Point-by-point response to Interactive comment on “Quantifying uncertaintyon sediment loads using bootstrap confidenceintervals” by J. I. F. Slaets et al.

T. Kumke (Referee)
thomas.kumke@ucb.com
Received and published: 15 August 2016

Original comments in italic
Responses in non-italic

Overall, this is a very well written paper on an important topic. The modelling approach using bootstrap estimates is an important approach and the authors nicely show the strength of the bootstrap.

We thank reviewer 1 for his positive evaluation of our paper. Specific comments are addressed below:

Here are some minor comments for consideration: (i) Introduction: the introduction seems to be quite exhaustive, for example there is a verylong introduction on uncertainties, this could be surely reduced. On the other hand, serialcorrelation is an important aspect of the modelling approach, this has been hardlymentioned. Although, specific aims of the paper were introduced, I would strongly recommendthat the authors state why their approach is very important for the estimationof loads.

The introduction section on sources of uncertainty for sediment and discharge rating curves has been shortened.
Furthermore, the aspect of serial correlation in sediment rating curves in literature has been given more emphasis in the introduction: “… For sediment concentration, however, flow-proportional sampling is often performed to obtain samples at the highest concentrations. Those observations are usually taken closely together during storms and thus most likely are not independent in time (Slaets et al., 2014). Linear mixed models that account for serial correlation provide an alternative to least squares regression to establish a sediment rating curve for this type of data. Lessels and Bishop (2013) similarly found that the inclusion of a temporal autocorrelation component improved the accuracy and decreased the bias in predictions of total phosphorus and nitrogen river loads. If there is serial correlation in the sediment data, it is necessary to use an adjusted version of the bootstrap that retains the serial correlation in the data intact (Lahiri, 2003). Such methods have already been explored in hydrology in relation to the discharge rating curve: Ebtehajet al. (2010) and Selle and Hannah (2010) uses block bootstrap methods to assess uncertainty on and improve robustness of model parameter estimates for discharge prediction.”
Additionally to the specific aims, we have added a section on the importance of our approach for load estimation: “Combining these aspects, the proposed method provides a means to assess uncertainty on any type of constituent load which was calculated from continuous constituent concentration and discharge predictions estimated with regression-type methods. The approach thus allows load estimates to be reported with an uncertainty assessment, rather than as a point estimate alone, making them informative to end users and decision makers.”

(ii) Bootstrap: I feel that the methodological aspects on the bootstrap couldbe reduced in length. I am sure that most readers are familiar with the basic principles.

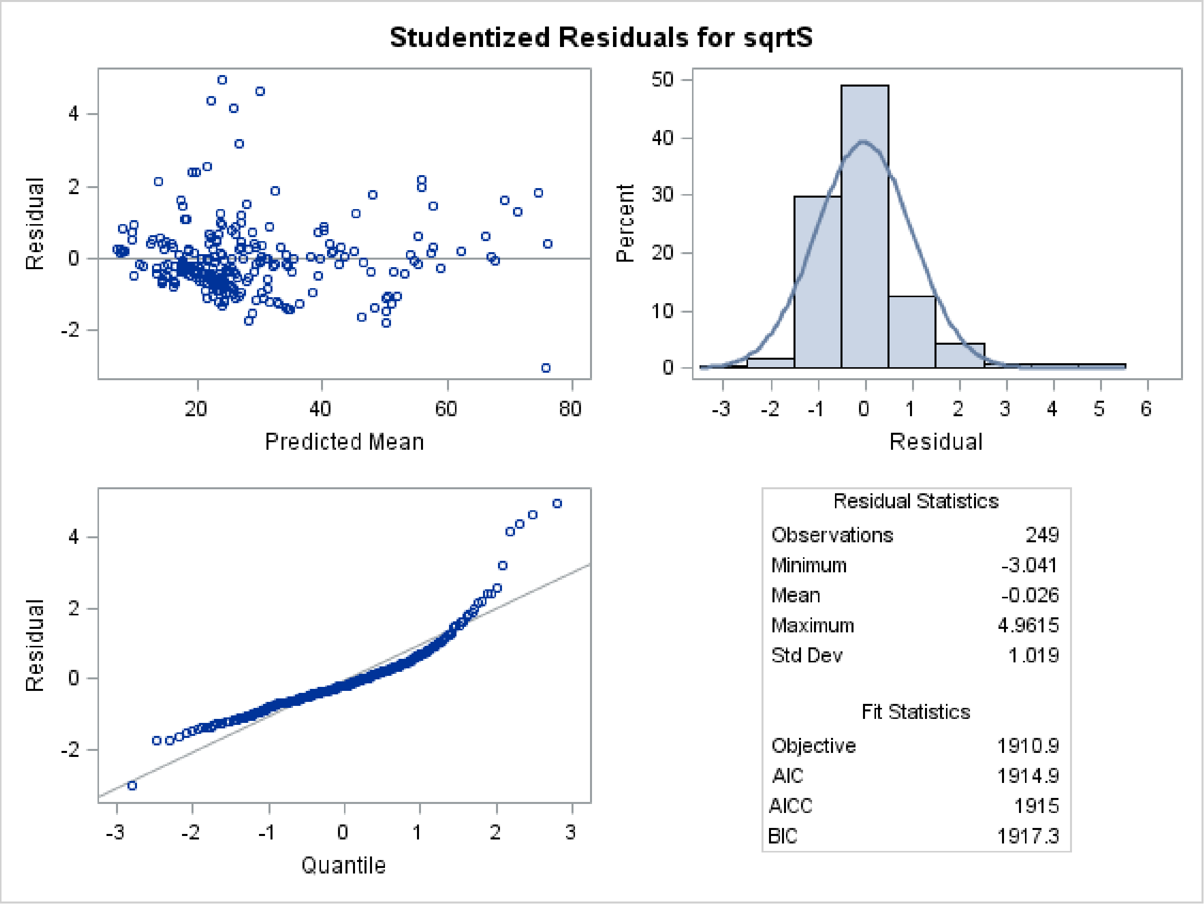
We have shortened the methodological section on the bootstrap and refer the reader to Efron and Tibshirani (1993) for further details.

I have a few issues with some of the used language, for example l. 215: no clear winner? I am not sure whether there are winners or losers in a scientific context. Perhaps rephrasing helps here, eg, it remains unclear which of those specialized methods......

The sentence has been rephrased to: “Among these specialized methods, no preferred method has emerged from literature. Furthermore, many of these methods require a vast set of decisions with regards to for example the block size for which no general recommendation exists. As a consequence, results from different methods are not straightforward to compare.”

(iii) Results: The results of the modelling are nicely summarized. However, a couple ofquestions remains open. The effect of transformations should be evaluated accordingto the introduction, however, only the log-transformation (and back-transformation) was analyzed. What is the impact of other transformations on the CIs (ie, transformations with simple backtransforms)?

The choice for a specific data transformation is driven by the need to stabilize the residual error variance. In our case, the log transformation was the one that was successful in doing so. The log transformation is very commonly used in load studies, and the most frequently used alternative data transformations in load estimation are typically other power transformations such as 1/Y, log(Y), square root, cube root, and fourth root, all of which are Box Cox transformations. These were tested but were not successful in obtaining normality and homoscedasticity of residuals, as is shown here in the diagnostic plots for the square root transformation:



Studentized Residuals for sqrtS

Residual Statistics	
Observations	249
Minimum	-3.041
Mean	-0.026
Maximum	4.9615
Std Dev	1.019

Fit Statistics	
Objective	1910.9
AIC	1914.9
AICC	1915
BIC	1917.3

As the diagnostic plots show that other commonly used transformations in the Box-Cox family are less appropriate for our dataset, we would rather not to present results for another transformation in the Box-Cox family. The exploration of alternative transformations has now been added to the Material and Methods section of the paper: "Other transformations, such as the square root, were inspected using residual plots and were found to be unsuitable for meeting the assumptions of normality and homoscedasticity."
Alternatives to Box Cox transformations such as the exponential family of Manly (for non-positive data) are also available, but far less frequently seen in load estimation: the main reason for data transformation in load studies is heterogeneity of variance, and depending on the strength of this heteroscedasticity, some form of the power Box-Cox family is usually successful in stabilizing the variance.
In order to make our discussion of the implications of the log transformation more generally applicable in the case of other transformations, we have added the following statements to show the similarities for the whole Box-Cox power family of transformations to the discussion section: “The most commonly used data transformations in load estimation are typically other members of the Box Cox power family, such as 1/Y, square root, cube root, and fourth root. Transformations in this family are usually required where the original data exhibit pronounced skewness and heteroscedasticity, which is generally the case in load studies. Therefore for all transformations in the Box Cox family, naïve back-transformation of estimates would similarly result in biased estimates of means on the original scale, as was illustrated with the log-transformation in our dataset.” To further emphasize the importance of the choice of the transformation, which needs to be appropriate to the data at hand, the following text has been added after explaining the various correction options: “Regardless of the chosen correction factor, it is important that homoscedasticity after the transformation is confirmed by visually inspecting the diagnostic plots, as was done in the case of this dataset.”

2000 bootstrap cycles were selected, however, it might be of interest, especially for readers not so familiar with the bootstrap, to explore the effect of the number of cycles on the estimates and CIs.

We thank the reviewer for this interesting suggestion. In order to assess the effect of the number of bootstrap replicates, we have re-run the load estimation for one year with 500, 1000 and 1500 bootstrap cycles. The resulting histograms are shown in Figure 8 and their implications discussed in the results section on the required number of bootstrap replicates. As the crucial point is the smoothness of the histograms in order to have reliable bootstrap estimates, especially in the tails as we are looking at confidence intervals via the percentile method, there is no straightforward relationship with number of bootstrap iterations and wider (or narrower) CI’s, or lower or higher estimates. Rather, the lack of smoothness especially in the tails makes the estimates unreliable, which we have clarified by adding the following text: “Before looking at the bootstrap confidence intervals, the histograms of the bootstrap load estimates were evaluated (Figure 7). The histogram of the 2000 bootstrap estimates looked reasonably smooth, so we concluded that sample size was adequate for the percentile bootstrap. When reducing the number of bootstrap replicates (Figure 8), the change in smoothness, especially in the right tail, becomes visible. Tail smoothness of the empirical distribution is a requirement when using the percentile method to obtain confidence intervals (Efron and Tibshirani, 1993). At 500 bootstrap replicates, the centre of the distribution displays lack of smoothness as well, thus not only affecting the confidence interval estimates, but the load estimates as well.”

The authors nicely explainedthe importance of including serial correlation, but in fact, it was only considered a firstorder autocorrelation. Did the authors explore at least a second order autocorrelation?

In Slaets et al. (2014), where the same dataset was used, we explored the use of several alternative variance-covariance structures to model the serial correlation. The Akaike Information Criterion (AIC) was used to compare various candidate models and a spatial power structure with time as the coordinate showed the best fit based on this criterion. In the bootstrap iterations, however, the spatial power structure caused convergence issues in a large number of the bootstrap replicates which would result in biased bootstrap estimates, and therefore we switched back to the AR(1) structure. The power model is an extension of the AR(1) model to accommodate for unequally spaced observations, and both structures are essentially different parameterizations of the same model (Piepho et al., 2015). In the spatial power model, the autocorrelation decays as a function of the distance between observations (in this case, the distance in time). The difference in AIC between the AR(1) and spatial power model was 4 points, indicating that while the spatial power model is most likely the best performing model, there is still considerable support for the AR(1) model (Burnham and Anderson, 2002). Unfortunately the AR(2) structure is not available in the Mixed procedure of SAS, and if it were available, we would expect to encounter convergence issues in running the bootstrap iterations. As the AIC of the AR(1) was comparatively close to that of the spatial power model, however, the AR(1) structure was an adequate approximation of the serial correlation structure in the data.

We have added the discussion of the spatial power model to the methods section: "In a previous model published in Slaets et al. (2014), we explored the use of several alternative variance-covariance structures to model the serial correlation. The selected spatial power model unfortunately caused non-convergence for a large number of the bootstrap replicates when using it for bootstrap load estimates, and therefore the AR(1) structure was implemented as it did not have convergence issues. The difference in AIC between the AR(1) and spatial power model was 4 points. Therefore the spatial power model is most likely the best performing model, but there is still considerable support for the AR(1) model (Burnham and Anderson, 2002)."

Finally, did the authors consider to compare the bootstrap results with results of different complex models, eg GAM?

The most common GAMS can be estimated using maximum likelihood methods. For more complex GAMS, we are not aware of any procedures in SAS that implement ML algorithms that can also fit random effects and serial correlation. An alternative is computing a profile likelihood but we believe that least squares and maximum likelihood methods are by far the most commonly used methods to establish sediment rating curves. One possible alternative would be to use the Glimmix procedure to fit B-splines, which are very similar to GAMs, to explore nonlinearities as the Glimmix procedure can model random effects. With the level of noise in the sediment rating curve, however, there is a danger of overfitting unless very clear irregular nonlinear shapes are seen in the data. Therefore we consider comparison to generalized additive models to be outside the scope of our paper, though we refer to their potential for further exploration of load estimation uncertainty in the conclusions: "Reporting uncertainty is especially important when water quality models are complex. There has been a great increase in the use of more complex predictive methods for water quality, for example the use of Artificial Neural Networks, Random Forests or Generalized Additive Models (Berk, 2008). The advent of these methods makes the consistent reporting of measures of uncertainty even more essential: the more complex a model is, the more prone it is to overfitting (Burnham and Anderson, 2002), as was demonstrated by the inflated confidence intervals when adding predictor variables to the sediment concentration model."