WORLD ENVIRONMENTAL AND WATER RESOURCES CONGRESS 2015

FLOODS, DROUGHTS, AND ECOSYSTEMS

PROCEEDINGS OF THE 2015 WORLD ENVIRONMENTAL AND WATER RESOURCES CONGRESS

May 17-21, 2015
Austin, Texas

SPONSORED BY
Environmental and Water Resources Institute (EWRI)
of the American Society of Civil Engineers

EDITED BY
Karen Karvazy, P.E.
Veronica L. Webster, Ph.D.

Published by the American Society of Civil Engineers

Published by American Society of Civil Engineers
1801 Alexander Bell Drive
Reston, Virginia, 20191-4382
www.asce.org/publications | ascelibrary.org

Errata: Errata, if any, can be found at http://dx.doi.org/10.1061/9780784479162

ISBN 978-0-7844-7916-2 (PDF)
Manufactured in the United States of America.

Quantification of Uncertainty in Sediment Provenance Model

A. Karim, M.S.[1], I. Ahmed, Ph.D., P.E., CPESC[2], T.W. Boutton, Ph.D.[3], K.B. Strom, Ph.D.[4] and J.F. Fox, Ph.D.[5]

[1]Former Graduate Student, Department of Civil & Environmental Engineering, Prairie View A&M University, TX. (Currently Ph.D. Student, Texas Tech)
[2]Assistant Professor, Department of Civil & Environmental Engineering, Prairie View A&M University, PO Box 519, MS 2510, Prairie View, TX 77446; PH: (936) 261-1659; Fax: (936) 261-1662; email: ifahmed@earthlink.net
[3]Regents Professor and Texas AgriLife Senior Faculty Fellow, Department of Ecosystems Science & Mgmt., Texas A&M University, College Station, TX.
[4]Associate Professor, Department of Civil & Environmental Engineering, University of Houston, TX. (Currently Associate Professor,Virginia Tech)
[5]Associate Professor, Department of Civil & Environmental Engineering, University of Kentucky

ABSTRACT

We quantify and compare model uncertainty derived from sediment provenance or fingerprinting models using mathematical and statistical formulation rooted in the traditional Optimization and Bayesian Markov Chain Monte Carlo Simulation (MCMC) schemes. An ensemble prediction of soil yield percentage estimation from sub-watersheds of the 60 square-mile urbanized Buffalo Bayou Watershed of Houston was accomplished by forcing Markov chain rainfall time series windows generated by USDA's CLIGEN weather generator module of the Water Erosion Prediction Project (WEPP) software. In doing so, we also attempt to present a decision support tool that allows us to tell how much of a land area may be considered in simulation such that the model resolution definitively captures the contribution of soil from different sub-watershed sources. This was done by forcing the Bayesian model with varying lead time rainfall time series. Results for a given watershed contribution area shows that the model uncertainty remains constant after a certain lead time forecast. This allows the user to decide on how much land area to consider and when to stop the simulation on reasonable ground.

INTRODUCTION

Water flow induced soil erosion occurs from sources in a watershed during high rainfall events. The eroded soil transports to the watershed outlet. Ensemble predictions of soil yield fraction estimates from sub-watersheds are prone to uncertainty. The two sources of uncertainty investigated in this work are: (1) the physically based water erosion model uncertainty and (2) the uncertainty derived from the variable duration of rainfall time series. This paper focuses on the first. The dependency on frequentist approach is investigated and compared with a scientifically and statistically sound Bayesian model. The research framework consisted of computer modules creation with the aid of field data to validate those models. The research elements are listed in Figure 1.

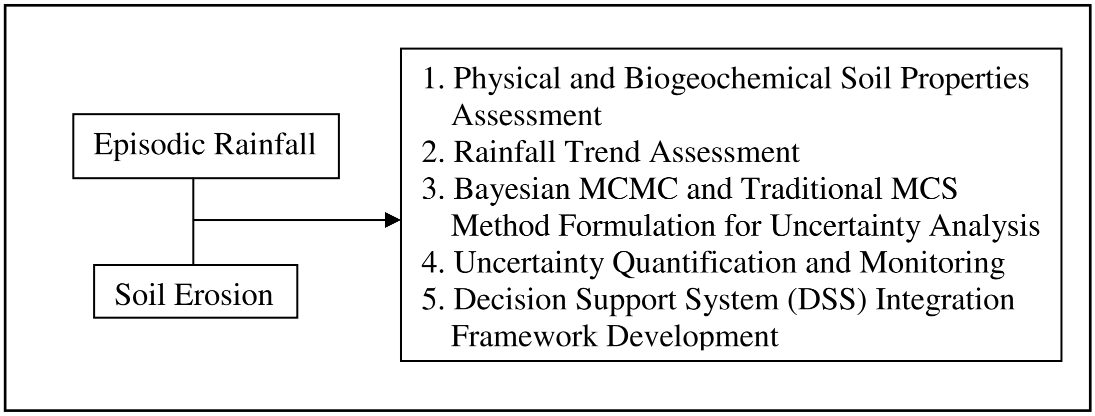

Figure 1. Research Elements

Specifically, the inter-disciplinary work attempted to find the effect of episodic rainfall on watershed-scale soil erosion yield fraction estimates from the North and the South of the Buffalo Bayou Harris County, Texas. This leads to statistically independent soil erosion model runs using WEPP with rainfall time series of various lengths to ascertain the effect of episodic nature of rainfall on the model prediction.

The work touches model parameter uncertainty associated with sediment fingerprinting technology in a Bayesian framework and compares that with tradition optimization based Monte Carlo Simulation (MCS). The motivational research hypothesis is outlined in the next section. The work builds on the noted recent work in this area by Fox and Papanicolaou (2008) where the authors attempted to identify eroded soil source from up to two land-use types by using Bayesian Markov Chain Monte Carlo (MCMC) simulation method using biogeochemical soil properties or tracers.

This study extends the analysis to include erosion model parameter uncertainty under episodic rainfall trends in the watershed. The Bayesian MCMC framework has proven to be a superior tool that is mathematically robust compared with the traditional MCS, and provides conservative response standard deviations (or uncertainty). Whereas, the MCS method is highly dependent on random selection of statistical properties (e.g., sample mean) and results in lower response standard deviations with "user-defined" increasing number of MCS iterations with no scientifically or statistically valid stopping criteria. The rainfall lead time based method in conjunction with the Bayesian model of Fox and Papanicolaou (2008) presented here has the advantage of informing the soil the user when to stop a model simulation run and how much land contribution area to work with for a given sediment source "un-mixing" model.

RESEARCH HYPOTHESIS

The variable nature of soil erosion, exacerbated by episodic rainfall trends on a large watershed, imposes a limit on model predictability as unavoidable errors in the initial state grow rapidly and render the model outputs useless. The most successful means of confronting this obstacle is to run a collection, or ensemble, of data or its statistics (mean and standard deviation), each starting from a slightly different initial state. The combined output can then be used to draw probabilistic inference about the prediction variable which in this study is the soil yield fractions from land sources.

A gap exists in the quantification and the comparison of estimated uncertainties from the Bayesian and the frequentist optimization methods or Monte Carlo Simulation. The study illustrates, through deterministic and stochastic model runs that the scatter in the deterministic soil yield data obtained from water erosion model of the USDA under various land slope conditions, lead to model uncertainty in the posterior distribution of the regression weights used in the stochastic Bayesian ensemble soil yield fraction prediction model. The regression weights are directly dependent on the statistically independent soil yield estimates from physical soil erosion model e.g., WEPP.

The study sought answers to the questions: (1) Is there a scientifically and statistically sound interdisciplinary approach utilizes the spatially variable soil biogeochemical properties to estimate soil yield fractions, and the corresponding uncertainty?; (2) If so, how is it superior to any frequentist approach that may require high resolution statistical definition of raw data which may not be available due to environmental and anthropogenic constraints?; and (3) What improvement is achieved to advance research in the subject?

RESEARCH FRAMEWORK

Figure 2 illustrates the concept of sediment fingerprinting [1]. The figure illustrates erosion occurance 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 (Phillips *et al*., 2000).

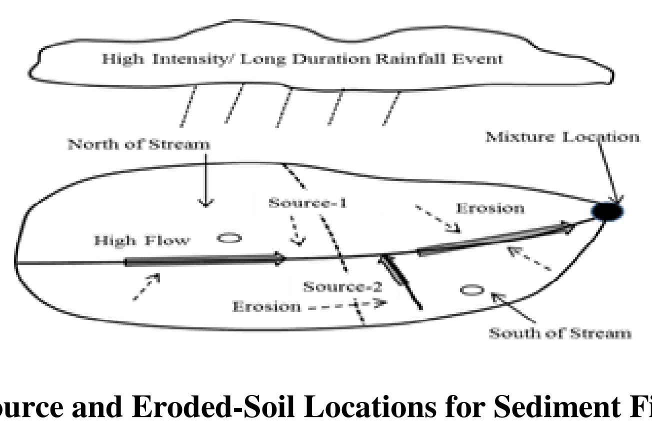

Figure 2. Source and Eroded-Soil Locations for Sediment Fingerprinting

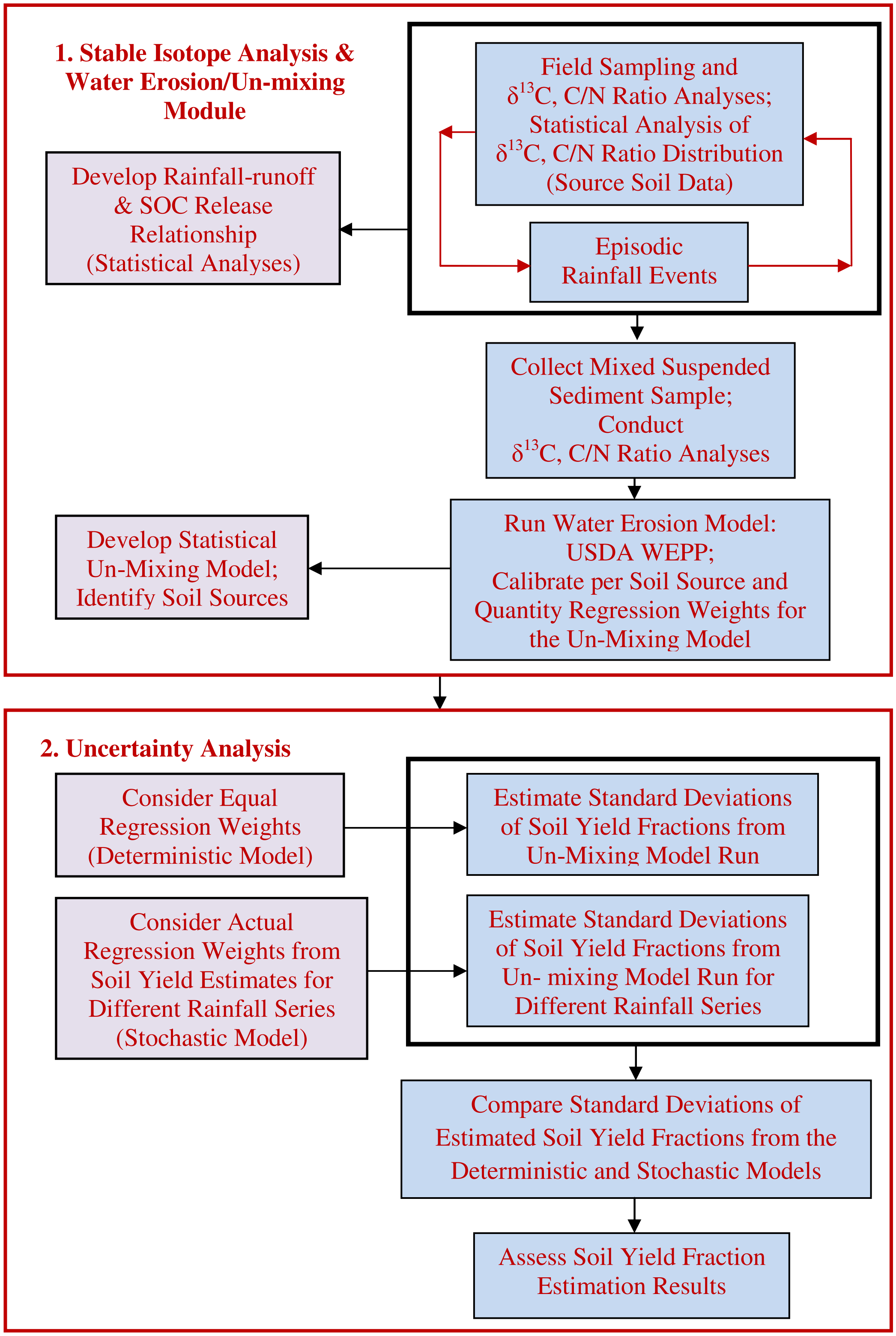

Figure 3. Research Framework (Karim, 2014; Ahmed *et al.*, 2011)

To apply fingerprinting method, two sources for tracers have been considered: (i) tracer data from land-use sources, and (ii) tracer data from eroded-soil. After analyzing the source- and eroded-soils for their tracer values, the contribution of eroded-soil from each source is estimated via Bayesian computational statistics.

Noteworthy, the decision on whether to use $\delta^{13}C$ or $\delta^{15}N$ in the statistical models depends on which of these two tracers show good correlation between the source and the mixture sample data. Denitrification can lead to enriched $\delta^{15}N$ in the mixture sample from riparian corridor compared with that from the source samples. In such cases, the correlation in isotope readings between the source and mixture sample location can be invalid. There was indication of such denitrification in this study that needs field work based validation that was beyond the scope of the study. $\delta^{13}C$ was therefore chosen.

The use of multiple tracers leads to an over-determined system of mass-balance equation matrix leading to multivariate distribution of the tracers. The problem can be conveniently solved using Bayesian Markov Chain Monte Carlo (MCMC) simulatin. Figure 3 illustrates the general integration framework. The efforts focused on the integration of Markov Chain rainfall series driven Water Erosion Prediction Project (WEPP) model (USDA, 1995) with its statistical relationship to spatial soil loss.

FORMULATION

Spatial variation of tracer data leads to model prediction uncertainty. Under the traditional nonlinear least square optimization method, due to the different ranges of the magnitudes of various tracer data, it is not statistically sound to conduct uncertainty analysis that considers simultaneous simulation of jointly distributed data. The "un-mixing" model for two sub-watershed sources and one tracer (e.g., $\delta^{13}C$) can be formulated as:

$$Z = X_1 P_1 + X_2 P_2$$
$$P_1 + P_2 = 1$$

where, Z represents mixture sample trace data, X stands for source sample mean, and P is the fraction contributed by the two 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) frequentist approach based on least square error minimization (Yu and Oldfeld, 1989; Collins *et al.*, 1997; Krause *et al.*, 2003), and (2) Bayesian Markov Chain Monte Carlo (MCMC) simulation approach (Fox and Papanicolaou, 2008).

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

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

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. The authors of [7] and [8] 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 sampling requires prior knowledge on distribution of sample data (Billheimer, 2001). It is not always possible to collect large number of samples from the field, especially in urban areas with access constraints. Therefore, it is assumed that the uncertainty in the population mean of each source property can be represented by Student's t-distribution with a confidence interval. To circumvent this constraint, Fox and Papanicolaou (2008) applied a Bayesian MCMC method with low informative prior by treating the parameters as random variables and then training the posterior distribution of all model parameters.

Bayesian MCMC framework not only allows multi-variate models but also facilitates the representation of more than one erosion process within the same sub-watershed or source by an erosion process parameter. 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 (Fox and Papanicolaou, 2008):

$$x^i_{jk} \sim \mathrm{MVN}_T\left[\mu_{jk}, \mathrm{COV}_{T\times T}\left(x_{jk}\right)\right]$$

In Bayesian statistics, the mean, μ and covariance matrix, COV(x) will have distribution of their own which are given multivariate normal (MVN) and Wishart distributions, respectively, to facilitate MCMC simulation using Gibbs sampling in WinBUGS (Ntzoufras, 2009):

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

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:

$$z_{mixture} \sim \mathrm{MVN}_T(\varphi, \Gamma)$$

with z the 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 (Fox and Papanicolaou, 2008):

$$\varphi = \textstyle\sum_k v_k P_k$$

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

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

SOURCE OF UNCERTAINTY

A set of *erosion process parameters*, α_{jk} considered by Fox and Papanicolaou (2008) are the weights of multiple linear regression equation for soil fraction yield. This weight, α_{jk}, is estimated by:

$$\alpha_{jk} = \frac{S_{jk}}{\Sigma_j S_{jk}}$$

where, *S* is the sediment yield. The erosion type identifier v_k is then $\alpha_{jk,}$ times x_{jk}. v_k is given a multivariate normal distribution and is a function of the *episodic erosion parameter, β* (Fox and Papanicolaou, 2008). In this study, the parameter of uncertainty is the *erosion process parameter, α*. 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 (Fox and Papanicolaou, 2008) that the entire watershed contributes to soil erosion and thus, a constant value of *β* is used in the un-mixing model. This significant assumption results in the same standard deviation and Monte Carlo (MC) errors for soil yield fractions from two sources (Fox and Papanicolaou, 2008).

The sediment yield, *S*, was estimated using physical process-based WEPP erosion model (USDA, 1995). 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 (Karim, 2014). This leads to uncertainty in soil fraction prediction that can only be tackled by coupling erosion prediction model with optimization routine which was beyond the scope of this work. Bayesian MCMC with Gibbs sampling (Bolstad, 2010) 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 (note the proportionality sign):

$$P\,(\textit{All model Parameters} \mid \mathrm{x_{jk}}, \mathrm{z}) \;\alpha\; P\,(\textit{All model parameters}) \times P\,(\mathrm{x_{jk}}, \mathrm{z} \mid \textit{All model parameters})$$

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

COMPARISON OF UNCERTAINTYFROM THE TWO MODELS

Land soil sampling on the urbanized Buffalo Bayou Watershed in Houston, Texas was designed to follow the drainage network managed by Harris County Flood Control District. Soil sampling was conducted in the vicinity of existing open or underground channels and roadside gutters. Where no such channel was found, soil samples were collected by the drainage gutters. Noteworthy, this leads to a sparse distribution of $\delta^{13}C$ and C/N ratio on the whole watershed but fairs well when viewed over sub-watershed scale, as validated using geostatistical analyses in ArcGIS (Karim, 2014).

The strategy was necessary to determine the percent soil erosion contribution from the sub-watersheds per Bayesian statistical "un-mixing" model algorithm of Fox and Papanicolaou (2008). In a nutshell, the project teams' interest has been to collect many samples in the whole watershed but not so many within a sub-watershed. Thus, the watershed-scale analysis led to *ensemble* soil loss predictions using 22 years of rainfall data based regression weights estimated from USDA Water Erosion Prediction Project (WEPP) model runs (Karim, 2014). No soil samples could be found in the neighborhoods with Storm Water Pollution Prevention Plans (SWPPP) in place. Figure 4 is a partial view of the watershed with land and Bayou sampling locations.

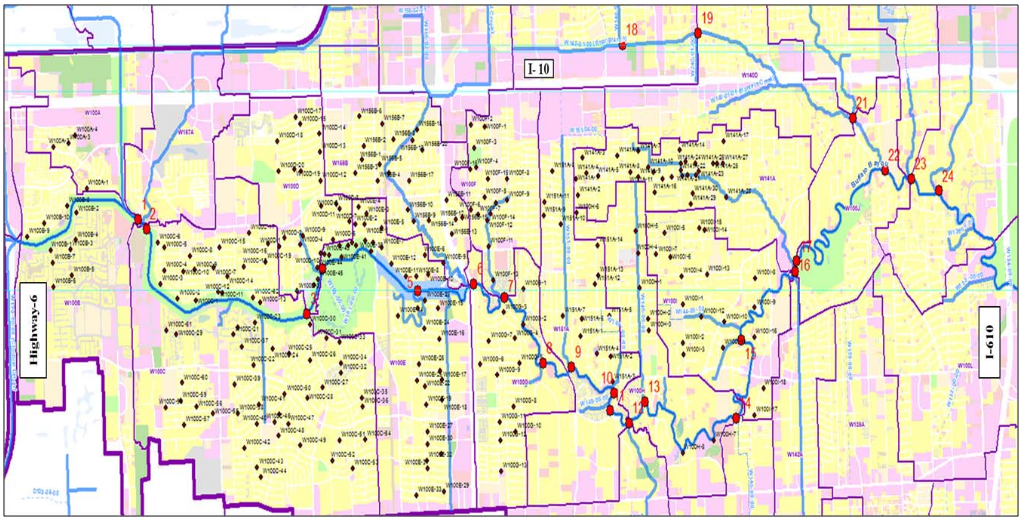

Figure 4. Partial Buffalo Bayou Watershed View with Land and Bayou Sampling Locations

Table 1. Standard Deviation of Soil Yield Fractions from Bayesian Simulation

Bayou Location	Upstream Land Sampling Points	6- & 11-year Rainfall Series	16-year Rainfall Series	22-year Rainfall Sereis
3	52	0.06386	0.07478	0.07478
4	79	0.05415	0.05415	0.05468
5	122	0.04828	0.04149	0.04364
6	148	0.04076	0.04076	0.04293

The output from the Bayesian model shows (Table 1) an increase in the standard deviation with increasing length of rainfall 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. The traditional Monte Carlo Simulation underestimates the uncertainty (Table 2) and the uncertainty decreases with number of iterations which is user-defined. There is no mathematical or statistical basis to know at which iteration the solution is acceptable.

Table 2. Standard Deviation of Soil Yield Fractions from Monte Carlo Simulation

Number of Iterations	Standard Deviation	Number of Iterations	Standard Deviation
100	0.033636	600	0.013760
200	0.025089	700	0.012845
300	0.019147	800	0.011430
400	0.015611	900	0.010985
500	0.015023	1000	0.010774

One other observation is in our ability to tell how much of a land area covered in the model results in what magnitude of model uncertainty. This is important because neither of the two methods have a memory to predict the contribution of upstream land area to soil yield. The use of variable lead time rainfall series sheds light on how the uncertainty (standard deviation) changes with larger spatial coverage. Table 1 shows the results for four different suspended sediment sample locations (3 to 6) on the Bayou. These location numbers increase in the downstream direction from west to east (Figure 4) with increasing land contribution areas. As is seen in Table 1 for Location 3 for example, the uncertainty estimate becomes constant at 0.07478 when 16- and 22-year long term rainfall time series were used. In this particular case, an 11-year rainfall time series would suffice to estimate the uncertainty. Bayesian approach allows us to make such prediction. The numbers in the tables are the standard deviations of the fractions of soil yield. They are to be multiplied by 100 to obtain the percent soil yield from a sub-source.

ACKNOWLEDGMENTS:

The research was funded by USDA/NIFA Award No. 2011-38821-30970.
The authors thank all local and federal government agencies and the citizens of the Buffalo Bayou Watershed for their support and patience during the course of this long term project development and implementation.

REFERENCES:

Ahmed, I., James, A.A., Boutton, T.W., and Strom, K.B. (2011). “Hydrologic Influences on Soil Organic Carbon Loss Monitoring Using Stable Isotopes,” Research Proposal to USDA/NIFA, Washington, D.C.

Billheimer, D. (2001). “Compositional receptor modeling,” *Environmetrics*, 12, pp. 451–467.

Bolstad, W.M. (2010). *Understanding Computational Bayesian Statistics*, Wiley, NY.

Collins, A.L., Walling, D.E., and Leeks, G.J.L. (1997). “Source type ascription for fluvial suspended sediment based on a quantitative composite fingerprinting technique,” *Catena*, 29, pp. 1–27.

Fox, J.F. and Papanicolaou, A.N. (2008). “An Un-mixing Model to Study Watershed Erosion Processes,” *Advances in Water Resources,* 31(1):96-108.

Franks, S. W. and Rowan, J. S. (2000). “Multi-parameter fingerprinting of sediment sources: Uncertainty estimation and tracer selection,” *Computational methods in water resources XIII*, Bentley et al., eds., Balkema, Rotterdam, pp.1067–1074.

Karim A. (2014). *Bayesian Ensemble Prediction of Watershed Scale Sediment Deliver*y, M.S. Thesis, Prairie View A&M University, Prairie View, TX.

Krause, A. K., Franks, S. W., Kalma, J. D., Loughran, R. J., and Rowan, J. S. (2003).“Multi-parameter fingerprinting of sediment deposition in a small gullied catchment in SE Australia,” *Catena*, 53, pp. 327–348.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*, Wiley, NY.

Phillips, J.M., Russell, M.A., and Walling D.E. (2000). “Time-integrated sampling of fluvial suspended sediment: a simple methodology for small catchments,” *Hydrol Process*, pp. 2589-2602.

USDA (1995). “WEPP User Summary,” National Soil Erosion Research Laboratory, W. Lafayette, IN.

Yu, L. and Oldfield, F. (1989). “A multivariate mixing model for identifying sediment source from magnetic measurements,” *Quater. Res.*, 32, pp. 168–181.