Composite measures of watershed health from a water quality perspective

Ganeshchandra Mallya[1], Mohamed Hantush[2], Rao S. Govindaraju[1,3,*]

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

[2] U.S. EPA National Risk Management Research Laboratory, Cincinnati, OH.

[3] Bowen Engineering Head and Christopher B. and Susan S. Burke Professor

[*] Corresponding author email: govind@purdue.edu

ABSTRACT

Water quality data at gaging stations are typically compared with established federal, state, or local water quality standards to determine if violations (concentrations of specific constituents falling outside acceptable limits) have occurred. Based on the frequency and severity of water quality violations, risk metrics such as reliability, resilience, and vulnerability (R-R-V) are computed for assessing water quality-based watershed health. In this study, a modified methodology for computing R-R-V measures is presented, and a new composite watershed health index is proposed. Risk-based assessments for different water quality parameters are carried out using identified national sampling stations within the Upper Mississippi River Basin, the Maumee River Basin, and the Ohio River Basin. The distributional properties of risk measures with respect to water quality parameters are reported. Scaling behaviors of risk measures using stream order, specifically for the watershed health (WH) index, suggest that WH values increased with stream order for suspended sediment concentration, nitrogen, and orthophosphate in the Upper Mississippi River Basin. Spatial distribution of risk measures enable identification of locations exhibiting poor

watershed health with respect to the chosen numerical standard, and the role of land use characteristics within the watershed.

Keywords: Reliability, resilience, vulnerability, watershed health, scaling, trend analysis, water quality, stream networks

1. INTRODUCTION:

Clean drinking water is not only essential for human health, good water quality is important for avoiding coastal eutrophication and ocean acidification, and for maintaining healthy riverine, marine, and coastal ecosystems which is one of the 17 goals of United Nation's Sustainable Development Goals 2015-2030. Risk-based assessments have been used by water resources planners for characterizing reservoirs (Hashimoto et al., 1982; Moy et al., 1986; Vogel and Bolognese, 1995; Jain and Bhunya, 2008) and urban-water distribution networks (Mondal et al., 2010). Risk analyses have also been popularly used in water quality assessment (Maier et al., 2001; Hoque et al., 2012, 2013, 2014, 2016), evaluating the impact of climate change on water resources systems (Asefa et al., 2014; Fowler et al., 2003; Mondal and Wasimi, 2007), hydrological impact of rain water harvesting systems (Glendenning and Vervoort, 2011) and for drought characterization (Maity et al., 2012). The risk measures that are commonly computed in most risk analyses are reliability, resilience and vulnerability (R-R-V). Reliability is the probability that the system is in compliance at a given time with respect to user specified water quality standard; resilience is defined as the probability of the system to recover to a compliant state given that it was non-compliant the previous time step; and vulnerability is a measure of the average severity of damage during a non-compliant event. While reliability and resilience have probabilistic definitions, vulnerability is often used to quantify the average magnitude of damage caused during a failed (or catastrophic) event. As a result, previously used vulnerability measures have not been comparable to reliability and resilience. In this study we introduce an objective framework for computing a vulnerability metric in water quality risk assessment that is dimensionless and ranges between zero to one. Though the proposed definition is not truly a probability measure, it nevertheless has distributional properties. We also propose a composite water quality-based watershed health measure that describes the overall health of the watershed with respect to any

chosen water quality constituent. By definition, the composite watershed health measure also scales between zero and one, with one indicating a very healthy watershed and vice-versa.

Since risk measures are computed at USGS stations located along the stream network, they may follow scaling laws. Scaling laws have been a popular topic in watershed hydrology. Specifically, Horton's scaling laws (Horton, 1945) provided valuable insights into the hierarchical organizational structure of stream networks within watersheds. These laws also provided an understanding on how the physical properties of streams (e.g. flow characteristics, slope, etc.) would change as a function of spatial scale. Building on this, Strahler (1952, 1957) and Shreve (1966) popularized the concept of stream orders where numbers are assigned to stream reaches based on their hierarchies. Streams that are farthest from the watershed outlet have lowest stream orders, and those that are closest to the outlet have higher stream orders. While some researchers have highlighted the limitations of stream orders and proposed modifications (Peckham and Gupta, 1999; Gangodagamage et al., 2011), others have proposed alternative approaches to scaling analysis (Rigon et al., 1996; Betz et al., 2010; Zaliapin et al., 2010; Gangodagamage et al., 2011). Because of its simplicity, stream order continues to be used in many scaling studies (King et al., 2005; Vondracek et al., 2005; Hoque et al., 2014). For example, Hoque et al. (2014) investigated the scaling behavior of watershed risk measures over four study watersheds in the U.S. Midwest using the Soil & Water Assessment Tool (SWAT) and two scaling measures – contributing upland area and stream order. Their study focused on finding the effective stream order threshold that would yield stable risk measures; however, additional research is needed using measured streamflow data across large river basins.

In this study we investigated scaling laws using actual streamflow measurements within large river basins. Stream order and drainage density were considered as candidate scaling measures.

Drainage density is defined as the ratio of total length of streams within a drainage basin divided by the total drainage area; i.e. it provides a measure of how well existing streams drain a basin. We specifically ask the following questions: What are the distributional properties of these risk measures, and are the means of these risk measures similar at stations of the same stream order or those that have the same drainage density? Do risk measures of reliability, resilience and newly-defined vulnerability increase (or decrease) as we move downstream (lower stream order to higher stream order) or how are they related to different values of drainage density, i.e. do they follow popular scaling laws within large river systems?

We further investigate the spatial distribution of risk measures over large river basins as a prelude to identifying potential source areas. We relate the spatial distribution of risk measures with dominant land use type of each drainage area. Such a comparison allows us to identify land use categories that influence risk measures for different water quality parameters. Mann-Kendall trend test and Sen's slope (Kendall, 1948; Sen, 1968; Hirsch, 1982; Hamed and Ramachandra Rao, 1998) are used to identify influences that are statistically significant. These insights can serve as a useful guide for watershed risk assessment and for implementing useful management plans.

2. STUDY AREA:

The Upper Mississippi River Basin (UMRB), the Ohio River Basin (ORB), and the Maumee River Basin (MRB) were chosen as the study area. The states within the UMRB study area include Minnesota (MN), Wisconsin (WI), Iowa (IA), Illinois (IL), and parts of Missouri (MO). The UMRB has dominant agricultural land use (64%) and drains into Gulf of Mexico. The ORB is spread over states of Indiana (IN), Ohio (OH), Kentucky (KY), and parts of Illinois (IL), Tennessee (TN), Pennsylvania (PA), West Virginia (WV), and New York (NY). It has dominant forest land use (46%) followed by agricultural land use (44%) and drains into the Mississippi River and

ultimately into the Gulf of Mexico. Due to intensive agricultural activities that involve application of fertilizers, both UMRB and ORB are considered to be among the primary sources of nutrients that reach the Gulf and cause eutrophication (Burkart and James, 1999; Alexander et al., 2008). The MRB is the largest Great Lakes watershed, draining all or parts of 17 Ohio (OH) counties, two Michigan (MI) counties and five Indiana (IN) counties into Maumee Bay and then to Lake Erie just east of Toledo, Ohio. The MRB is agriculturally intensive (53%) and the nutrients that get washed off from this river basin cause algal blooms in Lake Erie during the summer months (Michalak et al., 2013).

A total of 214 USGS stations (Figure 1) - 57 stations in UMRB, 99 stations in ORB, and 58 stations in MRB with available water quality (WQ) data were identified over the study area. The U. S. Geological Survey (USGS, http://waterdata.usgs.gov/nwis/rt) daily streamflow dataset was utilized. In general, streamflow data are subjected to human interference, and therefore data contain both natural and regulated flows. Only unregulated stations were included in this study.

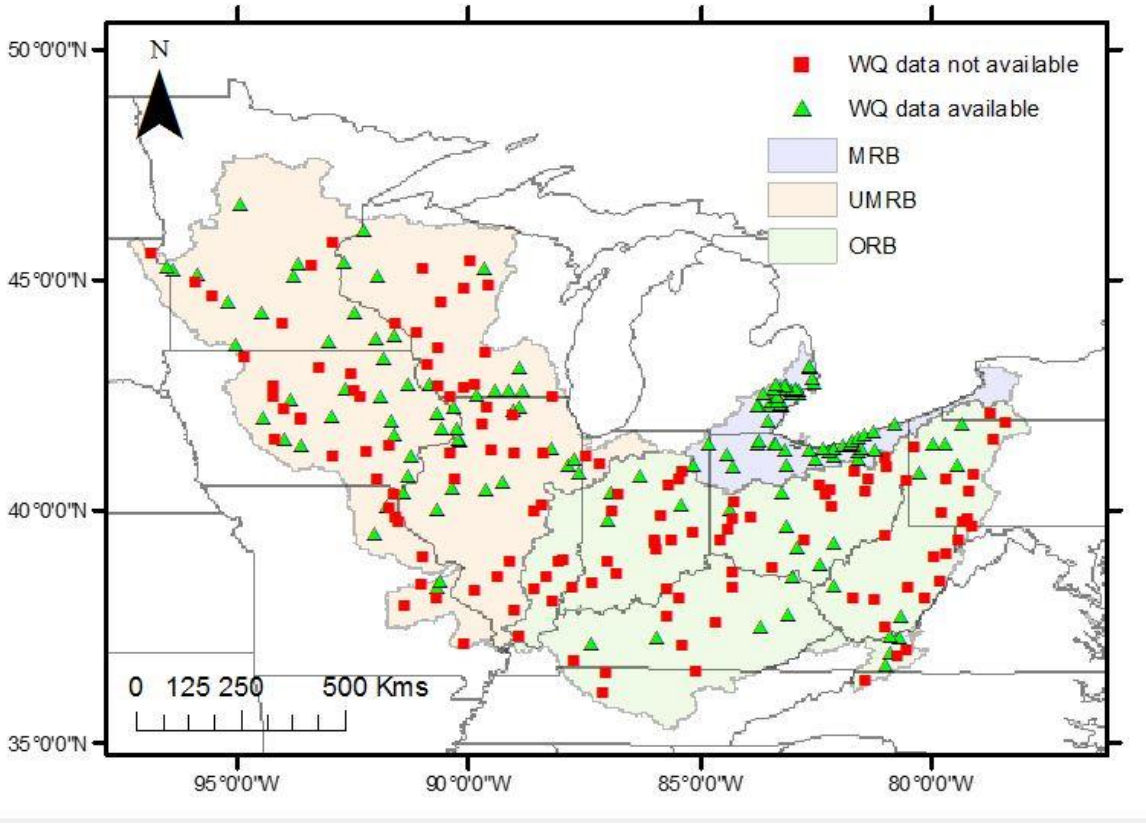

Figure 1: Study area showing unregulated USGS stations. Green markers denote stations where water quality data are available.

The U. S. Geological Survey (USGS, goo.gl/K1Th9D) daily water quality dataset and the USGS National Water Quality Assessment (NAWQA, goo.gl/Wq5dYi) data warehouse were used to collect chemical, biological and physical water quality data for the study area where available. We have a total of 151 stations with Suspended Sediment Concentration data (parameter code 80154), 70 stations with Nitrate + Nitrite data (parameter code 00631), and 49 stations with Orthophosphate data (parameter code 00671). These parameters were chosen based on the number of sampling stations with minimum 30 observations over the study period (1966 to current depending on data availability). The threshold of 30 observations was chosen to ensure a statistically robust model during reconstruction of the WQ time series.

3. METHODOLOGY:

While daily continuous records of streamflow data were available over the study area, water quality data are discontinuous in time. Using the data reconstruction method proposed in Hoque et al. (2012) water quality data at all stations were reconstructed as a function of daily streamflow measurements available at those stations using relevance vector machines (RVM; Bishop, 2006; Schölkopf and Smola, 2002; Tipping, 2001). Let X_t be the daily reconstructed time series of a water quality parameter with standard numerical target X^*. We then define compliance (S) and noncompliance states (F) as:

$S: \{X_t \leq X^*\}$ is compliance state ($X_t > X^*$, e.g., for Dissolved Oxygen)

$F: \{X_t > X^*\}$ is noncompliance state ($X_t \leq X^*$, e.g., for Dissolved Oxygen)

A compliance event is one where the reconstructed water quality data is below (or above in case of Dissolved Oxygen or DO) the standard numerical target for one or more successive days, and is noncompliant otherwise. Then using the definitions of risk measures given by Hashimoto et al. (1982) and Hoque et al. (2012), reliability (p) is defined as the probability of the system to be in compliant state. Mathematically it can be written as

$$p = 1 - P\{X_t \in F\} = 1 - \frac{1}{n}\sum_{t=1}^{n} z_t \quad (1)$$

where z_t = 1 when $X_t \in F$ and 0 when $X_t \in S$, and n is the total number of data points.

Similarly, resilience (r) is defined as the probability of the system to recover from a non-compliance state and can be mathematically written as below:

$$r = P\{X_{t+1} \in S | X_t \in F\} = \frac{P\{X_{t+1} \in S \cap X_t \in F\}}{P\{X_t \in F\}} = \frac{\sum_{t=1}^{n} y_t}{\sum_{t=1}^{n} z_t} = \frac{l}{m}$$

$r = 1$ when $X_t \in S$ for all t (2)

where $y_t = 1$ when $X_{t+1} \in S$ and $X_t \in F$ and 0 otherwise, and m is number of times where $X_t > X^*$ ($X_t < X^*$, e.g., in case of oxygen) or $m = \sum_{t=1}^{n} z_t$, and $l = \sum_{t=1}^{n} y_t$.

The standard definitions of vulnerability involve quantification of magnitude of damage during a noncompliant event. For water quality violations in stream networks there is no objective way in which magnitude of damage can be quantified. Previous studies on water supply and water quality have computed vulnerability by quantifying magnitude of violations with respect to a chosen standard (Hoque et al., 2012; Kjeldsen and Rosbjerg, 2004; Kundzewicz and Kindler, 1995; Moy et al., 1986). These resulting vulnerability measures did not scale from 0 to 1.

In this study we propose a new metric (v) – *robustness* – that scales between 0 and 1, and can be mathematically defined as the exponent of the negative of the sample mean of log normalized ratio of X_t to the standard value as follows

$$v = exp\left\{-\frac{1}{m}\sum_{t=1}^{n} ln\left[\frac{Q_t X_t \Delta t}{Q_t X^* \Delta t}\right] H[X_t - X^*]\right\} \qquad \left(ln\left[\frac{Q_t X^* \Delta t}{Q_t X_t \Delta t}\right] H[X^* - X_t] \text{ in case of DO}\right) \quad (3)$$

where m is number of times where $X_t > X^*$, $Q_t X_t \Delta t$ is the water quality load at time t, $Q_t X^* \Delta t$ is the standard water quality load at time t and $H[.]$ is the Heaviside function that accounts only for the noncompliant events. Note that when deviations of X_t from X^* are large $v \to 0$; when deviations are small $v \to 1$, and vice versa for a constituent whose concentration should be less than the standard (e.g., oxygen). This is consistent with definitions for reliability (p) and resilience (r). Vulnerability can now be quantified as:

$V = 1 - v$ (4)

Note that with these definitions, robustness and vulnerability have the same distributional properties as random variables. We now define a conservative composite measure of watershed health (h) as the geometric mean of the three indices:

$$h = (p\ r\ v)^{\frac{1}{3}} \tag{5}$$

If $p = r = v = 1$, $h = 1$, if either one is 0, $h = 0$, i.e. a watershed/catchment is healthy when all risk measures are high (close to 1), otherwise it is not.

Suspended sediment concentration (SSC, parameter code 80154, standard 30 mg/L) was chosen as the water quality parameter to monitor sediment erosion by runoff and turbidity in streams. Similarly, nitrite and nitrate (NO2 + NO3 in mg/L as N, parameter code 00631, standard 10 mg/L) were chosen as the water quality parameters of interest to monitor nitrogen concentrations within the study area. Orthophosphate (parameter code – 00671, standard of 0.1 mg/L) was chosen as the water quality parameter of interest to monitor Phosphorus concentrations over the study area. These standards were adopted from US EPA (1986). The RVM (Hoque et al., 2012) was implemented to reconstruct the water quality time series along with corresponding uncertainties. The latter was computed from an ensemble of 10,000 time series realizations generated by the Monte Carlo approach used in RVM. Reliability, resilience, robustness, vulnerability and composite watershed health measure were computed using each of these 10,000 reconstructed water quality time series realizations at all stations in the study area where the three water quality constituents and streamflow rates were monitored.

These 10,000 values of each risk metric were then used to study the distributional properties of these risk indicators. Using model selection metrics such as Akaike Information Criteria (AIC; Akaike, 1974) and Bayesian Information Criteria (BIC; Schwarz, 1978) several candidate

probability distributions such as Beta, Birnbaum-Saunders, Exponential, Extreme value, Gamma, Generalized extreme value, Generalized Pareto, Inverse Gaussian, Logistic, Log-logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t location-scale and Weibull distribution were tested. Distribution properties of the risk metrics were investigated over all stations in the study area for suspended sediment concentration, nitrogen and phosphorus, and for different water quality standards (e.g. US EPA (1986), Aquatic Life Criteria, Human Health Criteria, Maximum Contaminant Load for drinking water, etc.) of these chosen constituents. Further, to statistically confirm, if the risk measures (with respect to chosen WQ constituent) obtained for different numerical targets were different compared to commonly used standard (e.g. 30 mg/L for SSC; US EPA, (1986)), a Kolmogorov-Smirnov (KS) test was performed.

To study the scaling relationship of risk measures within a river basin, first, their box-plots were compared for different stream orders. Then, Mann-Kendall (MK) trend test was performed on their mean values and Sen's slope was reported. The risk measures were said to statistically increase with stream orders if the p-value of MK-test was less than the critical value ($\alpha = 0.05$) and Sen's slope was positive. Similarly, risk measures were said to statistically decrease with stream orders if p-value was less than the critical value ($\alpha = 0.05$) and Sen's slope was negative. The MK-test was also repeated for lower-order (≤ 2) and higher-order (> 2) streams within a river basin.

Finally, to study the spatial distribution of risk measures and explore their relationship with different land use types, a combination of spatial plots, scatter plots of risk measures versus percentage land use, and Mann-Kendall trend tests were used. The Mann-Kendall trend test was performed by considering stations over each river basin separately.

4. RESULTS AND DISCUSSION:

4.1 Distribution properties of risk measures

For any chosen water quality constituent at a USGS monitoring station, following Hoque et al. (2012) and equations (1)-(5), reliability, resilience, vulnerability and composite watershed health measures were computed for each of the 10,000 Monte-Carlo realizations of reconstructed water quality time series. It should however be noted that we have not included uncertainty in streamflow measurements when reconstructing water quality time series as information on measurement error was not available at USGS stream gauges used in this study. Figure 2 shows the empirical distribution of the risk metrics from the 10,000 simulations at USGS station 3015500 for suspended sediment concentration (SSC) along with the fit of four distributions according to best BIC scores. The numerical target for SSC was set to 30 mg/L. For this station, Beta distribution provided the best fit for reliability and watershed health measures. For resilience and robustness measures, Rician and Gamma distributions had best BIC scores respectively, while Beta distribution was still among top-6 distributions and the differences in BIC scores of the top-6 distributions were negligible. While we have investigated the distributional properties over all stations in the study area with respect to suspended sediment concentration, nitrogen, and phosphorus, for the purpose of brevity we have only included results from three randomly selected stations (each belonging to either UMRB, MRB or ORB, see Figure A3, Figure A4, and Figure A5) in Appendix. The Beta distribution provided fairly good description of risk measures for all three water quality constituents, as well as for all stations in the study area.

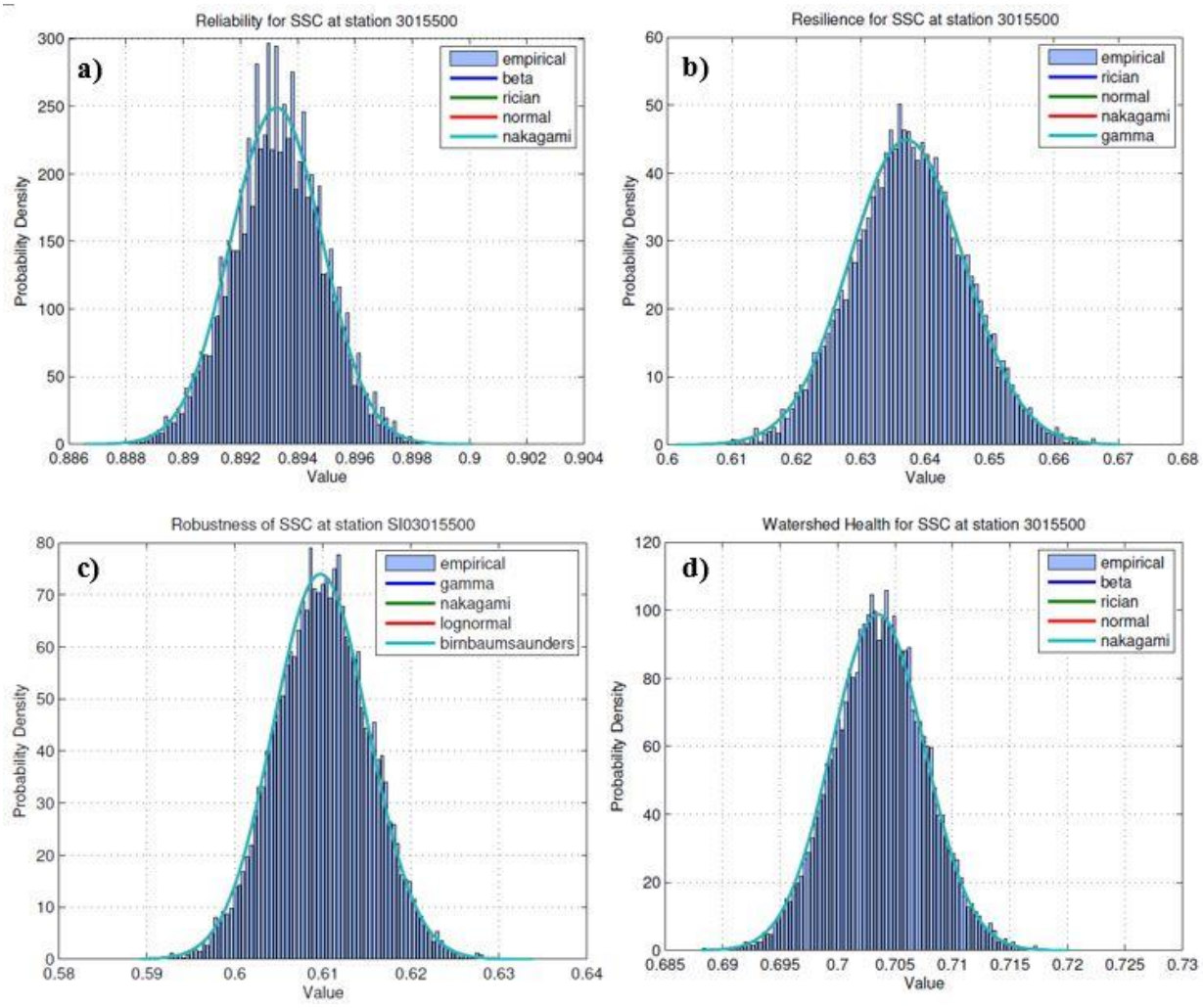

Figure 2: Four best fit distributions to 10000 realizations of a) reliability, b) resilience, c) robustness, and d) composite Watershed Health measure for Suspended Sediment Concentration (SSC) at USGS station 3015500.

A common feature of the distributional properties of the risk metrics is that they are tightly bound about the mean (see Figure A2 in Appendix). The tight bounds indicate that any chosen Monte-Carlo realization could provide risk estimates close to the true value (or the mean of the distribution). While the mean of the distribution shifted for other water quality standards, the values were found to be tightly bound similar to Figure A2. Further, the histograms of risk measures (Figure A6) with respect to WQ constituents for different choices of numerical targets

(e.g. 7.5 mg/L, 15 mg/L, 30 mg/L, 45 mg/L, 60 mg/L, and 120 mg/L for SSC) were well separated indicating that risk measures were sensitive to the choice of numerical target and the Kolmogorov-Smirnov (KS) tests (Figure A7 and Figure A8 in Appendix) at different significance levels α (e.g. 0.1, 0.05, 0.025, and 0.01) showed that the risk measures obtained for different numerical targets were statistically different compared to commonly used standards (US EPA, 1986). The study found that at stations where results were not sensitive, the reconstructed water quality time series never exceeded the numerical targets. Similar results may also be obtained when the magnitude of exceedances are large enough that raising or lowering the standard does not affect the risk measures, or when there is large uncertainty in the reconstructed water quality time series. Discussion of these results with respect to SSC, Nitrate + Nitrite, and Orthophosphate are provided in Appendix (Figure A7 and Figure A8).

4.2 Scaling behavior of risk measures

The scaling behavior of risk measures was investigated for suspended sediment concentration (77, 40, and 34 stations), nitrogen (38, 20, and 12 stations) and phosphorus (30, 9, and 10 stations) over UMRB, ORB and MRB respectively. Strahler number and drainage density were initially chosen as the scaling metrics. No scaling behavior was apparent with drainage density, and results with Strahler number, also known as stream order, are discussed below. For suspended sediment concentration the standard (or threshold) was set to 30 mg/L (US EPA, 1986), reliability, resilience, robustness and the composite watershed health measures were computed at each station in the study area where measurements were available.

The top row in Figure 3 shows the scaling relationship of reliability with stream order in UMRB, ORB and MRB, respectively. Mann-Kendall trend test was performed on the mean values of reliability for all stream orders, as well as for lower-order (≤ 2) and higher-order (> 2) streams

within a river basin. The results for the latter are included within brackets. When the results are statistically significant, the Sen's slope is shown as a red dashed line, and the slope values are shown in red color. The height of the box plots suggest that there is scatter in reliability measures for each stream order, but their median values show relatively flat (Sen's slope = -0.0001 for UMRB) to decreasing trends (Sen's slope -0.03 and -0.06 for ORB and MRB, respectively) with increase in stream order, although they were not statistically significant at $\alpha = 0.05$ based on the Mann-Kendall test. The scaling relationship of resilience with stream order for UMRB, ORB and MRB are shown in the second row of Figure 3. Median values of resilience show a relatively flat trend (Sen's slope = -0.0002) in UMRB and decreasing trends (Sen's slope of -0.035 and -0.035) over ORB and MRB with increase in stream order. However, these trends were not statistically significant according to Mann-Kendall trend test. When considering only higher stream orders (>2), the mean resilience values showed a statistically significant decreasing trend (Sen's slope = -0.05) for ORB, indicating deteriorating downstream conditions. Reliability and resilience are expected to have similar behavior (Hashimoto et al., 1982). The newly defined risk measure, robustness (v), shows a statistically significant positive trend (Sen's slope = 0.035) with increase in stream order for UMRB, and weak decreasing trends (Sen's slope of -0.028 and -0.042) for ORB and MRB (third row of Figure 3). Therefore, vulnerability which is defined as $(1 - v)$ would show a decreasing trend with stream order in UMRB. The statistically significant positive trends in robustness measures over UMRB are possibly due to dilution effects at higher stream orders – but this needs further investigation. Watershed health has a statistically insignificant decreasing trend (Sen's slope of -0.03 and -0.05) with increase in stream order over ORB and MRB, however this trend is positive (Sen's slope = 0.015) and significant for UMRB.

To attribute the reason for statistically significant improvements (according to MK trend test, see Figure 3) in mean watershed health (w.r.t. SSC) over UMRB another one-tailed KS test was performed. The KS test investigated whether watershed health values at all higher-order (stream order > 2) stations combined were statistically higher compared to all values computed at lower-order (stream order ≤ 2) stations. The p-values from this KS test (see Table A1) indicated that the null hypothesis of similar distributions of watershed health between stations located at lower-order and higher-order streams could not be rejected at 5% significance level (α). Table A2 shows the results from a one-tailed KS test with an alternate hypothesis that percentage area under agriculture land use for stations located along lower-order streams is statistically higher than for stations along higher-order streams. The p-values indicate that the null hypothesis (similar distributions) cannot be rejected at 5% significance level, and corroborates the results in Figure 6 (first row, first column) where stations from both lower- and higher-order streams are present in almost equal numbers at the lower (0-10%) and higher (70-95%) ends of the x-axis (percentage areas under agricultural land use). Finally, we also investigated whether forest land use may have played a role in improved watershed health (with respect to SSC) at higher stream orders. The p-values (see Table A3) indicate that percentage area under forest is statistically higher for higher-order streams, when compared to lower-order streams at 5% significance level in UMRB. These results support the theory of dilution effects (relatively pristine flows from forested land) at downstream locations in UMRB leading to improved watershed health.

Based on previous scaling studies (Hoque et al., 2014) which have used models, there is an expectation to observe trends in risk measures with stream order. But we were not able to find such trends in this study because we have not used models. Models impose these scales explicitly, whereas such rules are not being imposed here during data reconstruction. Reconstructed

observational data do not recognize that these stations (and their observations) are part of a stream network, or that certain laws have to be obeyed, so there is less chance of realizing such trends.

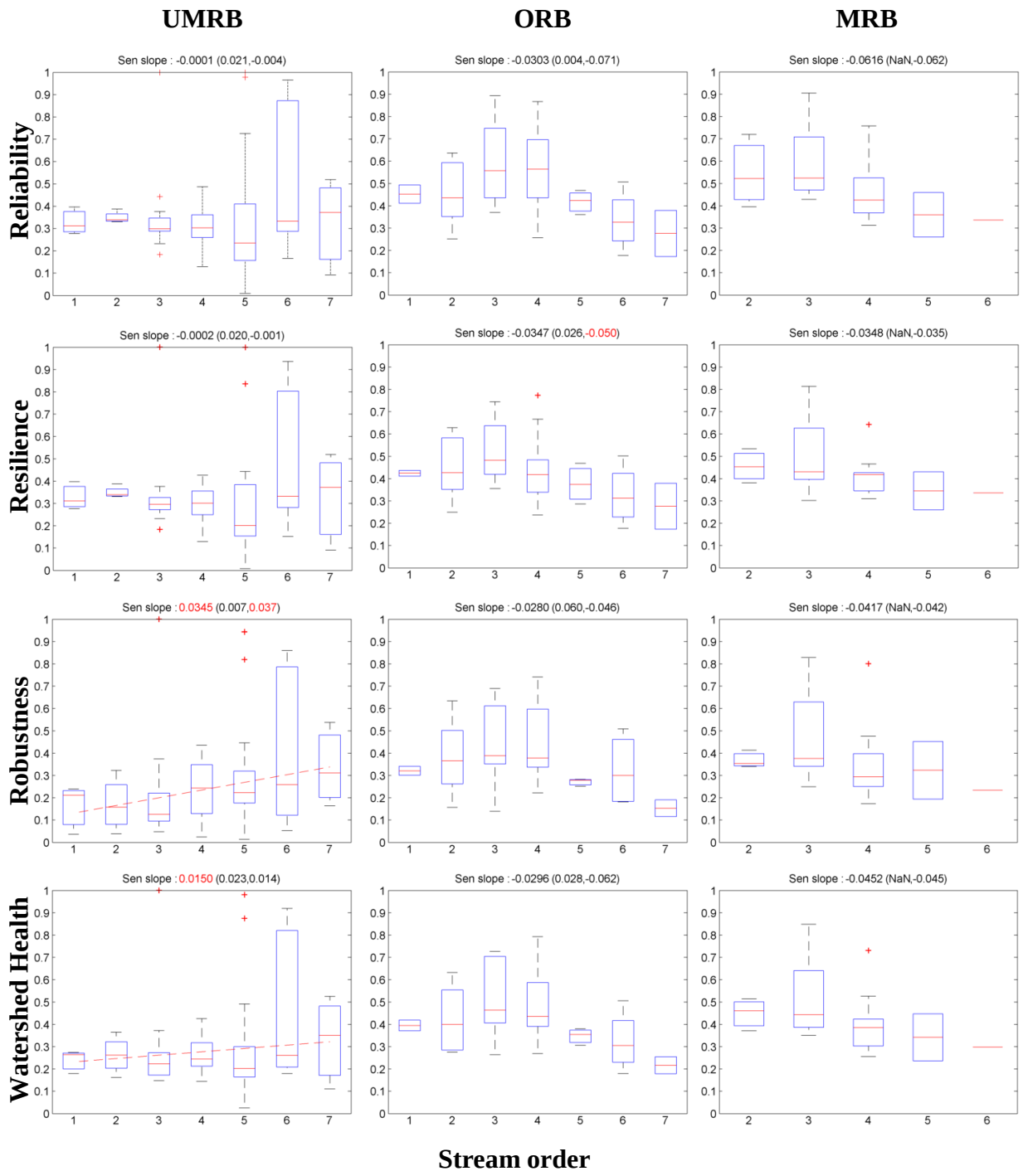

Figure 3: Scaling behavior of risk metrics for Suspended Sediment Concentration with respect to stream order at UMRB, ORB and MRB. Sen's slope estimate from Mann-Kendall trend test for mean is shown. The values within the brackets denote Sen's slope estimate from Mann-Kendall trend test for mean values of risk measures on lower- (≤ 2) and higher- (> 2) order streams respectively. Statistically significant trends $(\alpha = 0.05)$ are labeled in red color. Sen's slope is denoted as a red dashed line for cases where trends are significant.

The scaling relationships of different risk metrics for nitrogen were investigated by choosing stream order as the scaling metric. Nitrogen data records were available only at 20 stations in ORB and 12 stations in MRB. Therefore, scaling analysis over these basins were not performed. UMRB on the other hand had 38 nitrogen monitoring stations. Figure 4 shows the scaling results of risk metrics with respect to stream order over UMRB. While the slopes of the trends (Sen's slope estimates were 0.015, 0.015, 0.005, and 0.01 for reliability, resilience, robustness, and watershed health, respectively) were found to be positive, they were not statistically significant at $(\alpha = 0.05)$. Positive trends in risk measures indicate that conditions improve as we move downstream in UMRB. Dilution effects may have played a role in improving conditions (Ahearn et al., 2005; Anbumozhi et al., 2005). As in the case of SSC, one-tailed KS tests were carried out to test several hypotheses. The p-values from KS test shown in Table A1 indicated that the null hypothesis of similar distributions in watershed health (with respect to nitrogen) at lower- and higher- stream orders could not be rejected at 5% significance level (α). The p-values of KS test (Table A2), with an alternate hypothesis of statistically higher percentage area of agricultural land use at lower-order streams, indicate that the null hypothesis (similar distributions in percentage area between lower- and higher-order streams) cannot be rejected at 5% significance level, and this agrees with the results in Figure 6 (first row, second column) where the stations from both

lower- and higher-order streams are clustered around 85% mark. Next we investigated whether forest land use may have played a role in improved watershed health (with respect to nitrogen) at higher stream orders. Figure A1 (in Appendix) shows that the area under forest land use dominates as we move downstream, which implies smaller amounts of nutrients reach the receiving waters, and consequently may have caused dilution. The watershed health for nitrogen shows a statistically significant improvement with increase in forest land use (Figure 6; second row, second column). However, most of the markers tend to be clustered below 30% mark. Therefore, to test for statistical difference in percentage area of forest land use between lower-order and higher-order streams a one-tailed KS test was performed. The p-values in Table A3 indicate that in case of UMRB, the percentage area under forest is statistically higher for higher-order streams, when compared to lower-order streams at 5% significance level. These results lead us to believe that higher contribution of runoff from forested land at downstream locations in UMRB may have resulted in dilution and improved watershed health.

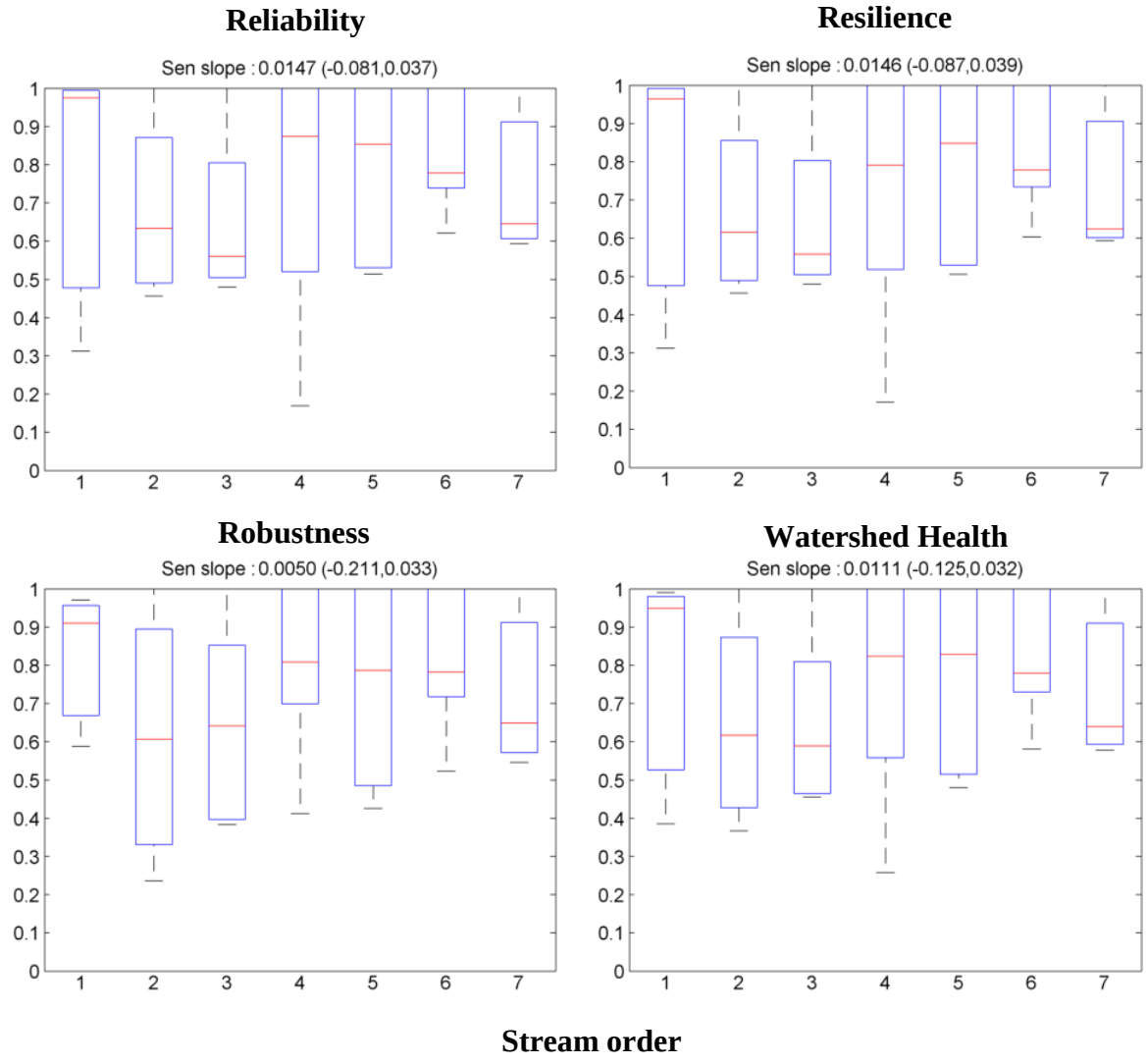

Figure 4: Scaling behavior of risk metrics for Nitrogen with respect to stream order at UMRB. Sen's slope estimate from Mann-Kendall trend test for mean is shown. The values within the brackets denote Sen's slope estimate from Mann-Kendall trend test for mean values of risk measures on lower- (≤ 2) and higher- (> 2) order streams, respectively. Statistically significant trends $(\alpha = 0.05)$ are labeled in red color. Sen's slope is denoted as a red dashed line for cases where trends are significant.

Given the strong affinity of phosphorus to sediments, the scaling relationships of risk measures for phosphorus as a function of stream order were found to be similar to SSC (see Figure A9 in Appendix), with risk measures showing positive trend with increasing stream order. The positive trend was found to be statistically significant only for watershed health index, and may be attributed to potential dilution effects caused by waters originating from forested lands within UMRB resulting in improved health conditions as we move downstream (see Table A3).

4.3 Spatial distribution of risk measures

Figure 5 shows the spatial distribution of mean risk metrics for suspended sediment concentrations at several USGS stations in the study area where this water quality constituent is monitored (readers are referred to Figure A12 for a zoomed map of spatial distribution of risk measures at MRB). At few interior locations (stream order 3-5 in UMRB, 3-4 in ORB, and 2-3 in MRB) of the three river basins in the study area, reliability, resilience, and watershed health values are in the range 0.8 to 1.0 indicating very few violations with respect to the chosen water quality standard value. The vulnerability values for these stations are close to zero as expected. As we move further downstream (notice these also include several lower-order streams with predominant agriculture land use) we notice that the watershed health (and other risk measures) deteriorates to a range of 0.3 to 0.8 in the middle reaches, and 0.1 to 0.3 in the lower reaches. The results from Figure 5 when analyzed simultaneously with Figure 6 offer possible explanation. First column (Figure 6, SSC) indicates that watersheds with dominant agricultural land use and smaller area under forest land use are the current hotspots for suspended sediment concentration at the chosen water quality standard value. Mann-Kendall trend test on values of watershed health indicates that the trend is negative (Sens's slope = -0.006) with respect to increasing agricultural land use area and is statistically significant. While watersheds (drainage areas) that are predominantly forests are generally pristine with respect to SSC (Mann-Kendall trend is positive and significant, with Sen's

slope of 0.007). Although more dispersed, the pattern for phosphorus mimicked that for SSC (third in row 1 and 2, Figure 6)

For nitrate, it can be observed that areas within the URMB drained by stream orders 1 to 3 and more than 80% agricultural land use have relatively low watershed health index values (second in first row, Figure 6). Drainage areas corresponding to higher stream orders show better watershed health values with decreasing % of agricultural land use. This pattern is accentuated by the visible separation of relatively lower health values (blue stations) from the relatively higher health stations (marked in yellow). This pattern is generally true for phosphorus where health values increase with the stream order but only with decreasing percentage of agricultural land use (third in first row, Figure 6). It appears that higher-order streams tend to be associated with less agriculture-dominated areas and are generally healthier when compared to head-watersheds. Higher-order streams whose drainage areas are predominantly agricultural have low health values, especially with respect to suspended sediment concentration.

A few water quality stations are available for predominantly urban catchments which occupy the smallest land-use fraction in the UMRB. The fourth row in Figure 6 shows that streams of order 1 to 3 draining areas that are 70% or more urban are in poor health with respect to SSC, good health with respect to nitrate, and mixed health for phosphorus.

Similar observations can be made with regards to WH values for SSC at ORB and MRB, but have not been shown here for brevity. For Nitrogen and Orthophosphate, stations (located in ORB and MRB) with agriculturally intensive drainage areas have high values for vulnerability, and low values for reliability, resilience, and watershed health (see Figure A10 and Figure A11 in Appendix).

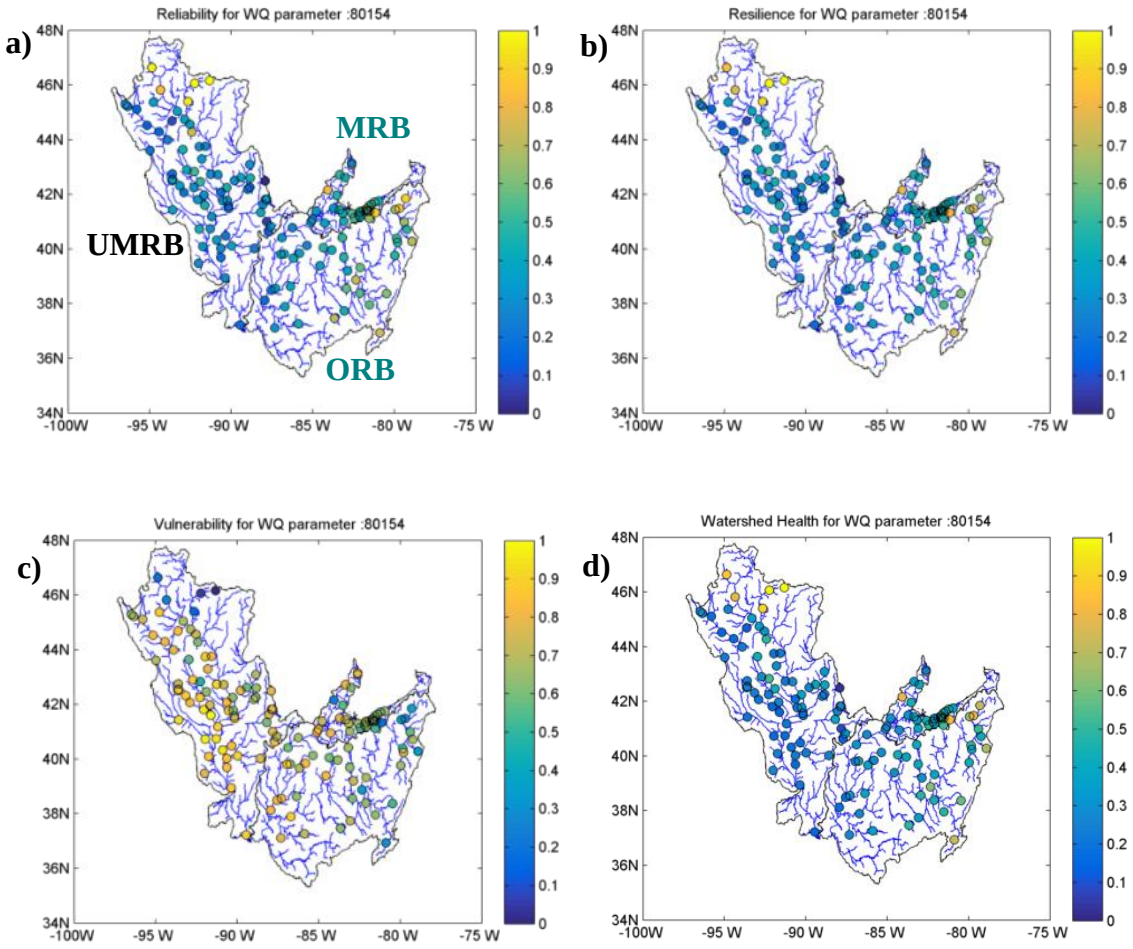

Figure 5: Spatial distribution of a) Reliability, b) Resilience, c) Vulnerability, and d) Watershed Health measures for Suspended Sediment Concentration (parameter code: 80154).

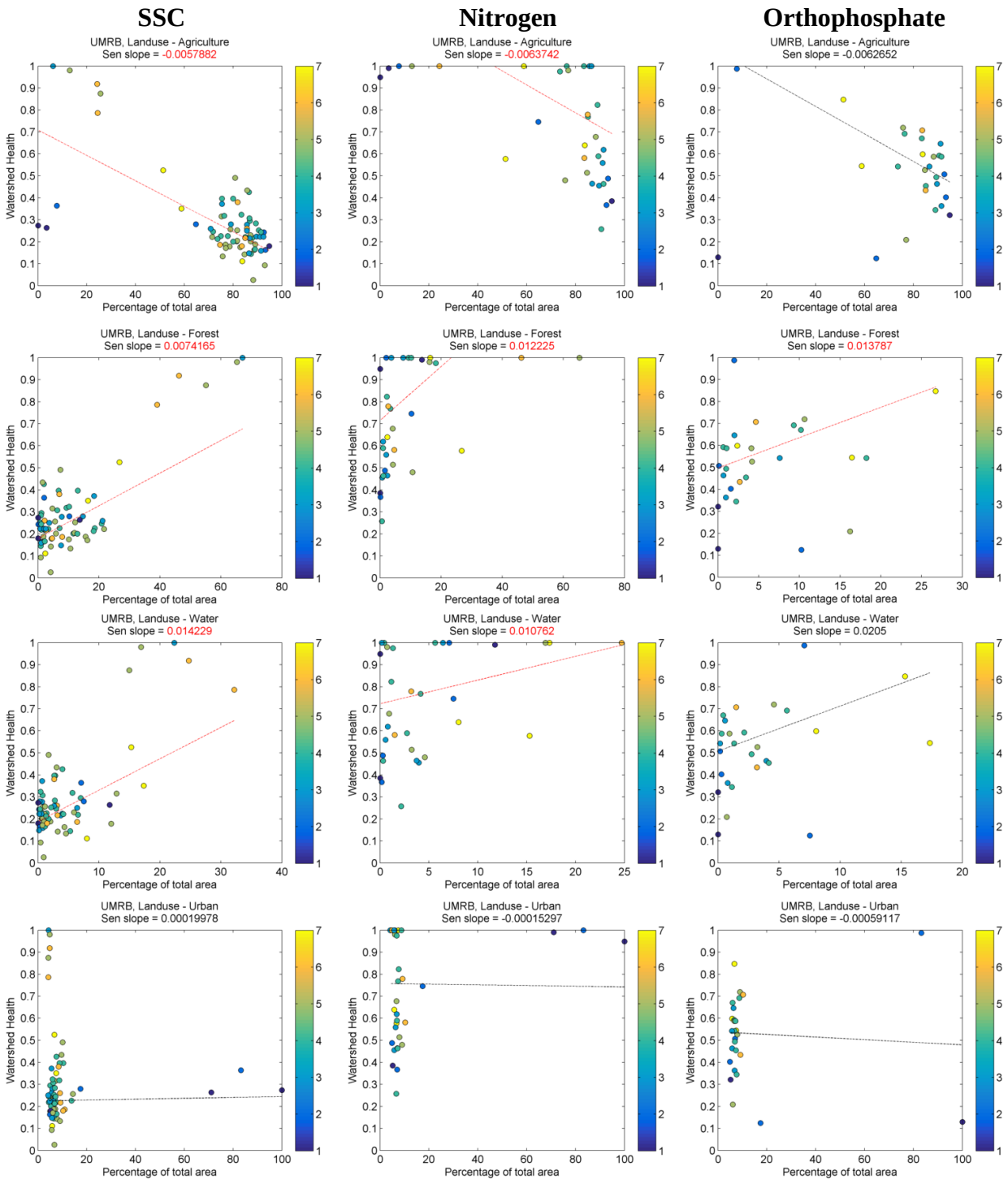

Figure 6: Relationship between watershed health index and area under agriculture land use (first row), forest land use (second row), water land use (third row), and urban land use (fourth row) for

SSC, Nitrogen, and Phosphorus at UMRB. The markers represent stations where these WQ constituents are measured, and they are color coded based on stream order.

5. CONCLUSIONS:

In this study, we proposed a new definition for *robustness* that scales conveniently between zero and one. A composite measure for water quality-based watershed health defined as the geometric mean of reliability, resilience and robustness was introduced. When the reliability, resilience and robustness measures are high, the resulting watershed health measure is high and vice-versa. The distributional properties of the risk metrics were briefly discussed. Among different candidate distributions, Beta distribution was recommended as a promising model. While the mean was found to shift based on the choice of water quality standard, the spread about the mean was tightly bound indicating that the risk measures computed from any of the 10000 Monte-Carlo simulations provides risk estimates that are close to the mean. Thus, for simplification purposes we can neglect the variation around the mean of the reconstructed series, and confidently use the risk value obtained from one MC simulation.

The sensitivity analysis of risk measures indicated that except for some stations (about 30% of the stations) that record nitrogen data, the risk measures were found to be sensitive to the choice of numerical targets particularly for most stations that monitor SSC and Orthophosphate data.

The nature of scaling was dependent on the choice of water quality parameter, and significant scatter in risk measures were observed at different stream orders. For SSC, there was significant increase in WH with stream order at UMRB. Similar trend results were obtained for WH using Orthophosphate at UMRB. However for Nitrogen, the computed trend with respect to stream orders was positive but not statistically significant. Positive trends in watershed health for the three

water quality constituents were attributed to dilution effects from drainage areas with predominantly forest land use. Based on earlier scaling studies there was an expectation to observe trends in risk measures with stream order. But such trends were not apparent in this study because the statistical reconstruction method used here does not recognize stations as part of any network, and therefore scaling laws are not imposed explicitly. However, this is not the case when hydrologic models are used for simulating flows or water quality time-series.

In terms of spatial distribution of SSC risk measures over the study area, WH was found to decrease at stations that drain larger agricultural land use areas, whereas WH increased at stations with larger forest land use areas within the UMRB. Similar trend results were observed for WH based on Nitrogen, with WH decreasing with increase in agricultural land use areas, and WH increasing with increase in forest land use areas. The results for SSC and Nitrogen were statistically significant at UMRB. For Orthophosphate the trends were similar to SSC and Nitrogen, however it was statistically significant for forested area, and not statistically significant for agricultural land use areas. The results were found to be consistent for other river basins (ORB and MRB) although the number stations where observations were available were relatively small compared to UMRB.

Watershed health with respect to water quality parameters were most impacted in agriculturally dominated watersheds, while those with highest percentage of forested areas (and thus most protected from changes) were not. The watershed health values at these relatively undisturbed (pristine) forested watersheds give us an estimate of what to expect as the upper limit of good watershed health.

We expect similar results for forested watersheds in other areas, which have perhaps seen less human influence. But for agriculture watersheds where cultivation practices such as irrigation and

fertilizer application have been used, we should expect that with time, conditions would have deteriorated.

We note that quality of reconstruction is affected by the number of observed water quality data points at a station. With fewer observations, there will be larger uncertainty (wider band around the median estimate) in our reconstructed series. There may be a critical number of data points needed for meaningful reconstruction.

6. ACKNOWLEDGEMENTS:

The U.S. Environmental Protection Agency through its Office of Research and Development funded and managed the research described here under EPA Contract #EP-C-15-010. The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

7. REFERENCES:

Ahearn, D.S., Sheibley, R.W., Dahlgren, R.A., Anderson, M., Johnson, J., Tate, K.W., 2005. Land use and land cover influence on water quality in the last free-flowing river draining the western Sierra Nevada, California. Journal of Hydrology 313, 234–247. https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2005.02.038

Akaike, H., 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19, 716–723.

Alexander, R.B., Smith, R.A., Schwarz, G.E., Boyer, E.W., Nolan, J.V., Brakebill, J.W., 2008. Differences in Phosphorus and Nitrogen Delivery to The Gulf of Mexico from the Mississippi River Basin. Environ. Sci. Technol. 42, 822–830. https://doi.org/10.1021/es0716103

Anbumozhi, V., Radhakrishnan, J., Yamaji, E., 2005. Impact of riparian buffer zones on water quality and associated management considerations. Ecological Engineering 24, 517–523. https://doi.org/http://dx.doi.org/10.1016/j.ecoleng.2004.01.007

Asefa, T., Clayton, J., Adams, A., Anderson, D., 2014. Performance evaluation of a water resources system under varying climatic conditions: Reliability, Resilience, Vulnerability and beyond. Journal of Hydrology 508, 53–65. https://doi.org/https://doi.org/10.1016/j.jhydrol.2013.10.043

Betz, R., Hitt, N., Dymond, R., Heatwole, C., 2010. A method for quantifying stream network topology over large geographic extents. Journal of Spatial Hydrology 10.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer New York.

Burkart, M.R., James, D.E., 1999. Agricultural-nitrogen contributions to hypoxia in the Gulf of Mexico. Journal of Environmental Quality 28, 850–859.
Fowler, H.J., Kilsby, C.G., O'Connell, P.E., 2003. Modeling the impacts of climatic change and variability on the reliability, resilience, and vulnerability of a water resource system. Water Resources Research 39, n/a–n/a. https://doi.org/10.1029/2002WR001778
Gangodagamage, C., Belmont, P., Foufoula-Georgiou, E., 2011. Revisiting scaling laws in river basins: New considerations across hillslope and fluvial regimes. Water Resources Research 47.
Glendenning, C.J., Vervoort, R.W., 2011. Hydrological impacts of rainwater harvesting (RWH) in a case study catchment: The Arvari River, Rajasthan, India: Part 2. Catchment-scale impacts. Agricultural Water Management 98, 715–730. https://doi.org/https://doi.org/10.1016/j.agwat.2010.11.010
Hamed, K.H., Ramachandra Rao, A., 1998. A modified Mann-Kendall trend test for autocorrelated data. Journal of Hydrology 204, 182–196. https://doi.org/10.1016/S0022-1694(97)00125-X
Hashimoto, T., Stedinger, J.R., Loucks, D.P., 1982. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. Water Resour. Res. 18, 14–20. https://doi.org/10.1029/WR018i001p00014
Hirsch, R.M., 1982. Techniques of trend analysis for monthly water quality data. WATER RESOUR RES 18, 107. https://doi.org/10.1029/WR018i001p00107
Hirschboeck, K.K., 1991. Climate and floods. US Geological Survey Water Supply Paper 2375, 67–88.
Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogramm. Eng. Remote Sens 81, 345–354.
Hoque, Y.M., Hantush, M.M., Govindaraju, R.S., 2014. On the scaling behavior of reliability–resilience–vulnerability indices in agricultural watersheds. Ecological Indicators 40, 136–146.
Hoque, Y.M., Raj, C., Hantush, M.M., Chaubey, I., Govindaraju, R.S., 2013. How Do Land-Use and Climate Change Affect Watershed Health? A Scenario-Based Analysis. Water Qual Expo Health 6, 19–33. https://doi.org/10.1007/s12403-013-0102-6
Hoque, Y.M., Tripathi, S., Hantush, M.M., Govindaraju, R.S., 2016. Aggregate Measures of Watershed Health from Reconstructed Water Quality Data with Uncertainty. Journal of environmental quality 45, 709–719.
Hoque, Y.M., Tripathi, S., Hantush, M.M., Govindaraju, R.S., 2012. Watershed reliability, resilience and vulnerability analysis under uncertainty using water quality data. Journal of Environmental Management 109, 101–112. https://doi.org/10.1016/j.jenvman.2012.05.010
Horton, R.E., 1945. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. Geological society of America bulletin 56, 275–370.
Jain, S.K., Bhunya, P.K., 2008. Reliability, resilience and vulnerability of a multipurpose storage reservoir / Confiance, résilience et vulnérabilité d'un barrage multi-objectifs. Hydrological Sciences Journal 53, 434–447. https://doi.org/10.1623/hysj.53.2.434
Kendall, M.G., 1948. Rank correlation methods.

King, R.S., Baker, M.E., Whigham, D.F., Weller, D.E., Jordan, T.E., Kazyak, P.F., Hurd, M.K., 2005. Spatial considerations for linking watershed land cover to ecological indicators in streams. Ecological applications 15, 137–153.

Kjeldsen, T.R., Rosbjerg, D., 2004. Choice of reliability, resilience and vulnerability estimators for risk assessments of water resources systems. Hydrological Sciences Journal 49, 767. https://doi.org/10.1623/hysj.49.5.755.55136

Kundzewicz, Z.W., Kindler, J., 1995. Multiple criteria for evaluation of reliability aspects of water resource systems. IAHS Publications-Series of Proceedings and Reports-Intern Assoc Hydrological Sciences 231, 217–224.

Maier, H.R., Lence, B.J., Tolson, B.A., Foschi, R.O., 2001. First-order reliability method for estimating reliability, vulnerability, and resilience. Water Resources Research 37, 779–790.

Maity, R., Sharma, A., Nagesh Kumar, D., Chanda, K., 2012. Characterizing drought using the reliability-resilience-vulnerability concept. Journal of Hydrologic Engineering 18, 859–869.

Michalak, A.M., Anderson, E.J., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K., Confesor, R., Daloğlu, I., DePinto, J.V., Evans, M.A., Fahnenstiel, G.L., He, L., Ho, J.C., Jenkins, L., Johengen, T.H., Kuo, K.C., LaPorte, E., Liu, X., McWilliams, M.R., Moore, M.R., Posselt, D.J., Richards, R.P., Scavia, D., Steiner, A.L., Verhamme, E., Wright, D.M., Zagorski, M.A., 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. PNAS 110, 6448–6452. https://doi.org/10.1073/pnas.1216006110

Mondal, M.S., Chowdhury, J.U., Ferdous, M.R., 2010. Risk-based evaluation for meeting future water demand of the Brahmaputra floodplain within Bangladesh. Water resources management 24, 853–869.

Mondal, M.S., Wasimi, S.A., 2007. Evaluation of Risk-Related Performance in Water Management for the Ganges Delta of Bangladesh. Journal of Water Resources Planning and Management 133, 179–187. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:2(179)

Moy, W.-S., Cohon, J.L., ReVelle, C.S., 1986. A programming model for analysis of the reliability, resilience, and vulnerability of a water supply reservoir. Water resources research 22, 489–498.

Peckham, S.D., Gupta, V.K., 1999. A reformulation of Horton's Laws for large river networks in terms of statistical self-similarity. Water Resources Research 35, 2763–2777.

Rigon, R., Rodriguez-Iturbe, I., Maritan, A., Giacometti, A., Tarboton, D.G., Rinaldo, A., 1996. On Hack's law. Water Resources Research 32, 3367–3374.

Schölkopf, B., Smola, A.J., 2002. Learning with kernels: support vector machines, regularization, optimization and beyond. the MIT Press.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464.

Sen, P.K., 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. Journal of the American Statistical Association 63, 1379–1389.

Shreve, R.L., 1966. Statistical law of stream numbers. The Journal of Geology 17–37.

Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. Civ. Eng 101, 1258–1262.

Strahler, A.N., 1952. Hypsometric (area-Altitude) Analysis of Erosional Topography. Geological Society of America Bulletin 63, 1117–1142. https://doi.org/10.1130/0016-7606(1952)63[1117:HAAOET]2.0.CO;2
Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244. https://doi.org/10.1162/15324430152748236
US EPA, 1986. Quality Criteria for Water [WWW Document]. URL http://water.epa.gov/scitech/swguidance/standards/criteria/aqlife/upload/2009_01_13_criteria_goldbook.pdf (accessed 12.7.12).
Vogel, R.M., Bolognese, R.A., 1995. Storage-Reliability-Resilience-Yield Relations for Over-Year Water Supply Systems. Water Resources Research 31, 645–654. https://doi.org/10.1029/94WR02972
Vondracek, B., Blann, K.L., Cox, C.B., Nerbonne, J.F., Mumford, K.G., Nerbonne, B.A., Sovell, L.A., Zimmerman, J.K., 2005. Land use, spatial scale, and stream systems: lessons from an agricultural region. Environmental Management 36, 775–791.
Zaliapin, I., Foufoula-Georgiou, E., Ghil, M., 2010. Transport on river networks: A dynamic tree approach. Journal of Geophysical Research: Earth Surface 115.

A. APPENDIX:

A.1 Climate over the study area:

The study area receives precipitation under the following scenarios (Hirschboeck, 1991): (a) when moisture laden southerly winds from the Gulf of Mexico meet the continental polar wind entering through Canada, (b) when occasionally moisture laden winds from the Pacific ocean make it through the Rocky mountains, and (c) due to the great lakes. As the study area has a large geographic extent, there is significant spatial and temporal variation in precipitation. The average annual precipitation for UMRB is around 950 mm, ORB receives around 1150 mm and MRB receives approximately 1000 mm. While February is typically the driest month over the study region, May and June are among the wettest. Winters are extremely cold with average high of 3.3℃ and average low of about -6℃, spring season is cool with average high of 16.8℃ and average low of 4.5℃, summers are warm with an average high temperature of 28.4℃ and average low temperature of 16.1℃, and finally during fall season, the study region has average high and low temperatures of 18.1℃ and 6℃, respectively.

A.2 Land use and Land cover over the study area:

The land use and land cover map for the study area is shown in Figure A1. The land use land cover data was first obtained from National Land Cover Database 2011 (Homer et al., 2015) and then reclassified into four broad classes – water, urban, agriculture, and forest. The distribution of land use and land cover classes in the three river basins are as follows. For UMRB water is 8%, urban is 9%, forest is 19% and agriculture is 64%. For ORB water is 2%, urban is 10%, forest is 46% and agriculture is 42%. Similarly for MRB, water is 6%, urban is 24%, forest is 17%, and agriculture is 53%. Therefore the chosen study area has agriculture and forests as the dominant class.

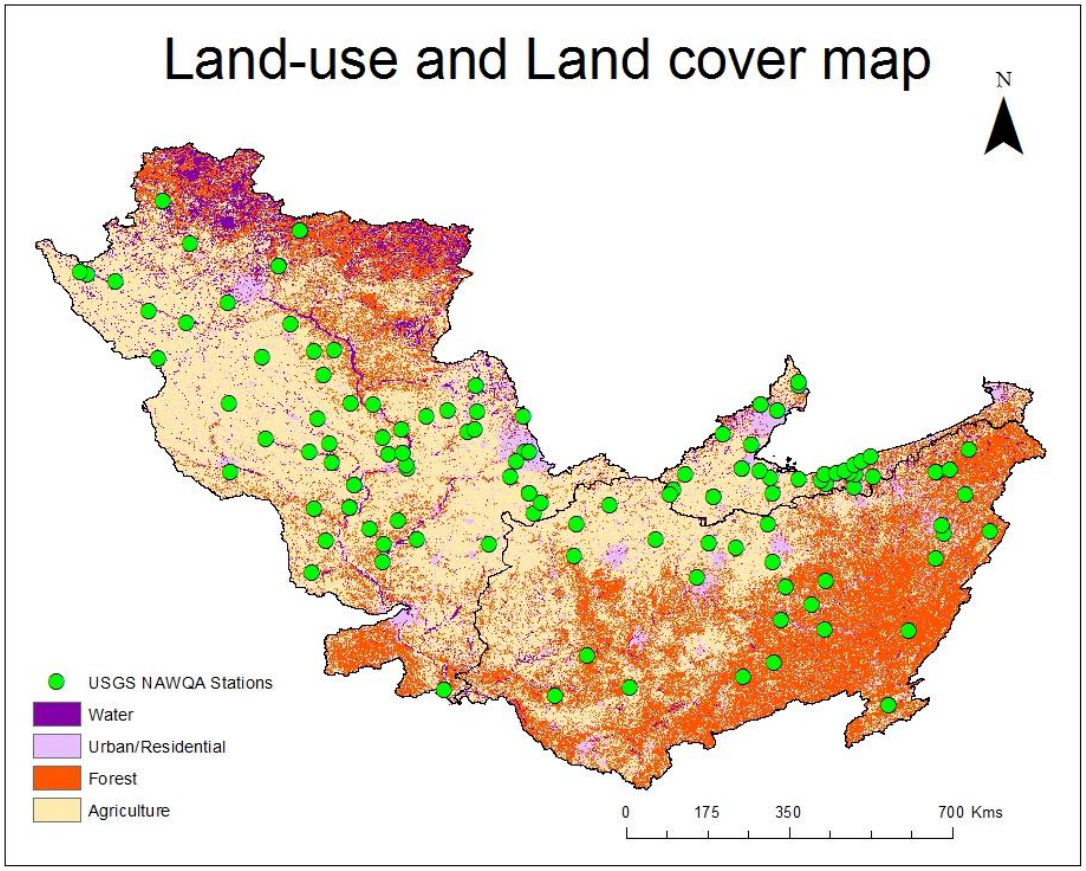

Figure A1: Land use and Land cover map of the study area showing four broad classes of land use – Water, Urban, Forest and Agriculture.

A.3 Distribution properties of risk measures:

A.3.1 Beta distribution:

Figure A2 shows the Beta distribution fit for all risk measures at USGS station 3015500, and we note that the spread is small. For example, reliability values have a spread of 0.01 [0.878 to 0.9], resilience values have a spread of 0.05 [0.61 to 0.66], robustness has a spread of 0.025 [0.598 to 0.624] and watershed health metric has a spread of 0.025 [0.69 to 0.715].

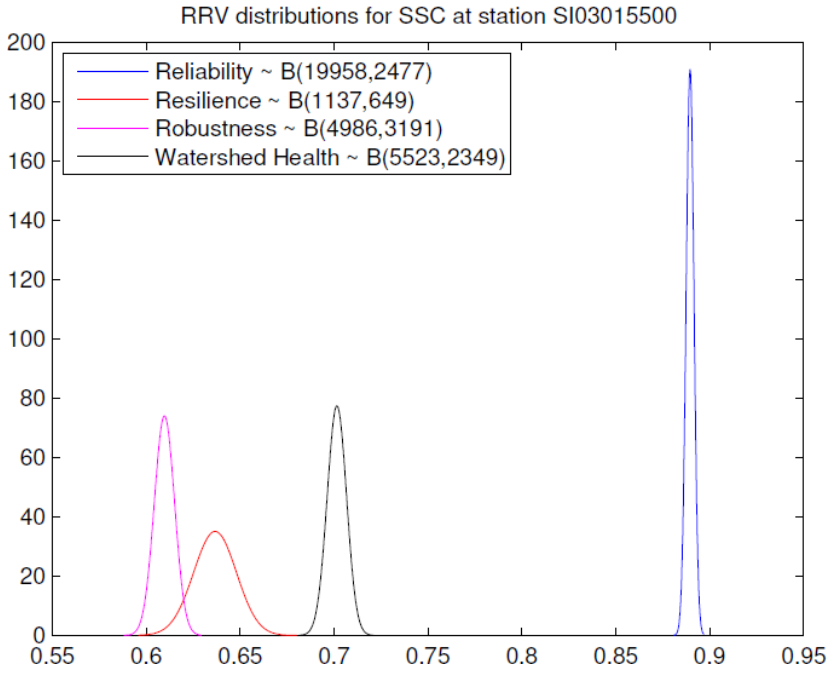

Figure A2: Beta distribution fit to 10000 realizations of reliability, resilience, robustness, and composite Watershed Health measure for Suspended Sediment Concentration (SSC) at USGS station 3015500.

The probability density function (pdf) of the beta distribution is

$$y = f(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x) \quad (1)$$

where $B(.,.)$ is the Beta function. The indicator function $I_{(0,1)}(x)$ ensures that only values of x in the range (0,1) have nonzero probability. The beta distribution has a functional relationship with the t distribution. If Y is an observation from Student's t distribution with v degrees of freedom, then the following transformation generates X, which is beta distributed.

$$X = \frac{1}{2} + \frac{1}{2}\frac{Y}{\sqrt{(v+Y^2)}} \quad (2)$$

i.e., if $Y \sim t(v)$, then $X \sim B\left(\frac{v}{2}, \frac{v}{2}\right)$

Some statistical properties of Beta distributions:

Mean: $E[X] = \frac{a}{a+b}$, also $E[lnX] = \psi(a) - \psi(a+b)$

Where ψ is the digamma function defined as: $\psi(x) = \frac{d}{dx} ln\,\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$, and $\Gamma(x)$ is the gamma function.

Median: $\frac{a-\frac{1}{3}}{a+b-\frac{2}{3}}$, when $a, b > 1$

Mode: $\frac{a-1}{a+b-2}$, when $a, b > 1$

Geometric mean: $\exp(E[lnX]) = \exp(\psi(a) - \psi(a+b))$

A.3.2 Distributional properties of risk measures for Suspended Sediment Concentration:

The distributional properties of risk measures for suspended sediment concentration (SSC) with respect to a numerical target of 30 mg/L were analyzed at all stations in the study area where SSC data were available. For the purpose of brevity we show sample plots from three stations (USGS 05526000, 04193500, 03320500), with each one of them belonging to UMRB, MRB, and ORB, respectively. The Beta distribution was found to be one of the best fit distribution for risk measures

at all stations. Since we are mostly dealing with probabilities, or measures (vulnerability and watershed health) that strictly scale between 0 and 1, we suggest the use Beta distribution to describe their properties. Figure A3a shows the histogram of composite watershed health, at USGS station 05526000 (Iroquois River near Chebanese, IL – located in the Upper Mississippi River Basin), obtained from 10,000 Monte-Carlo realizations. The mean watershed health is about 0.295 and has a tight spread 0.02. The minimum watershed health value is about 0.285 and the maximum watershed health value is about 0.304. Similarly, the distributional properties for R-R-V measures were analyzed for this station. Figure A3b shows the histogram of watershed health, at USGS station 04193500 (Maumee River at Waterville, OH - located in the Maumee River Basin), obtained from 10,000 Monte-Carlo runs. The mean watershed health is about 0.315, with a spread of 0.02 (minimum value = 0.306, maximum value of 0.324). Figure A3c shows the Beta distribution fits (along with parameters) for R-R-V and watershed health at USGS station 03320500 (Pond River near Apex, KY - located in the Ohio River Basin). The spread in the distribution for all risk measures are small, indicating smaller uncertainty in their estimation. This also indicates that we may not really need 10,000 MC runs to get a robust estimate of these risk measures.

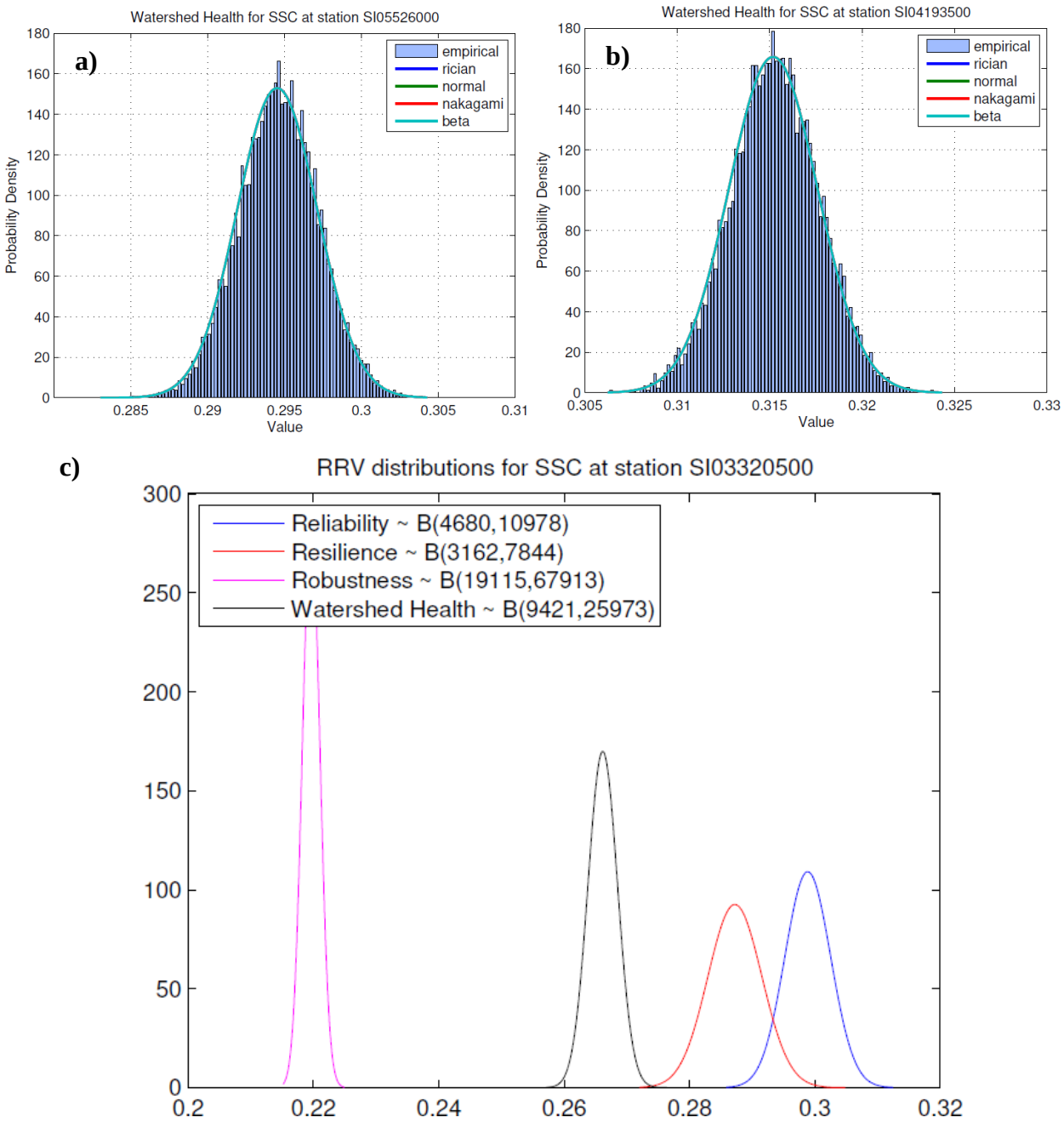

Figure A3 : Four best fit distributions to 10000 realizations of composite Watershed Health measure for Suspended Sediment Concentration (SSC) at USGS stations a) 05526000 in UMRB, and b) 04193500 in MRB. The beta distribution fit to 10000 realizations of reliability, resilience, robustness, and composite Watershed Health measure for SSC at USGS station 03320500 in ORB is shown in panel (c).

A.3.3 Distributional properties of risk measures for Nitrogen:

Distributional properties of risk measures for nitrogen (parameter code: 00631) were analyzed at all stations where this water quality constituent was measured. The numerical standard was set to 10 mg/L and risk measures were computed using reconstructed daily series for this parameter (i.e. for each of the 10,000 Monte-Carlo runs). Figure A4a-b shows the histogram of watershed health at USGS stations 05572000 (Sangamon River at Monticello, IL – located in the Upper Mississippi River Basin) and 04193500 (Maumee River at Waterville, OH - located in the Maumee River Basin). As in the case of SSC, the Beta distribution was found to be one of the best fit distributions (Beta was listed as the 5th best distribution for 05572000, and therefore not included in the figure legend). This was consistent across different metrics, as well as at other stations where this water quality constituent was measured. The probability density function for different risk measures at USGS station 03373530 (Lost River near Leipsic, IN – located in the Ohio River Basin) are shown in Figure A4c. Though the spread in the distribution for resilience is wider compared to other risk measures, it is still comparatively small (about 0.1, with a minimum of 0.8 and maximum of 0.9). The mean watershed health is around 0.88, indicating high level of compliance with respect to the chosen numerical standard of 10 mg/L.

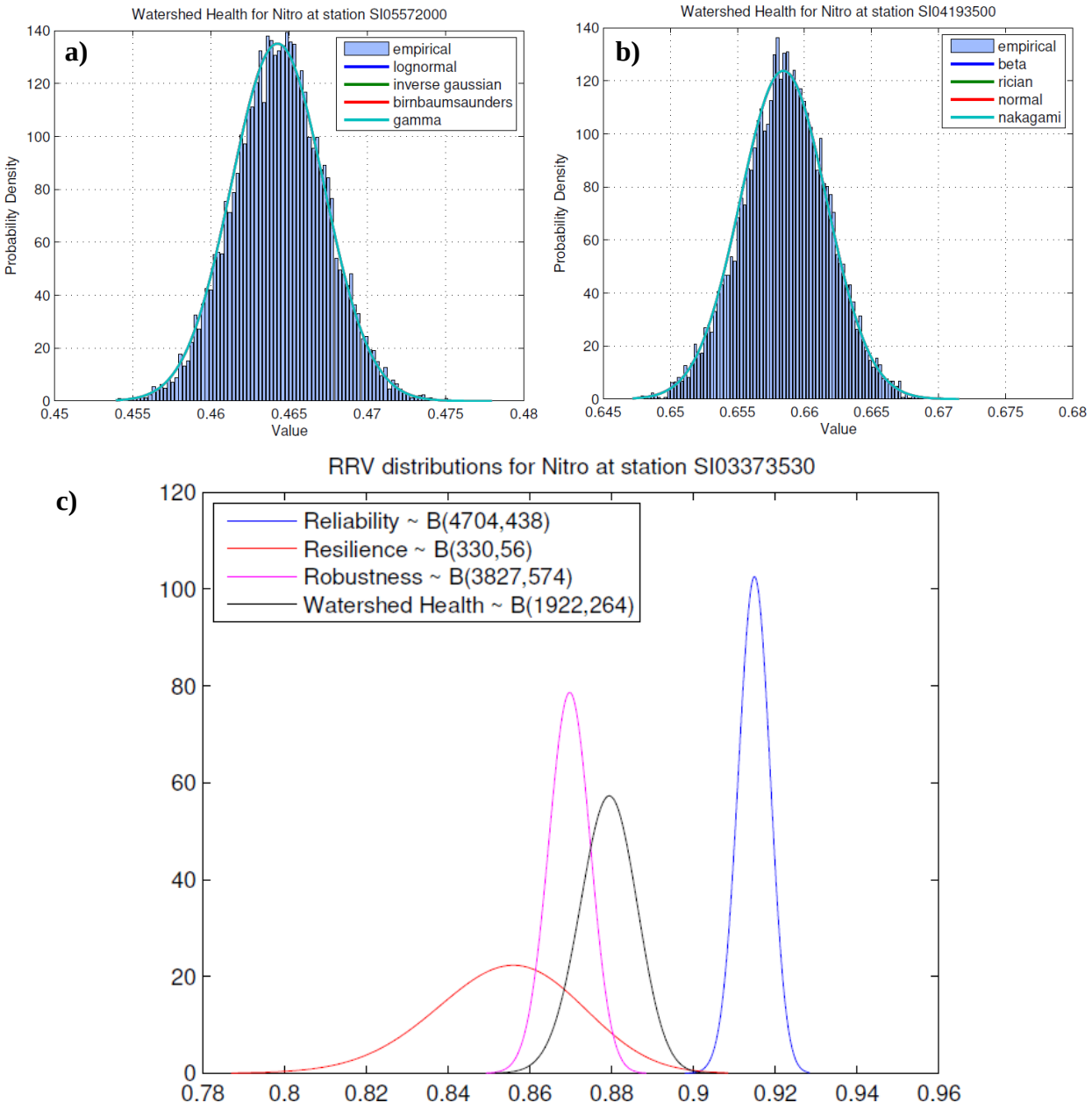

Figure A4: Four best fit distributions to 10000 realizations of composite Watershed Health measure for Nitrogen (parameter code: 00631) at USGS stations a) 05572000 in UMRB, and b) 04193500 in MRB. The beta distribution fit to 10000 realizations of reliability, resilience, robustness, and composite Watershed Health measure for Nitrogen at USGS station 03373530 in ORB is shown in panel (c).

A.3.4 Distributional properties of risk measures for Orthophosphate:

Distributional properties of risk measures for orthophosphate (parameter code: 00671) were also analyzed at all stations where measurements were available. Using a numerical standard of 0.1 mg/L, risk measures were computed for each of the 10,000 (MC-realizations) reconstructed daily series for this parameter. Figure A5a-b shows the histogram of watershed health at USGS stations 05584500 (La Moine River at Colmar, IL – located in the Upper Mississippi River Basin) and 04193500 (Maumee River at Waterville, OH - located in the Maumee River Basin). Though Beta distribution was not among the top-4 distributions that were tested, it still provided very good fit. Therefore, Beta distribution was used for describing the distributional properties of risk measures for orthophosphate. The probability density plot using Beta distribution for different risk measures at USGS station 03274000 (Great Miami River near Hamilton, OH – located in the Ohio River Basin) are shown in Figure A5c. The mean watershed health is around 0.405, indicating moderate level of compliance with respect to the chosen numerical standard of 0.1 mg/L.

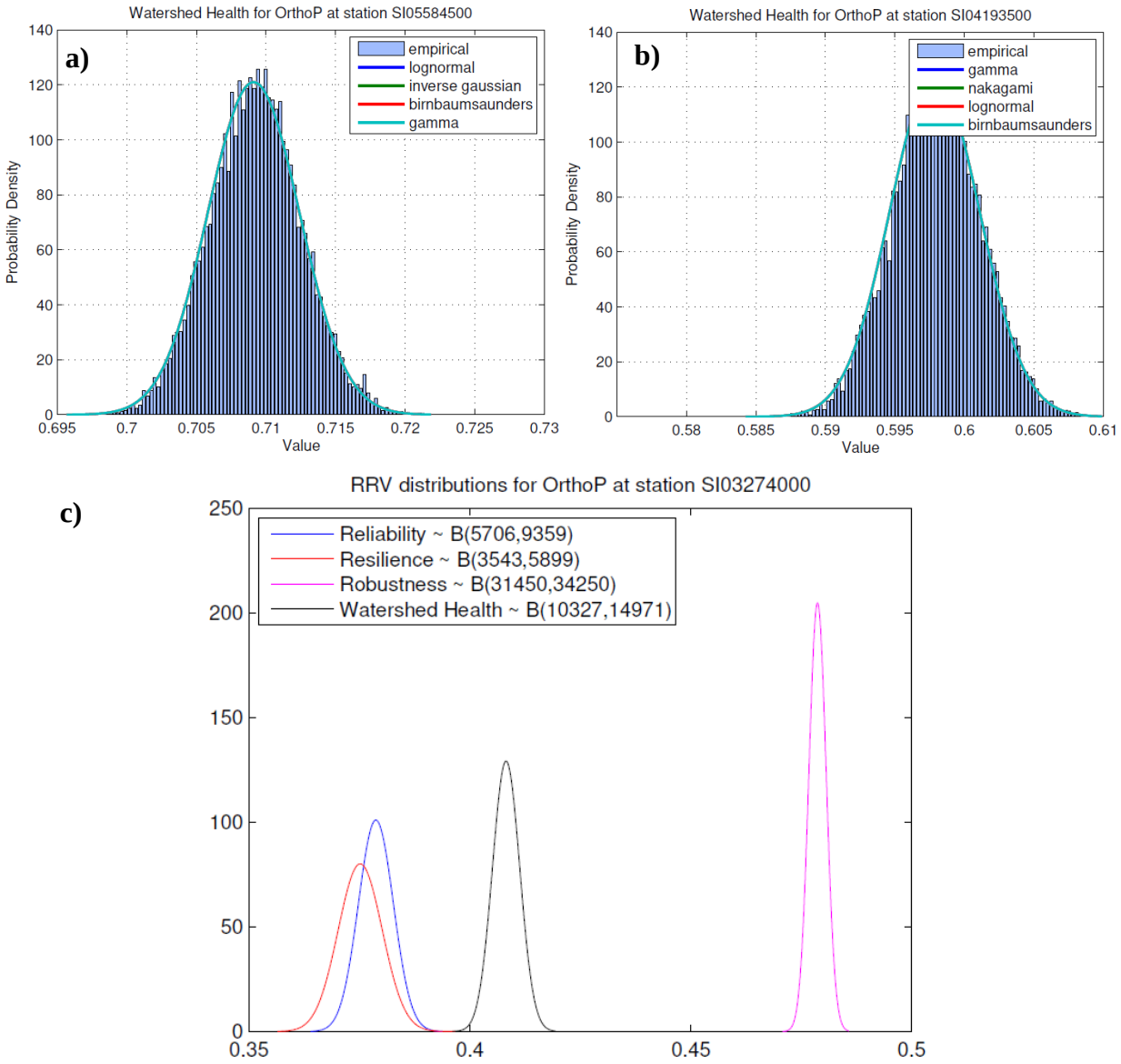

Figure A5: Four best fit distributions to 10000 realizations of composite Watershed Health measure for Orthophosphate (parameter code: 00671) at USGS stations a) 05584500 in UMRB, and b) 04193500 in MRB. The beta distribution fit to 10000 realizations of reliability, resilience, robustness, and composite Watershed Health measure for Orthophosphate at USGS station 03274000 in ORB is shown in panel (c).

A.3.5 Impact of varying WQ standard on risk measures:

The histograms of reliability (1000 MC simulations or repeated model runs with different inputs) with respect to SSC at USGS station 03015500, for different choices of numerical targets (i.e. 7.5 mg/L, 15 mg/L, 30 mg/L, 45 mg/L, 60 mg/L, and 120 mg/L) were then compared (Figure A6). Two features may be noted from Figure A6 namely, a) histograms are tightly bound which indicate that a single MC simulation can provide a good estimate of reliability measure, and b) they are well separated indicating that reliability measures at this station are sensitive to the choice of numerical target. A stringent target of 7.5 mg/L on SSC results in more number of violations causing the reliability measure to drop to a mean value of 0.54. On the other hand, a target of 120 mg/L means that there are fewer violations, leading to high reliability of 0.99 (mean).

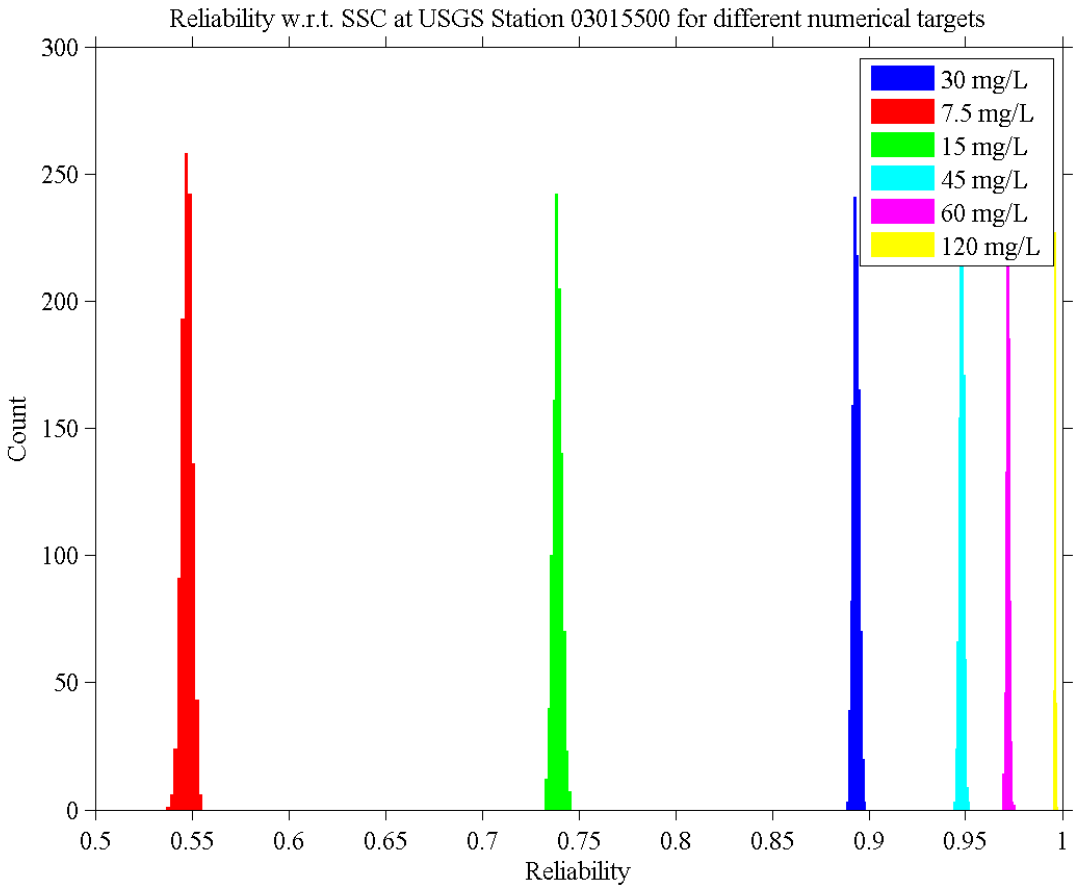

Figure A6: Histogram of reliability with respect to suspended sediment concentration at USGS station 03015500 for different choice of numerical standards (i.e. 7.5 mg/L, 15 mg/L, 30 mg/L, 45 mg/L, 60 mg/L, and 120 mg/L).

A Kolmogorov-Smirnov (KS) test was performed to investigate if the reliability measures (with respect to SSC) obtained for different numerical targets were significantly different compared to commonly used standard of 30 mg/L (US EPA, 1986). Figure A7 shows the results of KS test at USGS stations (denoted as circular markers) over the study region where SSC observations were available. The filled markers indicate that H_0 is rejected as the p-value for KS test was less than the statistical significance level (α) of 0.05. Results were found to be statistically significant at all stations in the study region for smaller numerical targets (7.5 mg/L and 15 mg/L) and at all but one station (USGS 05331833, located in UMRB) for larger numerical targets (45 mg/L, 60 mg/L, and 120 mg/L). Similar results were obtained, when the KS test was repeated for different values of α (i.e. 0.1, 0.025, and 0.01).

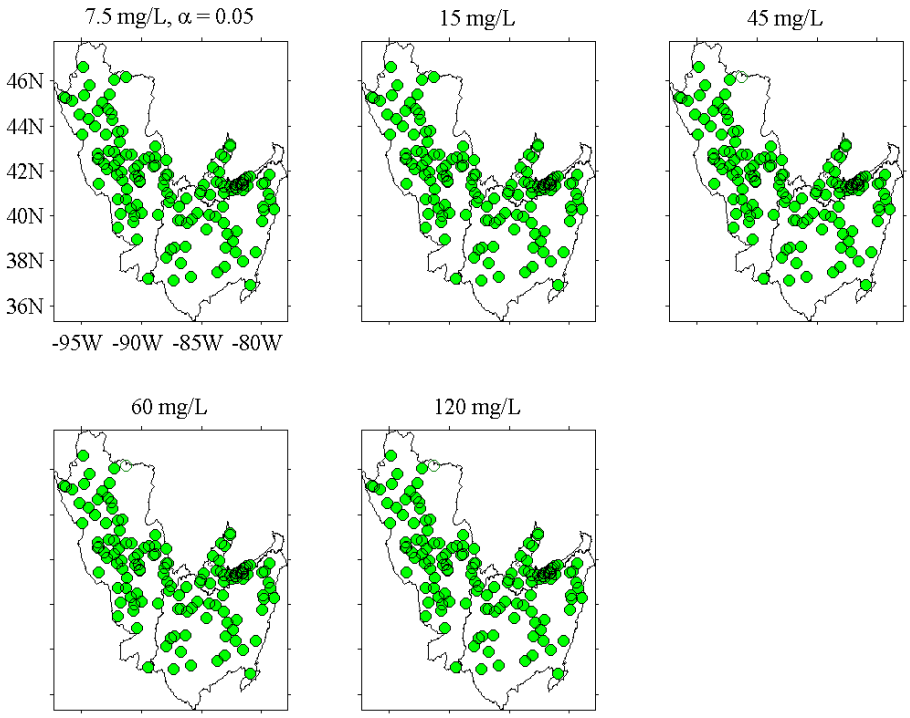

Figure A7: Kolmogorov-Smirnov (KS) test results show whether the samples of reliability for different choices of numerical targets (e.g. 7.5 mg/L, 15 mg/L, 45 mg/L, 60 mg/L, and 120 mg/L) were generated from the same distribution as that of 30 mg/L standard for SSC. The circular markers denote USGS stations where SSC measurements were available. Filled markers indicate that the distributions were different and hollow markers indicate that the distributions were same. Significance level, α, was set to 0.05.

The results of different analyses presented in this study is comparable even within watersheds with varying characteristics as long as the same WQ standards are being used to compute risk measures (e.g. 30 mg/L for SSC), and the results are being interpreted only with respect to water quality within the stream. The consequences of poor (or good) watershed health for example, may however be felt differently depending on the location and water use (e.g. agricultural use, industrial use,

domestic use, sustaining aquatic biodiversity etc). This is irrespective of heterogeneity in watershed characteristics

The KS-test results of reliability measures for nitrogen (see Figure A8) indicate that out of 70 stations where nitrogen data (parameter code 00631) were available, there were 14 stations where H_0 was not rejected when the numerical target was 2.5 mg/L. Similarly, the H_0 of KS-test was not rejected at 21, 28, 28, and 28 stations when the target was 5 mg/L, 15 mg/L, 20 mg/L, and 40 mg/L, respectively. When the numerical target is reduced to 5 mg/L, at several stations (shown as hollow markers in Figure A8) more cases of violations will be picked up when compared to commonly used standard of 10 mg/L (US EPA, 1986). Similarly, when the numerical target is increased from 10mg/L to say 40 mg/L, fewer cases of violations will be picked up. Further, investigation showed that at these stations the reconstructed water quality time series never exceeded the numerical targets (see Figure A10). Similar results were obtained when the KS test was repeated for different significance levels (i.e. α = 0.1, 0.025, and 0.01). There may be other scenarios where we may see similar results: a) when the exceedances are so high that raising or lowering the standard does not affect the results much, or b) the uncertainty of the reconstructed series may be so wide that even by altering the standard we still pick similar number of violations.

Sensitivity results for Orthophosphate (parameter code 00671) were similar to that of SSC. Distributions of reliability measures obtained for numerical targets 0.025 mg/L, 0.05 mg/L, 0.15 mg/L, 0.2 mg/L, and 0.4 mg/L were compared with distribution of reliability measures for the commonly used standard of 0.1 mg/L (US EPA, 1986). Except at one USGS station (04175600) in MRB, at all other stations the H_0 was rejected. Similar distributional findings were found for other risk measures – resilience, vulnerability, and watershed health.

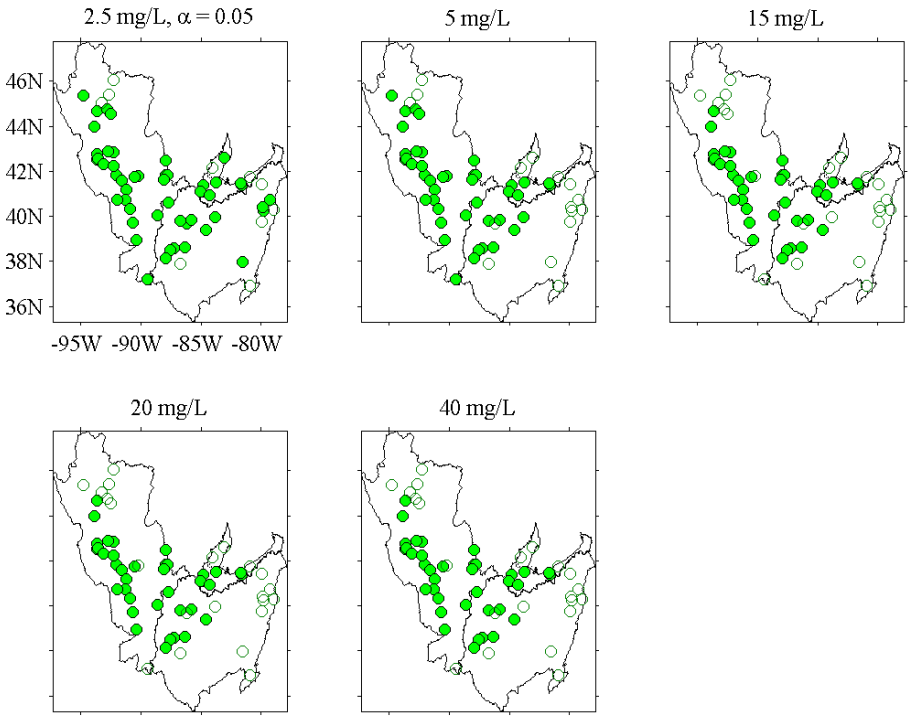

Figure A8: Kolmogorov-Smirnov (KS) test results show whether the samples of reliability for different choices of numerical targets (e.g. 2.5 mg/L, 5 mg/L, 15 mg/L, 20 mg/L, and 40 mg/L) were generated from the same distribution as that of 10 mg/L standard for Nitrate + Nitrite (parameter code 00631). The circular markers denote USGS stations where Nitrate + Nitrite measurements were available. Filled markers indicate that the distributions were different and hollow markers indicate that the distributions were same. Significance level, α, was set to 0.05.

A.4 Scaling behavior of risk measures for Orthophosphate

By setting the water quality standard for phosphorus as 0.1 mg/L (US EPA, 1986), reliability, resilience, vulnerability (and robustness) and the composite watershed health metrics were computed at each station where measurements were available. The scaling relationships of risk measures for phosphorus as a function of stream order was investigated (Figure A9). As only few stations have records of Orthophosphate over ORB (9 stations) and MRB (10 stations), results are only shown for UMRB (30 stations). All the risk measures at UMRB showed a positive trend with

increasing stream orders, although the trend was significant (at $\alpha = 0.05$) only for watershed health index (Sen's slope = 0.0342). The latter is consistent with observed positive trend for suspended sediment concentration given the strong affinity of phosphorus to sediments. The trend results indicate that potential dilution effects caused by increased potential of mixing with waters originating from forested lands within UMRB might be contributing to improved health conditions as we move downstream. This reasoning was confirmed using KS test, and the results are shown in Table A3.

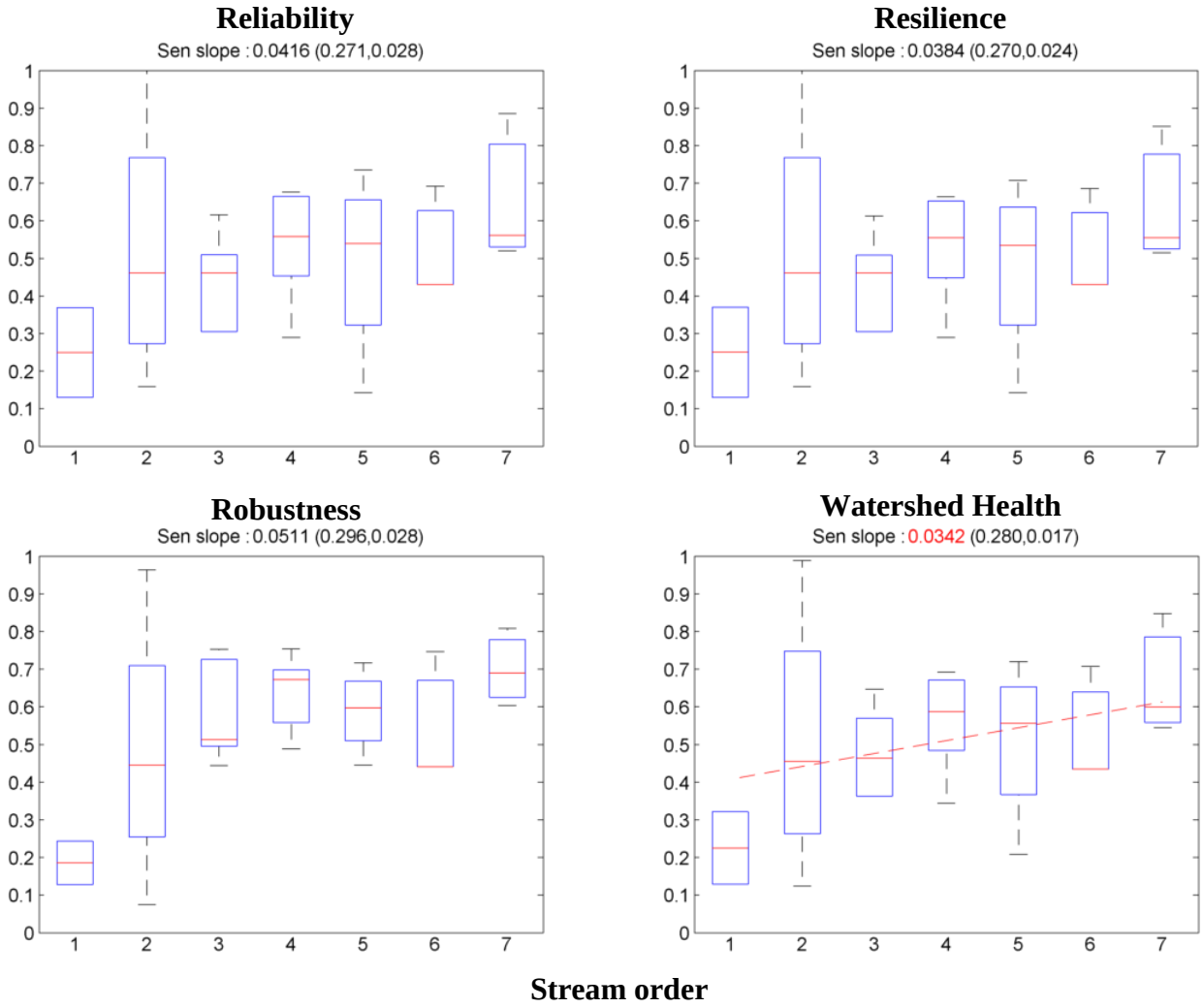

Figure A9: Scaling behavior of risk metrics for Orthophosphate with respect to stream order at UMRB. Sen's slope estimate from Mann-Kendall trend test for mean is shown. The values within the brackets denote Sen's slope estimate from Mann-Kendall trend test for mean values of risk measures on lower- (≤ 2) and higher- (> 2) order streams respectively. Statistically significant trends $(\alpha = 0.05)$ are labeled in red color. Sen's slope is denoted as a red dashed line for cases where trends are significant.

A.5 Spatial distribution of risk measures

Figure A10 shows the spatial distribution of risk metrics for Nitrogen (parameter code: 00631) which can be used to visually identify locations with low watershed health values. Several stations in the upper reaches of UMRB and ORB, and few stations in MRB have high reliability values (equal to 1.0) for nitrogen indicating zero violations with respect to the chosen standard. These stations drain areas with predominant forest land use (see Figure A1). Stations with agriculturally intensive drainage areas have lower reliability, resilience and watershed health (see second column in Figure 6 of main text), and higher values of vulnerability. Specifically, the Mann-Kendall trend test (Figure 6) shows significant negative slope, indicating that watershed health measure for Nitrogen decreases with increase in agricultural land-use. A significant positive trend was observed for watershed health measure versus forest areas, indicating better health metric for Nitrogen when forest land use is high. However, it must be noted that the results are sensitive to the choice of water quality standard (see Figure A6 and Figure A7). If the standard for nitrogen is lower than 10 mg/L, we expect to observe more stations with lower watershed health and vice-versa (Figure A8).

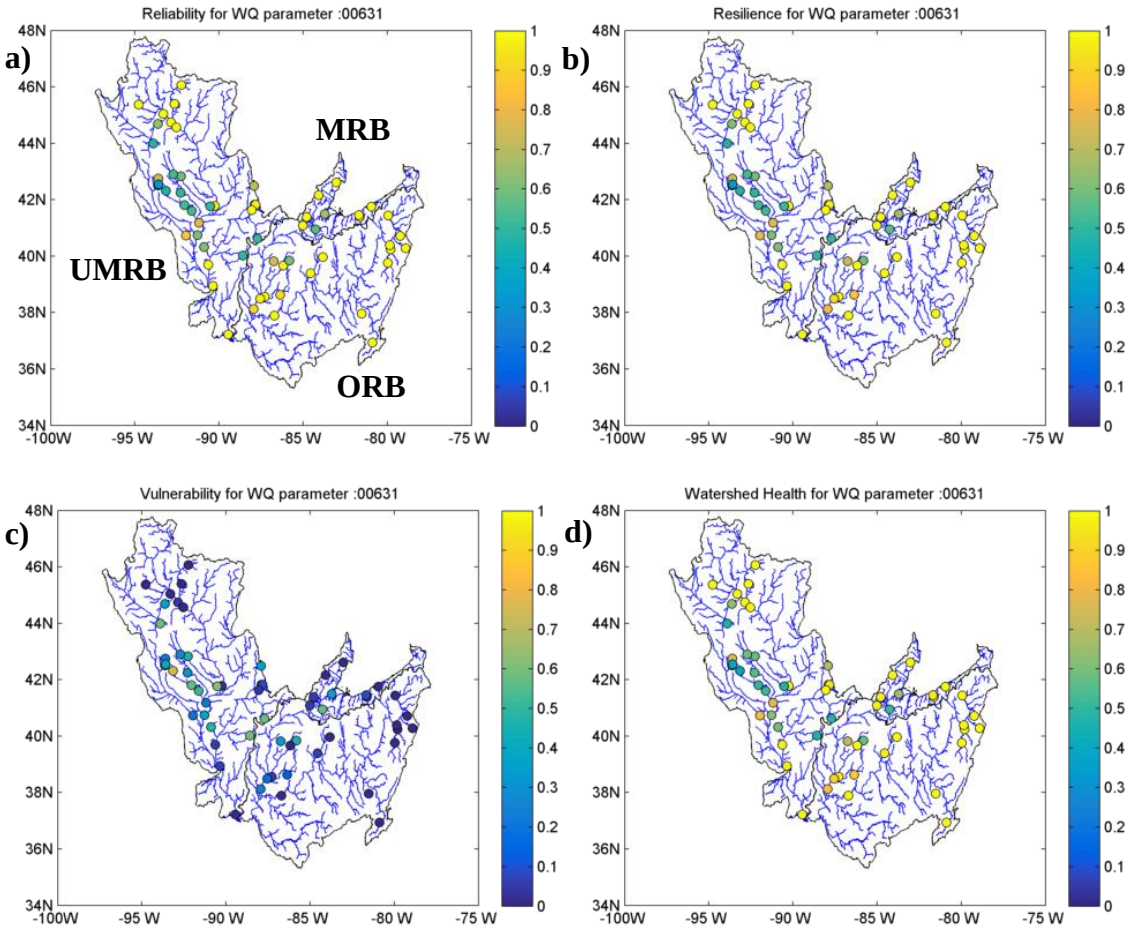

Figure A10: Spatial distribution of a) reliability, b) resilience, c) vulnerability, and d) Watershed Health measures for Nitrite + Nitrate (parameter code: 00631).

Figure A11 shows the spatial distribution of risk metrics for Orthophosphate. Few stations in the upper reaches of the study area with drainage areas that are predominantly of forest land use (see Figure A1), have high watershed health measures indicating smaller number of violations with respect to the chosen water quality standard. A plot of watershed health at stations where Orthophosphate is monitored versus drainage areas under agricultural and forest land use are shown in Figure 6 (third column). Mann-Kendall trend test was performed to ascertain the nature of relationship between watershed health and areas under agriculture and forest land use classes.

In the case of agriculture land use, the trend was negative (not statistically significant) indicating that watershed health measure for Orthophosphate decreases with increase in area under agricultural land use. On the other hand, the trend was positive and statistically significant when comparing watershed health measure and area under forest land use, indicating that watershed health measure typically improves when area under forest land use increases.

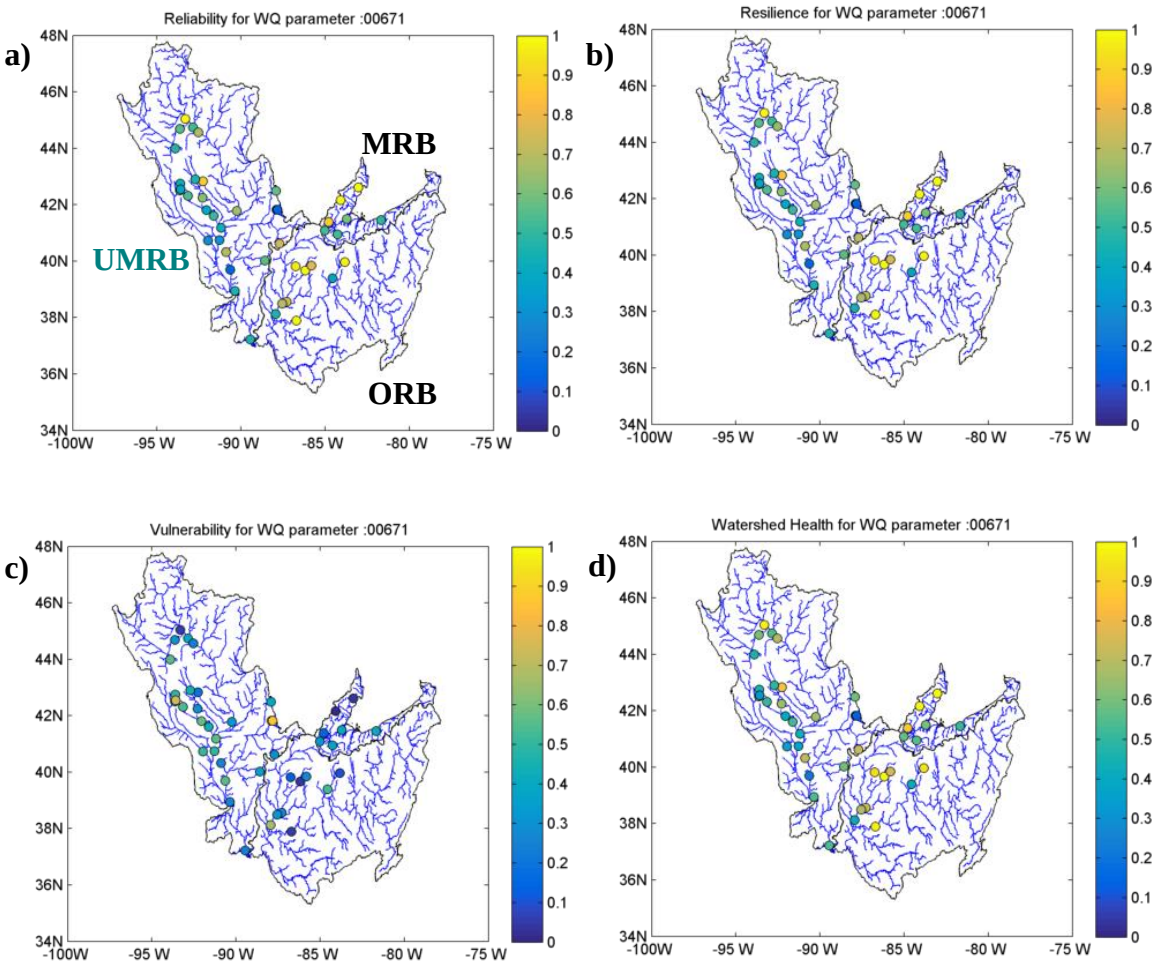

Figure A11: Spatial distribution of a) reliability, b) resilience, c) vulnerability, and d) Watershed Health measures for Orthophosphate (parameter code: 00671).

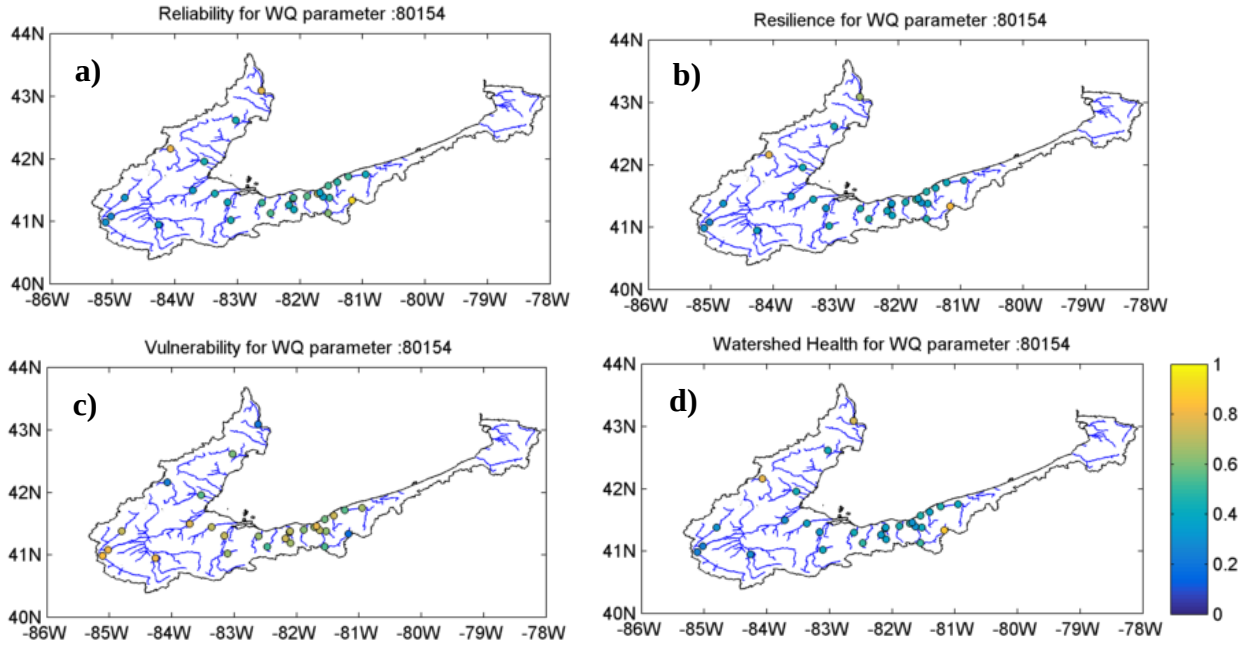

Figure A12: Zoomed map of spatial distribution of a) reliability, b) resilience, c) vulnerability, and d) Watershed Health measures over MRB with respect to Suspended Sediment Concentration (parameter code: 80154).

Table A1: Results of Kolmogorov-Smirnov one-tailed test to investigate if watershed health index of higher-order streams is statistically higher than lower-order streams at 5% significance level (α).

River Basin	**p-values**		
	SSC	**Nitrogen**	**Orthophosphate**
UMRB	0.534	0.432	0.062
ORB	0.627	0.840	0.866
MRB	0.880	NA*	NA*

Table A2: Results of Kolmogorov-Smirnov one-tailed test to investigate if percentage area under agricultural land use for stations located along lower-order streams is statistically higher than stations located along higher-order streams at 5% significance level (α).

River Basin	p-values		
	SSC	**Nitrogen**	**Orthophosphate**
UMRB	0.098	0.091	0.062
ORB	0.220	0.387	0.273
MRB	1.000	NA*	NA*

Table A3: Results of Kolmogorov-Smirnov one-tailed test to investigate if percentage area under forest land use for stations located along higher-order streams is statistically higher than stations located along lower-order streams at 5% significance level (α).

River Basin	p-values		
	SSC	**Nitrogen**	**Orthophosphate**
UMRB	0.030	0.029	0.023
ORB	0.104	0.290	0.099
MRB	0.241	NA*	NA*

*In Tables A1-A3, NA denotes KS test was not carried out because data for lower-order streams were not available at MRB.

A.6 Appendix Bibliography:

Hirschboeck, K.K., 1991. Climate and floods. US Geological Survey Water Supply Paper 2375, 67–88.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogramm. Eng. Remote Sens 81, 345–354.