Response to Statement of Findings

On May 18, 2015, EPA received the statement of findings from the two peer reviewers based on their review of the final reports developed by U.S. EPA contractors for the Agency's Illinois River Watershed (IRW) TMDL Modeling Program. After careful review of the comments, Baker Team (Michael Baker International, Dynamic Solutions, LLC, and Aqua Terra Consultants), in consultation with EPA has prepared the following responses.

Introductory Comments

Presented below are the technical findings developed by the two peer reviewers who have reviewed the key final reports developed by U.S. EPA contractors for the Agency's Illinois River Watershed (IRW) TMDL Modeling Program. As an introduction for these peer reviewers, EA prepared an introduction and summary report describing the TMDL model and how it was developed to help familiarize them with the essential elements and features of the models and the model development process. We also provided each of the eight final reports on EA's ftp site for the reviewers to download.

As part of this effort, EA was required to develop and provide specific issues (or focused "charge questions") to be addressed by the peer reviewers. The charge questions were to identify the technical and scientific issues relating to the US EPA's Illinois River Watershed TMDL modeling effort on which US EPA Region 6 would like feedback, and invite suggestions for improving the modeling effort as a whole. Although reviewers were encouraged to provide comments or opinions on any other issues they deemed noteworthy, the specific charge questions were as follows:

1. As developed, are the HSPF and EFDC models able to reasonably represent the physical conditions of the IRW?
2. Is the model sufficient to reasonably model water quality conditions for nutrients in the IRW, and be a useful tool for developing numeric TMDLs for nutrients in the watershed?
3. Does the interface between the watershed model (HSPF) and the reservoir model (EFDC) seem reasonable?
4. Given the point and non-point source location and loading data that have been identified to date for agricultural operations (e.g., poultry, hogs, cattle, manure), is this sufficient to run the model and develop alternate watershed management scenarios?
5. Does the model appear to be sensitive enough to allow compliance assessments relative to the State of Oklahoma's 0.037 mg/L TP standard?
6. Are there any overall concerns with the current model that would draw question to future water quality predictions for nutrients?
7. Are there suggested improvements to the model (or model inputs) that would improve its use in developing load allocations to meet water quality objectives?

Each of the peer reviewer's comments is presented below.

Comments by Peer Reviewer #1

Comment #1.

1) As developed, are the HSPF and EFDC models able to represent the physical conditions of the IRW?

First, it should be noted that this reviewer fully agrees with the conclusions of the Model Selection Memorandum (Donigian and Imhoff, 2011) that the HSPF and EFDC are appropriate models to apply to the Illinois River Watershed (IRW) and Lake Tenkiller and that these models have clear advantages in comparison to SWAT and AQUATOX for this study. Second, in order to properly answer this question the performance of the two models is reviewed separately in the following subsections.

Response #1. Concur with the comment

Comment #2.

Hydrological Simulation Program—Fortran (HSPF)

HSPF Segmentation and Input Data

The application of HSPF to the IRW divides the approximately 1600 mi^2 watershed into 133 model sub-basins (or segments). These sub-basins are further divided into pervious land segments (PERLNDs) and impervious segments (IMPLNDs). The PERLNDs are distinguished by land cover/land use into the following categories: Forest (42.78%); Pasture/Hay (41.08%) [further divided into 4 types depending on slope and whether or not poultry litter is applied to the pasture]; Grass/Shrub/Barren (4.33%); Developed, Open Space (5.93%); Developed, Low Intensity (2.42%); Developed, Medium/High Intensity (1.27%); Wetlands (0.60%); and Cultivated Crops (0.15%). A particularly important and appropriate aspect of the HSPF modeling for the IRW is the application of the AGCHEM subroutines to the Pasture PERLNDs to simulate the direct connection between nutrient application rates from chemical fertilizers, manure, and poultry litter, and subsequent buildup and potential runoff to rivers and streams, from applied pasture lands (Baker et al., 2013, p. 7). The developed areas were further divided into effective impervious areas (i.e. impervious areas that directly discharge to the drainage network) and pervious areas (which include the impervious areas that drain to pervious areas). Similarly, the Illinois River (160 mi) and its major tributaries are represented by 126 stream reaches in the HSPF model.

The runoff, sediment yield, and pollutant loads from the PERLNDs and IMPLNDs are driven by a detailed spatial representation of precipitation across the IRW. For the portion of the IRW in Oklahoma data from 11 precipitation gages were used—4 daily and 7 hourly or 15-min. For the portion of the IRW in Arkansas 28 "pseudo" stations were created from daily NEXRAD data, which were used because of sparse gage coverage in the Arkansas – only 3 gages near the eastern boundary of the watershed (Baker et al., 2013, p. 14). Baker et al. (2015b, p. 14) reported some problems with the NEXRAD derived precipitation data as follows:

> "Further analysis identified 10-15 events with rainfall totals at some of the NEXRAD 'pseudo' stations with extreme daily amounts in the range of 10-22 inches in a single day. … Consequently, for these selected events we adjusted the rainfall for the outlier site based on rainfall amounts at neighboring sites. This does raise questions regarding the accuracy of the NEXRAD data for other non-extreme events."

The extremely good flow durations curves shown for the State Line (Reach 630) in Figure 2-18 of Baker et al. (2015b) indicate that the adjustments to the NEXRAD precipitation data have worked very well.

In Summary, the segmentation of the watershed and river system and the meteorological input data used to "drive" the model is consistent with "best current engineering practice and scientific knowledge" and reasonably represents the physical conditions of the IRW.

Response #2. Concur with the comments.

Comment #3.

HSPF Calibration and Validation

In this section, the quality of the calibration and validation of the HSPF model applied to the IRW is reviewed to determine if the final, calibrated and validated model reasonably represents the physical conditions of the IRW. James and Burges (1982) recommended that graphical and statistical means be used to assess the quality of the fit between simulated results and measured values because trends and biases can be easily detected on graphs, and statistics provide an objective measure of whether one simulation is an improvement over another. This approach to calibration and validation was followed in the application of HSPF to the IRW. With respect to statistical measures the Modeling Quality Assurance Project Plan (QAPP) (Baker 2013, p. 43) states the following:

> "Given the uncertain state-of-the-art in model performance criteria, the inherent errors in input and observed data, and the approximate nature of model formulations, **absolute** criteria for watershed model acceptance or rejection are not generally considered appropriate by most modeling professionals. And yet, most decision makers want definitive answers to the questions – 'How accurate is the model?', 'Is the model good enough for this evaluation?', 'How uncertain or reliable are the model predictions?'. Consequently, we propose that targets or tolerance ranges, such as those shown above, be defined as general targets or goals for model calibration and validation for the corresponding modeled quantities. These tolerances should be applied to comparisons of simulated and observed mean flows, stage, concentrations, and other state variables of concern in the IRW TMDL effort, with larger deviations expected for individual sample points in both space and time. The values shown above *[Table 6.1 and Figure 6.12]* have been derived primarily from HSPF experience and selected past efforts on model performance criteria; however, they do reflect common tolerances accepted by many modeling professionals." (items in italics added)

Table 6-1 of Baker (2013, p. 44) gives the following specific calibration tolerances, primarily aimed at annual and total model outputs, Hydrology/Flow within ± 15%, Sediment Loadings/ Concentrations within ± 30%, Water Temperature ± 10%, and Nutrient Loadings/Concentrations within ± 25%. These correspond to the "Good" range of historic HSPF performance as documented in Table 2-16 which originated from Donigian et al. (1984, p. 111) and Donigian (2000). Further, Baker et al. (2015b, p. 53-54) stated the following with respect to correlation coefficients (R) and coefficients of determination (R^2, which is the equivalent of the Nash-Sutcliffe (1970) coefficient of model-fit efficiency):

> "Consequently, for the IRW modeling effort, we have proposed that the targets and tolerance ranges for **'Daily'** flows should correspond, at a minimum, to a **'Fair to Good'** agreement, and those for **'Monthly'** flows should correspond to **'Good to Very Good'** agreement for calibration. For the validation comparisons, we expect some decrease in model performance due to less dense gage coverage during that time period. Thus we expect the validation results to correspond to the **'Fair to Good'** ranges for both daily and monthly flows."

The range definitions are given in Figure 2-16 (Baker et al., 2015b, p. 53). In order to compare the results reported in Baker et al. (2015b) to specific targets, the targets for acceptable calibration and verification of monthly flows proposed in the first QAPP for an HSPF application to the watersheds affected by the proposed Crandon Mine in Wisconsin (USGS and Aqua Terra, 1998, p. 25) are considered here, namely correlation coefficients greater than 0.85 and coefficients of model-fit efficiency greater than 0.8.

HSPF Hydrologic Calibration and Validation: Baker et al. (2015b, p. 54) reports "Annual volume comparison shows a Very Good or better calibration, with all the calibration volume errors less than 10%. The validation volume errors are higher, as is expected, with all errors within 14%, except for Caney Creek which is an outlier at 40% error." It should be noted that the Caney Creek validation comparison is for only 3 years while the other comparisons are for longer periods. Further, all the calibration and validation errors except Caney Creek validation meet the quality targets identified in the modeling QAPP (Baker, 2013).

With respect to the monthly coefficient of model-fit efficiency and correlation coefficient the following results are found in Baker et al. (2015b). For the calibration period, the coefficient of model-fit efficiency values at 6 of 10 gaging stations exceeded 0.8 and the average over all 10 sites was 0.80. For the verification period, the coefficient of model-fit efficiency values at 4 of 10 gaging stations exceeded 0.8 and the average over all 10 sites was 0.69. For the calibration period, the correlation coefficient values at 8 of 10 gaging stations exceeded 0.85 and the average over all 10 sites was 0.90. For the verification period, the coefficient of model-fit efficiency the values at 7 of 10 gaging stations exceeded 0.85 and the average over all 10 sites was 0.85. Thus, the IRW hydrologic calibration and validation met the targets set for the QAPP for the Crandon Mine Project (USGS and Aqua Terra, 1998) with the exception of the monthly coefficient of model-fit efficiency during the validation period for which the lower quality of the results are related to the less dense gage network available during the validation period.

The flow duration curves shown in Figures 2-17 and 2-18 of Baker et al. (2015b) and the time series plots show in Figures 2-19 and 2-20 also indicate a very good hydrologic calibration has been achieved.

Baker et al. (2015b, p. 55) state: "In summary, the model results show a Fair to Good overall calibration and validation, and in some cases (i.e. sites) a Very Good simulation, confirming that the overall model should provide a sound basis for subsequent water quality simulations." I fully concur with this conclusion.

Perhaps the one concern readers might have is with respect to comment (d) on calibration and validation in Baker et al. (2015b, p. 54): "The Annual Flow Volumes in Tables 2-18 and 2-19 (for Stateline and Tahlequah), and those in Appendix A, show a wide range in year-to-year differences, with the year 2006 especially problematic, usually overestimated, for a number of sites." In particular, huge oversimulations of flows are reported for Water Year (WY) 2006: 62.36% at the State Line (Reach 630) and 54.15% at Tahlequah (Reach 870). However, such year-to-year variability and large oversimulations (or undersimulations) for one year in an otherwise well-simulated period of several years are not unusual in HSPF applications. In fact, the year-to-year variability of flow and precipitation is one of the reasons that the length of record recommended as the minimum for adequate model calibration is 3 to 5 years (Donigian et al., 1984; Linsley et al., 1982, p. 347). For example, for St. Joseph Creek in DuPage County, Illinois, WY 1987 was over simulated by 50.6%, whereas 4 of the other 6 years had errors less than 10% with the remaining two years having errors less than 25% (Duncker and Melching, 1998). Similarly, for Sawmill Creek in DuPage County, Illinois, WY 1987 was over simulated by 58.0%, whereas 3 of the other 6 years had errors less than 10% with two more years having errors less than 25% (Duncker and Melching, 1998).

In Summary, the hydrologic calibration of HSPF is consistent with "best current engineering practice and scientific knowledge" and reasonably represents the physical conditions of the IRW.

Response #3. Concur with the comments.

Comment #4.

HSPF Water Quality Calibration and Validation: Baker et al. (2015b, p. 64) states

> "Sediment, or TSS (Total Suspended Solids), is often considered the most difficult and challenging water quality constituents to model. Lack of adequate sediment data, especially during storm events, lack of bed characterization data which has a major influence on the model results, and lack of sediment particle size information for both bed materials and storm samples all contribute to the difficulties in accurately simulating TSS."

Figures 2-21 to 2-24 in Baker et al. (2015b) show generally good agreement between measured and simulated TSS concentrations at four gage locations in the IRW. Baker et al. (2015b, p. 67) also notes "that sediment validation was intended as part of the water-quality validation; however, the lack of litter and nutrient application data during the validation time period precluded that effort." This same lack of validation also applies to other aspects of the water-quality model for the same reason. This is not a serious limitation to the documentation of the

accuracy of the HSPF water quality model because the long calibration period (9 years) effectively is double the length of the recommended period for HSPF calibration (3-5 years) mentioned earlier. Thus, the calibration period includes a wide enough range of conditions to create confidence in the accuracy of the model in simulating sediment and nutrient loads and concentrations. Baker et al. (2015b, p. 67) states "the IRW model provides a good representation of the sediment/TSS behavior within the IRW and a sound basis for the subsequent water quality calibration." I fully agree with this assessment.

Baker et al. (2015b, p. 68) states "water temperature is well simulated by HSPF, and the high degree of agreement is essentially identical in both the calibration and validation periods." Further, Baker et al. (2015b, p. 91) states "The DO simulation shows a very good seasonal pattern consistent with the observed data, and the peaks and valleys are generally well represented." I fully agree with both of these assessments.

The quality of the HSPF model in simulating nutrient concentrations and loads is much more complex to evaluate. For example, Baker et al. (2015b, p. 96) states

> "The drought conditions in 2005-06 had a major impact on model results, causing significant over-estimation of nutrient forms, especially both P and N forms and DO. Part of the cause is the under simulation of flow during that drought which contributed to the over-estimation for many concentrations."

This statement makes no sense because the flows were actually over simulated in 2006 by 50-60% as previously discussed. If this is a simple typographical error, and the authors really meant to say that flow was over simulated which contributed to over-estimation of many concentrations it may be reasonable. Nutrient loads would definitely be over-estimated as a result of the over simulation of flows, whether concentrations would be over-estimated would depend on whether the increase in load was proportionally bigger than the increase in flow so that the concentration would increase.

The simulated concentrations of Total Nitrogen (TN), Orthophosphate (PO_4), and Total Phosphorus (TP) at the State Line and Tahlequah generally agree well with the measured concentrations as shown in Figures 2-29 to 2-32 in Baker et al. (2015b). The results shown in these figures support the following conclusions in Baker et al. (2015b):

p. 91) "Overall, the P components are generally better simulated that the N components as P was the major focus of this study due to the OK scenic rivers standard on TP."

p. 96) "In summary, the overall water quality calibration for the IRW demonstrates overall reasonable agreement with the majority of the observed data, especially for the IR mainstem sites, and for the two major sites of concern, at the AR/OK state line and at Tahlequah."

When the comparison shifts to monthly nutrient loads at the State Line and Tahlequah the results are less conclusive. For example, Table 2-27 in Baker et al. (2015b) indicates that the value from LOADEST at Tahlequah for TP is over-estimated by 15% on average, while Table 2-28 indicates that total P from HSPF may be 30% higher on average than that from LOADEST. Thus, the total P load may be 50% higher than the data immediately upstream of Lake Tenkiller, which would be a huge problem for the intended use of the models. Baker et al. (2015b, p. 99) further states "The high correlation coefficients (0.9 for TN and 0.8 for TP) suggest that the two models demonstrate significant agreement." This statement is misleading. On the following pages, I have added 1:1 lines to Figures 2-33 and 2-34 and from these modified figures much poorer agreement between the models can be seen than indicated by the correlation coefficients.

The modified Figure 2-33 shows that the monthly loads of TP estimated by HSPF and by Haggard (see p. 97 of Baker et al. (2015b)) at the State Line (i.e. Siloam Springs, AR) are widely distributed around the 1:1 line (in red) but the load comparison is unbiased. However, for TN the modified Figure 2-33 shows that the monthly load estimates from Haggard typically are substantially higher than the monthly load estimates from HSPF. The modified Figure 2-34 shows that the monthly loads of TP estimated by HSPF are consistently and substantially

higher than those estimated by LOADEST at the Tahlequah (immediately upstream of Lake Tenkiller) for loads less than about 30,000 lbs/mo. The modified Figure 2-34 shows that the monthly loads of TN estimated by HSPF are consistently and substantially higher than those estimated by LOADEST at the Tahlequah (immediately upstream of Lake Tenkiller) for loads less than about 300,000 lbs/mo. These results imply that the nutrient loads into Lake Tenkiller may be substantially over-estimated by the HSPF model of the IRW.

The monthly nutrient load comparisons in the modified Figures 2-33 and 2-34 are not necessarily proof of deficiencies in the HSPF model of the IRW because the comparison is between two different types of estimates of the monthly nutrient loads—statistical models (Haggard at the State Line and LOADEST at Tahlequah) and process based model (HSPF). A cleaner comparison of the HSPF estimated nutrient loads and the measured data can be done by developing measured and simulated nutrient load rating curves. The measured load is computed as the daily mean measured flow times the measured concentration converted to pounds per day, and the simulated load is computed as the daily mean simulated flow times the daily mean simulated concentration. Because the measured concentrations represent point-in-time grab samples, these observed loads represent uncertain estimates, but considered over all measurement dates in the loading curves, useful comparisons can be made. Figure 1 shows the simulated and observed sediment loading curves for the Menomonee River at sampling site RI-09 in Milwaukee, WI, the fact the observed loads fall within the range of the simulated loads indicates that the simulation model is performing well. Conversely, Figure 2 shows the simulated and observed nitrate loading curves for the Cottonwood River near New Ulm, MN, and it can be clearly seen that the model substantially overestimates the nitrate loads for flows less than 1000 cfs, and, thus, the model needs further adjustment to reliably simulate nitrate for this watershed. The possible use of such sediment rating curves in the calibration and verification of HSPF was discussed in Baker (2013, p. 36). Nutrient rating curves could be developed for the IRW HSPF model to confirm the usefulness of the model. This would be a powerful comparison because the water-quality management problems in Lake Tenkiller are more dependent on the pollutant loading from the IRW. Thus, demonstrating that the model reliably reproduces the observed pollutant loading is vital.

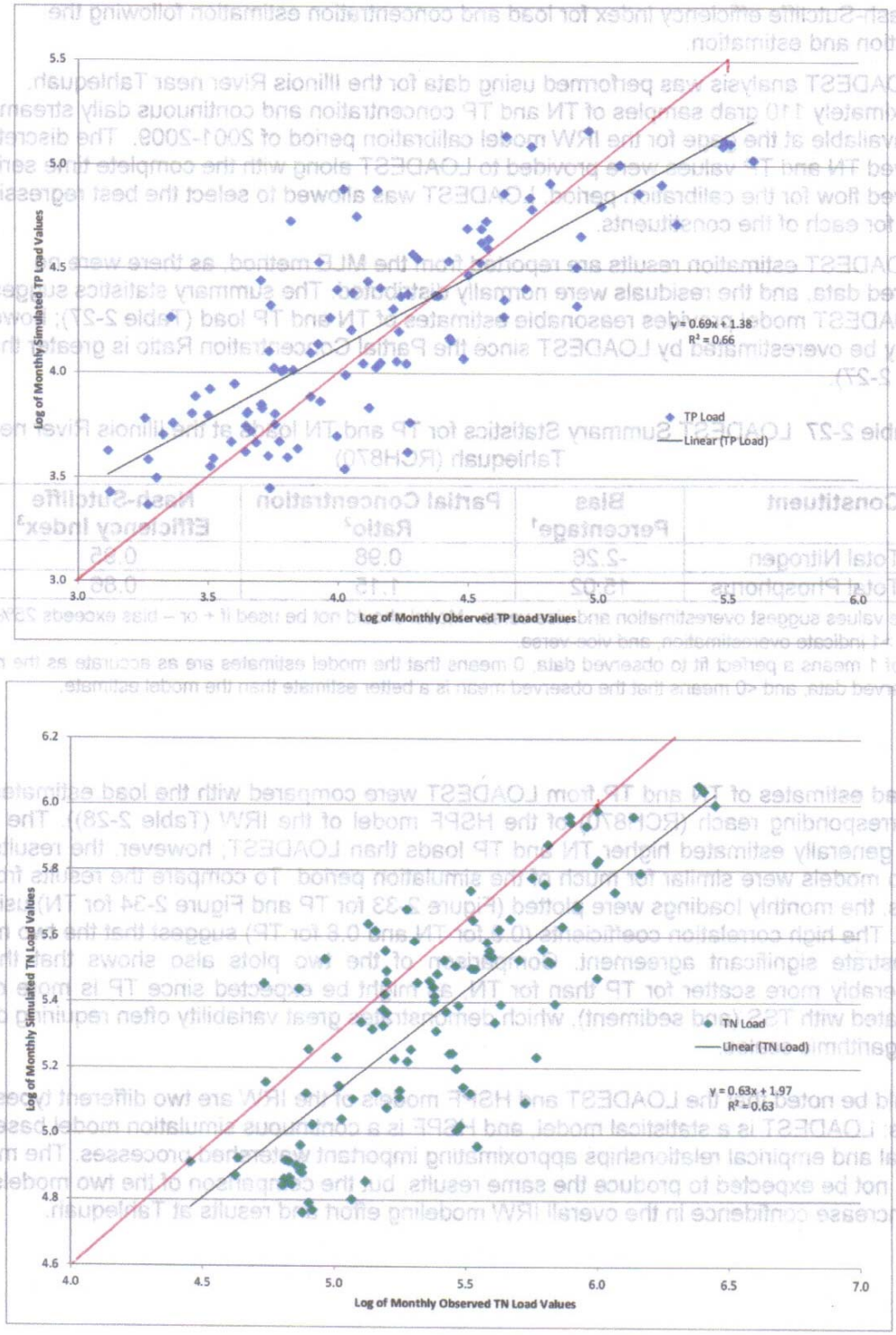


Figure 2-33 Scatterplots of TP (top) and TN (bottom) Monthly Loads at the IR South of Siloam Springs, AR

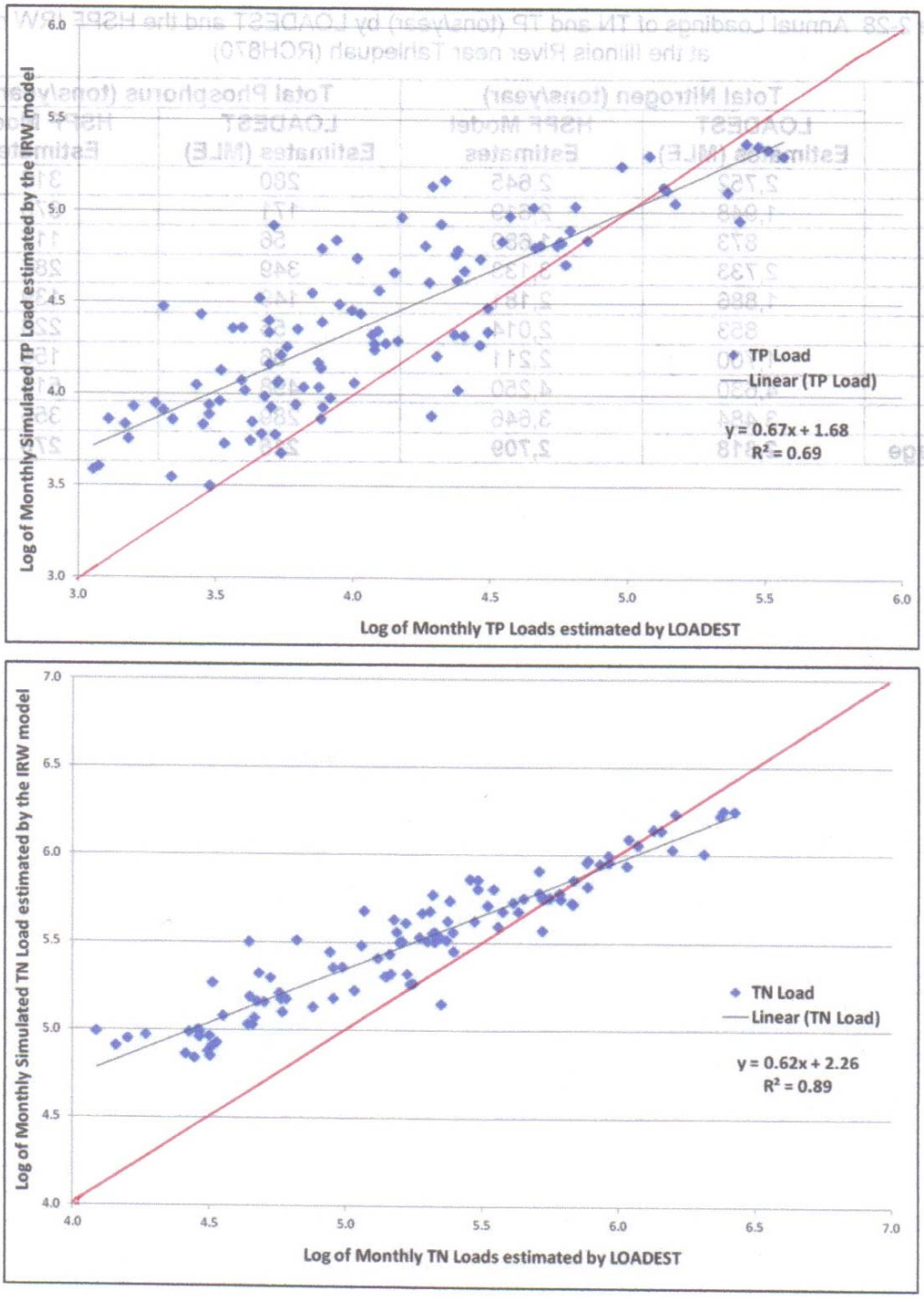


Figure 2-34 Comparison of Monthly Loading (lbs/mo) of TP (top) and TN (bottom) estimated with LOADEST and the HSPF IRW model at the Illinois River near Tahlequah (RCH870)

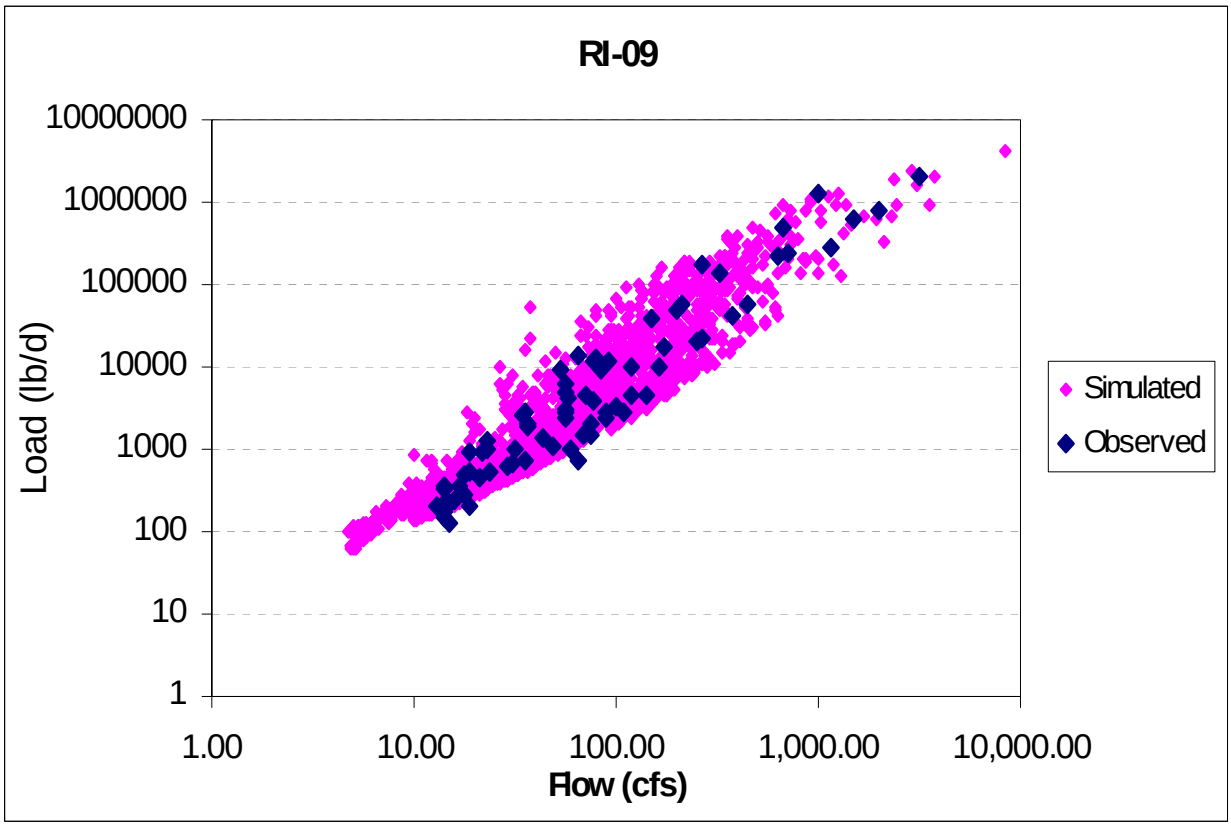
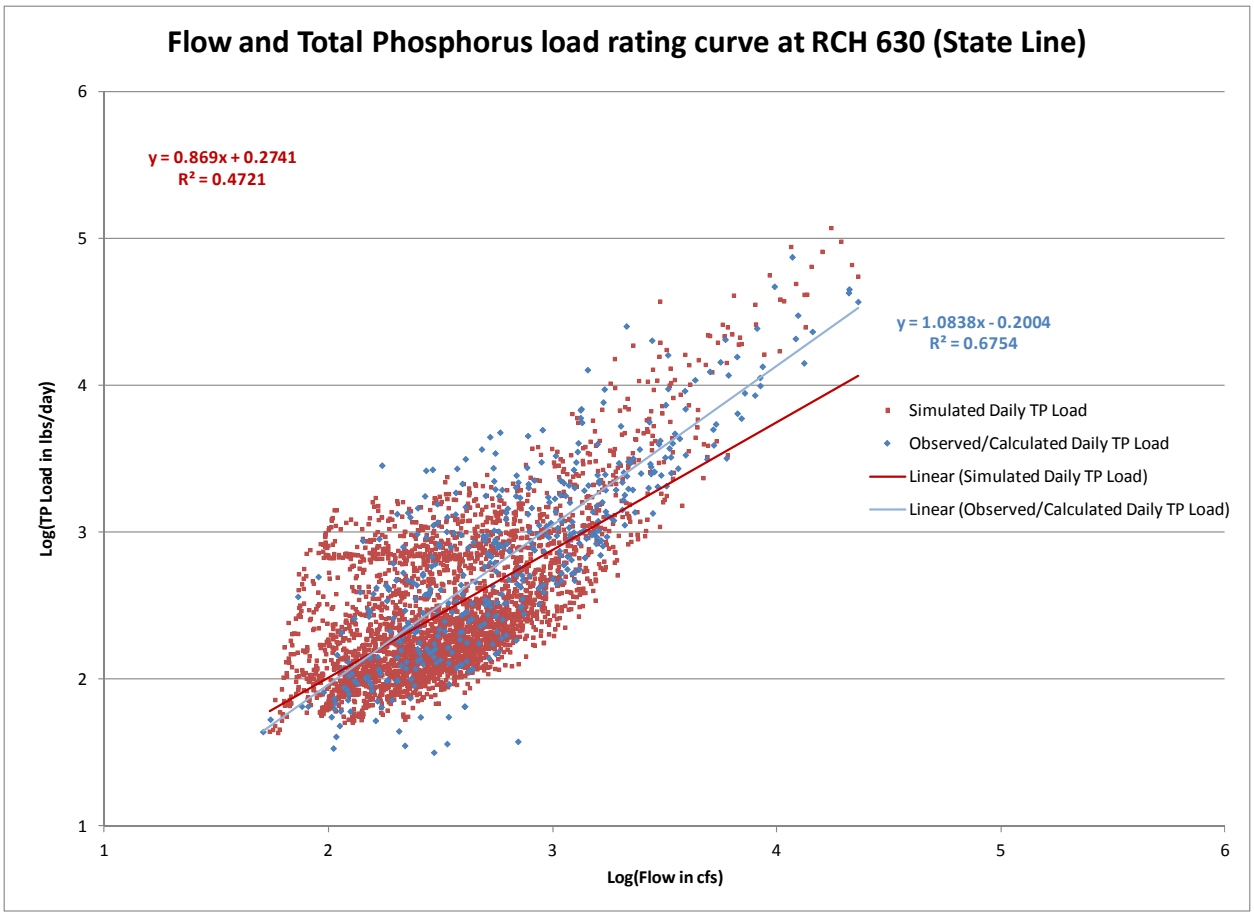
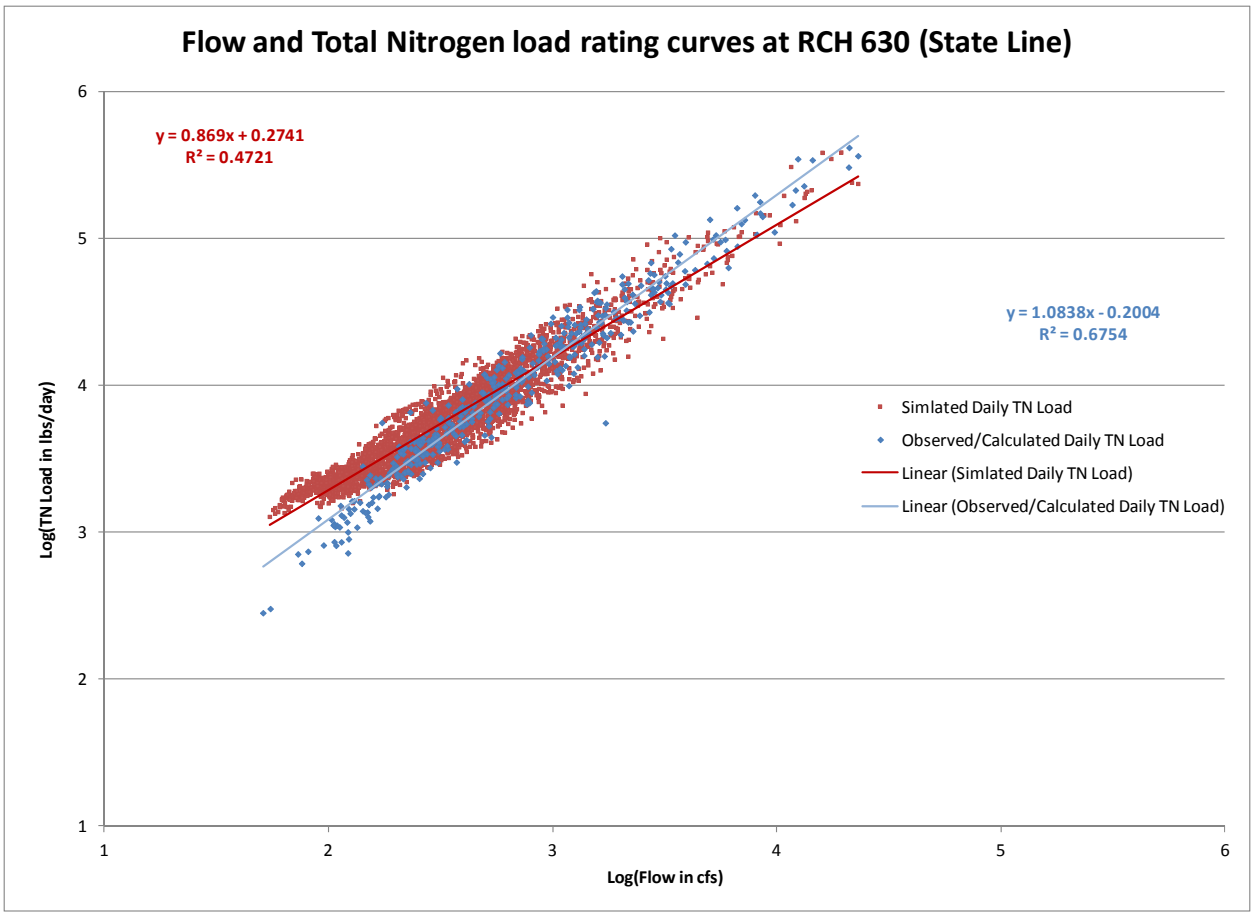


Figure 2. Simulated and observed sediment loading curves for the Menomonee River at sampling site RI-09 at Milwaukee for 1994-1998.

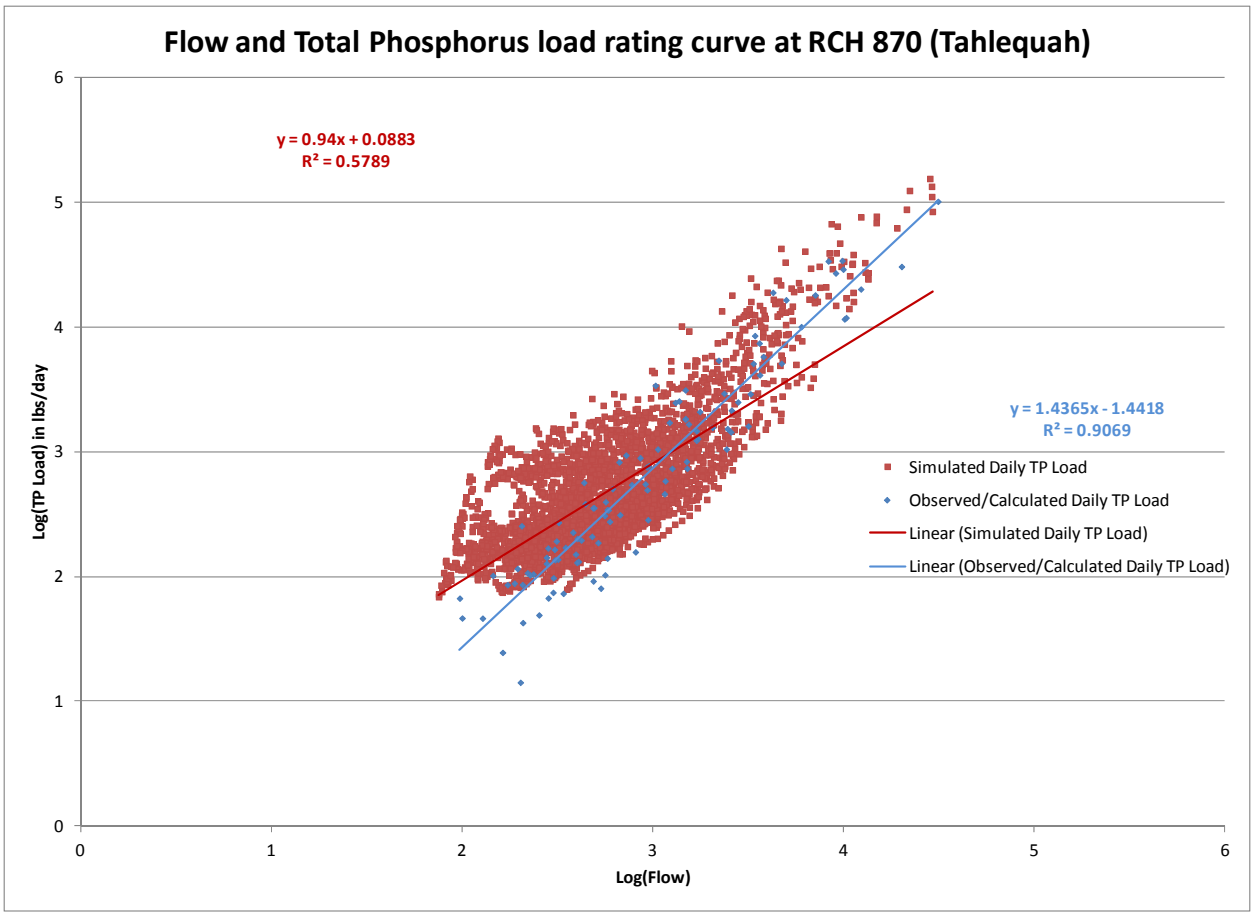
Response #4.

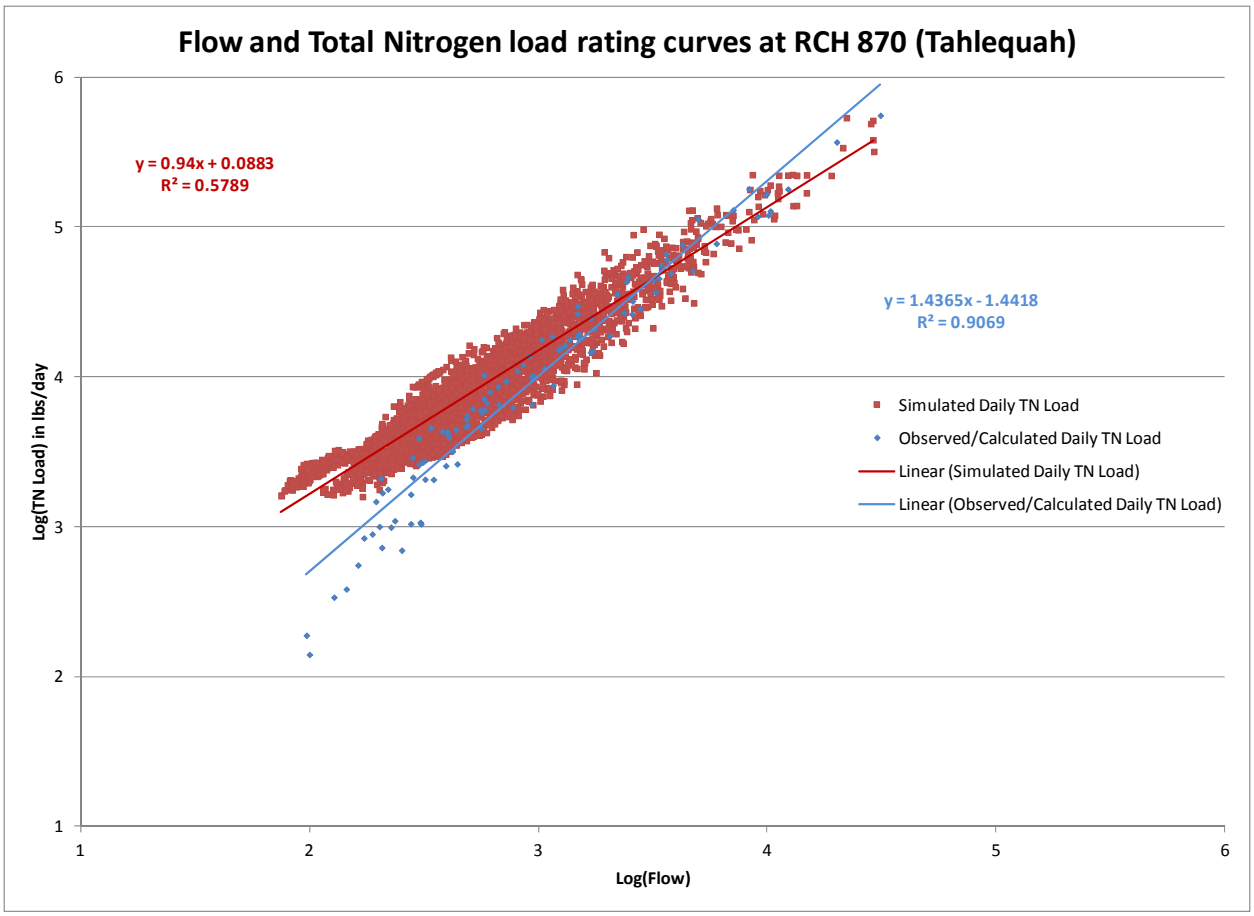
- The drought comment in the review comments (above), which the Reviewer #1 said "makes no sense", refers to the period from mid-2005 to late spring or early summer 2006, NOT to the entire WYs of 2005 and 2006. The daily flows with a log scale (Figures in Appendix A.1.4, log plots) clearly show the drought period extended over this time period, and ended with some small-moderate storms in spring-summer 2006; the WY ending in September 2006 was subsequently over-simulated due to some late summer/early fall storms, but it was still one of the dryer years in the calibration period.
- To address the nutrient rating curve issue as suggested by Reviewer #1, based on daily loads computed by Brian Haggard, rating curves were developed for the Stateline. The figures shown below provide those results for TP and TN at Reach 630, just upstream of the Stateline. The results show good consistency between the modeled and observed values for most of the range of the observed flows. For the TN plot the lower values appear to be slightly over-simulated below about 500 cfs.


Flow and Total Phosphorus load rating curve at RCH 630 (State Line)
y = 0.869x + 0.2741
R² = 0.4721
y = 1.0838x - 0.2004
R² = 0.6754
Simulated Daily TP Load
Observed/Calculated Daily TP Load
Linear (Simulated Daily TP Load)
Linear (Observed/Calculated Daily TP Load)
Log(TP Load in lbs/day)
Log(Flow in cfs)
1
2
3
4
5
6



The corresponding nutrient loading curves for Tahlequah are shown below. For Tahlequah there are fewer data points, as compared to the Stateline site, with most of the data points concentrated at lower flow levels.


Flow and Total Phosphorus load rating curve at RCH 870 (Tahlequah)
y = 0.94x + 0.0883
R² = 0.5789
y = 1.4365x - 1.4418
R² = 0.9069
Simulated Daily TP Load
Observed/Calculated Daily TP Load
Linear (Simulated Daily TP Load)
Linear (Observed/Calculated Daily TP Load)
Log(TP Load) in lbs/day
Log(Flow)
0
1
2
3
4
5
6



The 1:1 line in Figure 2-33 is in error, as the origin is NOT 0:0; below is the corrected figure with a proper 1:1 line.

INSERT CORRECTED Figure 2-33

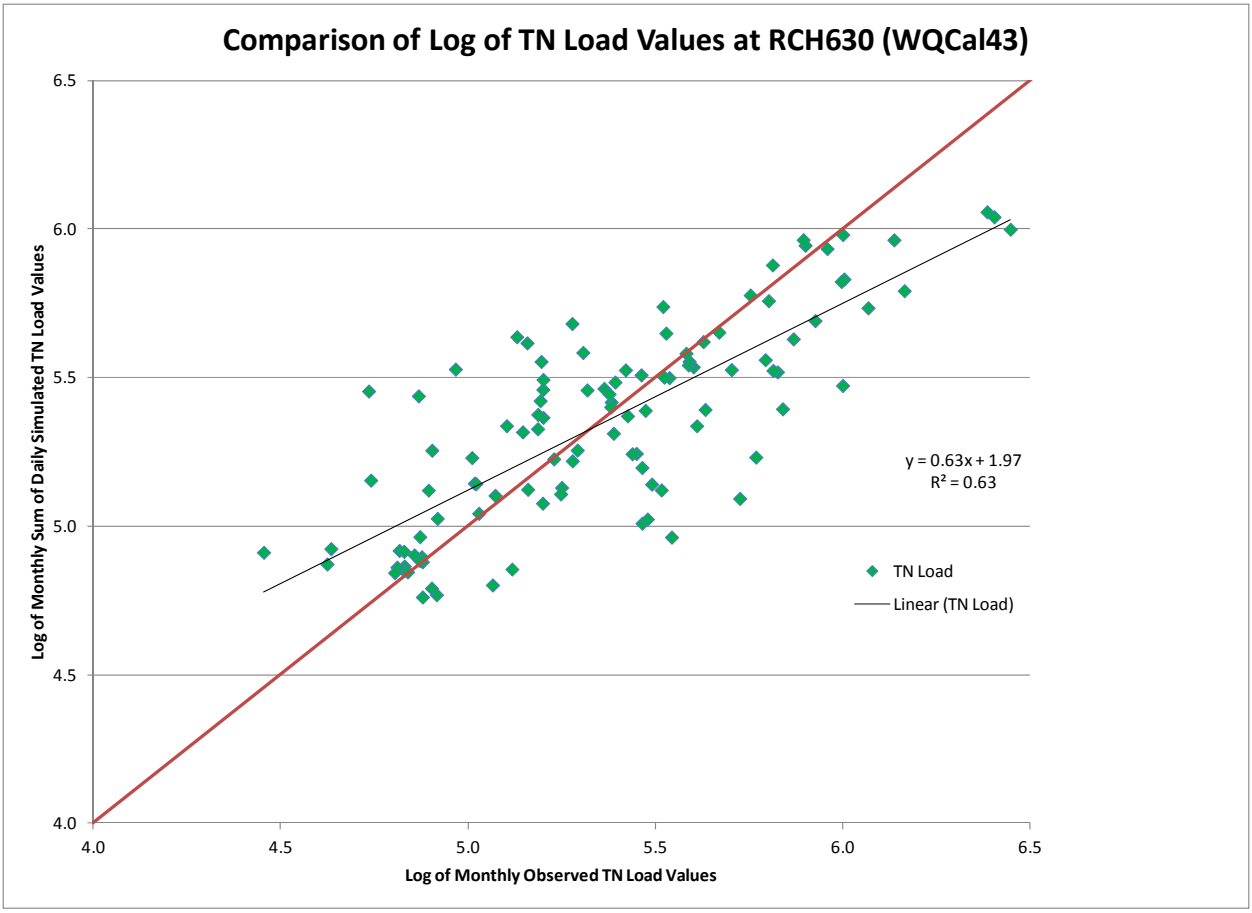


Figure 1. Corrected 1:1 line on Figure 2-33

Comment #5.

Environmental Fluid Dynamics Code (EFDC)

Lake Tenkiller comprises almost 13,000 acres of water and 130 miles of shoreline and its maximum depth exceeds 100 ft. Lake Tenkiller is represented in the EFDC model by 1443 horizontal cells with 16 even thickness vertical sigma layers used to represent vertical spatial resolution (Baker et al., 2015b, p. 106). This discretization of Lake Tenkiller is a substantial improvement from the 195 horizontal cells and 10 vertical layers applied in the original application of EFDC to Lake Tenkiller as described in Baker (2013, p. 21). It would seem that this discretization should reasonably represent the physical conditions of Lake Tenkiller. However, Baker et al. (2015b, p. 142) noted "The water temperature stratification simulated by EFDC in summer time is less than the observed data, which is caused by the artificial vertical numerical diffusion introduced by the sigma grid." Thus, even with the greatly increased discretization some numerical diffusion still occurs in the model results.

EFDC Calibration and Validation

If just the graphical and statistical comparisons of measured and simulated values in Lake Tenkiller presented in Baker et al. (2015b) are considered the EFDC model would appear to reasonably represent the physical conditions in Lake Tenkiller as discussed in the following paragraphs.

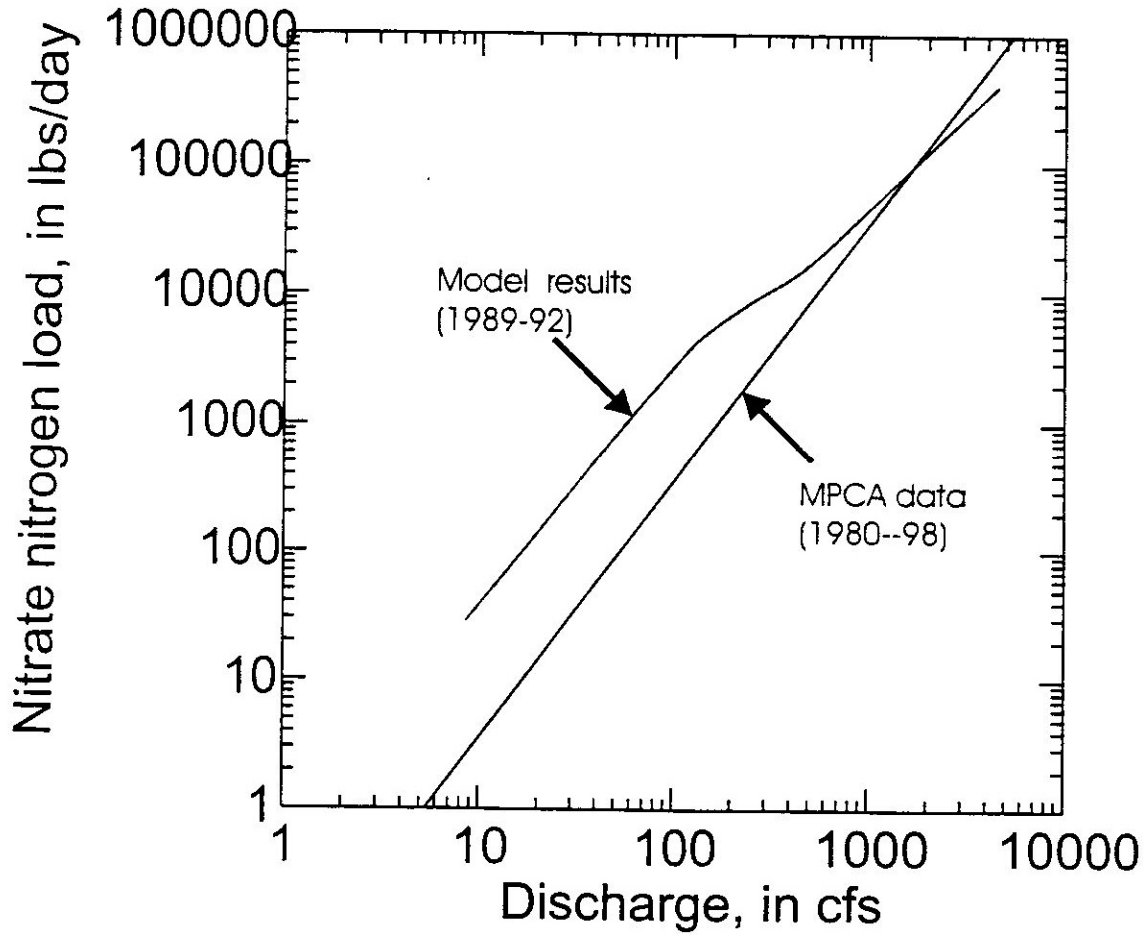


Figure 2. Simulated and observed nitrate loading curves for Cottonwood River near New Ulm, Minnesota.

For the simulated lake stages for the calibration period (2006) the calculated Root Mean Square (RMS) error was 0.029 m and the relative RMS error was 0.6% (Baker et al., 2015b, p. 138) and for the validation period (2005) the calculated RMS error was 0.022 m and the relative RMS error was 0.3% (Baker et al., 2015b, p. 139). For the simulated temperature in the surface layer the relative RMS errors ranged from 4.5-10.1% (considered a "Good" result) and in the bottom layer the relative RMS errors ranged from 15.0-27.1% (considered a "Fair" result) (Baker et al., 2015b, p. 141). For the simulated Dissolved Oxygen (DO) the calculated relative RMS errors ranged from 15.3% at the bottom layer of station LK-04 to 30.3% at the surface layer at station LK-03 (Baker et al., 2015b, p. 157). Some of these results are a little higher than the calibration target of 20% (Baker et al., 2015b, p. 137), but are still acceptable. For the simulated algae concentration the calculated relative RMS errors ranged from 25.2% at the surface layer of station LK-04 to 67.0% at the bottom layer of station LK-04 (Baker et al., 2015b, p. 167). These values are well within the calibration target of 100% (Baker et al., 2015b, p. 137). The majority of observed ammonia data were labeled as less than the detection limit of 0.1 mg/L, and, in most of the cases, the simulated ammonia values were less than or very close to the detection limit of 0.1 mg/L (Baker et al., 2015b, p. 177).

If the relative RMS errors are considered the results for TSS and Total Organic Carbon are much higher than the calibration targets given by Baker et al. (2015b, p. 137). These high percentage errors are a function of the low measured TSS and TOC concentrations in Lake Tenkiller and the subsequent small range in measured concentrations. A better way to gage the accuracy of the model for these constituents is to consider the actual RMS errors and differences in the averages of these constituents. For TSS, RMS errors range from 2.68 to 17.94 mg/L and the differences in the averages range from -2.302 to 0.327 mg/L (Baker et al., 2015b, p.152), which indicates good agreement between the measured and simulated values. For TOC RMS errors range from 0.742 to

1.154 mg/L and the differences in the averages range from -0.888 to 0.845 mg/L (Baker et al., 2015b, p.173), which indicates good agreement between the measured and simulated values.

Generally, the EFDC simulated nitrate (NO_3) concentrations are consistently higher than the observed NO_3 data. Baker et al. (2015b, p. 178) suggested "One likely reason for the model-data discrepancy might be the unusually high NO_3 inputs from the HSPF derived upstream boundaries for the Illinois River and Baron Fork Creek." This may also be related to the over-estimation of TN loads previously discussed with respect to the modified Figure 2-34. Baker et al. (2015b, p. 178) further concluded "Considering the fact that Tenkiller Ferry Lake is a phosphorus-limited eutrophication lake and the calculated relative RMS for NO3 are less than 100% in most cases, the model results of NO3 are deemed to be acceptable."

Baker et al. (2015b, p. 201) further noted:

> "Performance targets for surface and bottom layer total phosphorus (TP) were either met, or were very close to the performance target of 50% for the lacustrine stations LK-01 (39-87%) and LK-02 (29-60%). For the riverine station (LK-04) (37-117%) and the transition zone station (LK-03) (76-78%), however, the model results were somewhat higher than the 50% target for TP."

The worst performance of the EFDC water-quality model was obtained for PO_4. Baker et al. (2015b, p. 189) noted:

> "The very high Relative RMS errors of 533% and 571% estimated at station LK-03 result from two factors. The first factor is the over-prediction of the TPO4 loading into the lake and the modeled in-lake response to the high loading during April-May 2006. The second factor is the very low range of observed TPO4 (0.007 to 0.008 mg/L) that includes numerous measurements in both 2005 and 2006 that are at the detection limit for TPO4."

Again these phosphorus species results may be related to the over-estimation of TP loads previously discussed with respect to the modified Figure 2-34.

The foregoing results imply the EFDC model reliably simulates all properties and chemical constituents in Lake Tenkiller except for NO_3, PO_4, and TP, and the errors in NO_3, PO_4, and TP are more related to the overestimated inflow loads than to the simulation routines in EFDC. However, the calibration and validation of the EFDC model of Lake Tenkiller has a bigger problem. Baker et al. (2015b, p. 105) stated "There are more observed water quality data available in year 2006; therefore, the Tenkiller Ferry Lake EFDC model is calibrated for the period of 1 January 2006 through 31 December 2006 and validated for the period of 1 January 2005 through 31 December 2005." However, as discussed earlier the flow on the Illinois River at Tahlequah was over-estimated by 54.15% for WY 2006. Therefore, the inflows to Lake Tenkiller during the calibration period are over-estimated on the order of 50% and subsequently the pollutant loads into the lake must also be substantially over-estimated. Thus, the parameters of the EFDC model must have been adjusted to compensate for the high inflows and pollutant loads, and in the cases of NO_3, PO_4, and TP the parameters could not be adjusted enough to yield good agreement between simulated and measured concentrations.

A specific example of the adjustment of the EFDC model to compensate for the over-estimated inflow is discussed by Baker et al. (2015b, p. 110) as follows:

> "The water supply withdrawals from Tenkiller Ferry Lake were not available; therefore, a flow balance was estimated using all inflows including HSPF simulated watershed flows, and rainfall, and all outflows including evaporation and flow release at the dam. A flow balance was computed to ensure that the EFDC model simulated lake stage matched the observed lake stage."

This adjustment may have a very serious impact on the water quality simulation for Lake Tenkiller because the over-estimated inflow will have high nutrient concentrations, whereas the water withdrawn to achieve the flow

balance will have lower nutrient concentrations resulting in a build-up of nutrients Lake Tenkiller as has been observed in the model results in Baker et al. (2015b).

If the calibration year were one of the years when HSPF yielded flows similar to the measured flows, using the HSPF outputs as the inputs to the EFDC model for Lake Tenkiller would be a reasonable approach to calibration. However, in this case using the HSPF outputs for 2006 as the inputs to the EFDC model for Lake Tenkiller is not a reasonable approach to calibration. A better, more appropriate, calibration of the EFDC model of Lake Tenkiller would be obtained by using the measured inflows and concentrations as the inputs to the EFDC model. Flows from ungaged tributaries and overland areas could be estimated on an area ratio basis and concentrations extrapolated to similar watersheds. Of course, there will be errors and uncertainties as the monthly water quality measurements are extrapolated to estimate daily loads to Lake Tenkiller, but these errors are likely to be far smaller than the 50% over-estimate in flows and subsequent over-estimates in pollutant loads coming from the results of the HSPF model for 2006. Once the EFDC model for Lake Tenkiller has been properly calibrated, then it can be used with the output from the calibrated HSPF model to evaluate various scenarios of changed loads from the IRW in the development of TMDLs for phosphorus.

Response #5. This comment is in regards to EFDC lake model performance for the 2005-2006 validation and calibration years. The reviewer reiterates what was written in the lake model report that identifies discrepancies in model calibration performance for nutrients that can be traced to inflows from the HSPF watershed model that were higher than observed flows and observed concentrations of nutrients. The reviewer recommends that the lake model be re-calibrated to 2006 conditions using observations of flow and water quality data to drive external boundary inflows to the lake model as an alternative to the HSPF watershed model results. The reviewer makes the assertion that any errors associated with external boundary conditions being driven by observed streamflow and water quality data "… are likely to be far smaller than the 50% over-estimate in flows and subsequent over-estimates of pollutant loads coming from the results of the HSPF model for 2006". Further, the reviewer implies that the EFDC model parameterization has been unreasonably adjusted to achieve calibration-validation of the water quality model. While this statement is presented as quantitative it is speculative and not supported by a detailed quantitative analysis of the model parameterization or model results.

EPA and the Baker Team do, however, recognize that the reviewer's concerns about EFDC model calibration and the recommendation for an alternate approach using observed data are based on improving performance of the lake model. In recommending that observed flow and water quality data be used to develop external boundary conditions, the reviewer erroneously implies that the inflow and loading from the HSPF watershed model was significantly overestimated for the entire model calibration period of 2006. The greatest discrepancy between EFDC model results and observations, however, occurred during a single storm event in May 2006. After the storm event, streamflow decreased and remained much lower than long-term daily average flow conditions until late August 2006. The large discrepancy seen in the EFDC model results for May 2006 is directly related to the overestimation of streamflow and nutrient loading data provided by the HSPF model as the simulated response to the storm event.

With the exception of the storm-event driven model-data discrepancy in May 2006, the EFDC model-observed data comparison for all other sampling events during the two year (2005-2006) calibration and validation period was considered reasonable. In developing the EFDC model calibration based on observed 2006 water quality conditions, the Baker Team most definitely did not adjust water quality kinetic parameter values beyond acceptable ranges to achieve "… compensation for the high inflows and pollutant loads…" as implied by the reviewer. Model performance results were computed for the period of all water quality records available from June 2005 through September 2006. The EFDC model kinetic coefficients and model parameters developed for model calibration (documented in Appendix A of the EFDC model report) show that values developed for model calibration are within accepted ranges as reported in the literature.

The methodology used to develop the flow balance to ensure that the water level, surface area, volume and residence time of the simulated lake is accurate and is an accepted practice for developing hydrodynamic models of reservoirs. A flow balance for the Lake Tenkiller model is needed to account for the known variance of flow measurements provided by the USGS, unknown inflows, and unknown outflows such as leakage identified by the US Army Corp of Engineers at the dam and unknown water withdrawals for local water supply systems served by Lake Tenkiller. On a daily average basis, the volume accounted for by the flow balance represented less than 0.1%, of the average volume of the lake during 2005-2006. When flow balance water is withdrawn from the lake, the mass of nutrients removed from the lake will correspond to the local nutrient concentration simulated at lake locations assigned for the flow balance. Since flow balance withdrawals will export water with nutrient concentrations that may or may not be lower than watershed inflow concentrations, and the volume over the simulation period is relatively small, it is not correct to conclude that the flow balance adjustments will result in "a build-up of nutrients" and a "very serious impact" on the overall water quality simulation for Lake Tenkiller.

To further demonstrate the EFDC models capability to simulate the physics and WQ conditions for 23 of the 24 month simulation period, the Baker Team will revise the compilation of model performance statistics for the RMS error and the Relative RMS error by excluding observations in May 2006 when the HSPF model was not in good agreement with observed streamflow records. Our re-evaluation of model performance for water quality state variables will thus filter out paired lake model and observed data sets that represent the period of the few weeks in May 2006 when the HSPF streamflow results were much higher than the observed records. With May 2006 storm event driven-lake observations removed, lake model performance statistics will be re-computed, revised tables for model performance presented, and discussion of model performance will be edited as needed in the revised report.

Comment #6.

Summary

In general, the application of HSPF to the IRW is consistent with "best current engineering practice and scientific knowledge" and reasonably represents the physical conditions of the IRW. However, measured and simulated nutrient load rating curves should be developed and compared for the State Line and Tahlequah to confirm the reliability of the HSPF simulated nutrient loads. The EFDC model needs to be recalibrated using measured flows and concentrations (and appropriate extrapolations to ungagged tributaries and direct drainage areas) as inflows to Lake Tenkiller. Once the EFDC model for Lake Tenkiller has been properly calibrated, then it can be used with the output from the calibrated HSPF model to evaluate various alternative watershed management scenarios of changed loads from the IRW in the development of TMDLs for phosphorus.

Response #6.

Refer to the nutrient loading curves under Response#4.

The EPA and Baker Team do not agree that recalibration of the EFDC model is warranted. In recommending that observed flow and water quality data be used to develop external boundary conditions, the reviewer erroneously implies that the inflow and loading from the HSPF watershed model was significantly overestimated for the entire model calibration period of 2006. With the exception of the storm-event driven model-data discrepancy in May 2006 when the HSPF model did overestimate streamflow and nutrient loading, the EFDC model-observed data comparison for all other sampling events during the 2005-2006 calibration and validation period was considered reasonable.

Comment #7.

2) Is the model sufficient to reasonably model water quality conditions in the IRW, and be a useful tool for developing numeric TMDLs for nutrients in the watershed?

With respect to the application of the HSPF model to the IRW the following statement was made in the answer to Question 1: "The simulated concentrations of Total Nitrogen (TN), Orthophosphate (PO_4), and Total Phosphorus (TP) at the State Line and Tahlequah generally agree well with the measured concentrations as shown in Figures 2-29 to 2-32 in Baker et al. (2015b)." Thus, the HSPF model reasonably models the transport of nutrients from the IRW through the Illinois River system. Additional confidence in this statement can be gained through the comparison of measured and simulated nutrient load rating curves for the State Line and Tahlequah as suggested in the answer to Question 1.

With respect to the application of the EFDC to Lake Tenkiller, good calibration results were obtained for nearly all water quality constituents—including TP and ammonia but not PO_4, TN, and NO_3 (but it was reasoned that because phosphorus was the cause of algal problems in Lake Tenkiller the shortcomings in nitrogen species simulation were acceptable). These generally good results were obtained even though the inflow and related constituent loads to the lake were substantially over-simulated for the calibration year of 2006. Thus, it is likely that if the EFDC was recalibrated using the measured lake inflows and constituent concentrations similar calibration quality can be obtained, or even improved in the case of nutrients, while composing a more realistic and reasonable representation of water quality processes in Lake Tenkiller.

Thus, the calibrated HSPF model and recalibrated EFDC model should be useful tools for developing numeric TMDLs for nutrients in the watershed.

Response #7.

Probably refer back to Response #6.

The EPA and Baker Team do not agree that a recalibration of the EFDC model is warranted. In recommending that observed flow and water quality data be used to develop external boundary conditions, the reviewer erroneously implies that the inflow and loading from the HSPF watershed model was significantly overestimated for the entire model calibration period of 2006. With the exception of the storm-event driven model-data discrepancy in May 2006 when the HSPF model did overestimate streamflow and nutrient loading, the EFDC model-observed data comparison for all other sampling events during the 2005-2006 calibration and validation period was considered reasonable.

Comment #8.

3) ***Does the interface between the watershed model (HSPF) and the reservoir model (EFDC) seem reasonable?***

The interface between the watershed model (HSPF) and the reservoir model (EFDC) has been done in a manner consistent with "best current engineering practice and scientific knowledge" and reasonably represents the physical conditions of the IRW.

The technical aspects of the interface between the HSPF and EFDC models reasonably represents the physical conditions of the IRW, however, the linkage between the two models for the evaluation of scenarios related to the development of TMDLs needs some clarification. The HSPF model is run for 9-year calibration (2001-2009) and validation (1992-2000) periods, but the EFDC model is run for one year at a time in the calibration and validation. Thus, it is not clear how the changes in runoff and pollutant loads resulting from various phosphorus mitigation scenarios simulated for 9-year periods with HSPF will be entered into the EFDC model, i.e. will the EFDC model be run for the full 9-year periods or just "representative" years? Further, the Simulation Plan (Baker, 2013) and other documents talk about doing a spin-up analysis to determine how long it takes for quasi-equilibrium conditions to develop in Lake Tenkiller for various inflow conditions, e.g. (Baker, 2013, p. 72) states "When the calibration effort is completed, the lake model will be used to determine the "spin-up" time needed for the sediment flux model to attain quasi-equilibrium conditions driven by the existing watershed loads used for input to the lake model." Will similar spin-up analyses be done for the changed inflow conditions resulting from the various phosphorus mitigation scenarios? These functional aspects of how the two models will be used together must be better explained in the final model reports.

Response #8.

The Simulation Plan notes that the EFDC model of Lake Tenkiller will be used to assess lake impacts of the HSPF watershed load reduction scenarios based on a uniform percent reduction of TP, TN, and/or TSS (i.e., across the board). Specific BMP or wastewater discharge scenarios considered for phosphorus mitigation for the HSPF watershed model scenario assessments are not used to determine the lake model response.

In the technical memorandum "Nutrient Load Reduction Scenario Modeling of Tenkiller Ferry Lake, Oklahoma" dated February 2015, the Baker Team followed the approach outlined in the Simulation Plan for evaluation of the lake model response to a series of uniform reduction scenarios ranging from 20% to 75% removal of nutrients and sediment loading from the watershed. The HSPF nutrient and sediment loading generated for 2006 was used as the input to the lake model for the calibration year of 2006. The calibration year for the lake model was used as the baseline for assigning uniform reductions of nutrient and sediment concentrations from all HSPF tributary and distributed catchments.

The 65% removal scenario for the uniform reduction of sediment and nutrients was selected for the detailed spin-up simulation analysis of the long-term water quality response of the lake to reductions in watershed loads. The 65% removal scenario was used to simulate 16 years of sequential "spin-up" runs to evaluate the long-term response of water quality and sediment bed conditions in the lake to the reduction in external loads from the watershed. "Spin-up" runs are required with the sediment diagenesis model to allow the sediment bed to slowly attain a new periodic steady state condition that reflects the changes in external watershed loading from the 65% removal scenario. Watershed flow and reduced pollutant loading derived from the HSPF model (for the 2006 model calibration) were repeated for each of the 16 spin-up years. After 16 spin-up years, the lake water quality response is intended to represent sediment bed concentrations and fluxes under new equilibrium conditions.

Comment #9.

4) ***Given the point and non-point source location and loading data that have been identified to date for agricultural operations (e.g., poultry, hogs, cattle, manure), is this sufficient to run the model and develop alternate watershed management scenarios?***

As noted earlier, the HSPF application to the IRW took the wise and appropriate approach of applying the more agronomically detailed AGCHEM routines to simulate the nutrient balance for the pasture areas receiving fertilizer, manure, and poultry litter. In fact, 4 different kinds of pasture areas were defined depending on overland slope and whether or not poultry litter was applied. With respect to agricultural operations in using fertilizer, manure, and poultry litter on the various sub-areas in the IRW, the simulations done with HSPF focused to a large extent on understanding and properly characterizing the poultry litter—the amount and timing of its use and its nutrient content. This emphasis on poultry litter is reasonable considering the following from Baker et al. (2013, p. 83): "With more than 30 million birds in the watershed, generating more than 300,000 tons of litter each year, it is critical to attempt to represent the potential impacts of this source of P as part of the overall balance of P for this watershed."

Baker et al. (2015b, p. 84) state the following with respect to the available data on poultry litter and its spatial application:

> "The poultry litter nutrient applications were estimated primarily from data provided by the Arkansas Natural Resources Commission (ANRC, by E. Swaim and P. Fisk, multiple personal communications in 2011 through 2013) and the Oklahoma Department of Agriculture, Food and Forestry (ODAFF, Q. Pham) on litter generation, application, and export on both sides of the state line. While the OK data was provided for each 12-digit HUC, the AR data was almost exclusively for the entire IRW within AR. Only for 2009 to 2011 were litter application data provided by 12-digit HUC *[Hydrologic Unit Code]* for the AR side, and they were overlain with the 12-digit HUC coverage and the 2009 spatial distribution was used for the entire calibration period." (words in italics added)

Pages 84-86 of Baker et al. (2105b) further describe the various assumptions used to represent the application of the poultry litter in the IRW through the Special Actions block in HSPF. The assumptions detailed on these pages seem reasonable and sufficient for the calibration and validation of the HSPF model and to form a baseline condition to which to compare the model results for reduced nutrient inputs to the pasture areas of the watershed as part of alternative watershed management scenarios that might be considered in the development of TMDLs.

Response #9. Concur with the comment.

Comment #10.

5) ***Does the model appear to be sensitive enough to allow compliance assessments relative to the State of Oklahoma's 0.037 mg/L TP standard?***

First, it should be remembered that the HSPF simulated TP and PO_4 concentrations at the State Line (RCH 630) were found to be in good agreement with the measured values. Further, in the evaluation of the modified Figure 2-33 it was found that the monthly loads of TP at the State Line estimated with HSPF agreed with the monthly load estimates of TP from Haggard in an unbiased way. Thus, the estimated phosphorus delivered from Arkansas into Oklahoma in the simulations is reasonably accurate and unbiased.

The Sensitivity Analysis of the HSPF simulation done by Baker et al. (2015a, p. 7) indicates that TP phosphorus concentration is sensitive to precipitation, the infiltration parameter (INFILT), the critical shear stress for scour (TAUCS) and deposition (TAUCD), and the point source load. Further, the TP load is sensitive to precipitation, TAUCS and TAUCD, 6 hydrologic model parameters, point source load, and Non-Litter Pasture Loading. The sensitivity of TP to point source load and Non-Litter Pasture Loading is important because these are model inputs

that could be changed through various alternative watershed management scenarios to be considered in the development of TMDLs. Therefore, the model is sensitive enough to allow a compliance assessment of the State of Oklahoma's 0.037 mg/L TP standard relative to the application of various alternative watershed management scenarios.

It is interesting to find that the TP results were not substantially sensitive to changes in Litter Pasture Loading. I wonder if this is because "Pasture 1 – Litter" is a small fraction of the total pasture area that is the primary source of TP in the modeling. For example, Table 3 of Baker et al. (2015a, p. 7) indicates that TP load has a sensitivity coefficient of 22.6% for Non-Litter Pasture Loading while Figure 3 of Baker et al. (2015a, p. 6) indicates that TP load has a sensitivity coefficient a little less than 10% for Litter Pasture Loading, so Non-Litter Pasture Loading has more than double the impact of Litter Pasture Loading. However, if, for example, "Pasture 1 – Litter" represents 10% of the watershed area while the other 3 types of Pasture represent 30% of the watershed area the Litter Pasture Loading may have a bigger impact on TP than the Non-Litter Pasture Loading on a per area basis, and, thus, it can still be an important aspect of the various watershed management scenarios to be evaluated when developing TMDLs. It would be interesting to know what are the percentages of watershed area for the various types of pasture (I could not find it in any of the reports).

Response #10. The table below shows the acreage and '% of watershed' in each land use category, including the 4 different types of pasture. Pasture does make up more than 40% of the watershed area, while the litter pasture is about 4.5% of the total watershed. The litter pasture also comprises about 10% of the total pasture area in the watershed.

IRW Model Land Use Summarized

Category	Model (ac)	Model (mi^2)	Model %'s
Forest	451,843	706.0	42.8%
Pasture1	147,949	231.2	14.0%
Pasture2	67,500	105.5	6.4%
Pasture3	169,478	264.8	16.0%
Pasture1-Litter	47,606	74.4	4.5%
Total Pasture	432,533	675.8	40.9%
Grass/ Shrub/ Barren	44,751	69.9	4.2%
Wetlands	6,085	9.5	0.58%
Cropland	1,627	2.5	0.15%
Developed, Open (PLS)	61,679	96.4	5.8%
Developed, Low (PLS)	24,102	37.7	2.3%
Developed, Med/High PLS)	8,403	13.1	0.80%
Developed, Open (ILS)	1,254	2.0	0.12%
Developed, Low (ILS)	3,572	5.6	0.34%
Developed, Med/High (ILS)	5,359	8.4	0.51%
Total Developed	104,370	163.1	9.9%
Water	15,456	24.1	1.5%
Total	1,056,665	1,651.0	100.0%

6) Are there any overall concerns with the current models that would draw questions to future water quality predictions for nutrients?

See answer to Question 1.

7) ***Are there suggested improvements to the models (or model inputs) that would improve its use in developing load allocations to meet water quality objectives?***

See answer to Question 1.

Classification of Issues

Generally, in a Peer Review of a modeling effort done by or for U.S. EPA issues are expected to be classified into two categories in the report as follows:

- Significant issues impacting technical, system or usability of model (concerns with validity of results)
- Issues not impacting certification but recommended to be addressed in future revisions of the model

The majority of the "Significant Issues" are detailed in the foregoing answers to the specific questions evaluated earlier in this review. The remaining "Significant Issues" are related to the Sensitivity and Uncertainty Analyses of the EFDC model and are discussed in the following section.

Significant Issues

Comment #11.

Sensitivity and Uncertainty Analyses of the EFDC model

The sensitivity analysis of the EFDC model of Lake Tenkiller is incomplete and the uncertainty analysis is inadequate and inappropriate. The "uncertainty" of the EFDC model is computed from the statistics of model output variables computed from a sample of 17 simulations—the base, calibration/validation simulation and 16 sensitivity simulations involving one parameter-at-a time -50%, -25%, 25%, and 50% perturbations of each of the watershed TP load and three EFDC model parameters. This sample has no special statistical meaning. That is, the actual likelihood of having a case where three of the four considered input parameters are at their mean and the fourth input parameter is 25% or 50% high or low would be quite rare, and, thus, the meaning of the outputs of these simulations is not much different from any randomly generated Monte Carlo sample of 17 cases. As indicated in Baker et al. (2015a, p. 10) the Monte Carlo Simulation based uncertainty analysis of the HSPF model applied to the IRW required 500 simulations to achieve convergence with respect to the model output statistics of interest. Further, the EFDC model would, thus, require 500 simulations or maybe even more to achieve convergence for the model output statistics of interest. Thus, the 17 simulations considered for the EFDC model of Lake Tenkiller are an inadequate sample relative to what is needed to assess the model output uncertainty.

It is, of course, not realistic to generate 500 to 1000 or more Monte Carlo samples of output for a complex computational model like the EFDC because of the high computational time requirements. However, there are several uncertainty analysis methods that have a long history of successful application to water quality models that require far fewer model runs to obtain estimates of output uncertainty. For example, first-order uncertainty analysis (FOUA) was first applied in water-quality modeling by Burges and Lettenmaier (1975) for a Streeter-Phelps model of DO in a hypothetical stream. Its use quickly expanded to applications to much more complex water quality models, for example, Scavia et al. (1981) used it to identify the key uncertainty sources in the Saginaw Bay eutrophication model. By the late 1980s, it was recognized as a useful technique by the U.S. EPA with its inclusion in the uncertainty analysis routines (UNCAS) designed to work with QUAL2E (Brown and Barnwell, 1987). Melching and Bauwens (2000, 2001) showed that FOUA could identify key sources of uncertainty and measures of output uncertainty for a cascade of models used to compute DO in the Zenne River in Brussels, Belgium. The DO simulation involved the coupling of a nonpoint pollution load model (to estimate combined sewer flows and loads to treatment plants and directly to the river as overflows), a constant treatment efficiency model, and a river water-quality model (unsteady flow on an hourly time step) applied over a one year modeling period. The combination of models involved 53 uncertain parameters/inputs. Missaghi et al. (2013) applied FOUA to a complex three-dimensional (3-D) lake water quality model (ELCOM-CAEDYM) applied to Lake Minnetonka in Minnesota. The Estuary and Lake Computer Model (ELCOM) uses hydrodynamic and thermodynamic models to

simulate spatial and temporal variability of water temperature and velocity distributions over a 3-D computational grid applied to the lake. The Computational Aquatic Ecosystem Dynamics Model (CAEDYM) simulates biogeochemical and chemical water quality variables in lakes in 3-D. Thus, the study of Missaghi et al. (2013) demonstrates that FOUA can be applied to models similar to the EFDC application to Lake Tenkiller. The use of FOUA requires n+1 simulations, where n is the number of uncertain model parameters and inputs.

Another method that yields useful uncertainty analysis information with a relatively small computational burden is the Latin Hypercube Sampling (LHS) technique. LHS (McKay et al., 1979; McKay, 1988) is a stratified sampling approach that allows efficient estimation of the statistics of output. McKay (1988) suggested that the use of twice the number of involved uncertain variables for sample size ($M \geq 2n$) would be sufficient to yield accurate estimation of the statistics model output. Iman and Helton (1985) indicated that a choice of M equal to 4/3n usually gives satisfactory results. For a dynamic stream water-quality model over a 1-year simulation period, Manache and Melching (2007) compared results from LHS using M = 4/3n and M = 3n and found reasonable convergence in the identification of the most sensitive parameters but not in the calculation of the standard deviation of model output. Thus, if it is computationally feasible, the generation of larger number of samples would further enhance the accuracy of the estimation. The LHS technique has been widely applied to water quality modeling in the past (Jaffe and Ferrara, 1984; Melching and Bauwens, 2001; Sohrabi et al., 2003; Manache and Melching, 2004). The foregoing rules on the number of LHS samples are appropriate for complex models with many parameters. For a simpler model involving 4 uncertain parameters, Melching (1995) found that 50 LHS samples could yield reliable estimates of model output uncertainty.

Another means to simplify an uncertainty analysis, is to view spatially varying parameters as standardized variables as proposed by Melching and Bauwens (2000, 2001). In their study, eight of the parameters of the watershed pollutant-load model varied among the different sub-basins. If each of these eight parameters was considered independent in the uncertainty analysis, >200 variables would need to be considered and the uncertainty analysis would have become computationally prohibitive. Therefore, these eight parameters were considered as standardized variables in the uncertainty analysis. A standardized variable, Z_i, is computed as

$$Z_i = \frac{x_i - \overline{x_i}}{\sigma_i}$$

where x_i is the original variable value in subbasin i, $\overline{x_i}$ is the mean of variable x_i, and σ_i is the standard deviation of variable x_i. In this way, the parameters for each subbasin can have a mean and variance appropriate for that subbasin in the uncertainty analysis. This approach is similar to the ratio to the mean approach applied by Baker et al. (2015a) for the spatially varying parameters.

The sensitivity analysis of the EFDC model applied to Lake Tenkiller is incomplete because it considered only four parameters. Specifically, Baker et al. (2015a, p. 20) stated:

> "Based on the experience gained from numerous model runs during the calibration task, kinetic coefficients and model input parameters that significantly influenced the model results include the maximum algae growth rate, phosphorus half saturation constant for nutrient uptake by algae, and PO4 sorption enhancement factor for the sediment flux of phosphate from the sediment diagenesis model along with the watershed total phosphorus (TP) loads."

A more complete sensitivity analysis should be done for the EFDC model inputs and parameters. For example, for a model of similar complexity (ELCOM-CAEDYM), Missaghi et al. (2013) considered 29 model parameters in the sensitivity and uncertainty analyses. Further, if the FOUA or LHS uncertainty analysis methods are applied to the EFDC model of Lake Tenkiller, the results of these methods could be used directly in the sensitivity analysis of the EFDC model. If FOUA is applied to the EFDC model, the sensitivity analysis can be done as explained in Missaghi et al. (2013). If LHS is applied to the EFDC model, the sensitivity analysis can be done as explained in Melching and Bauwens (2001) or Manache and Melching (2004, 2008).

Also, in the uncertainty analysis of the EFDC model for Lake Tenkiller, the results of the Monte Carlo Simulation (MCS) based uncertainty analysis of the HSPF model of the IRW could be directly used in the evaluation of the uncertainty in the TP load to Lake Tenkiller. The MCS results could allow a mean, standard deviation, and an approximate probability distribution to be determined for the TP load to the lake for each day of the year. The resulting mean and standard deviation could be directly used in a FOUA. If LHS were applied for the uncertainty analysis, the mean, standard deviation, and probability distribution could be used to generate the uncertain daily TP loads using the standardized variable or mean ratio approaches (discussed earlier) to ensure each day has a similar deviation from its mean TP load for each individual Latin Hypercube sample.

The more complete and appropriate sensitivity and uncertainty analyses of the application of the EFDC to Lake Tenkiller will provide substantial insight for the development of TMDLs in the IRW.

Issues Not Impacting Certification

Response #11. The reviewer makes some excellent points and the Baker Team appreciates the brief literature review to make us aware of some recent studies of water quality model sensitivity and uncertainty analyses.

Sensitivity Analysis. The reviewer noted that the sensitivity analysis was incomplete because of the limited number of parameters evaluated (i.e., n=4). The reviewer was also critical of the ±25%, ±50% approach used to assign low and high values of the parameters selected for the sensitivity and uncertainty analyses. The Baker Team believes that our approach is a valid statistical expression of the Point Estimate Method originally developed by Rosenblueth (1981) and subsequently modified and applied by Harr (1989), Li (1992), and Christian and Baecher (1999). In the Point Estimate Method, three values -- low, middle and high-- of the perturbed parameter are required. The three values, usually taken to be the mean and ± 1σ or ± 2σ, for each input variable, are used to construct a pseudo-PDF from model outputs by joint probability calculations. The low and high value can be based on the middle value ± some percentage or the low and high values can be based on statistics for the model parameter (mean ± 1σ; mean ± 2σ). In the application of the Point Estimate Method to the Lake Tenkiller model, the Baker Team chose to use ±25%, ±50% of the model calibration parameter values to assign low and high parameter values around the middle calibrated parameter values.

In their sensitivity and uncertainty analysis for a lake model, Missaghi et al. (2013) identified the highest ranked kinetic parameters that contributed to most of the variance of the total lake model output uncertainty. Two of the three kinetic parameters selected for the sensitivity analysis by the Baker Team (benthic phosphate flux and half-saturation constant for phosphorus) were identified by Missaghi et al. as the highest ranked kinetic parameters for their lake model. Despite the limited number of model parameters selected for the lake model sensitivity analysis, the Baker Team did, in fact, choose parameters shown by Missaghi et al. to be very important model parameters for their lake model.

The Baker Team will not be performing any additional model runs to increase the number of model parameters evaluated for the lake model sensitivity analysis.

Uncertainty Analysis. The reviewer stated that the uncertainty analysis for the lake model was inadequate and inappropriate. The Baker Team thinks that there may be some confusion on the reviewer's part about the differences between the related concepts of uncertainty and sensitivity of a model. The Baker Team is basing our model prediction uncertainty for conditions similar to those of the calibration-validation period of 2005-2006 on the agreement between observed and modeled values of key water quality state variables.

Based on further discussions with EPA and other stakeholders, the Baker Team may update the uncertainty analysis for the lake model at some later date by providing joint probability-based calculations of model uncertainty based on the calibration and validation model performance statistics of both the HSPF watershed and EFDC lake models. Uncertainty of the HSPF watershed model would be included in an updated

uncertainty analysis of the lake model since errors in HSPF inflows and loads can propagate through the EFDC model calculations. If the Baker Team does update the Uncertainty Analysis, model joint probability error estimates would be developed for EFDC TSS, nutrients, dissolved oxygen and chlorophyll variables together with the correlated input variables from the HSPF watershed model. Since the Lake Tenkiller reservoir exhibits characteristic spatial gradients within riverine, transition and lacustrine zones, individual uncertainty estimates would be developed for stations within each of these spatial zones of the lake.

References cited in Response#11

Christian, J.T. and G.B. Baecher (1999) Point-estimate method as numerical quadrature. Jour. GeoTech. & GeoEnviron. Eng'r, ASCE, 125(9):779-786.

Harr, M.E. (1989) Probabilistic estimates for multivariate analyses. Appl. Math. Modelling, 13(5):313-318.

Li, K.S. (1992) Point-estimate method for calculating statistical moments. Jour. Eng'r Mech., ASCE, 118(7):1506-1511.

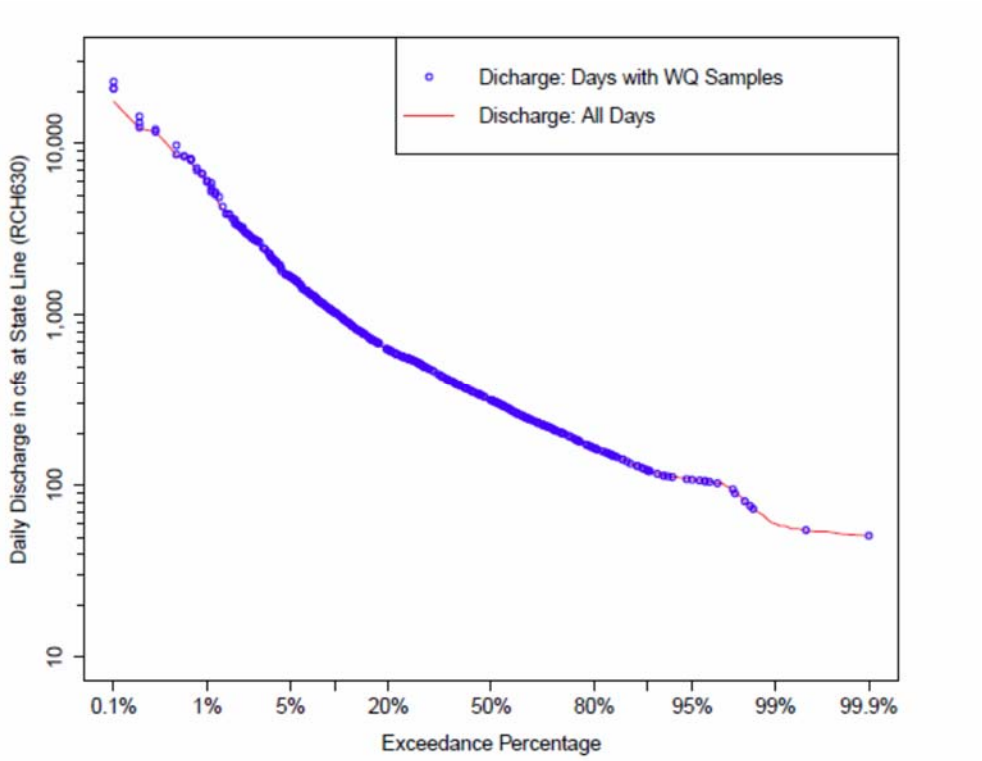
Missaghi, S., Hondzo, M., and Melching, C.S., (2013). Three-dimensional lake water quality modeling: Sensitivity and uncertainty analyses, Journal of Environmental Quality, 42, 1684-1698.

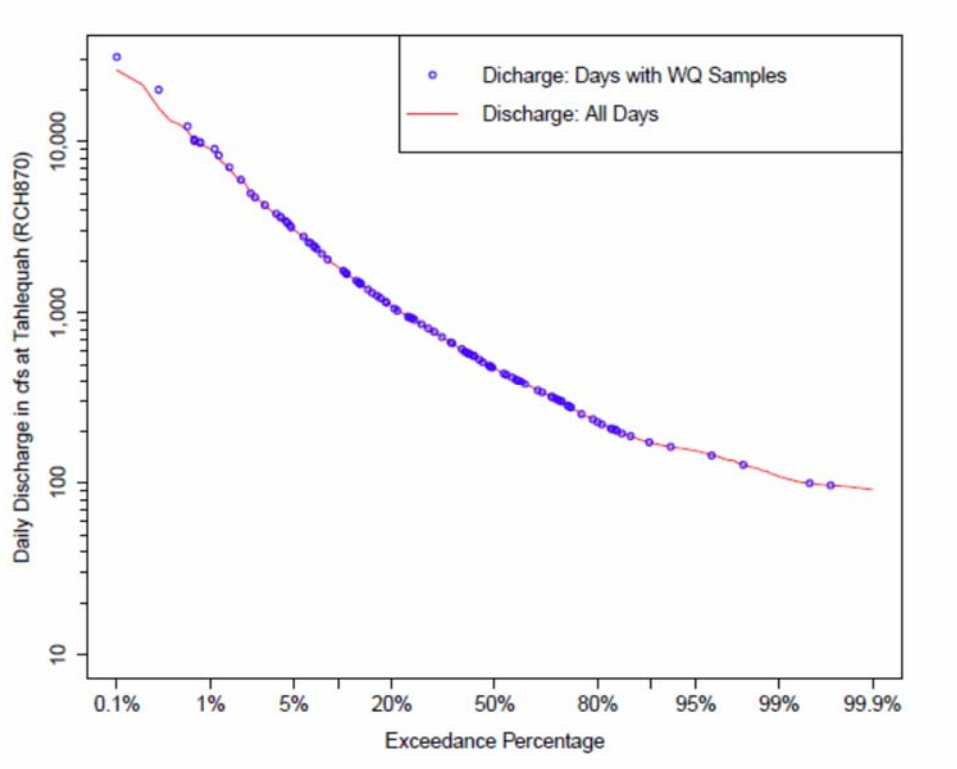
Rosenblueth, E. (1981) Two-point estimates in probabilities. Appl. Math. Modelling, 5(2):329-335.

Comment #12.

A. Baker et al. (2015b, p.64) state that lack of adequate sediment data, especially during storm events, is one of the primary difficulties in attaining a reliable calibration of the HSPF sediment yield simulations. Similar issues regarding lack of storm (or high flow) data for other chemical constituents for confirming the simulation accuracy for HSPF are mentioned at other locations in the various reports reviewed. One of the ways to confirm that the validation data are suitable (i.e. representative of the full range of flow conditions) for calibration of the water quality routines in HSPF is to compare the flow duration curve for all days at a nearby streamflow gage and a flow duration curve developed just considering the days on which a water quality measurement was made. If these two flow duration curves agree well, the measured water-quality constituent concentrations are representative of the range of flows likely at that location. Figure 3 compares the flow duration curve for all days at Reedy Creek near Loughman, FL, to the flow duration curve for days when total phosphorus samples were collected. From this figure it can be seen that the data available for calibration of total phosphorus simulation in HSPF is representative of the full range of flows at this site. Similar curves could be developed for the key water-quality calibration locations in the IRW and added to the model reports to illustrate the representativeness of the available water quality data.

Response #12. The flow duration (frequency) curves like those described by the reviewer for Reedy Creek in FL are shown below, first for the Stateline (Reach 630) and then Tahlequah (Reach 870). These curves show the standard flow-duration curve with the days on which samples were taken (as blue circles). This allows an assessment of whether the observations provide a reasonable representation of the full range of flows at the site. For the Stateline, 443 data points were available during the 9-year calibration period; the first figure shows a dense distribution of flows from less than100 cfs to about 8,000 cfs, with a few extreme low flow samples and a number of high flow samples. For the high flows above 5,000 cfs (~ 1% flow exceedance) there are 20+ samples (blue circles above 5,000 cfs), out of about 30-35 days with flows greater than 5,000 cfs. This is a good representation of high flow periods. However, for Tahlequah (second figure), there are only about 5 samples (blue circles) above the 1% flow of about 8,000 cfs. Thus, the available data at Tahlequah (amounting to 109 samples over 9 years) is much less representative of the full range of flows, and the density of the samples is also much less than at the Stateline, for the full range of flows. This indicates that the available data is less representative of all storm events than at the Stateline.





Comment #13.

B. Table 2-1 of Baker et al. (2015b) lists the meteorological stations in and near the IRW used in the HSPF simulation and their periods of record. The final model report should describe how Stilwell 5NNW and Lyons 2N (both discontinued 4/30/03) were extended to 2009. Similarly, how the record at the Mesonet stations was extended to 1992 and 1993 should be discussed in the final model report. Similarly, how the meteorological data series listed in Table 2-2 in Baker et al. (2015b) were extrapolated to the full calibration and validation time periods should be described in the final model report.

Response #13. In all cases those extensions were based on the closest met/precipitation station with a suitable record for the missing period. This is not often included in normal model documentation, but it has been added as Appendix G to Baker et al., (2015) – Summary of Precipitation and Meteorologic Data Development – in response to this Peer Review comment.

Comment #14.

C. With respect to the land cover/land use data used for the validation period Baker et al. (2013, p. 55) states:

> "We chose the 2001 coverage over the 1992 coverage due to the inconsistencies in classification noted in Section 3.3. Although the 2001 NLCD land use coverage is just outside the validation period, it is still expected to provide a good representation of conditions for the 1992-2000 time period."

This assumption conflicts with the statement in Baker et al. (2015b, p. 4) regarding the rapid population growth in Benton and Washington counties in Arkansas, namely: "the population of Benton and Washington Counties

increased by 45% between 1990 and 2000." Thus, the use of the 2001 coverage will overstate the developed land cover in these counties. The effects of this overstatement should be discussed in the final modeling report.

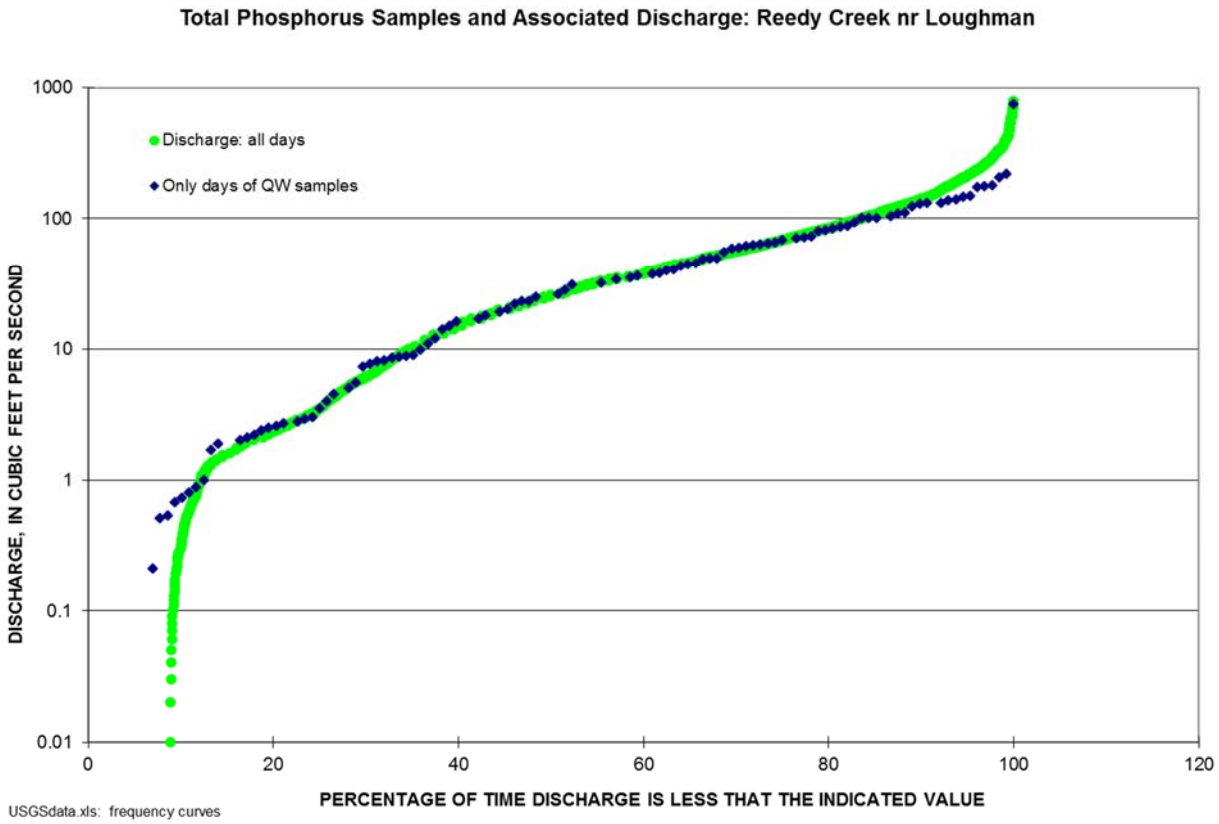


Figure 8. Flow duration curves for all days and days when total phosphorus samples were taken on Reedy Creek near Loughman, Florida.

Response #14. The 2001 NLCD coverage was used in this effort and is believed to be a reasonable representation of the actual land use for 2001, and was used for the validation period from 1992 - 2000. The general statement that there was a 45% increase from1990 to 2000 is simply a reflection of the fact that the census only occurred in those years, not that the increased development was uniform over that period. I believe our assumption was appropriate given that the alternative was to use the 1992 coverage whose categories were deemed to be somewhat inconsistent with the 2001 coverage.

References Cited

Michael Baker, Jr., Inc. 2013. Quality assurance project plan: Modeling QAPP, Illinois River watershed nutrient modeling development, Report prepared for U.S. EPA Region 6, Dallas, TX, 81 p.

Michael Baker, Jr., Inc., Aqua Terra Consultants, and Dynamic Solutions, LLC. 2013. Simulation plan for Illinois River watershed nutrient modeling development, Report prepared for U.S. EPA Region 6, Dallas, TX, 101 p.

Michael Baker, Jr., Inc., Aqua Terra Consultants, and Dynamic Solutions, LLC. 2015a. Procedures and methodologies for the sensitivity and uncertainty analyses for Illinois River watershed and Tenkiller Ferry Lake, Oklahoma, Report prepared for U.S. EPA Region 6, Dallas, TX, 108 p.

Michael Baker, Jr., Inc., Aqua Terra Consultants, and Dynamic Solutions, LLC. 2015b. Setup, calibration, and validation for Illinois River Watershed Nutrient Model and Tenkiller Ferry Lake EFDC water quality model, Report prepared for U.S. EPA Region 6, Dallas, TX, 351 p.

Brown, L.C. and Barnwell, T.O., Jr. 1987. The enhanced stream water quality models QUAL2E and QUAL2E-UNCAS: Documentation and user manual, *Report EPA/600/3-87/007*, U.S. Environmental Protection Agency, Athens, GA.

Burges, S.J. and Lettenmaier, D.P. 1975. Probabilistic methods in stream quality management, *Water Resources Bulletin*, **11**(1): 115-130.

Donigian, A.S., Jr. 2000. *HSPF Training Workshop Handbook and CD*, Lecture #19, Calibration and verification issues, Slide #L19-22, U.S. EPA Headquaters, Washington Information Center, 10-14 January 2000. Presented and prepared for the U.S. EPA Office of Water, Office of Science & Technology, Washington, DC.

Donigian, A.S., Jr. and Imhoff, J.C. 2011. Model selection for the Illinois River in AR/OK – Final Memorandum, prepared for U.S. EPA Region 6, Dallas, TX, 36 p.

Donigian, A.S., Jr., Imhoff, J.C., Bicknell, B.R., and Kittle, J.L., Jr., 1984, Application guide for hydrological simulation program-FORTRAN (HSPF), *EPA-600/3-84-065*, Environmental Research, 177 p.

Duncker, J.J. and Melching, C.S. 1998, Regional rainfall-runoff relations for simulation of streamflow for watersheds in Du Page County, Illinois, *U.S. Geological Survey Water-Resources Investigations Report 98-4035*, 80 p.

Iman, R.L. and Helton, J.C. 1985. A Comparison of Uncertainty and Sensitivity Analysis Techniques for Computer Models, *Report No. NUREGICR-3904, SAND 84-1461*, Sandia National Laboratories, Albuquerque, NM.

Jaffe, P.R. and Ferrara, R.A. 1984. Modeling sediment and water column interactions for hydrophobic pollutants, parameter discrimination and model response to input uncertainty, *Water Research*, **18**(9): 1169-1174.

James, L.D. and Burges, S.J. 1982, Selection, calibration and testing of hydrologic models, In: *Hydrologic Modeling of Small Watersheds*, Haan, C.T., Johnson, H.P., and Brakensiek, D.L., eds., American Society of Agricultural Engineers, St. Joseph, Mich., p. 437-472.

Linsley, R.K., Jr., Kohler, M.A., and Paulhus, J.L.H., 1982, *Hydrology for Engineers,* McGraw-Hill, New York, 508 p.

Manache, G. and Melching, C.S. 2004. Sensitivity analysis of a water-quality model using Latin hypercube sampling, *Journal of Water Resources Planning and Management*, **130**(3): 232-242.

Manache, G. and Melching, C.S., 2007. Sensitivity of Latin hypercube sampling to sample size and distributional assumptions, In: *Proceedings (CD-ROM)*, 32nd Congress of the International Association of Hydraulic Engineering and Research, Venice, Italy, July 1-6, 2007.

Manache, G. and Melching, C.S., 2008. Identification of reliable regression- and correlation-based sensitivity measures for importance ranking of water-quality model parameters," *Environmental Modelling & Software*, **23**(5), 549-562.

McKay, M.D. 1988. Chapter 4: Sensitivity and uncertainty analysis using a statistical sample of input values. In: *Uncertainty Analysis*, edited by Y. Ronen, CRC press, Inc., Boca Raton, FL.

McKay, M.D., Beckman, R.J., and Conover, W.J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, **21**(2), 239-245.

Melching, C. S. 1995. Reliability estimation, Chapter 3 in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, Water Resources Publications, Littleton, CO, p. 69-118.

Melching, C. S. and Bauwens, W. 2000. Comparison of uncertainty-analysis methods applied to simulation of urban water quality, In: *Stochastic Hydraulics 2000*, edited by Z.Y. Wang and S.X. Hu, A.A. Balkema, Rotterdam, The Netherlands, p. 717-725.

Melching, C.S. and Bauwens, W. 2001. Uncertainty in coupled non-point source and stream water-quality models, *Journal of Water Resources Planning and Management*, **127**(6), 403-413.

Missaghi, S., Hondzo, M., and Melching, C.S., 2013. Three-dimensional lake water quality modeling: Sensitivity and uncertainty analyses, *Journal of Environmental Quality*, **42**, 1684-1698.

Nash, J.E. and Sutcliffe J.V., 1970, River flow forecasting through conceptual models, Part 1- A discussion of principles, *Journal of Hydrology*, **10**, 282-290.

Scavia, D., Powers, W.F., Canale, R.P., and Moody, J.L. 1981. Comparison of first-order error analysis and Monte Carlo simulation in time-dependent lake eutrophication models, *Water Resources Research*, **17**(4), 1051-1059.

Sohrabi, T.M., Shirmohammadi, A., Chu, T.W., Montas, H., and Nejadhashemi, A.P. 2003. Uncertainty analysis of hydrologic and water quality predictions for a small watershed using SWAT2000, *Environmental Forensics*, **4**(4), 229-238.

U.S. Geological Survey (USGS) and Aqua Terra Consultants. 1998. Quality assurance project plan for rainfall-runoff simulation using the Hydrological Simulation Program-Fortran (HSPF) for the proposed Crandon Mine area, Crandon, Wisconsin, Report prepared for U.S. EPA Region 5, Chicago, IL, 50 p.

_____

Comments by Peer Reviewer #2

Comment #15.

This letter documents my peer review of the Illinois River watershed and lake models in response to your request for a peer review as described in your memo of March 6, 2015 “Introduction, Summary Report and Peer Review Tasks.” My review focused primarily on the document “Setup, Calibration, and Validation for Illinois River Watershed Nutrient Model and Tenkiller Ferry Lake EFDC Water Quality Model,” dated February 13, 2015. Appendices A through F were provided to me separately.

The objective of this modeling effort, as I understand it, is to develop a model that can lead to scientifically sound numeric TMDLs and a basin-wide water quality restoration plan. The primary water quality issues are that several stream segments in both Arkansas and Oklahoma do not meet water quality criteria for total phosphorus and water quality criteria for dissolved oxygen (DO), chlorophyll-a, and Trophic State Index in Lake Tenkiller are not being satisfied. Related to these challenges are that Oklahoma has a more stringent scenic river in-stream total phosphorus criterion of 0.037 mg/L. The model is intended to inform the efficacy of alternative management and regulatory strategies that might lead to better support of beneficial uses in both States. My comments are as follows.

The models selected for these efforts are well-regarded public domain models: 1) HSPF for the watershed modeling and 2) EFDC for the lake modeling. These are credible choices for achieving the modeling objectives.

Models, of course, are approximations of the real world and are limited not only by the methodologies within the model for simulating the real world, but by the availability of data to calibrate and validate the modeling, the available budget, and the skill and experience of the modeling team.

Overall, the watershed and lake modeling efforts appear to be comprehensive and competently executed. However, there are a few areas where the modeling (or its documentation) could be improved to better fulfill the objectives previously stated for the models. In some of these areas it may be that the work has been performed, but not documented in the reports I have reviewed. Or it may be documented, but in my review I did not come across the relevant explanations. For other areas, additional work could benefit the overall modeling effort.

Response #15. No response needed for this introductory statement.

Comment #16.

1. Use of NEXRAD precipitation data.

   NEXRAD precipitation data was used extensively in the Arkansas portion of the watershed with little explanation of how the data were ground-truthed. The identification of "phantom storms" was partially, but not wholly explained or corrected. The NEXRAD data were also not compared to the physical gages in the vicinity, for example, NEXRAD virtual stations 20 and 23 are adjacent to the Fayetteville BASINS station. It is not clear that use of these data was a net benefit for the model compared with using the more sparse physical gaging network.

Response #16. The NEXRAD data were provided by Dr Saraswat from UA which he processed and used in this SWAT model application. The extreme data values with the other NCDC stations, both in AR and OK, were compared and it was noticed that the largest single day rainfall recorded was 8.8 inches, whereas the NEXRAD data showed numerous instances of 10-15 inches and up to 22 inches in a single day (as discussed in the report, page 14) The use of NEXRAD was never intended when the project was started, but stakeholders and reviewers (see comments by Matlock) indicated that they felt the NEXRAD data was the best available and should be used; subsequently EPA agreed with this and requested to use NEXRAD data. In support of this, there were only 3 NCDC stations in the AR portion of the watershed, so there definitely was some benefit to use of that data to better define spatial variations. As noted in Comments #3 and #4, the overall hydrology calibration is very good in spite of the selected issues uncovered with the NEXRAD data.

Comment #17.

2. Precipitation/streamflow in the watershed model.

   The hydrology of the watershed model is a significant driver for the water quality analysis in the watershed and the hydrodynamics and water quality analysis in the lake model. With statistics summarized in Table 2-17, the report draws the conclusion that the hydrologic model results "show a Fair to Good overall calibration and validation, and in some cases (i.e. sites) a Very Good simulation, confirming that the overall model should provide a sound basis for subsequent water quality simulations." The table reports that overall annual flow volume errors at the calibration sites for the calibration data average 2.87 percent. However, table 2-18 (Illinois River south of Siloam Springs) and Table 2-19 (Illinois River near Tahlequah) shows a different picture with the average absolute annual volume error being 16 and 13 percent, respectively.

   Water year 2006 presents the greatest error at both locations at 62 and 54 percent, respectively. (It is assumed that when not explicitly stated in the report that a year designation is the water year.) In

addition to the large errors, no explanation is given for the apparent inconsistency between the precipitation and observed flow volume for this year. At both locations the precipitation is significantly higher than in 2005, but the observed flow is significantly lower. The report states that this is the second year of a drought, but this is not clear because precipitation is above normal for 2006.

Model credibility would be enhanced if annual flow volumes would match observed flow volumes more closely.

Response #17.

- Yes, the WY (Water Year), from Oct 1 through Sept 30, was used throughout the study; page 48 clearly shows the WYs assigned for calibration and validation. If desired, the labels can be revised accordingly in selected tables.
- Please refer to Comment # 3 where the reviewer discusses the year-to-year variations not being an accurate overall assessment of the model results and performance.

Comment #18.

3. Storm event/low flow calibration/validation

   Related to achieving a good hydrologic balance in the watershed is the examination of high flow events and low flow periods. (It is these parts of the flow duration curves where observed and simulated values tend to diverge.) Examining such events also informs the specification of infiltration and evapotranspiration parameters. Analysis of such events may also reveal insights into annual flow balance issues.

Response #18. A Weight-of-Evidence approach to calibration addresses all flow regimes, since significant emphasis was placed on matching the flow duration curves while calibrating to daily flow time series. All flow levels are addressed, and water balance components were evaluated as part of these procedures.

Comment #19.

4. Tenkiller Ferry Lake water level calibration/validation lake levels

   The hydrologic/hydrodynamic calibration and validation was only performed on two years: 2005 and 2006. As reported, the observed and simulated water levels matched well, but this short period for comparison is silent on the ability of the model to match overall hydraulic residence time. Water levels should be evaluated for the full period of calibration and validation of the watershed model (water years 1992 through 2009) assuming these observed values are available. This will also address the question of how the reservoir model performs outside of the identified 2005-2006 “drought period.”

Response #19. Reviewer #2 notes that the lake model report "… is silent on the ability of the model to match overall hydraulic residence time". In the revision of the report, the Baker Team will present estimates of observed hydraulic residence time for 2005-2006 for the lake based on estimates of annual average observed inflow and water level (lake volume) data. The observed estimates of hydraulic residence time for 2005 and 2006 will be compared to the model estimates developed from simulation data for HSPF inflow and EFDC water level (lake volume).

Reviewer #2 suggests that water levels in the lake model should be evaluated for the full period of calibration and validation of the watershed model from 1992-2009. Although the HSPF watershed model was setup for the years from 1992-2009, the EFDC lake model was setup, calibrated and validated only to water level and water quality observations collected during 2005-2006. The Baker Team used all available station data for the lake model calibration and validation years of 2005-2006. OWRB BUMP reports show that data was collected in Lake Tenkiller at quarterly intervals from 2001-2002; 2003-2004; 2005-2006; and in 2011-2012.

Comment #20.

5. Tenkiller Ferry Lake Flow Balancing.

   In the Boundary Conditions section (3.6), the report states that a flow boundary "to define a flow balance to account for water removed from the lake by water supply and other unaccounted flows such as leakage from the dam" was incorporated in the EFDC model. The report further states that the water supply withdrawals were not available; therefore "a flow balance was computed to ensure that the EFDC model simulated lake stage matched the observed lake stage."

   If I am interpreting this correctly, an unknown value – water supply withdrawals – was calculated to match observed; it is no surprise, then, that the calibration matched so well. If this interpretation is accurate, the model calibration and validation for lake level is not meaningful.

   In addition, I did not find any data summarizing the hydrologic balance on the lake for the calibration or validation period. This seems important considering that 2006 was the worst year for the HSPF watershed model in terms of matching observed flow volume.

   Finally, no data were presented on the estimated values for the water supply withdrawals/unaccounted losses computed for 2005 and 2006, nor was there any assessment about whether these values are reasonable. It is also unclear what would be assumed for these values when applying the lake model to a longer period as would be needed to evaluate TMDLs or other management actions.

Response #20. The methodology used to develop a flow balance to ensure that the water level, surface area and volume of the simulated lake is accurate and is an accepted practice for developing hydrodynamic models of reservoirs. The flow balance is needed to account for known variance of flow data measured by the USGS, unknown inflows, and unknown outflows such as leakage identified by the US Army Corp of Engineers at the dam and unknown water withdrawals for local water supply systems served by Lake Tenkiller. On a daily average basis, the volume accounted for by the flow balance represented less than 0.1%, of the average volume of the lake during 2005-2006.

Reviewer #2 notes that "It is also unclear what would be assumed for these values when applying the lake model to a longer period as would be needed to evaluate TMDLs or other management actions". The lake model will be applied for the TMDL determinations using the data that has been setup to calibrate and validate the model for the 2 year period from 2005-2006. The lake model will not be applied for a longer period for evaluation of the TMDLs.

Comment #21.

6. Instream total phosphorus data calibration/validation not clearly reported.

   Because total phosphorus is problematic in the Illinois River and many of its tributaries, as well as Tenkiller Ferry Lake, it is critical to have confidence in the simulations. As reported in Table 2-27, the modeling of total nitrogen is better than that for total phosphorus, with the watershed model exhibiting a tendency to overestimate phosphorus loading. The report does not address why this is or if adjustments could be made to the model. It also does not address whether the model statistically produces similar violation frequency of phosphorus standards. It may be helpful in understanding the overall utility of the model for assessing instream phosphorus by summarizing the success – or lack thereof – of previous watershed modeling efforts related to phosphorus.

Response #21.

- Table 2-27 compares results from 2 models, HSPF vs LOADEST. This is not a model-data comparison. For the calibration effort a greater reliance and emphasis was placed on the actual model-data comparisons. LOADEST was not used for calibrations, but HSPF predicted values were compared against them only for a supplementary comparison.
- The available data at a number of sites is primarily grab sample values during mostly low, or base flow conditions, especially at Tahlequah. Strict calibration to those values would underestimate the loads since they mostly ignore storm flow conditions that transport and deliver the majority of the load to Lake Tenkiller.

Comment #22.

7. EFDC modeling of TP, DO, and Chl-a.

   The figures in appendices I and J for total phosphorus, dissolved oxygen, and chl a (the parameters generally in violation of water quality criteria) show general agreement with observed data in many cases. However, there seems to be a lack of validation data during storm events when the simulations indicate a rapid value change. The report mentions OWRB stations (1994 – 2012), but very little discussion of these data was found. In addition, if these data are outside of the selected 2005 and 2006 calibration/validation years, but within the 1992 to 2009 years for the HSPF runs, they still may be useful for comparison.

Response #22. The reviewer notes that there is a lack of validation data during storm events when simulated lake concentrations spiked. The Baker Team used all available station data for the model calibration and validation years of 2005-2006. OWRB BUMP reports indicate that data was collected in Lake Tenkiller at quarterly intervals from 2001-2002; 2003-2004; 2005-2006; and 2011-2012. It is not clear to EPA and the Baker Team how water quality observations collected in years other than 2005-2006 can help to interpret the 2005-2006 lake model response to storm event conditions that occurred in other years if water quality observations happened to be collected during a storm event in Lake Tenkiller.

Comment #23.

8. Validation data from other lake modeling efforts.

   The potential availability or utility of water quality validation data from other lake modeling efforts should be discussed to enhance the credibility of the current modeling effort.

Response #23. The EPA and the Baker Team does not agree that water quality data from other lake modeling efforts would contribute to an evaluation of model performance for calibration and validation of the EFDC model of Lake Tenkiller.