Distribution Agreement

Signature: Muqi Fan

Name: Muqi Fan Date: April 5th, 2023

An Analysis of Sampling Methods and Uncertainty Propagation for Shallow Water Modeling

by

Muqi Fan

Talea Mayo
Adviser

Mathematics

Talea Mayo
Adviser

Bree Ettinger
Committee Member

Sue Mialon
Committee Member

2023

An Analysis of Sampling Methods and Uncertainty Propagation for Shallow Water Modeling

by

Muqi Fan

Talea Mayo

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Mathematics

2023

Abstract

An Analysis of Sampling Methods and Uncertainty Propagation for Shallow Water Modeling
By Muqi Fan

Numerical models depend on inputs and parameters that are often uncertain. This causes uncertainty in the output. In this work, we explore both sampling and uncertainty propagation methods. We aim to assess which sampling methods best represent uncertainties in inputs and also explore which uncertainty propagation methods best depict the uncertainty that results in the output. We assess these methods using a two-dimensional shallow water model.

An Analysis of Sampling Methods and Uncertainty Propagation for Shallow Water Modeling

by

Muqi Fan

Talea Mayo
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Mathematics

2023

Acknowledgments

I would like to thank my advisor Talea Mayo for her abundant supports throughout the process of completing the thesis. I would also like to thank Professor Sue Mialon and Professor Bree Ettinger for being on my honor committee and provided valuable suggestions regarding my thesis.

Contents

1 Introduction

The field of mathematical modeling has transformed since the introduction of computers. In recent decades, computing has experienced technological advancements that have grown at an unimaginable rate, giving researchers the ability to gather and process large quantity of data like never before. It has also allowed researchers to increase sample sizes for statistical analyses, which would otherwise be limited, using simulation. With the ability to gather and process data of unimaginable quantities and at unimaginable speeds, the computational model, or modeling using computation and simulation, revolutionized traditional mathematical modeling methods. For each model, the most common concern is the accuracy of simulated outputs. Issues surrounding the accuracy of inputs, however, are often overlooked.

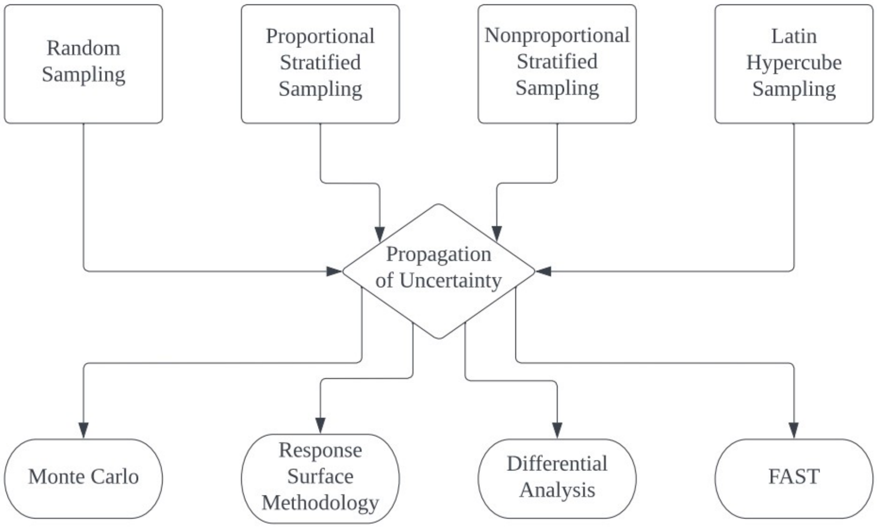
In most complex computational models, many parameters are involved in an effort to capture the main elements that play a role in the output. The parameters, however, usually contain certain degrees of uncertainties, such as the errors from measuring methods, errors due to limited sample sizes, and numerical errors. This leads to an issue: even if the model is able to predict the outputs accurately, the errors from the input parameters will still lead to errors in the output. Exploration of this issue leads to the analysis of the propagation of uncertainty, a crucial procedure needed for most computational models today, in order to assess the reliability of the modeling.

To describe the propagation of uncertainty in a brief notion, it is when the volatility in input parameters is passed on to the output results after being processed through a mathematical model. As the complexity of the model grows, it becomes more likely that a small uncertainty in the input will have some unpredictable impact on the results in the end, and thus such analysis is critical for almost all mathematical models.

The attempt to measure the mathematical error of a model has been one of the oldest topics in the history of mathematics, but the concern for the propagation of uncertainty was raised much more recently. One of the earliest prominent figures who raised the problem of propagation of error was Raymond T. Birge, who developed some of the earliest efficient and systematic methodologies measuring the propagation of error [1]. A few more papers regarding the issue were published after that, and propagation of uncertainty became one of the most important components in model development and validation today.

In this thesis, the variance will be the estimator measuring uncertainty, for it was one of the most commonly used parameters for the subject, suggested in JCGM [2]. A few sampling methods and propagation of uncertainty methods will be explored. Samples generated in each of the sampling methods will be applied to a shallow water model, and with the procedures measuring the propagation of uncertainty, we hope to quantify the propagation of uncertainty and decide which sampling method is most accurate in characterizing the propagation of uncertainty in the shallow water model in 2D.

Figure 1: Overview


Random Sampling
Proportional Stratified Sampling
Nonproportional Stratified Sampling
Latin Hypercube Sampling
Propagation of Uncertainty
Monte Carlo
Response Surface Methodology
Differential Analysis
FAST

2 Sampling Methodologies

With the law of large numbers, as long as our sample size is large enough, we can be infinitely close to truth. The problem is, when the collection of samples becomes too difficult and too expensive, we must explore other options. Therefore, prior to analyzing propagation of uncertainty, we will first explore some of the most popular sampling methods: Random Sampling, Stratified Sampling, and Latin Hypercube Sampling. We will then compare the advantages and disadvantages of each sampling method and discuss the ideal scenarios in which to use each.

2.1 Random Sampling

Let random variable X_r be a random variables with distributions D. The simple random sample suggest that if one want N samples from each random variables, one should select the samples based on some probabilistic function derived from D. The estimators for the mean and variance are:

$$E(X_r) = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

Where X_i denote the ith sample from X, and

$$Var(X_r) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}) \tag{2}$$

2.2 Stratified Sampling Method

The stratified sampling method suggest that we split the sample space into stratas, or intervals, and then draw set amount of random sample from each of the stratas. Let there be a random variable X_s, be a random variables with distributions D. Let the sample space of S be split into I disjoint stratas S_i. Two forms of stratified sampling can be conduct after the stratification: the proportional method and the nonproportional method. For the proportional case, the number of sample for each strata chosen must be based on the size of the strata in comparison to the size of the sample space, whereas for the nonproportional case there is no such restriction.

The reason behind the nonproportional sampling is when the given distribution is considered biased to begin with, especially when the extreme values are underrepresented. Sometimes approach of nonproportional sampling method such as the snowball sampling [3] is favored over the proportional approach. Consider the example with a normal distribution below:

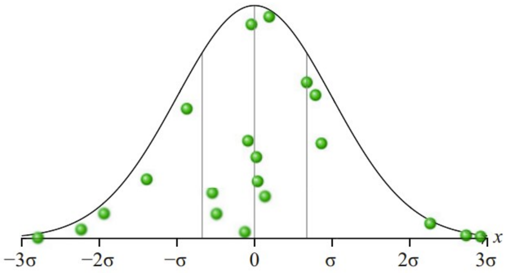


Figure 1: Unproportional Stratified Sampling with normal distribution split into four stratas and extract 5 data points from each strata. As the figure shows, such method allowed us to represent the data points from the the tails more. [4]

The mean and variance are derived as follows. let N_i be the number of elements in each stratum, and n_i be the number of elements chosen from N_i to be data point in each stratum. Then, let y_{ij} be the jth unit in ith stratum. Therefore the population mean can be expressed as [5]:

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{N_i} = \sum_{i=1}^{k} p_i \bar{y}_{N_i} \tag{3}$$

Where p_i represent the probability of a data point from ith stratum. As for the variance of the stratified sampling,

$$Var(\bar{y}_N) = Var(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i}) \tag{4}$$

since we assumed that samples are independent and identically distributed, we can establish that the variance of sum is the sum of variance:

$$Var(\bar{y}_N) = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 var(\bar{y}_{n_i}) \tag{5}$$

Then by the finite population of correction, we will rewrite the sample variance of the strata as proportion of the population variance of the strata:

$$Var(\bar{y}_N) = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 (\frac{N_i - n_i}{N_i}) \frac{var(\bar{y}_{N_i})}{n_i} \tag{6}$$

From the derivation by Mckay [6], suggested that the proportional stratified sampling variance can also be expressed as the following

$$Var(X_s) = Var(X_r) - \frac{1}{N} \sum_{i=1}^{N} p_i (\mu_i - \mu)^2 \tag{7}$$

Where $p_i = n_i/N$ represent proportional sampling and μ_i represent the sample average at corresponding strata.

2.3 Latin Hypercube Sampling

The Latin hypercube sampling suggest that instead of drawing some number of samples from potentially uneven stratas, one sample will be draw from each equally spaced stratas that covers the entire sample space. Let the input sample space be k-dimension such that there are k random variables X_1, X_2,...X_i...X_N. After dividing each of the X into equal-sized N stratas, we will then select the a random number from each of the stratas of different dimensions at random, and match them with one another to form a sample.

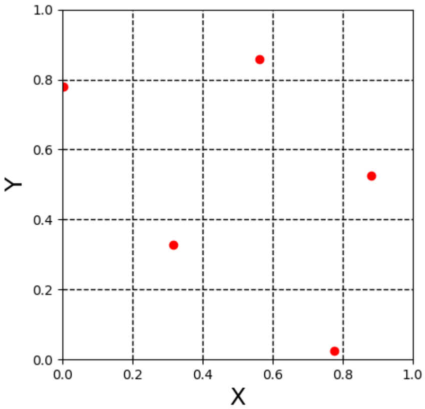


Figure 2: In a two variable LHS, after separating each vector of the sample space into equally spaced 4 stratas, we draw a random number for each strata for both vectors of the matrix, and the use the set of number to represent the actual sample to be drawn from the XY sample space. [7]

The LHS method is advantages when there is few number of input variables, as it represent each variable's distribution accurately, which is opposite toward the stratified method. The variance of the LHS is derived as following [6]:

$$Var(X_l) = Var(X_r) + \frac{N-1}{N} * \frac{1}{N^k(N-1)^k} \sum_{all i,j} (\mu_i - \mu)(\mu_j - \mu) \qquad (8)$$

2.4 Comparing Sampling Methods

By observing the relationship between variance estimations from each sampling method, we found out two points: One being that stratified sampling always have a lower variance than random sampling, and LHS have a lower variance than random sampling only if

$$\frac{N-1}{N} * \frac{1}{N^k(N-1)^k} \sum_{all i,j} (\mu_i - \mu)(\mu_j - \mu) \leq 0 \tag{9}$$

Which is true only when the covariance between cells with different coordinate are negative. This is true when the given function is monotonic. [6] Though simple random sampling was probably still the most commonly used method of sampling, its limitations was also obvious, as it usually has much higher variances and may be under-representative toward the extreme cases within a distribution. The stratified method was created so the extreme values won't be overlooked. As for the LHS, the method is favorable due to its lower computational cost, acknowledged by more and more researchers over the years [8]. Such feature is especially crucial for the Monte Carlo process done on a slower machines.

3 Procedures Measure Propagation of Uncertainty

The analysis of propagation of uncertainty is closely linked to the sensitivity analysis. The sensitivity analysis method generally falls into two categories: the local sensitivity test and the global sensitivity test. The prior examine the uncertainty in output by changing individual input parameters one at time, while the latter attempt to understand the sensitivity by examine the response of the output due to some total change of the input. There exists many procedures measuring how uncertainty propagate through a model. In this section, we will discuss some of the most popular methods used: Monte Carlo Analysis, Differential Analysis,Response Surface methodology, and Fourier Amplitude Sensitivity Test. For simplicity, we assume that the input variables are independently distributed.

3.1 Monte Carlo Analysis

One of the most popular methodology is the famous Monte Carlo Analysis. The general idea of this methodology is straightforward: first we will sample the input, or combination of inputs should the input parameters be plural. Afterward we will put the inputs through the given model, and thus obtain the corresponding output. By repeat the process countless times, we will obtain a sample for input population and a sample for output population. Through examine each of their variance, we will thus observe how uncertainty propagate through the model.

3.1.1 Monte Carlo Analysis

The Monte Carlo analysis is probably the most commonly used methodology in measuring the propagation of uncertainty when the computational cost is not a concern. The Monte Carlo Analysis procedures has following steps [9]:

1. Use the samples drawn to build probability distribution for every input variables
2. Obtain a of sample from every input variables
3. Put the set of input samples through the model
4. Repeat the step 2 and 3 numerous times, ideally larger than 10^6 iterations.

The real challenge of conducting a Monte Carlo efficiently comes from two aspect: the cost it takes to sample in a large scale, and the cost of running them through a model countless times. If the data collection is convenient and the choice of PDF is simple enough, we can use the direct sampling and generate samples from the inverse CDF. The exponential distribution, as an example, have a simple inverse CDF and thus easy to directly sample fromsince its PDF is

$$f(x) = \lambda * e^{-(\lambda)(x)} \tag{10}$$

And accordingly, the CDF become:

$$F(x) = \int_{x}^{0} \lambda * e^{-(\lambda)(x)} du = 1 - e^{-(\lambda)(x)} \tag{11}$$

Solve the inverse of the CDF $x = F^{-1}(y)$, we have

$$D(u) = \frac{1}{\lambda} ln(1-u) \tag{12}$$

Noted now u is can be generated as a random number with the range of [0,1], which is one of the most common function of many computational languages. Thus the sample data can be repeatedly generated from the original exponential PDF.

Such method, however, is not always available. For sometimes CDF of a distribution is not only very difficult to derive the inverse, but also very hard to draw sample from. One example being the inverse of Gaussian distribution. Though it is possible to obtain the inverse of the PDF as the following [10]:

$$f(x; m, \lambda) = \sqrt{\frac{\lambda}{2\pi\, x^3}} e^{-\frac{\lambda(x-\mu)^2}{2x\mu^2}} \tag{13}$$

Where m is the mean and λ is the shape parameter, it is extremely inefficient for computers to draw large amount of samples from such complex inverse CDF. Facing such challenge, two alternative approaches was introduced.

3.1.2 Rejection Sampling

The rejection sampling method suggest instead of trying to draw sample from the difficult probability distribution function f, we instead introduce another much simpler PDF g such that with some constant c [9]:

$$c = \sup_{allx} \frac{f(x)}{g(x)} < \infty \tag{14}$$

Where the support of f is contained within the support of g. After generating data through the inverse of G, the CDF of g just like what we discussed in the direct sampling method, we will then decide whether to accept or reject the data collected with the following procedure:

1. Generate uniform distribution $U\ [0, 1]$

2. Test if the generated data x is contained in both distributions, or

$$U \leq \frac{f(x)}{c * g(x)} \tag{15}$$

If true, accept the data, and if not, reject the data.

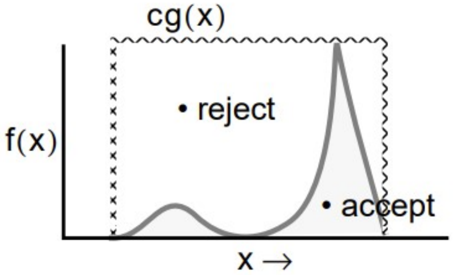


Figure 3: The uniform distribution represent cg(x), and samples drawn from such distribution is much easier than directly drawing from f(x) [9]

3.1.3 Importance Sampling

The method of importance sampling is very similar to rejection sampling, except of reject the data if it falls outside of the given distribution, a weight is assigned to each of the data drawn [9]. Let there be a population X with distribution of f, and let's say we are interested in calculating the mean of h(x), or

$$E_f(h(X))$$

by extract n samples. Instead of finding the mean directly, we introduce another distribution q(X) such that

$$E_f(h(X)) = E_g(\frac{f(X)}{g(X)}h(X)) \tag{16}$$

Expand the equation of mean:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} h(x_i) \tag{17}$$

With $\frac{f(x_i)}{g(x_i)}$ we can thus express the estimator for sample mean. The importance sampling has the advantage over rejection sampling since that instead of simply throw samples away when they are not contained in the distribution, at accept them after assigned some weight, thus require less computations.

3.2 Differential Analysis

The differential analysis started with an assumption: The mathematical model can be deduced to a first order Taylor series, taking the form as following:

$$f(X) = f(x_0) + \sum_{j=1}^{n} \frac{\partial f(x_0)}{\partial x_j} (x_j - x_{j0}) \tag{18}$$

Where X is the vector of inputs of $x_0, x_1, ...x_n$, and x_0 is the base vector selected. For convenience, we will let x_0 be the vector of means of each input. Once the above expansion is established, one can easily determine its estimators for the expected value and variance [11]:

$$E(f(x)) = f(x_0) + \sum_{j=1}^{n} \frac{\partial f(x_0)}{\partial x_j} E(x_j - x_{j0}) \tag{19}$$

But since $E(x_j - x_{j0}) = 0$,

$$E(f(x)) = f(x_0) \tag{20}$$

And as for variance or the uncertainty estimation:

$$Var(f(x)) = \sum_{j=1}^{n}(\frac{\partial f(x_0)}{\partial x_j})^2 Var(x_j) + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n}\frac{\partial f(x_0)}{\partial x_j}\frac{\partial f(x_0)}{\partial x_k}Cov(x_j, x_k) \tag{21}$$

With our assumption that the input variables are independent from each other,

$$Var(f(x)) = \sum_{j=1}^{n}(\frac{\partial f(x_0)}{\partial x_j})^2 Var(x_j) \tag{22}$$

Therefore, with knowledge of $Var(x_j)$, the uncertainty of each individual input variable, we will be able to estimate the variance of the output f(x). There is a major issue with this method, however.For some mathematical models is simply too complicated for us to assume any kind of linear relationship between input and output, it is then very difficult to obtain the accurate partial derivatives for the Taylor series. To solve such issue, I will use Regression Analysis in the next propagation of uncertainty method to determine the partial differentials.

3.3 Response Surface Methodology(RSM)

The method of RSM focus on examine how the output response to each individual input variables. The method is often used to solve optimization problem involving uncertainties, which is different than the traditional optimization problem that generally focus on one single, constant solution. For example, if one want to know what is the ideal output of CO2 for each car in a town that is best for both economy and environment, one would like to get a single constant answer. However, it would be impossible to maintain that standard with every car owner. Thus with RSM, we will obtain a range of ideal solutions that are considered acceptable. In our case, we will use method for a different purpose: to explore the propagation of uncertainty by breaking down the mathematical model into smaller pieces, such that every individual independent variable will produce certain degree of response for the output. Such method has another name called regression analysis. Consider the model as following [12]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_j X_n + \mu \tag{23}$$

Where there are $X_1, X_2, ...X_n$ independent variables, and each has m samples drawn. The μ stands for the error, and Y represent the vector of responses given each combination of input variables. It can be represented in the matrix form $Y = X\beta + \mu$ as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix} \tag{24}$$

And let $\hat{Y}$ be the estimated response called fitted value, such that $Y = \hat{Y} + \mu$. A few assumptions must be made for such regression:

1. The samples are chosen randomly
2. Zero Conditional Mean:
$$E(\mu|x) = 0$$
3. No perfect Collinearity, such that no x_i can be perfectly explained by some linear combination of some other x variable vectors.
4. As a result from (a) and (b), E(XU) = 0.

Let the sum of error squared be denoted as SSR. The βs in the regression is determined in least square method, such that SSR is minimized as follonwing:

$$min_\beta \sum_{i=1}^{n} \mu^2 = min_\beta(\boldsymbol{\mu}'\boldsymbol{\mu}) \tag{25}$$

Expanding $\boldsymbol{\mu}$ we then have

$$min_\beta(Y-\beta X)'(Y-\beta X)=Y'Y-2\beta X'Y+\beta' X'X\beta \tag{26}$$

Setting first order differential equal to 0 it becomes:

$$\partial_\beta SSR=-2X'Y+2X'X\beta=0 \tag{27}$$

Solving the system we will obtain the β estimators:

$$\hat{\beta}=(X'X)^{-1}X'Y \tag{28}$$

In certain cases it is possible that the relationship between input variable and output variable are not necessarily linear. In such case we can use polynomial regression to estimate the model:

$$Y=\beta_0+\beta_{11}X_1+\beta_{12}X_1^2+...+\beta_{1m}X_1^m+\beta_{21}X_2+...+\beta_{2m}X_2^m+...\beta_{n1}X_n+...\beta_{nm}X_n^m+\mu \tag{29}$$

With similar process of using using least square method from the linear case, we will obtain the estimators for βij by solve the system of first order condition equations:

$$\frac{\partial SSR}{\partial\beta_{ij}}=0 \tag{30}$$

for all i,j.

And accordingly the variance of Y to be

$$Var(Y) = \sum_{i=1}^{n}(\sum_{j=1}^{m}\beta ij^2 Var(X^j) + 2\sum_{m1}^{k=1}\sum_{m}^{j=k+1}\beta_{ik}\beta_{ij}Cov(X^k, X^j)) + Var(\mu) \tag{31}$$

As for the Variance and Covariance of powers of X:

$$Var(X_i^j) = E(X_i^{2j}) - E(X_i^j)^2 \tag{32}$$

$$Cov(X^k, X^j) = E(X^{k+j}) - E(X^k)E(X^j) \tag{33}$$

Which can be found using moment generating function

$$E(X_i^j) = \partial_t(M_x(0)) \tag{34}$$

For reference of the application later, the first four moment of normal distributions are [13]:

$$\mu$$

$$\mu^2 + \sigma^2$$

$$\mu^3 + 3\sigma^2$$

$$\mu^4 + 6\mu^3\sigma^2 + 15\mu\sigma^4$$

Theoretically, as the power of the polynomial grow, the less the sum of error squared will be, which is equivalent of saying that the variance of error will be smaller. Then by (31), we will exclude the variance of error term when we are trying to estimate the variance of the population output.

3.4 Fourier Amplitude Sensitivity Test(FAST)

The FAST is another widely used method in measuring the propagation uncertainty in many fields. The FAST method have similar idea approaching the measurement of uncertainty with the Differential Analysis and Response Surface methodology discussed before, but instead it tries to break down the variance into partial variances. The FAST method will first try to break down the output variance using Solbol's variance decomposition method, and then apply Fourier transformation and break down the total output variance into partial variances contributed by each individual input variables.

Consider the following model [14]:

$$y = g(x_1, x_2, ...x_n) \tag{35}$$

Where y is a model with n input parameters, and each input variable has distribution $f_i(x_i)$. Now consider the notion

$$V^{(x_i,x_j)} = V(E(y|x_i, x_j) \tag{36}$$

Be the joint partial variance of y that is due to uncertainty in x_i and x_j, and let $V_{(x_i,x_j)}$ Be the part of V(y) that result from the interaction of x_i and x_j. Assuming the input parameters are independent from one another, we can establish the Sobol's variance decomposition as following [15]:

$$V^{(x_i)} = V_{(x_i)} \tag{37}$$

$$V_{x_i,x_j} = V^{(x_i,x_j)} - V_{x_i} - V_{x_j} \tag{38}$$

......

$$V_{x_1,x_2,...x_n} = V^{(x_1,x_2,...x_n)} - \sum_{i=1}^{n} V_{x_i} - \sum_{i<j} V_{x_i,x_j} - \sum_{i<j<k} V_{x_i,x_j,x_k} - ..... \tag{39}$$

Noted how here $V^{(x_1,x_2,...x_n)}$ is equivalent to V(y), so reformulate the equation we have

$$V(y) = (\sum_{i=1}^{n} V_{x_i} + \sum_{i<j} V_{x_i,x_j} + \sum_{i<j<k} V_{x_i,x_j,x_k} + ...) + V_{x_1,x_2,...x_n} \tag{40}$$

Thus complete the decomposition of variance for y. Then, we will reconstruct y into a periodic search function

$$x_i = G(\theta_i) = F^{-1}(\frac{1}{2} + \frac{1}{\pi} arcsin(sin(\theta_i))) \tag{41}$$

Where θ_i is the random variable with uniform distribution from 0 to 2π, and F^{-1} is the inverse CDF of the x_i parameter Then the original model can be rewritten to

$$y = g(G(\theta_1), G(\theta_2), ...G(\theta_n)) \tag{42}$$

Note that the equation (40) still holds after the periodic transformation of the function [14]. After the transformation into the periodic parameters, We can apply multiple Fourier transformation on above equation:

$$g(G(\theta_1), G(\theta_2), ...G(\theta_n)) = \sum_{r_1, r_2, ...r_n = -\infty}^{\infty} C^{\theta}_{r_1, r_2, ...r_n} e^{r_1\theta_1 + r_2\theta_2 + ... + r_n\theta_n} \tag{43}$$

Where

$$C^{\theta}_{r_1, r_2, ...r_n} = (\frac{1}{2\pi})^n \int_0^{2\pi} ... \int_0^{2\pi} g(G(\theta_1), G(\theta_2), ...G(\theta_n)) e^{r_1\theta_1 + r_2\theta_2 + ... + r_n\theta_n} d\theta_1 d\theta_2 .....d\theta_n \tag{44}$$

This allow us to establish the fact that:

$$C^{\theta}_{r_1, r_2, ...r_n} = E(g(G(\theta_1), G(\theta_2), ...G(\theta_n)) e^{r_1\theta_1 + r_2\theta_2 + ... + r_n\theta_n}) \tag{45}$$

Which is

$$C^{\theta}_{r_1, r_2, ...r_n} = \frac{1}{N} \sum_{j=1}^{N} g(G(\theta_1^j), G(\theta_2^j), ...G(\theta_n^j)) e^{r_1\theta_1^j + r_2\theta_2^j + ... + r_n\theta_n^j} \tag{46}$$

Where the θ^j means the nth sample of chosen from the periodic space. With some prove we can also establish that the summation of Fourier amplitudes also estimate the corresponding partial variances:

$$V_{x_i} = \sum_{|r_i|=1}^{\infty} |C^{\theta}_{0,0,...,ri,...r_n}|^2 \tag{47}$$

$$V_{x_i x_j} = \sum_{|r_i|,|r_j|=1}^{\infty} |C^{\theta}_{0,0,\ldots,ri,\ldots rj,\ldots r_n}|^2 \tag{48}$$

......

$$V_{x_1,x_2,\ldots x_n} = \sum_{|r_1|,|r_2|,\ldots|r_n|=1}^{\infty} |C^{\theta}_{r1,r2,\ldots r_n}|^2 \tag{49}$$

Using the decomposition of variance formula, we can establish that

$$V(y) = V^{(x_1,x_2,\ldots x_n)} = \sum_{|r_1|,|r_2|,\ldots|r_n|=1}^{\infty} |C^{\theta}_{r1,r2,\ldots r_n}|^2 \tag{50}$$

With the detailed derivation of FAST see [14]

Thus complete the Fourier Amplitude Sensitivity Test and we have an estimate of the output variance based on Fourier Coeficients. In the application section, I will use SALib documentation's code in python to conduct FAST analysis.

4 Applications

4.1 Shallow Water Equation in 2D

The shallow water equation in 2D if a system of differential equations describing the behavior of fluid wave under certain impact. The system of equations is directly derived from conservation of mass and conservation of momentum [16].

$$\frac{\partial h}{\partial t} + \frac{\partial uh}{\partial x} + \frac{\partial vh}{\partial y} = 0$$

$$\frac{\partial uh}{\partial t} + \frac{\partial u^2h + \frac{1}{2}gh^2}{\partial x} + \frac{\partial ubh}{\partial y} = 0$$

$$\frac{\partial uh}{\partial t} + \frac{\partial u^2h + \frac{1}{2}gh^2}{\partial y} + \frac{\partial ubh}{\partial x} = 0$$

Where x and y represent coordinates, h(x,y,t) is the fluid height, u(x,y,t) and v(x,y,t) are velocity vectors, and g as the gravitational constant. We will rewrite the equation to be

$$\partial_t U + \partial_x F(U) + \partial_y G(U)$$

Where

$$U = \begin{pmatrix} h \\ uh \\ vh \end{pmatrix}$$

$$F(U) = \begin{pmatrix} uh \\ u^2h + \frac{1}{2}gh^2 \\ uvh \end{pmatrix}$$

$$G(U) = \begin{pmatrix} vh \\ uvh \\ v^2h + \frac{1}{2}gh^2 \end{pmatrix}$$

For simplicity, let the shallow water model have a initial condition of u = v = 0 at t = 0 for the entire surface, and with a reflexive boundary condition such that the waves bounce back after reaching the edge. Let there be certain initial impact landed at x_i,y_i with a radius of r.

To solve such problem we will use the finite difference method. We will first assign the water surface a mesh grid with equal intervals.

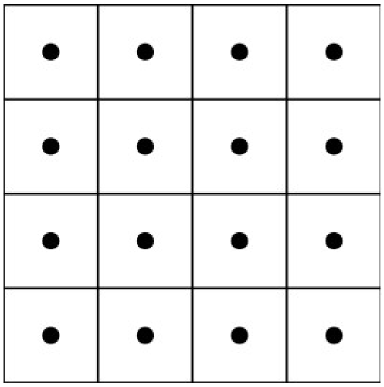


Figure 4: The stage 1 of the Runge Kutta process of the Shallow Water Model, where the center of each meshgrid intervals are known. [16]

Let the term $U_{i,j}^{n}$ denote the U vector at ith row, jth column, and nth time. The finite difference method consist of two steps: In the first step, we will use the Runge Kutta, or the shooting method to estimate what will happen at the midpoints on the mesh after $t = \frac{1}{2}$ has passed by solving the following equations:

$$U_{i+\frac{1}{2},j}^{n+\frac{1}{2}} = \frac{1}{2}(U_{i+1,j}^{n} + U_{i,j}^{n}) - \frac{\Delta t}{2\Delta x}(F_{i+1,j}^{n} - F_{i,j}^{n})$$

$$U_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}(U_{i,j+1}^{n} + U_{i,j}^{n}) - \frac{\Delta t}{2\Delta y}(G_{i,j+1}^{n} - G_{i,j}^{n})$$

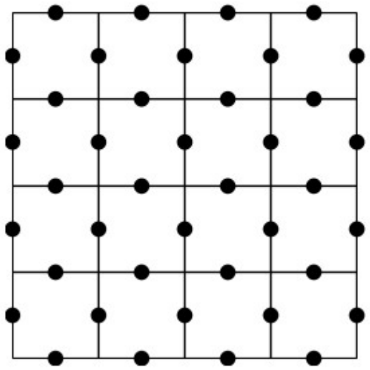


Figure 5: The stage 2 of the Runge Kutta process of the Shallow Water Model, where the information on the half points are directly derived from the information given in stage 1. [16]

And the second step will then compute the center of the cell at t = 1 using the values obtained from the step 1, using the following equation:

$$U_{i,j}^{n+1} = U_{i,j}^{n} - \frac{\Delta t}{\Delta x}(F_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{i,j-\frac{1}{2}}^{n+\frac{1}{2}}) - \frac{\Delta t}{\Delta y}(G_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - G_{i-\frac{1}{2},j}^{n+\frac{1}{2}})$$

Thus solves U at t=1. Repeat the method indefinite time will give the motion of the water surface over time.

Consider the following case: we want to know the height of the water level at some location and time, but we are uncertain about exactly where and when does the initial impact occur.Let such uncertainty be represented by x_i and y_i distributed normally with $\mu = 78$ and $\sigma = 2$, and r with normal distribution of $\mu = 3$ and $\sigma = 0.5$. Let the interested output be the height of the wave at $(x, y) = (60, 40)$ and $t = 30$. We will use the procedures of propagation of uncertainty to examine how the uncertainty in the initial condition will impact the height of the water surface at this location and time.

4.2 Results

In order to examine the propagation of uncertainty in the shallow water model problem, we will first draw a samples of $N = 10000$ for each of the input variables from the normal distributions mentioned above. This will serve as the population, and with Monte Carlo process we will obtain corresponding variance of the solutions, which will serve as the true value for the uncertainty of output. Then, we will draw a sample of $n = 600$ for each of the x,y,r using random, proportional stratified, nonproportional stratified, and Latin hypercube method. For the stratified sampling, we will have four equally spaced stratas across the range of the input variables. For the nonproportional sampling method, 10 samples will be redistributed from the each of the middle strata to the border stratas.

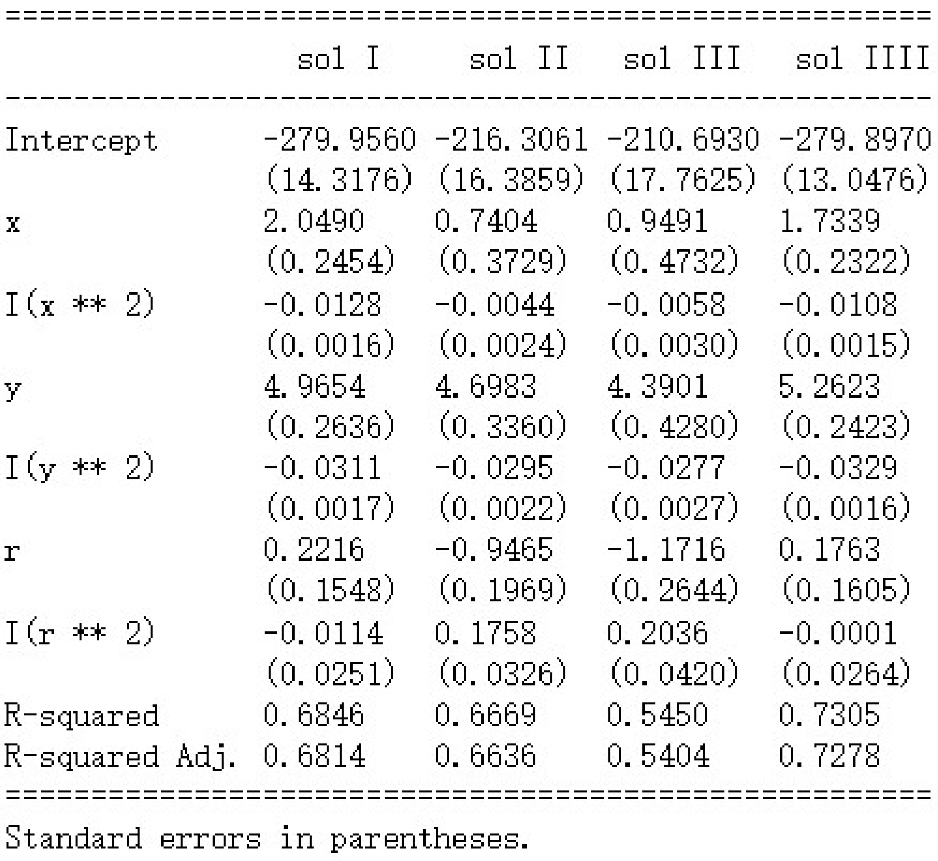
	V(x)	V(y)	V(r)
Random	4.0267	4.1399	0.2705
Proportional Stratified	3.8018	4.0704	0.2527
Nonproportional Stratified	4.8182	4.3020	0.2790
LHS	4.0128	4.0579	0.2461

Table 1: Measured Input Sample Variances

With population output generated by Monte Carlo over N=10000 iterations as our standard as truth, we found that the expected uncertainty of the output is about 0.163. Overall, the draws of x,y, and r from Random, proportional stratified, and LHS showed a similar variances toward the population variances. The nonproportional stratified method, on the other hand, showed a much higher variances in x and y.

4.2.1 Response Surface Method and Differential Analysis

Using the statsmodels module in python, we were able to do regression analyses on samples from all of four sampling methods.



```
=======================================================
                 sol I     sol II   sol III   sol IIII
-------------------------------------------------------
Intercept      -279.9560 -216.3061 -210.6930 -279.8970
               (14.3176) (16.3859) (17.7625) (13.0476)
x              2.0490    0.7404    0.9491    1.7339
               (0.2454)  (0.3729)  (0.4732)  (0.2322)
I(x ** 2)      -0.0128   -0.0044   -0.0058   -0.0108
               (0.0016)  (0.0024)  (0.0030)  (0.0015)
y              4.9654    4.6983    4.3901    5.2623
               (0.2636)  (0.3360)  (0.4280)  (0.2423)
I(y ** 2)      -0.0311   -0.0295   -0.0277   -0.0329
               (0.0017)  (0.0022)  (0.0027)  (0.0016)
r              0.2216    -0.9465   -1.1716   0.1763
               (0.1548)  (0.1969)  (0.2644)  (0.1605)
I(r ** 2)      -0.0114   0.1758    0.2036    -0.0001
               (0.0251)  (0.0326)  (0.0420)  (0.0264)
R-squared      0.6846    0.6669    0.5450    0.7305
R-squared Adj. 0.6814    0.6636    0.5404    0.7278
=======================================================
Standard errors in parentheses.
```

Figure 6: Response Surface Method Results

Where sol I,II,III,IIII each represent the results for Random, Proportional stratified, Nonproportional Stratified, and LHS sampling. The coeficients for the βs are the values without parentheses.

The regression results from different sampling methods showed a similar pattern, such that there is a strong linear relationship between the x,y and the output variables. The quadratic relationship, however, are much less significant. With the method of measuring the variance mentioned in the section 3.3 we thus obtained the output uncertainty:

	Var(U)
Random	0.1094
Proportional Stratified	0.1270
Nonproportional Stratified	0.1067
LHS	0.1216

Table 2: RSM Results

As expected the estimated variances are much lower than the true uncertainty since the variance of the error term is not accounted in.

4.2.2 The Differential Analysis

With the coefficients of regressions determined in the previous method, we will be able to construct the Taylor series representation of the model, and thus estimate the variances.

	Var(U)
Random	0.0675
Proportional Stratified	0.0827
Nonproportional Stratified	0.0330
LHS	0.0852

Table 3: Differential Analysis

4.2.3 The Fourier Amplitude Sensitivity test

With Fourier amplitude sensitivity test we found a common pattern such that the output variances caused by each individual input variable along are very small, and the variances resulted from the interaction between input variables played a much more significant role in determine the final uncertainty.

	S1	ST	S1_conf
x	0.0431	0.8668	0.0821
y	0.0454	0.8555	0.0669
r	0.0343	0.8775	0.0635

Table 4: Random Sampling

	S1	ST	S1_conf
x	0.0575	0.8726	0.0754
y	0.0208	0.4482	0.0776
r	0.0335	0.6725	0.0662

Table 5: Proportional Stratified Sampling

	S1	ST	S1_conf
x	0.0292	0.8870	0.0748
y	0.0125	0.3294	0.0628
r	0.0057	0.1919	0.0676

Table 6: Nonproportional Stratified Sampling

	S1	ST	S1_conf
x	0.0405	0.8472	0.0741
y	0.0101	0.7715	0.0776
r	0.0360	0.9057	0.0726

Table 7: Latin Hypercube Sampling

Where S1 represent the portion of V(U) that is caused by the input variable along, and ST represent the portion of the V(U) caused by the input variable as well as all of the interaction involves it. Using the formula (50) we will obtain the resulting output uncertainty for each sampling method in table below:

Table 8: FAST

	Var(U)
Random	0.5208
Proportional Stratified	1.3570
Nonproportional Stratified	1.3810
LHS	0.5474

4.3 Discussion

With the Monte Carlo simulation on the population data we obtained the supposed true variance of the output should be 0.1631, yet non of the measurements from the propagation procedures matched that. The RSM results revealed that there is a strong linear relationship between the output and x,y, but not as much with r, and not much of quadratic relations at all. Both the results from RSM and Differential Analysis showed a similar pattern: The LHS and the proportional stratified sampling resulted in a output uncertainty much closer than the true value of population V(U) compare to other two sampling methods. From results in Table 4 to 7, FAST revealed that the interaction between variables played a much bigger role on the V(U) than the variables just by themselves. The estimations of V(U) from FAST, however, are much conflicting with what we observed thus far, as all of the estimations are much larger than the expected truth.

4.4 Conclusion

Generally, the results from our methods did not meet our expectations. We found out that RSM method is best at characterizing the propagation of uncertainty, but as expected, its results shows lower value than the true variance of output due to the ignorance of the error variance. The differential analysis showed a less accurate prediction than the RSM method, but has shown a similar pattern. As for the Fourier Amplitude Sensitivity test, its results is highly erred form the expected output variances, but we can still observe how each input and their interactions influence the output variance. One possible explanation with the discrepancies showed in the differential analysis result may be the method used for measuring the individual partial differential's, which is a complex topic in the field of sensitivity analysis. The order of the polynomial regression may be still too low to truly capture the partial effect of each input toward the output by minimizing the SSR. The sample size of n=600 is also a source of error, for it may too small to capture the true propagation of uncertainty.

References

[1] R. T. Birge, The propagation of errors, American Journal of Physics 7 (6) (1939) 351–357.

[2] I. Iso, B. OIML, Guide to the expression of uncertainty in measurement, Geneva, Switzerland 122 (1995) 16–17.

[3] I. Etikan, R. Alkassim, S. Abubakar, Comparision of snowball sampling and sequential sampling technique, Biometrics and Biostatistics International Journal 3 (1) (2016) 55.

[4] R. G. McClarren, P. McClarren, Penrose, Uncertainty quantification and predictive computational science, Springer, 2018.

[5] P. V. Sukhatme, Sampling theory of surveys with applications, Journal of The Royal Statistical Society Series C-applied Statistics 3 (1954) 80–83.

[6] M. D. McKay, R. J. Beckman, W. J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 42 (1) (2000) 55–61.

[7] Latin hypercube sampling (lhs) (2017).
URL https://icme.hpc.msstate.edu/mediawiki/index.php/Latin_Hypercube_Sampling_%28LHS%29.html

[8] F. A. Viana, A tutorial on latin hypercube design of experiments, Quality and reliability engineering international 32 (5) (2016) 1975–1985.

[9] R. D. Peng, Advanced statistical computing, Work in progress (2018) 121.

[10] R. Chhikara, The inverse Gaussian distribution: theory: methodology, and applications, Vol. 95, CRC Press, 1988.

[11] J. C. Helton, F. J. Davis, Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliability Engineering & System Safety 81 (1) (2003) 23–69.

[12] J. M. Wooldridge, Introductory econometrics: A modern approach, Cengage learning, 2015.

[13] C. Clapham, J. Nicholson, J. R. Nicholson, The concise Oxford dictionary of mathematics, Oxford University Press, 2014.

[14] C. Xu, G. Gertner, Understanding and comparisons of different sampling approaches for the fourier amplitudes sensitivity test (fast), Computational statistics & data analysis 55 (1) (2011) 184–198.

[15] I. M. Sobol', On sensitivity estimation for nonlinear mathematical models, Matematicheskoe modelirovanie 2 (1) (1990) 112–118.

[16] C. B. Moler, Experiments with MATLAB, Society for Industrial and Applied Mathematics, 2011.