The published article can be found at: https://doi.org/10.1016/j.envsoft.2020.104735

Uncertainty Quantification in Reconstruction of Sparse Water Quality Time Series: Implications for Watershed Health and Risk-Based TMDL Assessment

Ganeshchandra Mallya[1], Abhinav Gupta[1,*], Mohamed M. Hantush[2], Rao S. Govindaraju[1]

[1]Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, USA

[2] Center for Environmental Solutions and Emergency Response, U.S., Cincinnati, OH, USA

*Corresponding Author:

Abhinav Gupta
+17654046047
550 W Stadium Ave
West Lafayette, IN 47907
gupta353@purdue.edu

Abstract

Despite the plethora of methods available for uncertainty quantification, their use has been limited in the practice of water quality (WQ) modeling. In this paper, a decision support tool (DST) that yields a continuous time series of WQ loads from sparse data using streamflows as predictor variables is presented. The DST estimates uncertainty due to residual errors using a relevance vector machine. To highlight the importance of uncertainty quantification, two applications enabled within the DST are discussed. The DST computes (i) probability distributions of four measures of WQ risk analysis- reliability, resilience, vulnerability, and watershed health- as opposed to single deterministic values and (ii) concentration/load reduction required in a WQ constituent to meet total maximum daily load (TMDL) targets along with the associated risk of failure. Accounting for uncertainty reveals that a deterministic analysis may mislead about the WQ risk and the level of compliance attained with established TMDLs.

Keywords: Decision support tool, Water quality risk analysis, TMDL, Relevance vector machine, Uncertainty quantification, LOADEST

Software and data availability

Name of the software	Web-based Decision Support Tool
Programming languages	MATLAB, JavaScript and PHP
Name of the Dataset available	**A Decision Support Tool for Water Quality Modeling**
Developer and contact information	Ganeshchandra Mallya, gmallya@purdue.edu
Year First available	2018
Software required	Any web browser
Availability	Through the URL https://engineering.purdue.edu/WaterDST/StandaloneTool/
User manual	Available online through the 'User Manual' tab of the Decision Support Tool
Cost	Free

1. Introduction

Environmental decisions are often based upon mathematical models of the system under consideration (Beven, 2007; Refsgaard et al., 2006). For example, in total maximum daily load (TMDL, developed by United States Environmental Protection Agency, USEPA) development, models such as SWAT (Soil and water assessment tool, Arnold and Allen, 1999) or HSPF (Hydrological Simulation Program Fortran; Jia and Culver, 2006) are frequently used for simulation of water quality (WQ) constituents (e.g., Indiana Department of Environmental Management, IDEM, 2017). Typically, the parameters of a model are calibrated against available observations, and many observations are required to calibrate a complex model like SWAT. In some applications, continuous time series of streamflow observations and WQ constituent concentrations are required to assess the health of an impaired water body. Examples include (a) identification of sources of pollution in a waterbody (Mallya et al., 2018) and (b) computation of load reduction required (LRR) to restore a waterbody to healthy conditions (Park et al., 2015). Whereas streamflows are measured frequently in a watershed, WQ constituent such as suspended-solids, nitrogen, and phosphorus concentrations are measured sparsely (e.g., biweekly and only in summer months). Sparse WQ data are not amenable to direct use in reliable decision making (Kjeldsen and Rosbjerg, 2004). Therefore, various models have been developed for temporal *reconstruction* of WQ data (e.g., load estimator (LOADEST) Runkel et al., 2004). In this study, reconstruction refers to the estimation of WQ constituent concentration/load values at time-steps where WQ data is unavailable using the observed WQ constituent data.

Any mathematical representation of an open system, such as the ones encountered in environmental modeling, incur uncertainties due to lack of complete knowledge about the system, inadequate representation of dominant processes through mathematical equations, erroneous data used for parameter estimation, and difficult-to-represent local characteristics of the system (Beven, 2007). These uncertainties in the modeling process should be considered to make informed management decisions (Beven, 2007). The quantification of uncertainties in hydrologic and WQ modeling is carried out using probabilistic methods (see Ahamadisharaf et al., 2019 for a review). In a typical probabilistic analysis, the residual time series (the difference between observed and simulated response of the system) is assumed to follow a probability distribution. The parameters of the probability distribution are estimated against observed residuals time series, and, subsequently, the calibrated probability distribution is used for uncertainty quantification. The residual time series is an aggregate of *measurement errors*, *structural errors*, and errors in the numerical implementation of the model. Measurement errors refer to errors in the measurement of streamflow and WQ data. Structural errors exist because a model is an approximation of reality. Calibrated model parameters also incur uncertainty which is referred to as parametric uncertainty. Parametric uncertainty exists due to measurement errors, structural errors, and limited information in the data to calibrate the parameters. A modeler should ensure that errors in the numerical implementation are negligible. Thus, the residual time series is an aggregate of measurement and structural errors. Despite growing awareness about the importance of uncertainty due to structural errors (Brynjsdottir and O'Hagan, 2014), measurement errors (Baldassarre and Montanari, 2009), unknown parameters (Melching and Bauwens, 2001) and residual errors (Beven and Binely, 1992; Borsuk and Stow, 2000; Borsuk et al., 2002; Chaudhary and Hantush, 2017; Hoque et al., 2012; Hantush and Chaudhary, 2014), uncertainty is rarely quantified in *practice* of WQ modeling. For example, in TMDL applications, the current practice is to use a margin of safety (MOS) to account for uncertainty in the relationship between the pollutant load and the quality of the receiving water body (Novotny, 2002). The MOS is typically assigned by making conservative assumptions or specified explicitly as a percentage (e.g., 5-10%) of the TMDL (NRC, 2001). Recently, Nunoo et al. (2020) found that, in 84% of the 37,841 TMDLs reported, uncertainty analysis was not carried out to select a margin of safety (MOS). Subjective or arbitrary specification of MOS might lead to overly conservative estimates and increased cost of implementation of pollution control measures (Zhang and Yu, 2004).

In WQ modeling, the pervasiveness of uncertainty has been long recognized (Beck, 1987) followed by several efforts to quantify it (e.g., Ahamadisharaf and Benham, 2020; Borsuk, 2003; Chaudhary and Hantush, 2017; Hoque et al., 2012; Jia and Culver, 2008; Reckhow, 2003; Shirmohammadi et al., 2006; Zhang and Yu, 2004; Zheng et al., 2011; Zheng and Han, 2016; and Zheng and Keller, 2008). Uncertainty quantification for complex models tend to be complicated, time-consuming, and computationally demanding (see Smith et al., 2014, chap. 2 for some examples). Thus, researchers have sought simpler statistical models for simulation of WQ time series (e.g., LOADEST).

For both physical and statistical models, a rich theory has been developed to quantify uncertainty by residual analysis (e.g., Smith et al., 2015). However, uncertainty quantification is frequently avoided in practice for the following reasons (Pappenberger and Beven, 2006): (1) it is subjective, too difficult to perform and cannot be incorporated into decision making; (2) it is not required if one uses physically realistic models; and (3) it is too difficult for policy-makers to understand and does not really matter in making the final decision. Pappenberger and Beven (2006) further argued that the reasons cited above are untenable, and they emphasized the importance of an open discourse of uncertainty in environmental models. Reckhow (2003) pointed out that modelers should clearly communicate the uncertainties associated with their models to decision-makers. A solution to the problem 'uncertainty analysis is too difficult to perform' is availability of easy-to-use software packages that can be used by practitioners with little effort (e.g., Gronewold and Borsuk, 2009). In this paper, we present one such software through a decision support tool (DST).

The DST reconstructs WQ constituent time series by using streamflow values as predictor variables and employing the state-of-the-art relevance vector machine (RVM; Tipping, 2001) that can accommodate nonlinear transformations between streamflow and WQ data. Moreover, the RVM provides uncertainty estimates that are conditioned on streamflows and can account for errors in streamflows (not explored here). The DST uses the reconstructed WQ time series along with the uncertainty estimates for the following two applications:

(1) WQ risk assessment by computing indices such as reliability, resilience, vulnerability, and composite watershed health index (Hoque et al., 2012, Mallya et al., 2018).
(2) Computation of LRR of a WQ constituent so that the TMDL criterion is met with an acceptable risk of noncompliance (Camacho et al., 2018).

Specifically, the DST provides a probabilistic estimate of watershed health and the required reduction in pollutant concentration/load as a function of the risk of violating TMDL criteria. Fig. 1 shows the overview of the three tasks carried out by the DST. In this study, using the St. Joseph River Watershed (spread over parts of Indiana, Michigan, and Ohio) as a test case, the two applications illustrate the importance of uncertainty in assessing watershed health and in conducting TMDL studies. While other investigators (e.g., Borsuk et al., 2002; Hantush and Chaudhary, 2014; and Camacho et al., 2018) have demonstrated the benefits of probabilistic uncertainty estimation in TMDLs, no study has implemented such a framework at the watershed scale, to the best of our knowledge.

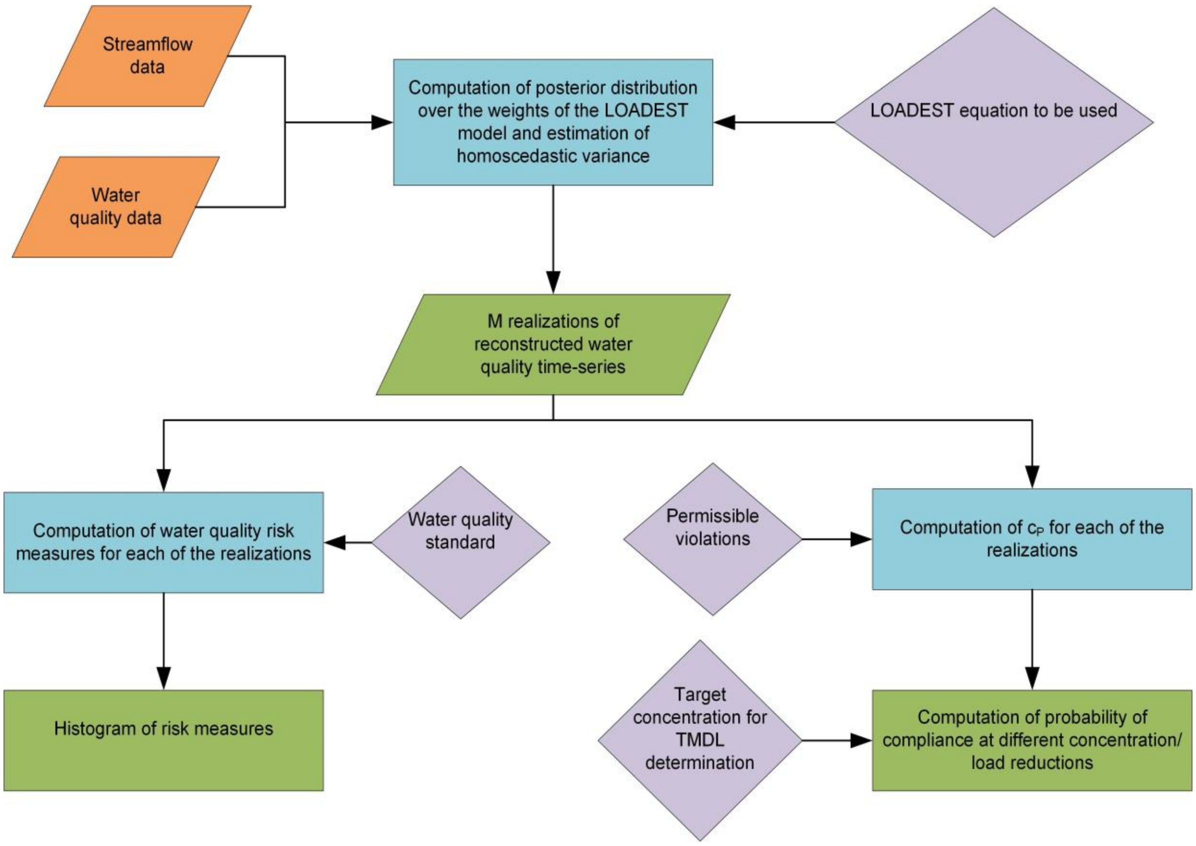

Fig. 1. A flowchart of the steps implemented in the Decision Support Tool

2. Theory

2.1 Reconstruction of water quality time series

Traditionally, simple regression equations LOADEST (Load Estimator, Runkel et al., 2004) have been used for reconstruction of WQ time series using available streamflows as predictors. However, the uncertainty associated with these estimates is rarely reported (or used). The DST uses RVM (Bishop, 2006 and Tipping, 2001) to estimate the uncertainty associated with the reconstruction. A general statistical model is used in the DST:

$$y_j = \sum_{i=1}^{N} w_i \phi_i(x_j) + \epsilon_j, \qquad j = 1,2,\dots,n \tag{1}$$

where y_j is log of load of WQ constituent at j^{th} time-step , ϕ_i is the i^{th} (linear or non-linear) basis function, x_j is streamflow at j^{th} time-step, N is the number of the basis functions, w_i is the weight of the i^{th} basis function, ϵ_j is the error in the estimation of y_j using $\sum_i w_i \phi_i(x_j)$ as a model, and n is the number of time-steps at which observed streamflows are available. The basis functions may be chosen to achieve a nonlinear transformations of streamflows into WQ constituents (Tripathi and Govindaraju, 2007). The current version of the DST uses the same transformations as the basis functions as are used in LOADEST due to the historical use of LOADEST in WQ literature and to enable a comparison with LOADEST. For example, in case of the LOADEST second equation (as listed in Table 1), the following three basis functions would be required: 1, $ln\, x$ and $ln\, x^2$.

The ϵ_j in Eq. (1) is assumed to be a zero-mean, Gaussian random variable with a homoscedastic variance σ_ϵ^2. Additionally, in RVM, each weight w_i is assigned a zero-mean Gaussian prior with variance α_i^{-1}. This specification of prior allows automatic determination of only the relevant basis functions in Eq. (1) leading to predictions that are potentially robust to errors in predictor variables. Subsequently, the posterior distribution of weight vectors $\boldsymbol{w}$ (conditioned on the vector $\boldsymbol{y}$, matrix $\Phi = [\boldsymbol{\phi}(x_1), \boldsymbol{\phi}(x_2), \dots, \boldsymbol{\phi}(x_n)]$ of predictors, parameters $\alpha_i's$, and the variance σ_ϵ^2) is computed using the Bayes theorem. The posterior distribution of $\boldsymbol{w}$ is found to be Gaussian with mean $\boldsymbol{\mu_w}$ and covariance-matrix $\Sigma_{\boldsymbol{w}}$ such that

$$\Sigma_w = (\sigma^{-2}\Phi\Phi^T + \mathrm{A})^{-1}, \tag{2}$$
$$\boldsymbol{\mu_w} = \sigma^{-2}\Sigma_w \Phi \boldsymbol{y}, \text{ and}$$
$$A = diag(\alpha_1, \alpha_2, \dots, \alpha_N).$$

The predictive distribution at time-step *t*, given streamflow x_t is then Gaussian with mean

$$\mu_{y_t} = \boldsymbol{\mu_w^T}\boldsymbol{\phi}(x_t), \tag{3}$$

and covariance between predicted log of load at t_1 and t_2

$$\Sigma_{t_1,t_2} = \sigma_\epsilon^2 \delta(t_1 - t_2) + \boldsymbol{\phi}(x_{t_1})^T \Sigma_w \boldsymbol{\phi}(x_{t_2}), \tag{4}$$

where $\delta(t_1 - t_2)$ is the Dirac-delta function. Equation (4) shows that the uncertainty in reconstruction at time-step *t* is dependent upon predictor x_t and covariance matrix $\Sigma_{\boldsymbol{w}}$ of weight vector $\boldsymbol{w}$. For convenience in applications, the posterior distribution over the parameters $\alpha_i's$ and σ_ϵ^2 is approximated by the Dirac-delta function $\delta(\boldsymbol{\alpha_{MAP}}, \sigma_{\epsilon,MAP}^2)$, where $\boldsymbol{\alpha_{MAP}}$ and $\sigma_{\epsilon,MAP}^2$ are the maximum posterior estimates of these parameters.

The DST requires daily streamflow time series and measurements of the WQ constituent as inputs by the user. If the streamflow gauge and WQ monitoring stations are not co-located, it uses the watershed area-ratio method (section 2.4) for the estimation of streamflow at the WQ monitoring station. It allows the users to select any one of the nine LOADEST equations (Table 1), or to pick the best LOADEST equation based on Akaike Information criterion (AIC; Akaike, 1973) if desired. To represent the uncertainty in reconstructed WQ time series, it draws 10000 Monte Carlo (MC) samples from the logarithm of load estimated as a multivariate Gaussian distribution (Fang and Zhang, 1990). The mean and covariance matrix of the Gaussian distribution are given by Eqs. (3) and (4), respectively. The 10000 MC simulations were found sufficient to obtain stable estimates of lower and upper bounds of 90% and 95% prediction intervals/credible regions, WQ risk measures (section 2.2), and TMDL compliance plots (section 2.3). Subsequently, the DST computes 90% and 95% credible regions as follows. The 90% credible region is the region bounded by 5^{th} and 95^{th} percentiles of MC samples; the percentiles are computed at each time-step. The 95% credible region is the region bounded by 2.5^{th} and 97.5^{th} percentiles of MC samples.

Even though DST uses the same basis-functions in RVM as those in LOADEST, a significant difference between the RVM (as employed in the DST) and LOADEST (as employed by Park et al., 2015) exists in the parameter estimation method. LOADEST uses adjusted-maximum-likelihood estimation (AMLE) to estimate the weight vector $\boldsymbol{w}$. In RVM, the choice of the prior $N(0, \alpha_i^{-1})$ on i^{th} weight w_i expresses a preference for smaller weights (Tripathi, 2009, pp. 13).: The smaller weights dampen observation errors in predictors as in LASSO and ridge regression (Friedman et al., 2001). In some cases, the parameter α_i^{-1} may converge to zero thus eliminating the i^{th} basis function from the set of predictor variables; the reduced set of predictor variables results in a computationally efficient model (Bishop, 2006 and Tipping, 2001).

Table 1. LOADEST equations

	LOADEST equations
1.	$ln\,L = w_0 + w_1 lnQ$
2.	$ln\,L = w_0 + w_1 lnQ + w_2 lnQ^2$
3.	$ln\,L = w_0 + w_1 lnQ + w_2 \delta t$
4.	$ln\,L = w_0 + w_1 lnQ + w_2\, sin(2\pi\delta t) + w_3\, cos(2\pi\delta t)$
5.	$ln\,L = w_0 + w_1 lnQ + w_2 lnQ^2 + w_3 \delta t$
6.	$ln\,L = w_0 + w_1 lnQ + w_2 lnQ^2 + w_3\, sin(2\pi\delta t) + w_4\, cos(2\pi\delta t)$
7.	$ln\,L = w_0 + w_1 lnQ + w_2\, sin(2\pi\delta t) + w_3\, cos(2\pi\delta t) + w_4 \delta t$
8.	$ln\,L = w_0 + w_1 lnQ + w_2 lnQ^2 + w_3\, sin(2\pi\delta t) + w_4\, cos(2\pi\delta t) + w_5 \delta t$
9.	$ln\,L = w_0 + w_1 lnQ + w_2 lnQ^2 + w_3\, sin(2\pi\delta t) + w_4\, cos(2\pi\delta t) + w_5 \delta t + w_6 \delta t^2$

δt = time in decimal units since first Julian of a year; for example, 31 Jan, 2005 is represented as 2005.085
L = load (Kg day^{-1})
Q =streamflow in $m^3 s^{-1}$ or $ft^3 s^{-1}$

2.2 Water quality risk analysis

The health of a watershed is quantified by using the following three measures (Hoque et al., 2012): reliability (R), resilience (R), and vulnerability (V). Additionally, a composite watershed health measure can be computed as a function of R-R-V (Mallya et al., 2018). Suppose, Y_t is the concentration or load of the reconstructed WQ constituent at time-step t with standard numerical target denoted by Y_t^* (concentration or load). The reliability (ρ) is defined as the probability of the waterbody being in the compliant state, that is,

$$\rho = P\{Y_t \in S\} = 1 - P\{Y_t \in F\}, \tag{5}$$

where $P\{\bullet\}$ denotes the probability of the event $\{\bullet\}$, S denotes the event $\{Y_t \leq Y_t^*\}$ denoting the safe/compliant state, and F denotes the event $\{Y_t > Y_t^*\}$ denoting failed/noncompliant state. The definitions of the compliant and non-compliant state will be reversed in case of dissolved oxygen, that is, $S = \{Y_t \geq Y_t^*\}$ and $F = \{Y_t < Y_t^*\}$. Given the time series of WQ constituent, the DST estimates ρ as

$$\rho = 1 - \frac{1}{n}\sum_{t=1}^{n} z_t, \tag{6}$$

where $z_t = 1$, when $Y_t \in F$ and $z_t = 0$, when $Y_t \in S$ and n is the total number of time-steps. Resilience (r) is defined as the probability of the system to recover from a non-compliant state, that is,

$$r = P\{Y_{t+1} \in S | Y_t \in F\}, \tag{7}$$

and can be estimated as

$$r = \frac{\sum_{t=1}^{n} u_t}{\sum_{t=1}^{n} z_t}, \tag{8}$$

where $u_t = 1$ when $Y_{t+1} \in S$ and $Y_t \in F$ and 0 otherwise, and z_t as defined above. In summary, a waterbody (or watershed) is resilient if it returns from a non-compliant state to a compliant state; the longer the waterbody takes to reach a compliant state from a non-compliant state, the less resilient the waterbody.

Vulnerability is defined as the magnitude of severity of violation during a noncompliant event. For WQ violations in a waterbody, there is no universal measure to quantify the severity of a violation. Mallya et al. (2018) proposed a new measure referred to as robustness – opposite of vulnerability – that scales between 0 and 1- as

$$v_o = \left\{ \Pi_{t=1}^{n} \left(\frac{Y_t^*}{Y_t} \right)^{H[Y_t - Y^*]} \right\}^{\frac{1}{m}}, \tag{9}$$

where m is number of time-steps at which $Y_t > Y_t^*$, $H[\bullet]$ is the Heaviside function so that (9) accounts only for the noncompliant events. When the deviations of Y_t from Y^* are large then $v_o \rightarrow 0$; when deviations are small then $v_o \rightarrow 1$, which is consistent with definitions for reliability (ρ) and resilience (r). Vulnerability (v) can now be defined as:

$$v = 1 - v_o \tag{10}$$

A composite measure of watershed health (h) is defined as (Mallya et al., 2018):

$$h = (\rho \, r \, v_o)^{\frac{1}{3}} \tag{11}$$

Clearly, if $\rho = r = v_0 = 1$ then $h = 1$, i.e., the drainage area is healthy with respect to the WQ constituent of interest. Similarly, if any one of the risk-measures is 0 then $h = 0$, i.e., the drainage area is impaired with respect to WQ constituent of interest.

Reliability, resilience and vulnerability can be used to design appropriate measures to improve the WQ of a waterbody. For instance, reliability should be used as a guiding measure if frequent violations of a WQ constituent are not allowed. In cases where durations of violations are more consequential than the frequency of violations, resilience is a useful measure. Similarly, vulnerability is a useful measure when the goal is to reduce the severity of violations. Moreover, spatial distribution of these WQ risk-measures over different streams can be used to identify the critical sources of pollution in a watershed (e.g., Mallya et al., 2018). The usefulness of different WQ risk measures also depend upon the timescale of analysis. For example, in some states, E. Coli concentration should be below 235 $\mathrm{cfu/100ml}$ at least 89.5% of the time-steps at daily timescale and should be below 126 $\mathrm{cfu/100ml}$ at 100% of time-steps at monthly timescale (Ahmadisharaf and Benham, 2020). Thus, in case of E. Coli, watershed managers need to ensure that waterbody has 0.895 and 1.00 reliability at daily and monthly timescales, respectively. Current version of the DST computes WQ risk measures at the daily timescale. A future version will include WQ-risk analysis at other timescales.

Uncertainty associated with reconstructed WQ data will carry over to the risk measures also. The DST uses MC method to estimate probability distributions of the R-R-V and the WH measures. It computes R-R-V and WH for each realization of reconstructed WQ time-series, and, subsequently, constructs the histograms of these measures from the ensemble values, and tabulates the mean and standard-deviation of the measures. The DST has a Google map interface which shows the locations of USGS WQ stations and National Water Quality Assessment (NAWQA) stations; a user can click on any one these WQ stations and fill out a form to reconstruct WQ data and compute WQ risk measures.

2.3 Risk-based total maximum daily load (TMDL) analysis

In this section, we present the theory behind the determination of the LRR to meet TMDL targets. Currently, the MOS is used to account for uncertainties associated with TMDL development. For example, suppose that the concentration of total phosphorus (TP) corresponding to a TMDL is 0.08 mg L^{-1} and a deterministic model predicts that to maintain this concentration in a waterbody, the maximum allowable load in the watershed is 600 Kg day^{-1}. Then an arbitrarily selected 10% of the allowable load can be reserved for MOS. The 10% MOS serves to acknowledge that the magnitude of uncertainties in the model simulated TP are such that 540 Kg day^{-1} in watershed may also correspond to maximum TP concentration of 0.08 mg L^{-1}. However, the protective cushion provided by the 10% of the allowable load remains at best unknown. A small MOS may result in non-attainment of the water quality standard, but a large MOS can be inefficient and costly (Novotny, 2003). Therefore, a realistic estimate of uncertainty is required. The DST accommodates the uncertainty through MC simulations that yield an ensemble of M realizations of the WQ time series.

Borsuk et al. (2002) presented a framework for a probabilistic TMDL development. Let C^* be the user-defined target TMDL concentration of a waterbody, and suppose that the waterbody violates the TMDL standard at most p fraction of the times during the period of analysis. Then, one can define the probability of compliance (κ_p) to permissible violations p as:

$$\kappa_p = P\{C_p \leq C^*\} = F_{C_p}(C^*), \tag{12}$$

where C_p denotes the concentration value that is exceeded p fraction of the times in a realization of WQ time series; p is the permissible fraction of violations (0.05, 0.10, …etc.). The quantity C_p

is a random variable that represents uncertainty associated with reconstructed WQ time series. The quantity κ_p is the fraction of WQ constituent time series generated by MC method that are compliant. The definition of κ_p would be reversed in case of dissolved oxygen, i.e. $\kappa_p = P\{C_p \geq C^*\}$. The distribution of C_p is determined by MC method as follows.

For a reconstructed WQ constituent time series, a value of c_p is determined using

$$c_p = G^{-1}(1 - p), \tag{13}$$

where c_p is the $100(1 - p)^{th}$ percentile of a reconstructed WQ time series, G is the empirical cumulative distribution function (CDF) of reconstructed WQ constituent time series, and p is the fraction of permissible violations (0.05, 0, 0.1, 0.15, 0.20 etc.). The quantity c_p may be interpreted as threshold value to be compared against the target TMDL concentration that would ensure compliance of a WQ time series realization. If c_p is below or equal to C^* then the WQ time series is already compliant; if the c_p is above C^* then a concentration reduction of $c_p - C^*$ is required for the WQ time series to be compliant. (Required concentration reduction is C^*- c_p for dissolved oxygen). Note that the G is estimated for each of the WQ concentration time series as if the concentration values at different time-steps are independent draws of random variable with the distribution function G. The M realizations of reconstructed WQ time series will yield M values of c_p and, in turn, a distribution of c_p. Subsequently, Eq. (12) is used to compute the probability of compliance. The probability of non-compliance (β_p) for a given fraction of permissible violations (p) is defined as

$$\beta_p = 1 - \kappa_p. \tag{14}$$

Figure 2 shows a graphical illustration of computation of κ_p for a given value of concentration/load reduction. Fig 2a shows the ensemble of distribution functions G of the 10000 reconstructed TP concentration time series without any concentration/load reduction. At $p = 0.05$, the value of c_p obtained for one of the realizations of TP time series is shown. Fig 2b shows the histogram of c_p values obtained in this manner from the ensemble of reconstructed WQ time series. Clearly, all the c_p values are above 0.08 mg L^{-1}; therefore, κ_p is zero. Fig. 2c shows the histogram of c_p values after concentration reduction; the dark green area corresponds to the fraction of the ensemble time series that violates the TMDL criterion 0.08 mg L^{-1} more than 0.05 fraction of of the times after the concentration reduction. The DST lists percentage compliance (κ) with different values of permissible violations (p) at different load and concentration reduction in two tables. Hereafter, we drop the subscript p from κ and β for brevity.

The DST computes load reduction required (LRR) at a daily timescale. If the waterbody is required to be compliant at monthly or annual timescales, then LRR should be computed at monthly or annual timescales, respectively, and the WQ and TMDL time series should be aggregated from daily to monthly or annual timescales. Another way of computing LRR is to compute the difference between *average daily load* and *average TMDL load*. Average daily load is the average of the load over the entire time-period of analysis, and, similarly, the average TMDL is the average of the TMDL over the entire the period of analysis. The DST also computes average pollutant load in the waterbody and the difference between average daily load and average TMDL load. We note that the web based LOADEST tool computes the LRR by computing the difference between average simulated load and TMDL load but not the LRR at daily timescale.

A TMDL is established based on a threshold concentration below which a waterbody is no longer impaired. The consequences of water pollution are generally tied to concentration of pollutant in the water column rather than the total load carried by the waterbody. The value of concentration in a waterbody, however, depends upon the load introduced into it. Based on the application, either load and/or concentration reductions could be important. For example, a lake (especially a closed lake) ecosystem will be affected by both the load of TP in the lake-bed and concentrations of TP in the water column; but a river draining into the lake is likely to be affected only by concentration of the TP in the water column. In a river, high load of TP with high streamflow may result in low concentrations which will not affect the river ecosystem but could result in high load of TP to a receiving lake which will affect the lake ecosystem. The DST reports reductions required both in terms of constant concentration and constant load. Constant load reduction implies that measures are taken to reduce the load in the waterbody by the same amount at each day, and same for constant concentration reduction. However, reductions would be required only during periods when the waterbody violates the TMDL criterion. If nutrient violations are seasonal, then targeted pollution control measures only during these periods might achieve compliance.

The DST allows different values of the WQ standard (used in WQ risk analysis) and TMDL concentration (used in TMDL development). The distinction might be useful when WQ standard, usually determined by the federal agencies such as USEPA, cannot be met with available resources so that a higher (in case of nutrients) or lower (in case of dissolved oxygen) TMDL concentration value must be used.

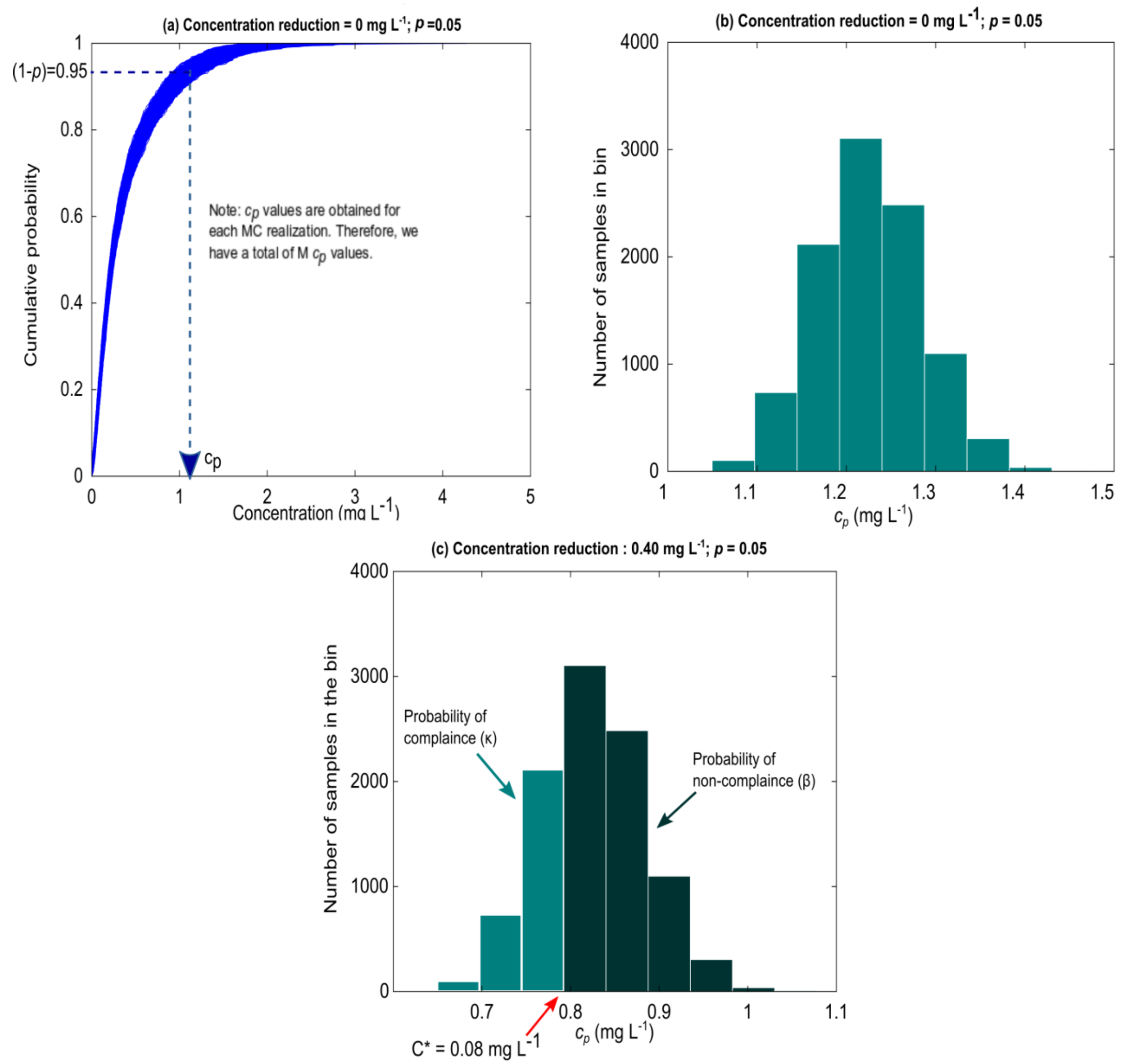

Fig. 2. Illustration of risk-based total maximum daily load (TMDL) analysis for total phosphorus (TP) with target TMDL concentration (C^*) of 0.08 mg L^{-1}: (a) calculating $c_p = F^{-1}(1-p)$ for one of the realizations of reconstructed water quality (WQ) series with no load or concentration reduction and $p = 0.05$, (b) histogram of c_p values before any load reduction, and (c) calculating the probability of compliance (κ) from c_p values that were obtained after a load reduction of 210 kg day^{-1} from the realizations of reconstructed WQ time series.

2.4 What if streamflow data is not available at a water quality monitoring station?

Often, WQ monitoring stations in a river-network are not co-located with a streamflow gauge, but a streamflow gauge may be available in proximity of a WQ monitoring station at a downstream or upstream location in the river-network. In this case, the DST estimates streamflows, Q_u, at the ungauged station as (Emerson et al., 2005; Ries, 2007)

$$Q_u = \left(\frac{A_u}{A_g}\right)^b Q_{g,} \tag{15}$$

where A_u and A_g are the areas draining to the ungauged and gauged stations; Q_g is the measured flow rate at the streamflow gauge;, and b is the exponent that varies with geographic region and climate but may be assumed to be equal to 1 when unknown. Emerson et al. (2005) reported the values of b equal to 0.85, 0.91, and 1.02 for winter, spring, and summer seasons, respectively, in Red River of the North Basin (in North Dakota and South Dakota). Typically, $b = 1$ is accurate for mean-annual flows (Rodriguez-Iturbe and Rinaldo, 2001, chap. 1). The factor $\left(\frac{A_u}{A_g}\right)^b$ in Eq. (15) is the watershed area ratio that must be supplied by the user. When streamflow measurements are available at the WQ sampling site, this conversion factor will be equal to unity. When the sampling site is located upstream of a streamflow gauge, the conversion factor will be less than unity; when the sampling site is located downstream of a streamflow gauge, the conversion factor will be greater than unity (Emerson et al., 2005; Ries, 2007).The watershed area ratio method assumes that ratio of total volume of water flowing through the ungauged station and gauged station is a function of drainage area of the two stations. At daily timescale, this assumption will be valid only if the drainage areas of the gauged and ungauged monitoring stations have *significant* overlap; that is, the streamwise distance between the two stations is small. This assumption is reasonable at annual timescale (Rodriguez-Iturbe and Rinaldo, 2001) but, at daily timescale, it is valid only in limited situations. At daily timescale, factors such as spatial variability of rainfall (Gabellani et al., 2007) and difference in land-use and topography will result in differences in the shape of streamflow time series. At daily timescale, a hydrologic model must be employed to estimate streamflows at ungauged stations. Gupta and Govindaraju (2019) showed that a simple hydrologic model calibrated against observations at a gauged location may entail significant uncertainties in estimated streamflows at ungauged locations. Nevertheless, the DST does not account for these uncertainties; this topic would be the scope of a different study. The DST operates at daily timescale; thus, streamflow and WQ data should be available at daily timescale. Typically, WQ data are collected as grab samples representing instantaneous values. An implicit assumption in the DST is that instantaneous values represent the average daily values.

3. Case study

3.1 Study area and data

The DST was used to conduct WQ risk and TMDL analyses in the St. Joseph River Watershed (SJRW), USA (Fig. 3). The watershed is spread over parts of Indiana, Ohio, and Michigan. A large part of SJRW is covered with agricultural fields (Fig. 3b). Therefore, agricultural runoff is expected to be the main contributor of total phosphorus (TP) in the SJRW river-network (IDEM, 2017, pp. 34). A few animal operations and point-sources, in urban areas and large villages, can potentially contribute TP in some parts of SJRW river-network. Total dissolved solids (TDS) are solid particles suspended in water column which may consist of nutrients, biological particles such as algae, and soil particles. Soil particles end up in water column because of channel erosion under high flow conditions, soil erosion from agricultural areas (livestock grazing, plowing), and urban areas (construction sites) and, to a limited extent, from forests (IDEM, 2017).

The daily-streamflow data at five stations in the watershed were made available by United States Geological Survey (USGS, 2016). The TP concentration data at 11 locations and TDS at 14 locations in the watershed were made available through St. Joseph River Watershed Initiative (SJRWI at http://wqis.ipfw.edu/, accessed: 26 Aug, 2018, Figs. 4 and 5). St. Joseph River (SJR) and its tributaries are designated for aquatic life use (ALU), recreational use (RU), and warm and cold habitats (IDEM, 2017). TP and TDS primarily affect ALU; IDEM (2017) reported that many portions of the SJR and its tributaries were impaired for ALU. They concluded that TP and TDS load reductions of up to 66% and 95%, respectively, would be required to meet the TMDL criterion. Figs. 4 and 5 show that all the stations violated TP concentration standards except stations 123, 128, and 150, and all the stations violated TDS standards.

In this study, the WQ risk and TMDL analyses were restricted to time-period 2000-2017, thus TP and TDS data were reconstructed using observed streamflow from 01/01/2000 to 12/31/2017. It should be noted that (a) selection of this time-period is user's choice as long as there is a time-period where both streamflow and WQ have concurrent measurements. For each TP and TDS measurement station, streamflow data available at the nearest streamflow measurement station were used as an input to the DST. For example, for station 122, streamflow data available at the station USGS 04180000 were used. The required area-ratios (section 2.4) were computed using

SWAT, though this ratio could be computed using any geographical information system or available drainage area information.

Concentrations values of 0.08 mg L^{-1} and 750 mg L^{-1} were used as WQ standards for calculation of WQ risk measures with respect to TP and TDS, respectively (SJRWI by http://wqis.pfw.edu/, accessed: 26 Aug, 2018); and concentration values of 0.30 mg L^{-1} and 30 mg L^{-1} were used for TMDL development for TP and TDS (IDEM, 2017), respectively. Out of the 9 LOADEST equations, the one with the best fit based on AIC was used to reconstruct WQ data. Subsequently, results corresponding to station 122 are discussed in detail and results corresponding to other stations are summarized for brevity. Henceforth, the analysis conducted using the DST described in this paper is referred to as rvm-LOADEST, and the analysis without uncertainty quantification is referred to as deterministic-LOADEST. The deterministic analysis was conducted using the web-based tool (by Park et al., 2015). The same LOADEST equation was used for reconstruction in both rvm- and deterministic-LOADEST. LOADEST equations used for TP and TDS reconstruction at different stations are listed in Appendix A. Note that the DST may also be used in a deterministic mode by using the expected value of the reconstructed WQ concentration/load time series and ignoring the information on uncertainty.

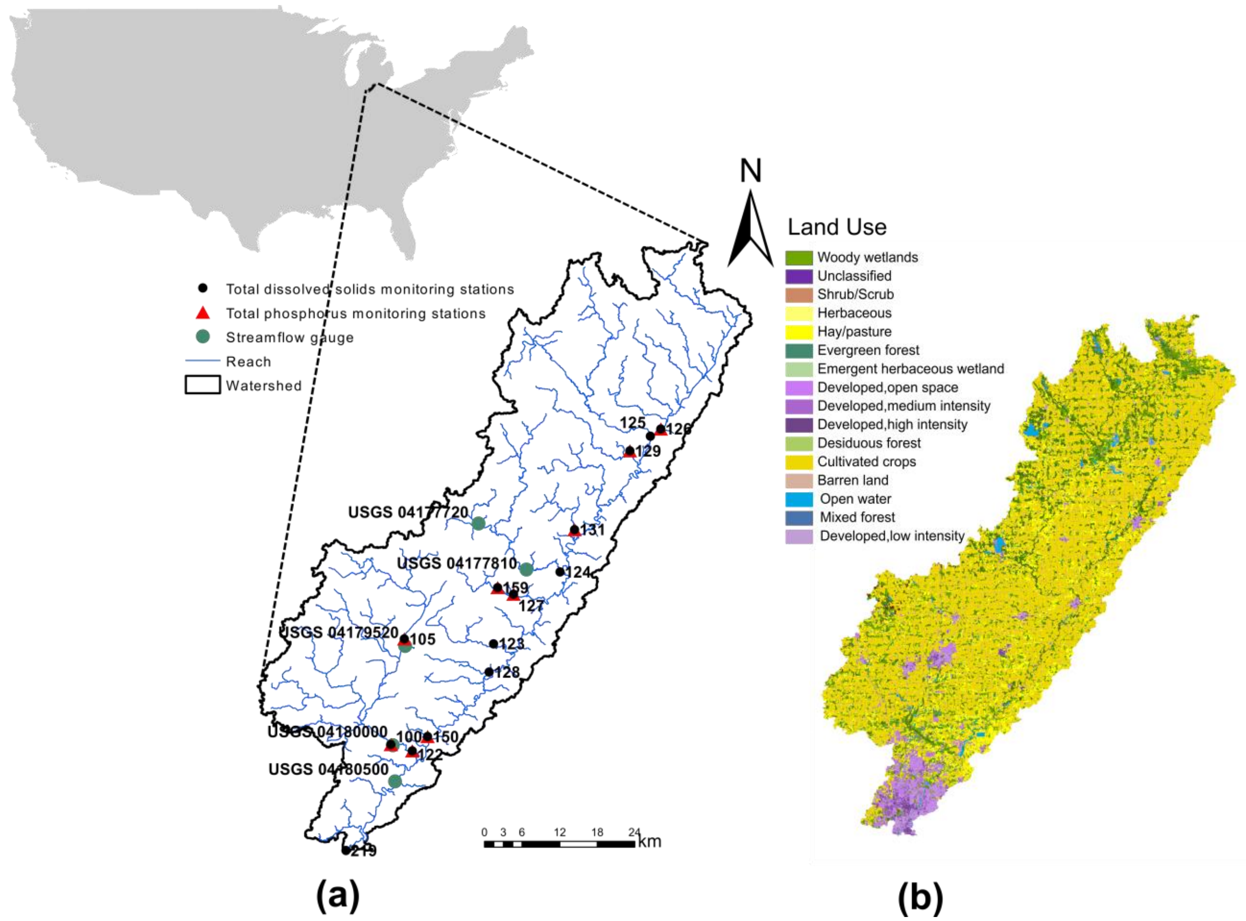

Fig. 3. (a) St. Joseph River watershed (SJRW) with a delineated stream network and locations of total phosphorus (TP), total dissolved solids (TDS), and streamflow data stations. For TP and TDS, station numbers as listed on SJRWI website are shown by red triangles and black dots, respectively, and for streamflow, USGS station numbers are shown by green colored dots. (b) Land use pattern of SJRW.

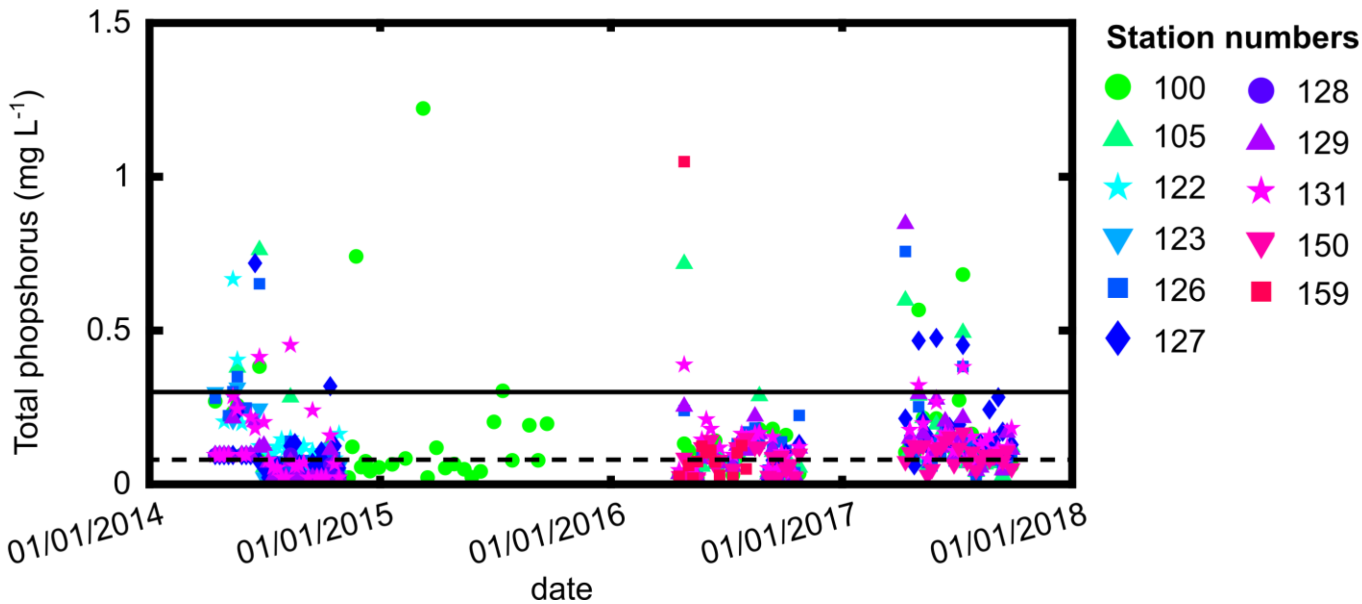

Fig. 4. Observed total phosphorus data at 11 monitoring stations. The solid black line represents TMDL concentration and dashed black line represents standard concentration.

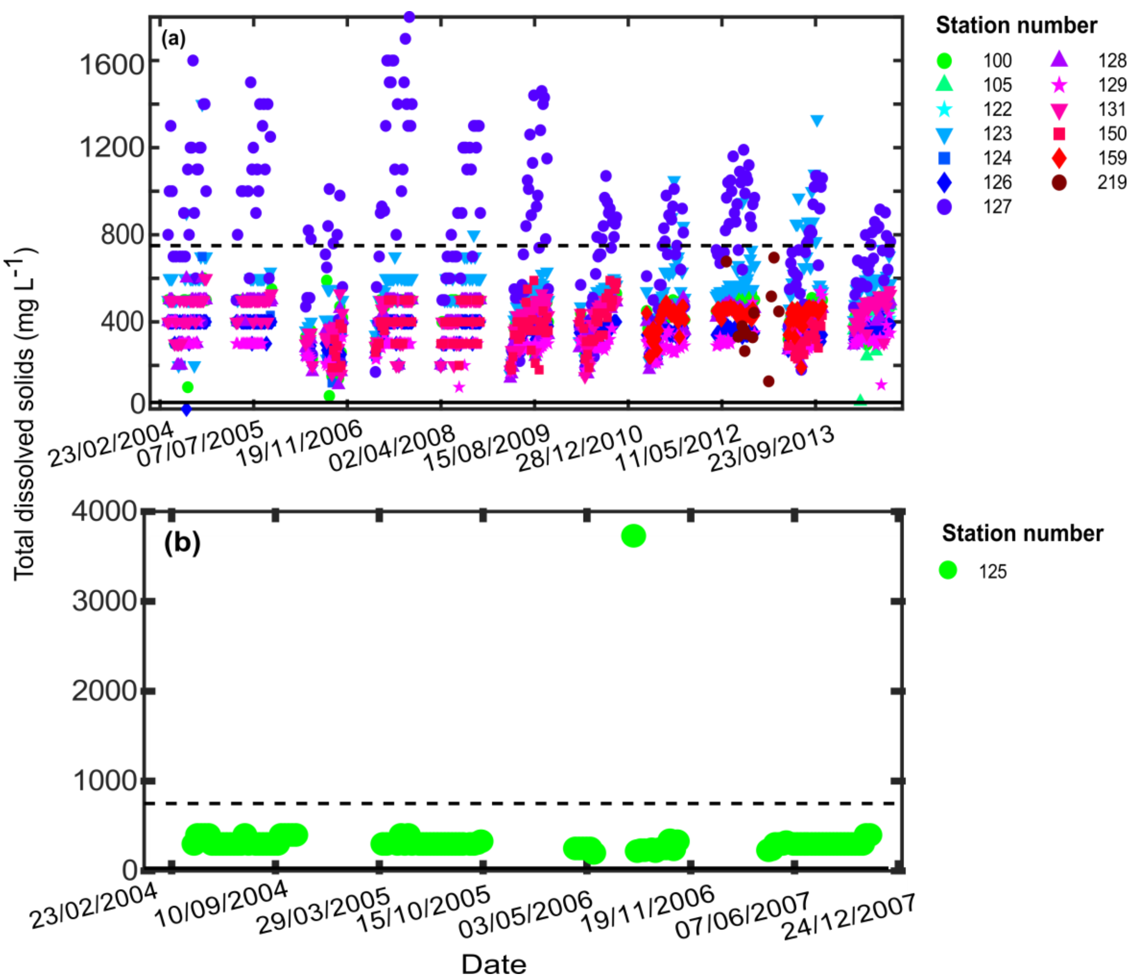

Fig. 5. Observed total dissolved solids (TDS) data at (a) all the monitoring stations except 125 (b) monitoring station 125. The observations at station 125 are shown separately because of one instance of exceptionally high value of TDS at this station. The solid black line represents TMDL concentration and dashed black line represents standard concentration.

3.2 Results

The run time of DST depends upon the time-period of analysis and number of observations available for reconstruction. The total runtime of the DST for one WQ monitoring station and one WQ constituent was approximately 2 minutes for this case study.

Total Phosphorus (TP)

Except three, all observations were enveloped by the 90% and 95% credible regions which is expected since all these observations were used for estimation of the weight vector and variance of residuals (Fig. 6). The uncertainty band was wide, especially in high concentration regions. Since the land-use in SJRW is dominated by agriculture, high concentrations of TP are expected to be associated with high agricultural runoff and high streamflows. According to Eq. (4), high streamflows will result in high prediction variance of TP concentrations. The reconstructed TP time series obtained by the computing mean of rvm-LOADEST time-series and obtained by deterministic-LOADEST were approximately the same (Fig. 6). The estimated weights by RVM were consistently smaller than those obtained by LOADEST, providing better hedging against errors in streamflow observations (Appendix A1).

The ranges of R-R-V and WH measures at station 122 were 0.245 to 0.278 (reliability), 0.235 to 0.278 (resilience), 0.572 to 0.593 (vulnerability) and 0.289 to 0.320 (watershed health), respectively (Table 2). The R-R-V and WH measures obtained by deterministic-LOADEST were 0.007, 0.002, 0.468 and 0.020, respectively. The mean values of the measures obtained by rvm-LOADEST are substantially different from those obtained by deterministic-LOADEST. In fact, the values of the risk measures obtained by deterministic-LOADEST are not even contained in the range of those obtained by rvm-LOADEST; this is due to consideration of uncertainty in rvm-LOADEST and the sensitivity of the nonlinear risk measures to slight differences in the WQ constituent time series. At station 122, the reliability and resilience are low, and the vulnerability is high. It implies that this station incurs frequent violations of WQ standard and takes a long time to recover, and severity of violations is also high. Note that if pollution control measures are used to increase the resilience of station 122, the reliability would also increase. However, if pollution control measures are taken to increase the reliability only, the resilience may not increase.

Figure 7 shows the LRR computed by deterministic- and by rvm-LOADEST at different compliance values and different permissible violations. At most of the stations, the LRR as computed by using deterministic-LOADEST did not achieve even 50% compliance after the uncertainty in reconstructed WQ time series was considered. For example, at station 122, the LRR as obtained by deterministic- and rvm-LOADEST at $p = 0.10$ and 50% compliance were 0 and 35.9 Kg day^{-1}, respectively; the LRR as obtained by deterministic- and rvm-LOADEST at $p = 0.05$ and 50% compliance were 0 and 211.6 Kg day^{-1}, respectively; the LRR as obtained by deterministic- and rvm-LOADEST at $p = 0.03$ and 50% compliance were 75.9 Kg day^{-1} and 458.5 Kg day^{-1}, respectively. The LRR values as obtained by rvm-LOADESTs are function of κ; as the κ increases, the LRR increases.

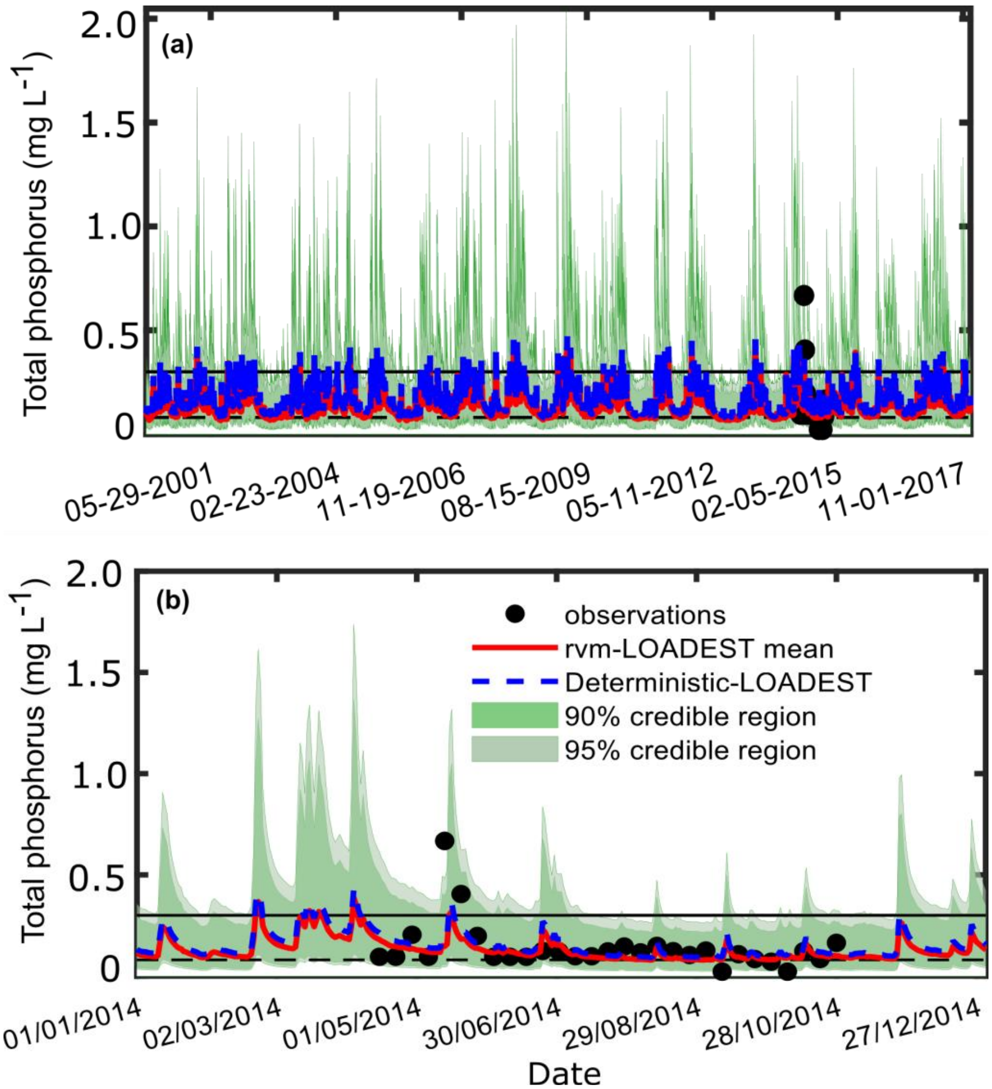

Fig. 6. Station 122, total phosphorus (TP). Observed and reconstructed daily TP during (a) 2000-2017 and (b) 2014 (the year in which water quality data was available)

Table 2. Total phosphorus (TP). The risk-measures obtained by deterministic (D)- and rvm-LOADEST. The measures are computed at daily timescale.

Station	Model Type	RVM statistic	Reliability	Resilience	Vulnerability	Watershed health
126	**D**	-	0.183	0.023	0.424	0.134
	rvm	(Mean, Median)	(0.363, 0.363)	(0.342, 0.342)	(0.557, 0.557)	(0.381 ,0.381)
		Range	(0.341-0.382)	(0.316-0.368)	(0.545-0.568)	(0.364-0.396)
150	**D**	-	0.962	0.008	0.202	0.182
	rvm	(Mean, Median)	(0.787, 0.787)	(0.660, 0.660)	(0.448, 0.448)	(0.659, 0.659)
		Range	(0.769-0.803)	(0.614-0.705)	(0.417-0.472)	(0.637-0.682)
159	**D**	-	0.001	0.001	0.300	0.009
	rvm	(Mean, Median)	(0.433, 0.433)	(0.433, 0.433)	(0.587, 0.587)	(0.426, 0.426)
		Range	(0.410-0.455)	(0.404-0.460)	(0.575-0.598)	(0.407-0.442)
127	**D**		0.720	0.034	0.412	0.244

	rvm	(Mean, Median)	(0.642, 0.642)	(0.453, 0.453)	(0.482, 0.482)	(0.532,0.532)
		Range	(0.622-0.659)	(0.418-0.490)	(0.465-0.498)	(0.515-0.551)
131	**D**	**-**	0.428	0.030	0.500	0.180
	rvm	(Mean, Median)	(0.596, 0.596)	(0.465, 0.465)	(0.470, 0.470)	(0.528, 0.528)
		Range	(0.578-0.617)	(0.431-0.498)	(0.454-0.484)	(0.510-0.546)
129	**D**	**-**	0.813	0.05 0	0.373	0.29
	rvm	(Mean, Median)	(0.776, 0.776)	(0.478, 0.478)	(0.448, 0.448)	(0.589, 0.589)
		Range	(0.760-0.791)	(0.428-0.529)	(0.428-0.473)	(0.564-0.615)
105	**D**	**-**	0. 807	0.132	0.476	0.382
	rvm	(Mean, Median)	(0.658, 0.658)	(0.493, 0.493)	(0.521, 0.521)	(0.537, 0.538)
		Range	(0.633-0.680)	(0.446-0.537)	(0.502-0.541)	(0.516-0.560)
122	**D**	**-**	0.007	0.002	0.47	0.020
	rvm	(Mean, Median)	(0.265, 0.266)	(0.256, 0.256)	(0.582, 0.582)	(0.305, 0.305)
		Range	(0.245-0.285)	(0.235-0.278)	(0.572-0.593)	(0.289-0.320)
100	**D**	**-**	0.464	0.062	0.451	0.250
	rvm	(Mean, Median)	(0.477, 0.477)	(0.398, 0.399)	(0.542, 0.543)	(0.443, 0.443)
		Range	(0.457-0.496)	(0.370-0.431)	(0.530-0.556)	(0.426-0.458)

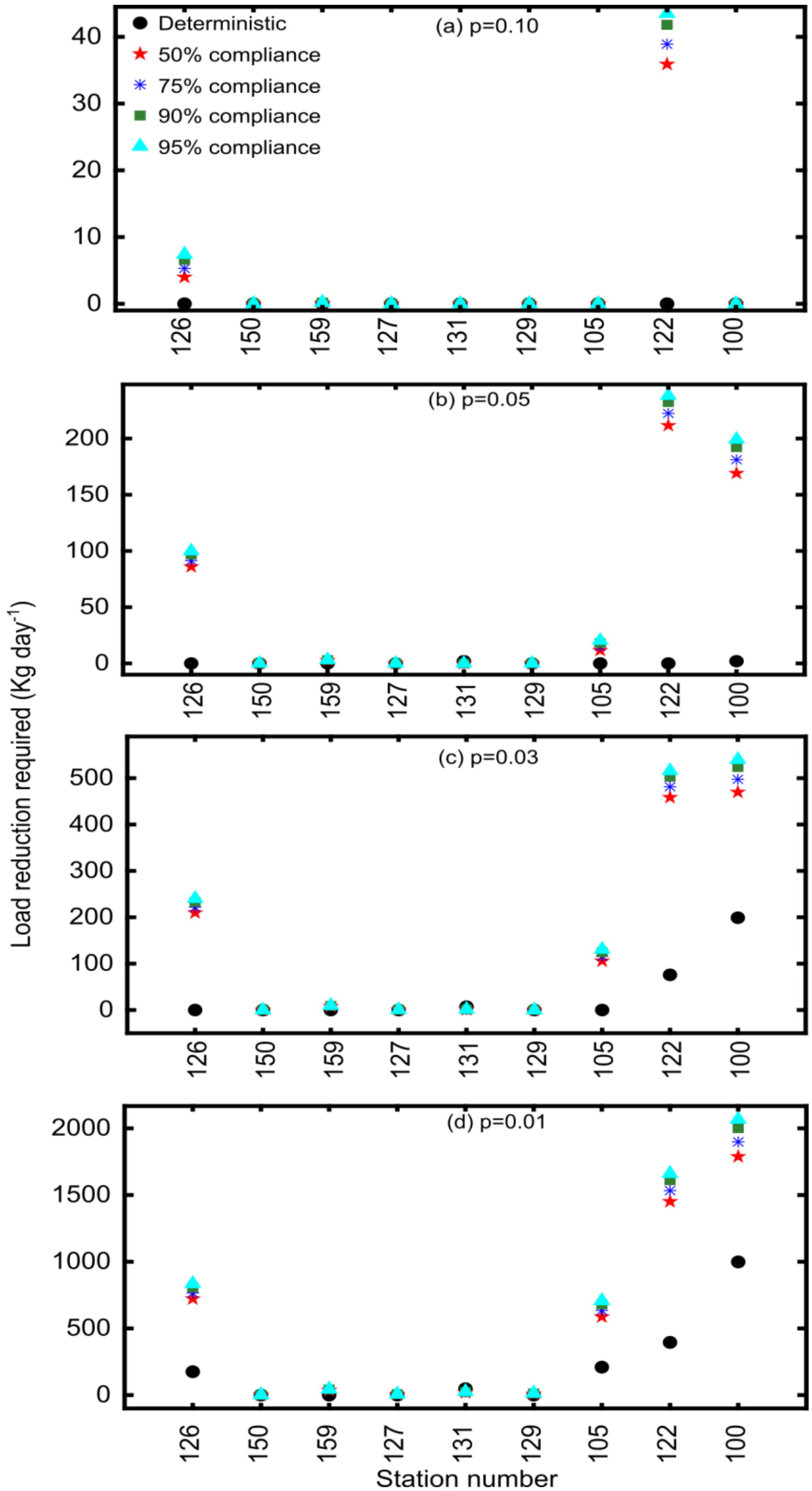

Fig. 7. Total phosphorus (TP). Daily timescale load reduction required at different station at permissible violations (p)

Total Dissolved Solids (TDS)

The general results obtained for TDS were same as those for TP. The reconstructed TDS time series obtained by mean of the rvm-LOADEST and deterministic-LOADEST were similar at most time-steps (Fig. 8b). All except one observation were enveloped by the 90% and 95% credible regions. Table 3 lists the statistics of R-R-V values obtained by using deterministic- and rvm-LOADEST models. As in case of TP, at many stations, the R-R-V values obtained by deterministic-LOADEST were outside the range of those obtained by rvm-LOADEST. The differences between the risk measured obtained by deterministic-LOADEST and the means of the risk measures obtained by rvm-LOADEST, however, was small. TDS violations occurred only at a few stations in the watershed. When violations did occur, the resilience of the station was low. For example, at station 127, the reliability was 0.604 (mean of rvm-LOADEST values) and the resilience was 0.353 implying if pollution control measures were to be put in place to increase the resilience of the waterbody, these will also increase the reliability of the waterbody. At most stations, the vulnerability was low implying that magnitude of violations was small. Overall, the watershed is in a healthy condition in terms of TDS.

At stations 105 and 122, the deterministic-LOADEST LRR values were larger than rvm-LOADEST LRR values computed at 50% and higher levels of compliance. At most of the stations, however, the LRR values computed by deterministic-LOADEST at various compliance percentages were smaller than those computed by deterministic-LOADEST (Table 5). In summary, if one carries out a deterministic analysis then the computed LRR may not be enough to comply with TMDL concentration, or may be too conservative as is the case for stations 105 and 122.

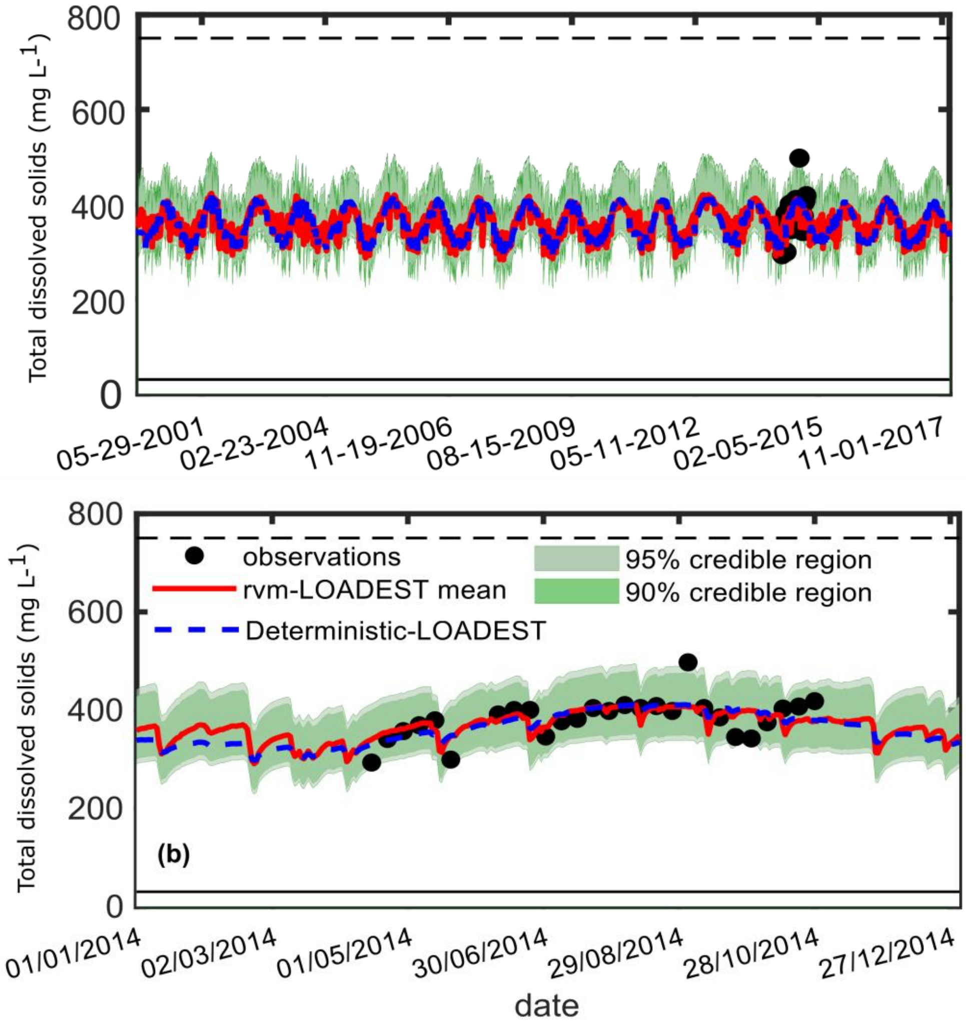

Fig. 8. Station 122, total dissolved solids (TDS). Observed and reconstructed daily TDS during (a) 2000-2017 and (b) 2014 (the year in which water quality data were available)

Table 3. Total dissolved solids (TDS). The risk-measures obtained by deterministic (D)- and rvm-LOADEST. The risk measures are computed at daily timescale.

Station	**Model type**	**RVM statistic**	**Reliability**	**Resilience**	**Vulnerability**	**Watershed health**
128	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(1.000, 1.000)	(0.999, 1.000)	(0.004, 0.000)	(0.999, 1.000)
		Range	(1.000)	(0.500-1.000)	(0.000-0.135)	(0786-1.000)
126	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(1.000, 1.000)	(1.000, 1.000)	(0.000, 0.000)	((1.000, 1.000))
		Range	(1.000)	(1.000)	(0.000)	(1.000)
150	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(0.999, 1.000)	(0.999, 1.000)	(0.015, 0.000)	(0.995, 1.000)
		Range	(0.999-1.000)	(0.500-1.000)	(0.000-0.221)	(0.769-1.000)
159	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(1.000, 1.000)	(1.000, 1.000)	(0.000, 0.000)	(1.000, 1.000)
		Range	(1.000)	(1.000)	(0.000)	(1.000)
127	**D**	-	0.581	0.053	0.210	0.29

	rvm	(Mean, Median)	(0.604, 0.605)	(0.354, 0.354)	(0.298, 0.298)	(0.532, 0.532)
		Range	(0.588-0.623)	(0.324-0.390)	(0.286-0.311)	(0.514-0.552)
131	**D**	-	0.996	0.115	0.020	0.483
	rvm	(Mean, Median)	(0.986, 0.986)	(0.860, 0.861)	(0.066, 0.066)	(0.925, 0.926)
		Range	(0.981-0.991)	(0.696-1.000)	(0.045-0.089)	(0.859-0.976)
123	**D**	-	0.957	0.094	0.055	0.44
	rvm	(Mean, Median)	(0.898, 0.899)	(0.713, 0.713)	(0.107, 0.107)	(0.830, 0.830)
		Range	(0.884-0.911)	(0.650-0.779)	(0.097-0.122)	(0.802-0.857)
129	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(0.999, 0.999)	(0.965, 1.000)	(0.0497, 0.048)	(0.971-0.982)
		Range	(0.998-1.000)	(0.333-1.000)	(0.000-0.229)	(0.680-1.000)
125	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(1.000, 1.000)	(1.000, 1.000)	(0.000, 0.000)	(0.1000, 1.000)
		Range	(1.000)	(1.000)	(0.000-0.058)	(0.980-1.000)
124	**D**		1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	(0.999, 1.000)	(0.998, 1.000)	(0.034, 0.029)	(0.987, 0.990)
		Range	(0.996-1.000)	(0.50-1.00)	(0.000-0.240)	(0.764-1.000)
105	**D**	-	0.999	0.125	0.054	0.491
	rvm	(Mean, Median)	(0.915, 0.915)	(0.769, 0.769)	(0.165, 0.165)	(0.837, 0.839)
		Range	(0.903-0.928)	(0.693-0.835)	(0.146-0.184)	(0.807-0.864)
122	**D**	-	1.000	1.000	0.000	1.000
	rvm	(Mean, Median)	1.000	1.000	0.000	1.000
		Range	(1.000)	(1.000)	(0.000)	(1.000)
100	**D**	-	1.000	1.000	0.000	1.00
	rvm	(Mean, Median)	(0.996, 0.998)	(0.981, 1.000)	(0.045, 0.044)	(0.978, 0.983)
		Range	(0.995-1.000)	(0.750-1.000)	(0.014-0.122)	(0.898-0.995)

Output not available from web-based LOADEST for station 219

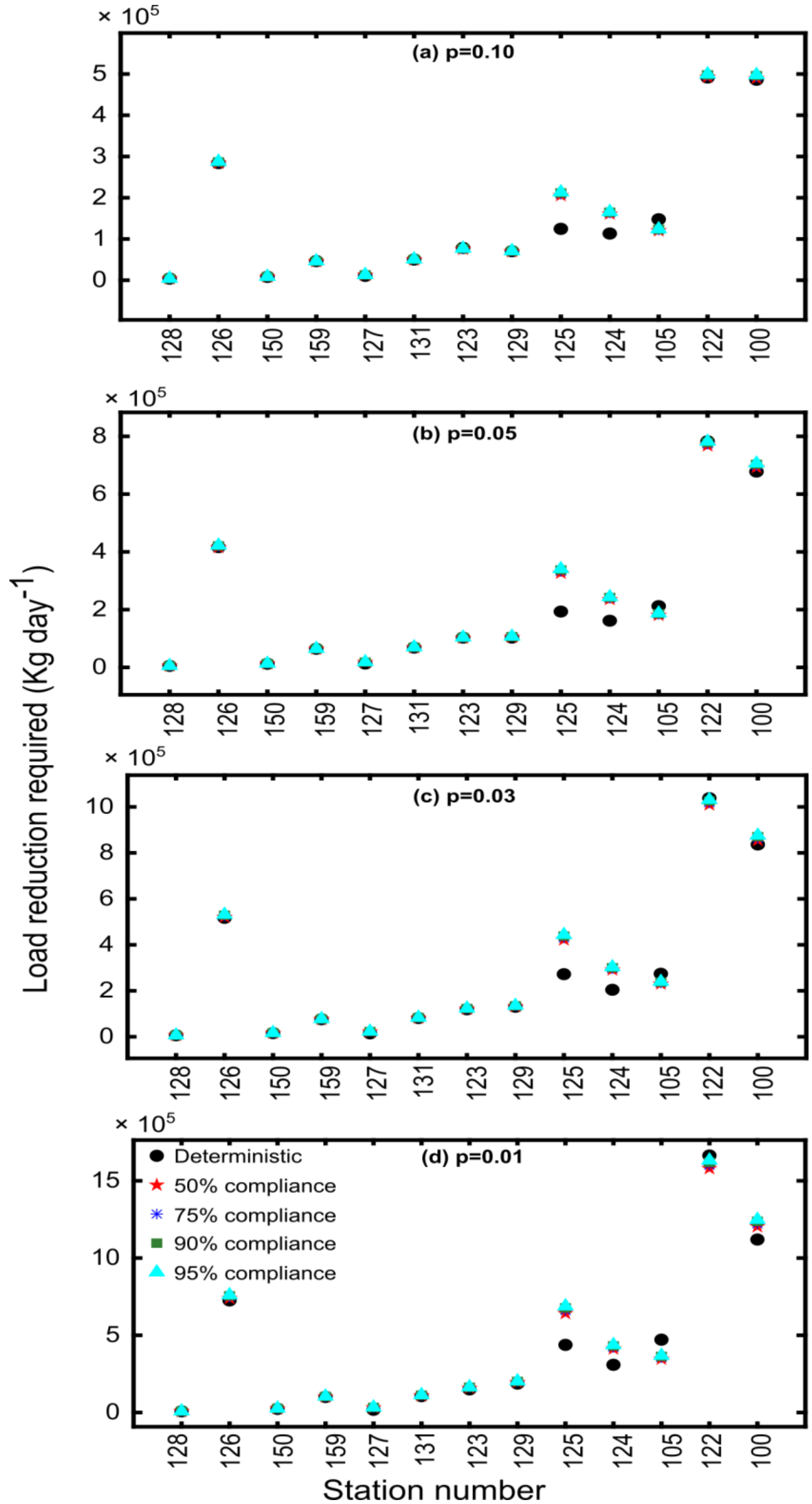

Fig. 9. Total dissolved solids (TDS). Daily time scale load reduction required at different station at permissible violations (p)

3.3 Discussion

In case of TP, the R-R-V measures obtained by deterministic-LOADEST were not even in the range of those obtained by rvm-LOADEST. This discrepancy is due to the assumption of statistical independence of errors in the reconstructed TP series at different time-steps and high sensitivity of these measures when the WQ concentration series is near the standard series. The discrepancy exists even though the reconstructed TP time series obtained by deterministic-LOADEST was approximately equal to the mean of the TP time series obtained by rvm-LOADEST. Out of 10,000 MC realizations of TP time series, there were many realizations that were close to the mean realization. But no realization was identical to the mean time series because of the added error-term in rvm-LOADEST. The difference between mean and actual realization of TP time series translated into large differences in estimated R-R-V measures because these measures are extremely sensitive to small changes in TP time series, especially when TP time series is close to the standard time series. Thus, if one were to simply take the mean values estimated by LOADEST and not account for uncertainty in these estimates, the assessment of WQ risk may be erroneous.

Since the discrepancy in R-R-V measures obtained by rvm-LOADEST and deterministic-LOADEST is due to statistical assumptions over the residual time series, this discrepancy also illustrates the importance of the assumptions in uncertainty analysis. This discrepancy can be resolved only if the statistical assumptions are such that they allow to draw the mean WQ constituent time series from the distribution. Moreover, evidence supports the hypothesis that a watershed that yields a WQ constituent time series with high positive autocorrelation is less resilient to perturbations (such as high pollutant load in the watershed) (Qi et al., 2016). Thus, introducing artificial correlations to reconstructed WQ constituent time series by means of statistical assumptions over the error-term may lead to misleading conclusions. There is no way to confirm if the statistical assumptions made are valid; QQ plots can tell us if the assumption are invalid, not if the assumptions are correct. One of the problems in probabilistic uncertainty quantification is the non-uniqueness of possible statistical assumptions that can fit the data.

To check the effectiveness of MOS value used in TMDL reports (e.g., IDEM, 2017), the MOS was computed using the same method as in IDEM (2017) (see SI). Specifically, five different TMDLs were computed for five different flow regions in each of the streams. The five flow regions were determined using daily exceedance probability (DEP) as follows : high flows ($0 - 0.10$ DEP), moist conditions ($0.10 - 0.40$ DEP), mid-range ($0.40 - 0.60$ DEP), dry conditions (0.60-0.90 DEP), and low flows ($0.90 - 1.00$ DEP). The results are discussed for high flow conditions only. As per rvm-LOADEST, up to 1000% of the TMDL should be allocated as MOS to achieve 50% compliance (Figs. S3 and S4). An MOS value greater than 100% implies that the maximum allowable load in the watershed cannot be determined reliably. IDEM (2017) report did not explicitly quantify uncertainties. The cases study illustrates how a deterministically estimated WQ times-series may lead to different conclusions than one that quantifies the effects of uncertainty. Authors of IDEM (2017) report used SWAT to simulate WQ constituent time series. SWAT is known to incur high uncertainties (Hollaway et al., 2018) but this information was not utilized in MOS specification.

To reconstruct WQ constituent time series, the DST assumes that logarithm of load and logarithm of streamflow are correlated. If the correlation is weak, the prediction accuracy of the model will be poor and the uncertainty band will be wide. To check the prediction accuracy of rvm-LOADEST, the observed and predicted mean values of TP and TDS were plotted; the results suggested that one can indeed reconstruct TP and TDS loads by using streamflow values as predictor variables (Figs. S5 andS6). In case of both TP and TDS, 90 to 100% of the observations were enveloped by the 90% credible region with a few exceptions (Tables S3 and S4). Moreover, the methodology adopted in the DST is valid only if the relationship between streamflows and pollutant loads remains unchanged during the period of analysis. If this relationship changes, model predictions will be poor. The relationship could change because of pollution control measures put in the drainage area during the period of analysis, but not during the observation period and change of rainfall-runoff-pollutant load relationship (due to climatic and/or land use changes), Thus, the DST cannot be used for future scenario analyses indiscriminately.

Further, the DST assumes that the residuals at different time-steps are statistically independent. If this assumption is invalid, the consequence would be over-estimation of information content in residuals which would result in under-estimation of uncertainty in $\boldsymbol{w}$ and, in-turn, an under-estimation of uncertainty in predicted loads. For convenience of analysis, the DST assumes that residuals are distributed according to Gaussian law with homoscedastic variance. To check the validity of these statistical assumptions, QQ plots were used (Figs. S7-S10). These plots revealed that the observed residuals did not satisfy the assumptions made by DST, implying that uncertainties in WQ reconstruction as reported in this study may be underestimated. One way of relaxing the assumption of independence is to model the residuals as an autoregressive (AR) process (e.g., Hantush and Chaudhary, 2014). In this process, the residual at a time-step is regressed against $(k-1)$ residuals at previous consecutive time-steps. Future versions of DST will be updated to accommodate this analysis.

Data limitations are ubiquitous in modeling exercises. In principle, there is no established restriction on the minimum number of observations to apply the DST. However, since the DST reconstructs WQ time series using a statistical regression method, very few observations may translate into over- or under-estimation of uncertainty. More observations imply better uncertainty estimates. For reference, Schwarz et al. (2006) suggested 15 observations to estimate *annual average* loads, and we suggest this number as a lower limit.

4. Summary and conclusions

A DST was developed to reconstruct WQ constituent time series and conduct risk-based WQ assessment and TMDL development. The DST uses RVM to incorporate uncertainty in reconstruction of WQ constituent time series. The tool estimates uncertainty due to residual errors in reconstructed WQ time series, allows users to propagate this uncertainty to R-R-V and WH risk assessment and TMDL estimation. These two applications of the DST were demonstrated for the SJRW. The following conclusion were drawn:

(1) The weights estimated by RVM are consistently smaller than those compared to web based LOADEST; the smaller weights are desirable because they hedge against errors in streamflows.

(2) Based on our experience, we expect WQ risk measures to be very sensitive to small changes in WQ constituent time series especially when realizations of loads/concentrations are close to the standard values; therefore, errors in the reconstruction of WQ constituent time series must be modeled to obtain a realistic estimate of WQ risk measures of a waterbody. This sensitivity, however, also illustrates the importance of realistic statistical assumptions over the error-term.

(3) At most stations, consideration of uncertainty in WQ risk measures led to very different conclusions about watershed health. Uncertainty analysis indicated a relatively poorer health at some WQ monitoring stations and a relatively better health at other WQ monitoring stations.

(4) The LRR values at daily timescale as yielded by a deterministic analysis may not be enough to achieve even 50% compliance. Uncertainty in LRR should be considered for effective

pollution control. The arbitrarily selected MOS values used in TMDL report may result in gross under-estimation of uncertainty. Therefore, MOS values should be based upon a systematic uncertainty analysis (as was also suggested by Reckhow, 2003).

As presented, the tool is restricted to quantifying uncertainty due to residual errors. It is possible to use the RVM methodology to explicitly incorporate measurement errors in streamflows, which will be the topic of future research.

Acknowledgments

This work was supported by the U.S. Environmental Protection Agency through its Office of Research and Development (Contract Number: EP-C-15-010). This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

References

Ahmadisharaf, E., Camacho, R. A., Zhang, H. X., Hantush, M. M., & Mohamoud, Y. M. (2019). Calibration and Validation of Watershed Models and Advances in Uncertainty Analysis in TMDL Studies. Journal of Hydrologic Engineering, 24(7), 03119001.

Ahmadisharaf, E., & Benham, B. L. (2020). Risk-based decision making to evaluate pollutant reduction scenarios. Science of The Total Environment, 702, 135022.

Akaikei, H. (1973). Information theory and an extension of maximum likelihood principle. In Proc. 2^{nd} Int. Symp. On Information Theory (pp. 267-281).

Arnold, J.G., & Allen, P.M. (1999). Automated methods for estimating baseflow and groundwater recharge from streamflow records. Journal of the Americam Water Resources Association vol 35(2): 411-424. https://doi.org/10.1111/j.1752-1688.1999.tb03599.x

Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: a quantitative analysis. Hydrology & Earth System Sciences, 13(6).

Beck, M. B. (1987). Water quality modeling: a review of the analysis of uncertainty. Water Resources Research, 23(8), 1393-1442.

Beven, K. (2007). Environmental modelling: An uncertain future?. CRC press.

Beven, K., & Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. Hydrological processes, 6(3), 279-298.

Bishop, C. M. (2006). Pattern recognition and machine learning (No. 681.326. 7 BIS)

Borsuk, M. E., and Stow, C. A. (2000). "Bayesian parameter estimation in a mixed-order model of BOD decay." Water Res., 34(6), 1830-1836.

Borsuk, M. E., Stow, C. A., & Reckhow, K. H. (2002). Predicting the frequency of water quality standard violations: A probabilistic approach for TMDL development.

Borsuk, M. E. (2003). A graphical probability network model to support water quality decision making for the Neuse River estuary, North Carolina.

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. Inverse Problems, 30(11), 114007. https://doi.org/10.1088/0266-5611/30/11/114007

Chaudhary, A., & Hantush, M. M. (2017). Bayesian Monte Carlo and maximum likelihood approach for uncertainty estimation and risk management: Application to lake oxygen recovery model. Water research, 108, 301-311.

Emerson, D. G., Vecchia, A. V., & Dahl, A. L. (2005). Evaluation of drainage-area ratio method used to estimate streamflow for the Red River of the North Basin, North Dakota and Minnesota (p. 1). US Department of the Interior, US Geological Survey.

Fang, K. T., & Zhang, Y. (1990). Generalized Multivariate Analysis. Springer- Verlag and Science Press, Berlin/Beijing.Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Gabellani, S., Boni, G., Ferraris, L., Von Hardenberg, J., & Provenzale, A. (2007). Propagation of uncertainty from rainfall to runoff: A case study with a stochastic rainfall generator. Advances in Water Resources, 30(10), 2061-2071.

Götzinger, J. & Bárdossy, A. (2008). Generic error model for calibration and uncertainty estimation of hydrological models. Water Resources Research, 44(12). https://doi.org/10.1029/2007WR006691

Gronewold, A. D., & Borsuk, M. E. (2009). A software tool for translating deterministic model results into probabilistic assessments of water quality standard compliance. Environmental Modelling & Software, 24(10), 1257-1262.

Gupta, A., & Govindaraju, R. S. (2019). Propagation of structural uncertainty in watershed hydrologic models. Journal of Hydrology, 575, 66-81.

Hantush, M. M., & Chaudhary, A. (2014). Bayesian framework for water quality model uncertainty estimation and risk management. Journal of Hydrologic Engineering, 19(9), 04014015.

Hashimoto, Tsuyoshi, Jery R. Stedinger, and Daniel P. Loucks. (1982). "Reliability, Resiliency, and Vulnerability Criteria for Water Resource System Performance Evaluation." Water Resources Research 18 (1): 14–20. Doi:10.1029/WR018i001p00014.

Hollaway, M. J., Beven, K. J., Benskin, C. M. H., Collins, A. L., Evans, R., Falloon, P. D., ... & Ockenden, M. C. (2018). The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability'uncertainty framework to a water quality model. Journal of hydrology, 558, 607-624.

Hoque, Yamen M., Shivam Tripathi, Mohamed M. Hantush, and Rao S. Govindaraju. (2012). "Watershed Reliability, Resilience and Vulnerability Analysis under Uncertainty Using Water Quality Data." Journal of Environmental Management 109 (0): 101–12. Doi:10.1016/j.jenvman.2012.05.010.

Indiana Department of Environmental Management. (2017). St. Joseph River Watershed TMDLs. Indianapolis (IN): IDEM, Office of Water Quality. 497 p.

Jia, Y., & Culver, T. B. (2006). Robust optimization for total maximum daily load allocations. Water Resources Research, 42(2).

Jia, Y., & Culver, T. B. (2008). Uncertainty analysis for watershed modeling using generalized likelihood uncertainty estimation with multiple calibration measures. Journal of Water Resources Planning and Management, 134(2), 97-106.

Kjeldsen, T. R., & Rosbjerg, D. (2004). Choice of reliability, resilience and vulnerability estimators for risk assessments of water resources systems. Hydrological sciences journal, 49(5).

Mallya, G., Hantush, M., & Govindaraju, R. S. (2018). Composite measures of watershed health from a water quality perspective. Journal of environmental management, 214, 104-124.

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

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. Water Resources Research, 48(9).

National Research Council. (2001). Assessing the TMDL approach to water quality management. National Academies Press.

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. Hydrological Sciences Journal, 61(9), 1666-1678. https://doi.org/10.1080/02626667.2016.1183009

Novotny, V. (2002). Water quality: diffuse pollution and watershed management. John Wiley & Sons.

Nunoo, R., Anderson, P., Kumar, S., & Zhu, J. J. (2020). Margin of Safety in TMDLs: Natural Language Processing-Aided Review of the State of Practice. Journal of Hydrologic Engineering, 25(4), 04020002.

Pappenberger, F., & Beven, K. J. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water resources research, 42(5).

Park, Y., Engel, B., Frankenberger, J., & Hwang, H. (2015). A web-based tool to estimate pollutant loading using LOADEST. Water, 7(9), 4858-4868.

Qi, M., Feng, M., Sun, T., & Yang, W. (2016). Resilience changes in watershed systems: A new perspective to quantify long-term hydrological shifts under perturbations. Journal of hydrology, 539, 281-289.

Reckhow, K. H. (2003). On the need for uncertainty assessment in TMDL modeling and implementation. J. Water Resour. Plan. Manage, 129(4), 245-246.

Refsgaard, J. C., Van der Sluijs, J. P., Brown, J., & Van der Keur, P. (2006). A framework for dealing with uncertainty due to model structure error. Advances in Water Resources, 29(11), 1586-1597.

Ries, K. G. (2007). The national streamflow statistics program: A computer program for estimating streamflow statistics for ungauged sites. DIANE Publishing.

Rodriguez-Iturbe, I., & Rinaldo, A. (2001). A view of river basins. In *Fractal river basins: chance and self-organization* (pp 01-95). Cambridge University Press

Runkel, Robert L, Charles Gene Crawford, and Timothy A Cohn. (2004). Load Estimator (LOADEST): A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers. US Department of the Interior, US Geological Survey.

Schwarz, G. E., Hoos, A. B., Alexander, R. B., & Smith, R. A. (2006). The SPARROW surface water-quality model: theory, application and user documentation. US geological survey techniques and methods report, book, 6(10), 248.

Shirmohammadi, A., Chaubey, I., Harmel, R. D., Bosch, D. D., Muñoz-Carpena, R., Dharmasri, C., ... & Graff, C. (2006). Uncertainty in TMDL models. Transactions of the ASABE, 49(4), 1033-1049.

Smith, R. C. (2014). Large-scale applications. In *Uncertainty quantification: theory, implementation, and applications* (pp. 11-44). Siam.

Smith, T., Marshall, L., & Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. Journal of Hydrology, 528, 29-37.

St. Joseph River watershed Initiative (SJRWI). http://wqis.pfw.edu/. Accessed: 26 Aug, 2018.

Tetra Tech, Inc., 2006. Wabash River Nutrient and Pathogen TMDL Development. Illinois Environmental Protection Agency and Indiana Department of Environmental Management.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(Jun), 211-244.

Tripathi, S., & Govindaraju, R. S. (2007). On selection of kernel parametes in relevance vector machines for hydrologic applications. Stochastic Environmental Research and Risk Assessment, 21(6), 747-764.

U.S. Geological Survey. (2016). National Water Information System data available on the World Wide Web (USGS Water Data for the Nation), accessed September 02, 2018, at URL http://waterdata.usgs.gov/nwis/.

United States Environmental Protection Agency. (2015). "Quality Criteria for Water 1986 [The Gold Book]| US EPA." Accessed June 16. http://yosemite.epa.gov/water/owrccatalog.nsf/9da204a4b4406ef885256ae0007a79c7/18888fcb7d1b9dc285256b0600724b5f!OpenDocument.

US EPA. (2009). "National Recommended Water Quality Criteria." http://water.epa.gov/scitech/swguidance/standards/criteria/current/index.cfm.

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V., & Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?. Stochastic environmental research and risk assessment, 23(7), 1011-1026.

Zhang, H. X., & Yu, S. L. (2004). Applying the first-order error analysis in determining the margin of safety for total maximum daily load computations. Journal of Environmental Engineering, 130(6), 664-673.

Zheng, Y., & Han, F. (2016). Markov Chain Monte Carlo (MCMC) uncertainty analysis for watershed water quality modeling and management. Stochastic environmental research and risk assessment, 30(1), 293-308.

Zheng, Y., & Keller, A. A. (2008). Stochastic Watershed Water Quality Simulation for TMDL Development–A Case Study in the Newport Bay Watershed 1. JAWRA Journal of the American Water Resources Association, 44(6), 1397-1410.

Zheng, Y., Wang, W., Han, F., & Ping, J. (2011). Uncertainty assessment for watershed water quality modeling: a probabilistic collocation method based approach. Advances in water resources, 34(7), 887-898

APPENDIX A

Table A1. List of LOADEST equations used for total phosphorus (TP) reconstruction at different monitoring stations in the study area

Station	TP	Weights	Variance of $\ln y$
100	4	D: $\ln y = 3.64 + 1.62\ln Q - 0.26\sin(2\pi\delta t) - 0.30\cos(2\pi\delta t)$	D: 0.52
		RVM: $\ln y = 3.67 + 1.58\ln Q - 0.17\sin(2\pi\delta t) - 0.22\cos(2\pi\delta t)$	RVM: 0.52
105	8	D: $\ln y = 1.15 + 1.39\ln Q + 0.13\ln Q^2 - 0.28\sin(2\pi\delta t) - 0.39\cos(2\pi\delta t) + 0.12\delta t$	D: 0.49
		RVM: $\ln y = 1.30 + 1.35\ln Q + 0.11\ln Q^2 - 0.11\sin(2\pi\delta t) - 0.24\cos(2\pi\delta t) + 0.10\delta t$	RVM:0.49
122	1	D: $\ln y = 3.65 + 1.35\ln Q$	D: 0.35
		RVM: $\ln y = 3.65 + 1.33\ln Q$	RVM: 0.34
126	1	D: $\ln y = 3.17 + 1.38\ln Q$	D: 0.50
		RVM: $\ln y = 3.16 + 1.37\ln Q$	RVM: 0.50
127	3	D: $\ln y = -0.85 + 1.28\ln Q + 0.17\delta t$	D: 0.46
		RVM: $\ln y = -0.85 + 1.27\ln Q + 0.15\delta t$	RVM: 0.45
129	3	D: $\ln y = 0.82 + 1.28\ln Q + 0.20\delta t$	D: 0.47
		RVM: $\ln y = 0.82 + 1.28\ln Q + 0.18\delta t$	RVM: 0.47
131	7	D: $\ln y = 0.013 + 1.30\ln Q - 0.51\sin(2\pi\delta t) - 0.61\cos(2\pi\delta t) + 0.09\delta t$	D: 0.45
		RVM: $\ln y = 0 + 1.30\ln Q - 0.50\sin(2\pi\delta t) - 0.62\cos(2\pi\delta t) + 0.05\delta t$	RVM: 0.44

150	7	D: $\ln y = -1.35 + 1.01\ln Q - 0.03\sin(2\pi\delta t) - 0.54\cos(2\pi\delta t) + 0.30\delta t$ RVM: $\ln y = -1.26 + 1.00\ln Q + 0\sin(2\pi\delta t) - 0.41\cos(2\pi\delta t) + 0.24\delta t$	D: 0.33 RVM: 0.32
159	1	D: $\ln y = 0 + 1.02\ln Q$ RVM: $\ln y = 0.000014 + 1.02\ln Q$	D: 1.11 RVM: 1.05

Table A2. TP. Estimated covariance matrix of weights at station 122 using RVM

	a_0	a_1
a_0	0.013380	0.005026
a_1	0.005026	0.016743

Table A3. List of LOADEST equations used for total dissolved solids (TDS) reconstruction at different monitoring stations in the study area

Station	TDS	Weights	Variance of $\ln y$
100	6	D: $\ln y = 12.04 + 0.82\ln Q - 0.02\ln Q^2 + 0.02\sin(2\pi\delta t) + 0.08\cos(2\pi\delta t)$ RVM: $\ln y = 12.05 + 0.83\ln Q - 0.02\ln Q^2 + 0.01\sin(2\pi\delta t) + 0.07\cos(2\pi\delta t)$	D: 0.034 RVM=0.034
105	1	D: $\ln y = 10.21 + 0.76\ln Q$ RVM: $\ln y = 10.21 + 0.75\ln Q$	D: 0.19 RVM: 0.19
122	4	D: $\ln y = 11.62 + 0.97\ln Q + 0.05\sin(2\pi\delta t) + 0.09\cos(2\pi\delta t)$ RVM: $\ln y = 11.65 + 0.94\ln Q + 0\sin(2\pi\delta t) + 0.04\cos(2\pi\delta t)$	D: 0.01 RVM: 0.01
123	9	D: $\ln y = 9.76 + 0.84\ln Q - 0.05\ln Q^2 - 0.04\sin(2\pi\delta t) + 0.03\cos(2\pi\delta t) + 0.01\delta t + 0\delta t^2$ RVM: $\ln y = 9.75 + 0.84\ln Q - 0.05\ln Q^2 - 0.02\sin(2\pi\delta t) + 0\cos(2\pi\delta t) + 0.01\delta t + 0\delta t^2$	D: 0.04 RVM: 0.04
124	3	D: $\ln y = 10.75 + 0.85\ln Q - 0.03\delta t$ RVM: $\ln y = 10.75 + 0.85\ln Q - 0.03\delta t$	D: 0.05 RVM: 0.05
125	1	D: $\ln y = 10.80 + 1.00\ln Q$ RVM: $\ln y = 10.80 + 1.00\ln Q$	D: 0.07 RVM: 0.07
126	9	D: $\ln y = 11.37 + 0.92\ln Q - 0.03\ln Q^2 + 0\sin(2\pi\delta t) + 0\cos(2\pi\delta t) - 0.01\delta t + 0\delta t^2$ RVM: $\ln y = 11.37 + 0.92\ln Q - 0.03\ln Q^2 + 0\sin(2\pi\delta t) + 0\cos(2\pi\delta t) - 0.01\delta t + 0\delta t^2$	D: 0.01 RVM: 0.01
127	7	D: $\ln y = 8.14 + 0.73\ln Q - 0.05\sin(2\pi\delta t) - 0.04\cos(2\pi\delta t) - 0.027\delta t$ RVM: $\ln y = 7.99 + 0.71\ln Q + 0\sin(2\pi\delta t) + 0\cos(2\pi\delta t) - 0.03\delta t$	D: 0.09 RVM: 0.10
128	2	D: $\ln y = 6.79 + 0.85\ln Q - 0.04\ln Q^2$ RVM: $\ln y = 6.80 + 0.85\ln Q - 0.04\ln Q^2$	D: 0.02 RVM: 0.02
129	9	D: $\ln y = 9.24 + 0.96\ln Q - 0.02\ln Q^2 - 0.13\sin(2\pi\delta t) + 0.09\cos(2\pi\delta t) + 0\delta t + 0.01\delta t^2$ RVM: $\ln y = 9.23 + 0.96\ln Q - 0.02\ln Q^2 - 0.12\sin(2\pi\delta t) + 0.09\cos(2\pi\delta t) + 0\delta t + 0.01\delta t^2$	D: 0.03 RVM: 0.03
131	9	D: $\ln y = 9.21 + 0.86\ln Q - 0.04\ln Q^2 + 0.06\sin(2\pi\delta t) + 0.07\cos(2\pi\delta t) + 0\delta t + 0\delta t^2$ RVM: $\ln y = 9.19 + 0.86\ln Q - 0.04\ln Q^2 + 0.05\sin(2\pi\delta t) + 0.06\cos(2\pi\delta t) + 0\delta t + 0.01\delta t^2$	D: 0.03 RVM: 0.03
150	9	D: $\ln y = 7.58 + 0.86\ln Q - 0.01\ln Q^2 - 0.11\sin(2\pi\delta t) - 0.09\cos(2\pi\delta t) - 0.02\delta t - 0.01\delta t^2$ RVM: $\ln y = 7.55 + 0.87\ln Q - 0.01\ln Q^2 - 0.09\sin(2\pi\delta t) - 0.07\cos(2\pi\delta t) - 0.02\delta t - 0.01\delta t^2$	D: 0.04 RVM: 0.04

159	6	D: $\ln y = 9.05 + 0.90\ln\mathrm{Q} - 0.04\ln\mathrm{Q}^2 - 0.04\sin(2\pi\delta\mathrm{t}) - 0.04\cos(2\pi\delta\mathrm{t})$ RVM: $\ln y = 9.04 + 0.91\ln\mathrm{Q} - 0.04\ln\mathrm{Q}^2 - 0.03\sin(2\pi\delta\mathrm{t}) - 0.02\cos(2\pi\delta\mathrm{t})$	D: 0.01 RVM: 0.01

Table A4. TDS. Estimated covariance matrix of weights at station 122 using RVM

	a_0	a_1	a_2
a_0	0.000705	0.000071	-0.000676
a_1	0.000071	0.000421	0.000128
a_2	-0.000676	0.000128	0.001286