Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-34-AC1, 2016

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on "SWAT Modeling of Water Quantity and Quality in the Tennessee River Basin: Spatiotemporal Calibration and Validation" *by* G. Wang et al.

G. Wang et al.

wangg@ornl.gov

Received and published: 10 March 2016

Response to Interactive comment on "SWAT Modeling of Water Quantity and Quality in the Tennessee River Basin: Spatiotemporal Calibration and Validation" by G. Wang et al.

G. Wang et al. wangg@ornl.gov

[Response—] We would like to thank Referee #2 for his/her constructive comments. While you will be able to see the changes on the manuscript, we would like to highlight the following points: (1) We explained why we have to use empirical datasets (LOADEST and SPARROW), not the directly-observed datasets for model calibration

and validation. (2) We changed the thresholds of correlation between two variables and explained why there was very low correlation between sediment/P and runoff in our spatial analysis. (3) We added statistical tests to compare the TN and TP estimated by SWAT and SPARROW. We added Supplementary Fig. S6(a-d) to show spatial comparison (PBIAS, i.e., %bias) between SWAT and SPARROW modeled TN and TP. (4) We added Table S3 to show calibrated SWAT parameter values. Please see our point-by-point "Response to Comments" between [Response—] and [—Response] following the reviewer's original comments. [—Response]

Anonymous Referee #2 Received and published: 1 March 2016

In this manuscript, the authors applied the SWAT model to the Tennessee River Basin to simulate water quantity and quality. Statistical modeling results were used to evaluate model performance. They found that model simulations were improved after parameter calibration. Correlation analyses were conducted to analyze the impacts of watershed attributes on water qualities. The authors have done lots of work in model simulation, calibration, and analysis. However, I think the manuscript needs to be substantially revised for publication. Here are my major concerns: First, I do not quite agree about the way how model performance evaluation is conducted. Although there are some difficulties in collecting observational data for model evaluation, field data should be the most valuable and reliable material for benchmarking. However, the authors mainly used estimates from statistic models to calibrate and evaluate their modeling results. Since there are significant uncertainties in these statistic models, particularly LOADEST, comparing you model with these modeling results introduces additional uncertainties to calibration and validation of this work. As a result, the authors should include comparison with streamflow records, and concentrations of different elements from the USGS gauges in their work. Temporally explicit observations are limited, particularly for water quality variables, but the authors should at least compare the long-term averages.

[Response—] Thanks for the reviewer's positive and thoughtful comments. We understand the reviewer's concerns on the calibration of hydrological model against empirical datasets, which differs from the traditional way to use observed data. At the beginning of this study, I also attempted to collect observations for SWAT calibration and validation. As mentioned in the manuscript, [Revised Manuscript Page 10 Line 216-220]" The utilization of streamflow (discharge) data for model calibration and validation in this study made little sense owing to the reservoir operation in the TRB. Because streamflow at a station within the TRB is largely a measure of the outflow from the upstream reservoir(s) and because observed reservoir outflow was used in this study, we calibrated hydrologic parameters based on runoff (i.e., total water yield) instead of streamflow." [Page 11-12 Line 242-253]" Nutrient measurements are sparse in rivers of the TRB. We have attempted to collect in-situ water quality monitoring data from over 6,000 USGS and EPA (Environmental Protection Agency) stations within the TRB through the National Water Quality Monitoring Council (NWQMC)'s online Water Quality Portal (WQP) (NWQMC, 2015). However, these observed data are not ready and useful for model calibration owing to the following reasons: (i) Although there are many measurement sites (stations), very few long-term time series are available within our study period (after 1980s); (ii) Not all of the water quality variables are measured at a specific site; and (iii) There is scaling issue regarding the water quality data. Due to limited sub-daily data points (i.e., one measurement in one month or several/many months), it is meaningless to do model calibration at daily scale. If we want to do model calibration at the monthly or yearly scale, we need to integrate the data from sub-daily to monthly or yearly scale, which is difficult when there are lots of data gaps." To our understanding, these are also the reasons why the LOADEST and SPARROW datasets were generated and they focused on temporal and spatial scale, respectively. The LOADEST dataset was time-series of monthly nutrient fluxes generated for a specific site, i.e., the Tennessee River near Paducah, KY, because this site had longer time-series of observation compared to the other sites; while the SPARROW dataset was the mean annual values of spatially-distributed nutrient loadings. These two published datasets were generated by the statistical approach and might underlie large

uncertainty, as pointed out the the reviewer. Thorough analysis of their quality might have been reported in relevant publications. However, we did not find such analysis pertaining to the TRB. We could do such analysis but we think it could be an independent study whereas it is not the focus of this manuscript. Without useful direct observations of water quality data, we have to use these two published datasets as reference to calibrate and validation the SWAT model for TRB. Through our preliminary communications with the USGS experts, they support the use of their empirical modeling as a way of getting the best of both worlds, empirical and process-based models. Also owing to the uncertainty in these empirical datasets, our calibration and validation performance was not satisfactory, which might NOT be regarded as unsuccessful. One could see that our SWAT simulations were capable of capturing the temporal patterns in the temporal LOADEST dataset and the spatial pattern of TN (total nitrogen) in SPARROW. The discrepancies in high or low values of water quality between SWAT and LOADEST resulted in the low model NSE values. Nevertheless, the mean values of mean annual loading (MAL) of SWAT-simulated TN and TP across the TRB were comparable to that of SPARROW estimated based on our statistical tests suggested by the reviewer: [Page 15 Line 335-337]: "The NSE values for model validation were not as good as the NSE for calibration, but the PBIAS values (Table S2) were satisfactory except for NO3+NO2 (−157%)." [Page 16 Line 339-341]: "SWAT-simulated water quality responses reproduced the seasonal patterns found in LOADEST data during both calibration and validation periods (See Fig. S4)." [Page 16 Line 347-354]: "The LSD test indicated that the mean MALs of TN across the 32 HUC8 units were not significantly different between SWAT and SPARROW (p-value > 0.05) . . .. . . The PBIAS values (between SWAT and SPARROW) for TN at 26 out of 32 HUC8 units were within the range of ±70%, and the PBIAS values at three HUC8 were higher than 80% (Fig. S6a)." [Page 16-17 Line 355-365]: "The LSD test showed that the mean MAL of SWAT_TP (1.32 kg P/ha) was not significantly different from that of SPARROW_TP (0.88 kg P/ha) (Fig. 5b) . . .. . . The PBIAS values between SWAT_TP and SPARROW_TP at 13 out of 32 HUC8 units were within the range of ±70% (Fig. S6b). In addition, the PBIAS values between SWAT_OrgP+MinP and SPARROW_TP at 50% of the HUC8 units fell into the range of ±70% (Fig. S6c), and the PBIAS values between SWAT_OrgP+SolP and SPARROW_TP at 59% of the HUC8 units were within ±70% (Fig. S6d)" Our SWAT simulations also showed different, however more reasonable results for TP than SPARROW: [Page 18, Line 390-393]: "the SPARROW-estimated spatial patterns of TN and TP were correlated with each other; however, the SWAT-simulated spatial distributions of TN and TP were decoupled because MinP contributed most (65%) to TP and TN was dominated by inorganic nitrogen in SWAT." Overall, the observations in water quality were available but NOT ready and directly useful for model calibration and validation. The empirical temporal LOADEST and spatial SPARROW datasets were the best ones we could find at present to provide reference for our SWAT calibration and validation in the TRB. Our SWAT simulations could be an alternative to these datasets to characterize the water quality in the TRB, which is also the reason why we need SWAT modeling. Overall, the observations in water quality were available but NOT ready or directly useful for model calibration and validation. USGS are the experts on these data and that their synthetic data products reflect the best available use of observational data. The empirical temporal LOADEST and spatial SPARROW datasets were the best ones we could find at present to provide reference for our SWAT calibration and validation in the TRB. Our SWAT simulations could be an alternative to these datasets to characterize the water quality in the TRB, which is also the reason why we need process-based (e.g., SWAT) modeling. [—Response]

Second, although the authors claimed they did 'spatiotemporal 'calibration and validation in this work, I do not think this is well achieved. Only one gauge station was used, and I did not see any regional comparison maps between this study and SPARROW/regional runoff products. Instead, only regional means (figure 5) were compared and the spatial distribution of the selected variables from this work and the previous studies were not presented, which make it hard to validate results of this study.

[Response—] Thanks for the reviewer's constructive comments! We validated the nutri-

ent yields using the spatial dataset SPARROW. We add Supplementary Fig. S6(a-d) to show spatial comparison (PBIAS, i.e., %bias) between SWAT and SPARROW modeled TN and TP. We added statistical tests to compare the TN and TP estimated by SWAT and SPARROW. [Page 13 Line 269-274] in "Materials and Methods": "We compared the TN and TP MALs between SWAT and SPARROW by (i) testing the significance of difference in mean MALs across TRB by the Fisher's least significant difference (LSD) method (De Mendiburu, 2015); (ii) testing the significance of difference in the probability distribution of the MALs by the Kruskal-Wallis (KW) test (Giraudoux, 2013); and (iii) calculating the PBIAS of MALs between SWAT and SPARROW at the HUC8 level. All statistical tests were conducted at the significance level of α = 0.05." [Page 16-17 Line 345-365] in "Results and Discussion": "The spatial distributions of SWAT-simulated MALs (1986–2013) of TN and TP were comparable to the SPARROW-estimated MALs (1975–2004) (Fig. 5 and Fig. S6). The LSD test indicated that the mean MALs of TN across the 32 HUC8 units were not significantly different between SWAT and SPARROW (p-value > 0.05), although the SWAT-simulated MAL (5.5 kg N/ha) was 12% lower than the SPARROW estimate (6.2 kg N/ha) (Fig. 5a). The 50% CIs of MAL of TN were 2.5–6.7 kg N/ha and 4.7–7.4 kg N/ha by SWAT and SPARROW, respectively (Fig. 5a). The KW test was significant, which implied that the MALs from the two models did not originate from the same probability distribution. The PBIAS values (between SWAT and SPARROW) for TN at 26 out of 32 HUC8 units were within the range of ±70%, and the PBIAS values at three HUC8 were higher than 80% (Fig. S6a). The SWAT-simulated TP (SWAT_TP) consisted of three components, i.e., organic P (OrgP), soluble P (SolP), and mineral P (MinP). The LSD test showed that the mean MAL of SWAT_TP (1.32 kg P/ha) was not significantly different from that of SPARROW_TP (0.88 kg P/ha) (Fig. 5b). The KW test indicated that the SPARROW_TP and SWAT_TP did not originate from the same probability distribution, but there was no evidence of stochastic dominance between SPARROW_TP and SWAT_OrgP+MinP or between SPARROW_TP and SWAT_OrgP+SolP (Fig. 5b). The PBIAS values between SWAT_TP and SPARROW_TP at 13 out of 32 HUC8 units were within the range of ±70% (Fig. S6b). In addition, the PBIAS values between SWAT_OrgP+MinP and SPARROW_TP at 50% of the HUC8 units fell into the range of ±70% (Fig. S6c), and the PBIAS values between SWAT_OrgP+SolP and SPARROW_TP at 59% of the HUC8 units were within ±70% (Fig. S6d)" [—Response]

Third, spatial correlation analysis was not clearly introduced. I am wondering how the authors calculated the correlation coefficient? Did they use bivariate correlation or multiple linear regression? Did they consider the collinearity among the independent variables? Why only r was used to measure significance of the correlation, not P values?

[Response—] see [Page 13 Line 280-281]: "The bivariate correlation analyses were conducted using the 'cor' function in R (R Development Core Team, 2011) and the correlation were plotted by the 'corrplot' package (Wei, 2013)." Thus the collinearity among the variables was not considered in our study. Both P-value and correlation coefficient were used to measure significance of the correlation, only those correlations with p-value <0.05 were shown in Fig.6. please see Page 13 Line 282-285: "For this study, two variables were considered (i) highly correlated if the absolute value of correlation coefficient (|r|) was greater than 0.6 and p-value < 0.05, (ii) moderately correlated if |r| was between 0.4–0.6 and p-value < 0.05, and (iii) lowly correlated if |r| was between 0.2–0.4 and p-value < 0.05." [—Response]

Finally, interpretation of the results, particularly the correlation analysis is insufficient. In addition to report significant correlations, the authors should explain the underlying mechanisms responsible for the correlation, and be cautious with non-causative correlations.

[Response—] We agree with the reviewer's suggestions. Please see more explanations of underlying mechanisms following the presentation of the correlation, e.g., [Page 17-18 Line 382-403]: "We found that SWAT-simulated MALs of MinP (mineral P attached to sediment) and TP were highly correlated with sediment, which confirms

that sediment plays an important role in watershed phosphorus dynamics (Fig 6a). The TN yield was highly correlated with NO3. TN loadings were dominated by NO3, i.e., the fraction of TN that was NO3 ranged from 37% to 99% with an average of 80%. TP was not correlated with TN, but OrgP (organic P) was moderately correlated with OrgN (organic N) and SolP (soluble P) was moderately correlated with NO3, which implies similarity between SolP and NO3 dynamics and similarity between OrgP and OrgN dynamics in SWAT (Neitsch et al., 2011). In addition, the SPARROW-estimated spatial patterns of TN and TP were correlated with each other; however, the SWAT-simulated spatial distributions of TN and TP were decoupled because MinP contributed most (65%) to TP and TN was dominated by inorganic nitrogen in SWAT. Nutrient (Sediment, P and N) loadings were not significantly correlated with runoff in our spatial correlation analysis (Fig. 6a). This is because we were conducting spatial correlation analysis. If we implemented temporal correlation analysis for a specific HRU, subbasin or HUC8, taking the sediment MUSLE equation (Neitsch et al., 2011) as an example, the landscape factors (e.g., soil erodibility, land cover and management, support practice, topographic, coarse fragment) would not change or vary slightly with time, runoff would become the most important factor influencing sediment yield. Thus we could expect high correlations between sediment/P and runoff. The non-correlation between nutrient yields and runoff in our spatial analysis suggested that nutrient point-source and non-point sources and other physical landscape variables (e.g., topography and land cover) controlled spatial variation in SWAT-simulated nutrient loadings in the TRB." [Page 19 Line 409-410]: "Sediment loadings were moderately and positively correlated with Elevation_Drop (r = 0.47), which verifies that the representation of topography and topology in this region drives sediment dynamics (Wellen et al., 2015)." [Page 19 Line 412-416]: "OrgP (organic P) was highly associated with Developed_Fraction (r = 0.64) that represented human activities in urban area (Hoos and McMahon, 2009); SolP (soluble P) was moderately correlated with Hay_Fraction (r = 0.43) indicating the influence of agricultural fertilization; and MinP (mineral P) was lowly correlated with Elevation_Drop (r = 0.37) that was the primary driver for sediment generation." [Page

19 Line 418-423]: “NO3 was highly correlated with Hay_Fraction (r = 0.63) and moderately correlated with Crop_Fraction (r = 0.48), mostly owing to the response of NO3 yield to agricultural fertilization. In addition, NO3 showed a moderate and negative correlation with Forest_Fraction (r = -0.54) and Subbasin_Slope (r = -0.44). Note that TRB subbasins with steeper slopes generally had more forest and less cropland. The primary drivers controlling TN were the same as those for NO3 as TN was dominated by NO3.” [—Response]

Specific comments: Page 5, Line 97: but later you mentioned that only one site, close to the outlet of the basin, was used for model calibration

[Response—] please see our response to the major concern #1. We calibrated runoff in 32 HUC8 units. We calibrated water quality in one site close to the outlet of TRB, but we validated TP and TN yields against the spatial dataset SPARROW. [—Response]

Page 5, Line 106: you already provide the full name of this acronym on page 1.

[Response—] We deleted ‘(TRB)’ and kept the full name of ‘Tennessee River Basin’ because it is the lead sentence of this paragraph. [—Response]

Page 7, Line 147: as far as I know daymet is a modeling dataset. Does it also provide original site level observation?

[Response—] Yes, daymet is a modeling dataset. See [Page 8 Line 166-167]: “We downloaded synthetic meteorological data from DAYMET (Thornton et al., 1997) for the center of each HUC8 (Fig. 1) over the period 1980–2014 (35 years)” [—Response]

Table 1, it will be more helpful if you provide your calibrated parameter values, rather than providing the input file and fortran code.

[Response—] Thanks for the good suggestion! We added Table S3 to show calibrated SWAT parameter values. [—Response]

Figure 6a is confusing. Consider to label variables in a different way

[Response—] sorry for the confusing. This label style has been used in R packages to simplify the correlation plot. We added statement in the figure caption to further explain the figure, e.g., “Numbers in (a) denote correlation coefficients between the two variables shown in corresponding row and column.” [—Response]

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-34, 2016.