Evaluation and Performance Assessment of Watershed Models

A.S. Donigian, Jr. and J.C. Imhoff
AQUA TERRA Consultants
2685 Marine Way, Suite 1314
Mountain View, CA 94043

ABSTRACT

Model performance assessment, also known as skill assessment, is a necessary and critical element in the application of watershed models for water resource and water quality management. A comprehensive assessment of model performance includes:

1. Consideration of how well a model is able to simulate observed data that describe the watershed's hydrologic and water quality response to its forcing functions (e.g., meteorology, land disturbance activities, point and nonpoint source loadings),
2. Measuring the relative sensitivity of model output to various model parameters in the specific setting in which the model is being applied, and
3. Assessing the potential uncertainty that is introduced into model output as a result of naturally occurring variability in the actual values of model parameters.

By means of model calibration and validation the modeler gains understanding of the response behavior of the watershed that is being modeled and the ability of the parameterized model to mimic that response. Sensitivity analysis and uncertainty analysis provide both the modeler and those who use the model results with a means of understanding a model's inherent strengths (or limitations) in accuracy. The combination of these individual assessments determines the level of confidence that model users should invest in a model's ability to represent current and alternative watershed responses.

KEYWORDS

Watershed models, model performance assessment, model calibration, model validation, sensitivity analysis, uncertainty analysis, error analysis, HSPF.

INTRODUCTION

Model performance criteria have been contentious topics for more than 30 years. The issues inherent in measuring performance have recently been thrust to the forefront in the environmental arena as a result of the need for, and use of modeling for exposure/risk assessments, TMDL determinations, and environmental assessments. Despite a lack of consensus on how they should be evaluated, in practice, environmental models are being applied, and their results are being used, for assessment and regulatory purposes.

This paper explores methods of measuring and evaluating model performance by means of model calibration and validation, parameter sensitivity analysis and parameter uncertainty analysis. Model calibration and validation are examined within the context of a 'weight of

evidence' approach which has evolved as a result of more than 25 years experience with the U.S. EPA Hydrological Simulation Program - FORTRAN (HSPF) watershed model (Bicknell et al., 2005). We describe example applications and include model results to demonstrate the graphical and statistical procedures used to assess model performance. In addition, quantitative criteria for various statistical measures are discussed as a basis for evaluating model results and documenting the performance of model applications.

Long term experience with the HSPF model and predecessor models has provided a strong foundation for identifying the most sensitive model parameters for most climatic, edaphic, and physiographic watershed settings. However, sensitivity of model results to parameters varies from watershed to watershed, with relative sensitivity in a given watershed depending on the combined impacts of climate and watershed conditions. In other words, sensitivity for a specific watershed is a function of the specific combination of parameter values that reflect climate and watershed characteristics which control the hydrologic response, along with the sediment and water quality behavior. In this paper we discuss parameter sensitivity analysis in general terms applicable to all watershed models, and we provide techniques and examples specific to HSPF applications to illustrate this discussion.

The computational demands of comprehensive, dynamic watershed models are such that formal and rigorous analyses with numerous iterations of long time-period simulation runs that are traditionally used to assess uncertainty are rare and often not feasible. We describe an approach to addressing uncertainty that utilizes the results of model sensitivity analyses to identify the most sensitive parameters in a particular model application setting, and then uses the results of that effort to help implement a more focused uncertainty investigation. Again, we provide an example approach and results generated by HSPF.

WATERSHED MODEL CALIBRATION AND VALIDATION

Calibration and validation, included in model performance assessment, have been defined by the American Society of Testing and Materials, as follows:

Calibration	a test of the model with known input and output information that is used to adjust or estimate factors for which data are not available.
Validation	comparison of model results with numerical data independently derived from experiments or observations of the environment.

Model calibration is the process of adjusting model inputs within acceptable limits until the resulting predictions give good correlation with observed data. Commonly, calibration begins with the best estimates for model input based on measurements and subsequent data analysis. Results from initial simulations are then used to modify the values of the model input parameters. Models are often calibrated through a subjective trial-and-error adjustment of model input data because a large number of interrelated factors influence model output. However, the experience and judgment of the modeler are a major factor in calibrating a model accurately and efficiently. Further, the model should meet pre-specified quantitative measures of accuracy to establish its acceptability in answering the principal study questions (Tetra Tech, 2009).

Model validation is in reality an extension of the calibration process. Its purpose is to assure that the calibrated model properly assesses all the variables and conditions which can affect model results. While there are several approaches to validating a model, perhaps the most effective procedure is to use only a portion of the available record of observed values for calibration; once the final parameter values are developed through calibration, simulation is performed for the remaining period of observed values and goodness-of-fit between recorded and simulated values is reassessed. However, model credibility is also based on the ability of a single set of parameters to represent the entire range of observed data. Therefore, if a single parameter set can reasonably represent a wide range of events, then this is a form of validation.

Calibration and validation are achieved by considering qualitative *and* quantitative measures, involving both graphical comparisons and statistical tests. For flow simulations where continuous records are available, all these techniques should be employed, and the same comparisons should be performed, during both the calibration and validation phases. Comparisons of values for simulated and observed state variables are often performed for daily, monthly, and annual values, in addition to flow-frequency duration assessments. Statistical procedures often include error statistics, correlation and model-fit efficiency coefficients, and goodness-of-fit tests, as appropriate. For water quality constituents, model performance is often based primarily on visual and graphical presentations as the frequency of observed data is often inadequate for accurate statistical measures.

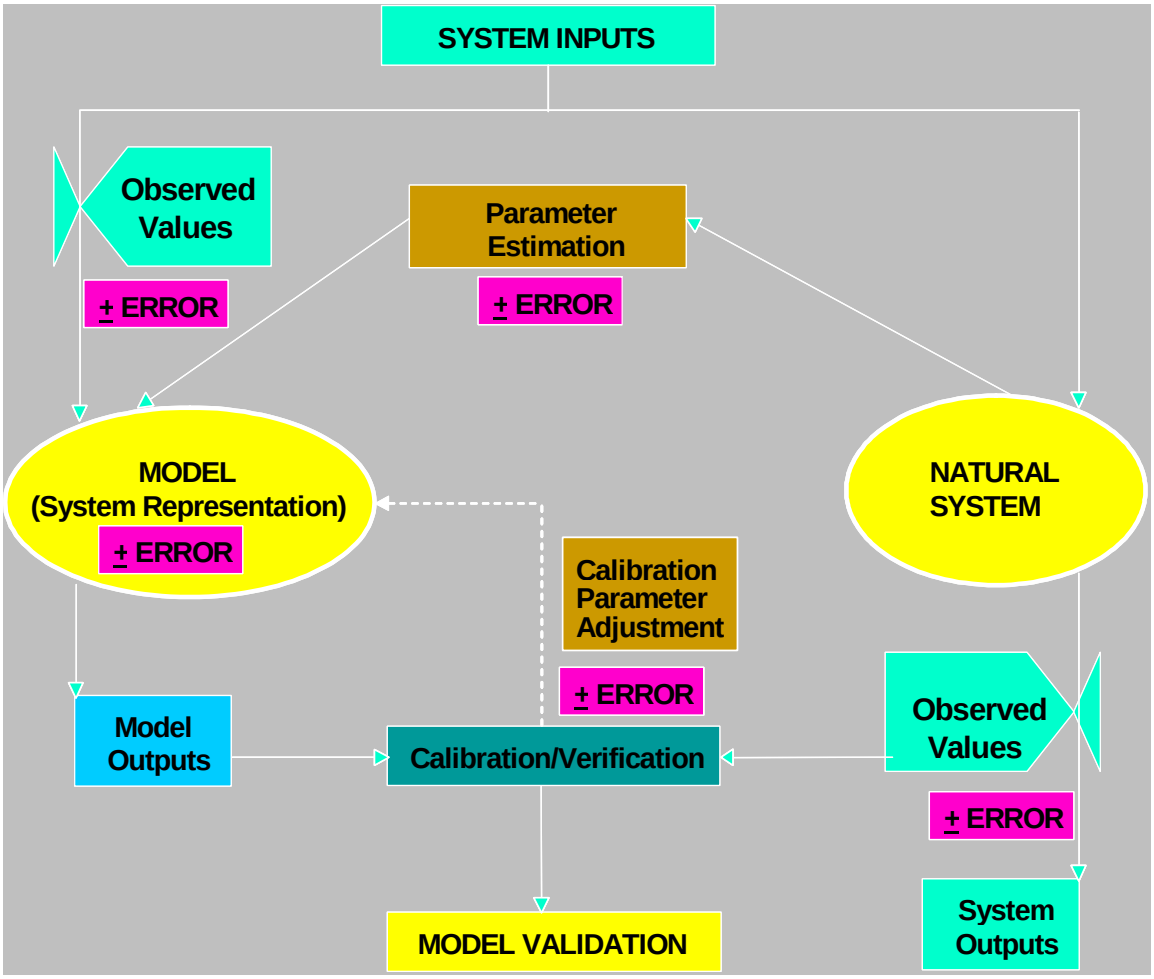
In practice, the model calibration/validation process can be viewed as a systematic analysis of errors or differences between model predictions and field observations. Figure 1 schematically compares the model with the 'natural system', i.e. the watershed, and identifies various sources of potential errors to be investigated. These types of analysis require evaluation of the accuracy and validity of the model input data, parameter values, model algorithms, calibration accuracy, and observed field data used in the calibration/validation. Clearly, the model user becomes a 'detective' in searching for the causes of the errors or differences, and potential remedies to improve the agreement and reduce the errors. A more complete discussion of these error sources is provided in Donigian and Rao (1990).

As noted above, model calibration and validation are necessary and critical steps in any model application. For HSPF, calibration is an iterative procedure of parameter evaluation and refinement, as a result of comparing simulated and observed values of interest. It is required for parameters that cannot be deterministically, and uniquely, evaluated from topographic, climatic, edaphic, or physical/chemical characteristics of the watershed and compounds of interest. Fortunately, the large majority of HSPF parameters do not fall in this category. Calibration is based on several years of simulation (at least 3 to 5 years) in order to evaluate parameters under a variety of climatic, soil moisture, and water quality conditions. Calibration should result in parameter values that produce the best overall agreement between simulated and observed values throughout the calibration period.

Calibration includes the comparison of both monthly and annual values, and individual storm events, whenever sufficient data are available for these comparisons. All of these comparisons should be performed for a proper calibration of hydrology and water quality parameters. In

addition, when a continuous observed record is available, such as for streamflow, simulated and observed values should be analyzed on a frequency basis and their resulting cumulative distributions (e.g. flow duration curves) compared to assess the model behavior and agreement over the full range of observations.

Figure 1 - Model versus Natural System: Inputs, Outputs, and Errors



Calibration is a hierarchical process beginning with hydrology calibration of both runoff and streamflow, followed by sediment erosion and sediment transport calibration, and finally calibration of nonpoint source loading rates and water quality constituents. When modeling land surface processes hydrologic calibration must precede sediment and water quality calibration since runoff is the transport mechanism by which nonpoint pollution occurs. Likewise, adjustments to the instream hydraulics simulation must be completed before instream sediment and water quality transport and processes are calibrated.

Performance Criteria for Calibration and Validation

Performance criteria to measure the success of the calibration and validation efforts have been contentious topics for more than 30 years (see Thomann, 1980; Thomann, 1982; James and

Burges, 1982; Donigian, 1982; ASTM, 1984). Although no **complete** consensus on model performance criteria is apparent from the past and recent model-related literature, a number of 'basic truths' are evident and are likely to be accepted by most modelers in modeling natural systems:

- Models are approximations of reality; they cannot precisely represent natural systems.
- There is no single, accepted statistic or test that determines whether or not a model is validated
- Both graphical comparisons and statistical tests are required in model calibration and validation.
- Models cannot be expected to be more accurate than the errors (confidence intervals) in the input and observed data.

All of these 'basic truths' must be considered in the development of appropriate procedures for model performance and quality assurance of modeling efforts. A **'weight of evidence'** approach is most widely used and accepted when models are examined and judged for acceptance. Simply put, the **weight-of-evidence** approach embodies the above 'truths', and demands that multiple model comparisons, both graphical and statistical, be demonstrated in order to assess model performance, while recognizing inherent errors and uncertainty in both the model, the input data, and the observations used to assess model acceptance.

Although individual watershed models will utilize different types of graphical and statistical procedures, they will generally include a subset of the following:

Graphical Comparisons**:**

1. Timeseries plots of observed and simulated values for fluxes (e.g. flow) or state variables (e.g. stage, sediment concentration, biomass concentration)
2. Observed vs. simulated scatter plots, with a 45^{o} linear regression line displayed, for fluxes or state variables
3. Cumulative frequency distributions of observed and simulated fluxes or state variable (e.g. flow duration curves)

Statistical Tests**:**

1. Error statistics, e.g. mean error, absolute mean error, relative error, relative bias, standard error of estimate, etc.
2. Correlation tests, e.g. linear correlation coefficient, coefficient of model-fit efficiency, etc.
3. Cumulative distribution tests, e.g. Kolmogorov-Smirnov (KS) test

These comparisons and statistical tests are fully documented in a number of comprehensive references on applications of statistical procedures for biological assessment (Zar, 1999), hydrologic modeling (McCuen and Snyder, 1986), and environmental engineering (Berthouex and Brown, 1994).

Time series plots are generally evaluated visually as to the agreement, or lack thereof, between the simulated and observed values. Scatter plots usually include calculation of a correlation

coefficient, along with the slope and intercept of the linear regression line; thus the graphical and statistical assessments are combined. For comparing observed and simulated cumulative frequency distributions (e.g. flow duration curves), the KS test can be used to assess whether the two distributions are different at a selected significance level. Unfortunately, the reliability of the KS test is a direct function of the population of the observed data values that define the observed cumulative distribution. Except for flow comparisons at the major USGS gage sites, there is unlikely to be sufficient observed data (i.e. more than 50 data values per location and constituent) to perform this test reliably for most water quality and biotic constituents. Moreover, the KS test is often quite easy to 'pass', and a visual assessment of the agreement between observed and simulated flow duration curves, over the entire range of high to low flows, may be adequate and even more demanding in many situations.

In recognition of the inherent variability in natural systems and unavoidable errors in field observations, the USGS provides the following characterization of the accuracy of its streamflow records in all its surface water data reports (e.g. Socolow et al., 1997):

Excellent Rating	95 % of daily discharges are within 5 % of the true value
Good Rating	95 % of daily discharges are within 10 % of the true value
Fair Rating	95 % of daily discharges are within 15 % of the true value

Records that do not meet these criteria are rated as 'poor'. Clearly, model results for flow simulations that are within these accuracy tolerances can be considered acceptable calibration and validation results, since these levels of uncertainty are inherent in the observed data.

Table 1 lists general calibration/validation tolerances or targets that have been provided to model users as part of HSPF training workshops over the past 10 years (e.g. Donigian, 2000). The values in the table attempt to provide some general guidance, in terms of the percent mean errors or differences between simulated and observed values, so that users can gage what level of agreement or accuracy (i.e. very good, good, fair) may be expected from the model application.

The caveats at the bottom of the table indicate that the tolerance ranges should be applied to **mean** values, and that individual events or observations may show larger differences, and still be acceptable. In addition, the level of agreement to be expected depends on many site and application-specific conditions, including the data quality, purpose of the study, available resources, and available alternative assessment procedures that could meet the study objectives.

Figure 2 provides value ranges for both correlation coefficients (R) and coefficient of determination (R^2) for assessing model performance for both daily and monthly flows. The figure shows the range of values that may be appropriate for judging how well the model is performing based on the daily and monthly simulation results. As shown, the ranges for daily values are lower to reflect the difficulties in exactly duplicating the timing of flows, given the uncertainties in the timing of model inputs, mainly precipitation.

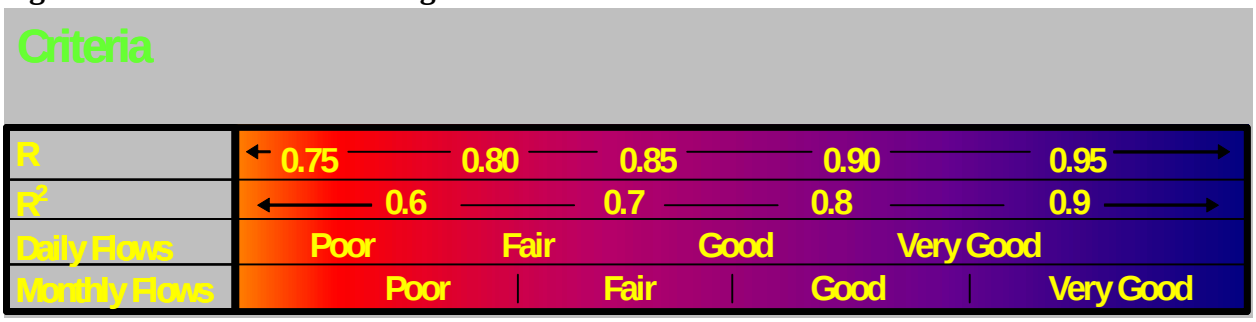
Given the uncertain state-of-the-art in model performance criteria, the inherent errors in input and observed data, and the approximate nature of model formulations, **absolute** criteria for watershed model acceptance or rejection are not generally considered appropriate by most

Table 1- General Calibration/Validation Targets or Tolerances for HSPF Application
Donigian (2000)

	% Difference Between Simulated and Recorded Values		
	Very Good	Good	Fair
Hydrology/Flow	< 10	10 - 15	15 - 25
Sediment	< 20	20 - 30	30 - 45
Water Temperature	< 7	8 - 12	13 - 18
Water Quality/Nutrients	< 15	15 - 25	25 - 35
Pesticides/Toxics	< 20	20 - 30	30 - 40

CAVEATS: Relevant to monthly and annual values; storm peaks may differ more
Quality and detail of input and calibration data
Purpose of model application
Availability of alternative assessment procedures
Resource availability (i.e. time, money, personnel)

Figure 2 - R and R^2 Value Ranges for Model Performance



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 specific study effort, with larger deviations expected for individual sample points in both space and time. The values shown above 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.

EXAMPLE PERFORMANCE ASSESSMENTS FOR CALIBRATION/VALIDATION

This section presents results from HSPF applications, (1) to the State of Connecticut for nutrient loadings to Long Island Sound, and (2) to the Housatonic River, MA for hydrology modeling, as examples of the types of graphical and statistical comparisons recommended for model calibration and validation.

The Connecticut Watershed Model (CTWM)

The Connecticut Watershed Model (CTWM), based on HSPF, was developed to evaluate nutrient sources and loadings within each of six nutrient management zones that lie primarily within the state of Connecticut, and assess their delivery efficiency to Long Island Sound (LIS). The CTWM evolved by first performing calibration and validation on three small test basins across the state (Norwalk, Quinnipiac, and Salmon – see Figure 3) representing a range of land uses, including urban, forest, and agricultural. The model was then extended to three major river calibration basins (Farmington, Housatonic, and Quinebaug) and subsequently expanded to a statewide model by using the most spatially applicable set of calibrated watershed parameters in non-calibrated areas. The user-friendly interface and framework of the CTWM was specifically designed to promote continuing use to assess multiple BMPs, implementation levels, and relative impacts of point source controls for nutrient reductions to LIS. Complete details of the study and the model development and application are provided in the Final Study Report (AQUA TERRA Consultants and HydroQual, 2001). Love and Donigian (2002) summarize the techniques and methods used in the CTWM model development and the "weight-of-evidence" approach used in the calibration and validation, while Donigian and Love (2002) discuss and present model results of alternative growth and BMP (Best Management Practice) Scenarios on nutrient loads to LIS.

The hydrologic calibration for the Test Watersheds and the Major Basins was performed for the time period of 1991-1995 while the period of 1986-1990 was used for validation. The available flow data include continuous flow records at the USGS gage sites shown in Figure 3 for the entire time period. Consistent with the calibration procedures discussed above, comparisons of simulated and observed flow were performed during the calibration and validation periods for daily, monthly, and annual values, as well as flow-frequency duration assessments. In addition, the input and simulated water balance components (e.g., precipitation, runoff, evapotranspiration) were reviewed for the individual land uses.

Calibration of the CTWM was a cyclical process of making parameter changes, running the model and producing the aforementioned comparisons of simulated and observed values, and interpreting the results. This process was greatly facilitated with the use of HSPEXP, an expert system for hydrologic calibration, specifically designed for use with HSPF, developed under contract for the USGS (Lumb et al., 1994). This package gives calibration advice, such as which model parameters to adjust and/or input to check, based on predetermined rules, and allows the user to interactively modify the HSPF Users Control Input (UCI) files, make model runs, examine statistics, and generate a variety of plots. The postprocessing capabilities of GenScn (e.g., listings, plots, statistics, etc.) were also used extensively during the calibration/validation effort.

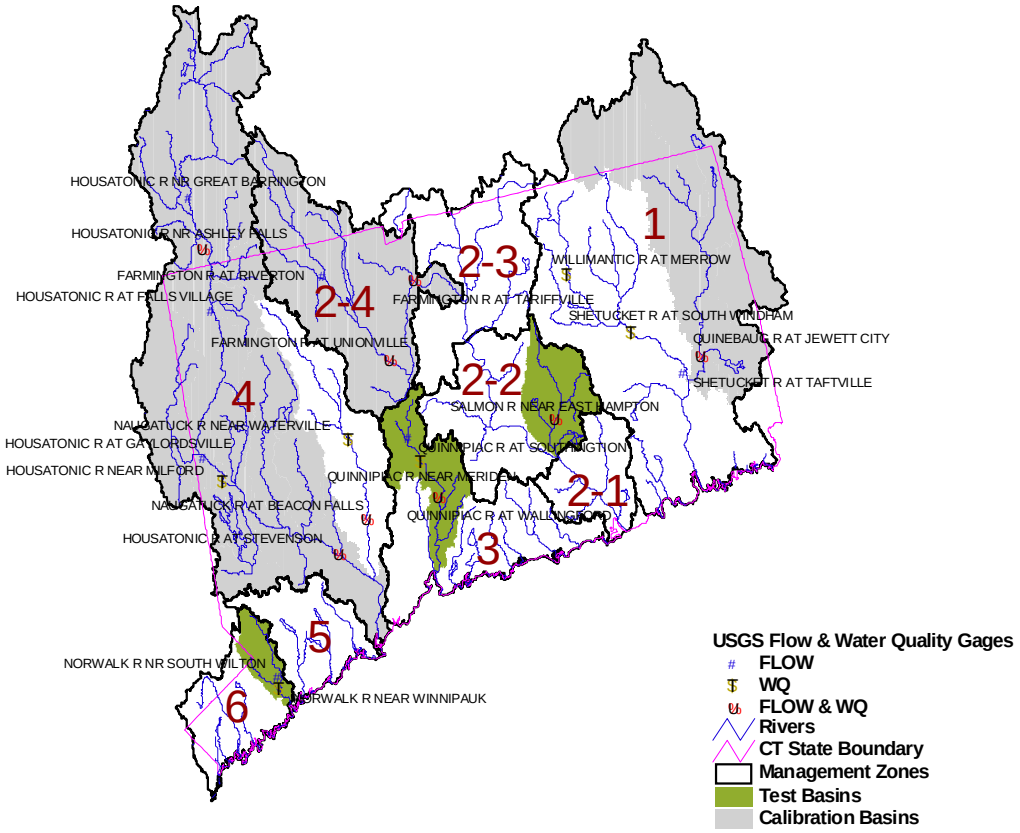


Figure 3 - USGS Flow and Water Quality Gages for the CTWM (AQUA TERRA Consultants and HydroQual (2001)

The hydrology calibration focused primarily on the monthly agreement of simulated and observed values as opposed to individual storm events, due to the greater sensitivity of LIS to long-term versus short-term nutrient loads (HydroQual, 1996).

The time period of the water quality calibration coincided with the hydrology calibration period, i.e. 1991-95. However, sufficient water quality data to support a validation were not available; the primary limitation being the lack of adequate point source data for the earlier period. In addition, both resource and data limitations precluded modeling sediment erosion and instream sediment transport and deposition processes, and their impacts on water quality. The calibration followed the steps discussed above for nonpoint and water quality calibration. The results presented here are a summary of the complete modeling results presented in the Final Project report with Appendices (AQUA TERRA Consultants and HydroQual, 2001).

Table 2 shows the mean annual runoff, simulated and observed, along with daily and monthly correlation coefficients for the six primary calibration sites. The CTWM hydrology results consistently show a good to very good agreement based on annual and monthly comparisons, defined by the calibration/validation targets discussed above. The monthly correlation coefficients are consistently greater than 0.9, and the daily values are greater than 0.8. The annual volumes are usually within the 10% target corresponding to a very good agreement, and always within the 15% target corresponding to a good agreement.

Table 2
Summary of CTWM hydrologic calibration/validation - annual flow and correlation coefficients

		Calibration Period (1991-1995)				Validation Period (1986-1990)			
Station Name	Station Number	Mean Observed Annual Flow (inches)	Mean Simulated Annual Flow (inches)	R Average Daily	R Average Monthly	Mean Observed Annual Flow (inches)	Mean Simulated Annual Flow (inches)	R Average Daily	R Average Monthly
Test Watershed Gages									
Salmon River nr East Hampton	01193500	23.6	24.4	0.83	0.92	26.3	25.8	0.79	0.92
Quinnipiac River at Wallingford	01196500	26.3	26.4	0.82	0.94	29.0	28.3	0.71	0.91
Norwalk River at South Wilton	01209700	21.4	21.7	0.84	0.93	25.9	25.2	0.75	0.91
Major Basin Gages									
Quinebaug River at Jewett City	01127000	23.8	23.6	0.82	0.93	27.2	24.7	0.86	0.95
Farmington River at Tariffville	01189995	26.2	26.0	0.85	0.92	26.2	29.1	0.87	0.94
Housatonic River at Stevenson	01205500	31.7	31.9	0.88	0.98	34.6	31.5	0.87	0.96

AQUA TERRA Consultants and HydroQual (2001)

Figures 4 and 5 present graphical comparisons of simulated and observed daily flows for the Quinnipiac River at Wallingford and the Farmington River at Tariffville, respectively. Figures 6 and 7 show flow duration plots for the same sites. Figures 8 and 9 show the scatterplots for daily flows at the Farmington gage for both the calibration and validation periods.

Based on the general 'weight-of-evidence', involving both graphical and statistical tests, the hydrology component of the CTWM was confirmed to be both calibrated and validated, and provides a sound basis for the water quality and loading purposes of the study.

Water Quality Results

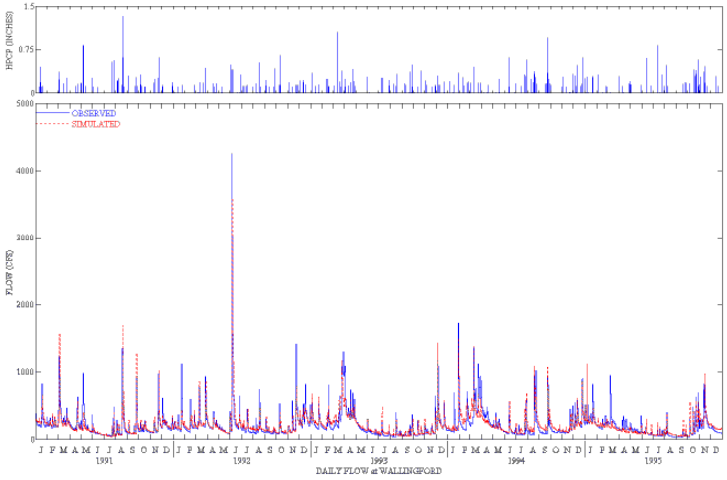
As noted above, the essence of watershed water quality calibration is to obtain acceptable agreement of observed and simulated concentrations (i.e. within defined criteria or targets), while maintaining the instream water quality parameters within physically realistic bounds, and the nonpoint loading rates within the expected ranges from the literature. The nonpoint loading rates, sometimes referred to as 'export coefficients' are highly variable, with value ranges sometimes up to an order of magnitude, depending on local and site conditions of soils, slopes, topography, climate, etc. Although a number of studies on export coefficients have been done for Connecticut, the values developed by Frink (1991) and shown below along with a 'standard error' term, appear to have the widest acceptance:

Frink's Export Coefficients (lb/ac/yr):

	Total Nitrogen	Total Phosphorus
Urban	12.0 ± 2.3	1.5 ± 0.2
Agriculture	6.8 ± 2.0	0.5 ± 0.13
Forest	2.1 ± 0.4	0.1 ± 0.03

Figure 4 - Observed and Simulated Daily Flow for the Quinnipiac River at Wallingford - Calibration and Validation
(Top curves are Daily Precipitation)

Calibration Period



Validation Period

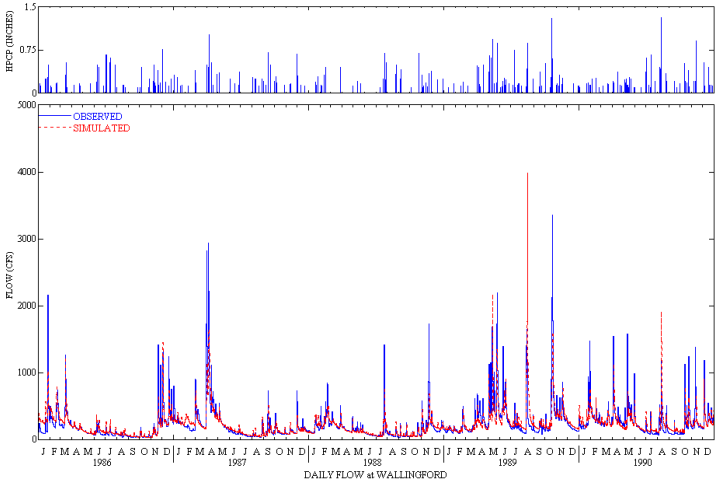
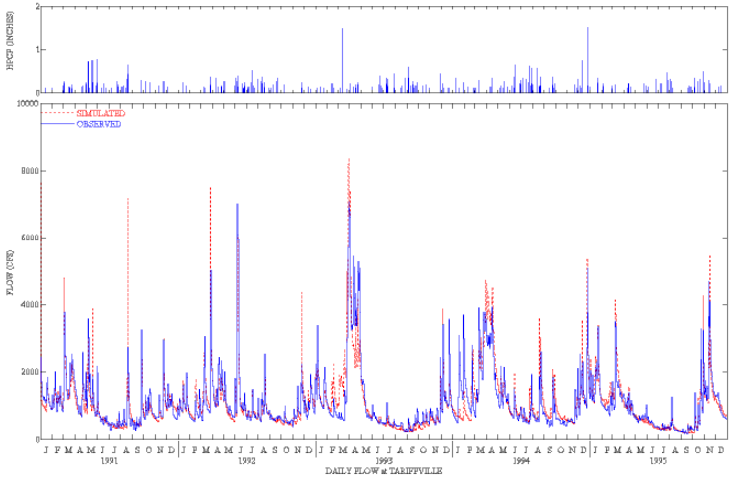


Figure 5 - Observed and Simulated Daily Flow for the Farmington River at Tariffville - Calibration and Validation
(Top curves are Daily Precipitation)

Calibration Period



Validation Period

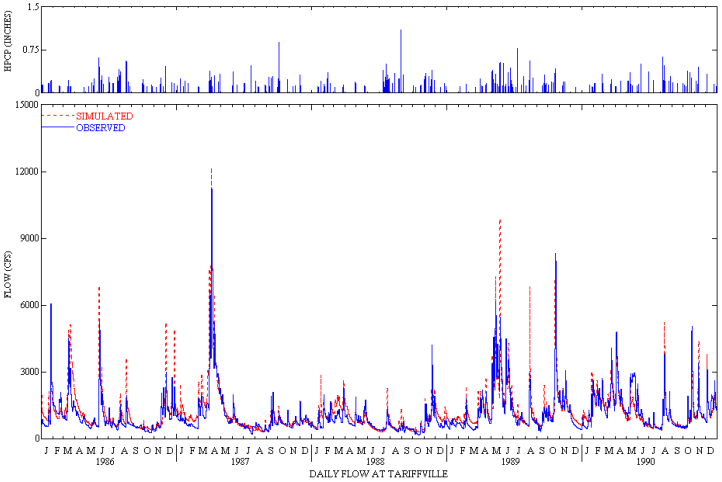


Figure 6 - Observed and Simulated Daily Flow Duration Curves for the Quinnipiac River at Wallingford - Calibration and Validation

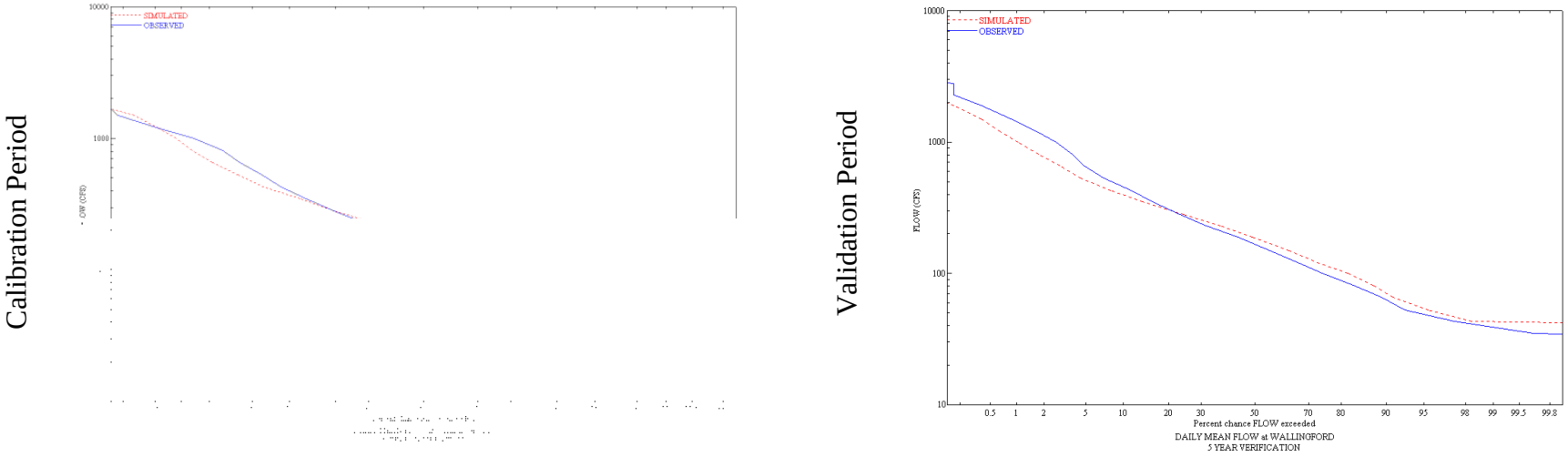


Figure 7 - Observed and Simulated Daily Flow Duration Curves for the Farmington River at Tariffville - Calibration and Validation

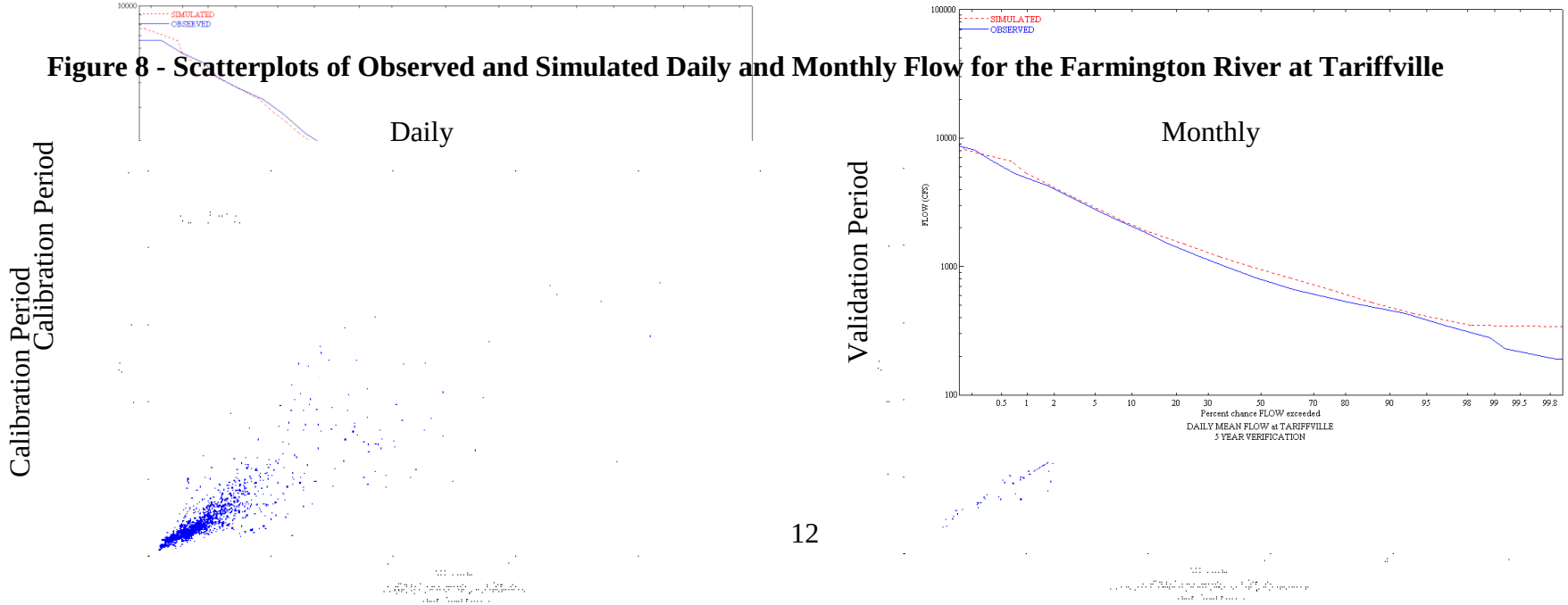
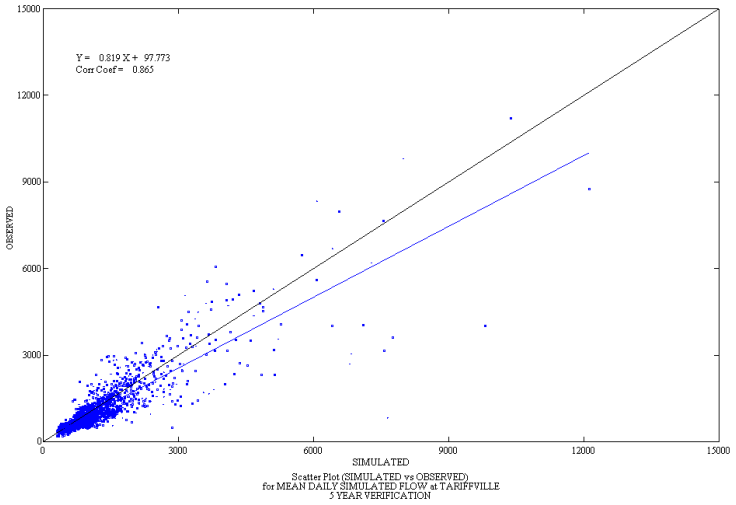


Figure 8 - Scatterplots of Observed and Simulated Daily and Monthly Flow for the Farmington River at Tariffville

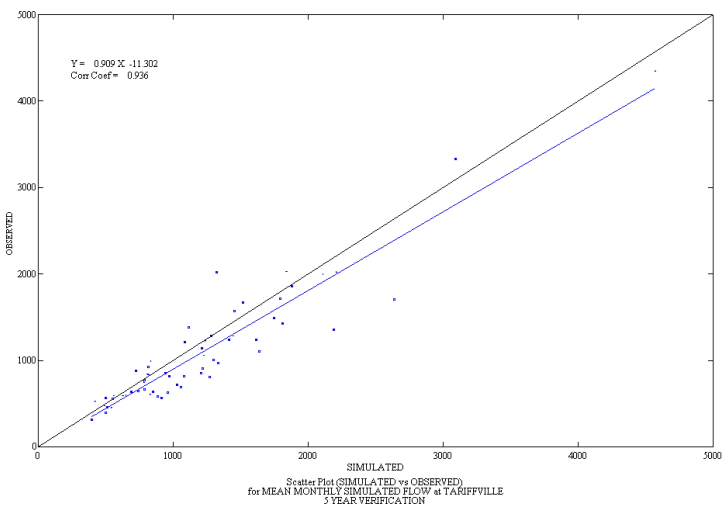
Figure 9 - Scatterplots of Observed and Simulated Daily and Monthly Flow for the Farmington River at Tariffville

Validation Period

Daily



Monthly



The above loading rates were used for general guidance, to supplement our past experience, in evaluating the CTWM loading rates and imposing relative magnitudes by land use type. No attempt was made to specifically calibrate the CTWM loading rates to duplicate the export coefficients noted above. The overall calculated mean annual loading rates and ranges for Total N and Total P for 1991-95, are summarized as follows:

	CTWM Loading Rates (lb/ac/yr) Mean (Range)	
	Total Nitrogen	Total Phosphorus
Urban - pervious	8.5 (5.6 - 15.7)	0.26 (0.20 - 0.41)
Urban - impervious	4.9 (3.7 - 6.6)	0.32 (0.18 - 0.36)
Agriculture	5.9 (3.4 - 11.6)	0.30 (0.23 - 0.44)
Forest	2.4 (1.4 - 4.3)	0.04 (0.03 - 0.08)
Wetlands	2.2 (1.4 - 3.5)	0.03 (0.02 - 0.05)

Considering the purposes of the study, and the assumptions in the model development (e.g. sediment not simulated), these loading rates were judged to be consistent with Frink's values and the general literature, and thus acceptable for the modeling effort (see Final Report for details and breakdown of TN and TP into components).

Tables 3 and 4 display the mean simulated and observed concentrations for the five-year period for all of the water quality stations where calibration was performed. The comparison of mean concentrations, and the ratios of simulated to observed values, demonstrate that simulated values are generally within 20% of observed, i.e. the ratios are mostly between 0.8 and 1.2, and often between 0.9 and 1.1. The biggest differences are for the phosphorus compounds, where the ratios range from 0.91 to 1.9. Considering all the sites (Table 4), the mean value for the ratios for DO, TOC and nitrogen forms are within a range of 0.89 to 0.99, while the phosphorus ratios are 1.33 to 1.40. Comparing these ratios to the proposed calibration targets indicates a 'very good' calibration of nitrogen, and a borderline 'fair' calibration of phosphorus.

Figures 10 and 11 present typical graphical comparisons made for simulated and observed water quality constituents. Figure 10 presents a comparison of simulated and observed Total Phosphorus for the Quinnipiac River at Wallingford. Figure 11 presents a similar comparison for Total Nitrogen at the Tariffville gage on the Farmington River.

Table 3
Average Annual Concentrations (mg/L) for the Calibration Period (1991-1995)

	Salmon River nr East Hampton			Quinnipiac River at Wallingford			Norwalk River at Winnipauk			Quinebaug River at Jewett City			Farmington River at Tariffville			Housatonic River at Stevenson		
Constituent	Observed	Simulated	Ratio * (sample size)	Observed	Simulated	Ratio * (sample size)	Observed	Simulated	Ratio * (sample size)	Observed	Simulated	Ratio * (sample size)	Observed	Simulated	Ratio * (sample size)	Observed	Simulated	Ratio * (sample size)
Dissolved Oxygen	10.9	10.5	0.96 (48)	10.4	10.3	0.99 (46)	11.6	10.4	0.90 (97)	10.4	10.3	0.99 (43)	10.2	10.8	1.06 (49)	9.5	9.5	1.01 (41)
Ammonia as N	0.03	0.02	0.82 (43)	0.19	0.18	0.92 (46)	0.04	0.04	1.18 (80)	0.08	0.06	0.73 (42)	0.10	0.09	0.82 (48)	0.06	0.06	1.10 (33)
Nitrite-Nitrate as N	0.22	0.27	1.21 (46)	2.82	2.45	0.87 (46)	0.39	0.40	1.03 (93)	0.44	0.37	0.84 (42)	0.71	0.59	0.83 (49)	0.36	0.41	1.15 (40)
Organic Nitrogen	0.31	0.25	0.80 (30)	0.50	0.60	1.20 (44)	0.33	0.28	0.86 (70)	0.45	0.39	0.86 (40)	0.31	0.28	0.90 (45)	0.33	0.28	0.84 (38)
Total Nitrogen	0.53	0.51	0.97 (30)	3.64	3.29	0.90 (44)	0.73	0.69	0.94 (70)	0.96	0.80	0.83 (40)	1.15	0.97	0.85 (45)	0.77	0.75	0.97 (38)
Orthophosphate as P	0.01	0.01	0.91 (48)	0.32	0.36	1.10 (46)	0.02	0.02	0.93 (94)	0.02	0.04	1.67 (43)	0.07	0.13	1.90 (49)	0.01	0.02	1.49 (32)
Organic Phosphorus	0.02	0.02	1.30 (48)	0.07	0.11	1.62 (46)	0.02	0.03	1.18 (94)	0.03	0.04	1.23 (43)	0.03	0.05	1.59 (49)	0.02	0.03	1.19 (33)
Total Phosphorus	0.02	0.03	1.35 (48)	0.39	0.47	1.19 (46)	0.04	0.05	1.10 (94)	0.06	0.08	1.44 (43)	0.10	0.18	1.82 (49)	0.03	0.05	1.47 (40)
Total Organic Carbon	3.9	2.8	0.71 (45)	4.5	4.8	1.06 (44)	4.0	3.2	0.81 (28)	5.6	4.9	0.86 (41)	3.9	3.3	0.84 (45)	3.8	2.9	1.06 (49)

* Ratios calculated from Simulated and Observed concentrations prior to rounding

Table 4
Average and Range of Simulated/Observed Concentration Ratios for all Sites

Constituent	Average	Range
Dissolved Oxygen	0.99	0.90 - 1.06
Ammonia as N	0.93	0.73 - 1.18
Nitrite-Nitrate as N	0.99	0.83 - 1.21
Organic Nitrogen	0.91	0.80 - 1.20
Total Nitrogen	0.91	0.83 - 0.97
Orthophosphate as P	1.33	0.91 - 1.90
Organic Phosphorus	1.35	1.18 - 1.62
Total Phosphorus	1.40	1.10 - 1.82
Total Organic Carbon	0.89	0.71 - 1.06

Figure 10 - Observed and Simulated Daily Total Phosphorus Concentrations for the Quinnipiac River at Wallingford

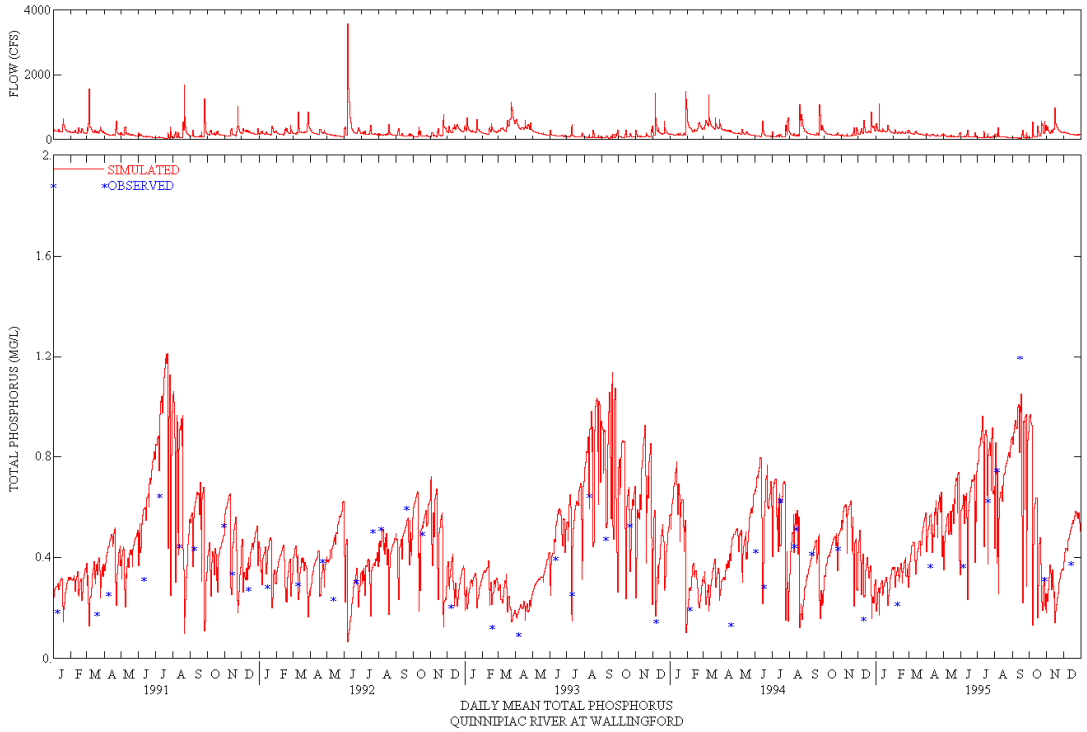
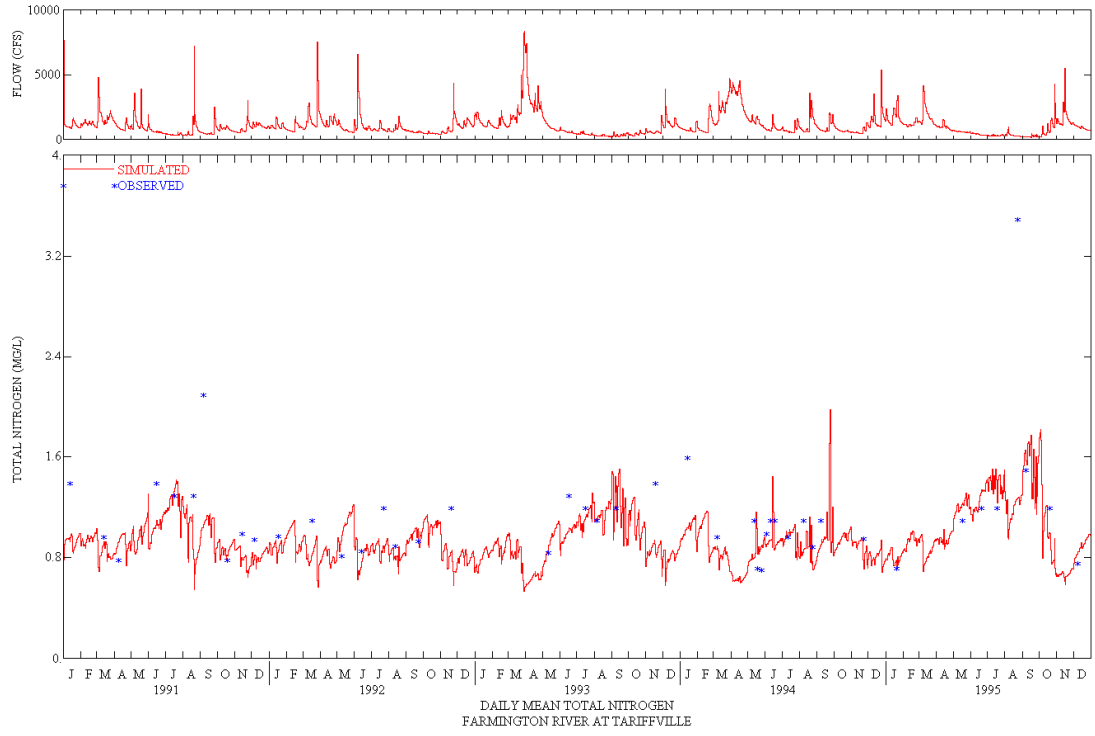


Figure 11 - Observed and Simulated Daily Total Nitrogen Concentrations for the Farmington River at Tariffville



CTWM Study Conclusions

Based on the general 'weight-of-evidence' of the hydrology and water quality simulation results, including the CTWM loading rates, the mean concentrations and ratios, and the timeseries comparisons of observed and simulated values, the CTWM was determined to be an acceptable representation of the Connecticut watersheds providing loadings to LIS. This evidence indicates that the predicted nitrogen and carbon loadings are a 'very good' representation of the observed data, based on the established calibration targets, and that the phosphorus loadings are a 'fair' representation. Clearly improvements can be made to better represent these loadings, especially for phosphorus, but the CTWM in its current form is a sound tool for examining loadings to LIS and providing the basis for developing and analyzing alternative watershed scenarios designed to improve the water quality of LIS.

Housatonic River Watershed

HSPF was applied to the almost 300 sq. mi. Housatonic River watershed in Massachusetts. The tables presented below demonstrate some additional types of comparisons for evaluating the hydrologic simulation results, in comparison with the targets shown in Table 1. Table 5 shows the annual simulated and observed runoff, along with annual precipitation, and percent error or difference for each year of the 10-year calibration. The total difference for the 10-years is less than 2%, while the annual differences are within about 15%, indicating a good to very good calibration.

Table 5
Annual Simulated and Observed Runoff (inches)

	Housatonic River Watershed			
	Precipitation	Simulated Flow	Observed Flow	Percent Error
1990	58.9	35.1	35.6	-1.4%
1991	47.0	23.3	22.8	2.1%
1992	45.7	23.7	20.1	15.2%
1993	47.6	27.6	26.0	5.8%
1994	46.3	25.9	25.5	1.5%
1995	44.0	20.7	21.0	-1.4%
1996	62.0	39.4	41.5	-5.3%
1997	42.2	21.4	23.2	-8.4%
1998	42.2	22.0	23.9	-8.6%
1999	46.9	21.6	24.8	-14.8%
Total	482.7	260.7	264.4	-1.4%
Average	48.3	26.1	26.4	-1.4%

Weston Solutions (2006)

Table 6 shows the statistical output available from HSPEXP for both the 'Watershed Outlet' and an 'Upstream Tributary' of about 60 sq. mi., while Table 7 shows a variety of statistics for both daily and monthly comparisons at the watershed outlet. The storm statistics in Table 6 are based

Table 6
Annual Flow Statistics from HSPEXP

	Upstream Tributary			Watershed Outlet	
	Simulated	Observed		Simulated	Observed
Average runoff, in inches	27.12	26.23		26.07	26.44
Total of highest 10% flows, in inches	10.88	10.72		8.56	8.94
Total of lowest 50% flows, in inches	4.22	4.19		5.09	5.13
Evapotranspiration, in inches	23.77	25.55[1]		23.41	26.09[1]
Total storm volume, in inches[2]	47.07	51.91		38.72	42.36
Average of storm peaks, in cfs[2]	710.84	791.88		2310.38	2287.19
	Calculated	Criteria		Calculated	Criteria
Error in total volume, %	3.40	10.00		-1.40	10.00
Error in 10% highest flows, %	1.50	15.00		-4.20	15.00
Error in 50% lowest flows, %	0.60	10.00		-0.60	10.00
Error in storm peaks, %	-10.20	15.00		1.00	15.00

1 – PET (estimated by multiplying observed pan evaporation data by 0.73)
2 – Based on 31 storms occurring between 1990 and 1999
Weston Solutions (2006)

Table 7
Daily and Monthly Average Flow Statistics

Unnamed Watershed					
	Daily			Monthly	
	Simulated	Observed		Simulated	Observed
Count	3652	3652		120	120
Mean, cfs	539.85	547.65		540.46	547.56
Geometric Mean, cfs	376.61	380.86		424.39	428.44
Correlation Coefficient (R)	0.86			0.93	
Coefficient of Determination (R^2)	0.74			0.87	
Mean Error, cfs	-7.80			-7.10	
Mean Absolute Error, cfs	152.97			101.22	
RMS Error, cfs	284.09			140.26	
Model Fit Efficiency (1.0 is perfect)	0.73			0.87	

Weston Solutions (2006)

on a selection of 31 events throughout the 10-year period, distributed to help evaluate seasonal differences. The correlation statistics in Table 7 indicate a 'good' calibration for daily values, and a 'very good' calibration of monthly flows, when compared to the value ranges in Figure 2.

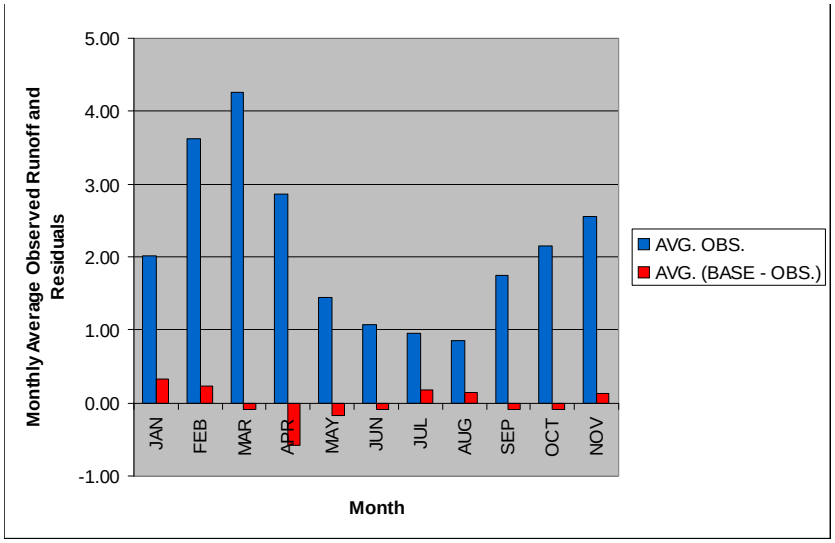
Table 8 shows the mean monthly observed and simulated runoff, along with their differences (or residuals) and '% error', as another assessment of the seasonal representation of the model; Figure 12 graphically shows the mean observed and the residuals from Table 8. This demonstrates a need to improve the spring and early summer results where the model undersimulates the monthly observations.

Table 8
Average Observed Monthly Runoff and Residuals

	Unnamed Watershed			
Month	Average Observed (in.)	Average Simulated (in.)	Average Residual (Simulated - Observed)	Percent Error
JAN	2.94	2.71	-0.24	-8.09%
FEB	2.01	2.34	0.33	16.46%
MAR	3.61	3.85	0.23	6.42%
APR	4.25	4.16	-0.09	-2.07%
MAY	2.86	2.28	-0.58	-20.19%
JUN	1.44	1.26	-0.18	-12.55%
JUL	1.07	0.97	-0.10	-9.03%
AUG	0.95	1.13	0.18	18.66%
SEP	0.85	0.98	0.14	16.39%
OCT	1.75	1.66	-0.08	-4.80%
NOV	2.15	2.05	-0.09	-4.38%
DEC	2.56	2.70	0.13	5.03%
Totals	26.46	26.08	-0.35	-1.32%

Weston Solutions (2006)

Figure 12 - Unnamed Watershed Observed Runoff and Residuals (inches)



Weston Solutions (2006)

Tables 9 and 10 respectively show the simulated and expected water balance for the watershed, and the separate water balances for each land use simulated by the model. As noted earlier, these comparisons are consistency checks to compare the overall simulation with the expected values from the literature, and to evaluate how well the model represents land use differences.

Table 9
Average Annual Expected and Simulated Water Balance

	Expected Ranges	Simulated
Moisture Supply	43 - 53	48
Total Runoff	23 - 27	24
Total ET	20 - 23	23
Deep Recharge	1 - 4	1

Table 10
Simulated Water Balance Components by Land Use

	Forest	**Agriculture**	**Urban Pervious**	**Wetland**	**Urban Impervious**
Moisture Supply	**48.6**	**48.4**	**48.5**	**48.5**	**48.3**
Total Runoff	**22.6**	**25.8**	**26.5**	**21.3**	**42.8**
Surface Runoff	1.0	4.6	4.6	0.3	42.7
Interflow	7.9	8.8	8.8	4.8	0.0
Baseflow	13.6	12.3	13.1	16.2	0.0
Total ET	**24.6**	**22.1**	**21.2**	**24.2**	**5.5**
Interception/Retention ET	9.6	6.1	6.3	4.6	5.5
Upper Zone ET	7.8	6.5	9.2	11.1	0.0
Lower Zone ET	6.6	9.2	5.3	4.6	0.0
Active GW ET	0.0	0.0	0.0	2.9	0.0
Baseflow ET	0.6	0.3	0.3	1.0	0.0
Deep Recharge	**1.4**	**0.5**	**0.8**	**3.0**	**0.0**

Weston Solutions (2006)

PARAMETER SENSITIVITY ANALYSIS

The sensitivity to variations in input parameter values is an important characteristic of a model. Sensitivity analysis is used to identify the most influential parameters in determining the accuracy and precision of model predictions. Sensitivity analysis quantitatively or semi-

quantitatively defines the dependence of the model's performance on a specific parameter or set of parameters. Sensitivity analysis can also be used to establish strategies for improving the efficiency of the calibration process.

Model sensitivity can be expressed as the relative rate of change of selected output caused by a unit change in the input. If the change in the input causes a large change in the output, the model is considered to be sensitive to that input parameter. Sensitivity analysis methods are mostly nonstatistical or even intuitive by nature. Sensitivity analysis is typically performed by changing one input parameter at a time and evaluating the effects on the distribution of the dependent variable. Nominal, minimum, and maximum values are specified for the selected input parameter.

It should be noted that informal sensitivity analyses (iterative parameter adjustments) provide the basis for model calibration and ensure that reasonable values for model parameters will be obtained and will in turn result in acceptable model results. The degree of allowable adjustment of any parameter is usually directly proportional to the uncertainty of its value and is limited to its expected range of values (Tetra Tech, 2009).

Sensitivity Analysis Procedures

In a paper reporting the results of modeling and assessing model performance in simulating flow, sediment and water temperature in the Housatonic River (MA), Donigian and Love (2007) recently outlined a generally applicable procedure for performing a sensitivity analysis for parameters used in HSPF calibration. They described the following steps:

1. Identify the critical model input and parameters, based either on past experience or specific calibration experience for the watershed.
2. Identify reasonable percent perturbations from the calibration values, increases and decreases, for each model input and parameter.
3. Assess the resulting changes to ensure the absolute differences in input and parameters are reasonable and appropriate. Perform a long-term model run (e.g., 25 years) using the calibration parameters as a baseline simulation.
4. Perform additional model runs for the entire period, with each run representing a single input/parameter change.
5. Process the model sensitivity run results to calculate the percent difference from the baseline and the sensitivity factor, defined as the percent change in model output divided by the percent change in input/parameter value.
6. Rank the model input and parameters by the sensitivity metric to establish those with the greatest impact on model results

Example Metrics and Results

In the above referenced sensitivity analysis performed for the HSPF application to the Housatonic River, the specific model output values for each model run which provided the basis for the sensitivity analyses include: (1) mean annual streamflow, cubic feet per second (cfs); (2)

mean annual runoff, inches; (3) 10 percent highest flows (i.e., mean flow exceeded 10 percent of the time), cfs; (4) 25 percent lowest flows (i.e., mean flow exceeded 75 percent of the time), cfs; (5) average peak flow (average of 62 selected storms), cfs; (6) mean annual total suspended solids (TSS) loadings, tons/year; (7) mean annual water temperature, °F; and (8) mean summer water temperature (June – August), °F. All of these quantities were produced by the model at six stream site locations throughout the Housatonic Watershed; additional details are provided in Weston Solutions, Inc. (2004; 2006).

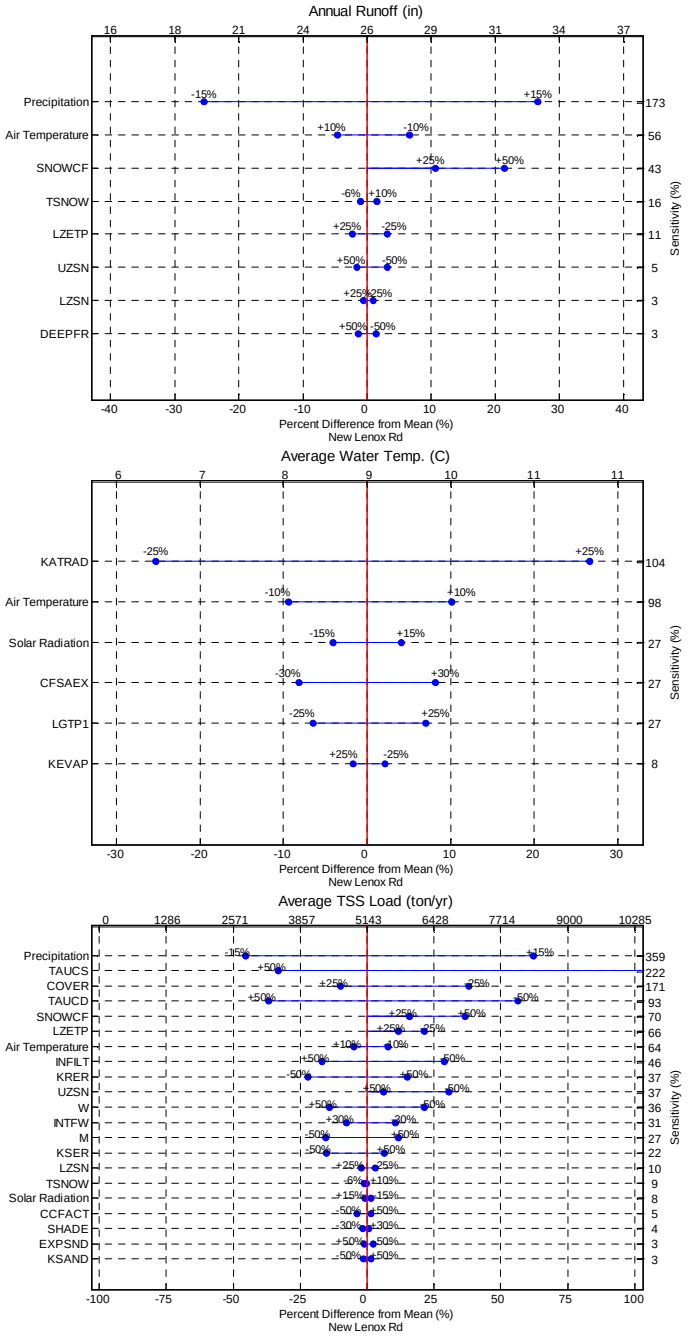
Sensitivity factors can be calculated as the ratio (expressed as a percentage) of the average absolute percent change in model output for the two model runs to the average absolute percent change in input/parameters. Values near 100% indicate a 1:1 sensitivity with the model producing a result in direct proportion to the input/parameter change; e.g., a 10% change in input/parameter produces a 10% change in model results. In a similar fashion, values of the sensitivity factor near 200% indicate a highly sensitive response of 2:1, whereas a value of 10% indicates relative model insensitivity of 0.1:1, where a 10% input/parameter change produces only a 1% model response.

Sensitivity results can be displayed in graphics such as Figure 13, referred to as "tornado diagrams." Within each diagram, the input/parameters are shown on the left ordinate, ranked by the sensitivity factor (highest to lowest) which is listed on the right ordinate. The bottom horizontal scale shows the "percent difference" from the baseline values, while the top horizontal scale shows the absolute values of the model results. Within the figures, the vertical center line is the mean value from the baseline run, with the width of the horizontal line for each model input/parameter representing the model results from the parameter perturbations. Of the 30 model input/parameters that were analyzed in the Housatonic study, only a subset had significant impacts and showed high sensitivity to each of the model output quantities; thus, the final tornado diagrams rank and display only the input/parameters with sensitivity factors greater than 2–3%.

Review of model sensitivity results can provide many useful insights. For example, the sensitivity analysis performed for the HSPF application to the Housatonic resulted in the following observations:

- All model results were the most sensitive to the input of precipitation and air temperature, reflecting the importance of accurately representing climate conditions for the watershed.
- Precipitation, as expected, dominated all the model output related to flow and sediment loads, with air temperature occupying the top spot for the water temperature results. Precipitation sensitivity factors approached or exceeded 200% for all the flow output and were in the range of 300–500% for annual TSS loads.
- For flow results, the next most sensitive parameters were related to the snow simulation, SNOWCF and TSNOW, reflecting the impact of snow accumulation and melt on the hydrology of the Housatonic River Watershed. The HSPF snow cover factor (SNOWCF) was used to increase the recorded gage precipitation for deficiencies in accurately recording snowfall amounts; thus, it had a direct impact on input precipitation during the

Figure 13 - Example Tornado Diagrams at New Lenox Road (Donigian and Love (2007)
(Conversion Factors: 1 in = 2.54 cm; 35.3 cfs = 1 cms; 1 short ton = 0.907 metric tons)



winter months. TSNOW, which is the threshold temperature between rain and snow, only affected the form of precipitation and not the amount.

- Soils-related parameters were next in importance, followed by sensitivity, including infiltration (INFILT), soil/plant evapotranspiration (LZETP), and soil moisture storages

(UZSN, LZSN). These parameters showed greater sensitivity for high and low flows values and for average storm peaks than for the mean annual runoff and flow.

- Average TSS loads were also sensitive to the forcing functions of precipitation and air temperature since they controlled the runoff and streamflow that supplies the source and transport mechanisms for TSS. However a larger number of parameters showed some degree of sensitivity for TSS since both hydraulic and sediment processes are important.
- For water temperature sensitivity, the forcing functions of air temperature and solar radiation were high on the list at all sites, along with instream heat exchange (KATRAD) and surface exposure (CFSAEX) parameters.

PARAMETER UNCERTAINTY

After calibration and parameter sensitivity analysis are completed, the uncertainty in the calibrated model caused by uncertainty in the estimates of the model input parameters can be assessed. Formal uncertainty analyses for watershed model applications have not historically been done, largely due to the complexity and computational demands of most watersheds that are modeled. To address the need for uncertainty analyses while recognizing such restrictions when complex codes are involved, one approach is to identify key parameters using a sensitivity analysis and then to focus on the model uncertainty associated with those parameters identified as most "sensitive."

For example, in the case of the Housatonic River HSPF application, sensitivity analysis results demonstrated the critical importance of accurate and representative climate forcing data to adequately perform watershed modeling. The results also revealed the significant impact of snow-related parameters and processes, the effects of instream scour and heat exchange parameters, and less sensitive impacts of soil parameters.

Typically in cases where uncertainty analysis is performed on watershed model applications, a Monte Carlo approach is utilized. For the Housatonic analysis, Monte Carlo simulation was performed for the watershed model that involved execution of 600 model runs, each for the 11-year time period of water years 1990–2000, with selected model parameters being randomly chosen from assigned Bounded Normal (NO) or Bounded Lognormal (LN) probability distributions. The parameters that were randomly varied were determined based on the results from the sensitivity analysis, i.e., those with high sensitivity factors. Bounded distributions were used for two reasons: to ensure parameter values stayed within physically realistic limits for the Housatonic Watershed and within computational limits imposed by HSPF. In assigning distributions, each parameter was first characterized in terms of whether it reflected soil, climate, vegetation, sediment, or general site characteristic, or some combination of these. Then an LN distribution was assigned for the soil- and sediment-related parameters and NO distributions were assigned for the others. A number of articles on soil hydrologic and hydraulic characteristics (see Weston Solutions Inc. (2006) for citations) clearly confirm a general consensus that soil properties more often demonstrate LN distributions, and the LN generally is preferred over an NO distribution. To address the issue of parameter correlation, major parameters were identified and grouped that were clearly related because they represented similar soil, sediment, or vegetation characteristics of the watershed. Consequently, these

parameters were correlated in terms of any perturbation performed as part of the uncertainty analysis, and an appropriate correlation structure was incorporated into the parameter perturbations generated for each model run. The Sandia Latin Hypercube Sampling (LHS) software (Wyss and Jorgensen, 1998) used in this effort implements a nonparametric technique known as rank correlation that allows the user to specify which model parameters to correlate within a sample. The method preserves the sampling scheme; i.e., the same numbers originally selected as input values are retained; only their pairing is affected to achieve the desired rank correlations (Iman and Conover, 1982). Thus for each correlated group, the parameters perturbations were correlated so that their values changed in the same direction and with similar magnitudes; i.e., they each increased or decreased together for each model run.

The model results were processed for the same output variables and locations as used in the sensitivity analysis to quantify the expected uncertainty in the model predictions. Quality assurance efforts focused on assessing both the plausibility of parameter distributions and model results and the stability of the Monte Carlo procedures. Model parameters generated for the Monte Carlo runs were plotted and checked for adherence to their assigned distributions and bounds, and model results were checked in comparison to the full range of calibration/validation results. Stability refers to the sensitivity of the outcome of interest to the sample size (i.e., number of runs) and was checked by analyzing the convergence behavior of the expected result as model runs were performed. After confidence was gained in the Monte Carlo methodology and procedures, uncertainty in the model predictions was expressed by calculating the 5th and 95th percentiles of the ranked output, representing the range for 90 percent of the model results. The differences between the mean value and the 5th and 95th percentiles values were calculated, divided by the mean and expressed as percentages, and averaged to express uncertainty as the percent deviation from the mean. Normalizing to the mean allowed for uncertainty comparisons to be made between the output variables (i.e., flow, TSS, temperature) and within specific percentiles of output variables; e.g., uncertainty in the 10 percent highest flows, 25 percent lowest flows, and throughout the flow duration curve.

Operationally, a number of different codes and software components were used to generate the parameter distributions, update the HSPF input file (UCI file), execute the run, and process the output. The overall process was organized within a Matlab© framework and used Sandia LHS software to generate the parameter values for the NO and LN distributions, in-house scripts to revise the UCIs, and Matlab© again to execute HSPF and process the output.

Following the quality assurance and stability checks indicating that the Monte Carlo simulation was stable and reasonable, the results of the model runs were analyzed to determine the 90 percent range representing the values between the 5th and 95th percentiles. These were determined using Matlab©, by rank ordering the 600 values, high to low, to identify the 30th value and the 571st value regions. For the majority of the output variables, the percentiles were calculated based on the annual mean values that resulted from each of the runs. For each of the six sites, the 5th and 95th percentile values were determined for each of the output variables; the results of this analysis are presented in Table 11. These values represent the range that encompasses 90 percent of the values produced by the 600 runs.

In addition, a flow duration curve was generated for each of the model runs at each site. This information was processed in the same manner as noted above, allowing generation of a mean along with the 5th and 95th percentile for selected flow-interval durations. Ultimately, this resulted in flow duration graphics for each site with three curves that show the mean bounded by the 5th and 95th percentile flow durations. An example flow duration curve for each site is shown in Figure 14.

A summary uncertainty statistic, referred to as "Percent Uncertainty" in the tables, was calculated as:

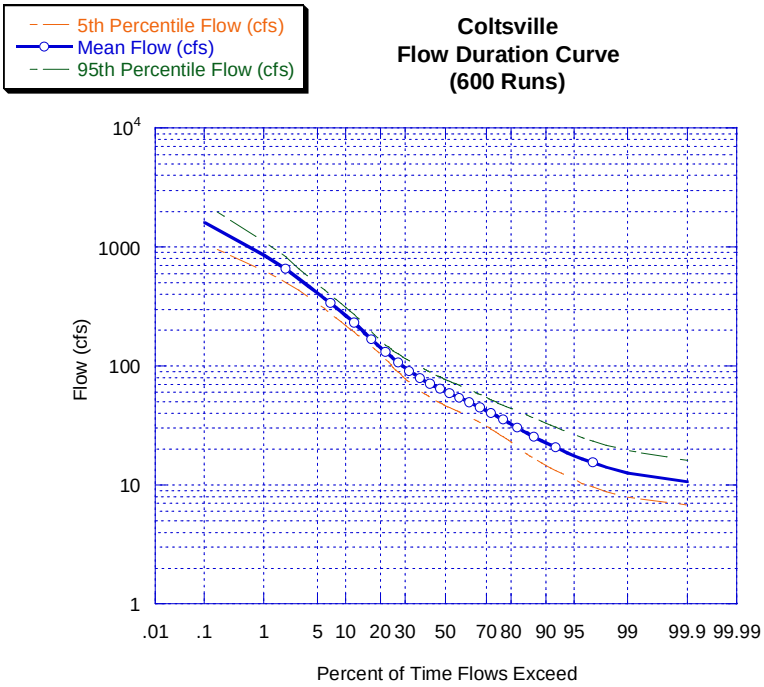
$$\text{Percent Uncertainty} = ((95\text{th Percentile} - 5\text{th Percentile}) / \text{Mean}) / 2 \times 100 \quad (1)$$

Table 11 - 5th to 95th Percentile Ranges and Percent Uncertainty of Output Variables (Donigian and Love, 2007)

	Mean Streamflow (cfs)	Mean Annual Runoff (in/yr)	10% High Flow (cfs)	25% Low Flow (cfs)	Mean Summer Water Temperature (C)	Mean Annual Water Temperature (C)	Mean Annual Sediment Load (ton/yr)
Coltsville - Reach 110							
5th Percentile	105	27	220	27	13.9	6.1	448
Mean	116	30	267	37	18.9	8.9	4,599
95th Percentile	129	33	311	49	23.3	12.2	9,785
Percent Uncertainty	10%	10%	17%	29%	13%	11%	102%
New Lenox–Reach 540							
5th Percentile	272	27	533	88	14.4	6.1	77
Mean	300	30	638	118	18.9	8.9	13,870
95th Percentile	332	34	738	152	22.8	11.7	47,654
Percent Uncertainty	10%	10%	16%	27%	12%	10%	172%
Great Barrington–Reach 900							
5th Percentile	519	27	1,040	149	15.0	6.1	165
Mean	576	30	1,273	211	21.1	10.0	15,522
95th Percentile	640	34	1,558	281	26.7	13.3	72,223
Percent Uncertainty	11%	11%	20%	31%	15%	13%	232%

(Conversion Factors: 1 in = 2.54 cm; 35.3 cfs = 1 cms; 1 short ton = 0.907 metric tons)

Figure 14 - Flow Duration Curves Showing Mean, 5th, and 95th Percentile Flows (Donigian and Love (2007))



This allowed comparison of uncertainty among the different variables analyzed in this effort, as follows:

- As shown in Table 11, the overall level of uncertainty was lowest for all the flow metrics, slightly higher for the water temperature metrics, and as expected, highest for the sediment metrics.
- The mean annual streamflow and runoff showed identical "percent uncertainty" values in the range of 10–11%, with high flows in the range of 17–20%, and low flows in the range of 27-31%. Greater uncertainty at the extreme flows–both high and low–was expected, and was demonstrated in the flow duration uncertainty results in Figure 14.
- The "Percent Uncertainty" for the output variables analyzed, as shown in Table 11, was very consistent among all the sites analyzed. The mean annual sediment load was an exception; it showed the largest variation and increases with increasing drainage area.
- The uncertainty levels for water temperature were low and similar for both the mean summer and mean annual values, in the range of 10–15%.

Analysis results showed that the uncertainty associated with flow and water temperature was in both cases similar to the level of agreement with available data for an acceptable calibration. Thus a high level of confidence could be placed in the representative nature of the HSPF

boundary conditions for flow and water temperature. However, the uncertainty for the boundary watershed sediment loads contained significantly higher levels of uncertainty.

CLOSURE

This paper has focused on presenting methods of measuring and evaluating model performance by means of model calibration and validation, parameter sensitivity analysis and parameter uncertainty analysis.

A 'weight-of-evidence' approach to watershed model calibration and validation has been described based on experience with the HSPF model. Examples have been provided to demonstrate some of the graphical and statistical comparisons that should be performed whenever model performance is evaluated. Although not all models will employ the identical procedures described above, it is clear that multiple tests and evaluations, not reliance on a single statistic, should be part of all watershed modeling studies.

Sensitivity analysis defines the dependence of the model's performance on a specific parameter or set of parameters. Sensitivity of model results to parameters varies from watershed to watershed, with relative sensitivity in a given watershed depending on the combined impacts of climate and watershed conditions. In this paper we have discussed parameter sensitivity analysis in general terms applicable to all watershed models, and we provide techniques and examples specific to HSPF applications to illustrate this discussion.

A final element of assessing and understanding a model's performance is achieved by means of uncertainty analysis. Formal uncertainty analyses for watershed model applications have not historically been done, largely due to the complexity and computational demands of most watersheds that are modeled. To address the need for uncertainty analyses while recognizing such restrictions when complex codes are involved, we have presented and discussed a new and inventive methodology to identify key parameters by using a preliminary sensitivity analysis and then focusing on the model uncertainty associated with those parameters identified as most "sensitive." Again, we have provided an example approach and results generated by HSPF applications.

By combining the lessons learned and metrics achieved by all three of these model performance exercises, model users and those dependent upon model results have a solid basis for establishing and expressing their level of confidence in a model's ability to represent current and alternative watershed responses.

REFERENCES

AQUA TERRA Consultants and HydroQual, Inc. (2001) Modeling Nutrient Loads to Long Island Sound from Connecticut Watersheds, and Impacts of Future Buildout and Management Scenarios. Prepared for Connecticut Department of Environmental Protection, Hartford, CT.

ASTM. (1984) Standard Practice for Evaluating Environmental Fate Models of Chemicals. Designation E978-84. American Society of Testing Materials. Philadelphia, PA. 8 p.

Bicknell, B.R., J.C. Imhoff, J.L. Kittle Jr., T.H. Jobes, and A.S. Donigian, Jr. (2005) Hydrological Simulation Program - Fortran (HSPF). User's Manual for Release 12.2 U.S. EPA National Exposure Research Laboratory, Athens, GA, in cooperation with U.S. Geological Survey, WRD, Reston, VA.

Berthouex, P. M. and L. C. Brown. 1994. *Statistics for Environmental Engineers.* Lewis Publishers, CRC Press, Boca Raton, FL. 335 p.

Donigian, Jr., A.S. (1982) Field Validation and Error Analysis of Chemical Fate Models. In: *Modeling Fate of Chemicals in the Aquatic Environment*. Dickson et al, (eds), Ann Arbor Science Publishers, Ann Arbor, MI. 303-323 p.

Donigian, Jr., A.S. (2000) HSPF Training Workshop Handbook and CD. Lecture #19. Calibration and Verification Issues, Slide #L19-22. EPA Headquarters, Washington Information Center, 10-14 January, 2000. Presented and prepared for U.S. EPA, Office of Water, Office of Science and Technology, Washington, D.C.

Donigian, Jr., A.S., and J.T. Love (2002) The Connecticut Watershed Model - A Tool for BMP Impact Assessment in Connecticut. Presented at WEF-Watershed 2002, February 23-27, 2002. Ft. Lauderdale, FL. CD-ROM Proceedings.

Donigian, A.S. Jr. and J.T. Love (2007) The Housatonic River Watershed Model: Model Application and Sensitivity/Uncertainty Analysis. 7th International IWA Symposium on System Analysis and Integrated Assessment in Water Management, May 7-9, 2007. Washington, DC. WATERMATEX Proceedings on CD-ROM.

Donigian, A.S. Jr. and P.S.C. Rao (1990) Selection, Application, and Validation of Environmental Models. Proceedings of International Symposium on Water Quality Modeling of Agricultural Nonpoint Sources. Part 2. June 19-23, 1988. Logan, UT. USDA-ARS Report No. ARS-81. D. G. Decoursey (ed). pp 577- 604

Frink, C. R. (1991) Estimating Nutrient Exports to Estuaries. J. Environ. Qual. 20(4): 717 - 724.

Hummel, P.R., J.C. Imhoff, R. Dusenbury and M. Gray (2000) TMDL USLE, A Practical Tool for Estimating Diffuse Sediment Source Loads within a Watershed Context. Prepared for: U.S. EPA National Exposure Research Laboratory, Athens, GA.

HydroQual, Inc. (1996) Water Quality Modeling Analysis of Hypoxia in Long Island Sound Using LIS 3.0. Conducted by direction of the Management Committee of the Long Island Sound Study through a contract with the New England Interstate Water Pollution Control Commission.

Iman, R. L. and W.J. Conover (1982) A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics B11,* **11,** 335–360.

Love, J. T. and A. S. Donigian, Jr. (2002) The Connecticut Watershed Model – Model Development, Calibration, and Validation. Presented at WEF-Watershed 2002, February 23-27, 2002. Ft. Lauderdale, FL. CD-ROM Proceedings.

Lumb, A.M., R.B. McCammon, and J.L. Kittle, Jr. (1994) Users Manual for an Expert System (HSPEXP) for Calibration of the Hydrological Simulation Program - FORTRAN. Water-Resources Investigations Report 94-4168, U.S. Geological Survey, Reston, VA. 102 p.

McCuen, R. H. and W. M. Snyder. 1986. *Hydrologic Modeling: Statistical Methods and Applications.* Prentice-Hall, Englewood Cliffs, NJ. 568 p.

Socolow, R.S., C. R. Leighton, J. L. Zanca, and L. R. Ramsey (1997) Water Resources Data, Massachusetts and Rhode Island: Water Year 1997. Water-Data Report MA-RI-97-1. U.S.G.S Water Resources Division, Marlborough, MA.

Tetra Tech (2009) Quality Assurance Project Plan for Watershed Modeling to Evaluate Potential Impacts of Climate and Land use Change on the Hydrology and Water Quality of Major U.S. Drainage Basins. Prepared for U.S. EPA Office of Research and Development Glogal Climate Research Program, Washington DC.

Thomann, R.V. (1980) Measures of Verification. In: Workshop on Verification of Water Quality Models. Edited by R.V. Thomann and T. O. Barnwell. EPA-600/9-80-016. U.S. EPA, ORD, Athens, GA. pp. 37-59.

Thomann, R.V. (1982) Verification of Water Quality Models. *Jour. Env. Engineering Div. (EED), Proc. ASCE, 108:EE5, October.*

Weston Solutions, Inc. (2004) Model Calibration: Modeling Study of PCB Contamination in the Housatonic River. Volume 2, Appendix A: Watershed Model Calibration . Prepared for U.S. Army Corps of Engineers and U.S. Environmental Protection Agency. DCN GE-100504-ACJS. Weston Solutions, Inc. West Chester, PA

Weston Solutions, Inc. (2006) Model Validation: Modeling Study of PCB Contamination in the Housatonic River. Volume 1. Section 6.1 HSPF Watershed Model Validation. Prepared for U.S. Army Corps of Engineers and U.S. Environmental Protection Agency. DCN GE-030706-ADBR. Weston Solutions, Inc. West Chester, PA

Wyss G. D. and Jorgensen K. H. (1998) *A User's Guide to LHS: Sandia's Latin Hypercube Sampling Software*, Report SAND98-0210, Sandia National Laboratories, Albuquerque, NM.

Zar, J.H. (1999) *Biostatistical Analysis*. 4th Edition. Prentice Hall, Upper Saddle River, NJ.