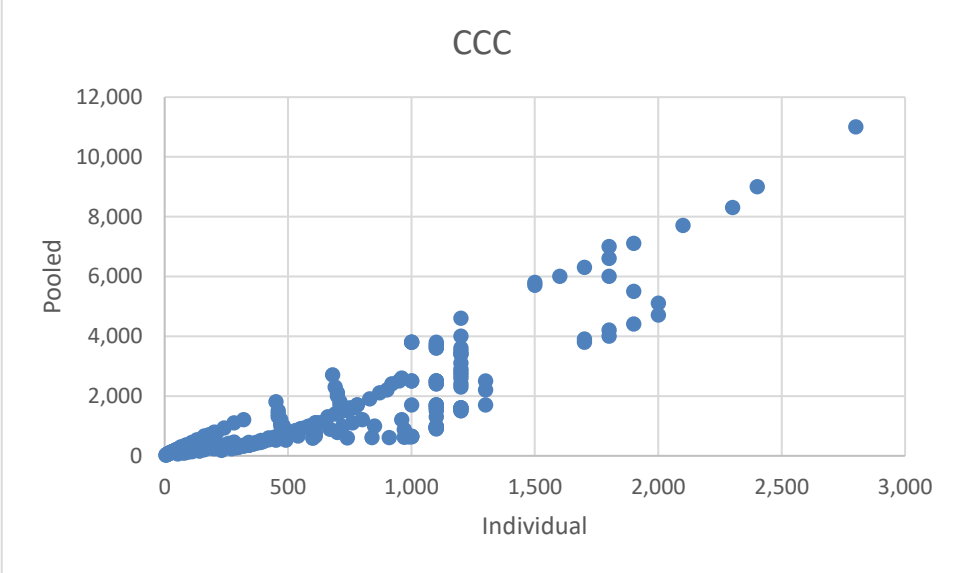
Table 5: Charge Question 2

Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

Charge Question 2a.

Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

REVEIWER NO.	REVIEWER COMMENT	EPA RESPONSE



These scatter plots possibly indicate relatively poor concordance of the output of the two models. Further comparison of the CMC and CCC results generated for the data of Appendix A input into the Non-Pooled Model and the Pooled Model, shown in the table below, yield the following:

- An average CMC Al concentration difference of 1.3 mg/L ranging from a minimum of 0.5 to 15.9 mg/L between the Non-Pooled Model and the Pooled Mode.
- An average CCC Al concentration difference of 0.81 mg/L ranging from a minimum of 0.36 to 8.2 mg/L between the Non-Pooled Model and the Pooled Mode.

Table 5: Charge Question 2

Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

Charge Question 2a.

Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

<table>
<tr><th>REVEIWER NO.</th><th>REVIEWER COMMENT</th><th>EPA RESPONSE</th></tr>
<tr><td></td><td>

An average CMC Al concentration ratio of 0.64 ranging from a minimum of 1.4 to 0.17 mg/L between the Non-Pooled Model and the Pooled Mode.
An average CCC Al concentration ratio of 0.58 ranging from a minimum of 1.6 to 0.20 mg/L between the Non-Pooled Model and the Pooled Mode.

<table>
<tr><th colspan="2">CMC Difference</th><th colspan="2">CCC Difference</th><th colspan="2">CMC Ratio</th><th colspan="2">CCC Ratio</th></tr>
<tr><td>-1,314</td><td>avg diff</td><td>-808</td><td>avg diff</td><td>0.640</td><td>avg ratio</td><td>0.580</td><td>avg</td></tr>
<tr><td>500</td><td>max</td><td>360</td><td>max</td><td>1.417</td><td>max</td><td>1.571</td><td>max</td></tr>
<tr><td>-15,900</td><td>min</td><td>-8,200</td><td>min</td><td>0.172</td><td>min</td><td>0.200</td><td>min</td></tr>
</table>
These analyses suggest that in practical use, the Non-Pooled Model and the Pooled Model would yield considerably different results, averaging 1.3 and 0.6 mg/L Al for the water conditions of Appendix A, potentially with up to five-fold differences in individual case analyses. This exercise demonstrates that practical application of the Pooled Model may not rise to the author's description "<u>Because the pooled MLR model performs well…"</u>
Thus, I can only conclude that in practical application, <u>if my use of the MLR models was not in error</u> (The user guide Readme was not particularly helpful in this regard), the Pooled Model results are uncomfortably different from the Non-Pooled Model.
</td><td></td></tr>
</table>

Table 6: Charge Question 2b. *Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	Results of these models show that use of the pooled model works as well or better than the individual models. However, I can hear the critics saying that there is no way that fish and aquatic invertebrate models should be combined because of the large difference in physiology between these two groups of organisms. I disagree because the results of the pooled model show their validity.	
2	The ratio plots indicate that the difference in prediction of the two models follows a U-shape or parabola of a second order polynomial model. The pH*pH term was included in the AIC regression model as mentioned on page 4 of the Memorandum (line 7 from the bottom) but this term was excluded in the final models on page 6. It is not clear to me whether the pH*pH term was included in the CMC and CCC calculations. The analysis of the relationship between Al toxicity and water quality parameters for individual species by DeForest et al. 2018 showed that the dependence of Al toxicity on pH for C. dubia followed a second order polynomial model (also for P. subcapitata although this was not included in the CMC and CCC calculations) while it was a linear model for P. promelas. Therefore, the pooled model will be either more represented C. dubia or P. promelas, depending on the inclusion or exclusion of pH*pH term.	
3	It makes sense to me to pool the data. Toxicity data are always sparse so expanding the data set makes sense in order to appropriately cover the range of DOC, pH and hardness required. DeForest comments on a similar issue in their original paper when they mention the uncertainty of applying MLR model for one species and endpoint to another species and endpoint but that this is an uncertainty common to hardness and BLM based approaches to bioavailability based adjusted species sensitivity distributions	

Table 6: Charge Question 2b. ***Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.***		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	(SSDs). Philosophically we are trying to protect the ecosystem so representing multiple species in the MLR seems a way to do this. In general it is not like one set of data is any more reliable than the next so including all the data is logical to me. But as you clearly asked in your charge question this is my opinion and I can certainly see the logic to use individual MLR results as well.	
4	From the comparisons here, the non-pooled model appears to have the "better" (or at least more logical) performance of the two. The exponential rise in the CCC in the pooled model with increasing pH is unexpected. The expectation is that total Al will be least toxic at circumneutral pH and start becoming more toxic at high pH. This is sort of captured in the BLM and non-pooled MLR. The magnitude of toxicity mitigation with DOC is much greater than that predicted by the BLM or non-pooled model, and the non-response to hardness in the pooled model suggests a glitch in this version.	
5	Knowing the degree of expertise of the MLR model authors, I was encouraged when they wrote: *"Because the pooled MLR model performs well, there no longer appears to be any benefit in using species-specific MLR models for ambient water quality criteria development.*" Furthermore, the model authors sufficiently back up this observation with performance metrics in their technical analysis memo. However, unless my use of the model was not correct (please better guide your users to where the inputs and outputs are), the Pooled Model does not seem to perform to the required level of "appropriateness," under the assumption that the model dynamics for the Individual or Non-Pooled Model is inherently more robust.	

Table 7: Charge Question 2c. *Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	I think the pooled model should be sufficiently robust and protective compared to the individual models and the results of this analysis show that.	
2	As discussed above, at pH 5 or between 8 and 9 the predicted criteria by the pooled MLR Model were approximately five times higher than the non-pooled MLR criteria. Therefore, at these environmental pH conditions, the pooled MLR criteria doesn't seem to be sufficiently robust and protective for low and high pH environment. pH values around 5 can be seen in metal contaminated sites, such as downstream of mine tailings. Water quality criteria for Al should be protective for this type of environment.	
3	For most waters the CMC is very similar for both approaches (in the range the model was calibrated – so excluding pH 5, 9 and 10 data from Appendix A). For many waters the pooled data will be the conservative model (DOC less than 5, Figure 4 for CMC). Inspection of the spreadsheet shows that the calculated CMC values in the pooled approach are less than the GMCV values. This should be sufficiently robust and protective. Similar to the DeForest paper if we consider the old 87 µg/L criteria and run simulations at 1 mg/L DOC, pH 6.5 and hardness of 14.7 with the pooled data we get a CCC of 120 and with the individual slopes spreadsheet we get a CCC of 130 µg/L. Not a dissimilar result to the old criteria and likely protective of aquatic life for this specific water chemistry.	
4	No, not consistently. It appears that the pooled MLR Aluminum criteria model would work well in waters with low to circumneutral pH and with relatively low DOC waters. In scenarios with high pH or high DOC the performance of the pooled	

Table 7: Charge Question 2c. *Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	model seems questionable, based on comparisons to the other two models. This is surprising, because the model fits are very similar between the species-specific and pooled MLRs in the DeForest 24August2018 memo and the data used in the model fitting covered the pH and DOC ranges of interest well (pH 6.3-8.7 and DOC 0.1 to 12 mg/L). This good agreement between the models and the protectiveness toward the sensitive taxa (C. dubia) used to develop it is illustrated in Figure 3. When the resultant CCCs from the species-specific models and the C. dubia EC10s from the updated toxicity data set (DeForest memo) are plotted together, the models fall on top of each other and the EC10s all fall at or just above the criteria values, just like they are supposed to (Figure 3). The textbook perfect behavior from the model data and the strange differences with the test "data" raises the specter that the MLRs may be overfit.	

Table 7: Charge Question 2c. *Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	However, the "data" from Appendix A and those used with the Santore ranges in Figure 1 are not "data" at all – they are contrived values selected to examine model calculations over a range of potential real world values. It is useful to compare real world data similarly. Figure 4 shows MLR CCC values for four streams for which appropriate time-series data could easily be found, and that might be close to the ranges of applicability (Figure 4). Data are from the U.S. Geological Survey's National Water Information System, http://waterdata.usgs.gov/nwis/. The relatively high pH, low DOC Snake River in Idaho showed good agreement between the two MLR approaches (Figure 4A). The other three streams are from low hardness, low pH waters in the Adirondacks and in Maine. The Wild River in Maine has variable and moderate DOC (1.4 to 12 mg/L) and the two Adirondack, New York streams have high DOC. The pooled MLR criterion values were consistently higher than the individual-slopes MLRs for these low pH, high hardness waters. The Adirondack streams also have extensive Al data, likely because of concerns of toxic episodes during acid rain episodes. For the period of record, the great majority of the total Al measurements were below both CCC models, with occasional exceedances of the lower, individual model (Figure 4). Finally, as noted in DeForest et al.'s (2018) initial presentation of the Al MLR approach, a chronic (60d) brook trout test was highly influential in EPA's older criterion document. This test had a NOEC of 88 µg/L and an LOEC of 169 µg/L, which was a 24% reduction in growth, and a growth reduction EC20 was calculated at about 156 µg/L. In DeForest et al.'s (2018) original MLR, the HC5 (the CCC by a different name) was calculated at 117 µg/L. This would seem a reasonable degree of protection for a sensitive	

Table 7: Charge Question 2c. *Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	species. At times when the Al approached criteria, the conditions were presumably stressful and result in reduced growth. However, such conditions presumably are only temporary during freshets and the fish populations would not be much harmed. In the updated criteria using the individual-slope MLR, for those conditions a CCC of 160 µg/L was calculated which is now as high as the EC20, which is a severe effect. The pooled slope MLR yields a CCC of 200 µg/L for the test conditions. This does not seem fully protective for a species that is of conservation concern in the southern Appalachians and other parts of its native range. *[Note: Figures are included in the section following Table 11 entitled "Five Figures and Reference from Reviewer 4's Review"]*	
5	With the experience and side-by-side data generated and outlined above, the Pooled MLR would not be sufficiently robust and typically over-protective.	

Table 8: Charge Question 2d. *Please provide suggestions of alternate approaches, if any.*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	One alternative approach would be the use of the HC5 (see Cardwell et al. Environmental Toxicology and Chemistry—Volume 37, Number 1—pp. 36–48, 2018). However, I am not sure that the HC5 is a better approach. Another alternative approach is the Biotic Ligand Model. Again, I am not sure that the BLM is a better approach than the MLR. I know something about the BLM when used for copper. It seems to me that the results of the BLM and the MLR may be similar but the MLR appears to be easier to use and is much more user friendly.	
2	I don't have alternative approaches and agree with the authors that the pooled model is more convenient for user because it is no more longer species specific. However, given the differences in relationship between Al toxicity and water quality parameters, such as pH (linear vs quadratic models) for different species, the pooled models would be biased and lead to less accurate prediction. In addition, the pooled and non-pooled approaches are basically statistical models. Three variables and interaction terms between them, including a quadratic term for pH were included in the models. The current available data don't seem to be strong for regression analysis of those many variables. To be more representative, more appropriate data are needed, especially data of factorial design experiments at low and high pH.	
3	I was on an earlier review of BLM based approaches. I do prefer BLM because of its mechanistic basis and the better behavior (at least in theory) during extrapolation. I think the MLR presented here is good though – but I think the pH range should be strictly restricted to the range of data used to calibrate it.	

Table 8: Charge Question 2d. *Please provide suggestions of alternate approaches, if any.*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	Also, I feel the reliance on lab tests is limiting and that real samples need to be evaluated. Total dissolved aluminum includes many potentially inert clay and other suspended particles that are not directly comparable to aluminum salt spiking in lab based trials. DeForest mentions this at the end of his paper, and that P. H Rodriguez is developing such a method, but there is no mention of this in the spreadsheets. The model predicts lab toxicity not field toxicity and this data gap will need to be filled.	
4	*Using the pooled model with caps on the questionable parameters might allow EPA to use the simpler pooled model-based criteria that would be easier for stakeholders to understand and use. Just where to set those caps would take a more careful examination of the model performance and data than is possible in the excessively short time allotted for this review. However, from figure 1 in particular, it looks like a cap for pH would be in the neighborhood of 8.5 and for DOC in the neighborhood of 2 mg/L. (Recall that a DOC of 2 in the pooled model may produce a CCC higher than that from a DOC of 12 in the non-pooled model (910 vs. 690 µg/L for hardness 25 mg/L, pH 7.5, Figure 1).* *[Note: Figures are included in the section following Table 11 entitled "Five Figures and Reference from Reviewer 4's Review"]*	
5	Unless I misused the models, only the Non-Pooled Model would be acceptable.	

Table 9: Charge Question 3 *Ease of Use:* Charge Question 3a. *Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	The fact that a calculator has been developed in Excel makes this one of the easiest methods I have ever seen. I can't come up with an easier approach than the one developed here.	
2	I found the instruction in "read me" tab to be useful. I don't know what will be included in the user manual but if someone want to determine the water quality criteria for Al based on pH, DOC, and hardness then the multiple scenarios and summary tabs are likely sufficient. I don't see the need to include the low ranks (1-4) in the multiple scenarios and over 20 scenarios or the acute and chronic data tabs.	
3	The spreadsheets are very easy to use. Very transparent – the DeForest equations are clearly available for all to see, as well as the source toxicity data. Adding the ReadMe tab in the proposed versions sent out as part of this review represents a significant improvement compared to the current online version of the MLR Aluminum Criteria Calculator. I do think it is unclear what the range should be for the MLR. The ReadMe states 6 to 9 pH but 9 is outside the range of the DeForest equations and I think is inappropriate. Also, as mentioned earlier it is unclear if outside the range data are simply flagged or if the computational approach is adjusted in some way. This needs to be clarified. When I first opened the spreadsheet the "multiple scenarios" and "over 20 scenarios" tab names confused me. I am not clear why the two tabs are needed. I guess for computational speed? This should be clarified in the ReadMe file. Otherwise why not use the	

Table 9: Charge Question 3 *Ease of Use:* Charge Question 3a. *Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	multiple scenarios all the time and just leave the unwanted fields blank? Also, it should be made clear what happens if you input less than the 20 or 500 water chemistries in those two tabs. They seem to just populate automatically with low default values – but the general user might be confused why data suddenly shows up that they didn't ask for. As already highlighted it is great that you can see the actual "DeForest" equations. Why not take it a step further and have the slope parameters in separate cells called by this equation. This would show the parameters to the end-user but also allow for ease of revision as new data modify the slopes for the equations. And ultimately since the DeForest papers actually calculate the effect concentrations it would be nice to have a column for the non-normalized EC20 results as well. I think that is a more relatable parameter than the normalized values. Now for a bigger "ask". It would be nice to link this spreadsheet to an equilibrium solver to predict solubility of common aluminum phases or even just amorphous gibbsite. This would not be a hard model to build. The results would be "just for information" but going forward it could help inform that question about inert and reactive solid aluminum. Linking the geochemistry predictions would also allow assessment of soluble versus particulate exposures.	
4	The care and skill that went into the macro enabled spreadsheets is obvious. However, for the "over 20 scenarios" runs, it took 5-10 minutes for a run. That was excruciating, trying to do multiple runs	

Table 9: Charge Question 3 *Ease of Use:* Charge Question 3a. ***Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?***		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	and it wasn't obvious whether it was running or had hung. Stakeholders will send EPA hate mail if their computers are locked up for 10 minutes after each time they click run. From the "Summary Sheet" tab, it looks as though once the modeling and criteria questions are set, it will no longer be necessary to normalize the entire SSD, and a straight "xlsx" equation will be sufficient? If not, I recommend striving for that; otherwise there will be endless complaints. Also, for those who work in organizations with centralized IT departments (a widespread malady), they may have trouble with macro-enabled Excel sheets. (I did, Figure 5). *[Note: Figures are included in the section following Table 11 entitled "Five Figures and Reference from Reviewer 4's Review"]*	
5	The guidance for the MLR spreadsheet to be used by stakeholders is far from complete and not particularly informative or useful in its present iteration. I found it frustratingly incomplete for a new user. The model only has a Readme page. For example, my environmental toxicology course students can work their way through California's LeadSpread 8 during risk assessment exam questions due to the quality of the associated manuals and user assistance. (https://www.dtsc.ca.gov/AssessingRisk/LeadSpread8.cfm). Employing spreadsheet comment fields, example calculations and a more intuitive user guide that may be a useful approach for the MLR when risk assessors access the aluminum aquatic toxicity model for the first time. As presented the MLR spreadsheets are	

Table 9: Charge Question 3 *Ease of Use:* Charge Question 3a. *Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
	not intuitive or easy to use. The model authors have attempted to insert some guidance, however this Readme guidance appears incomplete and only somewhat useful. It took me several hours to orient myself to understand the different input modalities (summary page, multiple, and over-20 multiple). In my experience most model software requires some familiarization time before user efficiency, however the supporting materials for the MLRs are below the median in quality and quantity of the materials provided. Other comments: The Readme page is not locked and is editable. Another approach to documentation and model use instruction may be better. The dual use of "Non-pooled" and "Individual" is confusing. The model seems to want to run all rows always in the multiple scenario worksheets, since the execution time was about the same for a few scenario entries, with the rest of the cells deleted. I was running the model on a Xeon processor workstation and it took about 5 minutes to run. Please use the word "output" or "result" to label the model endproduct better.	

Table 10: Charge Question 3b. *Do you have any other suggestions to improve the ease of use?*		
REVEIWER NO.	**REVIEWER COMMENT**	**EPA RESPONSE**
1	No. As mentioned above, the ease of use of the Calculator makes this very user friendly. I feel confident about the results developed from the MLR models in terms of developing aquatic life criteria for aluminum.	
2	Not really, I already see this approach easy to use compare to the BLM. However, I must say that BLM is more mechanistic approach. It takes chemical speciation and bioavailability into account, which can be applied for various environmental conditions. Given the limitation of the data and different relationships between Al toxicity and water quality parameters for different species as discussed above, the current pooled model might not be a robust approach. More data especially of factorial design experiments are needed for model calibration.	
3	I do not have any suggestions to improve ease of use. It is pretty easy to use. If you can use a spreadsheet you can use this calculator. The ReadMe needs some improved documentation, as I've indicated above, but this is a great tool.	
4	Not within the limited time available for review.	
5	Please see the comments above. I prefer models that clearly point me towards "Inputs" and "Outputs." After spending many hours with this model and supporting materials, I am still not entirely confident I am using it correctly. I had to teach myself what the summary page, multiple, and over-20 multiple inputs were by creating a small data set and applying it to each input mode so I could watch the output fields change to gain user confidence. Well developed tutorials such as the EPA Benchmark Dose support materials offer a template for excellence in user base training.	

Table 11. Specific Observations				
Reviewer No.	**Model Name**	**Tab**	**Cell**	**Comment or Question**
1				I have no additional observations or comments on the models other than what I have stated above.
2				Some specific observations and comments on the models were mentioned above.
3				[The reviewer did not provide any specific observations or comments on the models.]
4				[The reviewer did not provide any specific observations or comments on the models.]
5				No specific observations except those outlined above.

Seven Figures from Reviewer 3's Review

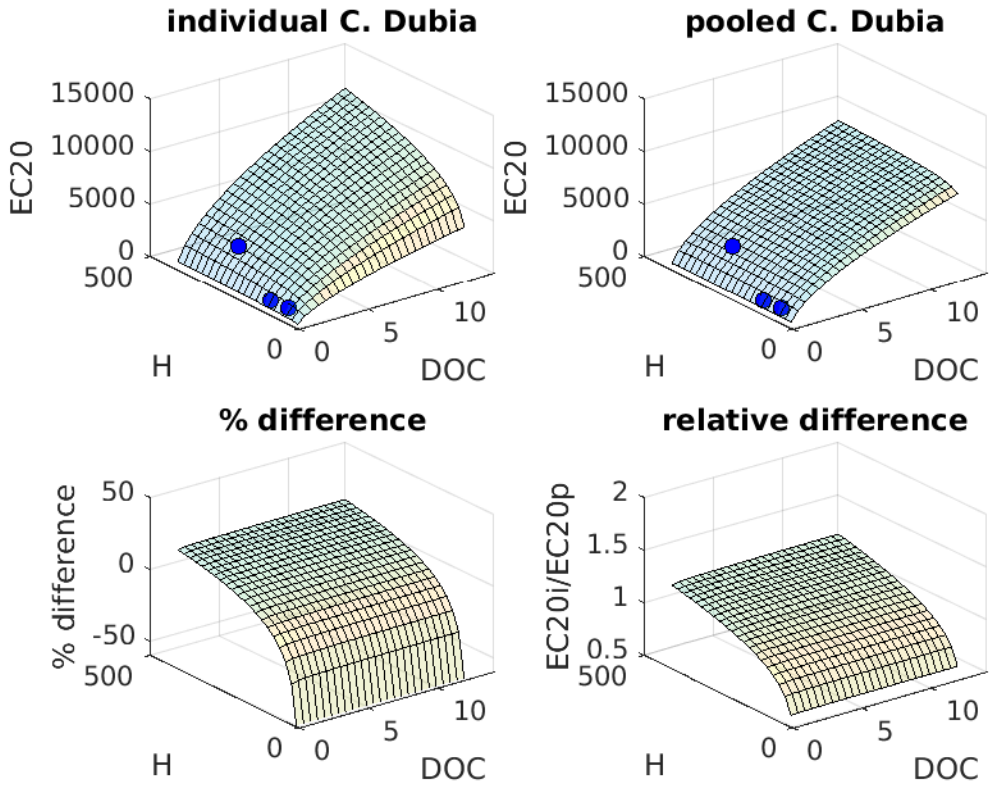


Figure 1: C. Dubia MLR predicted EC20 values at pH 6.3. The top left plot is determined using Equation 2 individual EC20 (EC20i) from the DeForest memo. The top right plot is determined using Equation 6 for pooled EC20 determinations (EC20p). The range of DOC and H were selected to match the calibration range of the MLR model. The blue dots correspond to chronic C. Dubia data from the chronic tab of the Criteria Calculator spreadsheet. The % difference plot corresponds to 100*(EC20i-EC20p)/EC20i and the relative difference is EC20i/EC20p.

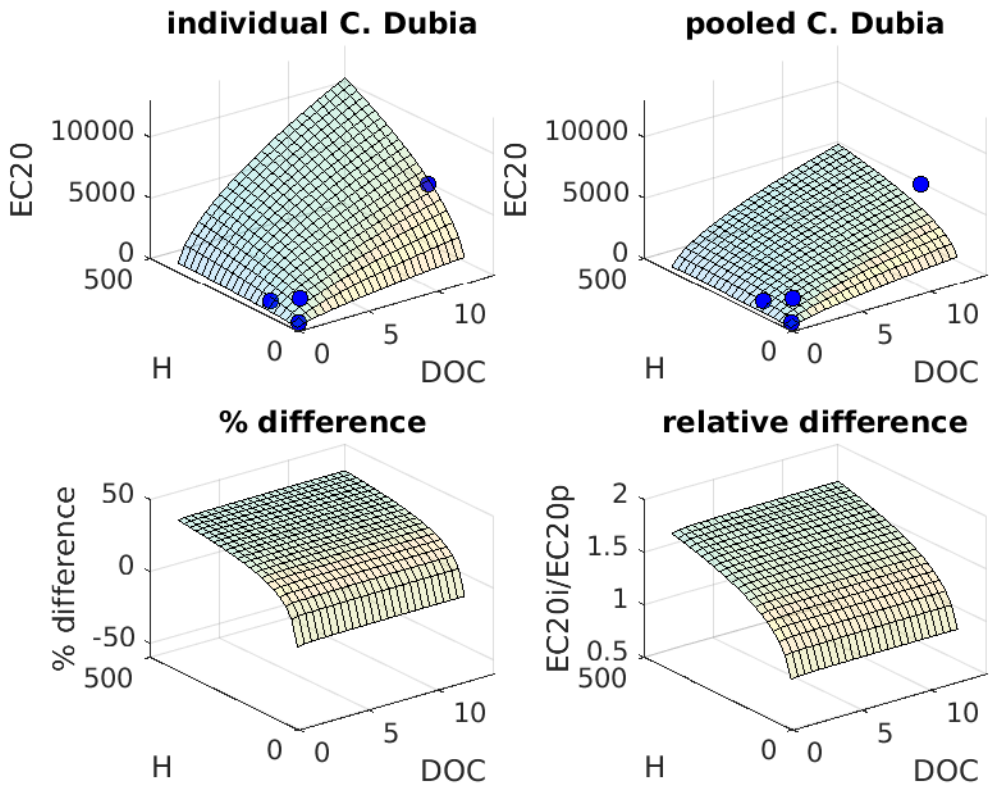


Figure 2: C. Dubia MLR predicted EC20 values at pH 7. The top left plot is determined using Equation 2 individual EC20 (EC20i) from the DeForest memo. The top right plot is determined using Equation 6 for pooled EC20 determinations (EC20p). The range of DOC and H were selected to match the calibration range of the MLR model. The blue dots correspond to chronic C. Dubia data from the chronic tab of the Criteria Calculator spreadsheet. The % difference plot corresponds to 100*(EC20i-EC20p)/EC20i and the relative difference is EC20i/EC20p.

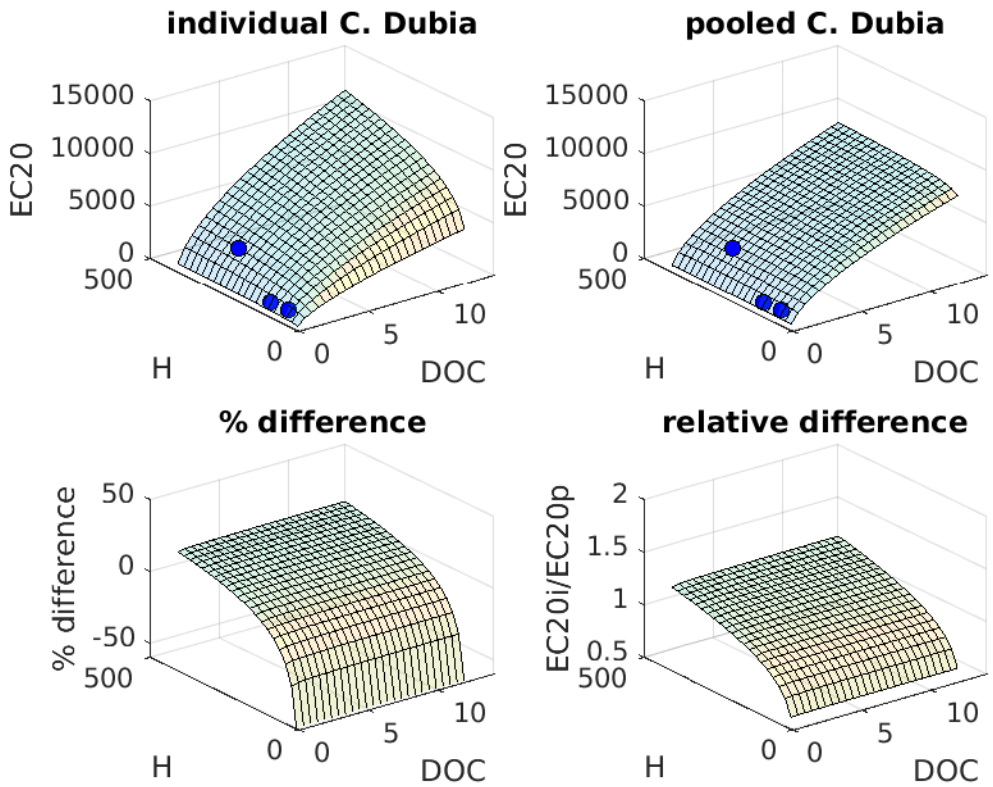


Figure 3: C. Dubia MLR predicted EC20 values at pH 8. The top left plot is determined using Equation 2 individual EC20 (EC20i) from the DeForest memo. The top right plot is determined using Equation 6 for pooled EC20 determinations (EC20p). The range of DOC and H were selected to match the calibration range of the MLR model. The blue dots correspond to chronic C. Dubia data from the chronic tab of the Criteria Calculator spreadsheet. The % difference plot corresponds to 100*(EC20i-EC20p)/EC20i and the relative difference is EC20i/EC20p.

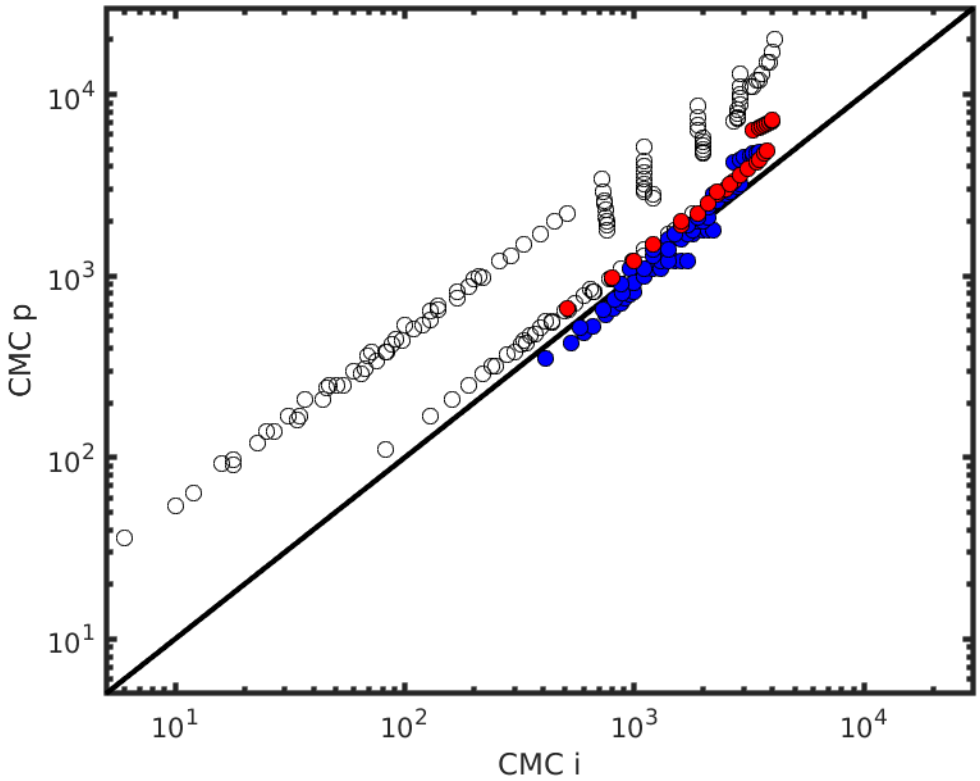


Figure 4: CMC determined using the individual spreadsheet (CMCi) and using the pooled approach (CMCp). The open circles represent all the calculations for the data in Appendix A. The closed symbols are for all the pH data in the range the model was calibrated. The red data are for high DOC (>5) and the blue data are for low DOC (<5).

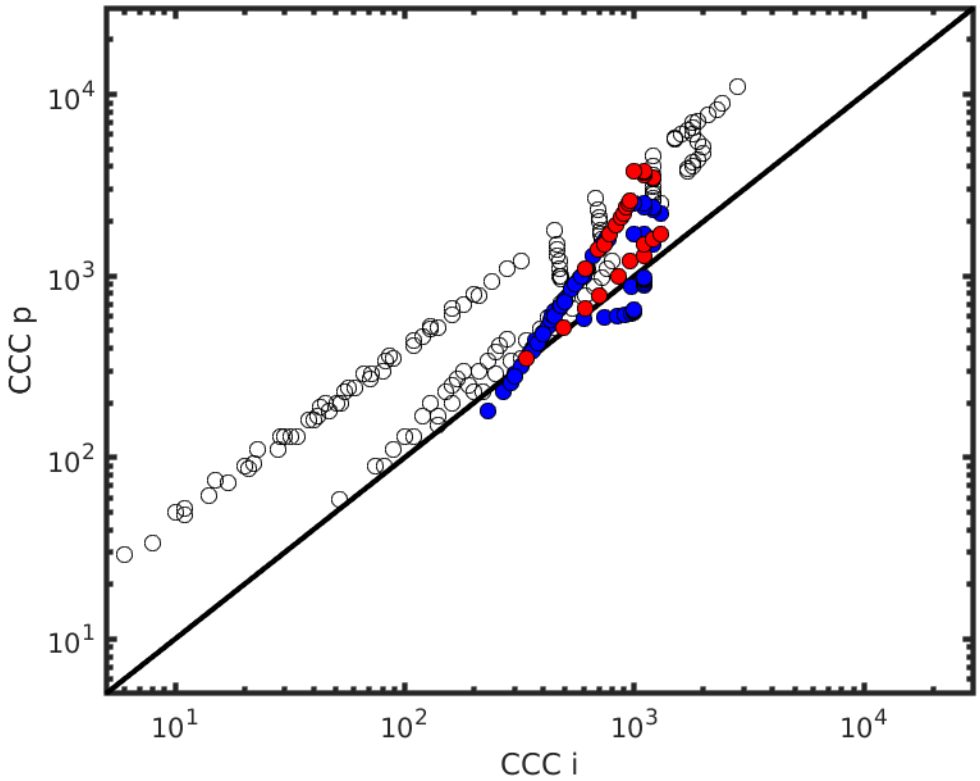


Figure 5: CCC determined using the individual spreadsheet (CCCi) and using the pooled approach (CCCp). The open circles represent all the calculations for the data in Appendix A. The closed symbols are for all the pH data in the range the model was calibrated. The red data are for high DOC (>5) and the blue data are for low DOC (<5).

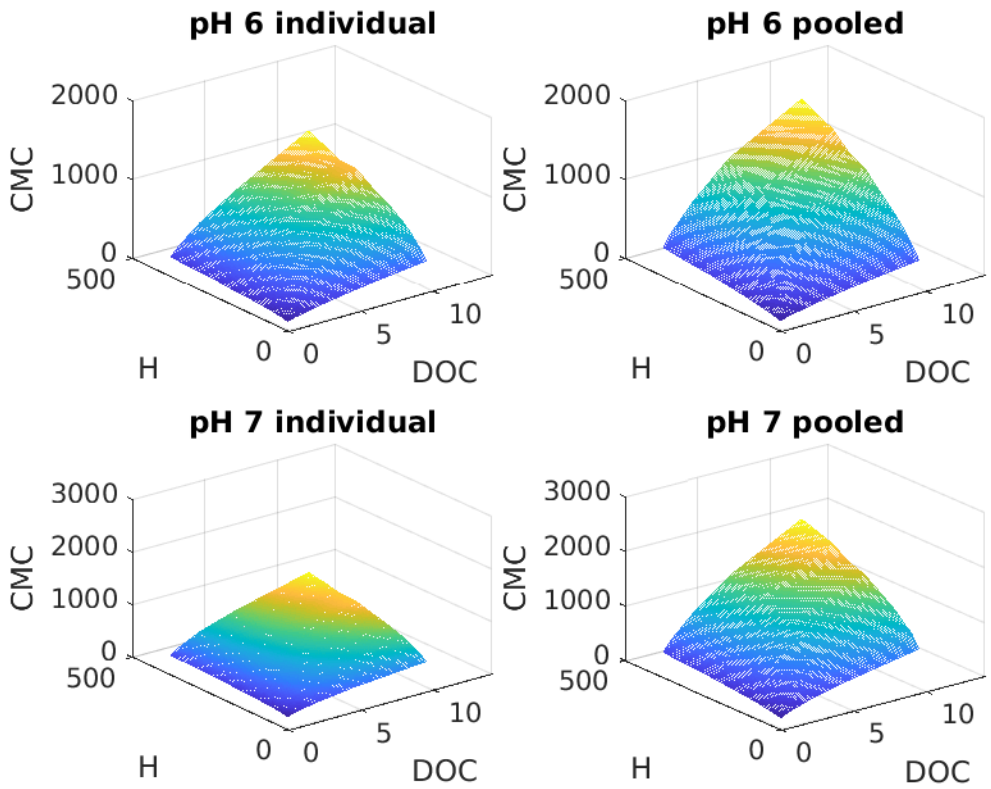


Figure 6: pH 6 and 7 Appendix A data used to derive CMC values as a function of hardness (H) and dissolved organic carbon (DOC). The results from the individual spreadsheet are shown on the left and for the pooled data are shown on the right.

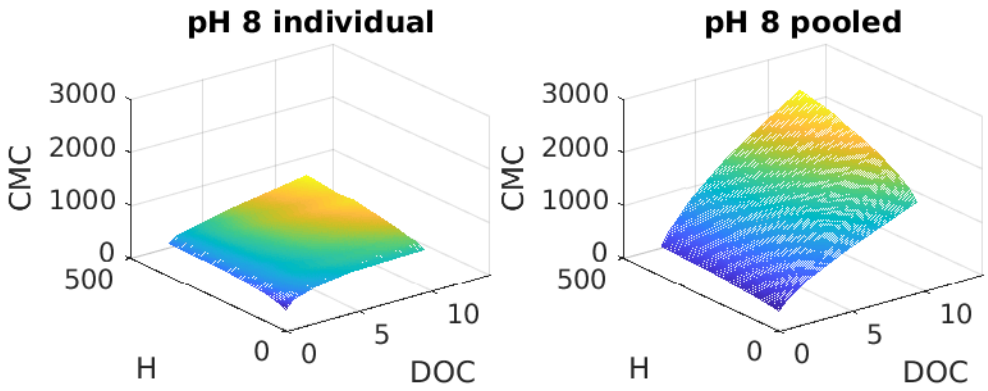


Figure 7: pH 8 Appendix A data used to derive CMC values as a function of hardness (H) and dissolved organic carbon (DOC). The results from the individual spreadsheet are shown on the left and for the pooled data are shown on the right.

Five Figures and Reference from Reviewer 4's Review

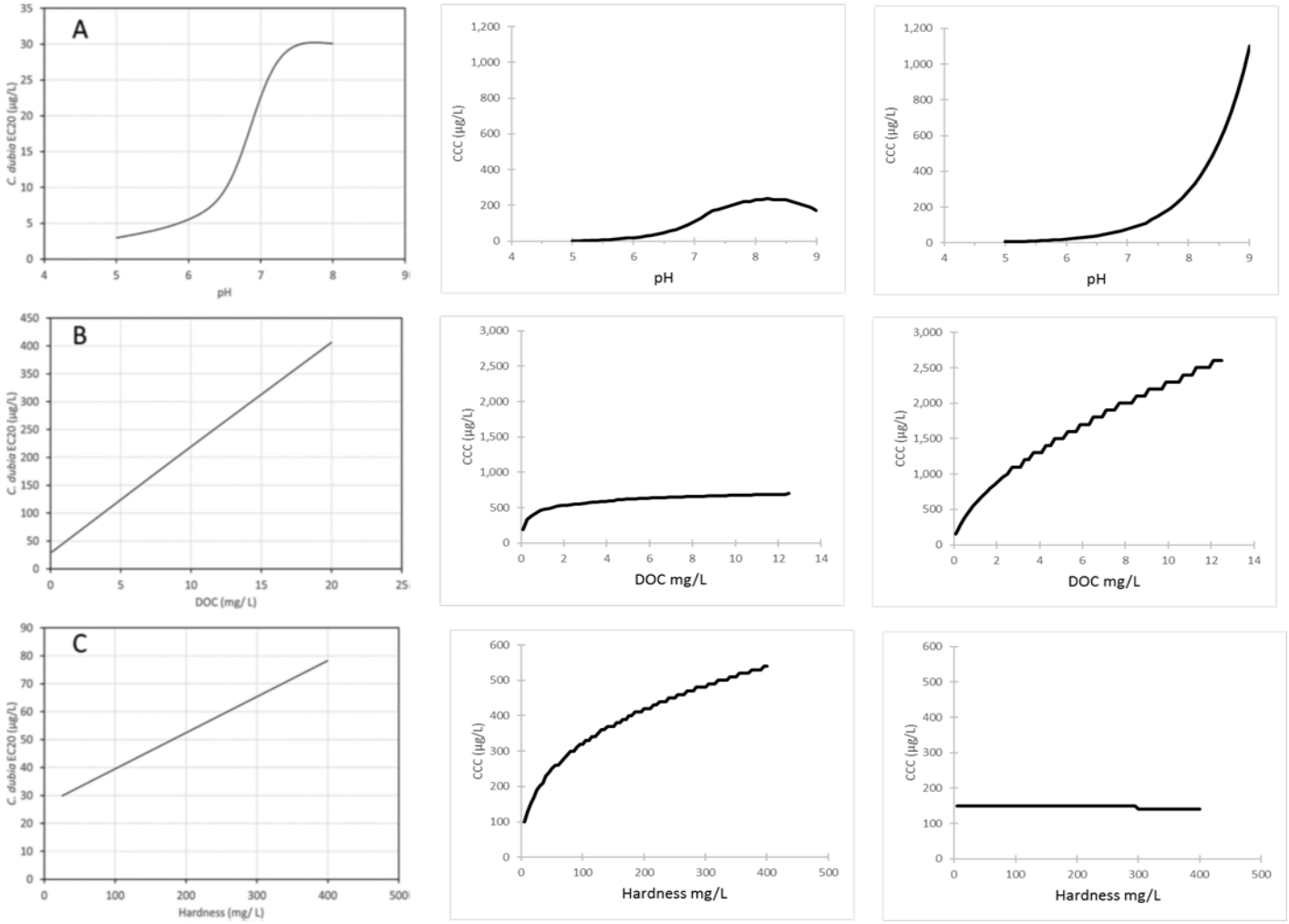


Figure 1. Variation in predicted toxicity patterns as a function of water quality showing the response in aluminum (Al) bioavailability for either the Al BLM (Santore et al. (2018), left); the individual slopes MLR (center), and the pooled slopes MLR (right) to changes in pH (A), dissolved organic carbon (DOC; B), and hardness (C). Base conditions for each simulation are temperature 20 8C, pH 7.5, DOC 0.1 mg/L, and hardness 25 mg/L. The response patterns between the models are disappointingly different (Warning – vertical axes scales are very different between the BLM and MLR plots.). Jittering is an artefact of the input values chosen for the MLR.

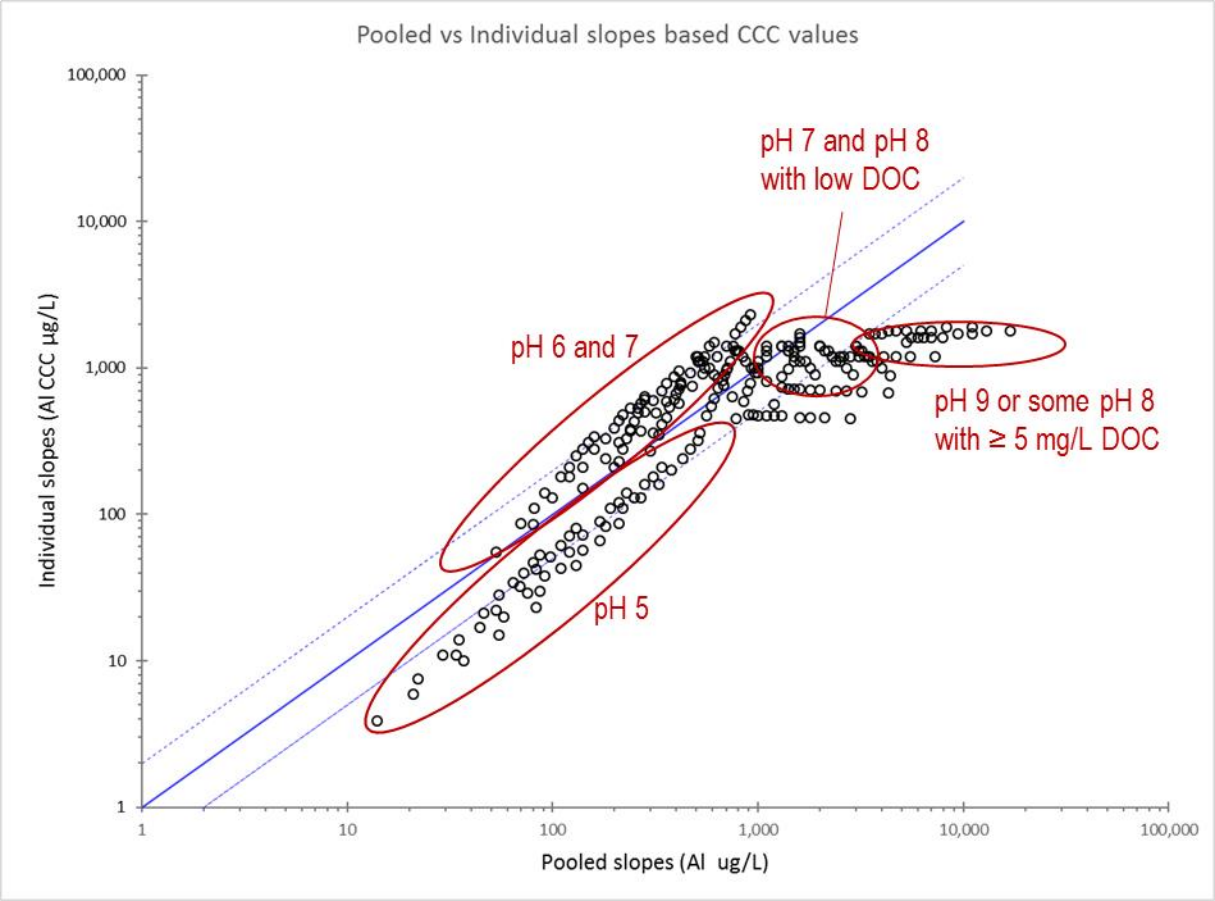


Figure 2. The 250 "Appendix A" test values covering a range of DOC, pH, and hardness values produced CCC values that were surprisingly divergent. 87 (35%) of the pairs differed by >2X and 37 (15%) differed by more than 3X. Poorest agreement was for the extreme values, especially for pH 9 combinations. Best agreement was for the pH 6 and 7 combinations, and pH 8 at low DOC.

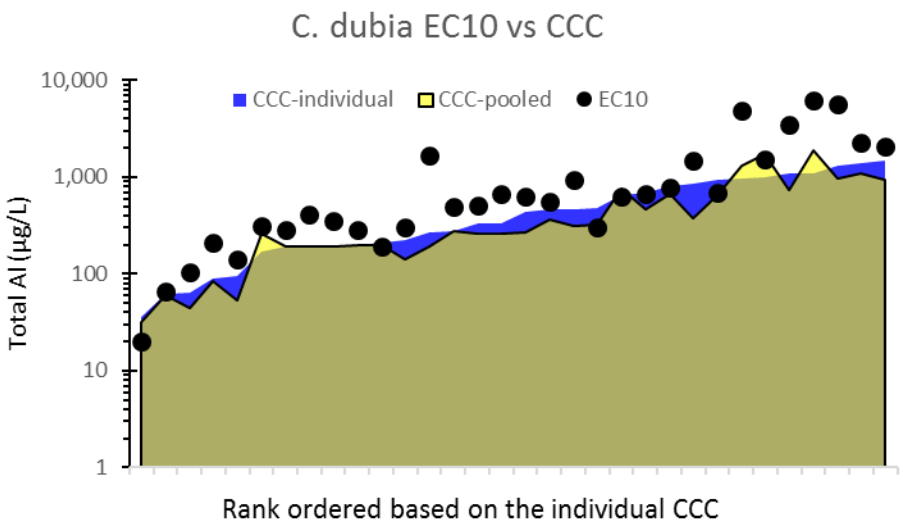


Figure 3. *Ceriodaphnia dubia* toxicity (EC10s) versus the non-pooled or pooled CCC versions. Data from DeForest memo,

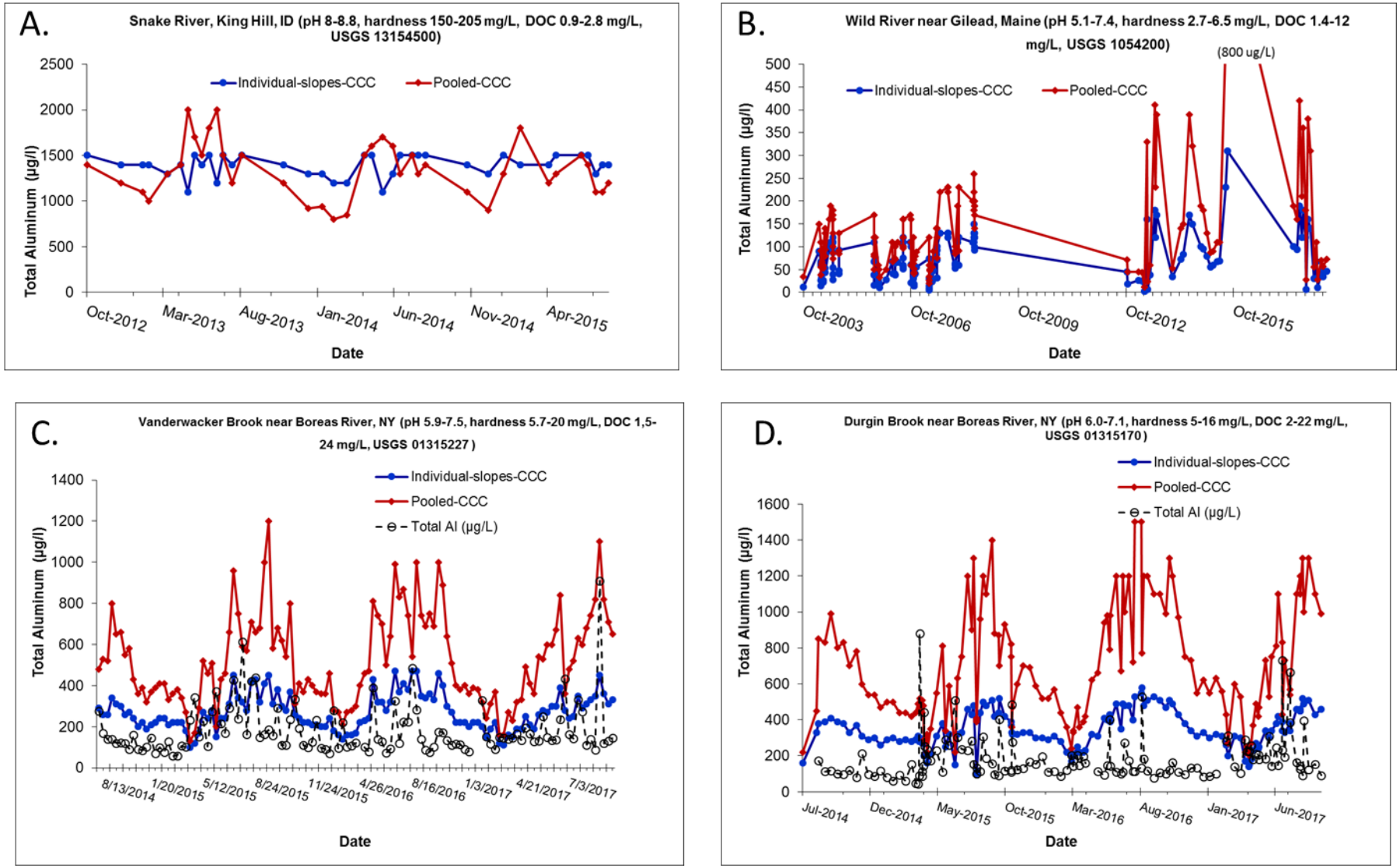


Figure 4. Comparisons of criteria in natural waters. In a river with moderately high pH and low DOC, the two MLR CCC versions were mostly similar; in the low pH waters in which aluminum toxicity is actually a real concern, the non-pooled MLR version tended to be lower.

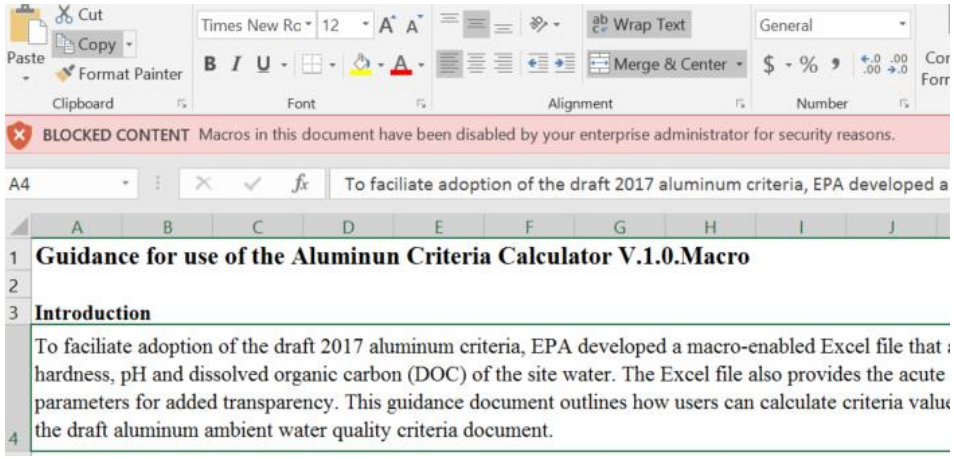


Figure 5. Corporate IT people don't like macro-enable Office files and may disable them just because they can. Reconfiguring to a simple equation would be much preferable for distribution to those who just want to calculate their number.

Reference

DeForest, D.K., K.V. Brix, L.M. Tear, and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. *Environmental Toxicology and Chemistry*. 37(1): 80-90. https://doi.org/10.1002/etc.3922

IV. INDIVIDUAL PEER REVIEWER COMMENTS

Review By:
Reviewer 1

Peer Review Comments on Aluminum Criteria Model

Reviewer 1

I. GENERAL IMPRESSIONS

Prior to agreeing to conduct this review, I have been working on an NAS panel on an update of the 2015 EPA Multi-Sector General Stormwater Permit (MSGP). Because aluminum is a stormwater benchmark monitoring requirement for some of the sectors in this permit, I have familiarized myself with the original aquatic life criteria developed for aluminum (1988). I have also briefly looked over the 2017 draft document. I therefore appreciate the difficulty of working with metal toxicity and risk assessments for aquatic ecosystems. As pointed out in the Deforest memorandum and other papers (see the special edition of ET&C 37(1) 2018 for a number of papers dealing with aluminum toxicity), including the 2017 draft, the editorial by Adams et al. 2018 (ET&C 37(1) 34–35, aluminum toxicity is dependent upon water quality characteristics (pH, hardness, DOC), not unlike other metals, including copper and zinc. The Biotic Ligand model has been used in the past but it is difficult to use. I found that the multiple linear regression (MLR) model approach outlined in the Deforest memorandum is well-thought out. I am particularly impressed with the Calculator as it produce excellent results and is easy to use. The additional studies (new toxicity data since the original ALC in 1988) included in this document are of great value as they increased all of the R^2 values. The MLR model is a great improvement over past models because it incorporates pH, DOC, and hardness as these values relate to bioavailability and hence toxicity. The MLR can be used to normalize acute and chronic toxicity data to a set of predetermined water quality conditions. The MLR was also used to determine what water quality parameters are of value and which are not as important in terms of R^2. Furthermore, the authors determined that a pooled MLR model had higher adjusted and predicted R^2 values compared to the species specific models. This conclusion was justified by the results of the individual and pooled models. I agree that the results of these models indicate that the pooled model should be used in place of individual models.

II. RESPONSE TO CHARGE QUESTIONS

1. Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.

1a. Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?

Yes. In fact, results of these MLR equations show that the addition of the new toxicity data improve the models.

1b. Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.

All of the MLRs are appropriately parameterized. I would not add anything to the model inputs. However, it was interesting to me that the ln(DOC) x pH term was excluded in the *C. dubia* model but retained in the *P. promelas* model. As a modeler, I have encountered scenarios like this in the past. Sometimes, this is just a matter of inadequate data sets.

1c. Does the pooled model behave similarly as the non-pooled models?

Yes. The pooled model does behave similarly to the non-pooled models. In fact, the R^2 were somewhat higher of the pooled model compared to the individual models. A strong case is made by DeForest et al. 2018, for the use of the pooled model over the use of the individual models.

2. Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

I compared the resulted of the non-pooled to the pooled results and found that the pooled results were similar to the individual results.

2a. Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

The Criterion Maximum Concentration (CMC) is the highest concentration of a chemical in water that aquatic organisms can be exposed to acutely without causing an adverse effect. The Criterion Continuous Concentration (CCC) is the highest concentration of a chemical in water that aquatic organisms can be exposed to indefinitely without resulting in an adverse effect. The CMC is usually higher than the CCC and this is exactly what the MLR models predict.

2b. Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.

Results of these models show that use of the pooled model works as well or better than the individual models. However, I can hear the critics saying that there is no way that fish and aquatic invertebrate models should be combined because of the large difference in physiology between these two groups of organisms. I disagree because the results of the pooled model show their validity.

2c. Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?

I think the pooled model should be sufficiently robust and protective compared to the individual models and the results of this analysis show that.

2d. Please provide suggestions of alternate approaches, if any.

One alternative approach would be the use of the HC5 (see Cardwell et al. Environmental Toxicology and Chemistry—Volume 37, Number 1—pp. 36–48, 2018). However, I am not sure that the HC5 is a better approach.

Another alternative approach is the Biotic Ligand Model. Again, I am not sure that the BLM is a better approach than the MLR. I know something about the BLM when used for copper. It seems to me that the results of the BLM and the MLR may be similar but the MLR appears to be easier to use and is much more user friendly.

3. Ease of Use:

3a. Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?

The fact that a calculator has been developed in Excel makes this one of the easiest methods I have ever seen. I can't come up with an easier approach than the one developed here.

3b. Do you have any other suggestions to improve the ease of use?

No. As mentioned above, the ease of use of the Calculator makes this very user friendly. I feel confident about the results developed from the MLR models in terms of developing aquatic life criteria for aluminum.

III. SPECIFIC OBSERVATIONS

Model Name	Tab	Cell	Comment or Question
			I have no additional observations or comments on the models other than what I have stated above.

Review By:
Reviewer 2

Peer Review Comments on Aluminum Criteria Model

Reviewer 2

I. GENERAL IMPRESSIONS

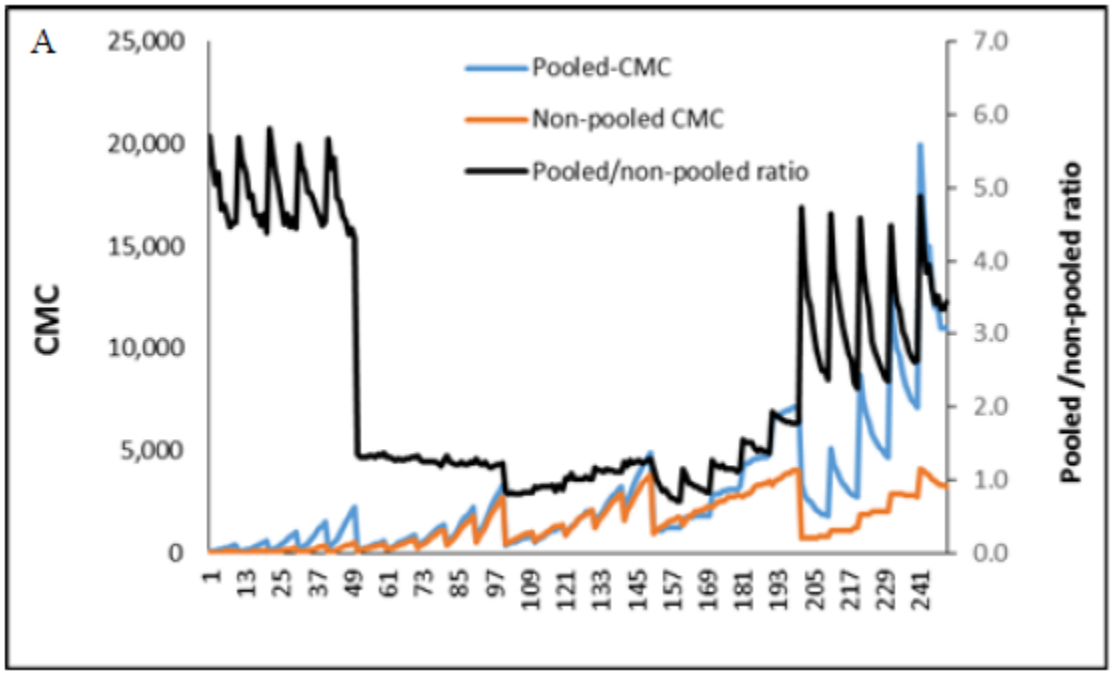
I have reviewed the documents provided by Versar that are presented in the below Table. An updated version of the Memorandum was provided on September 12. The Al criteria presented in these documents was developed based on multiple linear regression model approach. Two MLR criteria models were developed. One is for individual species (non-pooled model) and the other is for a combination of 2 species of C. dubia and P. promelas (pooled model). The model development was clearly described in DeForest et al. 2018 paper. The Memorandum presented an update to the models of DeForest et al. 2018 at which, new data for C. dubia and P. promelas were used for calculation of the model coefficients (slopes). A pooled model that combined data for C. dubia and P. promelas was also presented in the Memorandum. The provided scenarios of data that had a pH range of 5-9, a DOC range of 0.5-10 mg/L, and a hardness range of 25-400 mg/L as CaCO3 were used to run the models and calculate the CMC and CCC values. A relative site-by-site comparison of the CMC and CCC values of the pooled and non-pooled models was conducted by calculating the ratio of the CMC and CCC values predicted by the pooled model to those predicted by the non-pooled model (Fig A and B). Below are some general comments for the model development and performance. Some of these comments will be further discussed and presented in the answers to the charge questions.

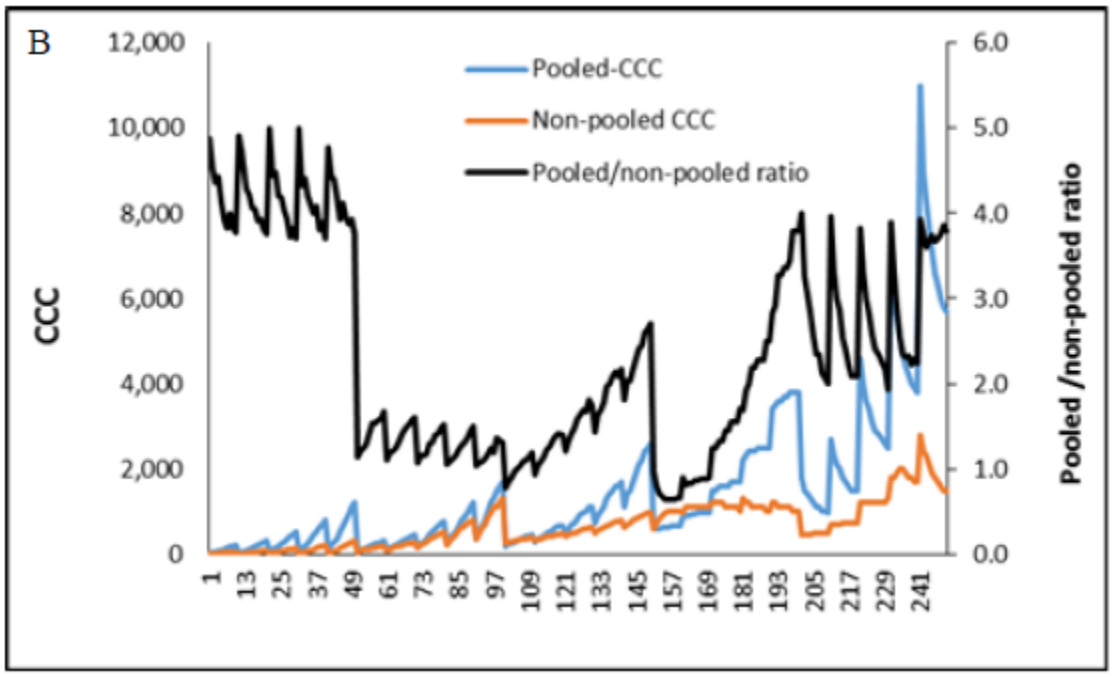
- The MLR model approach is for sure easier to use than the Biotic Ligand Model approach. However, the BLM takes metal speciation and bioavailability into account and can be applied for various environmental conditions. The MLR is a statistical approach and its application is logically limited the range of environmental conditions that was used for model development. Most of the data used for the model development were coming from laboratory research that used formulated water which is cleaner and less extreme than field waters. Given the complicated chemistry of Al, especially in different pH conditions, I am not sure how well the MLR model prediction will represent the natural environment.
- The current data (including the addition of the new data set) don't seem to be strong for a multiple regression analysis that get involved with at least 3 variables and interaction terms between them including a quadratic term, such as for pH (pH*pH). When such regression models are developed, data of factorial design experiments are more suitable for use. The limitation of data used for the model development might end up with a model that is less representative and hence less accurate prediction, especially for cases that the data are outside or at the boundary of the current range and for other species rather than the two species used for the model calibration.
- There are advantages and disadvantages between the pooled and non-pooled models. The non-pooled model clearly distinguish the dependence of Al toxicity on water quality. For examples, quadric model for pH and P. subcapitata and C. dubia but linear for P. promelas. The pooled model combined C. dubia and P. promelas data and likely excluded the quadratic term. This might make the model be biased to P. promelas. Since data for other fish species are not sufficient and the dependence of Al toxicity on pH for other fish species is unknown, the current pooled model might not be representative. The

conclusion of using the pooled model instead of non-pooled model for predicting Al criteria is less convincing. The pooled model predictions are much higher than the non-pooled model predictions for low and high pH cases. This doesn't sound that the pooled model criteria is protective although it is more convenient and preclude the need to recalculate genus species distribution.

- Given the MLR criteria- a statistical approach, 95% confidence intervals can be used instead of the acceptable prediction of 2-fold above and below the perfect prediction that has been used by the BLM approach.

File Name	Description
MLR Model_Pooled Slopes_Aluminum Criteria Calculator_8.29.18.xlsm	Pooled Slopes Aluminum Calculator
MLR Model_Individual_Slopes_Aluminum Criteria Calculator_8.29.18.xlsm	Individual Slopes Aluminum Calculator
Appendix A 9-5-18.xlsx	Appendix A file is to be used to check models for charge question #2
DeForest_et_al-2018-Environmental_Toxicology_and_Chemistry.pdf	DeForest et al. 2018 Paper
DeForest Aluminum MLR Models Update Memo (2018-08-24).pdf	DeForest Memo to EPA





II. RESPONSE TO CHARGE QUESTIONS

1. Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.

1a. Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?

Yes, the MLR models developed by DeForest et al. 2018 are basically statistical models. Therefore, the models will be more confident if more data are used for model calibration. The Memorandum mentioned the improvement (higher R2 values) when new data set was included. In addition, the new data set covered a wider range of water quality parameters. Therefore, the updated models logically can be used to predict the toxicity of Al for a wider range of water quality, such as hardness, pH, and DOC.

1b. Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.

The idea of combining fish and invertebrate data to develop a pooled model sounds reasonable because the model then can be used for predicting toxicity for both fish and invertebrate. However, it is not clear to me on how the sensitivity of each species was quantitatively taken into account. The Memorandum did mention that a species term and terms for each of the independent variables and their interactions were included in the pooled model but I don't see

them in the results and conclusion. Equations 5 to 8 are separately for C. dubia and P. promelas. No slope for species term and intercept value was presented for the pooled models on page 6 of the Memorandum.

1c. Does the pooled model behave similarly as the non-pooled models?

The predictions of the two models for various scenarios showed a similar trend (Fig A and B) but relatively the predictions of the two models at low and high pH are about 5 time different as discussed above.

2. Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

2a. Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

The predicted CMC and CCC values by the pooled and non-pooled models were plotted in Fig. A and B above. The first 50 data points are for pH 5 scenarios. The last 50 data points are for pH 9 scenarios. The ratio of the pooled to non-pooled CMC and CCC values were also plotted. It can be seen that the model predictions are not the same across the pH values and more pH dependent. At pH 5 and 9, the predicted CMC and CCC values by the pooled model were approximately 5 times higher than those by the non-pooled model. Both models seem to give similar predicted CMC and CCC values at pH between 6 and 8 (ratio ~ 1). This pH range captures most pH data used to develop the models (few data points with pH between 5 and 6). Outside of this pH range, especially at pH 5 and 9, the predictions are likely extrapolated because no pH 5 and 9 was used for model calibration. Therefore, the predictions might not be confident at these pH conditions.

2b. Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.

The ratio plots indicate that the difference in prediction of the two models follows a U-shape or parabola of a second order polynomial model. The pH*pH term was included in the AIC regression model as mentioned on page 4 of the Memorandum (line 7 from the bottom) but this term was excluded in the final models on page 6. It is not clear to me whether the pH*pH term was included in the CMC and CCC calculations. The analysis of the relationship between Al toxicity and water quality parameters for individual species by DeForest et al. 2018 showed that the dependence of Al toxicity on pH for C. dubia followed a second order polynomial model (also for P. subcapitata although this was not included in the CMC and CCC calculations) while it was a linear model for P. promelas. Therefore, the pooled model will be either more represented C. dubia or P. promelas, depending on the inclusion or exclusion of pH*pH term.

2c. Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?

As discussed above, at pH 5 or between 8 and 9 the predicted criteria by the pooled MLR Model were approximately five times higher than the non-pooled MLR criteria. Therefore, at these environmental pH conditions, the pooled MLR criteria doesn't seem to be sufficiently robust and protective for low and high pH environment. pH values around 5 can be seen in metal contaminated sites, such as downstream of mine tailings. Water quality criteria for Al should be protective for this type of environment.

2d. Please provide suggestions of alternate approaches, if any.

I don't have alternative approaches and agree with the authors that the pooled model is more convenient for user because it is no more longer species specific. However, given the differences in relationship between Al toxicity and water quality parameters, such as pH (linear vs quadratic models) for different species, the pooled models would be biased and lead to less accurate prediction. In addition, the pooled and non-pooled approaches are basically statistical models. Three variables and interaction terms between them, including a quadratic term for pH were included in the models. The current available data don't seem to be strong for regression analysis of those many variables. To be more representative, more appropriate data are needed, especially data of factorial design experiments at low and high pH.

3. Ease of Use:

3a. Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?

I found the instruction in "read me" tab to be useful. I don't know what will be included in the user manual but if someone want to determine the water quality criteria for Al based on pH, DOC, and hardness then the multiple scenarios and summary tabs are likely sufficient. I don't see the need to include the low ranks (1-4) in the multiple scenarios and over 20 scenarios or the acute and chronic data tabs.

3b. Do you have any other suggestions to improve the ease of use?

Not really, I already see this approach easy to use compare to the BLM. However, I must say that BLM is more mechanistic approach. It takes chemical speciation and bioavailability into account, which can be applied for various environmental conditions. Given the limitation of the data and different relationships between Al toxicity and water quality parameters for different species as discussed above, the current pooled model might not be a robust approach. More data especially of factorial design experiments are needed for model calibration.

III. SPECIFIC OBSERVATIONS

Model Name	Tab	Cell	Comment or Question
			Some specific observations and comments on the models were mentioned above.

Review By:
Reviewer 3

Peer Review Comments on Aluminum Criteria Model

Reviewer 3

I. GENERAL IMPRESSIONS

It is clear that the scope of this review is to evaluate different possible aluminum criteria calculators (excel spreadsheets) all based on multiple linear regression (MLR). The primary purpose of this review is to evaluate and provide written comments on EPA's Aluminum Criteria Calculator/Model and answer three charge questions. The focus of the review is on two Excel spreadsheets with multiple tabs that contain the aluminum model. A user's guide is included in the Excel spreadsheets as a ReadMe tab.

The starting place for this MLR process is the recent DeForest et al. (2017) paper along with more recent data and revised MLR models (memo from DeForest et al., 2018). From these MLR models, which predict ECx concentrations as a function of pH, hardness and DOC, spreadsheets were built to predict effect concentrations as a function of those 3 water chemistry variables and convert them to CCC and Criterion Maximum Concentration (CMC) for use by stake holders. Spreadsheets were built using old and new data (the old data spreadsheet is already available online, the new spreadsheets are what are being evaluated here). The new data spreadsheets include either pooled or non-pooled versions.

The initial impression of the proposed Criteria Calculator is that it was a good choice to use the familiar Excel software platform. Essentially all potential end-users (scientists, consultants, permit writers, …) will be familiar with Excel. This comfortable environment is a good choice for this tool. These models are designed for ease of use, using the common and familiar excel interface, and have been designed with the end user in mind. There is excellent transparency in how easy it is to find the underlying MLR equations within the spreadsheet, as well as seeing all the effects data that are used in the original MLR modelling.

The information presented is accurate (the spreadsheets seem to apply the DeForest equations correctly) and for the most part presented clearly (see some exceptions below). In terms of soundness of conclusions, there were no conclusions to evaluate. Just the software tools.

II. RESPONSE TO CHARGE QUESTIONS

1. Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.

1a. Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?

Yes it is appropriate to include the new toxicity data in the MLR equation. The original DeForest paper specifically mentions that data expanding the range of pH, DOC and hardness would be required to use the model for parameters outside the calibration range. A limitation of MLR models, because they are empirical, is that you cannot use them for waters outside the calibration range. Expanding the calibration range is exactly appropriate. Examination of Figures 1-4 in the DeForest memorandum clearly show that effect concentration predictions only negligibly change with this added data.

1b. Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.

The MLR method in the original DeForest paper is mathematically and scientifically sound. The parameters for both models were derived from this method so yes the parameters are sound. It is a limitation of empirical models that there is no theoretical basis for the values of the parameters so there is no theory to compare the values to. For this approach it is sufficient that the data points are described by the MLR parameters in a statistically best sense.

1c. Does the pooled model behave similarly as the non-pooled models?

Yes. There are three attached figures at the end of this document that demonstrate the same behavior of the pooled and non-pooled models (Figures 1 to 3). The individual (non-pooled) model and the pooled model both show protection (increasing EC20) as DOC increases and hardness increases for all 3 pHs plotted. *C. Dubia* was used as the example for these calculations. There are differences between the two models. The pooled model tends to show lower effect concentrations but the relative differences are never more than a factor of 2 and this only occurs at extremely low hardness values. The differences tend to be much smaller than that. More significantly it can be seen that by plotting the data used to calibrate the model (blue dots on Figures 1-3) the data and the model agree, although the pooled data does not agree as well as the individual data. This is to be expected because the pooled data has to satisfy more points simultaneously. The agreement between pooled and individual ECx predictions is also clearly shown by the four figures in the DeForest memo as mentioned in comment 1(a) above.

2. Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

Results of the side by side modelling are presented in the attached Figures 4 to 7.

2a. Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

Figure 4 demonstrates that the pooled spreadsheet often estimates higher CMC and CCC. It is unclear why Appendix A data were selected for this exercise though. Much of the pHs are outside the calibration range of the MLR. Unlike a mechanistic approach like a BLM, MLR cannot be extrapolated outside the calibration range. I am not clear on how this outside the range data was handled in the calculations. At one point in the instructions it just says it is flagged –

but it was not when I ran the spreadsheet. It seems the flag might only work when DOC is too high? Later in the "read me" tab it says the excel model will default to the maximum recommended conditions when parameters are outside the range. I do not know if this was done, or exactly what this means. For parameters outside the range, are they just flagged? Or is the computational approach modified in some way. Some clarity is needed.

In addition the documentation (read me) tab says that the range goes to pH of 9, but the DeForest memo states 8.1 is the calibration range. pH is of course on a log scale so 8 and 9 are an order of magnitude different.

If we focus on the data that is within the calibration range of DeForest's proposed equations the pooled and individual results are very similar (Figure 4 and 5 below) and cluster around the one to one line. The tendency is that at low DOC the pooled results are lower and for high DOC the pooled results are higher.

2b. Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.

It makes sense to me to pool the data. Toxicity data are always sparse so expanding the data set makes sense in order to appropriately cover the range of DOC, pH and hardness required. DeForest comments on a similar issue in their original paper when they mention the uncertainty of applying MLR model for one species and endpoint to another species and endpoint but that this is an uncertainty common to hardness and BLM based approaches to bioavailability based adjusted species sensitivity distributions (SSDs). Philosophically we are trying to protect the ecosystem so representing multiple species in the MLR seems a way to do this. In general it is not like one set of data is any more reliable than the next so including all the data is logical to me. But as you clearly asked in your charge question this is my opinion and I can certainly see the logic to use individual MLR results as well.

2c. Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?

For most waters the CMC is very similar for both approaches (in the range the model was calibrated – so excluding pH 5, 9 and 10 data from Appendix A). For many waters the pooled data will be the conservative model (DOC less than 5, Figure 4 for CMC).

Inspection of the spreadsheet shows that the calculated CMC values in the pooled approach are less than the GMCV values. This should be sufficiently robust and protective. Similar to the DeForest paper if we consider the old 87 µg/L criteria and run simulations at 1 mg/L DOC, pH 6.5 and hardness of 14.7 with the pooled data we get a CCC of 120 and with the individual slopes spreadsheet we get a CCC of 130 µg/L. Not a dissimilar result to the old criteria and likely protective of aquatic life for this specific water chemistry.

2d. Please provide suggestions of alternate approaches, if any.

I was on an earlier review of BLM based approaches. I do prefer BLM because of its mechanistic basis and the better behavior (at least in theory) during extrapolation. I think the

MLR presented here is good though – but I think the pH range should be strictly restricted to the range of data used to calibrate it.

Also, I feel the reliance on lab tests is limiting and that real samples need to be evaluated. Total dissolved aluminum includes many potentially inert clay and other suspended particles that are not directly comparable to aluminum salt spiking in lab based trials. DeForest mentions this at the end of his paper, and that P. H Rodriguez is developing such a method, but there is no mention of this in the spreadsheets. The model predicts lab toxicity not field toxicity and this data gap will need to be filled.

3. Ease of Use:

3a. Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?

The spreadsheets are very easy to use. Very transparent – the DeForest equations are clearly available for all to see, as well as the source toxicity data. Adding the ReadMe tab in the proposed versions sent out as part of this review represents a significant improvement compared to the current online version of the MLR Aluminum Criteria Calculator.

I do think it is unclear what the range should be for the MLR. The ReadMe states 6 to 9 pH but 9 is outside the range of the DeForest equations and I think is inappropriate. Also, as mentioned earlier it is unclear if outside the range data are simply flagged or if the computational approach is adjusted in some way. This needs to be clarified.

When I first opened the spreadsheet the "multiple scenarios" and "over 20 scenarios" tab names confused me. I am not clear why the two tabs are needed. I guess for computational speed? This should be clarified in the ReadMe file. Otherwise why not use the multiple scenarios all the time and just leave the unwanted fields blank? Also, it should be made clear what happens if you input less than the 20 or 500 water chemistries in those two tabs. They seem to just populate automatically with low default values – but the general user might be confused why data suddenly shows up that they didn't ask for.

As already highlighted it is great that you can see the actual "DeForest" equations. Why not take it a step further and have the slope parameters in separate cells called by this equation. This would show the parameters to the end-user but also allow for ease of revision as new data modify the slopes for the equations. And ultimately since the DeForest papers actually calculate the effect concentrations it would be nice to have a column for the non-normalized EC20 results as well. I think that is a more relatable parameter than the normalized values.

Now for a bigger "ask". It would be nice to link this spreadsheet to an equilibrium solver to predict solubility of common aluminum phases or even just amorphous gibbsite. This would not be a hard model to build. The results would be "just for information" but going forward it could help inform that question about inert and reactive solid aluminum. Linking the geochemistry predictions would also allow assessment of soluble versus particulate exposures.

3b. Do you have any other suggestions to improve the ease of use?

I do not have any suggestions to improve ease of use. It is pretty easy to use. If you can use a spreadsheet you can use this calculator. The ReadMe needs some improved documentation, as I've indicated above, but this is a great tool.

III. SPECIFIC OBSERVATIONS

Model Name	Tab	Cell	Comment or Question
			[The reviewer did not provide any specific observations or comments on the models.]

Figures

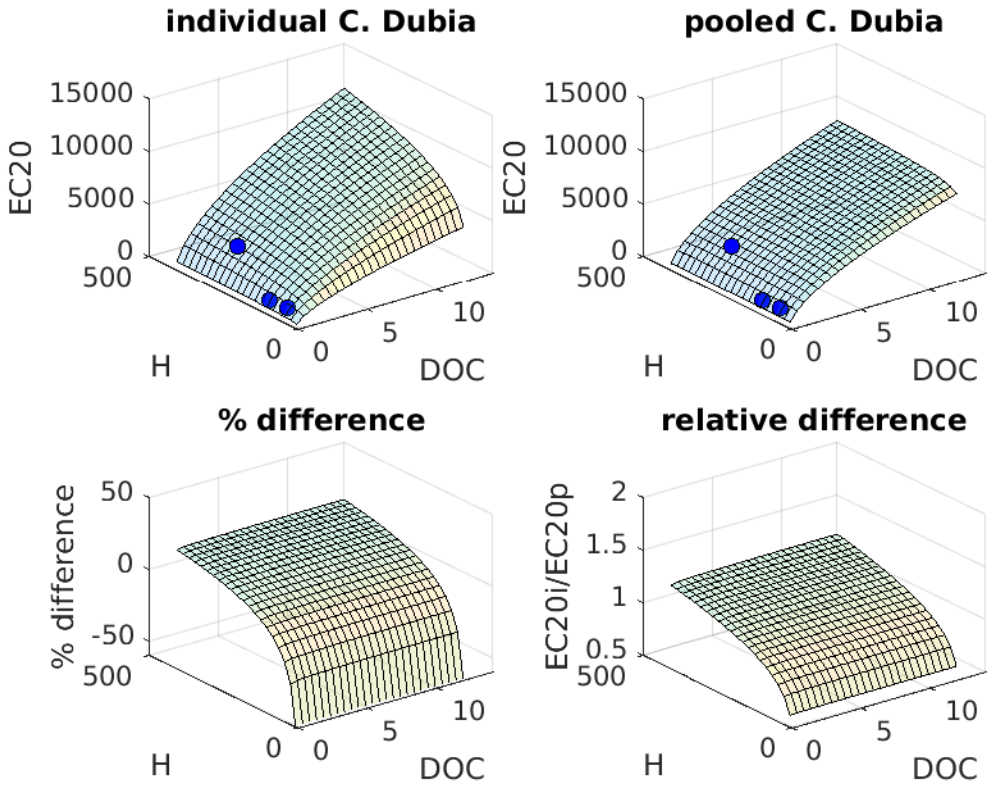


Figure 8: C. Dubia MLR predicted EC20 values at pH 6.3. The top left plot is determined using Equation 2 individual EC20 (EC20i) from the DeForest memo. The top right plot is determined using Equation 6 for pooled EC20 determinations (EC20p). The range of DOC and H were selected to match the calibration range of the MLR model. The blue dots correspond to chronic C. Dubia data from the chronic tab of the Criteria Calculator spreadsheet. The % difference plot corresponds to 100*(EC20i-EC20p)/EC20i and the relative difference is EC20i/EC20p.

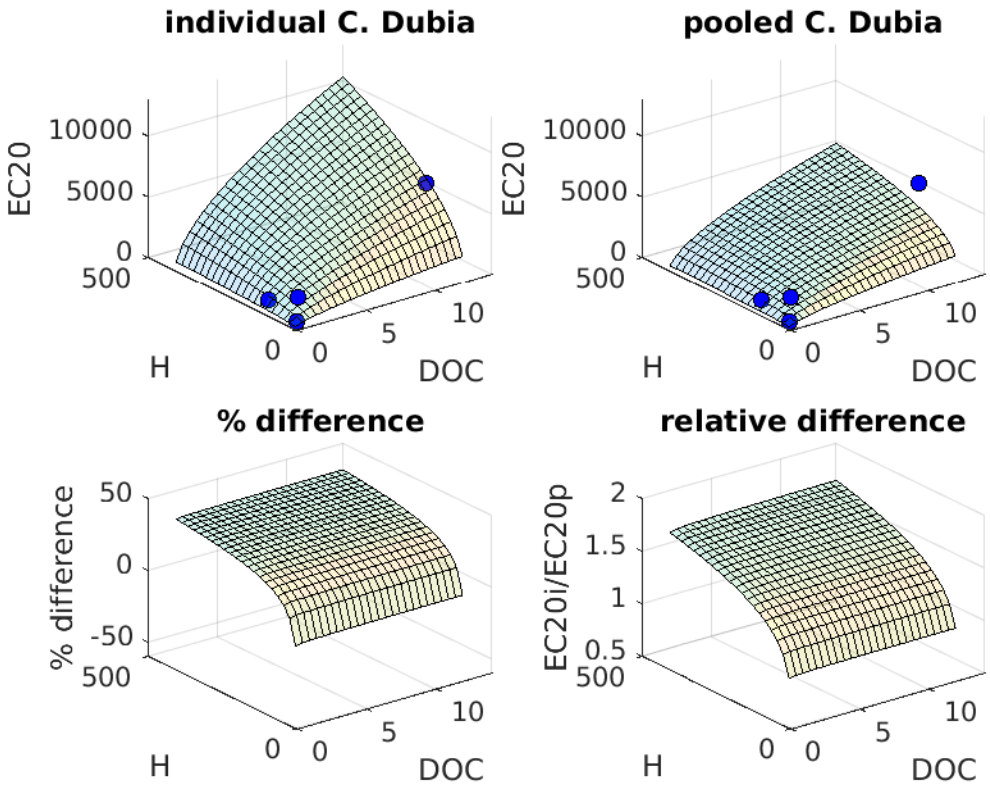


Figure 9: C. Dubia MLR predicted EC20 values at pH 7. The top left plot is determined using Equation 2 individual EC20 (EC20i) from the DeForest memo. The top right plot is determined using Equation 6 for pooled EC20 determinations (EC20p). The range of DOC and H were selected to match the calibration range of the MLR model. The blue dots correspond to chronic C. Dubia data from the chronic tab of the Criteria Calculator spreadsheet. The % difference plot corresponds to 100*(EC20i-EC20p)/EC20i and the relative difference is EC20i/EC20p.

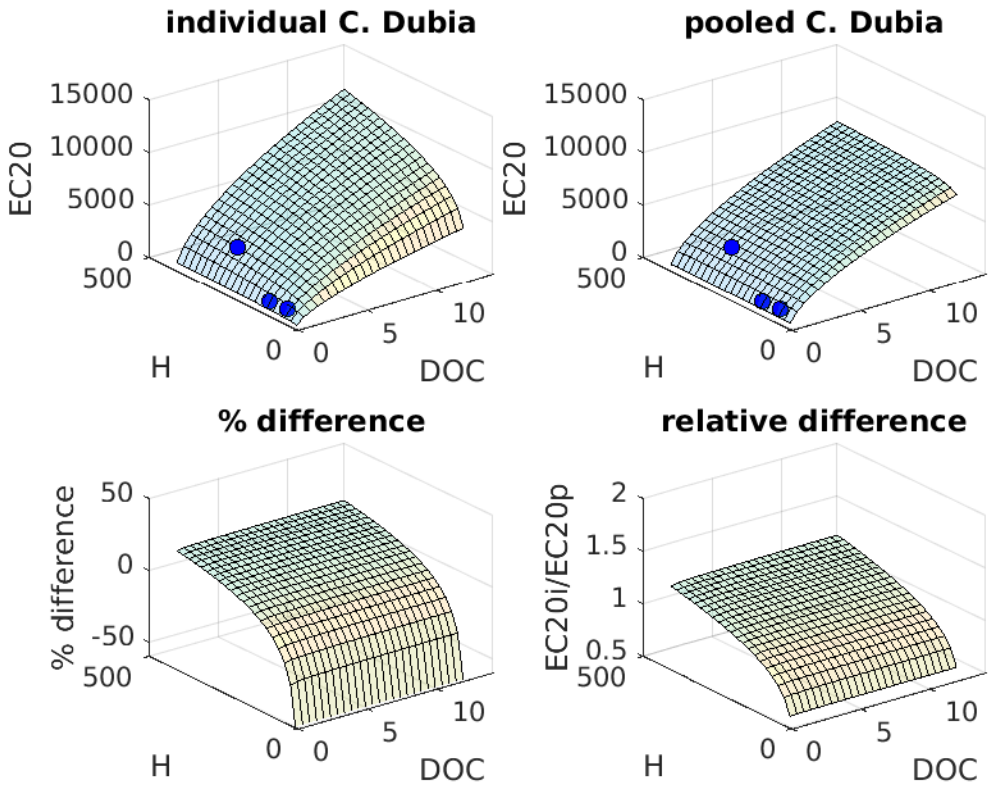


Figure 10: C. Dubia MLR predicted EC20 values at pH 8. The top left plot is determined using Equation 2 individual EC20 (EC20i) from the DeForest memo. The top right plot is determined using Equation 6 for pooled EC20 determinations (EC20p). The range of DOC and H were selected to match the calibration range of the MLR model. The blue dots correspond to chronic C. Dubia data from the chronic tab of the Criteria Calculator spreadsheet. The % difference plot corresponds to 100*(EC20i-EC20p)/EC20i and the relative difference is EC20i/EC20p.

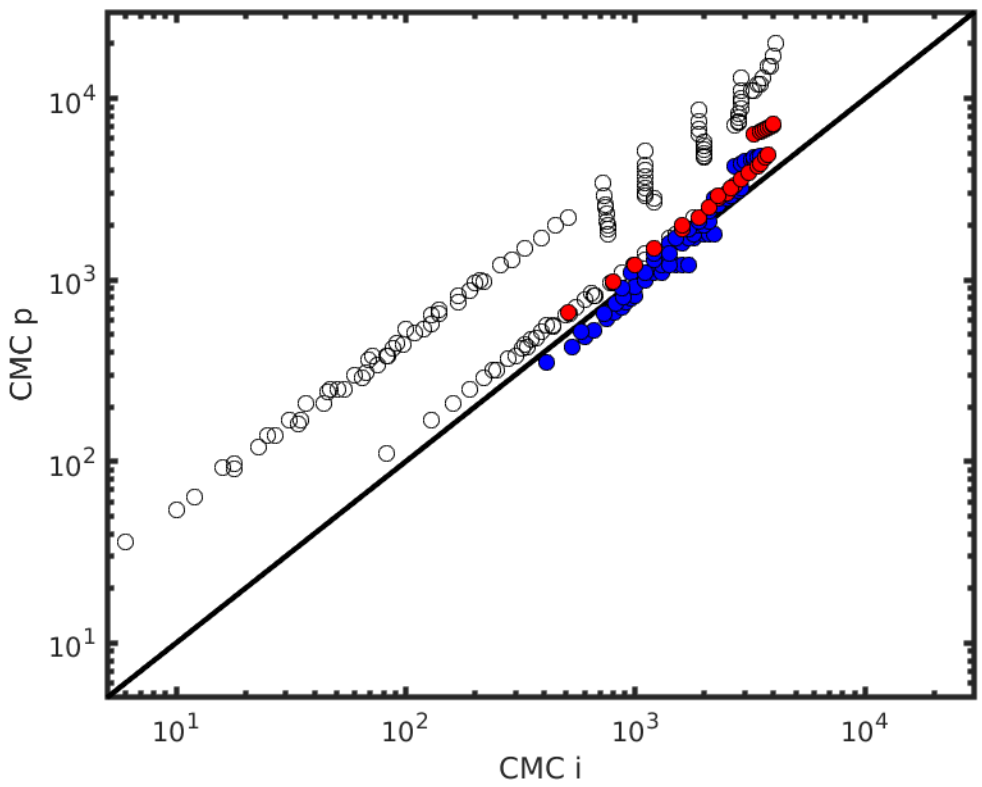


Figure 11: CMC determined using the individual spreadsheet (CMCi) and using the pooled approach (CMCp). The open circles represent all the calculations for the data in Appendix A. The closed symbols are for all the pH data in the range the model was calibrated. The red data are for high DOC (>5) and the blue data are for low DOC (<5).

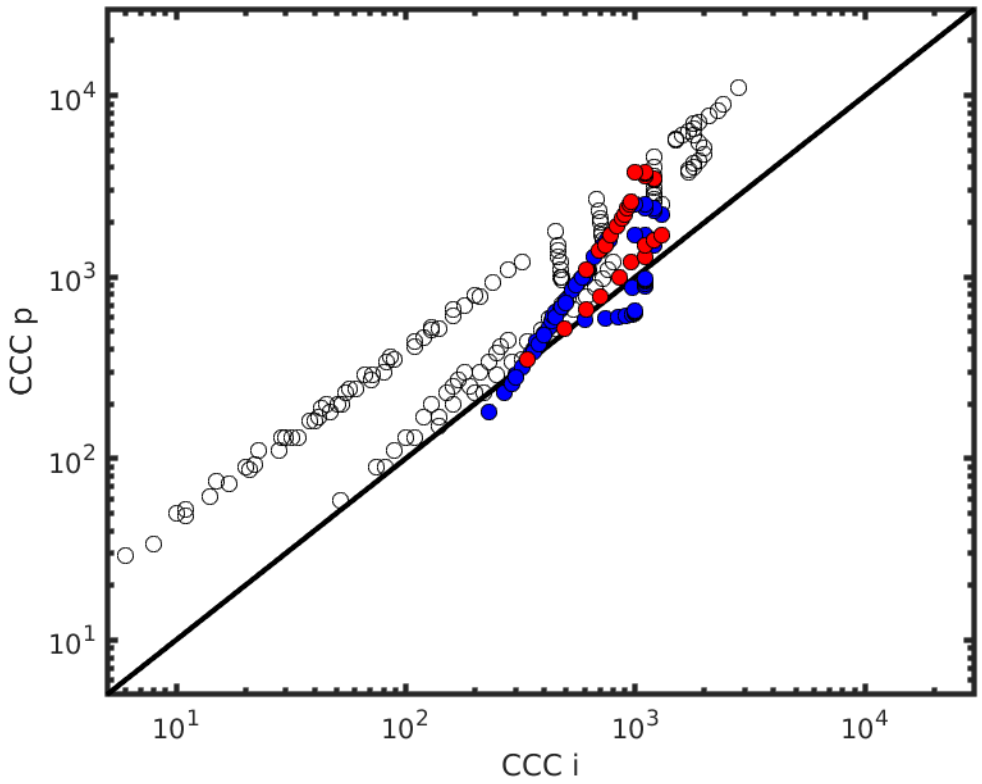


Figure 12: CCC determined using the individual spreadsheet (CCCi) and using the pooled approach (CCCp). The open circles represent all the calculations for the data in Appendix A. The closed symbols are for all the pH data in the range the model was calibrated. The red data are for high DOC (>5) and the blue data are for low DOC (<5).

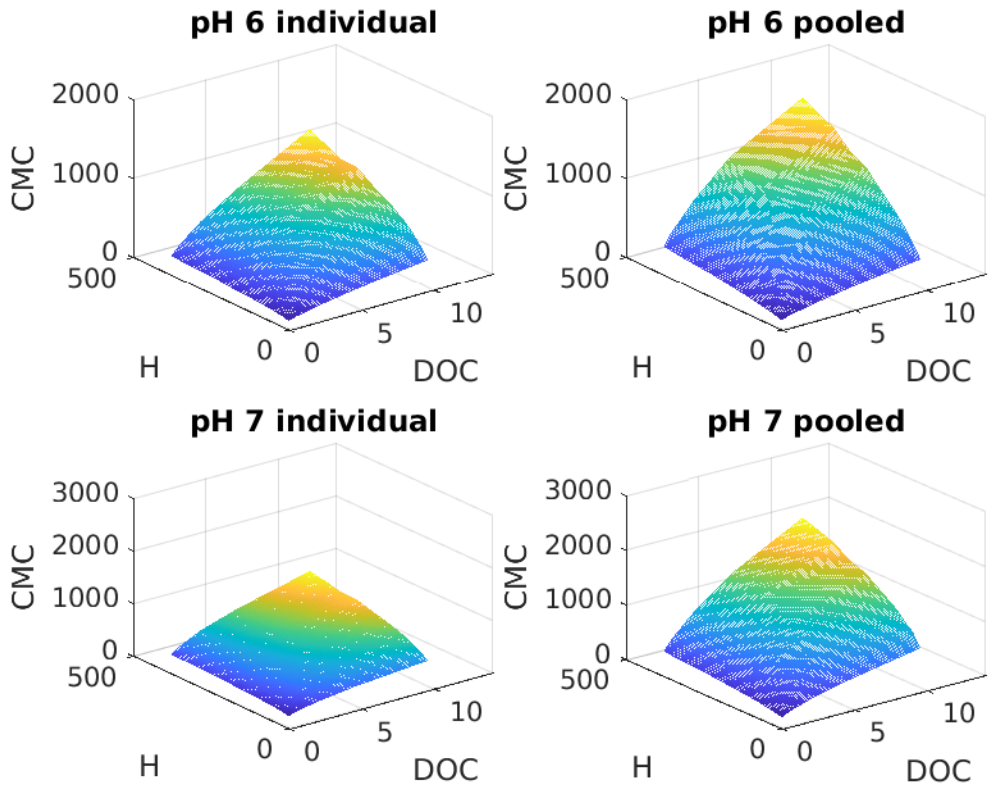


Figure 13: pH 6 and 7 Appendix A data used to derive CMC values as a function of hardness (H) and dissolved organic carbon (DOC). The results from the individual spreadsheet are shown on the left and for the pooled data are shown on the right.

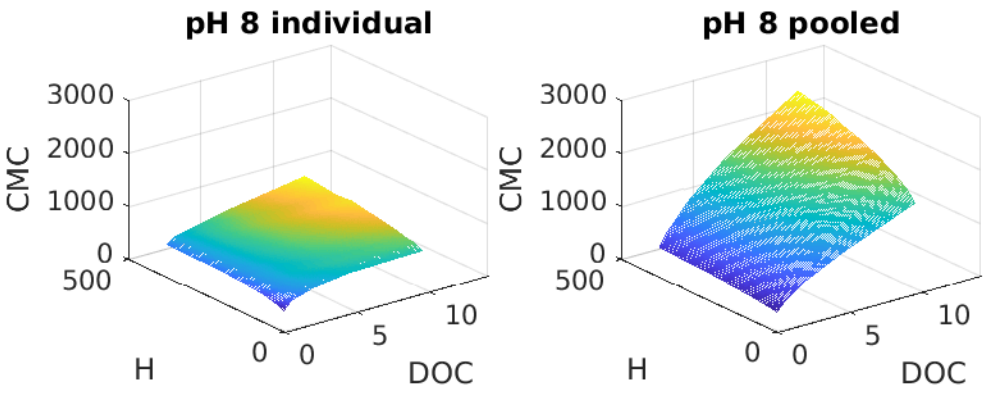


Figure 14: pH 8 Appendix A data used to derive CMC values as a function of hardness (H) and dissolved organic carbon (DOC). The results from the individual spreadsheet are shown on the left and for the pooled data are shown on the right.

Review By:
Reviewer 4

Peer Review Comments on Aluminum Criteria Model

Reviewer 4

I. GENERAL IMPRESSIONS

The use of multiple linear regression (MLRs) in metals criteria is an important step for translating the advances of biotic ligand modeling (BLMs) and related bioavailability research into functional criteria. Particularly with aluminum, they are a huge step forward from the old pH groups and can be both predictive of toxicity when exceeded, and protective of aquatic life uses when met. EPA has successfully used nonlinear regressions for many years with their ammonia criteria, and the educated public (i.e., dischargers, regulators) should have no problem working with these. The new toxicity dataset development and comprehensive data reduction and modeling are exemplary and hopefully harbingers for approaches with other outdated criteria.

This review focused on comparing the performance of two MLR models. The outputs of the two models were often dissimilar, which was not expected. Comparisons with BLM outputs and other comparisons of MLR outputs with test calculations and natural waters suggested that the individual or "non-pooled" MLR models has the better performance of the two. It was not clear that the pooled model would be as protective as intended by the guidelines for developing water quality criteria.

Unfortunately, the severely compressed review schedule and my overlapping field work prevented a more in-depth review of the underlying math, and precluded taking time to ask the developers if I was interpreting and using the model correctly. Some of my criticisms could well be off the mark owing to the haste of this review. I did see the 12 September 2018 email that there was a correction to the memo and model, but with my overlapping field work and the long processing times to run the model, I did not have opportunity to go back and repeat my analyses before the 20 September 2018 deadline.

II. RESPONSE TO CHARGE QUESTIONS

1. Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.

1a. Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?

Yes. The new toxicity data fills gaps in the tested water quality conditions that were lacking earlier.

1b. Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.

It's hard to say with confidence. Certainly, in the DeForest and others' update memo, the pooled model performs very well fitting the Ceriodaphnia and fathead minnow data. However, in comparisons

between the pooled model, the non-pooled model, and the aluminum BLM (Santore et al. 2018), the outputs were sometime quite different. Conceptually, these patterns should be similar between the models. They weren't. Unfortunately, in this type of comparison, while the comparisons are reassuring when they are similar, when they are dissimilar it is not obvious why or which model is more believable. However, some aspects of the pooled MLR do seem amiss, with the flat response for hardness and a much greater magnitude of change for the DOC than for the individual slopes MLR or the BLM. Generally, the performance looks better for the non-pooled model, but that would have to be weighed against any advantage of reduced complexity and possibly better response from stakeholders for the pooled model.

1c. Does the pooled model behave similarly as the non-pooled models?

Sometimes it is similar, but at other times the models are quite different. I looked at the patterns between the models in several ways – comparing to each other and the BLM (Figure 1), comparing their patterns in natural waters (Figure 2), comparing their performance with the test values provided here (Figure 3) and comparing back to the Ceriodaphnia toxicity data.

2. Using the data provided in the Appendix A, please complete a side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

2a. Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

The combinations of pH, DOC, and hardness values provided in Appendix A is a similar type of evaluation as that I used with the BLM responses in Figure 1. In Figure 2, the best agreement is with the water quality conditions most commonly represented in the datasets and used to develop the models (pH 6-7 and pH 8 at low DOC), so agreement in this range is expected.

The magnitude of difference between the models is substantial in some circumstances. For instance, with DOC the non-pooled model has toxicity sharply reduced (exponential increase in CCC) as DOC increases from 0.1 to about 2 mg/L, followed by a reduction in slope and slow increases. The non-pooled values steadily and steeply increase (Figure 1). The non-pooled CCC is about 500 µg/L by 2 mg/L DOC and only increases to 700 by 12 mg/L DOC. In contrast for the same values (2 and 12 mg/L DOC) the pooled model predicts much higher values, 900 and 2600 µg/L. The BLM predicts a linear reduction in toxicity (that is, a linear increase to the EC20 values) over this same range but the absolute values are much lower, about 70 to 250 µg/L for DOCs of 2 and 12 respectively (Figure 1). Granted it's not completely correct to compare CCC and Ceriodaphnia responses, but Ceriodaphnia are reasonably sensitive for the dataset (4th out 13 taxa) their EC20s should be slightly higher than the CCC for the same conditions. In figure 1, they generally were not higher.

2b. Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.

From the comparisons here, the non-pooled model appears to have the "better" (or at least more logical) performance of the two. The exponential rise in the CCC in the pooled model with increasing pH is

unexpected. The expectation is that total Al will be least toxic at circumneutral pH and start becoming more toxic at high pH. This is sort of captured in the BLM and non-pooled MLR. The magnitude of toxicity mitigation with DOC is much greater than that predicted by the BLM or non-pooled model, and the non-response to hardness in the pooled model suggests a glitch in this version.

2c. Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?

No, not consistently. It appears that the pooled MLR Aluminum criteria model would work well in waters with low to circumneutral pH and with relatively low DOC waters. In scenarios with high pH or high DOC the performance of the pooled model seems questionable, based on comparisons to the other two models. This is surprising, because the model fits are very similar between the species-specific and pooled MLRs in the DeForest 24August2018 memo and the data used in the model fitting covered the pH and DOC ranges of interest well (pH 6.3-8.7 and DOC 0.1 to 12 mg/L). This good agreement between the models and the protectiveness toward the sensitive taxa (C. dubia) used to develop it is illustrated in Figure 3. When the resultant CCCs from the species-specific models and the C. dubia EC10s from the updated toxicity data set (DeForest memo) are plotted together, the models fall on top of each other and the EC10s all fall at or just above the criteria values, just like they are supposed to (Figure 3). The textbook perfect behavior from the model data and the strange differences with the test "data" raises the specter that the MLRs may be overfit.

However, the "data" from Appendix A and those used with the Santore ranges in Figure 1 are not "data" at all – they are contrived values selected to examine model calculations over a range of potential real world values. It is useful to compare real world data similarly. Figure 4 shows MLR CCC values for four streams for which appropriate time-series data could easily be found, and that might be close to the ranges of applicability (Figure 4). Data are from the U.S. Geological Survey's National Water Information System, http://waterdata.usgs.gov/nwis/. The relatively high pH, low DOC Snake River in Idaho showed good agreement between the two MLR approaches (Figure 4A). The other three streams are from low hardness, low pH waters in the Adirondacks and in Maine. The Wild River in Maine has variable and moderate DOC (1.4 to 12 mg/L) and the two Adirondack, New York streams have high DOC. The pooled MLR criterion values were consistently higher than the individual-slopes MLRs for these low pH, high hardness waters. The Adirondack streams also have extensive Al data, likely because of concerns of toxic episodes during acid rain episodes. For the period of record, the great majority of the total Al measurements were below both CCC models, with occasional exceedances of the lower, individual model (Figure 4).

Finally, as noted in DeForest et al.'s (2018) initial presentation of the Al MLR approach, a chronic (60d) brook trout test was highly influential in EPA's older criterion document. This test had a NOEC of 88 µg/L and an LOEC of 169 µg/L, which was a 24% reduction in growth, and a growth reduction EC20 was calculated at about 156 µg/L. In DeForest et al.'s (2018) original MLR, the HC5 (the CCC by a different name) was calculated at 117 µg/L. This would seem a reasonable degree of protection for a sensitive species. At times when the Al approached criteria, the conditions were presumably stressful and result in reduced growth. However, such conditions presumably are only temporary during freshets and the fish populations would not be much harmed. In the updated criteria using the individual-slope MLR, for those conditions a CCC of 160 µg/L was calculated which is now as high as the EC20, which is a severe effect. The pooled slope MLR yields a CCC of 200 µg/L for the test conditions. This does

not seem fully protective for a species that is of conservation concern in the southern Appalachians and other parts of its native range.

2d. Please provide suggestions of alternate approaches, if any.

Using the pooled model with caps on the questionable parameters might allow EPA to use the simpler pooled model-based criteria that would be easier for stakeholders to understand and use. Just where to set those caps would take a more careful examination of the model performance and data than is possible in the excessively short time allotted for this review. However, from figure 1 in particular, it looks like a cap for pH would be in the neighborhood of 8.5 and for DOC in the neighborhood of 2 mg/L. (Recall that a DOC of 2 in the pooled model may produce a CCC higher than that from a DOC of 12 in the non-pooled model (910 vs. 690 µg/L for hardness 25 mg/L, pH 7.5, Figure 1).

3. Ease of Use:

3a. Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?

The care and skill that went into the macro enabled spreadsheets is obvious. However, for the "over 20 scenarios" runs, it took 5-10 minutes for a run. That was excruciating, trying to do multiple runs and it wasn't obvious whether it was running or had hung. Stakeholders will send EPA hate mail if their computers are locked up for 10 minutes after each time they click run. From the "Summary Sheet" tab, it looks as though once the modeling and criteria questions are set, it will no longer be necessary to normalize the entire SSD, and a straight "xlsx" equation will be sufficient? If not, I recommend striving for that; otherwise there will be endless complaints.

Also, for those who work in organizations with centralized IT departments (a widespread malady), they may have trouble with macro-enabled Excel sheets. (I did, Figure 5).

3b. Do you have any other suggestions to improve the ease of use?

Not within the limited time available for review.

III. SPECIFIC OBSERVATIONS

Model Name	Tab	Cell	Comment or Question
			[The reviewer did not provide any specific observations or comments on the models]

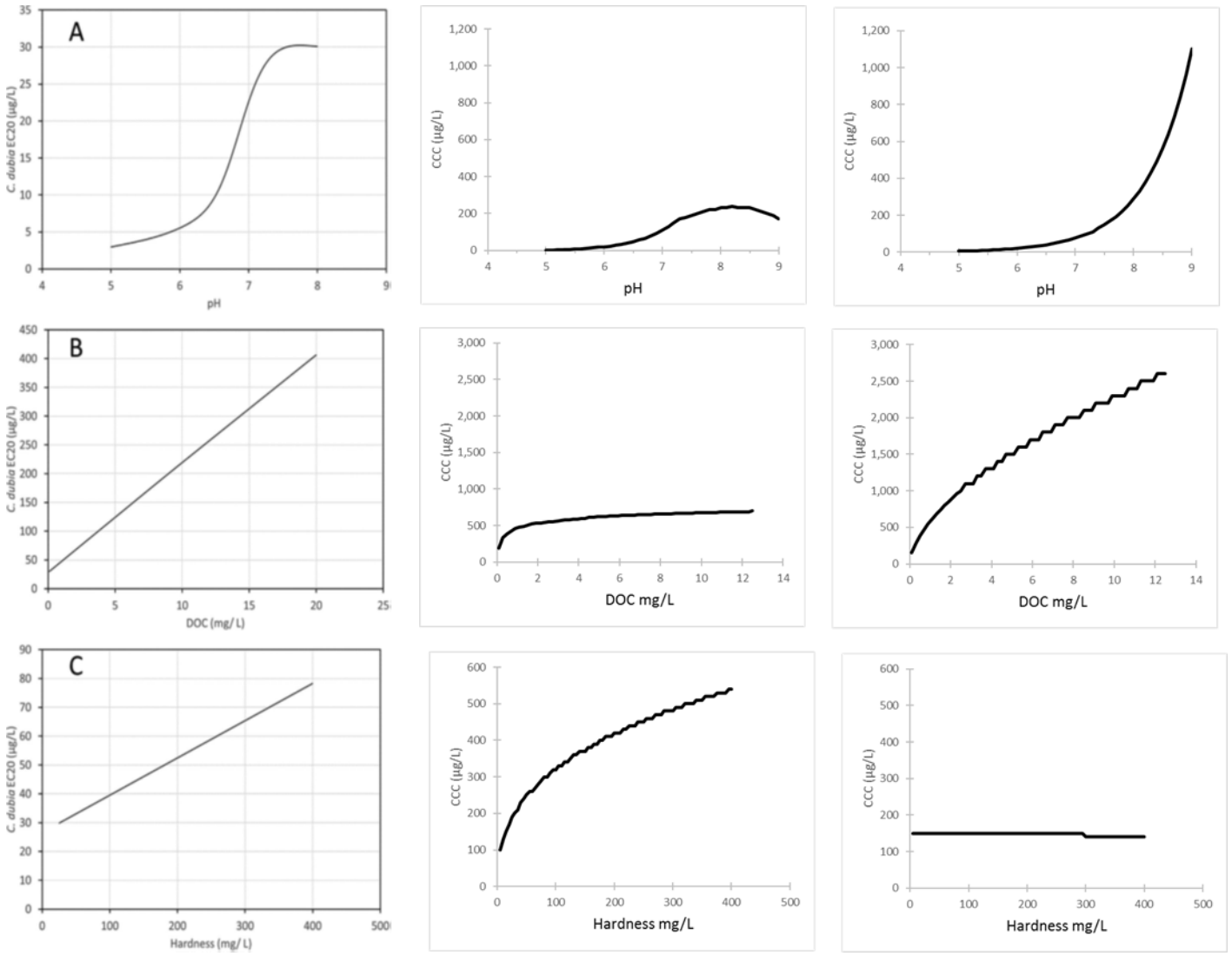


Figure 1. Variation in predicted toxicity patterns as a function of water quality showing the response in aluminum (Al) bioavailability for either the Al BLM (Santore et al. (2018), left); the individual slopes MLR (center), and the pooled slopes MLR (right) to changes in pH (A), dissolved organic carbon (DOC; B), and hardness (C). Base conditions for each simulation are temperature 20 8C, pH 7.5, DOC 0.1 mg/L, and hardness 25 mg/L. The response patterns between the models are disappointingly different (Warning – vertical axes scales are very different between the BLM and MLR plots.). Jittering is an artefact of the input values chosen for the MLR.

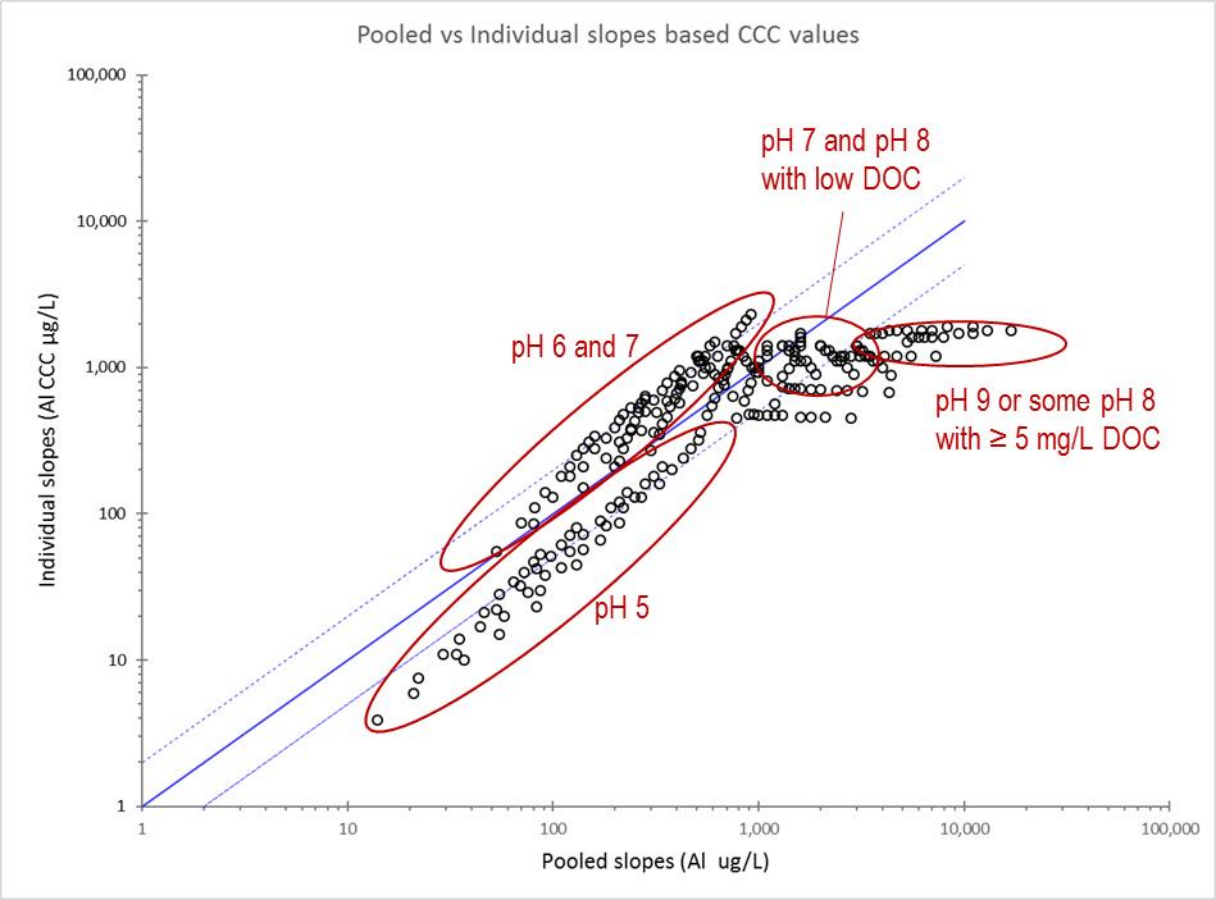


Figure 2. The 250 "Appendix A" test values covering a range of DOC, pH, and hardness values produced CCC values that were surprisingly divergent. 87 (35%) of the pairs differed by >2X and 37 (15%) differed by more than 3X. Poorest agreement was for the extreme values, especially for pH 9 combinations. Best agreement was for the pH 6 and 7 combinations, and pH 8 at low DOC.

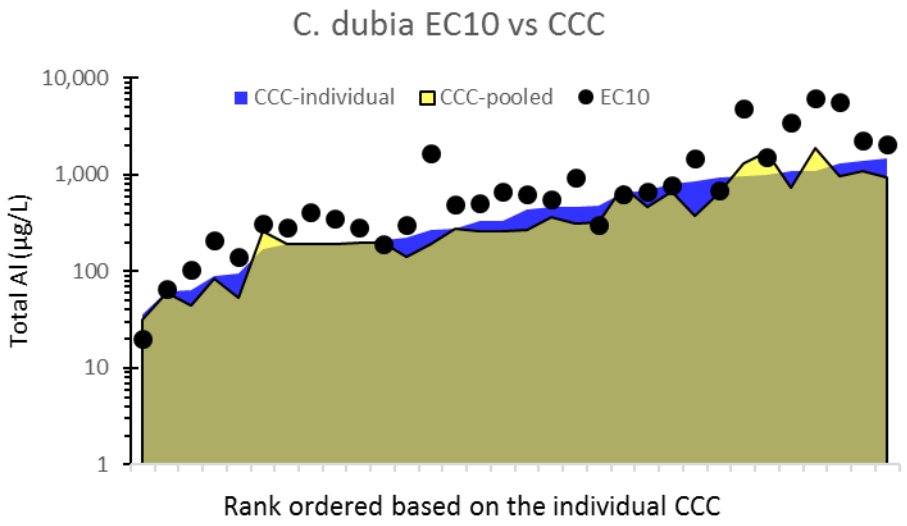


Figure 3. *Ceriodaphnia dubia* toxicity (EC10s) versus the non-pooled or pooled CCC versions. Data from DeForest memo,

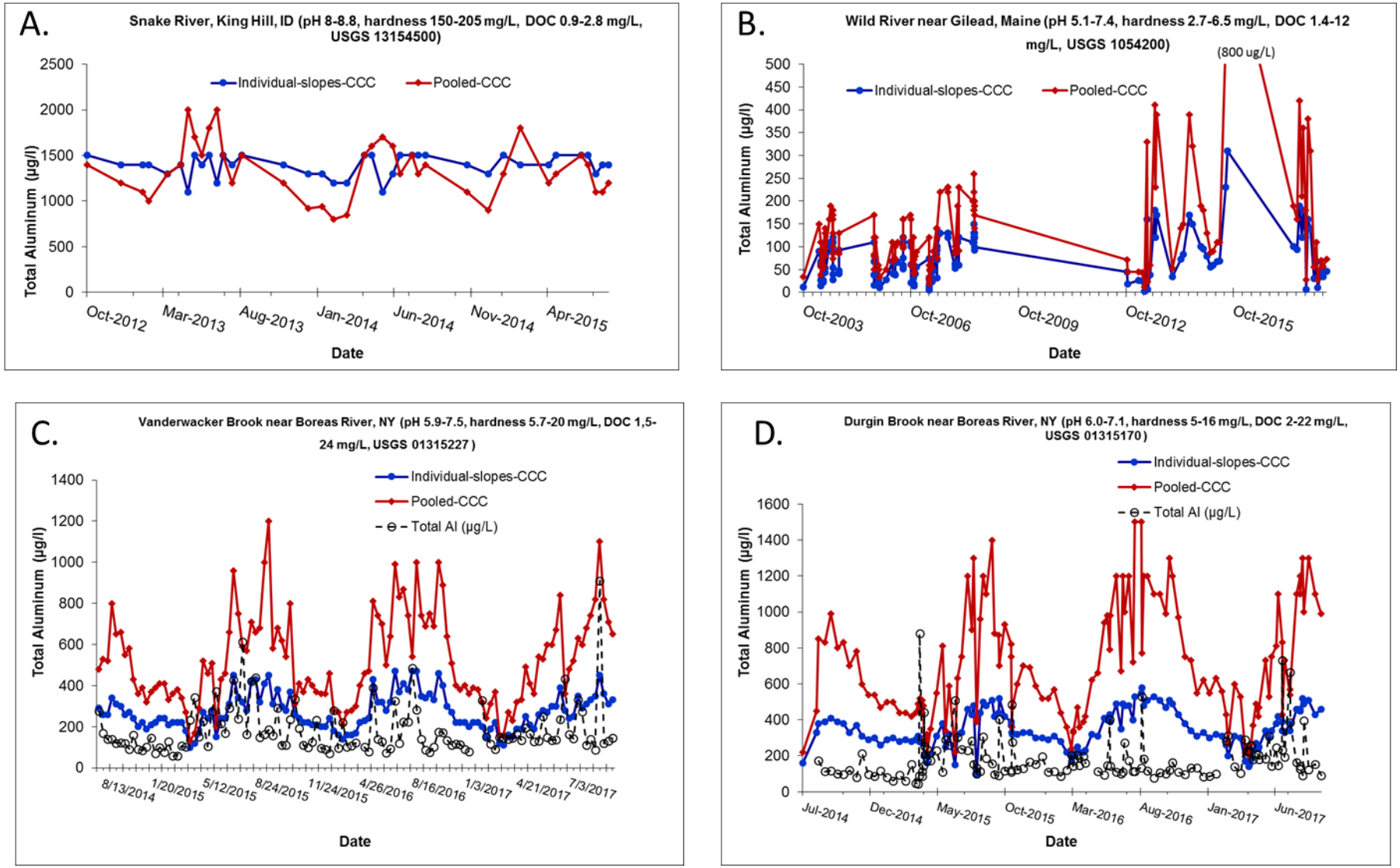


Figure 4. Comparisons of criteria in natural waters. In a river with moderately high pH and low DOC, the two MLR CCC versions were mostly similar; in the low pH waters in which aluminum toxicity is actually a real concern, the non-pooled MLR version tended to be lower.

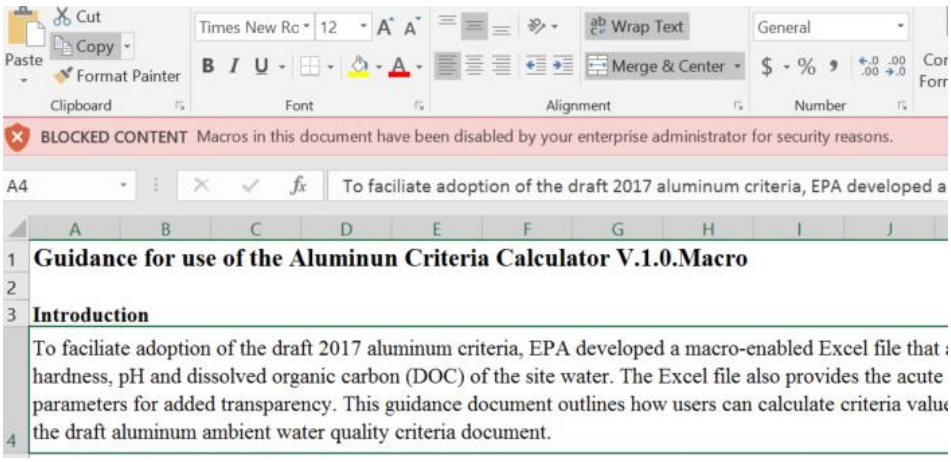


Figure 5. Corporate IT people don't like macro-enable Office files and may disable them just because they can. Reconfiguring to a simple equation would be much preferable for distribution to those who just want to calculate their number.

Reference

DeForest, D.K., K.V. Brix, L.M. Tear, and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. *Environmental Toxicology and Chemistry*. 37(1): 80-90. https://doi.org/10.1002/etc.3922

Review By:
Reviewer 5

Peer Review Comments on Aluminum Criteria Model

Reviewer 5

I. GENERAL IMPRESSIONS

The work is a very well-executed model development based on a highly-screened aquatic toxicity dataset that offers a significant advancement in environmental risk assessment of aluminum in freshwater. The authors of the DeForest et al. 2018 paper and the subsequent peer-reviewed citations represent experienced and qualified experts in the related fields. The enlarged dataset offered in the work of the OSU Aquatic Toxicology Lab has appropriately increased the value and usefulness of the MLR approach, and furthermore allows defendable pooled MLRs. The approach and dataset presented are peer-reviewed and represent our best available knowledge moving forward to update and improve the current three-decade-old approach to quantifying aluminum risk in aquatic ecosystems.

The papers, data, and technical memorandum used in the supporting material present a convincing case for moving forward. Although the actual model spreadsheet would be improved with better notation and comments fields for novice users, and a much better effort at user guidance, the overall MLR model appears well developed.

The model spreadsheet supporting documentation needs work before general distribution since the user base is less than familiar with this approach. The Readme appears written by experts for an audience of users with similar expertise and that is most often not the case at the state regulatory level, especially in smaller states. General release of the criteria calculating model with its present level of documentation may lead to confusion and frustration with many users.

The guidance for this review was somewhat challenging as well. For example the use of “Non-pooled” and “Individual” for the same thing was confusing. The models pre-loaded with scenarios was also somewhat mysterious at first, because I would assume you want the user base to fill in water quality scenarios of concern and run the model for specific results related to their management concerns.

The Pooled Model does not appear to produce results consistent with the output of Non-pooled Model when comparing a side-by-side scenario data set. Hence, unless there is a reason for the rather large non-concordance of the two output sets, possibly due to user error, the Pooled Model would not be appropriate for use and appears to be generally overprotective.

II. RESPONSE TO CHARGE QUESTIONS

1. Please review the DeForest et al. 2018 paper (DeForest, D.K., K.V. Brix, L.M. Tear and W.J. Adams. 2018. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environ. Toxicol. Chem. 37(1): 80-90) and the Memorandum "Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas" dated 8/24/18.

1a. Is it appropriate to integrate the new toxicity data into the MLR equations? If not, why not?

The DeForest et al. 2018 ETC paper is the most comprehensive attempt at developing a model of the aquatic toxicity of aluminum in three decades. The paper develops a multiple linear regression model based on DOC, pH, and hardness conditions that are derived from a robust, screened aquatic toxicity data set. The regression analysis was on data from P. subcapitata, C. dubia, and P. promelas. The predictive MLR model demonstrated the ability to predict chronic toxicity with variable DOC, pH, and hardness conditions within a factor of two for 91% of the tests explored. There have been four citations of this paper in the very short period since its publication – achieving a highly cited notation. However, most of these have one of the authors as a co-author, and two contain the additional Al aquatic toxicity data of Gensemer et al. The additional co-authors on these papers as well as their publication in the leading journals in the field suggest the research is if the highest quality. The MLR approach thus demonstrates in this peer-reviewed paper, its viability for use in a regulatory science arena related to risk management of the freshwater aquatic toxicity of aluminum.

It is appropriate and necessary to integrate the new toxicity data into the MLR equations. The OSU Aquatic Toxicology Lab data completes and enhances the MLR robustness specifically because of the targeted test quality and range of water quality conditions of the data set. The regulatory science community is fortunate that this data set became available during the review phase of the 2017 Draft Aquatic Life Criteria for Aluminum in Freshwater. As demonstrated in the September 12, 2018, updated August 24, 2018, Memorandum, Updated Aluminum Multiple Linear Regression Models for Ceriodaphnia dubia and Pimephales promelas, the integration of the new toxicity data expands the DOC, pH and hardness ranges where the MLR can be reliably used.

1b. Please comment on whether the pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized.

The pooled (fish and invertebrate captured in one equation) and non-pooled (fish and invertebrate captured by separate equations) MLRs are appropriately parameterized. The published DeForest et al. 2018 paper, and the subsequent works that cite this paper, develop a significant level of background in the peer-reviewed literature about the dominant water quality characteristics influencing aluminum aquatic toxicity. In the MLRs, ln(DOC), pH, and ln(Hard) are used in a common and defendable manner to define probability distributions in the scope of this risk assessment. The ground-truthing of the model with toxicity testing results suggests robustness.

"…the updated dataset supported development of a pooled MLR model that had comparably high adjusted and predicted R2 values compared to the species-specific MLR models. The pooled models also provided a similar level of accuracy in predicted EC10s and EC20s compared to the species-specific models."

1c. Does the pooled model behave similarly as the non-pooled models?

No, see Question 2 results below. When the conditions of Appendix A are copied into fields C, D, and E the CMC and CCC results generated in columns H and I for the Non-Pooled and Pooled models are quite different.

The model authors state in their technical memoranda:

"…the updated dataset supported development of a pooled MLR model that had comparably high adjusted and predicted R2 values compared to the species-specific MLR models. The pooled models also provided a similar level of accuracy in predicted EC10s and EC20s compared to the species-specific models."

"The pooled aluminum MLR models provided a similar level of accuracy in EC10 and EC20 predictions for C. dubia and P. promelas as the species-specific MLR models. For C. dubia, the percentage of predicted EC10s and EC20s within a factor of two of observed was unchanged (94% and 97%, respectively) (Figure 3). For P. promelas, the percentage of predicted EC10s and EC20s within a factor of two of observed decreased from 94% to 90% for EC10s and from 97% to 94% for EC20s (Figure 4)."

"Because the pooled MLR model performs well, there no longer appears to be any benefit in using species-specific MLR models for ambient water quality criteria development.(my emphasis) Use of the pooled model would preclude the need to recalculate the aluminum genus sensitivity distribution for each water chemistry of interest. Instead, chronic aluminum criteria could be condensed to a single equation, such as the existing hardness-based criteria for several metals or the pooled MLR-based criteria for copper described in Brix et al. (2017). The slopes from the recommended pooled models are:

- *Pooled slopes from EC10 model:*
  - *ln(DOC) = 0.645*
  - *pH = 1.995*
  - *ln(Hard) = 2.255*
  - *ln(Hard)×pH = -0.284*
- *Pooled slopes from EC20 model:*
  - *ln(DOC) = 0.592*
  - *pH = 1.998*
  - *ln(Hard) = 2.188*
  - *ln(Hard)×pH = -0.268"*

C. dubia

$$ln(EC10) = -8.618 + 0.645 \times ln[DOC] + 1.995 \times pH + 2.255 \times ln[Hard] - 0.284 \times ln[Hard] \times pH \quad (5)$$

$$ln(EC20) = -8.555 + 0.592 \times ln[DOC] + 1.998 \times pH + 2.188 \times ln[Hard] - 0.268 \times ln[Hard] \times pH \quad (6)$$

P. promelas

$$ln(EC10) = -7.606 + 0.645 \times ln[DOC] + 1.995 \times pH + 2.255 \times ln[Hard] - 0.284 \times ln[Hard] \times pH \quad (7)$$

$$ln(EC20) = -7.500 + 0.592 \times ln[DOC] + 1.998 \times pH + 2.188 \times ln[Hard] - 0.268 \times ln[Hard] \times pH \quad (8)$$

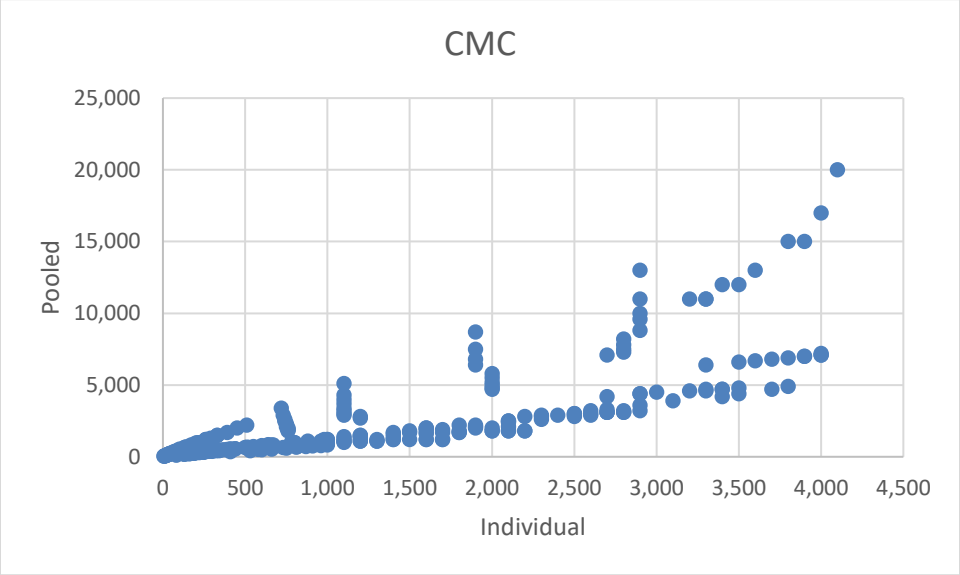
In these analyses, the authors appear to successfully defend use of a pooled MLR model in large part due to the expanded OSU data set made available in 2018. However, when same pH, DOC and Hardness field scenarios are loaded into the Non-pooled and Pooled models, the CMC and CCC results appear considerably different (see #2 below).

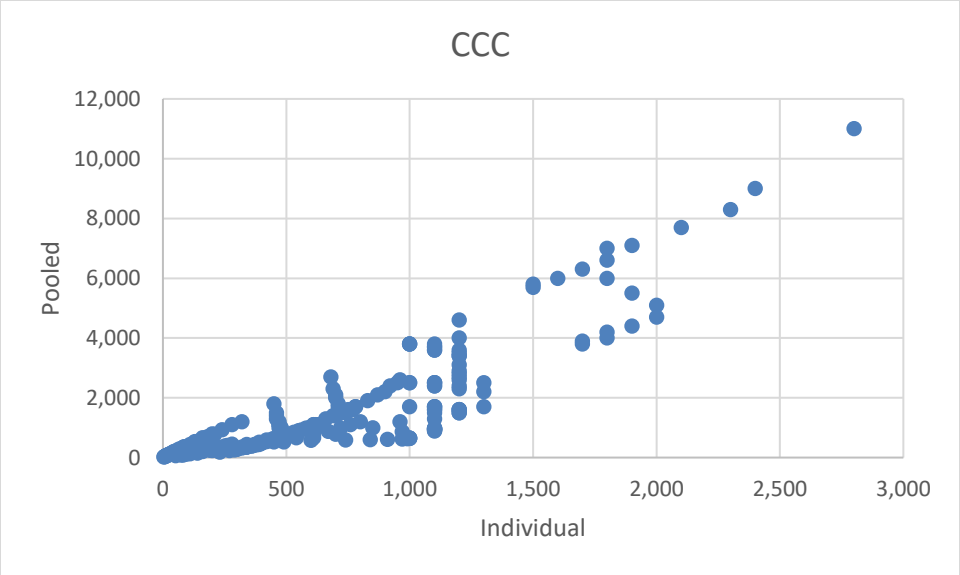
2. Using the data provided in the Appendix A, please complete side-by-side comparison of the results of the Non-pooled Aluminum Criteria Model and the Pooled Aluminum Criteria Model criteria derivations.

2a. Please draw conclusions regarding the differences in the values (CMC and CCC) generated and explain your rationale.

The water conditions listed in Appendix A were pasted into columns C, D, and E of the Non-Pooled Model (individual slopes) and the Pooled Model (pooled slopes). The model calculated CCC and CMC were copied into a self-constructed Side-by-Side comparison spreadsheet for analysis and inspection. The data were plotted in a scatter graph for visual trend analysis and were further analyzed by fundamental statistical analyses. I did not attempt to quantify or analyze the difference any further.

Upon generation of CCC and CMC values for the range of water conditions in Appendix A, there appears to be a significant positive bias for the pooled model result over the individual model result. The positive bias is generally smallest at higher water hardness levels, although more advanced multiparameter analyses may yield a different outcome.





These scatter plots possibly indicate relatively poor concordance of the output of the two models. Further comparison of the CMC and CCC results generated for the data of Appendix A input into the Non-Pooled Model and the Pooled Model, shown in the table below, yield the following:

- An average CMC Al concentration difference of 1.3 mg/L ranging from a minimum of 0.5 to 15.9 mg/L between the Non-Pooled Model and the Pooled Mode.

- An average CCC Al concentration difference of 0.81 mg/L ranging from a minimum of 0.36 to 8.2 mg/L between the Non-Pooled Model and the Pooled Mode.

- An average CMC Al concentration ratio of 0.64 ranging from a minimum of 1.4 to 0.17 mg/L between the Non-Pooled Model and the Pooled Mode.

- An average CCC Al concentration ratio of 0.58 ranging from a minimum of 1.6 to 0.20 mg/L between the Non-Pooled Model and the Pooled Mode.

CMC		CCC		CMC		CCC	
Difference				Ratio			
-1,314	avg diff	-808	avg diff	0.640	avg ratio	0.580	avg
500	max	360	max	1.417	max	1.571	max
-15,900	min	-8,200	min	0.172	min	0.200	min

These analyses suggest that in practical use, the Non-Pooled Model and the Pooled Model would yield considerably different results, averaging 1.3 and 0.6 mg/L Al for the water conditions of Appendix A, potentially with up to five-fold differences in individual case analyses. This exercise demonstrates that practical application of the Pooled Model may not rise to the author's description "*Because the pooled MLR model performs well…"*

Thus, I can only conclude that in practical application, if my use of the MLR models was not in error (The user guide Readme was not particularly helpful in this regard), the Pooled Model results are uncomfortably different from the Non-Pooled Model.

2b. Please evaluate the scientific appropriateness of using a pooled model vs. non-pooled model and explain the rationale of your opinion.

Knowing the degree of expertise of the MLR model authors, I was encouraged when they wrote: *"Because the pooled MLR model performs well, there no longer appears to be any benefit in using species-specific MLR models for ambient water quality criteria development.*" Furthermore, the model authors sufficiently back up this observation with performance metrics in their technical analysis memo. However, unless my use of the model was not correct (please better guide your users to where the inputs and outputs are), the Pooled Model does not seem to perform to the required level of "appropriateness," under the assumption that the model dynamics for the Individual or Non-Pooled Model is inherently more robust.

2c. Would the pooled MLR Aluminum Criteria Model be sufficiently robust and protective to use as the underlying basis for the aluminum aquatic life water quality criteria?

With the experience and side-by-side data generated and outlined above, the Pooled MLR would not be sufficiently robust and typically over-protective.

2d. Please provide suggestions of alternate approaches, if any.

Unless I misused the models, only the Non-Pooled Model would be acceptable.

3. Ease of Use:

3a. Please provide any suggestions of how to make an approach easier for a stakeholder (e.g., states) to use, such as improvements to user manual, better upfront input design, etc.?

The guidance for the MLR spreadsheet to be used by stakeholders is far from complete and not particularly informative or useful in its present iteration. I found it frustratingly incomplete for a new user. The model only has a Readme page. For example, my environmental toxicology course students can work their way through California's LeadSpread 8 during risk assessment exam questions due to the quality of the associated manuals and user assistance. (https://www.dtsc.ca.gov/AssessingRisk/LeadSpread8.cfm). Employing spreadsheet comment fields, example calculations and a more intuitive user guide that may be a useful approach for the MLR when risk assessors access the aluminum aquatic toxicity model for the first time. As presented the MLR spreadsheets are not intuitive or easy to use. The model authors have attempted to insert some guidance, however this Readme guidance appears incomplete and only somewhat useful. It took me several hours to orient myself to understand the different input modalities (summary page, multiple, and over-20 multiple). In my experience most model software requires some familiarization time before user efficiency, however the supporting materials for the MLRs are below the median in quality and quantity of the materials provided.

Other comments:
The Readme page is not locked and is editable. Another approach to documentation and model use instruction may be better.

The dual use of "Non-pooled" and "Individual" is confusing.

The model seems to want to run all rows always in the multiple scenario worksheets, since the execution time was about the same for a few scenario entries, with the rest of the cells deleted. I was running the model on a Xeon processor workstation and it took about 5 minutes to run.

Please use the word "output" or "result" to label the model endproduct better.

3b. Do you have any other suggestions to improve the ease of use?

Please see the comments above. I prefer models that clearly point me towards "Inputs" and "Outputs." After spending many hours with this model and supporting materials, I am still not entirely confident I am using it correctly. I had to teach myself what the summary page, multiple, and over-20 multiple inputs were by creating a small data set and applying it to each input mode so I could watch the output fields change to gain user confidence. Well developed tutorials such as the EPA Benchmark Dose support materials offer a template for excellence in user base training.

III. SPECIFIC OBSERVATIONS

Model Name	Tab	Cell	Comment or Question
			No specific observations except those outlined above.