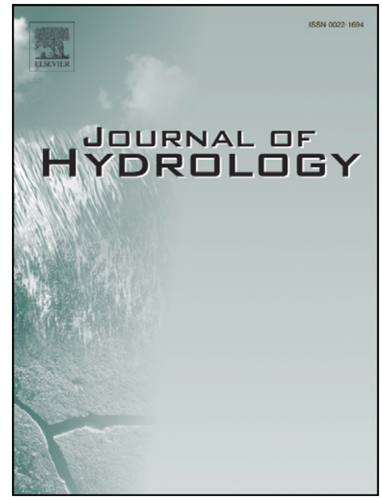
Accepted Manuscript



Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging

Xuesong Zhang, Raghavan Srinivasan, David Bosch

PII: S0022-1694(09)00354-0
DOI: 10.1016/j.jhydrol.2009.06.023
Reference: HYDROL 16642

To appear in: *Journal of Hydrology*

Received Date: 12 December 2008
Revised Date: 15 April 2009
Accepted Date: 16 June 2009

Please cite this article as: Zhang, X., Srinivasan, R., Bosch, D., Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging, *Journal of Hydrology* (2009), doi: 10.1016/j.jhydrol.2009.06.023

Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging

Xuesong Zhang[1], Raghavan Srinivasan[2], David Bosch[3]

[1] Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD 20740, USA

Xuesongzhang2004@gmail.com

[2] Spatial Sciences Laboratory, Department of Ecosystem Sciences and Management, Texas A&M University, College Station, TX 77843, USA.

[3] Southeast Watershed Research Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Tifton, GA 31793, USA.

Abstract In this paper, the Genetic Algorithms (GA) and Bayesian model averaging (BMA) were used to simultaneously conduct calibration and uncertainty analysis for the Soil and Water Assessment Tool (SWAT). In this combined method, several SWAT models with different structures are first selected; next GA is used to calibrate each model using observed streamflow data; finally, BMA is applied to combine the ensemble predictions and provide uncertainty interval estimation. This method was tested in two contrasting basins, the Little River Experimental Basin in Georgia, USA, and the Yellow River Headwater Basin in China. The results obtained in the two cast studies show that this combined method can provide deterministic predictions better than or comparable to the best calibrated model using GA. 66.7% and 90% uncertainty intervals estimated by this method were analyzed. The differences between the percentage of coverage of observations and the corresponding expected coverage percentage are within 10% for both calibration and validation periods in these two test basins. This combined methodology provides a practical and flexible tool to attain reliable deterministic simulation and uncertainty analysis of SWAT.

Key words optimization; modeling; basin; uncertainty; SWAT

1 INTRODUCTION

In recent years, hydrologic models are more and more widely applied by hydrologists and resource managers as a tool to understand and manage ecological and human activities that affect basin systems. Traditionally, the hydrologic models are calibrated to find one optimal hydrologic model with the optimum objective functions (e.g. sum square error). The optimized model is then used to assess water resources practices. The inferences based on a single model implicitly assumes that the probability that the single model generates the data accurately is 1, and neglects the uncertainty inherent in the model selection process (Montgomery and Nyhan, 2008; Raftery and Zheng, 2003). Uncertainty within model output is a major concern, particularly when modeling results are used to set policy. Because of uncertainties associated with input, model structure, parameter, and output, the model predictions are not a certain value, and should be represented with a confidence range (Beven and Binley, 1992, Gupta et al., 1998; Beven, 2000; Beven and Freer, 2001; Beven, 2006; Van Griensven, 2008). Reasonable estimates of prediction uncertainty of hydrologic processes are valuable to water resources and other relevant decision making processes (Liu and Gupta, 2007). Uncertainty estimates are routinely incorporated into Total Maximum Daily Load (TMDL) estimates and are an important part of the TMDL implementation plan (Shirmohammadi et al., 2006). Usually, water management projects are planned and designed using scenarios that fall at the conservative end of the range of plausible outcomes. Over estimation of uncertainty can result in over design of mitigation measures, while under estimation of uncertainty can lead to inadequate preparation for potential situations. In order to successfully apply hydrological models in practical water resources investigations, careful calibration and prediction uncertainty analysis are required (Duan et al., 1992; Beven and Binley, 1992; Vrugt et al., 2003; Yang et al., 2008; Van Griensven et al., 2008).

As a physically based hydrologic model that can simulate most of the key hydrologic processes at basin scale, the Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998) has been applied world wide for assessing water resources management (Gassman et al., 2007). In order to efficiently and effectively apply the SWAT model, different calibration and uncertainty analysis methods have been developed and applied to improve the prediction reliability and quantify prediction uncertainty of SWAT simulations (Eckhardt and Arnold, 2001; Bekele and Nicklow, 2007; Yang et al., 2007; Harmel and Smith, 2007; Arabi et al., 2007; Kannan et al., 2008). For example, Van Griensven et al. (2003) incorporated the shuffled complex evolution (SCE) algorithm for parameter calibration of SWAT, which was later extended to an uncertainty analysis method known as Sources of Uncertainty Global Assessment using Split SamplES (SUNGLASSES) (Van Griensven et al., 2008). Muleta and Nicklow (2005) combined Genetic Algorithms (GA) and Generalized Likelihood Uncertainty Estimation (GLUE) methods to conduct parameter calibration and uncertainty analysis of SWAT. Yang et al. (2008) compared four uncertainty analysis algorithms, that is GLUE (Beven and Binley, 1992), Sequential Uncertainty Fitting SUFI-2 (Abbaspour et al., 2004), Parameter solutions (ParaSol) (van Griensven and Meixner, 2004), and Markov Chain Monte Carlo (MCMC) based Bayesian analysis techniques for assessing the uncertainty of SWAT predictions. These uncertainty analysis algorithms are differing in philosophy, assumptions, and sampling strategies. Yang et al. (2008) suggested that, if computationally feasible, Bayesian Markov Chain Monte Carlo (MCMC) approaches are most recommendable because of their solid conceptual basis. It is worth noting that the MCMC method requires a large number of SWAT runs. For example, 45,000 runs of SWAT were performed in Yang et al. (2008). Zhang (2008b) test an evolutionary Monte Carlo based MCMC method for SWAT, which took about 200,000 model runs for convergence. Applying the MCMC based methods to assess water resources under future scenarios (e.g. best management practices, and land use/climate change) is very computationally intensive. In the previous uncertainty studies using SWAT, model prediction uncertainty is mainly attributed to parameter values. It is worth noting that the bias and uncertainty result from model structures selection can exert important impact on model prediction (Neuman, 2003; Butts et al., 2004a, 2004b). Butts et al. (2004a) presented an evaluation of model structure on hydrologic modeling uncertainty by selecting different plausible model structures within a general hydrological modeling tool, and emphasize the importance of exploring different model structures as part of the overall modeling approach. The SWAT model provides a hydrologic modeling tool that allows different model structures to be selected for representing different hydrological processes (e.g. potential evapotranspiration, snow routing, and flood routing). The major purpose of this study is to explore ensemble hydrologic simulation and uncertainty analysis using several model structures within the SWAT model framework.

Recently, Bayesian Model Averaging (BMA), a method for averaging over different competing models, has been applied to allow incorporating model uncertainty into model prediction. BMA possesses a range of theoretical optimality properties and has shown good performance in reliable prediction and uncertainty analysis in a variety of simulated and real data situations (e.g. weather forecast and hydrologic predictions) (Raftery et al., 2005; Ajami et al., 2006; Duan et al., 2007; Vrugt et al., 2007; Montgomery and Nyhan, 2008). The BMA can be used to examine several competitive models for hydrologic modeling and assessment. In practical applications of SWAT, modelers usually select one or several model structures and choose the best among them. To the best of the authors' knowledge, seldom studies have been conducted to jointly use multiple structures within the SWAT model. In this study, a combined method, which implements the Genetic Algorithms (GA) and BMA, was proposed to conduct calibration and uncertainty analysis of the SWAT model through jointly using multiple model structures. The general procedures for applying GA and BMA to conduct ensemble hydrologic predictions applied here are: 1) select the specific model components of SWAT to be examined, here we examined different snow, potential evapotranspiration and flow routing methods; 2) calibrate the parameters for each combination of model components using GA to provide competing models and model results; 3) use BMA to combine the ensemble predictions and provide uncertainty interval estimation. The examination was limited to the snow, potential evapotranspiration and flow routing to present a manageable number of modeling options for illustration purposes. Compared with running thousands of models for assessing management practices or climate / land use change scenarios using MCMC based method, the BMA has the potential to save a large number of runs of SWAT. Two basins were used to test the validity of this framework for providing accurate hydrologic prediction and uncertainty intervals estimation using SWAT. The combination of GA and BMA is expected to provide a practical tool for implementing calibration and uncertainty analysis of computationally intensive hydrologic models.

2. MATERALS AND METHODS

2.1. Study area description

Two basins, the Little River Experimental Basin (LREB) in the Southeastern USA and Yellow River Headwater Basin (YRHB) in central China were used in this study (Figure 1). The basins were selected to offer a contrast in hydrology for testing purposes. The basic characteristics of the two basins are introduced as follows.

The LREB (Figure 1) is the upper 334 km^2 of the Little River in Georgia, USA, and is the subject of long-term hydrologic and water quality research by USDA-ARS and cooperators (Sheridan, 1997; Bosch et al., 2007). The LREB is located in the Tifton Upland physiographic region, which is characterized by intensive agriculture in relatively small fields in upland areas and riparian forests along stream channels. The region has low topographic relief and is characterized by broad, flat alluvial floodplains, river terraces, and gently sloping uplands (Sheridan, 1997). Climate in this region is characterized as humid subtropical with an average annual precipitation of about 1167 mm based on data collected by USDA ARS from 1971 to 2000. Soils on the basin are predominantly sands and sandy loams with high infiltration rates. Since surface soils are underlain by shallow, relatively impermeable subsurface horizons, deep seepage and recharge to regional ground water systems are impeded (Sheridan, 1997). Land use types include forest (50%), cropland (31%), pasture (10%), water (2%), and urban (7%) (Bosch et al., 2006).

The YRHB (Figure 1) is an 114,345 km^2 mountainous river basin, which is located in the northeastern part of Tibetan plateau in China. This area is the primary source of water availability for the Yellow River Basin (Liu, 2004). The average elevation is about 4,217 m, and ranges between 2,600 and 6,266 m. The area slopes downward from west to east, ranging from a combined landform of low-mountains and wide valleys with lakes to smooth plateaus. The headwater area has a typical continental alpine cold and dry climate. The annual precipitation amount is around 600 mm and the average annual temperature for the YRHB is near 0°C. In winter the average temperature is below 0°C for most of the weather stations, while in summer the average temperature is above 0°C. This seasonal temperature variation makes snowmelt an important process in this area (Zhang et al., 2008a). This basin is characterized by gently sloping upland, river bed, and swamp and wetland. The major types of soils in this area are clay and loam with relatively low infiltration rates. The major land cover in the study area is grassland, which accounts for approximately 90% of the total area. Other land use/land cover (forest land, rangeland, agriculture land, and bare area) accounts for the remaining 10% of the area.

2.2 SWAT model description

SWAT is a continuous time, physically based hydrological model. SWAT subdivides a basin into sub-basins connected by a stream network, and further delineates Hydrologic

Response Units (HRUs) consisting of unique combinations of land cover and soils in each sub-basin. SWAT allows a number of different physical processes to be simulated in a basin. The hydrologic routines within SWAT account for snow fall and melt, vadose zone processes (i.e., infiltration, evaporation, plant uptake, lateral flows, and percolation), and ground water flows. The hydrologic cycle as simulated by SWAT is based on the water balance equation:

$$SW_t = SW_0 + \sum_{i=1}^{t}(R_{day} - Q_{surf} - E_a - w_{seep} - Q_{gw}) \quad (1)$$

where SW_t is the final soil water content (mm H_2O), SW_0 is the initial soil water content on day i (mm H_2O), t is the time (days), R_{day} is the amount of precipitation on day i (mm H_2O), Q_{surf} is the amount of surface runoff on day i (mm H_2O), E_a is the amount of evapotranspiration on day i (mm H_2O), w_{seep} is the amount of water entering the vadose zone from the soil profile on day i (mm H_2O), and Q_{gw} is the amount of return flow on day i (mm H_2O). Precipitation in SWAT is divided into rainfall and snowfall. There are three snow routing algorithms available in SWAT, which include the degree day (DD), DD plus elevation band (Fontaine et al., 2002), and the energy balance based SNOW17 models (Zhang et al., 2008a). Surface runoff volume is estimated using a modified version of the Soil Conservation Service (SCS) Curve Number (CN) method (Neitsch et al., 2005a). For evapotranspiration estimation, three options are available in SWAT, that is, Penman-Monteith, Priestley-Taylor, and Hargreaves methods (Neitsch et al., 2005a). A kinematic storage model is used to predict lateral flow, whereas return flow is simulated by creating a shallow aquifer (Arnold et al., 1998). The Variable Storage and Muskingum methods are used for channel flood routing. Outflow from a channel is adjusted for transmission losses, evaporation, diversions, and return flow (Arnold et al., 1998).

In the SWAT model, there are numerous parameters to be calibrated to match the simulated and observed flow. Van Liew et al. (2007) tested the suitability of SWAT for the

Conservation Effects Assessment Project in several USDA Agricultural Research Service basins. In the study conducted by Van Liew et al. (2007), eleven parameters were identified as sensitive for the LREB. These eleven parameters (Table 1) were adjusted by the GA for the LREB in this study. In the YRHB, five parameters (i.e. CN2, ESCO, SURLAG, GW_REVAP, and ALPHA_BF) were adjusted for the calibration according to Zhang et al. (2008a). The general description of the parameters used for the calibration is shown in Table 1. The parameters' ranges were limited according to van Griensven et al. (2006) and Neitsch et al. (2005b).

2.3 Genetic Algorithms

Zhang et al. (2009c) compared five global optimization algorithms for parameter calibration of SWAT in four basins, and their results show the advantage of GA over other algorithms for calibrating SWAT. Genetic Algorithms are stochastic search procedures inspired by evolutionary biology of natural selection and genetics (Holland, 1975; Goldberg, 1989), such as inheritance, mutation, selection, and crossover. The implementation of GA starts with initializing a population of candidate solutions (called chromosomes) which are randomly sampled from the feasible parameter space. In each generation, the individual chromosomes are selected through a fitness-based process, where the more fit chromosomes in the population are preferred to be selected to reproduce new promising offspring. Next, a new generation population of chromosomes is generated from these selected ones using crossover and mutation operations. The crossover operator chooses "parent" solutions and exchange important building blocks of two parent chromosomes to generate new "offspring" solutions. The "offspring" solutions are then randomly mutated to increase the diversity of new population. Through a steady-state-delete-worst plan (Reca and Martinez, 2006), the fitter chromosomes among the old and new population are input into next generation for evolution. This generational evolution of the parameter solutions is repeated until a maximum number of model evaluations are reached. With flexibility and robustness, GAs have been successfully applied to solve complex nonlinear programming problems in many science and engineering branches (Reca and Martinez, 2006). Following Schaffer et al. (1989) and Reca and Martinez (2006), the crossover rate was set to 0.5 and mutation rate was the reciprocal of the parameter dimension. Settings of population size and maximum model runs can substantially affect the performance of GA for calibrating SWAT, a small population size of 50 and a maximum number of SWAT runs of 5000 were selected in this study following Zhang et al. (2009c).

2.4 Bayesian Model Averaging

In hydrologic modeling, there are many ensemble based methods that can merge information from multiple sources (e.g. modeling results from different models and observed data from different sources). One simple method is the arithmetic mean method, which simply averages the predictions from several sources equally to obtain the ensemble mean prediction. This method has shown more reliable prediction than single model prediction (Raftery et al. 2005; Hsu et al., 2008). Recently, advanced BMA was proposed to combine multiple weather and hydrologic models results to provide more reliable predictions (e.g. Raftery et al. 2005; Ajami et al., 2006; Duan et al., 2007; Vrugt et al., 2007). BMA is a standard approach to inference in the presence of multiple competing models (Raftery and Zheng, 2003). This approach has been used to infer probabilistic predictions that possess more skill and reliability than the original ensemble members produced by several competing models (Duan et al., 2007). In BMA, the probabilistic distribution of a hydrologic prediction y is the weighted average of the posterior distribution of each model under consideration. Raftery et al. (2005) extended BMA from statistical models to weather forecast models. In the following, the BMA framework developed by Raftery et al. (2005) was introduced. The BMA prediction probability distribution can be represented as

$$p(y \mid f_1, f_2, \ldots, f_K) = \sum_{k=1}^{K} w_k g(y \mid f_k) \tag{2}$$

where K is the number of competing models and k is the index of each model. f_k denote the bias corrected prediction of a candidate model M_k. w_k is $p(f_k \mid D)$, the posterior probability of model prediction f_k, also known as the likelihood of model prediction f_k being the correct prediction given the observational data, D. w_k is nonnegative and with a sum ($\sum_{k=1}^{K} w_k$) of 1. $g(y \mid f_k)$ represents the conditional probability distribution function (PDF) of y conditional on f_k. Usually, the conditional distribution $g(y \mid f_k)$ can be represented as a normal distribution, $N(a_k + b_k f_k, \sigma_k^2)$, where a_k and b_k are regression coefficients obtained through least square linear regression. Following Raftery et al. (2005) and Duan et al. (2007), the BMA predictions mean and variance can be calculated as

$$E(y \mid f_1, f_2, \ldots f_K) = \sum_{k=1}^{K} w_k (a_k + b_k f_k) \tag{3}$$

$$Var(y \mid f_1, f_2, \ldots f_K) = \sum_{k=1}^{K} w_k \left[(a_k + b_k f_k) - \sum_{i=1}^{K} w_i (a_i + b_i f_i) \right]^2 + \sum_{k=1}^{K} w_k \sigma_k^2 \tag{4}$$

where σ_k^2 is the variance associated with model prediction f_k with respect to calibration data D. The BMA prediction mean is the weighted average of individual predictions weighted by the likelihood $p(f_k \mid D)$. It can be viewed as a deterministic prediction and compared with other individual predictions in the ensemble and ensemble mean. The two terms of the right-hand side of equation (4) represent the between-prediction variance and within-prediction variance, respectively. The BMA predicts spread-error correlation, and also accounts for the possibility that ensembles may be underdispersive, which is usually the case in ensemble predictions (Raftery et al., 2005).

In order to apply the BMA method, the weights w_k and variance σ_k^2 need to be estimated. In this study, the maximum likelihood estimation (MLE) method was adopted following Raftery et al. (2005). Let $\theta = \{w_1, w_2, \ldots, w_K, \sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\}$. The log form of the likelihood needs to be maximized is

$$\ell(\theta) = \log\left[\sum_{k=1}^{K} w_k g(y \mid f_k)\right] \tag{5}$$

It is difficult to analytically maximize this log likelihood. In this study, the Expectation and Maximization (EM) was used to find the maximum likelihood estimator. EM algorithm is iterative. It starts with a initial guess of θ^0. Then the EM algorithm alternates between the Expectation step and Maximization step to update the estimation of θ^{Iter}, where *Iter* is the iteration number. Finally, the Expectation step and Maximization step converge and are stopped when there is no significant change, measured by a small tolerance value, between two consecutive iterative log likelihood estimations. Following Raftery et al. (2005) and Duan et al. (2007), the procedures of applying EM algorithm to estimate w_k and σ_k^2 are briefly described in Appendix A. The probabilistic predictions of the variable of interest can be derived based each individual deterministic prediction and its weight and variance. The procedures used in this study to generate probabilistic predictions at each time step t are briefly described as follows (Gelman et al., 2003): i) select an individual competing model (M_k) with the probability proportional to its weight; ii) draw a replication y^{rep} from $g(y_t \mid f_{k,t})$; iii) repeat steps i and ii to obtain 1000 values that represent the distribution of y_t, with which the uncertainty intervals can be derived. For example, the 90% interval is within the range of the 5% and 95% quartiles. Similarly, other uncertainty intervals with different expected coverage percentage can be derived straightforward.

2.5 Generating competing hydrologic predictions of SWAT

Hydrologic environments are open and complex, rendering them prone to multiple interpretations and mathematical descriptions (Neuman, 2003). In practical application of hydrologic model, modelers typically select a single model among the several choices that is assumed to best represent the hydrologic system. The major advantage of BMA is to jointly use several model structures identified as plausible by the modelers. For the selection of candidate models for BMA, it is suggested to use previous research and theory to specify the set of model structures that are plausible and supported by data (Gelman and Rubin, 1995; Raftery et al., 2005; Duan et al., 2007; Vrugt et al., 2007; Montgomery and Nyhan, 2008). In this study, we followed the methodology used in previous literature on model structures selection. In the selection of model structure, we used the information provided in previous literature on SWAT (Neitsch et al., 2005a, 2005b) and the actual watershed characteristics. The purpose of this paper is to illustrate the application of GA and BMA for combining several plausible model structures within SWAT framework. It is out of the scope of this study to explore all possible model structures.

The SWAT model has several options for setting its model structures. Different evapotranspiration, snow accumulation and melt, and flow routing algorithms are available within the SWAT model system. In the LREB, as snowfall and melt is not an important process, we set up SWAT model structures by selecting different evapotranspiration and flow routing algorithms. For the potential evapotranspiration, Penman-Monteith “PM”, Priestley-Taylor “PT”, and Hargreaves “HG” were selected. For flow routing, Variable Storage “VS” and Muskingum “MK” were selected. Thus, SWAT_PM_VS denotes SWAT with Penman-Monteith potential evapotranspiration estimation and variable storage flow routing. A total of six model structures were defined, that is, SWAT_PM_VS, SWAT_PM_MK, SWAT_PT_VS, SWAT_PT_MK, SWAT_HG_VS, and SWAT_HG_MK. The evaluation time scale selected for the LREB was day. In the YRHB, we only choose three models with different snowfall and melt algorithms, because snow processes are significant (Zhang et al., 2008a) and the evaluation time scale was month. Previous studies (e.g. Fontaine et al., 2002; Zhang et al., 2008a) have shown that the SWAT model simulation is sensitive to snow routing methods in mountainous basin. The snow routing methods used in this study include the degree day "DD", DD plus elevation band "ELEV", and the energy based SNOW17 methods. The SWAT models with different snow modules are represented as SWAT-DD, SWAT-ELEV, and SWAT-SNOW17, respectively. The GA was applied to optimize the SWAT models with different structures in the LREB and YRHB. In the LREB, daily streamflow from 1999 to 2000 was used to calibrate model and daily streamflow from 2001 to 2002 was used to validate the model. Watershed weighted annual precipitation for this period for LREB varied from a high of 1085 mm observed in 2000 to a low of 884 mm observed in 1999. Precipitation and flow for both the calibration and validation periods were slightly below long term means. For the YRHB, monthly flow from 1976 to 1985 was used to calibrate model and monthly flow from 1986 to 1990 was used to validate the model. Precipitation for this period varied from 653 mm to 482 mm. Precipitation and flow of the selected periods in the YRHB are very close to long term average conditions. The calibrated models with smallest sum square error in the LREB and YRHB serve as competing models for the BMA, and the BMA mean and prediction uncertainty interval are calculated.

2.5 Statistical criteria for evaluating the performance of hydrologic prediction

Different statistical criteria were used to evaluate the individual SWAT model predictions, ensemble mean, BMA mean, and the uncertainty intervals obtained by the BMA. Following Santhi et al. (2001) and Moriasi et al. (2007), the evaluation coefficient for deterministic predictions include Percent Bias (*PBIAS*), Coefficient of Determination (R^2), and Nash-Sutcliffe Efficiency (*NSE). PBIA*S is calculated as

$$PBIAS = (\sum_{t=1}^{T}(f_t - y_t) \Big/ \sum_{t=1}^{T} y_t) \times 100 \tag{6}$$

where f_t is the model simulated value at time unit t, y_t is the observed data value at time unit t, and $t = 1,2,\ldots,T$. *PBIAS* measures the average tendency of the simulated data to be larger or smaller than their observed counterparts (Gupta et al., 1999). *PBIAS* values with small magnitude are preferred. Positive values indicate model overestimation bias, and negative values indicate underestimation model bias (Gupta et al., 1999).

The formula for calculating coefficient R^2 is

$$R^2 = \left\{ \sum_{t=1}^{T}(y_t - \bar{y})(f_t - \bar{f}) \Big/ \left[\sum_{t=1}^{T}(y_t - \bar{y})^2\right]^{0.5} \left[\sum_{t=1}^{T}(f_t - \bar{f})^2\right]^{0.5} \right\}^2 \tag{7}$$

where $\bar{y}$ is the mean of observed data value for the entire time period of the evaluation, $\bar{f}$ is the mean of simulated data value for the entire time period of the evaluation. The other symbols have the same meaning defined above. R^2 is equal to the square of the Pearson's product-moment correlation coefficient (Legates and McCabe, 1999). It represents the proportion of the total variance in the observed data that can be explained by the model. R^2 ranges between 0.0 and 1.0. Higher values mean better performance.

NSE is calculated as

$$NSE = 1.0 - \sum_{t=1}^{T}(y_t - f_t)^2 \Big/ \sum_{t=1}^{T}(y_t - \bar{y})^2 \tag{8}$$

NSE indicates how well the plot of the observed value versus the simulated value fits the 1:1 line, and ranges from $-\infty$ to 1 (Nash and Sutcliffe, 1970). The larger the *NSE* values, the better model performance.

In hydrologic modeling, different types of uncertainty limits can be recognized (e.g. Beven, 2006; Liu and Gupta, 2007). In this study, we are concerned with the modeling uncertainty and predictive uncertainty (Liu and Gupta, 2007). The modeling uncertainty limits, obtained through calibrating hydrologic models to match observed streamflow data, were expected to include a specified proportion of the calibration data set. The predictive uncertainty limits, obtained through applying the calibrated models to another independent data set, were expected to contain a specified proportion of future observations. In this study, the percentage of coverage (POC) of observations in the uncertainty interval was used to evaluate the uncertainty intervals obtained by the BMA scheme. The smaller difference between POC and the expected coverage percentage of an uncertainty interval indicate better performance of the estimated uncertainty interval. For a 90% uncertainty interval, which is expected to include 90% of the observed data, the POC value closer to 90% indicate the better performance of the uncertainty interval estimation.

3. RESULTS AND DISCUSSION

3.21Calibration and uncertainty analysis results in the LREB

The evaluation coefficients of the simulated daily streamflow by different prediction techniques in the LREB are listed in Table 2. The two sample Kolmogorov–Smirnov test (K–S test) (Chakravarti et al., 1967) reveals that the difference between the simulated results by models with default input and those calibrated by GA is significant at a significant level of 0.05. This indicates that model calibration can substantially improve model simulation. The calibrated parameters for the six models in the LREB are shown in Table 3, which clearly show that different model structure prefer different parameter values. For example, the calibrated values of CN range between -17% and 20%. For illustration purpose, the simulated daily streamflow by the different methods in March, 1999 and in March, 2001 are shown in Figures 2 and 3 for calibration and validation periods, respectively. The ensemble mean and BMA mean predictions were also plotted for comparison purpose. From Figures 2 and 3, there is obvious difference between the hydrographs simulated by different models, especially in the validation period. At a significant level of 0.05, the K-S test results show that there is significant difference between different model simulation results. The evaluation coefficients in Table 2 confirmed the difference between different models. For example, in calibration period, SWAT-HG-VS obtained *PBIAS* of -0.72%, while the *PBIAS* value of SWAT-PM-VS was 24.9%. The performance of calibrated models in validation period is different from that in calibration period. For example, the *PBIAS* values of SWAT-PT-VS increased from 22.94% in calibration period to 46.82% in validation period. Analysis of other evaluation coefficients also shows the difference between model performance in calibration and validation period (Table 2). The difference between model performance in calibration and validation periods is because the hydrologic conditions in validation period may change and do not look exactly like the hydrologic conditions during the calibration period (e.g. Beven, 2006; Liu and Gupta, 2007; Zhang et al., 2009a). The different properties exhibited by various models were combined by the arithmetic ensemble mean and Bayesian model averaging methods. The comparison of the evaluation coefficients of each single model and those of the ensemble based methods indicate the obvious superiority of applying ensemble based methods. Compared with single models predictions, the simple arithmetic ensemble mean obtained better results in terms of R^2, and N*SE* during both calibration and validation period. The BMA outperformed all the other seven methods in terms of all four evaluation coefficients in both calibration and validation periods. The above analysis clearly illustrates the advantage of using ensemble based methods to obtain reliable deterministic streamflow simulation, especially the Bayesian Model Averaging.

The 66.7% and 90% uncertainty intervals estimated by the BMA are shown in Figures 4 and 5 for calibration and validation periods, respectively. The estimated 66.7% and 90% uncertainty intervals cover about 76.04% and 91.14% of the observed data, respectively, in calibration period, and about 74.41% and 96.53% of the observed data, respectively, in validation period. The absolute differences between the POCs values computed from the uncertainty intervals estimated by the BMA and expected coverage percentages are within 10%. In general, the POC values estimated by BMA are matching well with the expected coverage percentage.

3.2 Calibration and uncertainty analysis results in the YRHB

The evaluation coefficients of the simulated monthly streamflow by different prediction techniques in the YRHB are listed in Table 4 for different prediction techniques. The K-S test results indicate that the difference between the simulated results by models with default input and those calibrated by GA is significant at a significant level of 0.05, which emphasize the importance of parameter calibration. The calibrated parameters (Table 5) for the three models also exhibit very different values in the YRHB. Using different snow routing methods can lead to variation of calibrated CN values from 2% to 14%. For illustration purpose, the simulated monthly streamflow by the different methods in 1976 and in 1986 are shown in Figures 6 and 7 for calibration and validation periods, respectively. Similar to the case in the LREB, the hydrographs simulated by the three models with different snow routing algorithms have pronounced differences. The SWAT-DD model consistently underestimates the streamflow, with *PBIAS* values of -17.71% and -17.98% for calibration and validation periods, respectively. The SWAT-SNOW17 model obtained positive *PBIAS* values less than 10% for both calibration and validation periods. The arithmetic ensemble mean and BMA mean predictions consistently obtained better performance in terms of R^2, and *NSE* than single model based predictions. In terms of *PBIAS*, BMA mean outperformed all the other methods in calibration period, while it performed less than SWAT-ELEV in validation period. But BMA mean still predicted small *PBIAS* value (less than 5%) in the validation period. In the YRHB test case, BMA provided better deterministic prediction than the best ensemble number in calibration period and similar results in validation period.

The uncertainty intervals with different expected coverage percentages estimated by BMA are shown in Figures 8 and 9 for the calibration and validation periods, respectively. The differences between the estimated POC values by BMA and the corresponding expected coverage percentages are within 6% for both calibration and validation periods. The estimated 66.7% and 90% uncertainty intervals cover about 64.2% and 87.5% of the observed data, respectively, in calibration period, and about 68.67% and 91.67% of the observed data, respectively, in validation period. This good match indicates the validity of using only three ensemble members to estimate the uncertainty of hydrologic predictions.

3.4 Discussion

The test results in the two contrasting basins indicate the combination of GA and BMA holds promise to be an efficient and effective technique to calibrate SWAT model and provide reasonable estimation of prediction uncertainty. The numbers of model runs of SWAT are 30000 and 15000 in LREB and YRHB, respectively. These numbers of model runs reported in this study is much less than those reported in previous studies. For example, two previous studies that applied MCMC for SWAT reported 45000 (Yang et al., 2008) and 200000 (Zhang, 2008b) model runs. In addition, in contrast to MCMC methods which usually require thousands of SWAT runs, one only needs to run several competing SWAT models with different model structures to assess water resources effect of different management and global change scenarios. For the computationally intensive SWAT model, the method used in this study has the potential to save enormous computational resources and time. It is still important to note that the time consumed by calibrating one model structure is intensive. We calibrated the candidates SWAT models on a computer with Pentium IV 3 GHZ and 1GB RAM. In the LREB, calibration of each of six model structures took about 3 days. A total of 18 days were spent on model calibration for the six model structures in the LREB. In the YRHB, calibration time consumed by SWAT-DD, SWAT-ELEV, and SWAT-SNOW17 was

3 days, 5 days, and 25 days, respectively. Given the enormous time consumed by constructing candidate model structures for BMA, using as small number of candidates as possible is very important. We tested the effect of reducing number of models on BMA prediction. In the LREB, we eliminated the candidate model with less *NSE* in sequence until there were only two models remaining. The calculated *PBIAS*, R^2, *NSE*, and *POC* values for each combination of model structures are listed in Table 5. The difference between these evaluation coefficients is very small. For example the *NSE* values range between 0.8 and 0.81 in calibration period and between 0.84 and 0.86 in validation period. The difference between *POC* values are less than 5% for both 66.7% and 90% intervals. It is worth noting that the *PBIAS* value reached 10% in validation period when using 2 candidate models. This compares to the *PBIAS* values less than 5% for the other combinations of candidate models. Further test in the YRHB show that the evaluation coefficients obtained with two candidate models (SWAT-ELEV and SWAT-SNOW17) are also very close to those calculated using all three models. Overall, reducing number candidate models does not have substantial effect on the performance of BMA in the two case studies. This result is similar to that in Raftery et al. (2005). Considering the relatively large *PBIAS* value obtained by using two candidate models in LREB, it is suggested that three or more model structures are needed for BMA. As hydrologic conditions are varying from site to site, much care should be taken when transfer the results to other basins.

There are several limitations of the method used in this study. It is also worth noting that the BMA mean prediction can not always outperform the other models predictions for all the evaluation coefficients. For example, in the YRHB, the BMA mean predicted larger PBIAS than SWAT-ELEV and performed almost the same as the simple arithmetic ensemble mean in validation period. The K–S test results show that the BMA mean prediction is significantly different from all ensemble members in LREB at a significance level of 0.05.

While in YRHB, the BMA mean is significantly different from all ensemble members at significance level of 0.2. As significance level of 0.05 is commonly used in hydrologic modeling, the results indicate that the relatively complex BMA analysis did not necessarily show significant improvement. The discrepancy between POC values obtained by the BMA and the expected coverage percentage, which reached about 10% and 6% respectively in the LREB and YRHB, respectively, also shows the BMA methods can be further improved. These inadequacies of the BMA method may be caused by several reasons: i) the uncertainty associated with the input data was not explicitly accounted for. For example, the precipitation uncertainty may have important effect on uncertainty estimation (Kavetski et al., 2006); ii) the residuals between simulated and observed streamflow data are assumed to independent, which may not be true in real world problems (Kuczera and Parent, 1998; Yang et al., 2007); iii) the prior knowledge of different uncertainty sources, which may affect the uncertainty estimation (Zhang et al., 2009a), was not explicitly considered in the BMA scheme. In the future, incorporating more sources of uncertainty into account (Kuzera et al., 2006) may improve the performance of this method. Methods on incorporating input data uncertainty, obtaining prior knowledge of model, and considering correlation between residuals deserve further research for improving the reliability of SWAT predictions. Another limitation of this method is that the application of GA for parameter estimation took very long time. The expensive computational cost is limiting the use of this method. In the future, incorporating surrogate model (e.g. Zhang et al., 2009b) and parallel computing techniques (e.g. Vrugt et al., 2006) into the model calibration process is a promising research topic.

For water resources investigations essential for relevant decision making processes, the predictive uncertainty estimation associated with hydrologic simulation is valuable. Predictive uncertainty limits are dependent on and different from modeling uncertainty. This is because when the calibrated hydrological models are applied to another set of data independent of the calibration data, the hydrologic conditions may change and therefore impact the predictive interval estimation (Beven, 2006; Liu et al., 2008; Zhang et al,. 2009a). The results obtained in the two test basins show that the percentage of coverage values of modeling and predictive uncertainty intervals can be different from each other. In the YRHB, the predictive uncertainty interval included more observed data than the modeling uncertainty intervals. For example, POC value of the 90% interval is 4% less in calibration period than that in validation period. In the LREB, the modeling uncertainty intervals in calibration period included more observed data for 66.7% interval than the corresponding predictive uncertainty intervals in validation period, while the 90% modeling uncertainty interval included about 5% less observed data than the 90% predictive uncertainty interval. Because of the future uncertainties due to natural and anthropogenic factors, the predictive uncertainty limits are also uncertain, which means that we are unable to estimate predictive uncertainty limits even if our estimation of modeling uncertainty limits are accurate. Hence in application of uncertainty analysis for hydrologic prediction, how to extend modeling uncertainty limits to predictive uncertainty limits remains a challenge for applying hydrologic models to water resources-related management and design problems.

4. CONCLUSIONS

In this paper, we presented the application of GA and BMA to simultaneously conduct calibration and uncertainty analysis of SWAT. The methodology provides a practical and flexible tool for jointly using multiple model structures within the SWAT model system. This method was tested in two basins. In the LREB, we selected six SWAT models with different evapotranspiration and flow routing algorithms, and tested this method using daily streamflow. In the YRHB, we selected three SWAT models with different snow routing modules, and used monthly streamflow data to test this method. The test results show that this combined method can provide deterministic predictions better than or comparable to the best calibrated model using GA. Further inspection of the 66.7% and 90% uncertainty intervals show that the combination of GA and BMA can provide reasonable uncertainty estimation. The differences between the computed percentage of coverage values and the corresponding expected coverage percentages are within 10% for both calibration and validation periods in these two test basins. It is anticipated that the combination of GA and BMA methods will have significant implications related to policy development. The method reduces the uncertainty associated with selecting any single model, thereby increasing the level of confidence in the simulation results. This is a critical component of policy assessments which are based upon modeling results and one which will become more routine in the future.

Acknowledgements

The authors would like to thank the associate editor and two anonymous reviewers for the constructive and valuable comments and suggestions, which greatly enhanced the quality of the manuscript.

Appendix A:

1. Initialization:

Set $Iter = 0$, $w_k^{Iter} = \dfrac{1}{K}$, and $\sigma^2_{k,Iter} = \dfrac{1}{K}\sum_{t=1}^{T}(\sum_{k=1}^{K}(y_t - f_{k,t})^2 / T)$, and fit the regression coefficients a_k and b_k for each candidate model using linear regression.

where T is the total number of data points in the calibration period, and *Iter* is the iteration number.

2. Computing the initial likelihood:

$$\ell(\boldsymbol{\theta}^{Iter}) = \log\left(\sum_{k=1}^{K} w_k g(y \mid f_k)\right) \qquad \text{A1}$$

where $g(y \mid f_k)$ is calculate as $\sum_{t=1}^{T} g(y_t \mid f_{k,t}, \sigma^2_{k,Iter})$. $g(y_t \mid f_{k,t}, \sigma^2_{k,Iter})$ represents a normal distribution center at $a_k + b_k f_{k,t}$ with variance of $\sigma^2_{k,Iter}$

3. Executing the expectation step

Set $Iter = Iter + 1$

For $k = 1,2,\ldots,K$ and $t = 1,2,\ldots,T$, $\hat{z}^{Iter}_{k,t} = g(y_t \mid f_{k,t}, \sigma_{k,Iter-1}) \Big/ \sum_{k=1}^{K} g(y_t \mid f_{k,t}, \sigma_{k,Iter-1})$

4. Executing the maximization step

Compute the weight for each model: $w_k^{Iter} = \dfrac{1}{T}\sum_{t=1}^{T}\hat{z}^{Iter}_{k,t}$

Update the variance of each model: $\sigma^2_{k,Iter} = \sum_{t=1}^{T}\hat{z}^{Iter}_{k,t}(y_t - f_{k,t})^2 \Big/ \sum_{t=1}^{T}\hat{z}^{Iter}_{k,t}$

5. Update the likelihood $\ell(\boldsymbol{\theta}^{Iter})$ using equation A1

6. Checking convergence:

If $\ell(\boldsymbol{\theta}^{Iter})$ - $\ell(\boldsymbol{\theta}^{Iter-1})$ is less than or equal to a pre-specified tolerance level (10^{-6}), stop; else go back to Step 3.

References

Abbaspour, K.C., Johnson, C. A., van Genuchten, M.T., 2004. Estimating uncertain flow and transport parameters using a sequential uncertainty fitting procedure. *Vadose Zone Journal* 3(4), 1340–1352.

Ajami, K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multi-model combination techniques for hydrological forecasting: application to distributed model intercomparison project results. *Journal of Hydrometeorology* 8, 755–768.

Arabi, M., Govindaraju, R.S., Hantush M. M., 2007. A probabilistic approach for analysis of uncertainty in the evaluation of watershed management practices. *Journal of Hydrology* 333, 459–471.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large-area hydrologic modeling and assessment: Part I. Model development. *Journal of the American Water Resources Association* 34(1): 73-89.

Bekele, G. E., Nicklow,W.J., 2007. Multi-objective automatic calibration of SWAT using NSGA-II. *Journal of Hydrology* 341: 165-176.

Beven, K.J. 2006. A manifesto for the enquiringly thesis, *Journal of Hydrology* 320, 18-36.

Beven, K., Binley, A., 1992. The future of distributed models – model calibration and uncertainty prediction. *Hydrological Processes* 6 (3), 279–298.

Beven, K.J., 2000. *Rainfall-runoff Modeling: The Primer*. John Wiley & Sons Press: New York.

Beven, K.J., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems. *Journal of Hydrology* 249, 11-29.

Bosch, D.D., Sheridan, J.M., Lowrance, R.R., Hubbard, R.K., Strickland, T.C., Feyereisen, G.W., Sullivan, D.G., 2007. Little River Experimental Watershed database. *Water Resources Research 43*, W09470, doi:10.1029/2006WR005844.

Bosch, D.D., Sullivan, D.G., Sheridan, J. M., 2006. Hydrologic impacts of land-use changes in coastal plain watersheds. *Transactions of the ASABE* 49(2): 423−432.

Butts, M.B., Payne, J.T., Kristensen, M. and Madsen, H., 2004a. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology* 298: 222-241.

Butts, M.B., Payne, J.T. and Overgaard, J., 2004b. Improving streamflow simulations and flood forecasting with multimodel ensemble. In: P.B. Liong (Editor), 6th International Conference on Hydroinformatics. World Scientific Publishing, Singapore.

Chakravarti I.M., Laha R.G., and Roy J. 1967. *Handbook of Methods of Applied Statistics*. John Wiley and Sons, New York, USA.

Duan, Q., Ajami, N. K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources* 30(5), 1371-1386.

Duan, Q., Sorooshian, S., Gupta, V.K., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* 28(4): 1015-1031.

Fontaine, T.A., Cruickshank, T.S., Arnold, J.G., Hotchkiss, R.H., 2002. Development of a snowfall-snowmelt routine for mountainous terrain for the soil water assessment tool (SWAT). *Journal of Hydrology* 262, 209-223.

Gassman, P.W., Reyes, M., Green, C.H., Arnold, J.G., 2007. The Soil and Water Assessment Tool: Historical development, applications, and future directions. *Transactions of the ASABE* 50(4): 1212-1250.

Gelman, A., Carlin J. B., Stern H. S., Rubin D. B., 2003. Bayesian Data Analysis (2nd edition). Chapman & Hall/CRC: Boca Raton, Florida, USA.

Gelman, A. Rubin D. B., 1995. Avoiding model selection in Bayesian social research. Sociological Methodology 25: 165-173.

Goldberg, D., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, Massachusetts, USA.

Gupta, H. V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurate measures of information. *Water Resources Research* 34(4): 751–763.

Gupta, H.V., Sorooshian, S., Yapo, P. O., 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering* 4(2), 135-143.

Harmel, R.D., Smith, P. K., 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *Journal of Hydrology* 337, 326–336.

Holland, J., 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan, USA.

Hsu, K., Moradkhani, H., Sorooshian, S., 2009. A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research* doi:10.1029/2008WR006824.

Kannan, N., Santhi, C., Arnold, J.G., 2008. Development of an automated procedure for estimation of the spatial variation of runoff in large river basins. *Journal of Hydrology* 359, 1–15.

Kavetski, D., Kuczera G., Franks S. W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research 42*, W03407, doi:10.1029/2005WR004368.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M. 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology* 331, 161– 177.

Kuczera, G., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology* 211, 69-85.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35(1): 233-241.

Liu, Y., Gupta,V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research* 43, W07401, doi:10.1029/2006WR005756.

Montgomery, J., Nyhan, B., 2008. *Bayesian Model Averaging: Theoretical developments and practical applications*. Available at http://www.duke.edu/~bjn3/montgomery-nyhan-bma.pdf. Accessed on Oct. 8, 2008.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R. D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* 50(3), 885−900.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., King, K.W., Williams, J.R., 2005a. Soil and Water Assessment Tool (SWAT) theoretical documentation. Blackland Research Center, Texas Agricultural Experiment Station, Temple, Texas, BRC Report 02-05.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J. R., 2005b. *Soil and Water Assessment Tool (SWAT) users manual*. Blackland Research Center, Texas Agricultural Experiment Station, Temple, Texas, BRC Report 02-06.

Neuman, S. P., 2003. Maximum Likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment* 17: 291-305.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* 113: 1155–1174.

Raftery, A.E., Zheng, Y., 2003. Discussion: performance of Bayesian model averaging. *Journal of the American Statistical Association* 98 (464), 931–938.

Reca, J., Martínez, J., 2006. Genetic algorithms for the design of looped irrigation water distribution networks. *Water Resources Research* 42, W05416, doi:10.1029/2005WR004383.

Santhi C., Arnold, J.G., Williams, J.R., Dugas, W.A., Hauck L., 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. *Journal of the American Water Resources Association* 37(5), 1169-1188.

Schaffer, J.D., Caruana, R.A., Eshelman, L.J., Das, R., 1989. A study of control parameters affecting online performance of genetic algorithms for function optimization. In: *Proceedings of the Third International Conference on Genetic algorithms* (ed. By Schaffer, J. D.), 51-60. Morgan Kaufmann, San Mateo, California, USA.

Sheridan, J.M., 1997. Rainfall-streamflow relations for coastal plain watersheds. *Transactions of ASAE* 13(3): 333-344.

Shirmohammadi, A., Chaubey, I., Harmel, R.D. Bosch, D.D. Munoz-Carpena, R.C. Dharmasi, A. Arabi, S.M., Wolfe, M.L. Frankenberger, J., Graff, C., Sohrabi. T.M., 2006. Uncertainty in TMDL Models. *Transactions of the ASABE* 49(4):1033-1049.

Van Griensven, A., Meixner, T., 2006. Methods to quantify and identify the sources of uncertainty for river basin water quality models. *Water Science Technology* 53 (1), 51–59.

Van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Di luzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. *Journal Hydrology* 324: 10-23.

Van Griensven, A., Meixner, T., Srinivasan, R., Grunwals, S., 2008. Fit-for-purpose analysis of uncertainty using split-sampling evaluations. *Hydrological Sciences Journal* 53(5), 1090-1103.

Van Liew, M.W., Arnold, J.G., Bosch, D.D., 2005. Problems and Potential of Autocalibrating a Hydrologic Model. *Transactions of the ASAE* 48(3), 1025-1040

Van Liew, M. W., Veith, T. L., Bosch, D. D., Arnold, J. G., 2007. Suitability of SWAT for the Conservation Effects Assessment Project: A comparison on USDA ARS watersheds. *Journal of Hydrologic Engineering* 12(2), 173-189.

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research* 39(8):1201, doi:10.1029/2002WR001642.

Vrugt, J. A., Nuallain B., Robinson B. A., Bouten W., Dekker S.C., Sloot P. M. A., 2006. Application of parallel computing to stochastic parameter estimation in environmental models. Computers & Geosciences, 32(8), 1139 - 1155

Vrugt, J. A., Robinson, B. A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research* 43, W01411, doi:10.1029/2005WR004838.

Yang, J., Reichert, P., Abbaspour, K. C., Xia, J., Yang, H., 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. *Journal of Hydrology* 358, 1–23.

Yang, K., Reichert, P., Abbaspour, K. C., Yang, H., 2007. Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference. *Journal of Hydrology* 340, 167–182.

Zhang, X, Srinivasan, R, Debele, B., Hao, F., 2008a. Runoff simulation of the Headwaters of the Yellow River using the SWAT model with three snowmelt algorithms. *Journal of the American Water Resources Association* 44(1), 48-61. DOI: 10.1111⁄j.1752-1688.2007.00137.x.

Zhang, X., 2008b. Evaluating and developing parameter optimization and uncertainty analysis methods for a computationally intensive distributed hydrologic model. Ph. diss. Texas A&M University, College Station, Texas, USA.

Zhang, X., Liang, F., Srinivasan, R., Van Liew, M., 2009a. Estimating Uncertainty of Streamflow Simulation using Bayesian Neural Networks. *Water Resources Research* doi:10.1029/2008WR007030.

Zhang, X., Srinivasan, R., Van Liew, M., 2009b. Approximating the SWAT Model Using Artificial Neural Network and Support Vector Machine. *Journal of the American Water Resources Association* 45(2): 460-474.

Zhang, X., Srinivasan, R., Zhao, K., Van Liew, M., 2009c. Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model. *Hydrological Processes* 23(3): 430-441.

List of Tables

727

728 Table 1 Parameters for calibration in SWAT model.

	Parameter	Description	Range
1	CN2	Curve Number	±20%
2	ESCO	Soil Evaporation compensation factor	0-1
3	SOL_AWC	Available soil water capacity	±20%
4	GW_REVAP	Ground water re-evaporation coefficient	0.02-0.2
5	REVAPMN	Threshold depth of water in the shallow aquifer for re-evaporation to occur (mm).	0-500
6	GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)	0-5000
7	GW_DELAY	Groundwater delay (days)	0-50
8	ALPHA_BF	Base flow recession constant	0-1
9	RCHRG_DP	Deep aquifer percolation fraction	0-1
10	CH_K2	Effective hydraulic conductivity in main channel alluvium (mm/hr)	0.01-150
11	SURLAG	Surface runoff lag coefficient (day)	0-10

729 Table 2 Evaluation coefficients for the six SWAT models, arithmetic mean, and BMA mean
730 in the LREB for both calibration and validation periods.

Coefficients / Models	Calibration			Validation		
	PBIAS	R^2	NSE	PBIAS	R^2	NSE
SWAT-HG-MK	-0.72%	0.76	0.74	-8.24%	0.82	0.71
SWAT-HG-VS	6.66%	0.76	0.75	27.07%	0.81	0.76
SWAT-PM-MK	23.90%	0.77	0.76	35.13%	0.82	0.8
SWAT-PM-VS	24.04%	0.72	0.71	39.77%	0.8	0.75
SWAT-PT-MK	11.26%	0.79	0.78	23.49%	0.85	0.74
SWAT-PT-VS	22.94%	0.71	0.7	46.82%	0.77	0.5
Ensemble Mean	14.60%	**0.81**	0.79	27.34%	0.86	0.84
BMA mean	**0.00%**	**0.81**	**0.81**	**3.07%**	**0.87**	**0.86**

Table 3 Calibrated parameter values for the six models in LREB.

Parameter \ Model	SWAT_ HG_MK	SWAT_ HG_VS	SWAT_ PM_MK	SWAT_ PM_VS	SWAT_ PT_MK	SWAT_ PT_VS
CN	9%	-17%	8%	-17%	6%	20%
ESCO	0.46	0.89	0.88	0.91	0.38	0.78
Surlag	9.99	2.8	9.78	1.1	9.69	2.3
ALPHA_BF	0.23	0.61	0.17	0.45	0.37	0.55
GW_REVAP	0.15	0.15	0.2	0.2	0.08	0.1
SOL_AWC	7%	-20%	18%	16%	18%	25%
CH_K2	144	147	146	130	131	147
GW_DELAY	22.57	3.7	18.87	2.19	22.8	3.07
RCHRG_DP	0.79	0.01	0.66	0.45	0.33	0.68
GWQMN	9.16	103.44	45.91	103.69	95.14	168.81
REVAPMN	215.14	24.59	486.46	402.32	263.62	190.1

Table 4 Evaluation coefficients for the three SWAT models, arithmetic mean, and BMA mean in the YRHB for both calibration and validation periods.

Models \ Coefficients	Calibration			Validation		
	PBIAS	R^2	NSE	PBIAS	R^2	NSE
SWAT-DD	-17.71%	0.82	0.77	-17.98%	0.84	0.78
SWAT-ELEV	-4.63%	0.85	0.84	**-0.31%**	0.83	0.83
SWAT-SNOW17	4.76%	0.87	0.84	7.12%	0.85	0.78
Ensemble Mean	-5.86%	**0.88**	0.87	-3.72%	**0.87**	**0.87**
BMA Mean	**0.00%**	**0.88**	**0.88**	3.71%	**0.87**	**0.87**

Table 5 Calibrated parameter values for the three models in YRHB.

Model \ Parameter	CN	ESCO	Surlag	ALPHA_BF	GW_REVAP
SWAT-DD	14%	0.28	4.90	0.16	0.03
SWAT-ELEV	7%	0.36	3.80	0.33	0.04
SWAT_SNOW17	2%	0.18	7.40	0.51	0.08

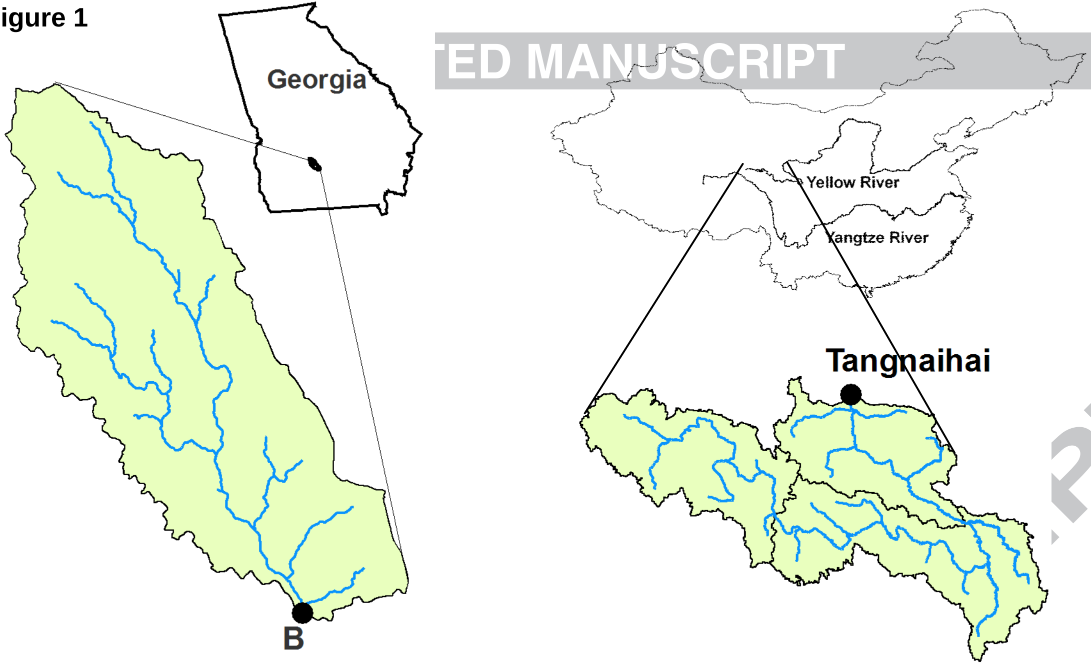
Table 6 Evaluation coefficients obtained using different number of candidate models in BMA in the LREB.

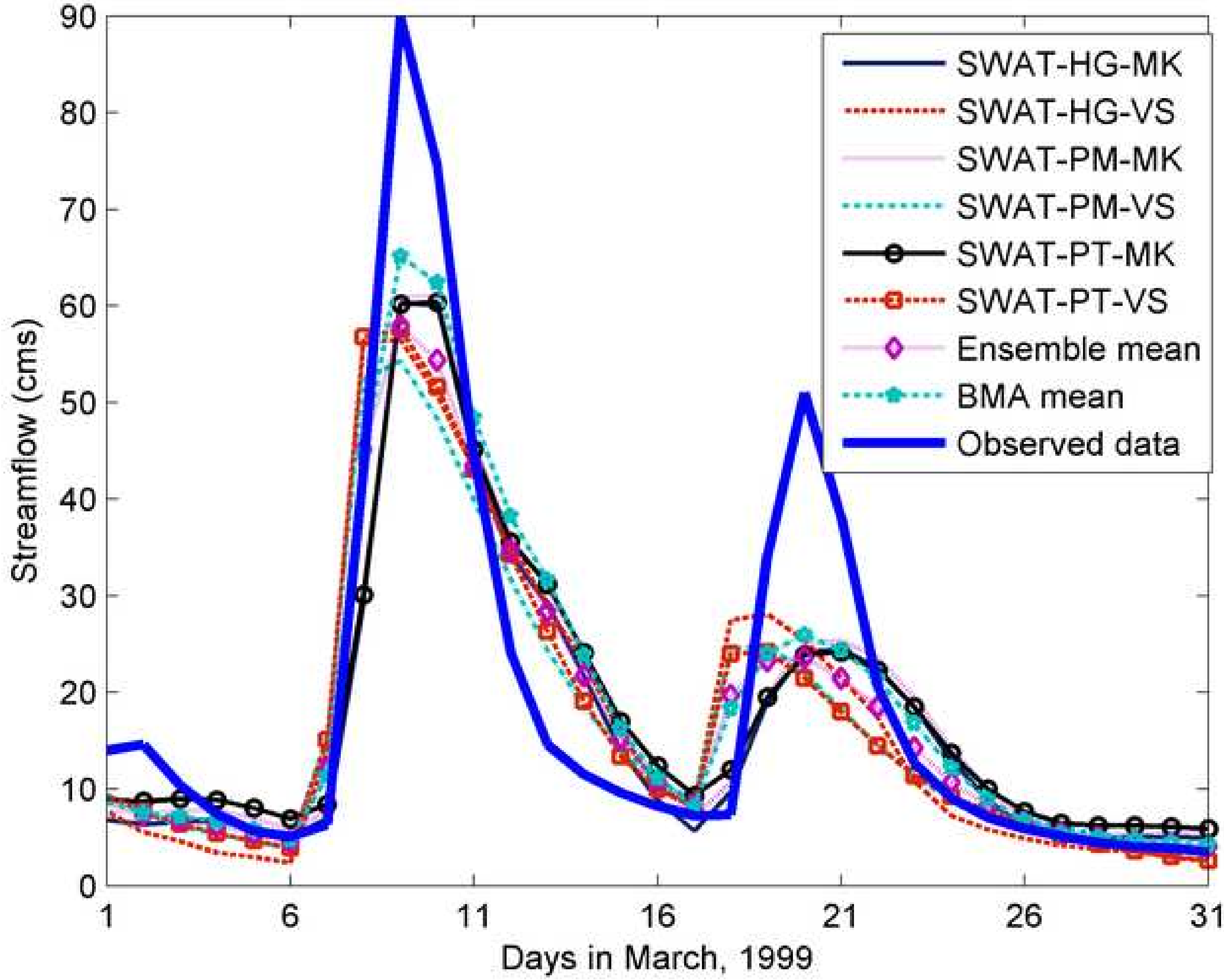
Coefficients	Calibration					Validation				
Number of Candidate models	PBIAS	R^2	NSE	66.7% POC	90% POC	PBIAS	R^2	NSE	66.7% POC	90% POC
6	0.00%	0.81	0.81	76.04%	91.14%	3.07%	0.87	0.86	74.41%	96.53%
5	0.00%	0.81	0.81	75.34%	91.32%	1.02%	0.86	0.86	74.15%	92.93%
4	0.00%	0.81	0.8	74.33%	90.99%	1.32%	0.86	0.86	73.89%	94.94%
3	0.00%	0.80	0.8	73.96%	91.71%	3.21%	0.86	0.85	73.36%	94.01%
2	0.00%	0.80	0.8	75.89%	93.15%	10.58%	0.85	0.84	77.02%	95.08%

ACCEPTED MANUSCRIPT

List of Figures

Figure 1




SWAT-HG-MK
SWAT-HG-VS
SWAT-PM-MK
SWAT-PM-VS
SWAT-PT-MK
SWAT-PT-VS
Ensemble mean
BMA mean
Observed data
Streamflow (cms)
Days in March, 1999
0
10
20
30
40
50
60
70
80
90
1
6
11
16
21
26
31

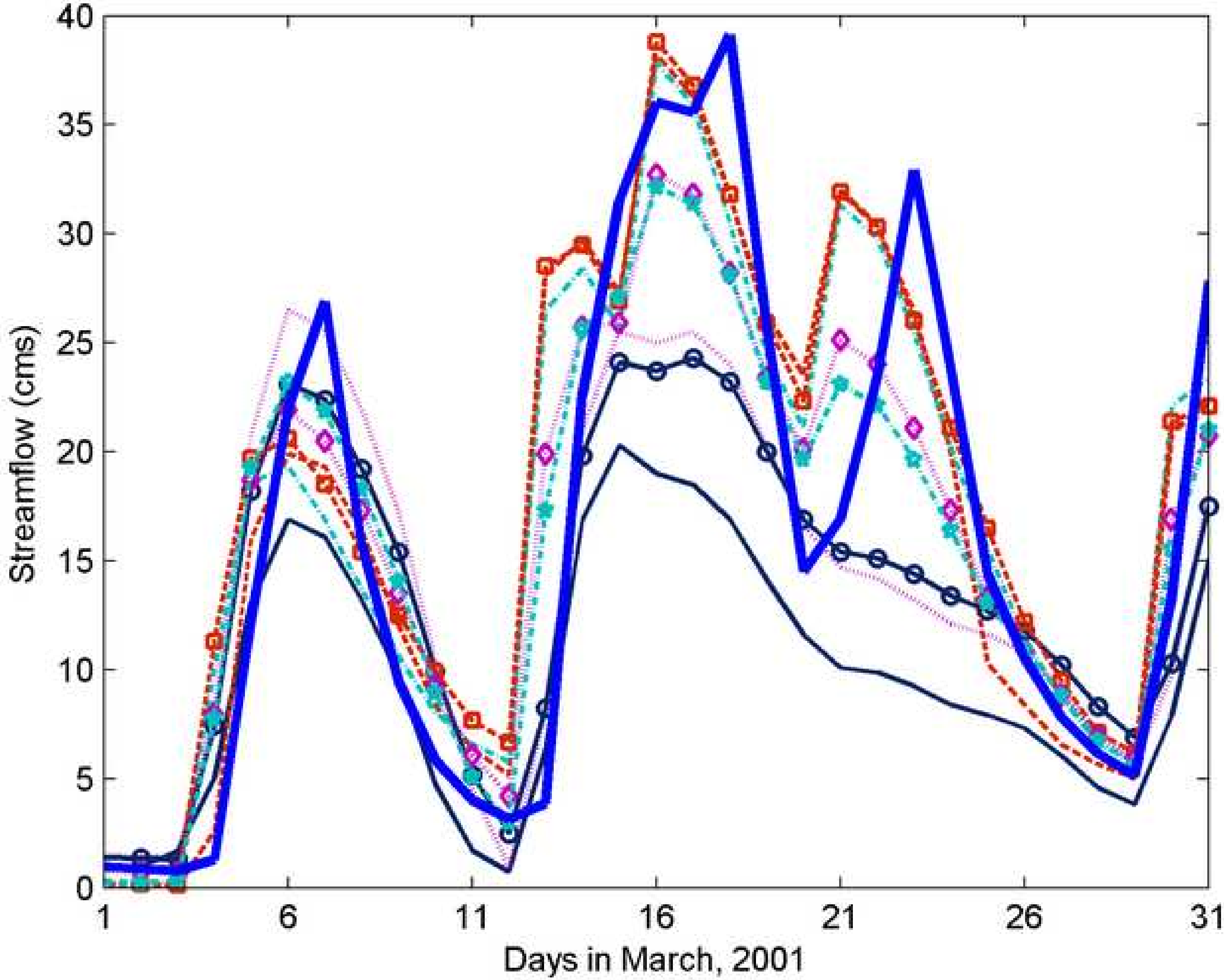
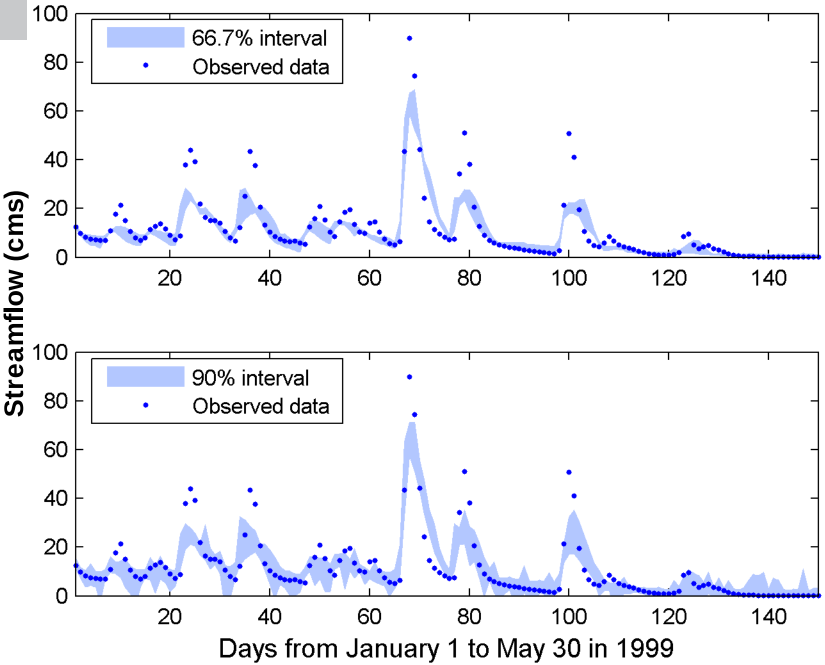
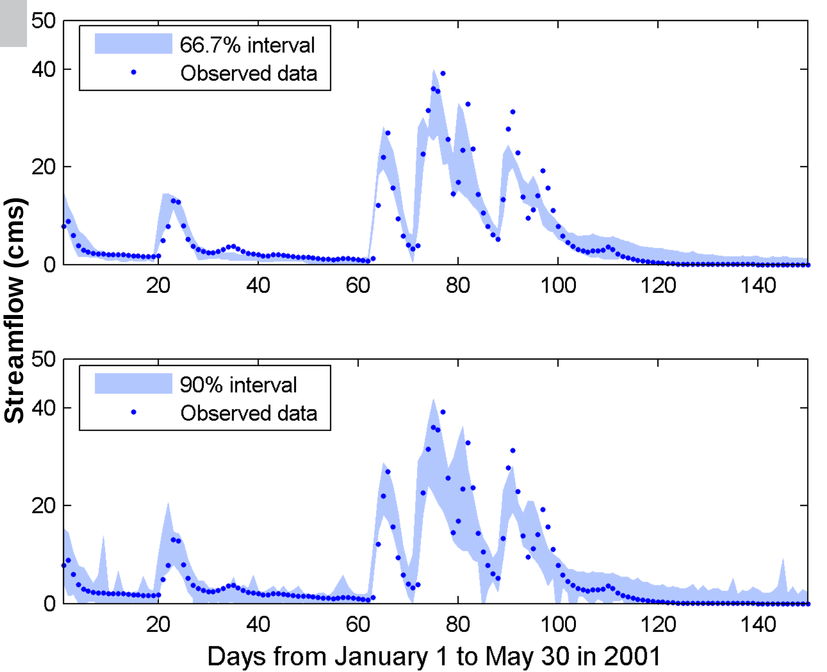

Streamflow (cms)
Days in March, 2001
0
5
10
15
20
25
30
35
40
1
6
11
16
21
26
31

Figure 4

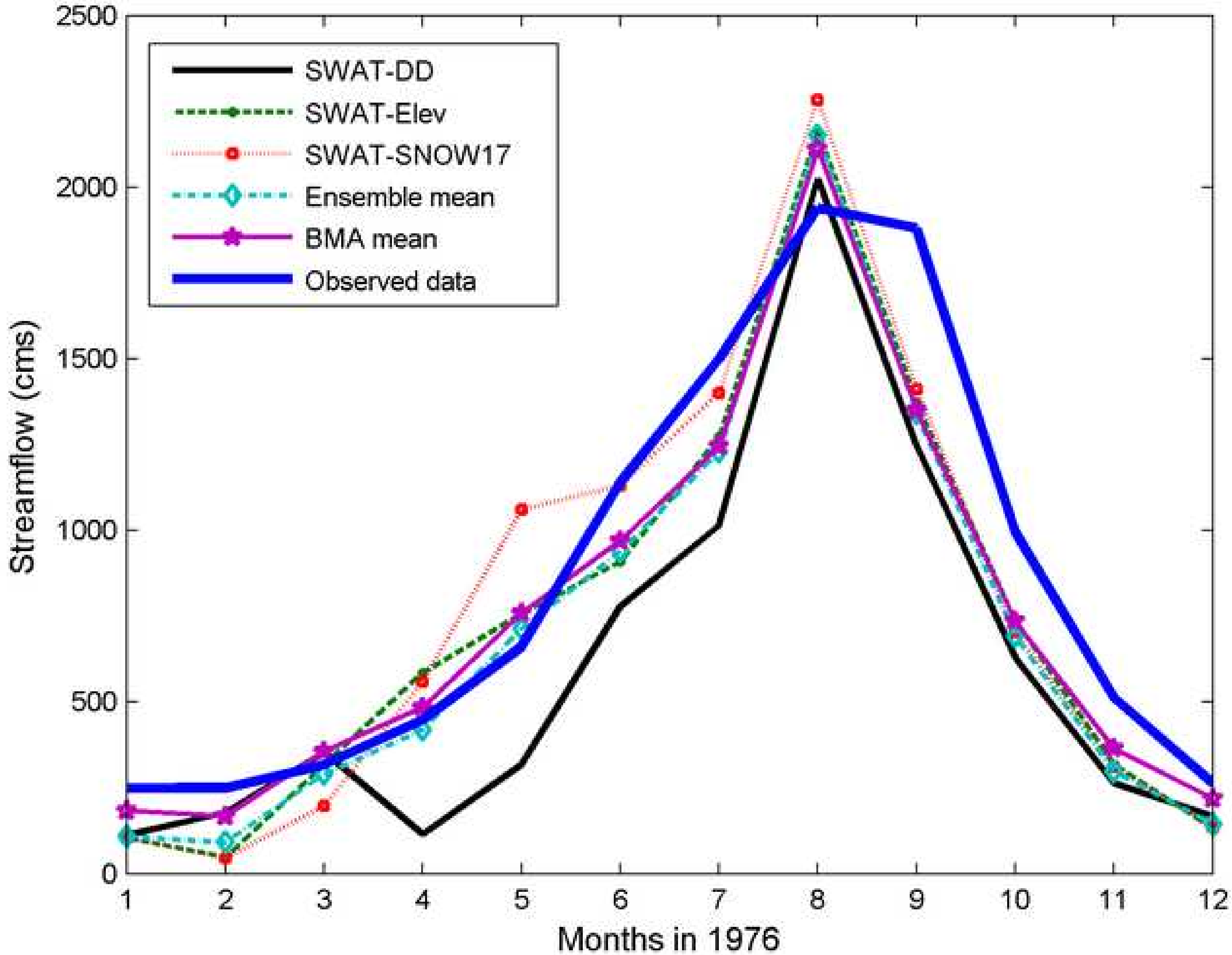


66.7% interval
Observed data
90% interval
Observed data
Streamflow (cms)
Days from January 1 to May 30 in 1999
0
20
40
60
80
100
20
40
60
80
100
120
140

Figure 5



66.7% interval
Observed data
90% interval
Observed data
Streamflow (cms)
Days from January 1 to May 30 in 2001
0
20
40
50
60
80
100
120
140


SWAT-DD
SWAT-Elev
SWAT-SNOW17
Ensemble mean
BMA mean
Observed data
Streamflow (cms)
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7
8
9
10
11
12
Months in 1976

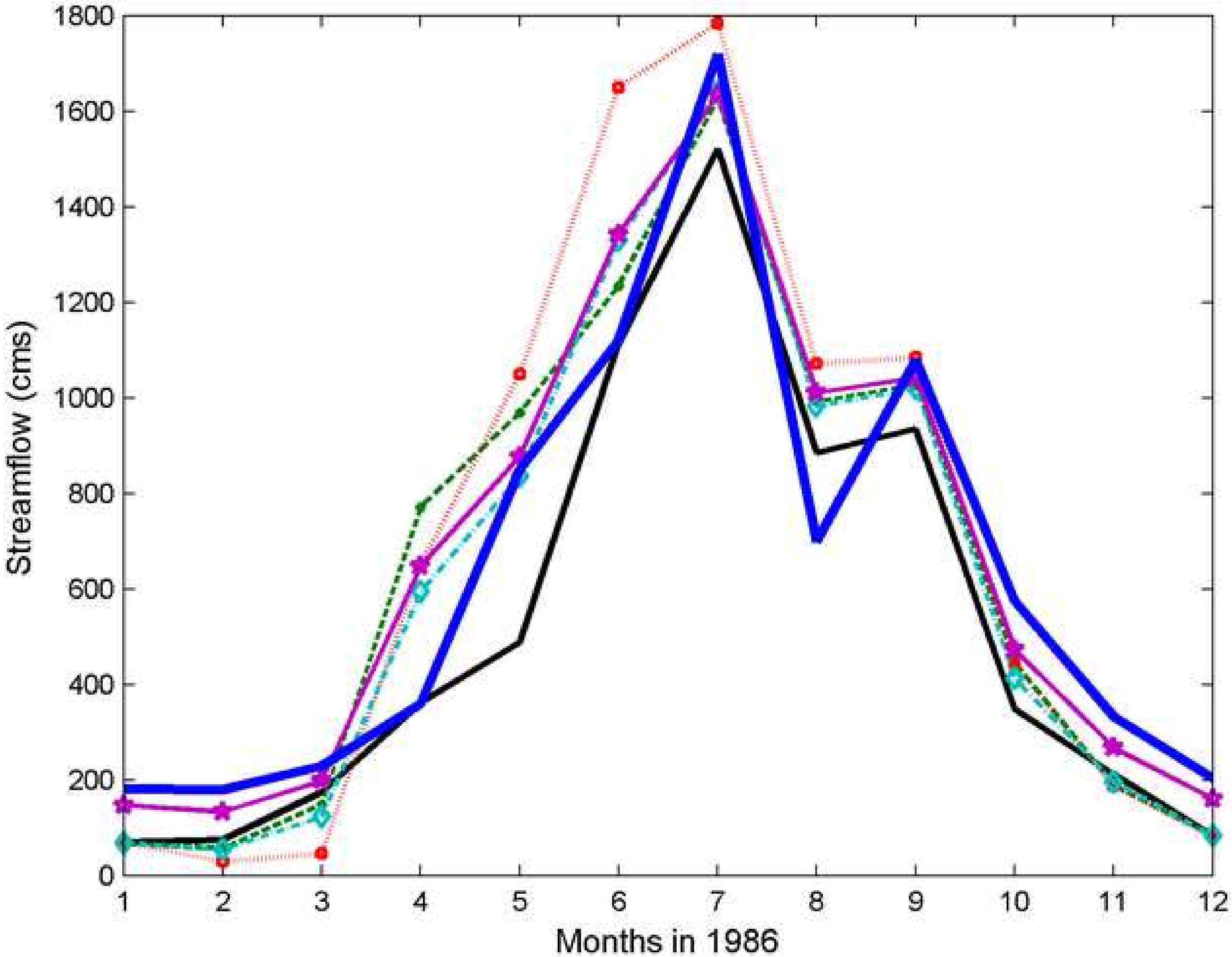
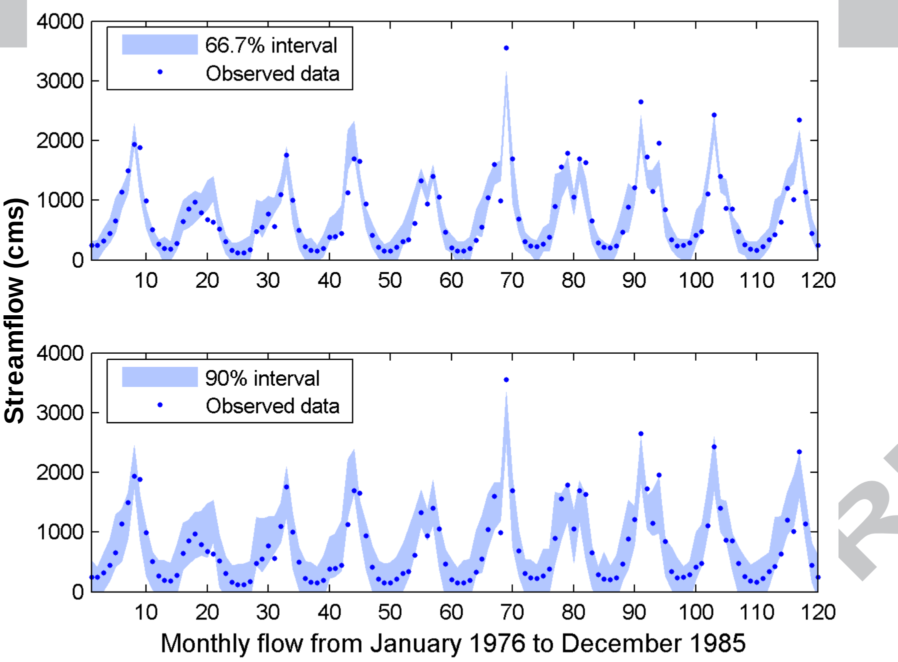
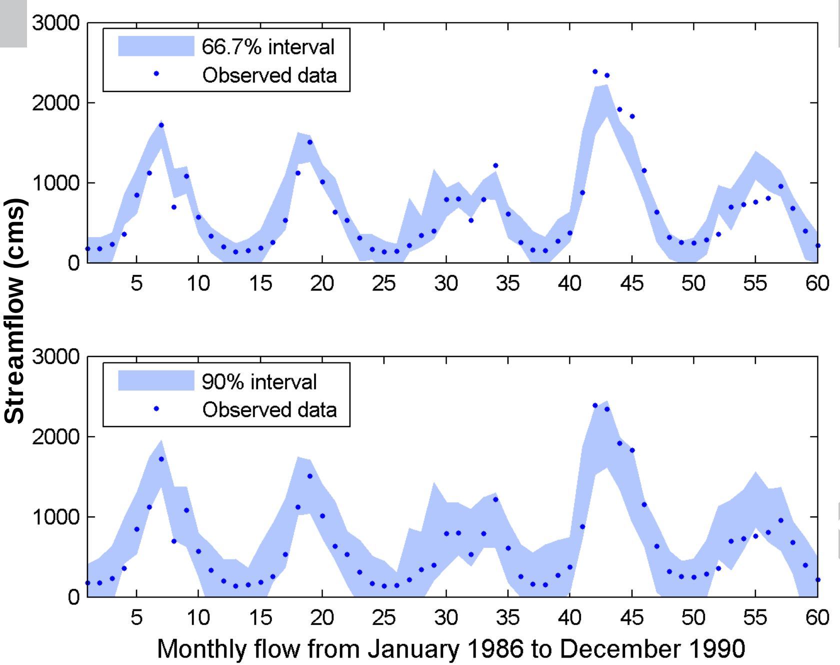

1800
1600
1400
1200
1000
800
600
400
200
0
Streamflow (cms)
1
2
3
4
5
6
7
8
9
10
11
12
Months in 1986

Figure 8



66.7% interval
Observed data
90% interval
Observed data
Streamflow (cms)
0
1000
2000
3000
4000
10
20
30
40
50
60
70
80
90
100
110
120
Monthly flow from January 1976 to December 1985

Figure 9



3000
2000
1000
0
66.7% interval
Observed data
5
10
15
20
25
30
35
40
45
50
55
60
90% interval
Observed data
Streamflow (cms)
Monthly flow from January 1986 to December 1990