APPENDIX B – STATISTICAL ANALYSIS METHODS

TABLE OF CONTENTS

AppB-1 Distribution Function

The distribution function describes the probability of variate, X, taking values less than or equal to a number x. It is defined as (for continuous):

$$F(x) = P(X \le x) \equiv \int_{x_{\min}}^{x} P(x)dx \qquad \text{Eq - 1}$$

where *P* is probability density function (PDF) for the continuous distribution, or probability mass function (PMF) in discrete case; and F(x) is the probability distribution function. F(X) is sometimes called the cumulative density function(CDF).

Three major properties of distribution function above are:

1. F is non-decreasing, $F(x) \le F(y) \quad \forall x \le y$
2. F is right continuous, i.e., $F(x) \downarrow F(y), \forall x \downarrow y$
3. $F(x) \to 1$ and $F(y) \to 0$ as x$\to+\infty$ and y $\to-\infty$,respectively

PDFs may take a number of distributions, such as normal, poisson, uniform, exponential, gamma, and Chi-Squared distributions, just name a few. These distributions can be described by parametric functions.

The PDF of the normal distribution is:

Normal PDF $$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < +\infty \qquad \text{Eq - 2}$$

Where variate x varies from negative infinity to positive infinity,
μ is the location parameter estimated as the sample mean of x,
σ is the scale parameter estimated as sample standard deviation.

When μ = 0 and σ = 1, the distribution is called Standard Normal Distribution.

The normal PDF is most widely applied to describe real data distributions because of two mainly reasons:

- Repeating experiments and monitoring of regular processes generally have results in normal distributions;
- The normal distribution assumption leads to tests that are simple, mathematically tractable, and powerful in comparison with tests that do not take normality assumption.

In this study, Eq-2 is used to perform the normality test on all datasets in order to determine the statistic analysis approach.

AppB-2 Normality Test

The normality test is conducted first on field monitoring datasets to determine whether the datasets generally have normal distributions. Many statistical analysis approaches are established based on the normality test results. If datasets are confirmed to have

normal distribution, the parametric methods become applicable; otherwise the non-parametric methods should be applied for analysis.

Depending on the univariate or multivariate characteristics, methods of conducting normality tests are different. In this study, the univariate tests were conducted for load and flow. Symbol representing the normality is N(0,1) for univariate, and $N_P(\mathbf{0},\mathbf{I})$, where **bold** is vector or matrix.

A few methods can be applied to test the normality. Selection of the normality test results depends on the data characteristics.

AppB-2.1 Q-Q Plot (Visual Approach)

The principles of Q-Q plot are to compare the quantile ($Q_{sample,i}$) of the sample dataset with the standard normal distributed quantile (Q_{std}). When the sampled datasets are significantly similar to the normal distribution, i.e., Q_{sample} is closely around the distribution of Q_{std}, the sampled dataset is considered as normally distributed. Otherwise the sampled data would be skewed from normal distribution.

The calculation of sample quantiles is to rank the sample data ascending first, then calculate the ranked value of

$$Q_{sample,i} = (i - \frac{3}{8}) / (N + \frac{1}{4}) \qquad \text{Eq-3}$$

as quantile list (some literatures use ½ instead 3/8 and 0 other than ¼). The parameter 3/8 is included to make the correction of continuity.

Using the following steps to construct this Q-Q distribution test. Considering a dataset includes 19 data points as shown in the table below:

- Ordering the dataset from the smallest to the largest;
- Computing $Q_{sample,i}$ for $i = 1,\ldots,19$, then determining the X_i value so that the area under the bell-shape curve (Eq-2) in the region from $-\infty$ to X_i is equal to Q_{std}.

Data	Rank	Q	X-coordinate	Y-Coordinate	
				Reference Line	Circle Point
51.30	1	0.03	-1.85	52.90	51.30
56.30	2	0.08	-1.38	55.30	56.30
56.50	3	0.14	-1.10	56.70	56.50
57.30	4	0.19	-0.88	57.80	57.30
57.50	5	0.24	-0.71	58.70	57.50
59.00	6	0.29	-0.55	59.50	59.00
59.80	7	0.34	-0.40	60.30	59.80
62.50	8	0.40	-0.26	61.00	62.50
62.50	9	0.45	-0.13	61.69	62.50
62.80	10	0.50	0.00	62.30	62.80
63.50	11	0.55	0.13	63.00	63.50
64.30	12	0.60	0.26	63.70	64.30
64.80	13	0.66	0.40	64.40	64.80
65.30	14	0.71	0.55	65.11	65.30
66.50	15	66.00	0.71	66.00	66.50
66.50	16	66.90	0.88	66.90	66.50
67.00	17	68.00	1.10	68.00	67.00
69.00	18	69.40	1.38	69.40	69.00
72.00	19	71.80	1.85	71.80	72.00
62.34	Mean				
5.13	STD				

In order to compare the sample datasets with the normal distribution, we plot the continuous normal distribution curve. The abscissa value X is determined as:

$$X = CDF^{-1}(Q_i) \text{ and } CDF = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \qquad \text{Eq-4}$$

For any value of x_0 in $(-\infty, \infty)$, computing CDF to be equal to Q_i. The curve is plotted in the Q-Q plot (Figure 1) as a straight line.

From the Q-Q plot Figure 1, it is observed that the selected dataset ($i = 1, 2, \cdots, 19$) has nearly normal distribution.

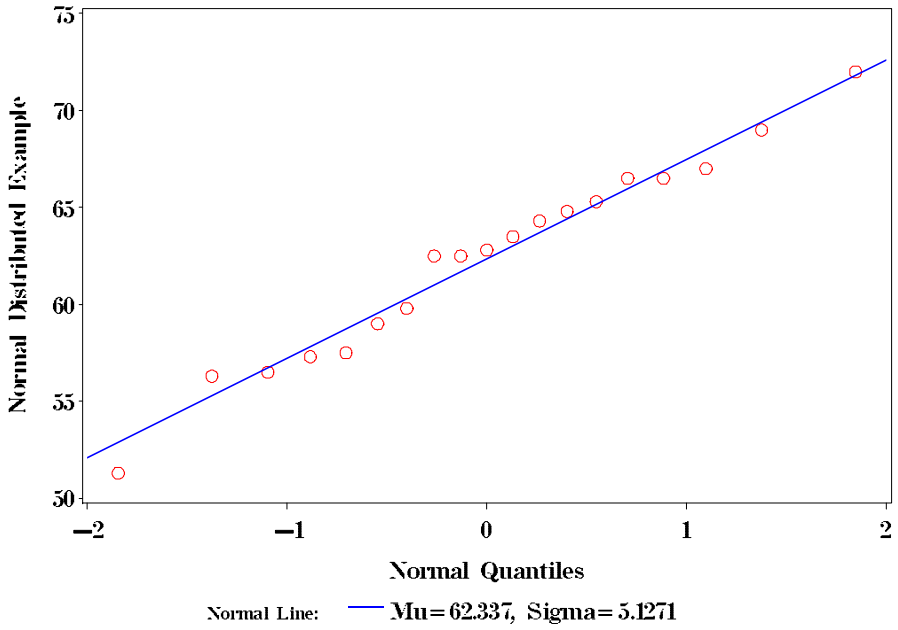

Figure 1. Sample data showing normal distribution

However, when considering the real farm flow and water quality data from this study, the Q-Q plots show that those datasets are distinct from the normal distributions. The flow, P load, P concentration and rainfall collected at Farm GISID 26-010-01 are analyzed in Q-Q plot.

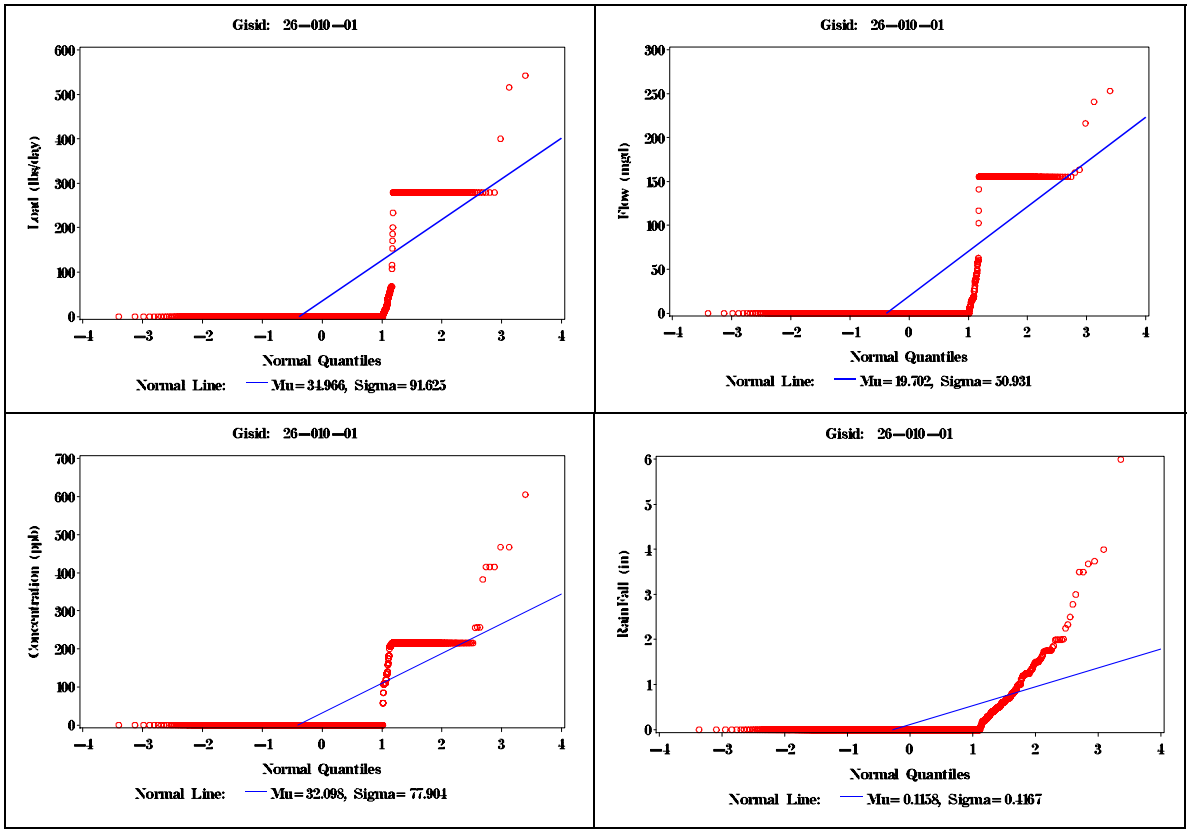

Figure 2. Real Flow, P load, P concentration and Rainfall data showing non-normal distribution

AppB-2.2 Skewness Indicator (Visual and Numerical)

Skewness measures the symmetry of the of sample data to the location of the maximum PDF value. Samples of normal distribution should be symmetric to the location of the mean, although the symmetry doesn't guarantee the normality. However, non-symmetry must be non-normal. (Kurtosis section will handles the symmetry, but non-normal case). Skewness calculation is shown as following:

$$Skewness = \frac{\sqrt{N}\sum_{i=1}^{N}\left(x'_i - \overline{x}'\right)^3}{\left[\sum_{i=1}^{N}\left(x' - \overline{x}'\right)^2\right]^{3/2}} \qquad \text{Eq - 5}$$

where, N = sample size
X_i =individual sample data
X_i-Bar =mean of sample set

Depending on the sample size, the standard critical skewness can be found from statistic tables. Generally, if the skewness is less than 0, we have negative skewness, otherwise we have positive skewness.

Kurtosis Indicator measures whether the sample data is peaked or flatted relative to a normal distribution. The over-peaked or over-flatted sample is not normal even it has symmetrical shape. The equation calculating the Kurtosis is as following:

$$Kurtosis = \frac{N\sum_{i=1}^{N}\left(x' - \overline{x}'\right)^4}{\left[\sum_{i=1}^{N}\left(x' - \overline{x}'\right)^2\right]^2} - 3 \qquad \text{Eq - 6}$$

Similar to Skewness, Kurtosis has its own critical value depending on sample size which can be found from statistic tables.

Using the same examples as shown above to explain how to apply the skewness indicator during the normality test.

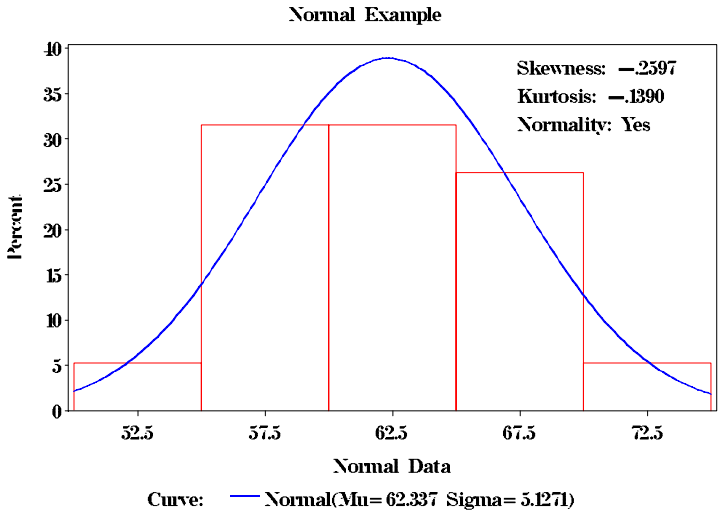

Figure 3. The skewness of the sample dataset showing nearly normal distribution

For the real data of flow, P load, P concentration and rainfall taken at GISID 26-010-01 Farm, Skewness and Kurtosis tests shows the non-normal distribution.

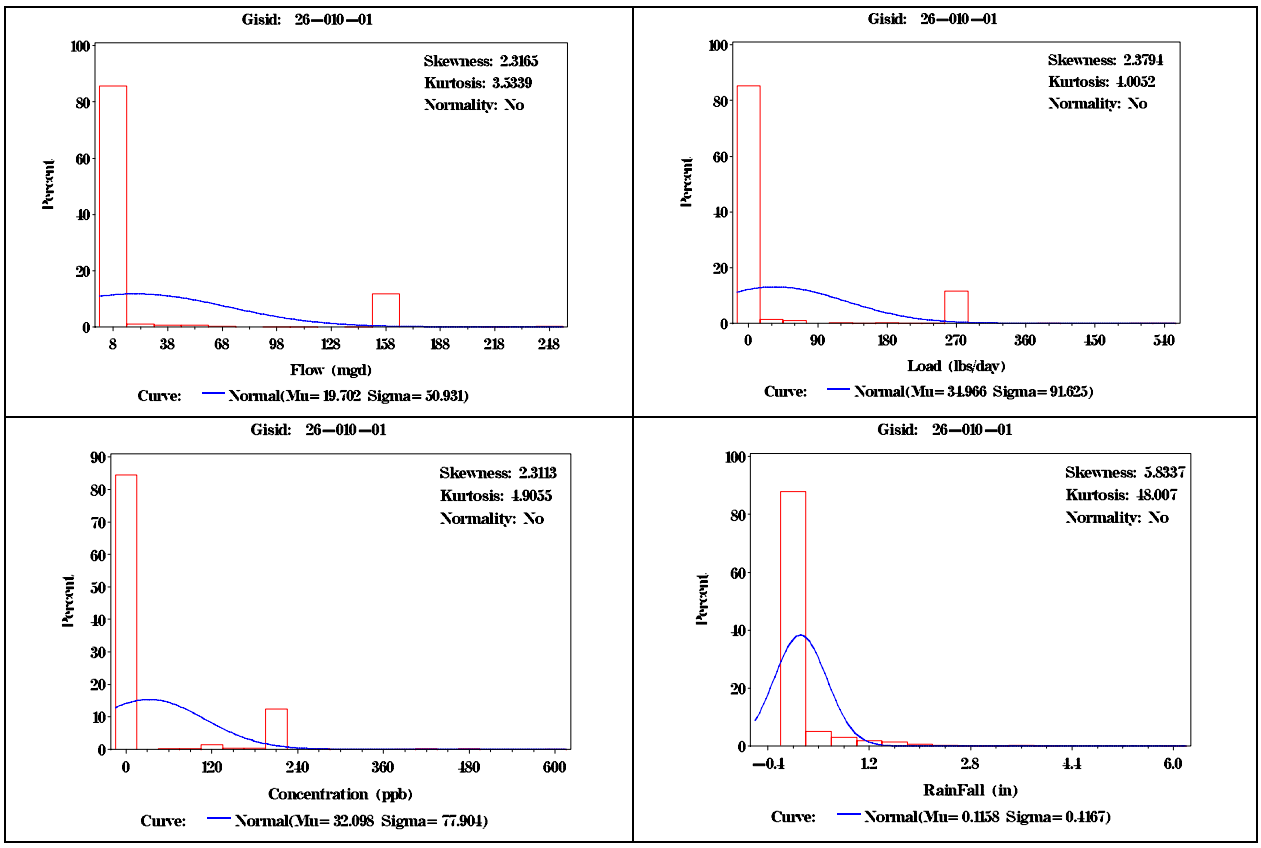

Figure 4. Real data from GISID 26-010-01 showing non-normal distribution

AppB-2.3 Komogorov-Smirnov Test (K-S Test, Numerical Method))

Komogorow-Smirnov test (K-S test) is applied to decide whether a sample dataset has its specific distribution. The K-S test is based on *empirical distribution function* (EDF). For an ascending dataset ($x_1,x_2,\ldots,x_n$), the EDF is defined as $F_n(x)=n(i)/N$, where n(i) is the number of data no more than x_i. Thus the $F_n(x)$ is a step function by 1/N increment.

To perform the K-S test on a given dataset, the compared cumulative distribution is calculated by using the normal PDF, which is

$$F(x)=\int_{-\infty}^{x}\frac{1}{s\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2}dx \qquad \text{Eq - 7}$$

For practical application, discrete summation is used other than integral to calculate F(x). From equations $F_n(x)$ and F(x), it is clear that both of them is less or equal to 1. K-S test is to calculate the maximum distance between EDF $F_n(x)$ assuming the normal F(x) for each pair, then compared with critical value which is sample size and alpha level predetermined related. The maximum distance equation is listed as following

$$\begin{aligned} D^+ &= \max\left(F_n(x_i)-F(x_i)\right) \\ D^- &= \max\left(F(x_i)-F_n(x_{i-1})\right) \\ D &= \max(D^+,D^-) \end{aligned} \qquad \text{Eq - 8}$$

where D^+ is the largest distance between EDF and the distribution function if EDF is greater than distribution function, D^- is the distance between EDF and the distribution function if EDF is less than distribution function.

H_0 The population has a normal distribution H_1: The population distribution does not have a normal distribution Reject H0 if the D value is greater than critical D value

Table 1. Example of this distance calculation

Rank	Data	Fn(x)	F(x)	D+	D-
1	51.3	0.053	0.016	0.037	
2	56.3	0.105	0.120		0.067
3	56.5	0.158	0.127	0.030	
4	57.3	0.211	0.163	0.048	
5	57.5	0.263	0.173	0.090	
6	59	0.316	0.258	0.058	
7	59.8	0.368	0.310	0.058	
8	62.5	0.421	0.513		0.144
9	62.5	0.474	0.513		0.092
10	62.8	0.526	0.536		0.062
11	63.5	0.579	0.590		0.063
12	64.3	0.632	0.649		0.070
13	64.8	0.684	0.685		0.053
14	65.3	0.737	0.718	0.018	
15	66.5	0.789	0.792		0.055
16	66.5	0.842	0.792	0.051	
17	67	0.895	0.818	0.076	
18	69	0.947	0.903	0.044	
19	72	1.000	0.970	0.030	
Mean	62.337		D	0.144	**Normality**
ndard Devia	5.127		D-critical	0.301	**YES**

The p-Value determining the normality can be found from the same table . The p value can be done by mathematical equations, but different equations are given from different literature. The equations used by SAS are not available to the description, and sometimes, the value calculated are different from each other, p-Value used in this document is cited from “D’Agostino, R.B., and Stephens, M. (1986) ‘Goodness of Fit’ New York Marcel Dekker”.

Table 2 shows the real datasets and their p-Values. It is seen all datasets used in the study do not have normal distribution.

Table 2. Farm Normality Assessment (K-S Test, All No-Normal)

GISID	Conc		Flow		Load	
	D_Value	pValue	D_Value	pValue	D_Value	pValue
26-003-01	0.438396	< 0.0100	0.347217	< 0.0100	0.444884	< 0.0100
26-004-01	0.508269	< 0.0100	0.361385	< 0.0100	0.476635	< 0.0100
26-006-01	0.340283	< 0.0100	0.396384	< 0.0100	0.334565	< 0.0100
26-009-01	0.331061	< 0.0100	0.385334	< 0.0100	0.386137	< 0.0100
26-010-01	0.504854	< 0.0100	0.495577	< 0.0100	0.493643	< 0.0100
26-010-02	0.3345	< 0.0100	0.332005	< 0.0100	0.353739	< 0.0100
50-002-01	0.467518	< 0.0100	0.44879	< 0.0100	0.440852	< 0.0100
50-002-02	0.452029	< 0.0100	0.431041	< 0.0100	0.410788	< 0.0100
50-005-05	0.42957	< 0.0100	0.482244	< 0.0100	0.421682	< 0.0100
50-006-01	0.491391	< 0.0100	0.489397	< 0.0100	0.467767	< 0.0100
50-007-02	0.439551	< 0.0100	0.425135	< 0.0100	0.403824	< 0.0100
50-007-03	0.282433	< 0.0100	0.299085	< 0.0100	0.328418	< 0.0100
50-008-01	0.442957	< 0.0100	0.438783	< 0.0100	0.4134	< 0.0100
50-010-03	0.391184	< 0.0100	0.428259	< 0.0100	0.393111	< 0.0100
50-010-05	0.450381	< 0.0100	0.433436	< 0.0100	0.398709	< 0.0100
50-011-02	0.244825	< 0.0100	0.290523	< 0.0100	0.332306	< 0.0100
50-011-05	0.309572	< 0.0100	0.233738	< 0.0100	0.305597	< 0.0100
50-011-06	0.368237	< 0.0100	0.356898	< 0.0100	0.385298	< 0.0100
50-013-01	0.411698	< 0.0100	0.433049	< 0.0100	0.38076	< 0.0100
50-015-01	0.443112	< 0.0100	0.419815	< 0.0100	0.391746	< 0.0100
50-015-02	0.428657	< 0.0100	0.407318	< 0.0100	0.387737	< 0.0100
50-016-01	0.47157	< 0.0100	0.448807	< 0.0100	0.429465	< 0.0100
50-018-01	0.450831	< 0.0100	0.421395	< 0.0100	0.411347	< 0.0100
50-018-02	0.418328	< 0.0100	0.393268	< 0.0100	0.360252	< 0.0100
50-018-03	0.383015	< 0.0100	0.374494	< 0.0100	0.344987	< 0.0100
50-018-04	0.463637	< 0.0100	0.4555	< 0.0100	0.435273	< 0.0100
50-018-05	0.449769	< 0.0100	0.44446	< 0.0100	0.427886	< 0.0100
50-018-06	0.424574	< 0.0100	0.452052	< 0.0100	0.399416	< 0.0100
50-018-07	0.391258	< 0.0100	0.382169	< 0.0100	0.392116	< 0.0100
50-018-08	0.370162	< 0.0100	0.372445	< 0.0100	0.365644	< 0.0100
50-018-09	0.467226	< 0.0100	0.458057	< 0.0100	0.43165	< 0.0100
50-018-12	0.41553	< 0.0100	0.428207	< 0.0100	0.391221	< 0.0100
50-018-13	0.41656	< 0.0100	0.416843	< 0.0100	0.371436	< 0.0100
50-018-21	0.204924	< 0.0100	0.199836	< 0.0100	0.304501	< 0.0100
50-018-22	0.353467	< 0.0100	0.360207	< 0.0100	0.372296	< 0.0100
50-018-23	0.370174	< 0.0100	0.392818	< 0.0100	0.406821	< 0.0100
50-018-24	0.42884	< 0.0100	0.423616	< 0.0100	0.399774	< 0.0100
50-025-01	0.428985	< 0.0100	0.480918	< 0.0100	0.443086	< 0.0100
50-026-01	0.472401	< 0.0100	0.473422	< 0.0100	0.437915	< 0.0100
50-033-01	0.405692	< 0.0100	0.343495	< 0.0100	0.362489	< 0.0100
50-035-03	0.260679	< 0.0100	0.262118	< 0.0100	0.275723	< 0.0100
50-036-01	0.415407	< 0.0100	0.470385	< 0.0100	0.41684	< 0.0100
50-037-01	0.370456	< 0.0100	0.378078	< 0.0100	0.349444	< 0.0100
50-038-01	0.474714	< 0.0100	0.464417	< 0.0100	0.43022	< 0.0100
50-040-01	0.464112	< 0.0100	0.4684	< 0.0100	0.43003	< 0.0100
50-040-02	0.494017	< 0.0100	0.48025	< 0.0100	0.45028	< 0.0100
50-043-01	0.384963	< 0.0100	0.398061	< 0.0100	0.413708	< 0.0100
50-044-01	0.499366	< 0.0100	0.472277	< 0.0100	0.461558	< 0.0100

Table 2. Farm Normality Assessment (K-S Test, All No-Normal) (Cont.)

GISID	Conc		Flow		Load	
	D_Value	pValue	D_Value	pValue	D_Value	pValue
50-047-06	0.457325	< 0.0100	0.461547	< 0.0100	0.439382	< 0.0100
50-047-07	0.42612	< 0.0100	0.431429	< 0.0100	0.408624	< 0.0100
50-052-01	0.440746	< 0.0100	0.418357	< 0.0100	0.388873	< 0.0100
50-054-01	0.376682	< 0.0100	0.388613	< 0.0100	0.351907	< 0.0100
50-054-02	0.484409	< 0.0100	0.485739	< 0.0100	0.466015	< 0.0100
50-054-03	0.474887	< 0.0100	0.466089	< 0.0100	0.43635	< 0.0100
50-054-04	0.375756	< 0.0100	0.387861	< 0.0100	0.357141	< 0.0100
50-054-05	0.426276	< 0.0100	0.440092	< 0.0100	0.408876	< 0.0100
50-057-01	0.460995	< 0.0100	0.489262	< 0.0100	0.450145	< 0.0100
50-059-01	0.2713	< 0.0100	0.308673	< 0.0100	0.33464	< 0.0100
50-059-02	0.444504	< 0.0100	0.426079	< 0.0100	0.38856	< 0.0100
50-059-03	0.470276	< 0.0100	0.475539	< 0.0100	0.444916	< 0.0100
50-059-04	0.515801	< 0.0100	0.507382	< 0.0100	0.498158	< 0.0100
50-061-08	0.482918	< 0.0100	0.518113	< 0.0100	0.472609	< 0.0100
50-061-09	0.457337	< 0.0100	0.452245	< 0.0100	0.433113	< 0.0100
50-061-10	0.34351	< 0.0100	0.345761	< 0.0100	0.32836	< 0.0100
50-061-14	0.45721	< 0.0100	0.467651	< 0.0100	0.445509	< 0.0100
50-061-15	0.464115	< 0.0100	0.468564	< 0.0100	0.434893	< 0.0100
50-061-16	0.481777	< 0.0100	0.494129	< 0.0100	0.466739	< 0.0100
50-061-17	0.413287	< 0.0100	0.400841	< 0.0100	0.371598	< 0.0100
50-064-01	0.228453	< 0.0100	0.289675	< 0.0100	0.330838	< 0.0100
50-064-02	0.244825	< 0.0100	0.290523	< 0.0100	0.332306	< 0.0100
50-064-03	0.228453	< 0.0100	0.289675	< 0.0100	0.330838	< 0.0100
50-064-04	0.228453	< 0.0100	0.289675	< 0.0100	0.330838	< 0.0100
50-065-01	0.305414	< 0.0100	0.357557	< 0.0100	0.336963	< 0.0100
50-065-03	0.228453	< 0.0100	0.289675	< 0.0100	0.330838	< 0.0100
50-065-04	0.244825	< 0.0100	0.290523	< 0.0100	0.332306	< 0.0100
50-065-05	0.238401	< 0.0100	0.293438	< 0.0100	0.329079	< 0.0100
50-065-06	0.238401	< 0.0100	0.293438	< 0.0100	0.329079	< 0.0100
50-065-08	0.228453	< 0.0100	0.289675	< 0.0100	0.330838	< 0.0100
50-065-09	0.244825	< 0.0100	0.290523	< 0.0100	0.332306	< 0.0100
50-067-01	0.463778	< 0.0100	0.467124	< 0.0100	0.442085	< 0.0100
50-067-02	0.37634	< 0.0100	0.39645	< 0.0100	0.353065	< 0.0100
50-067-03	0.368829	< 0.0100	0.39344	< 0.0100	0.353797	< 0.0100
50-067-04	0.331657	< 0.0100	0.342689	< 0.0100	0.313383	< 0.0100
50-067-05	0.365955	< 0.0100	0.358947	< 0.0100	0.318948	< 0.0100
50-067-06	0.303458	< 0.0100	0.301304	< 0.0100	0.337448	< 0.0100
50-067-07	0.343328	< 0.0100	0.259883	< 0.0100	0.338168	< 0.0100
50-067-08	0.316251	< 0.0100	0.36617	< 0.0100	0.321811	< 0.0100
50-067-09	0.435053	< 0.0100	0.325613	< 0.0100	0.367022	< 0.0100
50-067-10	0.338824	< 0.0100	0.311046	< 0.0100	0.350523	< 0.0100
50-067-11	0.330756	< 0.0100	0.273756	< 0.0100	0.339324	< 0.0100
50-067-12	0.330379	< 0.0100	0.219306	< 0.0100	0.315204	< 0.0100
50-067-13	0.446233	< 0.0100	0.410457	< 0.0100	0.401041	< 0.0100
50-068-01	0.334349	< 0.0100	0.353559	< 0.0100	0.347378	< 0.0100
50-068-02	0.45808	< 0.0100	0.450243	< 0.0100	0.421558	< 0.0100
50-071-01	0.382505	< 0.0100	0.426967	< 0.0100	0.371856	< 0.0100

AppB-2.4 Shapiro-Wilk Test (S-W Test applicable only for data size less than 2000, Numerical Method)

S-W test calculates a W statistic that a datset x1,x2,…xn comes (especially) from a normal population distribution. Small value of W is a evidence of departure from normal distribution. Basically, W statistic is proportional to the ratio of two estimates of variance. The numerator is the square of linear estimate of sigma based on order statistics. The denominator is the usual sum of square. The numerator coefficient comes from statistic table even though the analytical approximation can be applied. Equation of W statistic is as following

$$W = \frac{\left[\sum_{j=1}^{k} a_j \left(x_{n-j+1} - x_j\right)\right]^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad \text{Eq - 3}$$

$k = \frac{n}{2}$ if n is even

$k = \frac{n-1}{2}$ if n is odd

a_i *comes from statistic table*

H_0: The population has a normal distribution
H_1: The population does not have a normal distribution
Reject H0 if $W<W_{criticl,0.05}$

Table 3. Example dataset shows normal distribution

Rank	Data	(Data-DataBar)^2	a_j	$a_j * (x_{n-j+1}-x_j)$
1	51.3	121.8118837	0.4808	9.95256
2	56.3	36.4434626	0.3232	4.10464
3	56.5	34.06872576	0.2561	2.68905
4	57.3	25.36977839	0.2059	1.89428
5	57.5	23.39504155	0.1641	1.4769
6	59	11.13451524	0.1271	0.80073
7	59.8	6.435567867	0.0932	0.466
8	62.5	0.026620499	0.0612	0.11016
9	62.5	0.026620499	0.0303	0.0303
10	62.8	0.214515235		
11	63.5	1.352936288		
12	64.3	3.85398892		
13	64.8	6.067146814		
14	65.3	8.780304709		
15	66.5	17.33188366		
16	66.5	17.33188366		
17	67	21.74504155		
18	69	44.39767313		
19	72	93.3766205		
Mean	62.33684			
Denominator	473.1642			
Numerator	463.3093			**Normality**
W-statistic	0.979172	W-Critical	0.901	**YES**

Table 4. Farm Normality Assessment (S-W Test, All No-Normal)

GISID	Conc		Flow		Load	
	W_Value	pValue	W_Value	pValue	W_Value	pValue
26-003-01	0.548175	<0.0001	0.701028	0.0000	0.498223	<0.0001
26-004-01	0.379128	<0.0001	0.590816	0.0000	0.279789	<0.0001
26-010-01	0.443795	<0.0001	0.409052	0.0000	0.404115	<0.0001
50-006-01	0.465709	<0.0001	0.463816	0.0000	0.390012	<0.0001
50-007-03	0.691487	<0.0001	0.739913	0.0000	0.474336	<0.0001
50-011-02	0.774819	<0.0001	0.610196	0.0000	0.46638	<0.0001
50-011-05	0.412107	<0.0001	0.808979	0.0000	0.550855	<0.0001
50-026-01	0.345269	<0.0001	0.349702	0.0000	0.220153	<0.0001
50-035-03	0.621521	<0.0001	0.717306	0.0000	0.645093	<0.0001
50-036-01	0.216046	<0.0001	0.457531	0.0000	0.211036	<0.0001
50-040-02	0.377228	<0.0001	0.309393	0.0000	0.199975	<0.0001
50-043-01	0.446932	<0.0001	0.514799	0.0000	0.204237	<0.0001
50-044-01	0.42023	<0.0001	0.332233	0.0000	0.291438	<0.0001
50-047-06	0.395027	<0.0001	0.409633	0.0000	0.328996	<0.0001
50-052-01	0.561498	<0.0001	0.503814	0.0000	0.398493	<0.0001
50-057-01	0.307035	<0.0001	0.404539	0.0000	0.270356	<0.0001
50-061-14	0.400096	<0.0001	0.440516	0.0000	0.363736	<0.0001
50-064-02	0.774819	<0.0001	0.610196	0.0000	0.46638	<0.0001
50-065-01	0.606287	<0.0001	0.737593	0.0000	0.447278	<0.0001
50-065-04	0.774819	<0.0001	0.610196	0.0000	0.46638	<0.0001
50-065-09	0.774819	<0.0001	0.610196	0.0000	0.46638	<0.0001
50-067-12	0.646122	<0.0001	0.792967	0.0000	0.521955	<0.0001
50-068-02	0.478372	<0.0001	0.456368	0.0000	0.352457	<0.0001
50-071-01	0.399139	<0.0001	0.534758	0.0000	0.355246	<0.0001

AppB-3 Nonparametric Test

Nonparametric tests are often used when certain assumptions underlying of population are questionable. The well-developed ANOVA comparing two samples are basically based on normal distribution assumption. However, the data sample not taking this feature cannot be analyzed by ANOVA.

In this study, all normality tests farm data are non-parametric. Like K-S test described above, it is one kind non-parametric technology. Typically used nonparametric methods are Wilcoxon Mann Whitney test, Wiocxon Signed Rank test, and Sign test. The reason that parametric is called is because it requires estimation of the population parameters such as mean or standard deviation. Nonparametric methods don't need this estimation It should be pointed out that the exact definition of nonparametric varies from literature to literature.

Wilcoxon M-W test (also called Wilcoxon Rank Sum test or Mann Whitney U test) is used to compare whether or not two samples are drawn from the same populations, or the population are not different and shift to each other. Wilcoxon Sign Rank test is used to test the population location (median), and often involve paired data sets. It is called Sign is because the first step handling the data is rank them assuming no sign, and then give the original sing on it. For example, the data we have is –1 ,-8, –3, 5, 4, and 7, the first step is rank them as 1,3,4,5,7, 8 then add the sign to produced the ranked data as –1, -3, 4 , 5, 7, -8.

The Wilcoxon M-W treats the data as –8, -3, -1, 4, 5, 7, in other words, the data is ranked by themselves. Wilcoxon Sign Rank test is often to used to test whether data's median is 0. Sign test is similar with Wilcoxon Singed rank. In our application, we used Wilcoxon M-W test because the data in comparing might not be paired which implicates the data have the same size, and our goal is to compare the two population are the same or not.

To use Wilcoxon W-M, the first thing is to rank the all the data (n1+n2), and assign the rank 1 to the smallest, 2 to the second smallest, and so on so forth, till n1+n2 to the largest. Tied observations are assign the averaged rank, for example, if we have the second data, and the third data tired, we assign 2.5 to each of them [(2+3)/2]. Then we count the sum of the rank numbers for each group (we are talking the sample comparison, thus two group data). Let's say the sum of rank for data group 1 is T1, and T2 for data group 2. Obviously, if T1 and T2 are big different, let's say T1 is much bigger than T2, it indicates the distribution of data group 1 is right shift to the data group 2, thus their distribution is different. The hypothesis is

H_0: Two populations have the same probability distribution
H_1: The probability of population is shifted to either right or left of population 2
Test statistic(two tails): T1 if n1≤n2 or T2 if n2≤n1. we name T1 or T2 as T
Rejection Region: $T \geq T_u$ or $T \leq T_L$

T_U or T_L can be obtained from the statistic table.

Example of Wilcoxon W-M

Sample-1	Rank S1	Sample-2	Rank S2
17	16.5	10	5.5
14	12.5	15	14
12	8.5	7	2
16	15	6	1
23	19	13	10.5
18	18	11	7
10	5.5	12	8.5
8	3	9	4
13	10.5	17	16.5
		14	12.5
n1= 9	T1= 108.5	n2= 10	T2= 81.5
$Mean_1$= 14.56		$Mean_2$= 11.4	
$Median_1$= 14		$Median_2$= 11.5	

n1< n2	T= T1= 108	**Same?**
	$T_{U,9,10}$= 114	YES
	$T_{L,9,10}$= 66	$T_L < T < T_U$

Rank	Sorted data
1	6
2	7
3	8
4	9
5.5	10
5.5	10
7	11
8.5	12
8.5	12
10.5	13
10.5	13
12.5	14
12.5	14
14	15
15	16
16.5	17
16.5	17
18	18
19	23

Poisson PDF: $$P(x,\lambda)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$ for x=0,1,2… Eq - 4

Exponential PDF: $$P(x)=\begin{cases}0 & \forall x<0\\ \lambda e^{-\lambda x} & \forall x\geq 0\end{cases}$$ Eq - 5

Gamma PDF: $$P(x)=\frac{\left(\frac{x-\mu}{\beta}\right)^{\lambda-1}e^{-\frac{x-\mu}{\beta}}}{\beta\Gamma(\lambda)}$$ Eq - 6

Chi-Squared PDF: $$P(x)=\frac{e^{-\frac{x}{2}}x^{(\frac{\nu}{2}-1)}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}$$ Eq - 7