OCIMUM

Ahmed I et al. Res Adv Environ Sci 2019(1): 45-60.

Research and Advances: Environmental Sciences

Research Article

On the Prediction of Uncertainty in a Sediment Provenance Model

Iftekhar Ahmed[1*], Abdullah Karim[1], Thomas W Boutton[2] and Kyle B Strom[3]

[1]Department of Civil and Environmental Engineering, Prairie View A&M University, Prairie View, Texas, USA
[2]Department of Ecosystem Science and Management, Texas A&M University, College Station, Texas, USA
[3]Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia, USA

Received Date: 22 February, 2019
Accepted Date: 14 March, 2019
Version of Record Online: 23 March, 2019

Citation

Ahmed I, Karim A, Boutton TW, Strom KB (2019) On the Prediction of Uncertainty in a Sediment Provenance Model. Res Adv Environ Sci 2019(1): 45-60.

Correspondence should be addressed to
Iftekhar Ahmed, USA
E-mail: ifahmed@pvamu.edu

Abstract

A Bayesian framework is created to tackle uncertainty in sediment provenance or fingerprint model. Two sources of uncertainty have been identified. One is from the physically based watershed sediment yield model runs due to spatial variation in the ground slopes. The other is from the use of long-range simulated episodic rainfall time series modeled using Markov Chains. The paper extends the Bayesian Markov Chain Monte Carlo (MCMC) algorithms of Fox and Papanicolaou (2008) for ensemble prediction of soil yield fraction or percentage from the floodplains adjacent to a stream, and the associated uncertainty. An erosion process parameter is the uncertain parameter of focus in this study because of its direct link with the physically based water erosion model. This link is identified in the Bayesian MCMC simulation runs. The study finds that less uncertainty is associated with sediment fraction yield estimation with increasing number of spatially distributed soil $\delta^{13}C$ carbon isotope and Carbon/Nitrogen (C/N) atomic ratio tracer data, and low-range episodic rainfall time series when the Bayesian MCMC method is used. The drawbacks of the frequentist Monte Carlo Simulation method are discussed. The work compliments that of Fox and Papanicolaou (2008) via the introduction of prediction parameter uncertainty comparison based on the two aforementioned methods.

Keywords

Bayesian MCMC; Carbon/Nitrogen Atomic Ratio; Gibbs Sampling; Monte Carlo Simulation; Rainfall; Sediment Yield; Stable Isotope; Uncertainty; Watershed

Introduction

Physically based watershed erosion model estimates contain high uncertainty when predicting soil erosion contributions from more than one source within the watershed. In addition, physically based watershed models are typically unable to account for more than one erosion process and are limited to uniform erosion down-cutting across the soil surface rather than accounting for the episodic nature of erosion during high rainfall events. To circumvent the problems with conventional modeling approaches, watershed field-based studies can be considered to estimate soil erosion from more than one land-use type during high rainfall events [1]. The procedure is termed "sediment fingerprinting," and is well documented in the literature. It is defined as a field-based technology that estimates the contribution of eroded-soil from each land-use source through the use of tracer measurements and application of a statistical "un-mixing" model [2-5]. The un-mixing model allows discrimination or separation of the amount of suspended sediment from a mixed field sample according to land-use types and thus, the term "un-mixing." Biogeochemical tracers

$\delta^{13}C$, $\delta^{15}N$, and C/N atomic ratios are typically used to differentiate the land-use sources based on contrasting pedologic and anthropogenic history (e.g. C3 versus C4 plant photosynthetic pathways and crop management, respectively) of land uses [6]. Figure 1 illustrates the concept of sediment fingerprinting [7]. The figure 1 shows erosion occurring over a watershed during a high rainfall event. Soil erodes from two land-use sources, source 1 and source 2, and the soil is transported to the watershed outlet where eroded-soil is collected throughout the duration of the event using *in situ* suspended eroded-soil trap. The traps function as integrated samplers and soil settles within the traps over the duration of the erosion period [8].

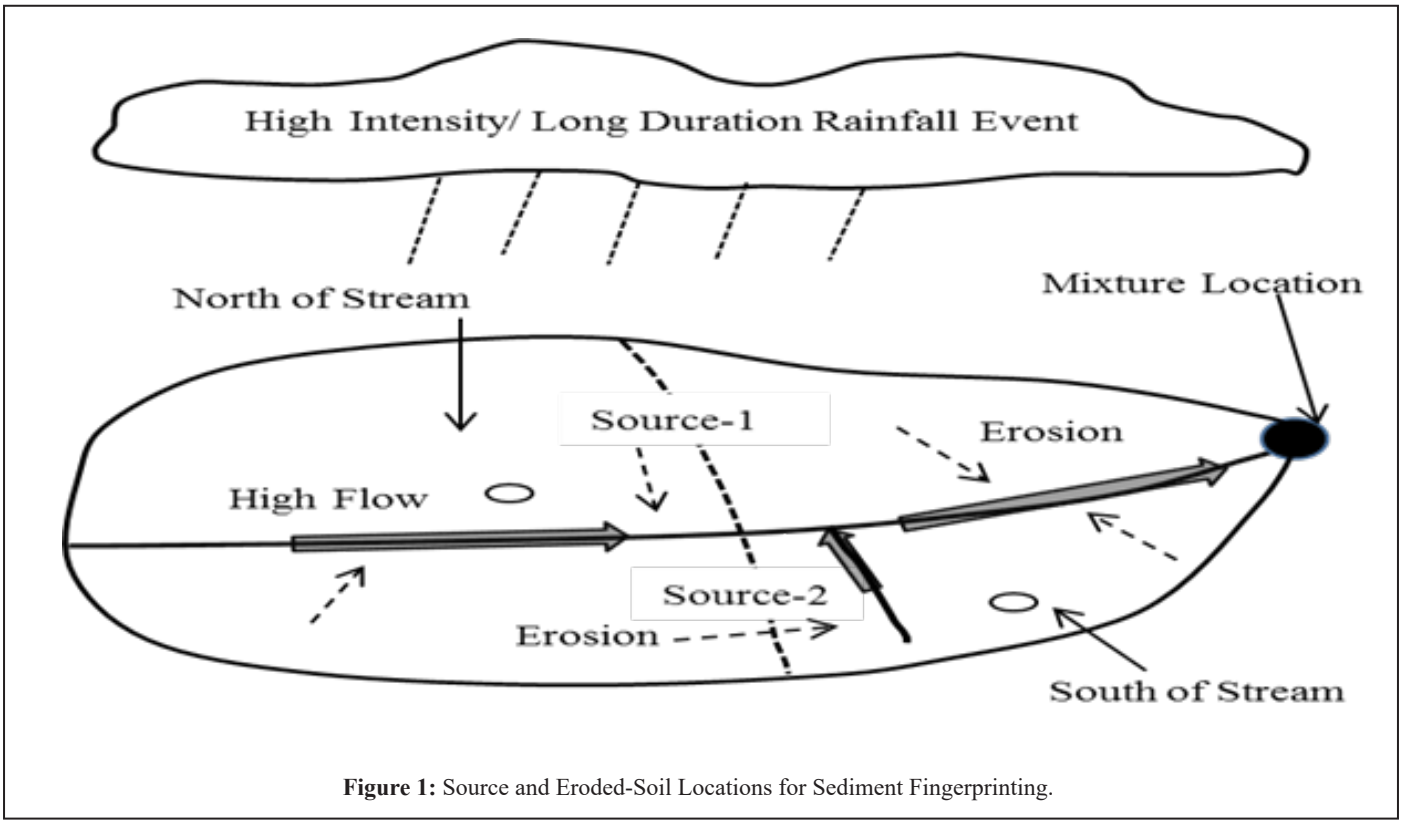

Figure 1: Source and Eroded-Soil Locations for Sediment Fingerprinting.

Two types of tracer are considered: (i) tracer data from land sources, and (ii) tracer data from eroded-soil at the watershed outlet. The use of multiple tracers leads to an over-determined system of mass-balance equation matrix with multivariate distribution of the tracers. The problem can be conveniently solved using Bayesian Markov Chain Monte Carlo (MCMC) simulation [9]. Fox and Papanicolaou [1] extended research to include erosion process parameter that is dependent on sediment yield estimates from physical water erosion models. Bayesian approach has advantage over optimization methods because specifying prior parameter distributions relaxes the need for full characterization which is not practical [10]. Figure 2 illustrates the research framework.

Sediment Fingerprint Mass Balance Model

Statistical un-mixing model can be applied to estimate the contribution of eroded soil percentages (as fractions) from each source sub-watershed by analyzing the tracer values of the source soil and the eroded-soil. An un-mixing model for the simplest scenario (two sub-watershed sources and one tracer) can be formulated as:

$$Z = X_1 P_1 + X_2 P_2 \quad (1)$$

$$P_1 + P_2 = 1 \quad (2)$$

where, Z represents mixture sample trace data, X stands for source sample mean, and P is the fraction contributed by the two eroded sediment sources. Subscript 1 and 2 stand for source 1 and 2, respectively. This un-mixing mass balance model becomes an over-determined system when more than two tracers are considered because this leads to more equations than unknowns. Two statistical solution schemes can be considered to solve the over-determined system: (1) a frequentist approach

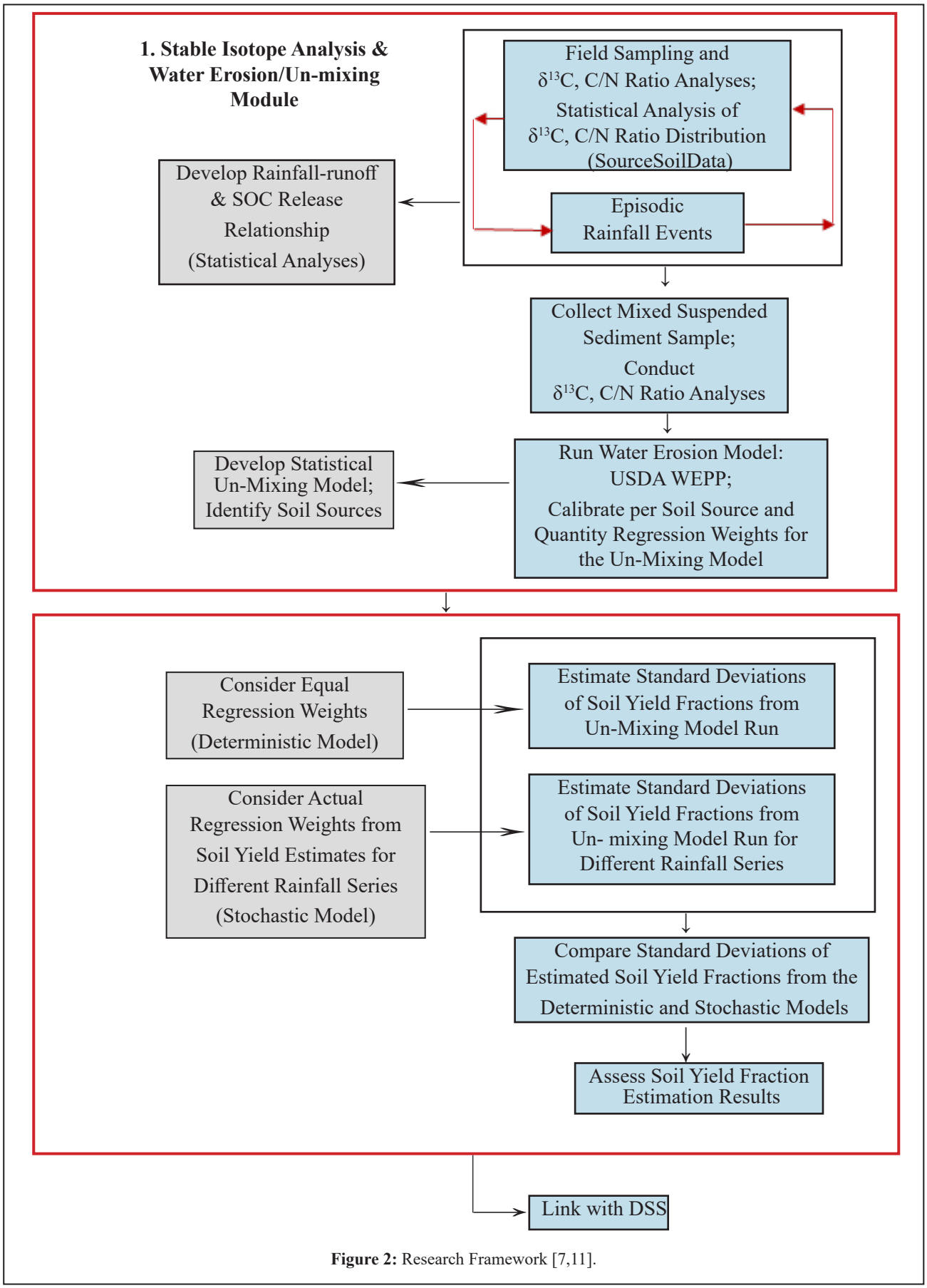

Figure 2: Research Framework [7,11].

based on least square error minimization [12-14], and (2) a Bayesian approach with MCMC simulation [1].

The matrix form of the mass balance equation with error terms to compensate for over-determined system is:

$$\begin{bmatrix} X_1^1 & X_2^1 & \dots & X_K^1 \\ X_1^2 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_1^{T-1} & \vdots & \vdots & \vdots \\ X_1^T & X_2^T & \dots & X_K^T \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix} = \begin{bmatrix} Z^1 \\ Z^2 \\ \vdots \\ Z^{T-1} \\ Z^T \end{bmatrix} + \begin{bmatrix} \varepsilon^1 \\ \varepsilon^2 \\ \vdots \\ \varepsilon^{T-1} \\ \varepsilon^T \end{bmatrix}$$

where, T is the total number of tracers, and K the number of sources. P denotes sources and ε^t is an error term introduced to solve the over-determined condition. Z is the mixture tracer data vector. Krause et al., [14] and Franks and Rowan [15] derived confidence intervals for the estimated fractions from each source, P_k, by using Monte Carlo sampling to draw from tracer source and mixture distributions. But Monte Carlo Simulation requires prior knowledge on distribution of sample data [10]. It is not always possible to collect desired number of soil samples in the nature. This leads to poor distribution of data. It is assumed that the uncertainty in the population mean of each source property can be represented by Students *t*-distribution with a confidence interval. To circumvent this problem, Fox and Papanicolaou [1] proposed Bayesian MCMC simulation method with low informative prior by treating the parameters as random variables and then training the posterior distribution of all model parameters. Joint distribution of random data sets can be tackled by the Bayesian MCMC simulation method. The two data sets used in this study are $\delta^{13}C$, and C/N atomic ratio.

Bayesian MCMC framework can facilitate the representation of more than one erosion process within the same sub-watershed or source by an erosion process parameter, and the use of an extra tracer distribution, to represent the episodic nature of the erosion, using the episodic erosion parameter. Two kinds of tracer data, from sediment sources, x, and from eroded-soil (as suspended sediment at a confluence), z are collected. Analysis of tracer values of the source soil and the eroded-soil estimate the contribution of eroded soil fractions from each source using the un-mixing model, based on the following mass balance equation [1]:

$$Z^T = \sum_K (x^T{}_k \times P_K), \sum_k P_k = 1 \qquad (4)$$

where, p is the fraction of eroded-soil contributed by the source, k. The tracer data from sediment sources, x, can be statistically represented by multivariate normal distribution with each tracer data value, i, and the index of soil erosion process, j, and the source type, k:

$$x_{jk}{}^i \sim MVN_T [\mu_{jk}, COV_{T\times T} (x_{jk})]$$

In Bayesian statistics, the mean, μ and covariance matrix, $COV(x)$ will have distribution of their own which are Multivariate Normal (MVN) and Wishart distributions, respectively, to facilitate MCMC simulation using Gibbs sampling in WinBUGS [16,17]:

$$\mu_{jk} \sim MVN\ (\theta, \tau)\ \ COV(x_{jk}) \sim Wishart\ (\omega, \rho)$$

where, θ, τ, ω and ρ can be specified as non-informative priors in the model. Similarly, the tracer data at all confluences can be represented by multivariate normal distribution:

$$z_{mixture} \sim MVN_T\ (\varphi, \Gamma)$$

with z being a vector of soil mixture tracer values, and $\Gamma \sim$ Wishart(Λ,ς). The parameter φ is specified in the deterministic equation for the mass balance inversion as [1]:

$$\varphi = \sum_k v_k P_k \qquad (5)$$

where, v_k is the soil erosion type identifier, and P_k has a Dirichlet distribution with parameter λ_k:

$$P_k \sim \text{Dirichlet}(\lambda_k)$$

An erosion process parameter, α_{jk} is considered by Fox and Papanicolaou [1] that includes the weights applied to each soil section where an erosion process is identified. It is a regression weight, estimated from soil yield estimates as the ratio of soil yield from any one sub-source to the summation of soil yields from all sub-sources within the same source type. This regression weight, α_{jk}, is estimated as:

$$\alpha_{jk} = \frac{S_{jk}}{\sum_j S_{jk}} \qquad (6)$$

where, S is the sediment yield. The erosion type identifier v_k is then α_{jk}, times x_{jk}. v_k is given a multivariate normal distribution and is a function of the episodic erosion parameter, β [1]. In this study, the parameter of uncertainty is the erosion process parameter, α_{jk}. The episodic erosion parameter is related to any grab sample after an episodic rainfall event. Such a grab sample is considered a member of the distribution of the soured sample tracer data. Over time, it is assumed that the entire watershed contributes to soil erosion and thus, a constant value of β is used in the un-mixing model. This assumption results in the same standard deviation and Monte Carlo (MC) errors for

soil yield fractions from two sources.

The sediment yield, *S,* is estimated using physical process-based WEPP erosion model [18] of the US Department of Agriculture. WEPP produces sediment yields from each sub-source (or sub-watershed). The summation of regression weights in a Bayesian multiple linear regression models should theoretically sum up to one. However, due to statistical independence of each WEPP run, the summation of weights may fall short of being one or could even be greater than one. This leads to uncertainty in soil fraction prediction. Bayesian MCMC with Gibbs sampling was applied to determine the probabilistic solution to the statistical un-mixing model for all parameters. The posterior distribution of all model parameters based on data is given by Bayes theorem:

$$P(\textit{All model Parameters} \mid x_{jk}, z) = P(\textit{All model parameters}) \times P(x_{jk}, z \mid \textit{All model parameters}) \quad (7)$$

The solutions to this model are the percentages of soils contributed by different sub-watersheds or sources.

Multivariate Continuous PDFs

Probability density functions for the mean, the variance or the covariance for Dirichlet, Multivariate normal, and Wishart distributions are found in Ntzoufras [16]. The Dirichlet distribution is continuous multivariate probability distribution. It is parameterized by a vector of positive real numbers. The multivariate normal distribution generalizes the uni-variate normal distribution to multivariate. The Wishart distribution generalizes the univariate chi-square distribution to multivariate. It is a distribution for positive definite symmetric matrices, typically for covariance matrices. These distributions serve important purpose in multivariate Bayesian analysis. Table 1 summarizes these distributions and their function in the present study.

Process	Probability Density Function	Details
Tracer Distribution	Dirichlet distribution: $f(x) = \Gamma(\lambda)[\prod_{k=1}^{M}\Gamma(\lambda_k)]^{-1}\prod_{k=1}^{M}P_k^{\lambda_k-1}$ $E(X_k)=\frac{\lambda_k}{\lambda}$ $Cov(X_k,X_l)=\frac{-\lambda_k\lambda_l}{[\lambda^2(\lambda+1)]}$	$f(x)$, $E(X_k)$, and $Cov(X_k,X_l)$ are probability density function, mean, and covariance, respectively. $\boldsymbol{P}$ and $\boldsymbol{\lambda}$ are vectors of dimension M with elements $P[k]=P_k \in (0,1)$ and $\lambda[k]=\lambda_k>0$ with $\sum_k^M P_k=1$ and $\lambda=\sum_{k=1}^{M}\lambda_k$
Fraction eroded soil distribution	Multivariate normal distribution: $f(x)=[(2\pi)^{-M/2}\lvert\Sigma\rvert^{1/2}\exp[-1/2\,(x-\mu)^{\Sigma}$ $\times\Sigma(x-\mu)]$ $E(X)=\mu$ $V(X)=\Sigma^{-1}$	$\boldsymbol{x}$ and $\boldsymbol{\mu}$ are vectors of dimension M with elements $x[i]=x_i \in (0,1)$ and $\mu[i]=\mu_i \in R$. $\boldsymbol{\Sigma}$ is $M \times M$ symmetric precision matrix
Covariance Matrix of tracers	Wishart distribution: $f(x)$ $=\lvert\Omega\rvert^{\rho/2}\lvert\Sigma\rvert^{(\rho-M-1)/2}\exp[-1/2\ Tr(\Omega\Sigma)]$ $E(X_{jk})=\rho A_{jk}$ $Cov(X_{jk},X_{lm})=\rho(A_{jl}A_{km}+A_{jm}A_{kl})$	$\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are $M \times M$ matrix of positive-definite (symmetric) matrices with elements Σ_{jk} and Ω_{jk}. A_{jk} are elements of matrix $\boldsymbol{A=R^{-1}}$, and $\rho>0$. Also mean, μ_i is the mean of X_i. $Cov(X_j,X_k)$ is covariance between X_j and X_k

Table 1: Multivariate Continuous PDFs for Bayesian Analysis in WinBUGS [16].

Transition Probability and Gibbs Sampling

Markov Chain Monte Carlo (MCMC) simulation is an iterative process that generates a chain of values for every parameter specified in the model. The samples are drawn from posterior distribution that is generated by a Markov Chain as its limiting distribution. Gibbs sampling, an extension of Metropolis-Hastings algorithm is used to accomplish this. When a Markov Chain is run for a long time till its limiting distribution, a random draw from the posterior is approximated by any selected parameter value since the initial run.

The solution of the linear system below gives steady state probability which ultimately yields long-run distribution of ergodic Markov Chain:

$$\pi = \pi P \quad (8)$$

where, π is the row vector of steady-state probabilities which is unconditional within a state space, and P is the one-step transition matrix for a Markov Chain. The right-hand side of the equation becomes the weighted sum of the one-step probabilities of entering that state from all other states which are all weighted by its respective long-run probabilities. Equation 8 becomes continuous when the long-run distribution of Markov Chain is in a continuous state space and the following equation for all measurable subsets of parameter space holds [19]:

$$\int_A \pi(\theta)d\theta = \int \pi(\theta)\, P(\theta, A)d\theta \quad (9)$$

where, θ is a vector of parameters namely, μ (mean) and σ^2 (variance), $\pi(\theta)$ is the long-run distribution of Markov Chain, A is the subset that measures parameter space, and $P(\theta,A)$ is the transition kernel of the chain. According to Equation 9, total steady state probability of set A is equal to steady state flow of probability into set A.

The explanation on how to find a Markov Chain that has the posterior as its long-run distribution is found in Bolstad [19] and briefly presented below for brevity. The interest is to find a probability transition kernel that satisfies the following equation for all A:

$$\int f(\theta|y)\, P(\theta, A)d\theta = \int_A f(\theta|y)\, d\theta \quad (10)$$

where, $f(\theta \mid y)$ is posterior density (without the scale factor to make it an exact density), and $P(\theta,A)$ is transition kernel of the chain. Suppose a transitional kernel is found that balances the steady state flow between every possible pair of states, then the long-run distribution for the ergodic Markov chain with that transition kernel is equal to the posterior density. The reversibility condition states that probability of moving from the starting value to candidate value of a parameter is equal to the probability of moving in the reverse direction. It can be expressed for all θ and θ' by Bolstad [19]:

$$f(\theta|y)\, q(\theta, \theta') = f(\theta' \mid y)\; q(\theta', \theta) \quad (11)$$

where, θ' is a candidate value, θ is the starting value, $f(\theta \mid y)$ is the target density of current value, and $f(\theta' \mid y)$ is the target density of candidate value, and $q(\theta, \theta')$ is a candidate distribution that generates θ' given θ. If candidate distribution satisfies Equation 11, then $f(\theta| y)$ becomes the long-run distribution of Markov chain. Then the probability kernel can be expressed as Bolstad [19]:

$$P(\theta, A) = \int_A q(\theta, \theta')d\theta' + r(\theta)\, \delta_A(\theta) \quad (12)$$

where, $P(\theta,A)$ is the transition kernel of the chain, $q(\theta, \theta')$ is a candidate distribution that generates θ' given θ, r (θ) is the probability the chain remains at θ and equals to $[1 - \int q(\theta, \theta')d\theta']$, and $\delta_A(\theta)$ is indicator function of set A. $\delta_A(\theta)$ can be further defined as:

$$\delta_A(\theta) = 1 \text{ (if } \theta \in A\text{)}$$
$$= 0 \text{ (if } \theta \notin A\text{)} \quad (13)$$

Combining Equations 11 and 13 yield

$$\int f(\theta|y)\, P(\theta, A)d\theta = \iint_A f(\theta|y)q(\theta, \theta')d\theta'\, d\theta + \int f(\theta|y)\, r(\theta)\delta_A(\theta)\, d\theta$$

and, by reversing the order of integration,

$$= \iint_A f(\theta|y)\, q(\theta, \theta')d\theta d\theta' + \int_A f(\theta|y)\, r(\theta)d\theta$$

Following the reversibility condition,

$$= \iint_A f(\theta'|y)\, q(\theta', \theta)d\theta d\theta' + \int_A f(\theta|y)\, r(\theta)d\theta$$

and by substitution,

$$= \int_A f(\theta'|y)\, (1 - r(\theta'))\, d\theta' + \int_A f(\theta|y)\, r(\theta)d\theta$$

as both θ and θ' are dummies of integration,

$$= \int_A f(\theta|y)d\theta \quad \textbf{qed.} \quad (14)$$

The steps of Blockwise Metropolis-Hastings algorithm for parameter sampling can be found in Bolstad [19]. In this system the parameter vector is partitioned into blocks as follows:

$$\boldsymbol{\theta = \theta_1, \theta_2, \ldots\ldots\ldots\ldots\ldots, \theta_J} \quad (15)$$

where, $\boldsymbol{\theta_j}$ is block of parameters when $\boldsymbol{\theta_{-j}}$ are all other parameters not in block j. The steps of Blockwise Metropolis-Hastings algorithm are Bolstad [19]:

1. Start at point in parameter space $\boldsymbol{\theta_1^0, \theta_2^0 \ldots\ldots\ldots\ldots\ldots \theta_J^0}$
2. For step $n = 1, \ldots\ldots\ldots\ldots, N$

For $j = 1,\ldots\ldots\ldots J$

- draw candidate from

$$q\left(\boldsymbol{\theta}_j^{(n-1)}, \boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_1^{(n)}, \ldots\ldots\ldots\ldots, \boldsymbol{\theta}_{j-1}^{(n)}, \boldsymbol{\theta}_{j+1}^{(n-1)} \ldots\ldots\ldots\ldots, \boldsymbol{\theta}_J^{(n-1)}\right)$$

- calculate the acceptance probability from

$$\alpha\left(\boldsymbol{\theta}_j^{(n-1)}, \boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_1^{(n)}, \ldots\ldots\ldots\ldots, \boldsymbol{\theta}_{j-1}^{(n)}, \boldsymbol{\theta}_{j+1}^{(n-1)} \ldots\ldots\ldots\ldots, \boldsymbol{\theta}_J^{(n-1)}\right)$$

- draw u from Uniform (0,1)
- if $u < \alpha\left(\theta_j^{(n-1)}, \theta_j'\right)$ then let $\theta_j^n = \theta_j'$, else let $\theta_j^n = \theta_j^{(n-1)}$

where, α is the acceptance probability. The procedure is illustrated in figure 3.

At step n, the candidate for the j^{th} block of parameters, θ_j is drawn, given the most recently updated values of the other blocks of parameters. Since the cycle is carried out through the parameters, the 1^{st} through the $(j-1)^{th}$ blocks of parameter values have already been updated to the nth step. The $(j+1)^{th}$ through the J^{th} blocks have not been updated and are still at $(n-1)^{th}$ step. u is greater than α for up to the $(j-1)^{th}$ block. At the j^{th} block, u is less than α, and the candidate is accepted (Figure 3). The advantage of Gibbs sampling over Metropolis-Hastings algorithm is that in Gibbs sampling, u is always less than α and thus the candidate parameter is always accepted.

Gibbs sampler, a special extension of the Metropolis-Hastings Algorithm, is applied when the full conditional distributions of each component in the multivariate distribution given all other components are known. This enables the use of true conditional density as the candidate density at each step for every block of parameters given the others. This situation is described by Bolstad [19]:

$$q\left(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_{-j}\right) = f\left(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y}\right) \qquad (16)$$

where, $\boldsymbol{\theta}_j$ is the block of parameters when $\boldsymbol{\theta}_{-j}$ are all other parameters not in block j, $q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_{-j})$ is the candidate density that generates candidate $\boldsymbol{\theta}_j'$ given starting value $\boldsymbol{\theta}_j$, and $f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})$ is the target density. At step n for block $\boldsymbol{\theta}_j$, the

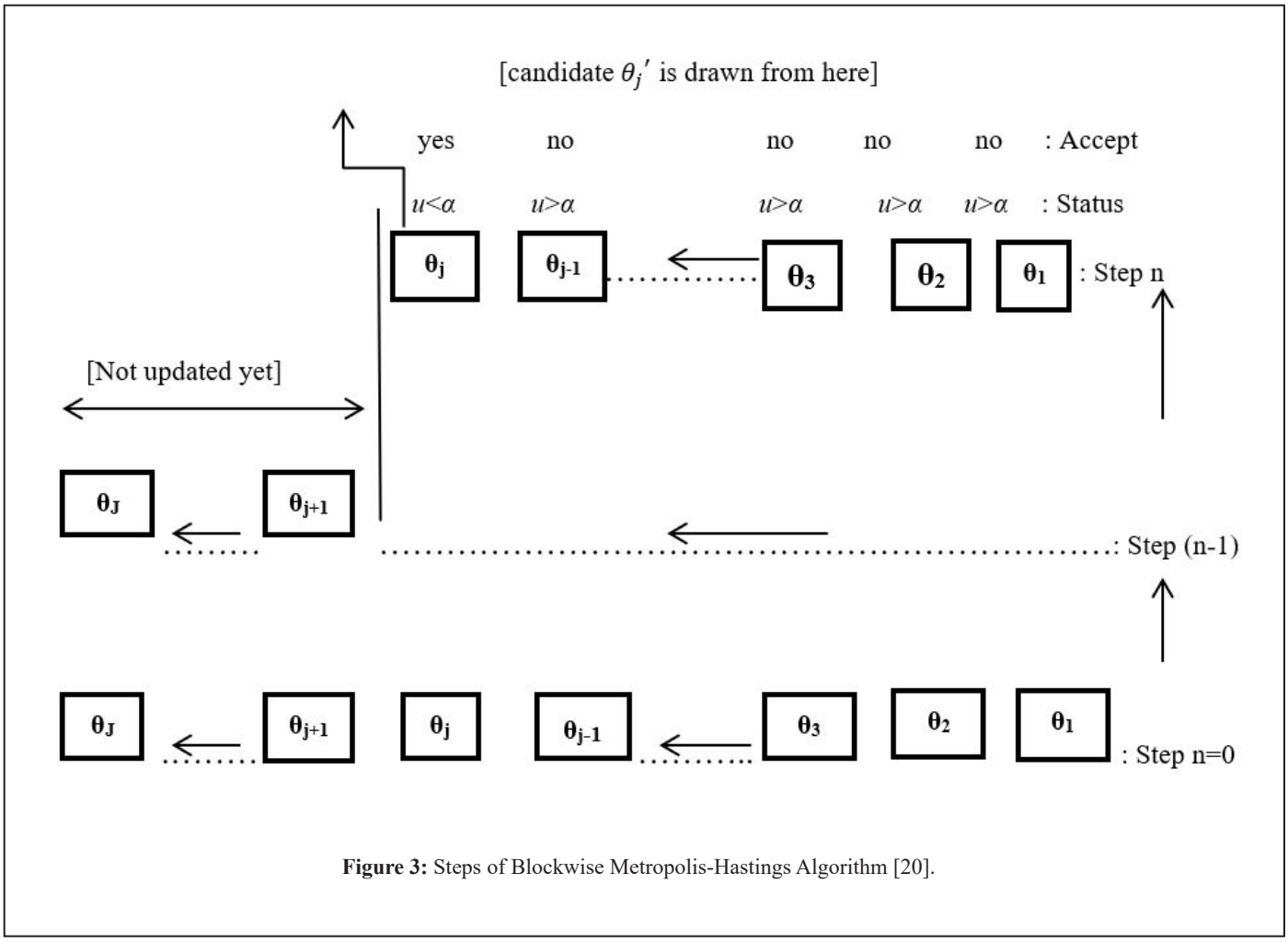

Figure 3: Steps of Blockwise Metropolis-Hastings Algorithm [20].

acceptance probability is given by Bolstad [19]:

$$\alpha\ (\theta_j^{(n-1)},\theta_j' \mid \theta_1^{(n)},\ldots\ldots\ldots\ldots,\theta_{j-1}^{(n)},\theta_{j+1}^{(n-1)}\ldots\ldots\ldots\ldots,\theta_J^{(n-1)}$$

$$=\min\left[1,\frac{f(\theta_j' \mid \theta_{-j},y)\ q(\theta_j',\theta_j \mid \theta_{-j})}{f(\theta_j \mid \theta_{-j},y)\ q(\theta_j,\theta_{-j} \mid \theta_{-j})}\right]$$

$$=1 \qquad (17)$$

Hence, all samples are accepted in Gibbs sampling method in each step leading to improved computational efficiency. Thus, Gibbs sampling is achieved in the case where each candidate block is drawn from its true conditional density given all other blocks at their most recently drawn values.

The Two-Step Simulation Process: Regression and "Un-Mixing" Model

The two-step regression and un-mixing modeling procedure is shown in the following flowcharts (Figure 4 and 5). The first flowchart (the regression module) discusses the use of the raw tracer data. The distributions are input as prior statistics to generate the posterior distributions of all model parameters. The second flowchart (the "un-mixing" module) uses the mean and standard deviation information from the regression module to generate percent sediment yield contribution from two sub-sources given the mixture sample's tracer data at a confluence of a Bayou and its tributary. The Regression model constitutes three major parts: Model, Initial Values and Data. The Model part includes model definition, likelihood, prior distribution,

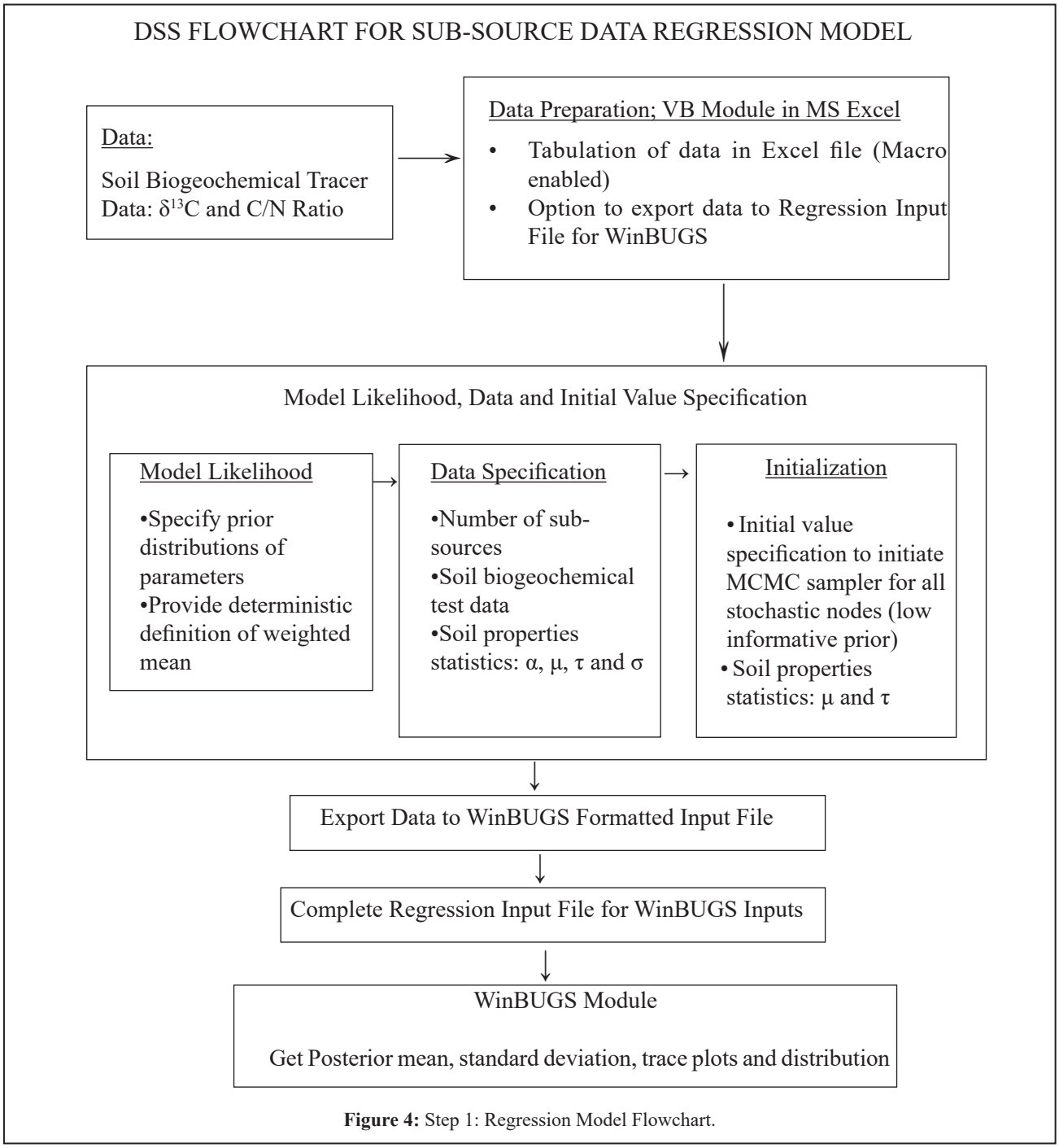

Figure 4: Step 1: Regression Model Flowchart.

and deterministic definitions of model parameters. The Initial Values part contains starting values of model parameters and the Data part holds data specifications and field data. The Un-mixing model flowchart is self-explanatory and uses posterior results from the regression model.

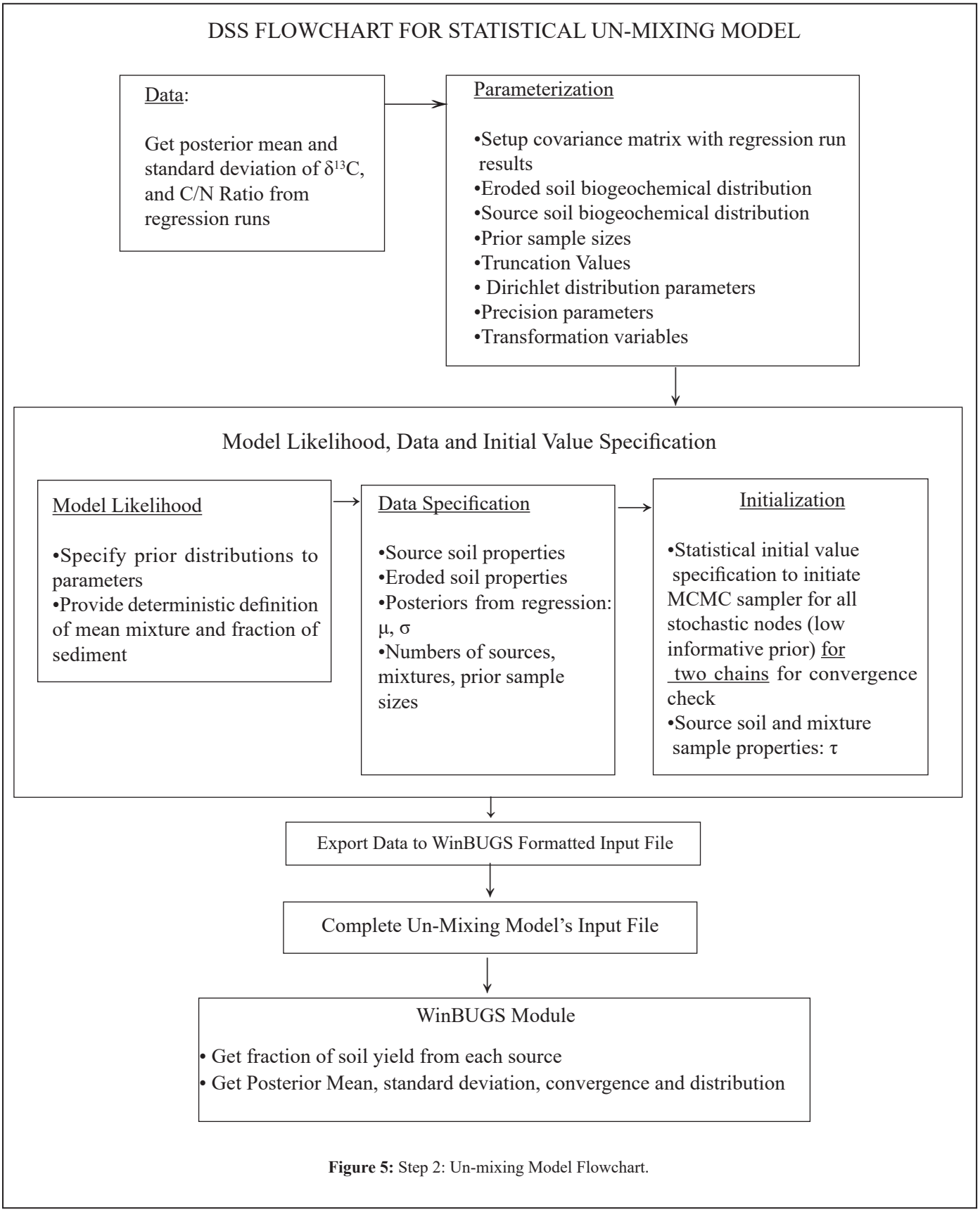

Figure 5: Step 2: Un-mixing Model Flowchart.

Study Area

The land area investigated is the highly urbanized Buffalo Bayou Watershed of Houston, Texas. Soil biogeochemical properties ($\delta^{13}C$ tracer isotope and C/N atomic ratio) were analyzed at Texas A&M University's (TAMU) Stable Isotopes for Biosphere Science (SIBS) laboratory. The project area was divided according to the University of Houston (UH) team's Buffalo Bayou *in-situ* suspended sediment sampling locations where mixture samples were collected during high flow events. The North and the South of the Buffalo Bayou are regarded as two sources of soil erosion for any given UH suspended sample mixture at a Bayou-Tributary confluence. Sensitivity of soil yield response to the episodic Markov rainfall structure, sub-watershed slopes, vegetation cover, and soil properties were analyzed by the physically based Water Erosion Prediction Project (WEPP) model of the US Department of Agriculture. Periodic rainfall time series was simulated using the CLIGEN weather generator of WEPP for 6-, 11-, 16-, and 22-year rainfall time series windows. Soil yield values generated by WEPP provided the regression weights or parameter α_{jk}.

The Buffalo Bayou watershed in Houston city limits extends from West limit (95°38´48″W) to East limit (95°21´30.36″W), from Highway-6 to the confluence of the Buffalo and the White Oaks Bayous. Its North-South limits are 29°49´22.17″N (at North) and 29°43´10.17″N (at South), respectively. The limits are marked in figure 6. The methods were tested on 15 square miles of the watershed starting from the western limit to UH sampling location number 6 along the Buffalo Bayou (Figure 7). Future work will look at localized areas within the watershed.

Land sampling was designed to follow the drainage network of watershed dictated by a map provided by the Harris County Flood Control District (HCFCD). All land samplig were done in the vicinity of an open or underground channel. Where no such channel was found, soil samples were collected by the drainage gutters. Noteworthy, this leads to a sparse distribution of soil on the whole watershed but fairs well when viewed over a smaller sub-watershed within it. This was validated using geostatistical analysis in ArcGIS. This strategy was necessary to be able to determine the percent soil erosion contribution from sub-watersheds per the Bayesian statistical un-mixing model of Fox and Papanicolaou [1]. In a nutshell, the project teams' purpose was to collect many samples in the whole watershed but not so many within a sub-watershed. The updates in MCMC simulation eventually results in ample realizations to produce a posterior confidence interval [19].

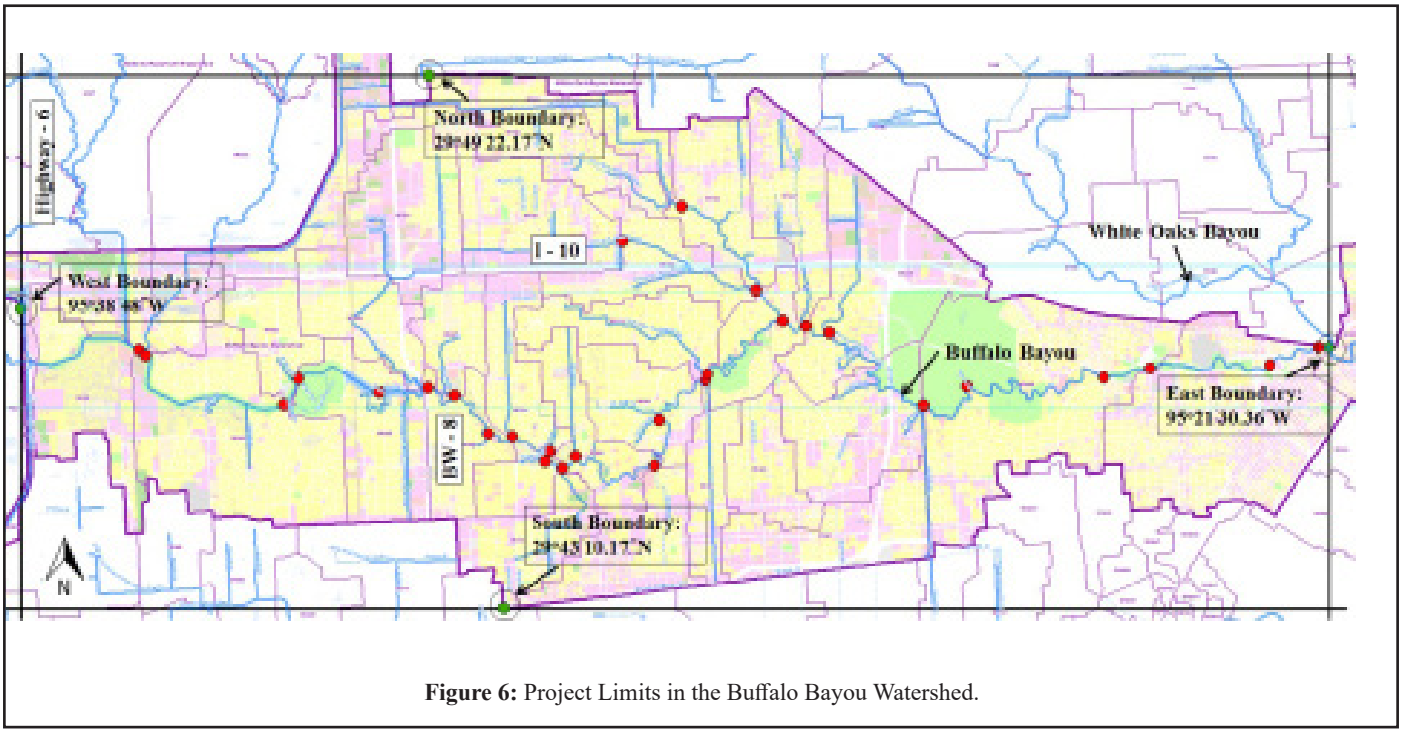

Figure 6: Project Limits in the Buffalo Bayou Watershed.

To quantify uncertainty in sediment provenance model, the project area was divided according to the Buffalo Bayou *in-situ* suspended sampling locations UH-1, UH-2, UH-3, UH-4, UH-5, UH-6, etc. moving downstream (Figure 7). Tracer data for land samples to the west of these UH sampling points were considered for analyses. These points were then categorized based on their positions in North and South of the Buffalo Bayou watershed. These areas are further subdivided equally (sampling points wise) into two sub-sources at both North and South of the Bayou. This set up was previously explained graphically in figure 1.

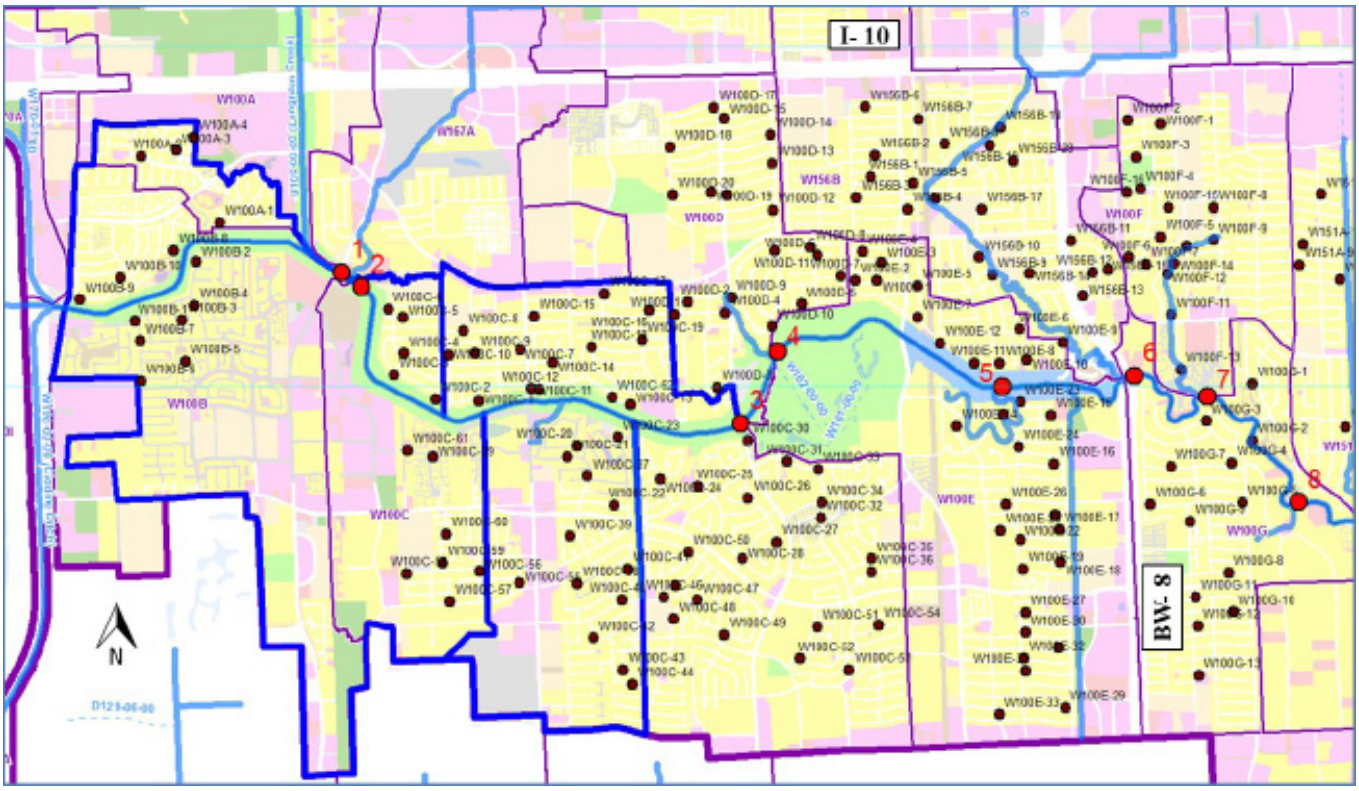

Figure 7: Land and Bayou Sampling Locations along the Buffalo Bayou Corridor.

Table 2 lists the contributing areas and the number of sampling point data used in the models. The number of sampling points increases with increasing UH station number moving downstream.

UH Station ID	Contributing Area (sq. miles)	No. of $\delta^{13}C$ and C/N Ratio Data to the Left (North)	No. of $\delta^{13}C$ and C/N Ratio Data to the Left (South)
3	9.2	27	25
4	9.8	42	37
5	13.3	71	51
6	15.4	80	68

Table 2: Contributing Area and the Number of Sampling Point Data Used in the Model.

Deterministic versus Stochastic Model

The soil fraction yield model was tested using deterministic model with equal regression weights on the two tracer data sets ($\delta^{13}C$ and C/N atomic ratio). As noted earlier, the summation of regression weights in a Bayesian multiple linear regression models should theoretically sum up to one. However, due to the statistical independence of each Water Erosion model (WEPP) run, the summation of weights may fall short of being one or could even be greater than one as is noted in table 3. This can lead to greater uncertainty (in terms of standard deviation) in soil fraction prediction. Choice of equal weights assumes the sediment yield prediction to be independent of rainfall pattern. The regression weights are represented by the erosion process parameter, α_{jk}, which is written in terms of soil yield estimates from WEPP (see equation 6) and allows the link between WEPP generated rainfall distribution and soil yield. A stochastic problem evolves when α_{jk} is no longer equal for the two data sets used in the Bayesian MCMC regression model to obtain posterior mean and variance used in the Bayesian un-mixing model.

Table 4 summarizes the posterior statistics from a Bayesian un-mixing model run with equal regression weights. P_1 and P_2 from the un-mixing model run considering equal regression weights are 0.875 and 0.125, respectively. This means 87.5% of the eroded soil is from the North of the Buffalo Bayou. The South contributes 12.5% of the eroded soil. The standard

deviation is 0.05088. The length of confidence interval is 0.1972. Also, its median of 0.8806 suggests that the posterior density plot would be close to normally distributed. The length of the confidence interval is 0.1972. The MC Error is 0.001448. Bayesian un-mixing model with variable regression weights was previously discussed. The mean sediment yield fractions and standard deviations with regression weights corresponding to the 6-, 11-, and 16-year rainfall series, and the 22-year rainfall series for UH-4 station are given in table 5, including the results from the deterministic model run with equal regression weights.

The standard deviations are 0.05415 and 0.05468 for the 6-, 11-, and 16-year rainfall series and the 22-year rainfall series-based runs, respectively, with different weights. When equal regression weights of 0.5 are considered, the standard deviations dropped to 0.05088 (Table 4). The estimated P_1 and P_2 of 0.875 and 0.125 with equal regression weights may not be deemed realistic since the uncertainty associated with the erosion process parameter (α_{jk}) are not considered in the model.

The same value of the standard deviations and the MC Errors for North and South of the Bayou in table 4 is due to the use of constant episodic erosion parameter (β) in Fox and Papanicolao [1]. The assumption is that over time, the entire watershed contributes to soil erosion and thus, a constant value of β is used in the un-mixing model. This assumption results in the same standard deviation and Monte Carlo (MC) errors for soil yield fractions from two sources.

Monte Carlo Simulation with Single Tracer Distribution

The advantage of a Bayesian MCMC model is in its ability to tackle multiple tracer data distributions to output posterior statistics of all tracers considered [1]. The traditional methods use Monte Carlo Simulation (MCS) to find confidence intervals on the prediction parameters (e.g. soil fraction yield) based on a single tracer data distribution at a time. This distribution for use in the MCS model also needs to be accurately known and have normal distribution which is rarely found in the nature and specifically in the urbanized Buffalo Bayou Watershed. MCS was applied on the distribution of the $\delta^{13}C$ data to see how the uncertainty results compare with that from the Bayesian MCMC model run. Noteworthy, the two methods are completely different. Table 6 summarizes the t-distribution statistics for $\delta^{13}C$ tracer data for all sampling points to the west (upstream) of UH-4 station: sample mean, standard deviation (St. Dev.), Degree of Freedom (DOF), and upper and lower limit of 95% confidence interval. Degree of Freedom (DOF) = (Number of samples) - 1.

Case	Source	Sub-source-1	Sub-source-2
Equal regression weights	North	0.5	0.5
(Deterministic)	South	0.5	0.5
The 6-, 11-, and 16-year rainfall series	North	0.44	0.51
(Stochastic)	South	0.49	0.42
The 22-year rainfall series	North	0.44	0.51
(Stochastic)	South	0.48	0.42

Table 3: Regression Weights (Deterministic and Stochastic) Summary for UH-4.

Source	Mean	St. Dev.	MC Error	2.50%	Median	97.50%
North	0.875	0.05088	0.001448	0.7598	0.8806	0.957
South	0.125	0.05088	0.001448	0.04298	0.1194	0.2402

Table 4: Bayesian Un-mixing Posterior Statistics with Equal Regression Weights for UH-4 (Deterministic Model).

Source	The 6-, 11-, and 16-Year Rainfall Series		The 22-Year Rainfall Series	
	Mean	St. Dev.	Mean	St. Dev.
North	0.8914	0.05415	0.8833	0.05468
South	0.1086	0.05415	0.1167	0.05468

Table 5: Bayesian Un-mixing Posterior Statistics with Actual Regression Weights for UH-4 (Stochastic Model).

North of the Buffalo Bayou					South of the Buffalo Bayou				
Sample Mean	Sample St.Dev.	DOF	Lower Limit	Upper Limit	Sample Mean	Sample St.Dev.	DOF	Lower Limit	Upper Limit
-22.13	2.05	41	-21.48	-22.78	-22.1	2.27	36	-21.33	-22.87

Table 6: *t*-Distribution Statistics of $\delta^{13}C$ Tracer (‰) Data to the Left of UH-4 Station.

Sample means of the two $\delta^{13}C$ tracer data are -22.13 ‰ and -22.10 ‰ for North and South of the Buffalo Bayou, respectively which are nearly the same. ‰ or "per mil" is the standard unit for the carbon isotope signature measurements. There were total 42 land sample data from the North and 37 from the South. Sample standard deviation was 2.05 ‰ for the North which is slightly lower than the standard deviation of 2.27 ‰ from the South of the Buffalo Bayou. That is why the range of confidence limits (difference between the upper and the lower limit of the confidence interval) of $\delta^{13}C$ tracers at North of the Buffalo Bayou is 1.29 ‰, which is lower than that range in the South of the Buffalo Bayou which is 1.53 ‰. Thus, distribution of $\delta^{13}C$ tracer data is tighter to the North of the Buffalo Bayou.

The MCS model was run for 100 to 1000 iterations for the same number of pairs of soil yield fractions from the North (P_1) and the South (P_2) of the Buffalo Bayou. A *t*-distribution was conducted on these generated pairs of fraction yield values from the North and the South with 95% Confidence Interval (CI) to monitor the uncertainty. The confidence intervals from the t-distribution statistics is shown in figures 8 and 9 for various iteration numbers to demonstrate the confidence interval reduction with increasing number of iterations.

Comparison of Uncertainty from Bayesian MCMC and Monte Carlo Simulation

A comparison was made on the standard deviations from the Bayesian MCMC and those from the increasing order of MCS iterations. Note that MCS does not have the capability to update distributions but is only dependent on the randomly generated sample means and a confidence interval on the sample data. Tables 7 and 8 summarize the estimated standard deviations from the two types of simulations. It is evidently clear that the Bayesian method outputs higher standard deviations compared with the MCS method. The Bayesian MCMC method is conservative and hints towards a risk aversive decision algorithm. The MCS method is dependent on the user defined number of iterations with no Markov Chain memory of the

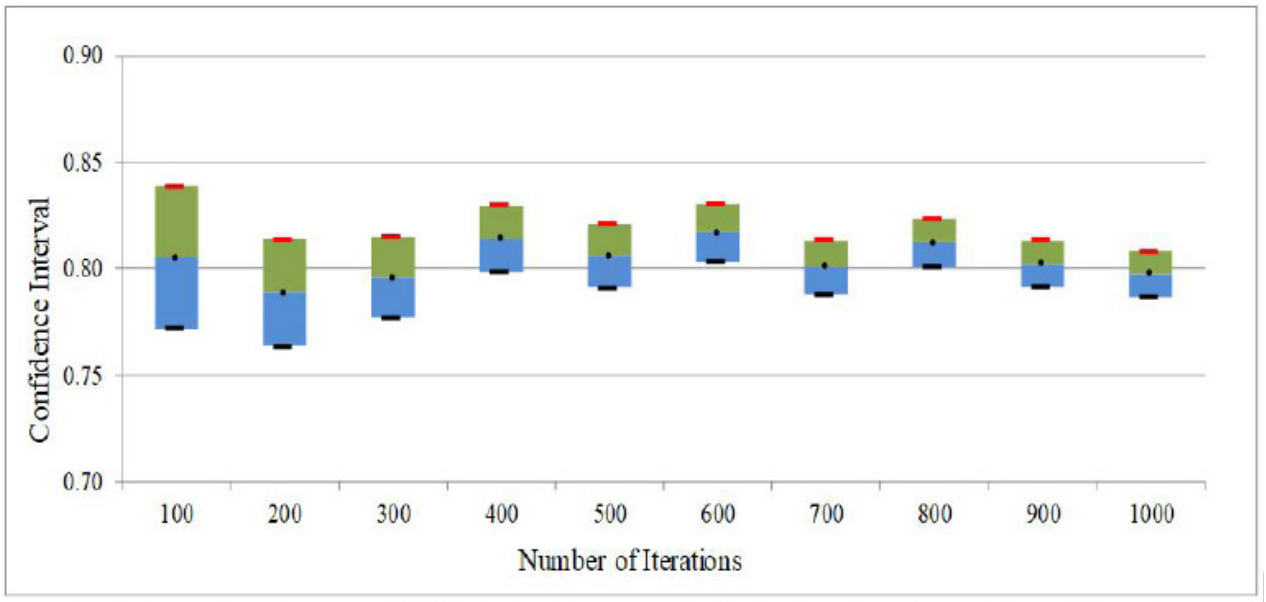

Figure 8: CI Plots Explaining *t*-Distribution Statistics on Soil Yield Fractions from North of the Bayou (P_1).

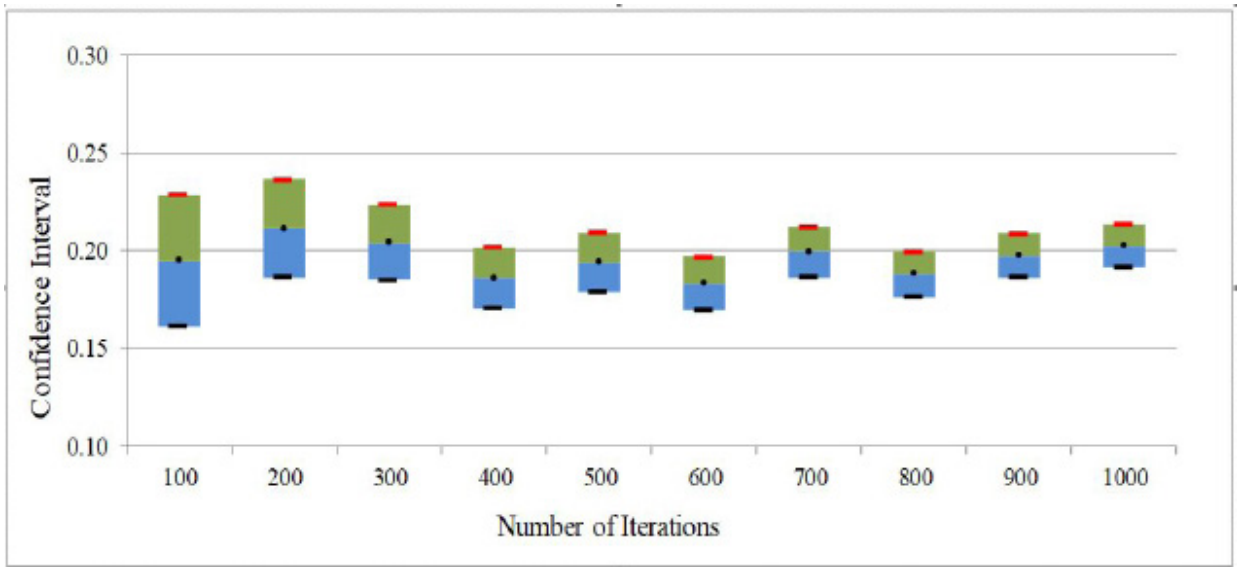

Figure 9: CI Plots Explaining *t*-Distribution Statistics on Soil Yield Fractions from South of the Bayou (P_2).

immediate past iteration; this is indicative of the mathematical robustness of the Bayesian MCMC approach. As expected, the MCS method based standard deviation is decreased with higher iteration number (Table 8). The main drawback with the MCS method in the context of this research is that it cannot be linked to the Markov Chain rainfall series which drives the water erosion of soil.

The output from the Bayesian MCMC method shows (Table 7) an increase in the standard deviation with increasing range of rainfall time series. This 'trend' is expected and is a good sign because long-range rainfall series typically does not preserve the original statistics or in other words, loses all statistical memory of the initial conditions. Therefore, the use of long-range forecast does not necessarily attribute to acceptable solutions but may provide insight on the system's sensitivity to weather variation. Exceptions can be seen in table 7 with nonlinear behavior across the table for UH sampling location number 5. The non-linear system behavior was an extension of an episodic rainfall pattern on the watershed which showed a mild sinusoidal pattern from month to month [21].

UH sampling ID	6-, and 11-Year Rainfall Series	16-Year Rainfall Series	22-Year Rainfall Series
3	0.06386	0.07478	0.07478
4	0.05415	0.05415	0.05468
5	0.04828	0.04149	0.04364
6	0.04076	0.04076	0.04293

Table 7: Standard Deviations of Soil Yield Fractions from Bayesian MCMC Simulations.

No. of Iterations	Standard Deviation	No. of Iterations	Standard Deviation
100	0.033636	600	0.01376
200	0.025089	700	0.012845
300	0.019147	800	0.01143
400	0.015611	900	0.010985
500	0.015023	1000	0.010774

Table 8: Standard Deviations of Soil Yield Fractions from Monte Carlo Simulations.

Conclusion

The main source of uncertainty in the variable pattern of soil erosion was captured by the use of erosion process parameter (the regression weight) that provided the link between physically based weather supported erosion model WEPP and the Bayesian MCMC framework. Such link cannot be established when MCS is used stand-alone. The episodic rainfall trends on the large watershed introduced limitation on the predictability of soil yield fraction estimation. Analysis was completed with the Bayesian MCMC sediment fingerprinting algorithm of Fox and Papanicolaou [1] to methodically estimate land area average soil yield fraction from the North and the South of the Bayou. Reasonable initialization of sample data statistics was very important for the desirable convergence of the Bayesian MCMC model given the sparse nature of the tracer data on the urbanized watershed. The posterior statistics obtained from successful initialization ensured the credibility of soil yield fraction estimation using the biogeochemical soil properties: the $\delta^{13}C$ stable isotope and the C/N (Carbon/Nitrogen) atomic ratio.

This work re-affirms that the Bayesian MCMC algorithm of Fox and Papanicolaou [1] is a unique and mathematically robust approach that is scientifically and statistically sound and supports the interdisciplinary approach to field-based validation of the sediment fingerprinting technology. It takes into account the spatially variable soil biogeochemical properties to estimate source soil yield contribution with the uncertainty associated with it. One advantage of the Bayesian MCMC over traditional MCS was that it can work with low informative priors or low resolution statistics of the raw data since it generates millions of iterations in less than an hour on a DELL Optiplex 1080 PC to converge to a posterior distribution and thus, eliminates the gross tendencies or variations that would otherwise be seen due to initialization errors. An MCS runs with 1000 iterations may take up to 4 hours on the same computer.

The rainfall data distribution from the local rain gages showed episodic pattern that could be an influence of continental scale physical and climatic processes. This needs future investigation. Rainfall statistics were generated by using WEPP's stochastic CLIGEN weather generator and needs to be compared with other rainfall simulation algorithms that suit the climate of the coastal South Texas.

The Bayesian MCMC framework used in WinBUGS required prior knowledge for model parameters which was non-informative. The posterior statistics of this study were based on this prior information and could be biased. More advanced and informative knowledge on tracers and erosion processes could help to improve the prior information. Investigation of spatial and temporal correlation between collected suspended sediment mixture samples are of future interest. Mixture sample data from downstream may not be representative for watershed areas far upstream due to lack of such memory in the Bayesian MCMC algorithm of Fox and Papanicolaou [1]. Future theoretical research can be directed to look for options to incorporate such effect in Bayesian or any other suitable framework.

Under episodic rainfall pattern that is typical of South Texas climate, uncertainty quantification presented in this paper holds the potential for effective sediment and watershed management in relation to soil conservation. The Bayesian framework is a move away from the physically based watershed erosion models that are typically unable to account for more than one erosion process and are limited to uniform erosion down-cutting across the soil surface rather than accounting for the episodic nature of erosion during high rainfall events. The uncertainty, presented as the standard deviation, can provide the watershed managers a reasonable bound to work with when making long term soil conservation decisions.

Acknowledgement

The work was supported by USDA/NIFA Award No. 2011-38821-30970. The authors are grateful to Dr. James F. Fox of the University of Kentucky for insights, and to the students and associates who provided outstanding field and laboratory support from Prairie View A&M University, the University of Houston, and Texas A&M University.

References

1. Fox JF, Papanicolaou AN (2008) An un-mixing model to study watershed erosion processes. Adv Water Resour 31: 96-108.

2. Motha JA, Wallbrink PJ, Hairsine PB, Grayson RB (2002) Tracer properties of eroded sediment and source material. Hydrol Process 16: 1983-2000.

3. Gruszowski KE, Foster IDL, Lees JA, Charlesworth SM (2003) Sediment sources and transport pathways in a rural catchment, Herefordshire, UK. Hydrol Process 17: 2665-2681.

4. Collins AL, Walling DE, Leeks GJL (1998) Use of composite fingerprints to determine the provenance of the contemporary suspended sediment load transported by rivers. Earth Surf Process Landforms 23: 31-52.

5. Fox JF, Papanicolaou AN, Abaci O (2005) The impact of agricultural erosion processes upon $\delta^{15}N$, $\delta^{13}C$, and C/N signatures of eroded-soil. Proceedings of RCEM Conference, IL, USA.

6. Boutton TW, Archer SR, Midwood AJ, Zitzer SF, Bol R (1998) $\delta^{13}C$ values of soil organic carbon and their use in documenting vegetation change in a subtropical savanna ecosystem.

Geoderma 82: 5-41.

7. Karim A (2014) Bayesian Ensemble Prediction of Watershed Scale Sediment Delivery. M.S. Thesis. Civil Engineering, Prairie View A&M University, Prairie View, TX, USA.

8. Phillips JM, Russell MA, Walling DE (2000) Time-integrated sampling of fluvial suspended sediment: a simple methodology for small catchments. Hydrol Process 14: 2589-2602.

9. Brewer MJ, Soulsby C, Dunn SM (2002) A Bayesian Model for Compositional Data Analysis. In: Härdle W, Rönz B (eds). Compstat, Physica, Heidelberg, Germany. Pg no: 105-110.

10. Billheimer D (2001) Compositional receptor modeling. Environmetrics 12: 451-467.

11. Ahmed I, James AA, Boutton TW, Strom KB (2011) Hydrologic Influences on Soil Organic Carbon Loss Monitoring Using Stable Isotopes. Research Proposal to: USDA/NIFA, Washington, DC, USA.

12. Yu L, Oldfield F (1989) A multivariate mixing model for identifying sediment source from magnetic measurements. Quater Res 32: 168-181.

13. Collins AL, Walling DE, Leeks GJL (1997) Source type ascription for fluvial suspended sediment based on a quantitative composite fingerprinting technique. Catena 29: 1-27.

14. Krause AK, Franks SW, Kalma JD, Loughran RJ, Rowan JS (2003) Multi-parameter fingerprinting of sediment deposition in a small gullied catchment in SE Australia. Catena 53: 327-348.

15. Franks SW, Rowan JS (2000) Multi-parameter fingerprinting of sediment sources: Uncertainty estimation and tracer selection. In: Bentley LR, Sykes JF, Gray WG, Brebbia CA, Pinder GF, et al. (eds.). Computational methods in water resources XIII. Balkema, Rotterdam, Netherlands. Pg no: 1067-1074.

16. Ntzoufras I (2009) Bayesian Modeling Using WinBUGS. Wiley, New York, USA.

17. Woodward P (2011) Bayesian Analysis Made Simple: An Excel GUI for WinBUGS. CRC Press, Taylor & Francis Group, New York, USA.

18. USDA (1995) WEPP user summary. National Soil Erosion Research Laboratory (NSERL) W. Lafayette, IN, USA.

19. Bolstad WM (2010) Understanding Computational Bayesian Statistics. Wiley, New York, USA.

20. Ahmed I, Boutton TW, Karim A, Strom KB, Fox JF (2013) Monitoring soil organic carbon loss from erosion using stable isotopes. Proceedings of International Conference on Soil Carbon Sequestration for Climate, Food Security, and Ecosystem Services, May 26-29, Reykjavik, Iceland.

21. Irvin-Smith N, Chatman T, Ahmed I (2012) Rainfall Trend Assessment for the Buffalo Bayou Watershed. Proceedings of 10th TAMUS Pathways Symposium, November, Galveston, TX, USA.