# WATERSHED MODEL CALIBRATION AND VALIDATION: THE HSPF EXPERIENCE

**by**
**A. S. Donigian, Jr.**
**AQUA TERRA Consultants**
**2685 Marine Way, Suite 1314**
**Mountain View, CA 94043**

Model calibration and validation are  necessary and critical steps in any model application. For most all watershed models, calibration is an iterative procedure of parameter evaluation and refinement, as a result of comparing simulated and observed values of interest. 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, and demonstrate the ability to predict field observations for periods separate from the calibration effort.

Model performance and calibration/validation are evaluated through 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.  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.

Model performance criteria, sometimes referred to as calibration or validation criteria, have been contentious topics for more than 20 years.  These issues have been recently 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.  A **'weight of evidence'** approach is most widely used in practice when models are examined and judged for acceptance for these purposes.

This paper explores the **'weight of evidence'** approach and the current practice of watershed model calibration and validation based on more than 20 years experience with the U. S. EPA Hydrological Simulation Program - FORTRAN (HSPF).  Example applications are described and model results are shown 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 model application efforts.

## MODEL CALIBRATION AND VALIDATION

The modeling, or model application, process can be described as comprised of three phases, as shown in Figure 1 (Donigian and Rao, 1990).  Phase I includes data collection, model input preparation, and parameter evaluation, i.e. all the steps needed to setup a model, characterize the watershed, and prepare for model executions. Phase II is the model testing phase which involves calibration, validation (or verification, both terms are used synonymously in this paper), and, when possible, post-audit. This is the phase in which the model is evaluated to assess whether it can reasonably represent the watershed behavior, for the purposes of the study. Phase III includes the ultimate use of the model, as a decision support tool for management and regulatory purposes.

Although specific application procedures for all watershed models differ due to the variations of the specific physical, chemical, and biological systems they each attempt to represent, they have many steps in common. The calibration and validation phase is especially critical since the outcome establishes how well the model represents the watershed, for the purpose of the study. Thus, this is the 'bottom-line' of the model application effort as it determines if the model results can be relied upon and used effectively for decision-making.

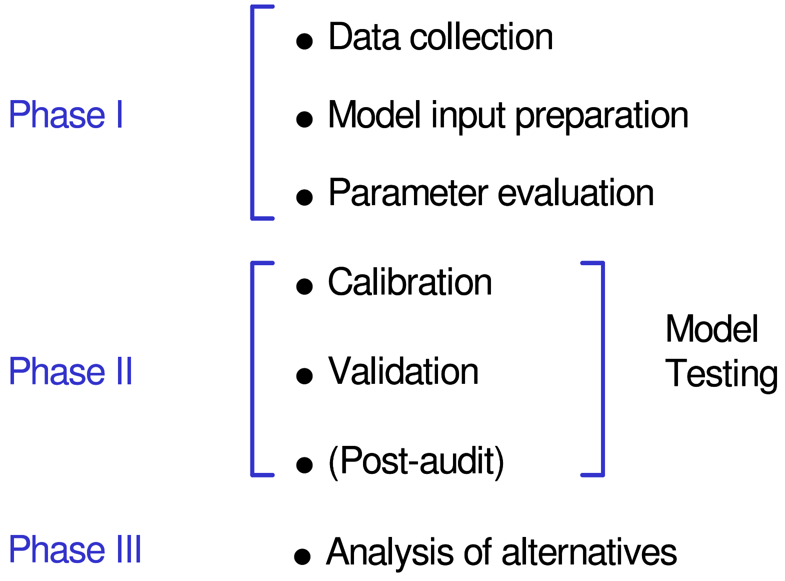


**Figure 1. The Modeling Process**

Calibration and validation have been defined by the American Society of Testing and Materials, as follows (ASTM, 1984):

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 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. This type of **split-sample calibration/validation** procedure is commonly used, and recommended, for many watershed modeling studies. Model credibility is based on the ability of a single set of parameters to represent the entire range of observed data. If a single parameter set can reasonably represent a wide range of events, then this is a form of validation.

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 2 schematically compares the model with the 'natural system', i.e. the watershed, and identifies various sources

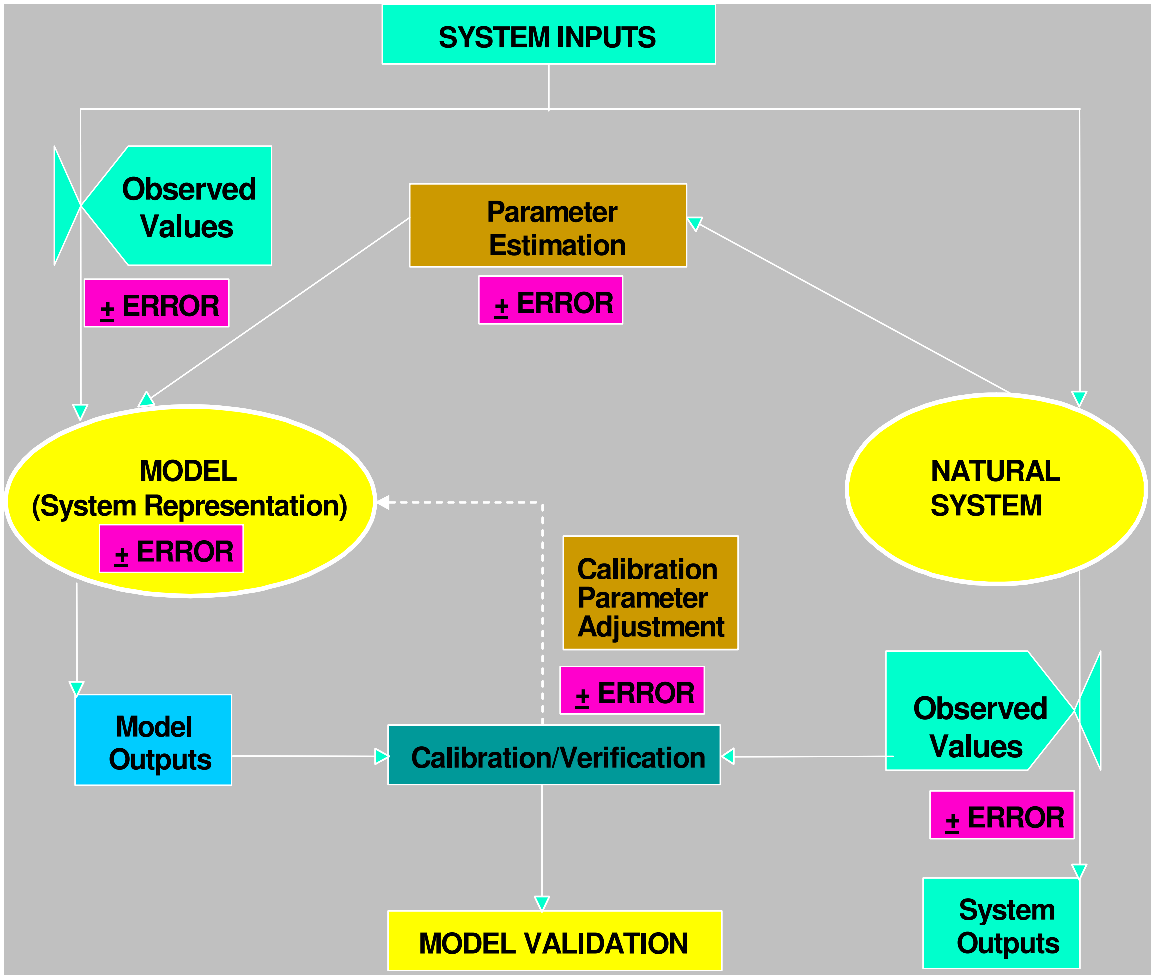


**Figure 2. Model versus Natural System: Inputs, Outputs, and Errors**

of potential errors to be investigated. These types of analysis requires 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).

Model performance, i.e. the ability to reproduce field observations, and calibration/validation are most often evaluated through both qualitative and quantitative measures, involving both graphical comparisons and statistical tests. For flow simulations where continuous records are available, all these techniques will be employed, and the same comparisons will be performed, during both the calibration and validation phases. Comparisons of simulated and observed state variables will be performed for daily, monthly, and annual values, in addition to flow-frequency duration assessments. Statistical procedures include error statistics, correlation and model-fit efficiency coefficients, and goodness-of-fit tests.

For sediment, water quality, and biotic constituents, model performance will be based primarily on visual and graphical presentations as the frequency of observed data is often inadequate for accurate statistical measures. However, we will investigate alternative model performance assessment techniques, e.g. error statistics and correlation measures, consistent with the population of observed data available for model testing.

## HSPF CALIBRATION AND VALIDATION PROCEDURES

Model application procedures for HSPF have been developed and described in the HSPF Application Guide (Donigian et al, 1984), in numerous watershed studies over the past 20 years (see HSPF Bibliography - Donigian, 2000), and most recently in HSPF applications to the Chesapeake Bay watershed (Donigian et al., 1994) and the Long Island Sound watersheds within the state of Connecticut (Love and Donigian, 2002). Model application procedures for HSPF include database development, watershed segmentation, and hydrology, sediment, and water quality calibration and validation.

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.

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. Each of these steps are described briefly below.

Since parameter evaluation is a key precursor to the calibration effort, a valuable source of initial starting values for many of the key calibration parameters is the recently-developed parameter database for HSPF, called HSPFParm (Donigian et al., 1999). HSPFParm is an interactive database (based on MS-Access) that includes calibrated parameter values for up to 45 watershed water quality studies across the U.S.

## Hydrologic Calibration

Hydrologic simulation combines the physical characteristics of the watershed and the observed meteorologic data series to produce the simulated hydrologic response. All watersheds have similar hydrologic components, but they are generally present in different combinations; thus different hydrologic responses occur on individual watersheds. HSPF simulates runoff from four components: surface runoff from impervious areas directly connected to the channel network, surface runoff from pervious areas, interflow from pervious areas, and groundwater flow. Since the historic streamflow is not divided into these four units, the relative relationship among these components must be inferred from the examination of many events over several years of continuous simulation.

A complete hydrologic calibration involves a successive examination of the following **four** characteristics of the watershed hydrology, in the following order: (1) annual water balance, (2) seasonal and monthly flow volumes, (3) baseflow, and (4) storm events. Simulated and observed values for each characteristic are examined and critical parameters are adjusted to improve or attain acceptable levels of agreement (discussed further below).

The annual water balance specifies the ultimate destination of incoming precipitation and is indicated as:

$$\begin{aligned}&\textbf{Precipitation - Actual Evapotranspiration - Deep Percolation}\\&\qquad\qquad\textbf{-}\ \boldsymbol{\Delta}\textbf{Soil Moisture Storage = Runoff}\end{aligned}$$

HSPF requires input precipitation and potential evapotranspiration (PET), which effectively ‘drive’ the hydrology of the watershed; actual evapotranspiration is calculated by the model from the input potential and ambient soil moisture conditions. Thus, both inputs must be accurate and representative of the watershed conditions; it is often necessary to adjust the input data derived from neighboring stations that may be some distance away in order to reflect conditions on the watershed. HSPF allows the use of spatial adjustment factors that uniformly adjust the input data to watershed conditions, based on local isohyetal, evaporation, and climatic patterns. Fortunately, evaporation does not vary as greatly with distance, and use of evaporation data from distant stations (e.g. 50 to 100 miles) is common practice.

In addition to the input meteorologic data series, the critical HSPF parameters that affect components of the **annual water balance** include soil moisture storages, infiltration rates, vegetal evapotranspiration, and losses to deep groundwater recharge (see the BASINS web site, www.epa.gov/ost/basins/bsnsdocs/html, for information on HSPF parameters, including Tech Note #6 which provides parameter estimation guidance).

Thus, from the water balance equation, if precipitation is measured on the watershed, and if deep percolation to groundwater is small or negligible, actual evapotranspiration must be adjusted to cause a change in the long-term runoff component of the water balance. Changes in soil moisture storages (e.g. LZSN in HSPF) and vegetation characteristics affect the actual evapotranspiration by making more or less moisture available to evaporate or transpire. Both soil moisture and infiltration parameters also have a major impact on percolation and are important in obtaining an annual water balance. In addition, on extremely small watersheds (less than 200-500 acres) that contribute runoff only during and immediately following storm events, surface detention and near-surface soil moisture storages can also affect annual runoff volumes because of their impact on individual storm events (described below). Whenever there are losses to deep groundwater, such as recharge, or subsurface flow not measured at the flow gage, the recharge parameters are used to represent this loss from the annual water balance.

In the next step in hydrologic calibration, after an annual water balance is obtained, the **seasonal or monthly distribution of runoff** can be adjusted with use of the infiltration parameter. This seasonal distribution is accomplished by dividing the incoming moisture among surface runoff, interflow, upper zone soil moisture storage, and percolation to lower zone soil moisture and groundwater storage. Increasing infiltration will reduce immediate surface runoff (including interflow) and increase the groundwater component; decreasing will produce the opposite result.

The focus of the next stage in calibration is the **baseflow component**. This portion of the flow is often adjusted in conjunction with the seasonal/monthly flow calibration (previous step) since moving runoff volume between seasons often means transferring the surface runoff from storm events in wet seasons to low flow periods during dry seasons; by increasing the infiltration parameter runoff is delayed and occurs later in the year as an increased groundwater or baseflow. The shape of the groundwater recession, i.e. the change in baseflow discharge, is controlled by the groundwater recession rate which controls the rate of outflow from the groundwater storage. Using hydrograph separation techniques, these values are often calculated as the slope of the receding baseflow portion of the hydrograph; these initial values are then adjusted as needed through calibration.

In the final stage of hydrologic calibration, after an acceptable agreement has been attained for annual/monthly volumes and baseflow conditions, simulated **hydrographs for selected storm events** can be effectively altered by adjusting surface detention and interflow parameters. These parameters are used to adjust the shape of the hydrograph to better agree with observed values; both parameters are evaluated primarily from past experience and modeling studies, and then adjusted in calibration. Also, minor adjustments to the infiltration parameter can be used to improve simulated hydrographs. Examination of both daily and short time interval (e.g. hourly or 15-minute) flows may be included depending on the purpose of the study and the available data.

In addition to the above comparisons, the water balance components (input and simulated) should be reviewed for consistency with expected literature values for the study watershed. This effort involves displaying model results for individual land uses for the following water balance components:

- ➢ Precipitation
- ➢ Total Runoff (sum of following components)
  - Overland flow
  - Interflow

  - Baseflow
- ➢ Total Actual Evapotranspiration (ET) (sum of following components)
  - Interception ET
  - Upper zone ET
  - Lower zone ET
  - Baseflow ET
  - Active groundwater ET
- ➢ Deep Groundwater Recharge/Losses

Although observed values are not be available for each of the water balance components listed above, the average annual values must be consistent with expected values for the region, as impacted by the individual land use categories. This is a separate consistency, or reality, check with data independent of the modeling (except for precipitation) to insure that land use categories and overall water balance reflect local conditions.

In recent years, the hydrology calibration process has been facilitated with the aide of HSPEXP, an expert system for hydrologic calibration, specifically designed for use with HSPF, developed under contract for the U. S. Geological Survey (Lumb, McCammon, and Kittle, 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.

**Snow Calibration**

Since snow accumulation and melt is an important component of streamflow in many climates, accurate simulation of snow depths and melt processes is needed to successfully model the hydrologic behavior of the watershed. Snow calibration is actually part of the hydrologic calibration. It is usually performed during the initial phase of the hydrologic calibration since the snow simulation can impact not only winter runoff volumes, but also spring and early summer streamflow.

Simulation of snow accumulation and melt processes suffers from two main sources of user-controlled uncertainty: representative meteorologic input data and parameter estimation. The additional meteorologic time series data required for snow simulation (i.e. air temperature, solar radiation, wind, and dewpoint temperature) are not often available in the immediate vicinity of the watershed, and consequently must be estimated or extrapolated from the nearest available weather station. Snowmelt simulation is especially sensitive to the air temperature and solar radiation time series since these are the major driving forces for the energy balance melt calculations. Fortunately, additional nearby stations are available with air temperature data. The spatial adjustment factors, noted above, is used to adjust each of the required input meteorologic data to more closely represent conditions on the watershed; also, the model allows an internal correction for air temperature as a function of elevation, using a ‘lapse’ rate that specifies the change in temperature for any elevation difference between the watershed and the temperature gage.

In most applications the primary goal of the snow simulation will be to adequately represent the total volume and relative timing of snowmelt to produce reasonable soil moisture conditions in the spring and early summer so that subsequent rainfall events can be accurately simulated. Where observed snow depth (and water equivalent) measurements are available, comparisons with simulated values should be made. However, a tremendous variation in observed snow depth values can occur in a watershed, as a function of elevation, exposure, topography, etc. Thus a single observation point or location will not always be representative of the watershed average. See the BASINS Tech Note #6 for discussion and estimation of the snow parameters.

In many instances it is difficult to determine if problems in the snow simulation are due to the non-representative meteorologic data or inaccurate parameter values. Consequently the accuracy expectations and general objectives of snow calibration are not as rigorous as for the overall hydrologic calibration. Comparisons of simulated weekly and monthly runoff volumes with observed streamflow during snowmelt periods, and observed snow depth (and water equivalent) values are the primary procedures followed for snow calibration. Day-to-day variations and comparisons on shorter intervals (i.e. 2-hour, 4-hour, 6-hour, etc.) are usually not as important as representing the overall snowmelt volume and relative timing in the observed weekly or bi-weekly period.

## Hydraulic Calibration

The major determinants of the routed flows simulated by the hydraulics section of HSPF, section HYDR are the hydrology results providing the inflows from the local drainage, inflows from any upstream reaches, and the physical data contained in the FTABLE, which is the stage-discharge function used for hydraulic routing in each stream reach. The FTABLE specifies values for surface area, reach volume, and discharge for a series of selected average depths of water in each reach. This information is part of the required model input and is obtained from cross-section data, channel characteristics (e.g. length, slope, roughness), and flow calculations. Since the FTABLE is an approximation of the stage-discharge-volume relationship for relatively long reaches, calibration of the values in the FTABLE is generally not needed. However, if flows and storage volumes at high flow conditions appear to be incorrect, some adjustments may be needed. Since HSPF cannot represent bi-directional flow, e.g. estuaries, linkage with hydrodynamic models is often needed to simulate tidal conditions and flow in rivers and streams with extremely flat slopes.

## Sediment Erosion Calibration

Sediment calibration follows the hydrologic calibration and must precede water quality calibration. Calibration of the parameters involved in simulation of watershed sediment erosion is more uncertain than hydrologic calibration due to less experience with sediment simulation in different regions of the country. The process is analogous; the major sediment parameters are modified to increase agreement between simulated and recorded monthly sediment loss and storm event sediment removal. However, observed monthly sediment loss is often not available, and the sediment calibration parameters are not as distinctly separated between those that affect monthly sediment and those that control storm sediment loss. In fact, annual sediment losses are often the result of only a few major storms during the year.

Sediment loadings to the stream channel are estimated by land use category from literature data, local Extension Service sources, or procedures like the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1972) and then adjusted for delivery to the stream with estimated sediment delivery ratios. Model parameters are then adjusted so that model calculated loadings are consistent with these estimated loading ranges. The loadings are further evaluated in conjunction with instream sediment transport calibration (discussed below) that extend to a point in the watershed where sediment concentration data is available. The objective is to represent the overall sediment behavior of the watershed, with knowledge of the morphological characteristics of the stream (i.e. aggrading or degrading behavior), using sediment loading rates that are consistent with available values and providing a reasonable match with instream sediment data. Recently a spreadsheet tool, TMDLUSLE, has been developed based on the USLE for estimating sediment loading rates as target values for calibration, and as a tool for sediment TMDL development; it is available from the U.S EPA website (www.epa.gov/ceampubl/tmdlusle.htm).

## Instream Sediment Transport Calibration

Once the sediment loading rates are calibrated to provide the expected input to the stream channel, the sediment calibration then focuses on the channel processes of deposition, scour , and transport that determine both the total sediment load and the outflow sediment concentrations to be compared with observations. Although the sediment load from the land surface is calculated in HSPF as a total input, it must be divided into sand, silt, and clay fractions for simulation of instream processes. Each sediment size fraction is simulated separately, and storages of each size are maintained for both the water column (i.e. suspended sediment) and the bed.

In HSPF, the transport of the **sand (non-cohesive) fraction** is commonly calculated as a power function of the average velocity in the channel reach in each timestep. This transport capacity is compared to the available inflow and storage of sand particles; the bed is scoured if there is excess capacity to be satisfied, and sand is deposited if the transport capacity is less than the available sand in the channel reach. For the **silt and clay (i.e. non-cohesive) fractions**, shear stress calculations are performed by the hydraulics (HYDR) module and are compared to user-defined critical, or threshold, values for deposition and scour for each size. When the shear stress in each timestep is greater than the critical value for scour, the bed is scoured at a user-defined erodibility rate; when the shear stress is less than the critical deposition value, the silt or clay fraction deposits at a settling rate input by the user for each size. If the calculated shear stress falls between the critical scour and deposition values, the suspended material is transported through the reach. After all scour and/or deposition fluxes have been determined, the bed and water column storages are updated and outflow concentrations and fluxes are calculated for each timestep. These simulations are performed by the SEDTRN module in HSPF, complete details of which are provided in the HSPF User Manual (Bicknell et a., 1997; 2001).

In HSPF, sediment transport calibration involves numerous steps in determining model parameters and appropriate adjustments needed to insure a reasonable simulation of the sediment transport and behavior of the channel system. These steps are usually as follows:

1. Divide input sediment loads into appropriate size fractions

2. Run HSPF to calculate shear stress in each reach to estimate critical scour and deposition values
3. Estimate initial parameter values and storages for all reaches
4. Adjust scour, deposition and transport parameters to impose scour and deposition conditions at appropriate times, e.g. scour at high flows, deposition at low flows
5. Analyze sediment bed behavior and transport in each channel reach
6. Compare simulated and observed sediment concentrations, bed depths, and particle size distributions, where available
7. Repeat steps 1 through 5 as needed

Rarely is there sufficient observed local data to accurately calibrate all parameters for each stream reach. Consequently, model users focus the calibration on sites with observed data and review simulations in all parts of the watershed to insure that the model results are consistent with field observations, historical reports, and expected behavior from past experience. Ideally comprehensive datasets available for storm runoff should include both tributary and mainstem sampling sites. Observed storm concentrations of TSS should be compared with model results, and the sediment loading rates by land use category should be compared with the expected targets and ranges, as noted above.

## Nonpoint Source Loading and Water Quality Calibration

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 following steps are usually performed at each of the calibration stations, following the hydrologic calibration and validation, and after the completion of input development for point source and atmospheric contributions:

1. Estimate all model parameters, including land use specific accumulation and depletion/removal rates, washoff rates, and subsurface concentrations
2. Superimpose the hydrology and tabulate, analyze, and compare simulated nonpoint loadings with expected range of nonpoint loadings from each land use and adjust loading parameters when necessary to improve agreement and consistency
3. Calibrate instream water temperature
4. Compare simulated and observed instream concentrations at each of the calibration stations
5. Analyze the results of comparisons in steps 3 and 4 to determine appropriate instream and/or nonpoint parameter adjustments, and repeat those steps as needed until calibration targets are achieved; Watershed loadings are adjusted when the instream simulated and observed concentrations are not in full agreement, and instream parameters have been adjusted throughout the range determined reasonable

Calibration procedures and parameters for simulation of nonpoint source pollutants will vary depending on whether constituents are modeled as sediment-associated or flow-associated. This refers to whether the loads are calculated as a function of sediment loadings or as a function of the overland flow rate. Due to their affinity for sediment, contaminants such metals, toxic organics, and phosphorous are usually modeled as sediment-associated, whereas BOD, nitrates, ammonia, and bacteria are often modeled as flow-associated.

Calibration of **sediment-associated pollutants** begins after a satisfactory calibration of sediment washoff has been completed. At this point, adjustments are performed in the contaminant **potency factors**, which are user-specified parameters for each contaminant. Potency factors are used primarily for highly sorptive contaminants that can be assumed to be transported with the sediment in the runoff. Generally, monthly and annual contaminant loss will not be available, so the potency factors will be adjusted by comparing simulated and recorded contaminant concentrations, or mass removal, for selected storm events. For nonpoint pollution, mass removal in terms of contaminant mass per unit time (e.g., gm/min) is often more indicative of the washoff and scour mechanisms than instantaneous observed contaminant concentrations.

Calibration procedures for simulation of **contaminants associated with overland flow** are focused on the adjustment of parameters relating to daily accumulation rates (lb/acre/day), accumulation limits (lb/acre), and washoff parameters (in/hr). As was the case for sediment-associated constituents, calibration is performed by comparing simulated and recorded contaminant concentrations, or mass removal, for selected storm events. In most cases, proper adjustment of corresponding parameters can be accomplished to provide a good representation of the washoff of flow-associated constituents. The HSPF Application Guide (Donigian et al., 1984) includes guidelines for calibration of these parameters, and the HSPFParm Database includes representative values for selected model applications for most conventional constituents.

In study areas where pollutant contributions are also associated with subsurface flows, contaminant concentration values are assigned for both interflow and active groundwater. The key parameters are simply the user- defined concentrations in interflow and groundwater/baseflow for each contaminant. HSPF includes the functionality to allow monthly values for all nonpoint loading parameters in order to better represent seasonal variations in the resulting loading rates.

In studies requiring detailed assessment of agricultural or forested runoff water quality for nutrients or pesticides, the mass balance soil module within HSPF, referred to as AGCHEM may need to be applied. Model users should consult the HSPF User Manual, the Application guide, and recent studies by Donigian et al (1998a, 1998b) that discuss application, input development, and calibration procedures.

Instream HSPF water quality calibration procedures are highly dependent on the specific constituents and processes represented, and in many ways, water quality calibration is equal parts art and science. As noted above, the goal 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 specific model parameters to be adjusted depend on the model options selected and constituents being modeled, e.g. BOD decay rates, reaeration rates, settling rates, algal growth rates, temperature correction factors, coliform die-off rates, adsorption/desorption coefficients, etc. Part of the 'art' of water quality calibration, is assessing the interacting effects of modeled quantities, e.g. algal growth on nutrient uptake, and being able to analyze multiple timeseries plots jointly to determine needed parameter adjustments. The HSPF Application Guide and other model application references noted above are useful sources of information on calibration practices, along with model parameter compendiums published in the literature (e.g. Bowie et al., 1985).

## MODEL PERFORMANCE CRITERIA

Model performance criteria, sometimes referred to as calibration or validation criteria, have been contentious topics for more than 20 years (see Thomann, 1980; Thomann, 1982; James and Burges, 1982; Donigian, 1982; ASTM, 1984). These issues have been recently 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. Recently (September 1999), an EPA-sponsored workshop entitled "Quality Assurance of Environmental Models" convened in Seattle, WA to address the issues related to problems of model assessment and quality assurance, development of methods and techniques, assurance of models used in regulation, and research and practice on model assurance (see the following web site for details: http://www.nrcse.washington.edu/events/qaem/qaem.asp). This workshop spawned a flurry of web-based activity among a group of more than 50 recognized modeling professionals (both model developers and users) in various federal and state agencies, universities, and consulting firms that clearly confirms the current lack of consensus on this topic.

Although no 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 can not 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. 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. A **'weight of evidence'** approach is most widely used and accepted when models are examined and judged for acceptance for these purposes. 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 all watersheds, and other environmental systems, models will utilize different types of graphical and statistical procedures, they will generally include some 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.

**Table 1 General Calibration/Validation Targets or Tolerances for HSPF Applications (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)

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.

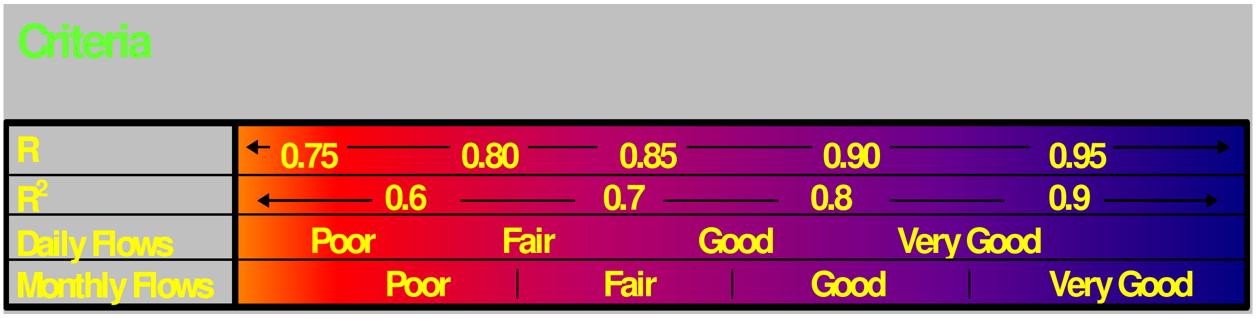


**Figure 3. R and R$^2$ Value Ranges for Model Performance**

Figure 3 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 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 HSPF CALIBRATION/VALIDATION COMPARISONS

This section presents results from recent HSPF applications, (1) to the State of Connecticut for nutrient loadings to Long Island Sound, and (2) to an Unnamed Watershed in the Northeastern U. S. 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 2) 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 2 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, evapotransipiration) 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, McCammon, and Kittle, 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.

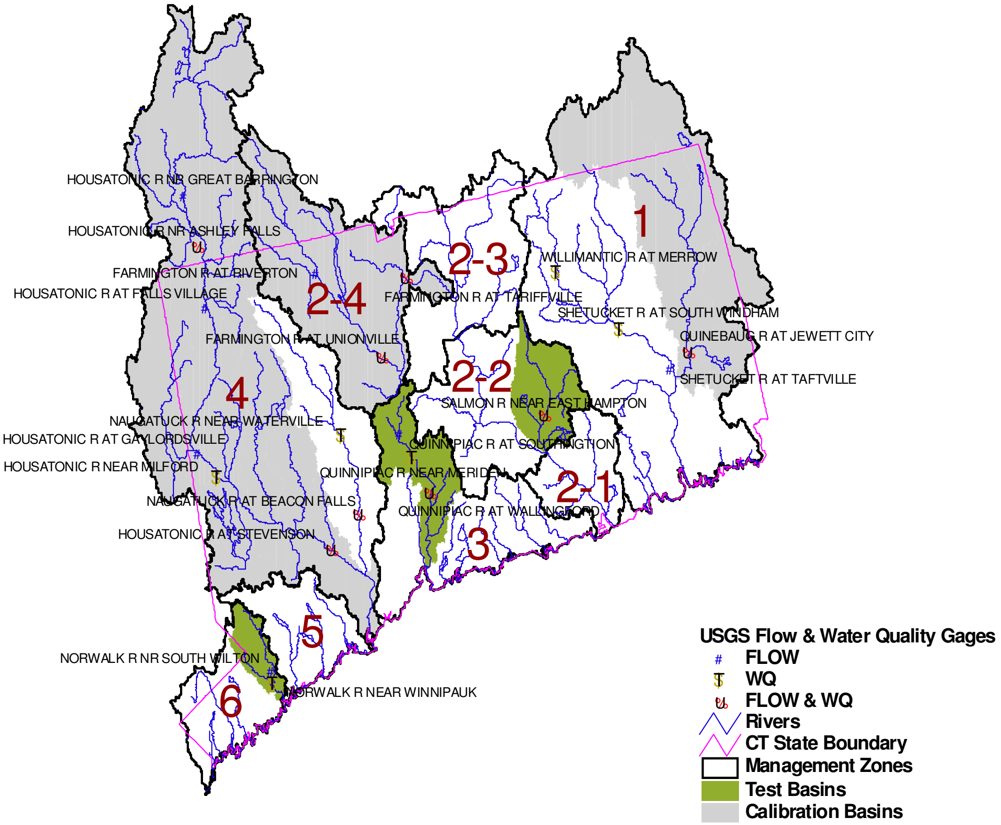

HOUSATONIC R NR GREAT BARRINGTON
HOUSATONIC R NR ASHLEY FALLS
FARMINGTON R AT RIVERTON
HOUSATONIC R AT FALLS VILLAGE
2-4
FARMINGTON R AT TARIFFVILLE
2-3
1
WILLIMANTIC R AT MERROW
SHETUCKET R AT SOUTH WINDHAM
QUINEBAUG R AT JEWETT CITY
SHETUCKET R AT TAFTVILLE
FARMINGTON R AT UNIONVILLE
2-2
SALMON R NEAR EAST HAMPTON
4
NAUGATUCK R NEAR WATERVILLE
HOUSATONIC R AT GAYLORDSVILLE
HOUSATONIC R NEAR MILFORD
QUINNIPIAC R AT SOUTHINGTON
QUINNIPIAC R NEAR MERIDEN
NAUGATUCK R AT BEACON FALLS
2-1
QUINNIPIAC R AT WALLINGFORD
HOUSATONIC R AT STEVENSON
3
5
NORWALK R NR SOUTH WILTON
NORWALK R NEAR WINNIPAUK
6
USGS Flow & Water Quality Gages
FLOW
WQ
FLOW & WQ
Rivers
CT State Boundary
Management Zones
Test Basins
Calibration Basins

**Figure 4. USGS flow and water quality gages for the CTWM**

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 correlation daily and monthly 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 for a very good agreement, and always within the 15% target for a good agreement.

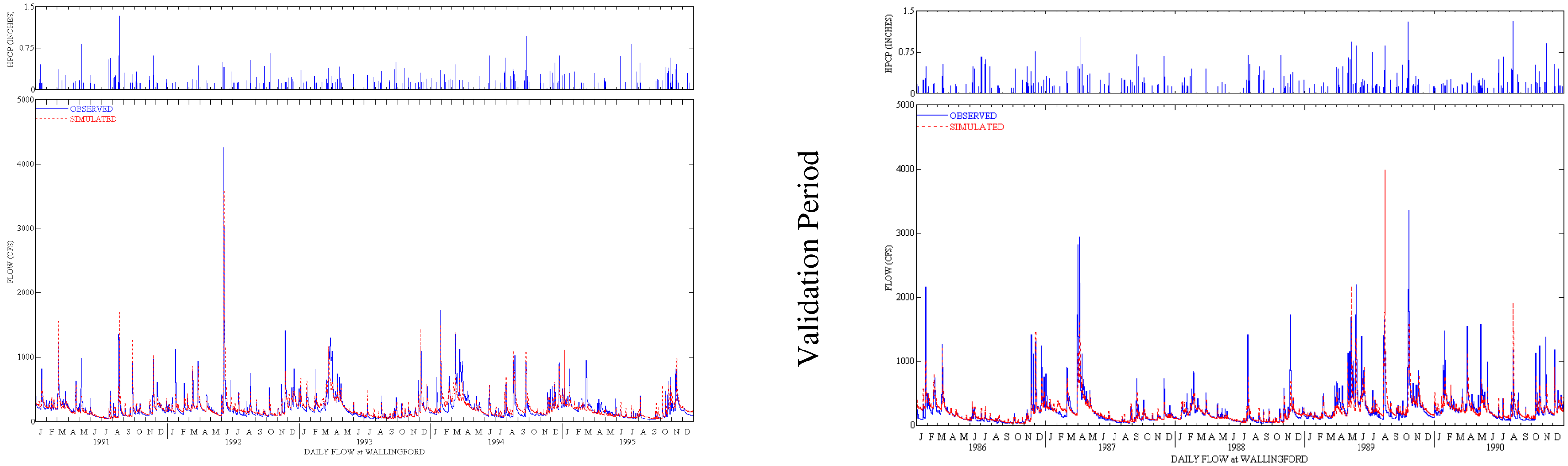
**Table 2**
**Summary of CTWM hydrologic calibration/validation - annual flow and correlation coefficients**

| | | 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 |

Figures 5 and 6 present graphical comparisons of simulated and observed daily flows for the Quinnipiac River at Wallingford and the Farmington River at Tariffville, respectively. Figures 7 and 8 show flow duration plots for the same sites. Figures 9 and 10 show the scatterplots for daily flows at the Farmington gage for both the calibration and validation periods.

**Figure 5.** **Observed and Simulated Daily Flow for the Quinnipiac River at Wallingford - Calibration and Validation**

(Top curves are Daily Precipitation)

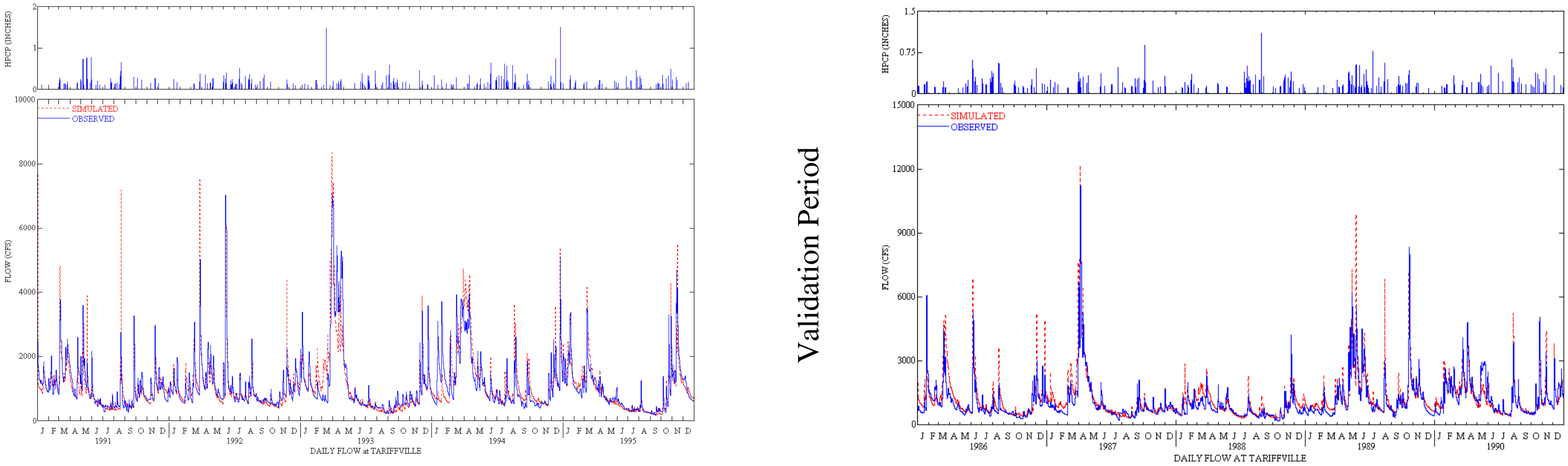


Calibration Period

Validation Period

**Figure 6.** **Observed and Simulated Daily Flow for the Farmington River at Tariffville - Calibration and Validation**

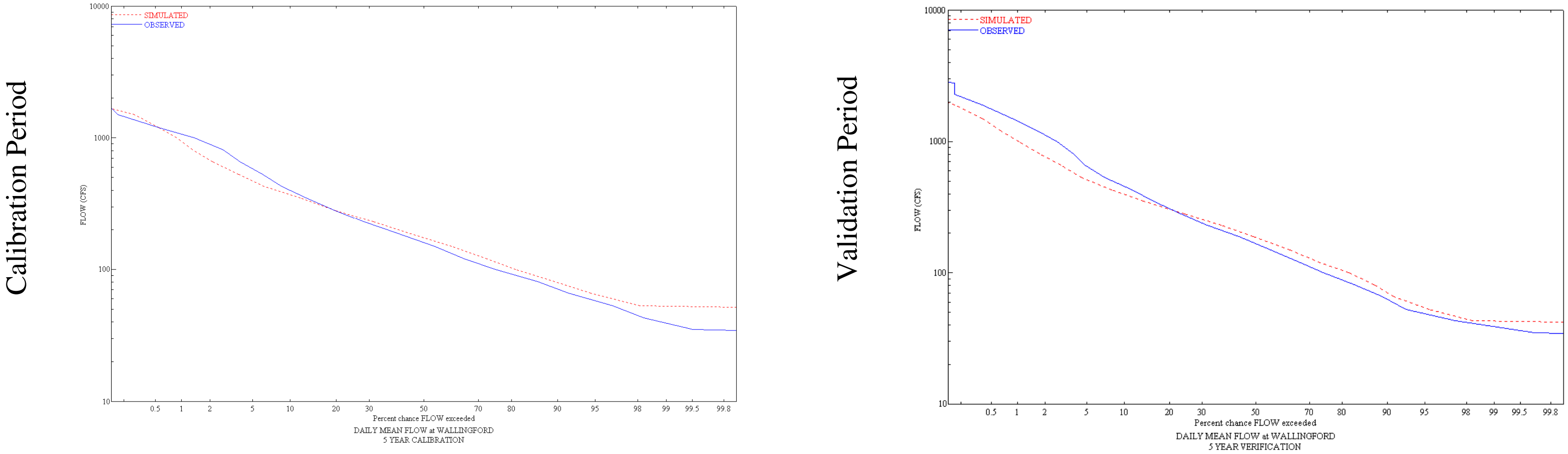
(Top curves are Daily Precipitation)



Calibration Period

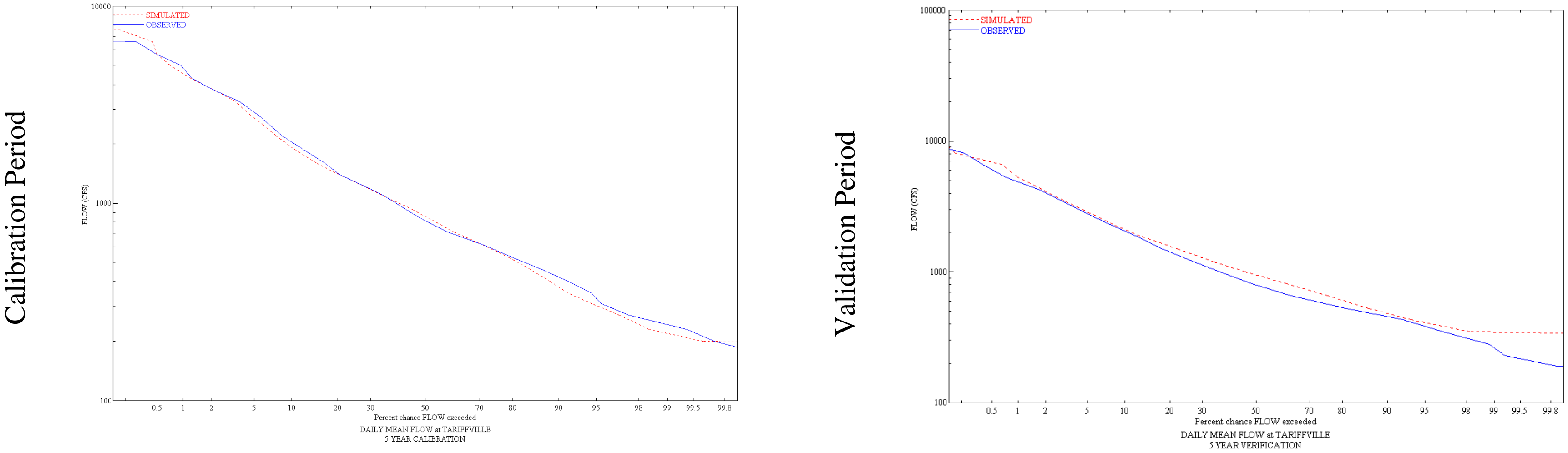
Validation Period

**Figure 7. Observed and Simulated Daily Flow Duration Curves for the Quinnipiac River at Wallingford - Calibration and Validation**


Calibration Period

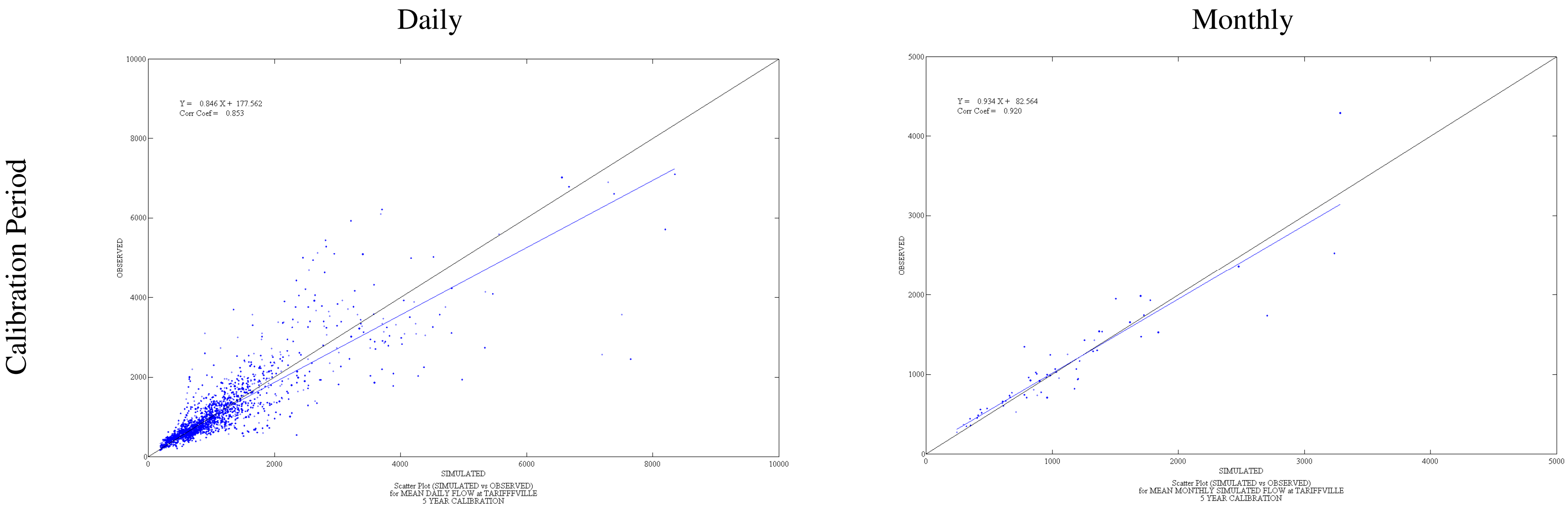
Validation Period

**Figure 8. Observed and Simulated Daily Flow Duration Curves for the Farmington River at Tariffville - Calibration and Validation**


Calibration Period

Validation Period

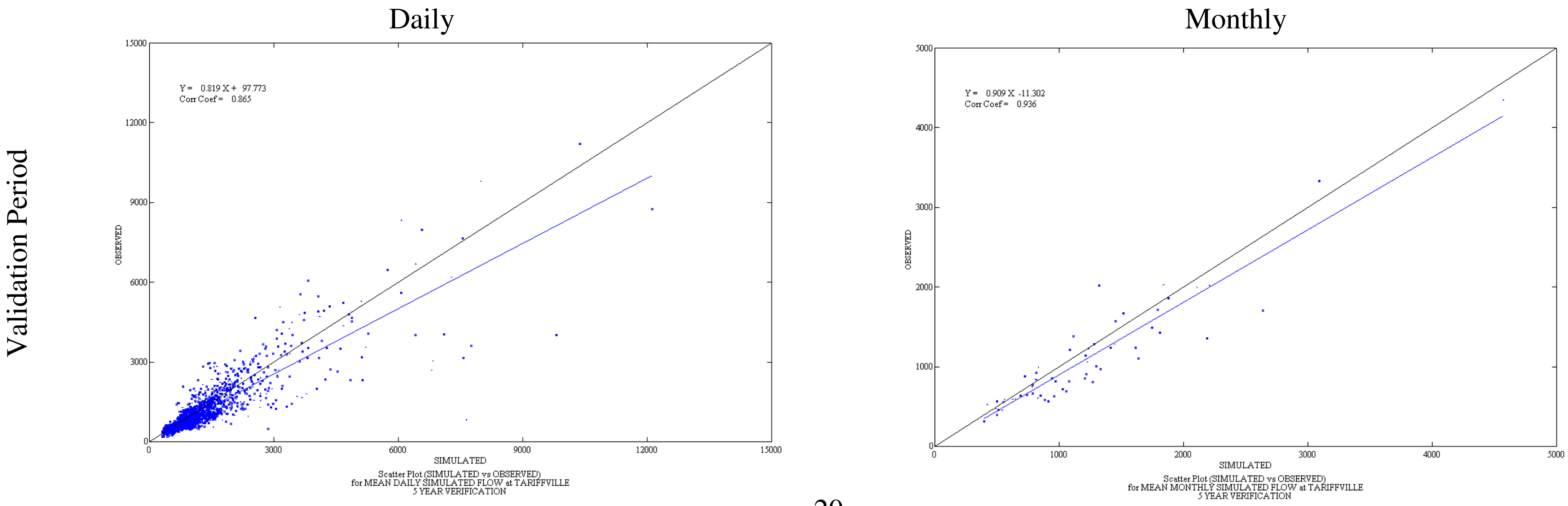
**Figure 9. Scatterplots of Observed and Simulated Daily and Monthly Flow for the Farmington River at Tariffville**



Daily

Monthly

**Figure 10. Scatterplots of Observed and Simulated Daily and Monthly Flow for the Farmington River at Tariffville**



Daily

Monthly

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 this 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 |

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.

### 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<br>(48) | 10.4 | 10.3 | 0.99<br>(46) | 11.6 | 10.4 | 0.90<br>(97) | 10.4 | 10.3 | 0.99<br>(43) | 10.2 | 10.8 | 1.06<br>(49) | 9.5 | 9.5 | 1.01<br>(41) |
| Ammonia as N | 0.03 | 0.02 | 0.82<br>(43) | 0.19 | 0.18 | 0.92<br>(46) | 0.04 | 0.04 | 1.18<br>(80) | 0.08 | 0.06 | 0.73<br>(42) | 0.10 | 0.09 | 0.82<br>(48) | 0.06 | 0.06 | 1.10<br>(33) |
| Nitrite-Nitrate as N | 0.22 | 0.27 | 1.21<br>(46) | 2.82 | 2.45 | 0.87<br>(46) | 0.39 | 0.40 | 1.03<br>(93) | 0.44 | 0.37 | 0.84<br>(42) | 0.71 | 0.59 | 0.83<br>(49) | 0.36 | 0.41 | 1.15<br>(40) |
| Organic Nitrogen | 0.31 | 0.25 | 0.80<br>(30) | 0.50 | 0.60 | 1.20<br>(44) | 0.33 | 0.28 | 0.86<br>(70) | 0.45 | 0.39 | 0.86<br>(40) | 0.31 | 0.28 | 0.90<br>(45) | 0.33 | 0.28 | 0.84<br>(38) |
| Total Nitrogen | 0.53 | 0.51 | 0.97<br>(30) | 3.64 | 3.29 | 0.90<br>(44) | 0.73 | 0.69 | 0.94<br>(70) | 0.96 | 0.80 | 0.83<br>(40) | 1.15 | 0.97 | 0.85<br>(45) | 0.77 | 0.75 | 0.97<br>(38) |
| Orthophosphate as P | 0.01 | 0.01 | 0.91<br>(48) | 0.32 | 0.36 | 1.10<br>(46) | 0.02 | 0.02 | 0.93<br>(94) | 0.02 | 0.04 | 1.67<br>(43) | 0.07 | 0.13 | 1.90<br>(49) | 0.01 | 0.02 | 1.49<br>(32) |
| Organic Phosphorus | 0.02 | 0.02 | 1.30<br>(48) | 0.07 | 0.11 | 1.62<br>(46) | 0.02 | 0.03 | 1.18<br>(94) | 0.03 | 0.04 | 1.23<br>(43) | 0.03 | 0.05 | 1.59<br>(49) | 0.02 | 0.03 | 1.19<br>(33) |
| Total Phosphorus | 0.02 | 0.03 | 1.35<br>(48) | 0.39 | 0.47 | 1.19<br>(46) | 0.04 | 0.05 | 1.10<br>(94) | 0.06 | 0.08 | 1.44<br>(43) | 0.10 | 0.18 | 1.82<br>(49) | 0.03 | 0.05 | 1.47<br>(40) |
| Total Organic Carbon | 3.9 | 2.8 | 0.71<br>(45) | 4.5 | 4.8 | 1.06<br>(44) | 4.0 | 3.2 | 0.81<br>(28) | 5.6 | 4.9 | 0.86<br>(41) | 3.9 | 3.3 | 0.84<br>(45) | 3.8 | 2.9 | 1.06<br>(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 |

Figures 11 and 12 present typical graphical comparisons made for simulated and observed water quality constituents. Figure 11 presents a comparison of simulated and observed total phosphorus for the Quinnipiac River at Wallingford. Figure 12 presents a similar comparison for total nitrogen at the Tariffville gage on the Farmington River.

**Figure 11 Observed and Simulated Daily Total Phosphorus Concentrations for the Quinnipiac River at Wallingford**

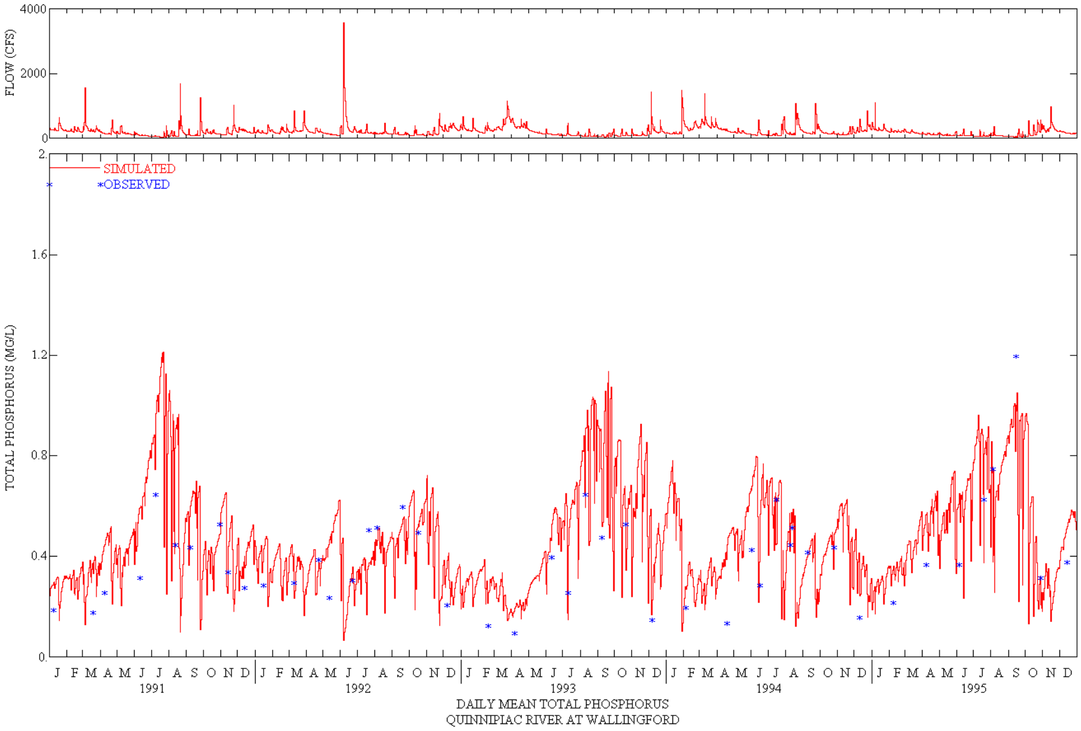
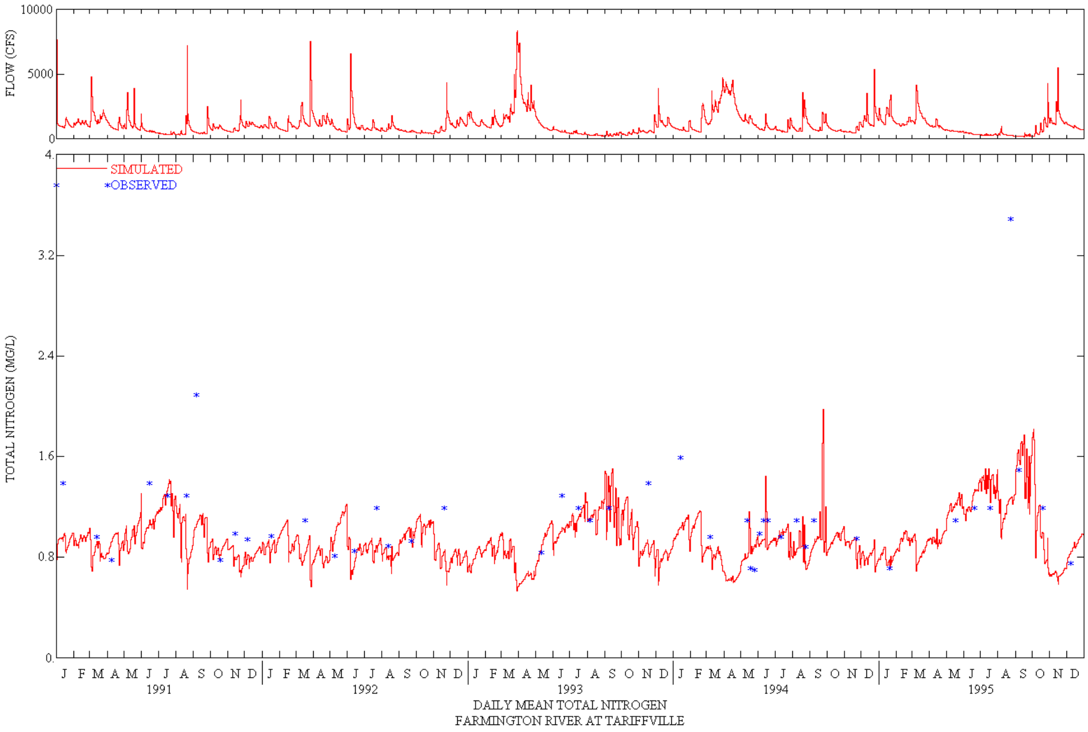


**Figure 12 Observed and Simulated Daily Total Nitrogen Concentrations for the Farmington River at Tariffville**



## 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.

## Unnamed Northeastern U. S. Watershed

HSPF is currently being applied to an almost 300 sq. mi. watershed in the Northeastern U. S. 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)**

| | Unnamed 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 | 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% |

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 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,

## 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

## 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 | |

and a 'very good' calibration of monthly flows, when compared to the value ranges in Figure 3.

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 13 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% |

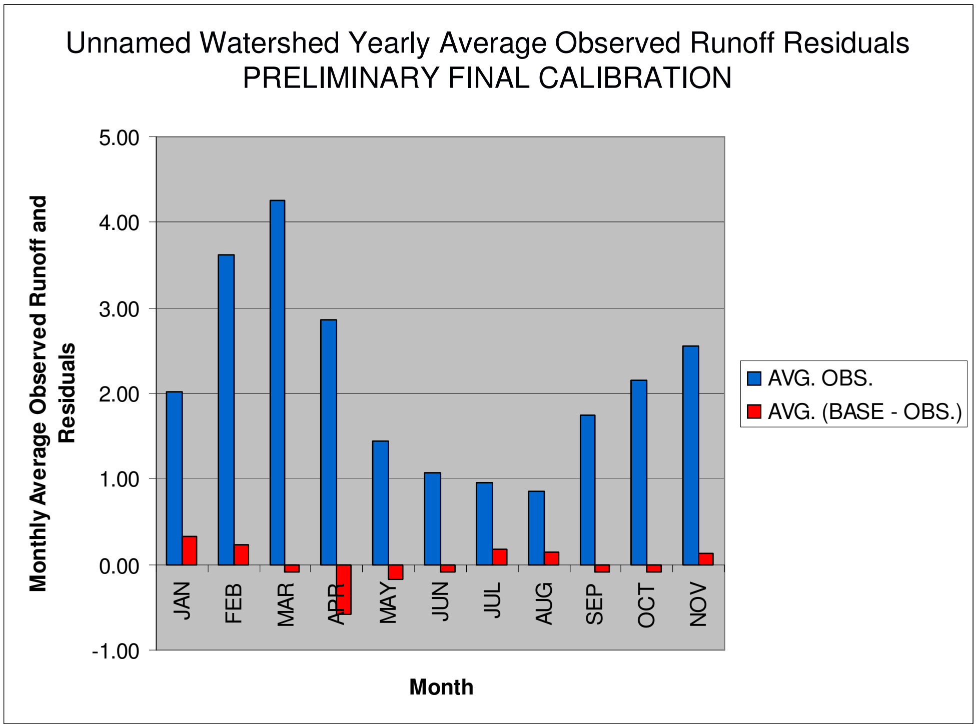


**Figure 13 Unnamed Watershed Observed Runoff and Residuals (inches)**

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** |

## CLOSURE

This paper has focused on presenting a ‘Weight-of-Evidence’ approach to watershed model calibration and validation 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.

## REFERENCES

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

AQUA TERRA Consultants and HydroQual, In., 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.

Bicknell, B.R., J.C. Imhoff, J.L. Kittle Jr., A.S. Donigian, Jr, and R.C. Johanson. 1997. Hydological Simulation Program - FORTRAN, User's Manual for Version 11. EPA/600/R-97/080. U.S. EPA, National Exposure Research Laboratory, Athens, GA. 763 p.

Bicknell, B.R., J.C. Imhoff, J.L. Kittle Jr., A.S. Donigian, Jr., T.H. Jobes, and R.C. Johanson. 2001. Hydological Simulation Program - FORTRAN, User's Manual for Version 12. U.S. EPA, National Exposure Research Laboratory, Athens, GA.

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

Bowie, et al. 1985. Rates, Constants, and Kinetics Formulations in Surface Water Quality Modeling (2$^{nd}$ Addition). EPA/600/3-85/040. U. S. Environmental Protection Agency, Washington D. C.

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, A.S. Jr., J.C. Imhoff, B.R. Bicknell and J.L. Kittle. 1984. Application Guide for Hydrological Simulation Program - Fortran (HSPF), prepared for U.S. EPA, EPA-600/3-84-065, Environmental Research Laboratory, Athens, GA.

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

Donigian, A.S. Jr., B.R. Bicknell, A.S. Patwardhan, L.C. Linker, C.H. Chang, and R. Reynolds. 1994. Chesapeake Bay Program - Watershed Model Application to Calculate Bay Nutrient Loadings: Final Findings and Recommendations (FINAL REPORT). Prepared for U.S. EPA Chesapeake Bay Program, Annapolis, Maryland.

Donigian, A.S. Jr., B. R. Bicknell, and J.C. Imhoff. 1995. Chapter 12. Hydrological Simulation Program - FORTRAN. In: Computer Models of Watershed Hydrology. V.P. Singh (ed). Water Resources Publications, Highland Ranch, CO. pp. 395-442.

Donigian, A.S. Jr., B. R. Bicknell, R.V. Chinnaswamy, and P. N Deliman. 1998a. Refinement of a Comprehensive Watershed Water Quality Model. Final Report. Prepared for U.S. Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS. Technical Report EL-98-6. 244 p.

Donigian, A.S. Jr., R.V. Chinnaswamy, and P. N Deliman. 1998b. Use of Nutrient Balances in Comprehensive Watershed Water Quality Modeling of Chesapeake Bay. Final Report. Prepared for U.S. Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS. Technical Report EL-98-5. 118 p.

Donigian, Jr., A. S., J. C. Imhoff, and J. L. Kittle, Jr. 1999. HSPFParm, An Interactive Database of HSPF Model Parameters. Version 1.0. EPA-823-R-99-004. Prepared for U. S. EPA, Office of Science and Technology, Washington, D. C. 39 p.

Donigian, A.S. Jr. 2000. Bibliography of HSPF and Related References. AQUA TERRA Consultants, Mountain View, CA.

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.

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

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.

Kittle, J.L. Jr., A. M. Lumb, P.R. Hummel, P.B. Duda, and M.H. Gray. 1998. A Tool for the Generation and Analysis of Model Simulation Scenarios for Watersheds (GenScn). Water Resources Investigation Report 98-4134. U.S. Geological Survey, Reston, VA. 152 p.

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.

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.

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.

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.*

U. S. EPA. 1997. Technical Guidance Manual for Developing Total Maximum Daily Loads – Book 2: Streams and Rivers, Part 1: Biochemcial Oxygen Demand/Dissolved Oxygen and Nutrients/Eutrophication. U. S. EPA Office of Water, Washington, D. C.

Wischmeier, W. H. and D. D. Smith. 1972. Predicting Rainfall Erosion Losses – A Guide to Conservation Planning. U. S. Dept. of Agriculture. Agriculture Handbook No. 537. Washington, D. C. 58 p.

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