Point-by-point response to Interactive comment on "Quantifying uncertainty on sediment loads using bootstrap confidence intervals" by J. I. F. Slaets et al.

Anonymous Referee #2
Received and published: 20 September 2016

Original comments in italic
Responses in non-italic

This paper is a very good contribution to the literature about load estimation and the uncertainty of load estimates. The consideration of the relative role of discharge uncertainty and concentration versus discharge uncertainty is a valuable contribution.

We thank reviewer 2 for his positive evaluation of our paper. Specific comments are addressed below:

I do have concerns about the quality of the regression relationships that were used in the analysis. See my comments on figures 3 and 4. Also, they need to be clear abouthow they view year-to-year variability. Do they consider each year to be a separate population or are they each a different sample from the same population. Are estimates from the two years done separately? There are two schools of thought about how concentration prediction models should be built: using just data from the year of interestor using data from many years, with some consideration for the possibility that there may be a temporal trend in that relationship. They should be explicit about this issue.

We consider the two-year period to be one population and use all samples from the two year study period to build the prediction models. Both years are thus predicted from the same model, the parameters of which are estimated from the same data. This information has been added now to the Materials and Methods section as follows: "All samples from the two year study period were used to build the concentration prediction model, and load estimates from both years are thus predicted from the same model with the same parameter estimates."

line 222-229. It is not clear why the base-flow samples should be considered to be independent. My own experience is that they are not (I am thinking here about the residuals from a concentration versus discharge relationship). They used a first order autoregressive process but it seems to me that the process may be more complex than that, with memory that lasts for many days duration.

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 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. The variance-covariance parameter estimates from this model showed that the autocorrelation becomes nearly zero for samples taken more than 80 min apart– which coincides with the average duration of rainfall events, and was the basis for the independence assumption for the base-flow samples. The reviewer rightly points out that in many natural catchments, base-flow samples are not necessarily independent. The different result for our dataset might be attributable to the irrigation management present in the paddy-containing watershed, which disturbs natural hydrographs and changes the memory effect.
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).
We have clarified the independence assumption for base-flow samples in the manuscript: “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). The spatial power structure with time as the coordinate showed that the autocorrelation becomes nearly zero for samples taken more than 80 min apart– which coincides with the average duration of rainfall events. Therefore the base-flow samples were considered to be independent.”

line 302. It appears to me that the method of removing transformation bias is similar to the approach called the “smearing estimate” proposed by Duan. I realize that Duan’s smearing estimator is mentioned later in the paper (line 472) but I think it needs tobe mentioned here as well. Furthermore, the question of whether the residuals are homoscedastic needs to be considered. If it is not, then this approach becomes problematic.

The two methods used for removing transformation bias in our approach are both parametric (adding half the residual error variance or simulating an AR(1) term). Duan's smearing estimator would be a third option and a non-parametric alternative, as it is a correction where the sample average of the exponentiated residuals from the model is used as the correction factor. Duan's smearing estimator assumes iid errors. As such, the procedure is not a suitable alternative in this study as serial correlation in the data is present. We have added these clarifications to the Methods section on data transformations: “With nonlinear data transformations (the log-transformation and the Box-Cox transformation being prime examples), predicted means cannot be naively back-transformed and interpreted as means on the original scale. Correction factors can be applied that compensate for the underestimation of SSC that arises from doing the predictions on the transformed scale. A commonly used non-parametric correction factor is Duan’s smearing estimator (Duan, 1983), where the sample average of the exponentiated residuals from the model is used as the correction factor. Duan's smearing estimator assumes independent and identically distributed errors, however, and is therefore also not a suitable alternative when serial correlation in the data is present. Alternatively, as pointed out by Rustomji and Wilkinson (2008), adding the modeled residual error removes the need to apply a correction factor and is therefore the recommended approach.”
The reviewer correctly points out that, irrespective of the chosen correction factor, it must be confirmed that the transformation was successful in stabilizing the variance. This point has been emphasized 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.”

Figure 3. A sample size of 15 is very small for constructing a rating curve. It is disturbing that for the lowest two and highest three predicted values the residuals are all negative and fairly substantially so. My reaction to this plot is that there is some significant lack of fit to the proposed rating curve model. Even with this small number of observations, perhaps a higher order or non-linear model should have been considered.

As the reviewer points out, the lowest two and highest three studentized residuals are negative, however they are not larger than two (and in fact smaller than 1.5). Furthermore, as can be seen from Figure 2, the 95% confidence intervals for the regression line as well as for new predictions are narrow and show stability over the whole range of flows. Nonlinearity was explored by adding a quadratic term for the log-transformed water level, but the adjusted R^2 changed only marginally (from 0.978 to 0.985), therefore we opted for the more parsimonious linear model.
The reviewer makes a good point that there is an effect of sample size on the bootstrap confidence intervals: the sample size in this study is small, and a larger sample size could thus result in a lower uncertainty on the load estimates of the discharge rating curve. Several aspects come into consideration when assessing sample size in the context of bootstrapping a stage-discharge rating curve. In very small sample sizes, the bootstrap could fail due to the set of possible bootstrap samples not being rich enough. For sample sizes as small as 8 and up, however, the number of distinct bootstrap samples gets large enough very quickly to result in consistent estimates (Chernick, 2011; Hall, 2013). Our sample of fifteen is thus large enough to bootstrap. As for the number of observations required to obtain a reliable discharge rating curve: no fixed sample size recommendation exists, as the accuracy of the discharge rating curve is determined not solely by the sample size, but also the stability of the river bed, shape of the cross-sectional profile, the range of discharges, and foremost the spread of the samples, as all ranges of flow need to be covered. The studied catchment has a small range of discharges (i.e. from 0.15 m^3/s to 4.30 m^3/s corresponding to minimum water levels of 1.1 m and maximum water levels of 1.6 m), a stable river bed and a short duration of the study (two years). And importantly, the dataset has a good spread over the ranges of discharge (Figure 2) covering the lowest and highest water levels measured throughout the 2010-2011 period for that specific location. These points regarding the effect of sample size and limitation of our dataset have now been added to the discussion: “Even though the discharge rating curve has a high accuracy, an estimate of Q is used as a predictor variable for concentration, and the concentration then gets multiplied with the estimate of Q, and so the effect is not as small as one would expect based on the R^2 of both rating curves. It is possible that the sample size of the discharge rating curve, which is relatively small (n=15) plays a role here, as a bootstrap iteration that does not contain the largest discharge values would result in a wider confidence interval for the estimated load.”

Figure 4 is even more disturbing in terms of fit. The observations should be on the y-axis so that the residuals can be visualized as the vertical distance from the 1:1 line. For virtually every observed value greater than about 600 mg/L the residuals were all positive or only very slightly negative. Conversely, for the vast majority of the observed values below about 100 mg/L the residuals were almost all negative and in some cases the predictions were as much as 10 times greater than the observed. This is a very flawed model to be used as the basis for this experiment. A study of errors needs to start with a fitted model that does not have such a high degree of bias.

Thank you for this good suggestion, we have now changed the figure to display the observed values on the Y-axis as the reviewer suggested. Figure 4 displays the observed versus the predicted values of the sediment rating curve after five-fold cross validation, and as such, it does not show model residuals or allows for the checking of model assumptions. Residuals of the sediment concentration model can be seen in Figure 5 plotted against the predicted values. As we use a linear mixed model with two quantitative predictors (discharge and turbidity) for our concentration predictions, we are not able to show a plotted regression equation; and plotting observed versus predicted values as we do in Figure 4 is an alternative way to visualize predictive accuracy of the model. As we show the predicted values after cross validation, this plot is a stronger measure of accuracy than a classic regression line. Figure 4 illustrates, as the reviewer correctly points out, that the concentration model for new predictions tends to over predict low concentrations and under predict high concentrations. This tendency of regression towards the mean is not a flaw of the model, but

typically seen when models are fitted to very noisy data and is also well documented in erosion studies (Nearing, 1998). We thank the reviewer for bringing up this important trend in Figure 4 and have added the discussion of this phenomenon in the text: “The sediment rating curve tends to over predict low concentrations and under predict high concentrations for new data, as is visible in Figure 4. This tendency of regression towards the mean is typically seen when models are fitted to very noisy data, and is also well documented in erosion studies (Nearing, 1998).”
While we therefore do not see model bias confirmed in the residual plots, the trend in Figure 4 demonstrates that the predictive power of our model has limitations, as is also clear from the reported Pearson’s r^2 between observed and predicted values after cross validation of 0.56. For datasets similar to our own, in fragmented landscapes, with heterogeneous terrain, soils, geology and land use, a model that explains half or more of the variation after validation would be difficult to improve without overfitting with a more complex model. That being said, in other catchments with less heterogeneity where a much more accurate sediment rating curve can be obtained, the resulting load estimates will be more accurate and it is a limitation of our dataset that we cannot assess confidence interval width for such a scenario. Therefore we have addressed these limitations in the discussion: “The accuracy of the sediment rating curve in this study (Pearson’s r^2=0.56 after cross validation) is reasonable for catchments with large heterogeneity in relief, land use, soil types and rainfall event characteristics. In more homogeneous settings, however, much more accurate sediment rating curves have been obtained, which can be expected to result in more narrow confidence intervals on their resulting load estimates.”

485-489. The issue is not whether concentrations or loads are log-normally distributed.The issue is the normality of the residuals from the fitted model. This is a common error in analysis of such data sets. The adequacy of the estimation method should be based on the distribution properties of the residuals.

The reviewer correctly points out that there is no need for concentrations or loads to be log-normally distributed, but rather model assumptions need to be checked on residuals in the case of regression type models. The point we were trying to make here was regarding the use of the delta-method as an alternative to obtain an estimate of the load variance. In this approach, log-normality of the loads was required, and therefore in the case of our dataset, the method would not have yielded valid results. To clarify this for the reader, the corresponding section has been changed as follows: “Regarding the data transformation, while the sediment concentration was log-normally distributed, the log-transformed load estimates were not normally distributed (Figure 7, right panel). This non-log-normality of our loads does not affect the viability of the bootstrap approach, as regression type methods do not require the concentration or load data but rather normality of the residuals. It does, however, limit the applicability of methods that use the log-normality assumption of the load to estimate a variance for the load, as was done for example by Wang et al. (2011) in an approach that used the delta-method as an alternative way to assess uncertainty on annual sediment load estimates.”

500-518. These points about overly complex models are very good. This is an important concern and I’m glad the authors emphasize it here.
We thank the reviewer for his positive evaluation of this point.

References

Burnham, K. P., and Anderson, D. R.: Model Selection and Inference: A Practical Information-Theoretic Approach, Springer-Verlag GmbH, New York, 1998.

Chernick, M. R.: Bootstrap methods: A guide for practitioners and researchers, John Wiley & Sons, 2011.

Duan, N.: Smearing estimate: a nonparametric retransformation method, Journal of the American Statistical Association, 78, 605-610, 1983.

Hall, P.: The bootstrap and Edgeworth expansion, Springer Science & Business Media, 2013.

Nearing, M. A.: Why soil erosion models over-predict small soil losses and under-predict large soil losses, Catena, 32, 15-22, 1998.

Rustomji, P., and Wilkinson, S. N.: Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves, Water Resources Research, 44, W09435, 2008.

Slaets, J. I. F., Schmitter, P., Hilger, T., Lamers, M., Piepho, H. P., Vien, T. D., and Cadisch, G.: A turbidity-based method to continuously monitor sediment, carbon and nitrogen flows in mountainous watersheds, Journal of Hydrology, 513, 45-57, 10.1016/j.jhydrol.2014.03.034, 2014.

Wang, Y. G., Kuhnert, P., and Henderson, B.: Load estimation with uncertainties from opportunistic sampling data - A semiparametric approach, Journal of Hydrology, 396, 148-157, 2011.