Statistical power of trend detection methods and sampling schedules for *Escherichia coli* concentrations

Michael Schramm 1,a, Lucas Gregory 2

1 Texas Water Resources Institute, Texas AM AgriLife Research

a Corresponding author, [michael.schramm@ag.tamu.edu](mailto:michael.schramm@ag.tamu.edu)

2020-11-23

 This is the abstract.  
  
 It consists of two paragraphs.

**Keywords** trend detection; *E. coli*, statistical power

**Abbreviations**

|  |  |
| --- | --- |
| Short Name | Descriptive Name |
| *E. coli* | *Escherichia coli* |
| GAM | generalized additive model |
| mL | milliliter |
| MPN | most probable number |
| SWQM | Surface Water Quality Monitoring |
| TCEQ | Texas Commission on Environmental Quality |

# Introduction

Fecal indicator bacteria are used to assess the sanitary quality of water for recreational and water supply purposes. Fecal indicator bacteria themselves are not dangerous but are utilized as an indicator of potential health risks associated with exposure to pathogens associated with fecal matter. *Escherichia coli* (*E. coli*) is used as a fecal indicator bacteria in Texas to assess if streams and other freshwater bodies meet numeric criteria for contact recreation. *E. coli* is a non-host specific bacteria found in the gut of warm-blooded animals, the presence of *E. coli* is used to indicate the potential for recent fecal contamination.

In-stream fecal indicator bacteria concentrations typically follow a log-normal distribution (Novotny 2004). As a result, the Texas Commission on Environmental Quality (TCEQ) biennially evaluates compliance with the in-stream criterion of 126 most probable number (MPN)/ 100 milliliters (mL) using the geometric mean over a seven-year assessment period. The geometric mean is simply a measure of central tendency calculated as the exponential of the arithmetic mean of logarithms:

Simplified, the geometric mean computes the arithmetic mean of *log(y)* and exponentiation returns the mean to the original scale. An alternative approach is to take the *n*th root of the product of *y\_i*. The current assessment approach requires a sample size of 20 over the previous 7-years with an 80% confidence interval that exceeds the 126 MPN/100 mL criterion at the lower bound in order to be determined impaired (TCEQ 2019a). Delistings require 20 samples and the geometric mean below the 126/100 mL criterion. TCEQ (2019a) does not specify how the confidence interval should be calculated. Traditional methods multiply a critical value (obtained from the standard normal distribution or Student’s t-distribution) by the standard error. Alternatively, the confidence intervals can be obtained by parametric bootstrap methods (Wilcox 2013). At 20 samples the method is fairly robust for estimating exceedance of the water quality criterion.

As of 2018, TCEQ identified 237 water bodies impaired due to elevated fecal indicator bacteria (TCEQ 2019b). Total Maximum Daily Loads and Implementation Plans or Watershed Protection Plans are developed for these impaired water bodies to address potential fecal indicator bacteria sources. As part of these plans, trend analysis is typically conducted to assess if bacterial concentrations have increased or decreased over time. Two common methods for assessing statistical significance of monotonic trends are the modified Mann-Kendall test and linear regression on fecal indicator bacteria concentration values (Helsel and Hirsch 2002; Yue and Wang 2002).

Yue and Wang (2002) described the calculation of the Mann-Kendall test and the modifications for correlated data. In short, when the Mann-Kendall test statistic, *S* is negative, newer values tend to be smaller than older values and indicate a downward trend. A small absolute value of *S* indicates no trend. The *P* value of the test statistic is estimated using the normal cumulative distribution function. The null hypothesis of the Mann-Kendall test is that there is no trend.

Simple linear regression on log-transformed *E. coli* concentrations are also suitable for identifying trends. In order to assess presence of a trend, the following linear regression is used:

where *y* is *E. coli* concentration, *β0* is the intercept, *β1* is the coefficient of time variable *x*, and ε are model residuals assumed normally distributed around mean zero. If linear regressions are utilized to assess *E. coli* trends, the analyst should plot residuals to ensure model residuals meet assumptions of heterogeneity and normal distribution around mean zero.

Both the Mann-Kendall test and linear regression are straight forward methods for water quality analysts to apply and assess trends in *E. coli* concentrations. They are well accepted and have routines available in most statistical software. However, general guidance is not available for the number of samples required to detect given effect sizes. Current guidance for assessment of attainment of the water quality criterion (20 samples over 7-years) is adequate given the ability to estimate confidence intervals for the geometric mean calculation. As a result, many monitoring programs across the state utilize quarterly routine sampling regimes, which equate to approximately 4 samples per year or 28 samples over a 7-year assessment period. Reporting the results of trend detection test implies the test has the statistical power to detect trends of certain magnitudes. However, that information is rarely reported and unlikely that it is routinely calculated by water quality analysts. Therefore, there is considerable uncertainty if monitoring schedules (especially those designed around quarterly monitoring) used across the state are adequate for detecting trends in fecal indicator bacteria.

Statistical power refers to the probability that a statistical test rejects the null hypothesis when the alternative hypothesis is actually true. In the case of the discussed trend tests, power is the probability that the null hypothesis of no trend is rejected when there is in fact a trend in the data. Statistical power is a function of pre-assigned significance level (α), effect size, sample size, and variance within the time series (Yue et al. 2002). First, a meaningful effect size must be determined. The effect size might be biologically meaningful or informed by stakeholder input. Statistical power can be determined for a range of sample size, significance levels, effect sizes and sample variance. The purpose of this article is to provide some guidance and context in determining monitoring frequency for trend analysis of fecal indicator bacteria, specifically *E. coli*. First, we estimate the statistical power of Mann-Kendall and linear regression trend tests at sampling sites across the state using Monte Carlo simulation. Second, we describe functional relationships between the statistical power of the two trends tests, effect size, variance, and number of samples across sampling sites in Texas.

# Methods

## Data

TCEQ Surface Water Quality Monitoring (SWQM) site information and associated *E. coli* samples collected during the 7-year period from January 2012 through December 2019 were obtained from the Water Quality Portal (<https://www.waterqualitydata.us/>) using the dataRetrieval package in R (De Cicco et al. 2018; R Core Team 2019). Data was restricted to river or stream sampling sites, and SWQM sites with fewer than 1 sample per year were removed from analysis. In total, *E. coli* data was assessed from 984 SWQM sites.

## Statistical Power Computation

The significance level, α, is the probability of rejecting the null hypothesis when it is true (Type I error). The probability of accepting the null hypothesis when it is false is a Type II error (*β*). The statistical power of a test is the probability of rejecting the null hypothesis when the alternative hypothesis is true and is equal to 1 - *β*. A power of 0.80 is typically considered appropriate, which equates to a 20% likelihood of encountering a Type II error. If sampling from a population where the null hypothesis is false, power is calculated as:

where *N* is the total number of tests and *Nrejected* are the total number of times the test rejected the null hypothesis.

For each SWQM site, Monte Carlo simulation was used to observe the statistical power of the Mann-Kendall and linear regression test for detecting trends (Sigal and Chalmers 2016). The simulation generates 1000 independent log-normal distributed time series samples per evaluated effect size for every SWQM site using the site specific log-transformed mean and standard deviation. Effect sizes were induced by reducing the annual log-transformed mean over the 7-year sampling period by 5, 10, 20, 40, and 80 percent. In total, 4.92 million simulations were run per trend detection method. Significance level, α, was set at 0.10. The Mann-Kendall test and linear regression is applied to each simulation sample and the number of times the tests correctly reject the null hypothesis (*Nrejected*) are tabulated.

## Describing Functional Relationships

The functional observed relationships between variance, sample size, and effect size at SWQM sites across Texas were described and visualized using generalized additive models (GAMs). GAMs are a flexible semi-parametric approach for modeling non-linear relationships between variables (Wood 2011; Wood 2017). GAMs were fit for each set of effect sizes with the mgcv package in R (Wood 2017). The general form of the GAM equation was:

where *s* is a spline based smoothing function applied to each covariate, *cv* is the sample Coefficient of Variation (CV), and *sample size* is the annual number of samples collected at the site. CV is a method of measuring the spread of a distribution relative to the size of the mean; specifically, it is ratio of the standard deviation to the mean. The response variable *Power* is the statistical power within the interval [0,1]. GAMs were fit using the beta regression family and the logit link function. GAMs were not used to specifically test for relationships but only to describe and visualize the known functional relationships that influence statistical power of trend tests as they apply to SWQM stations.

# Results

## Monitoring Frequency

Out of the 984 evaluated SWQM sites, 329 were located in water bodies with a TMDL. SWQM sites located on water bodies without a TMDL were generally sampled 3 to 4 times per year (Figure ). SWQM sites with a TMDL skewed higher, with a peak at 9 times per year and smaller peaks at 4 and 6 times per year. This suggests that increased monitoring efforts are being targets towards sites where planning efforts have been implemented. Similarly, the *E. coli* geometric mean skewed higher at sites with a TMDL (Figure ).

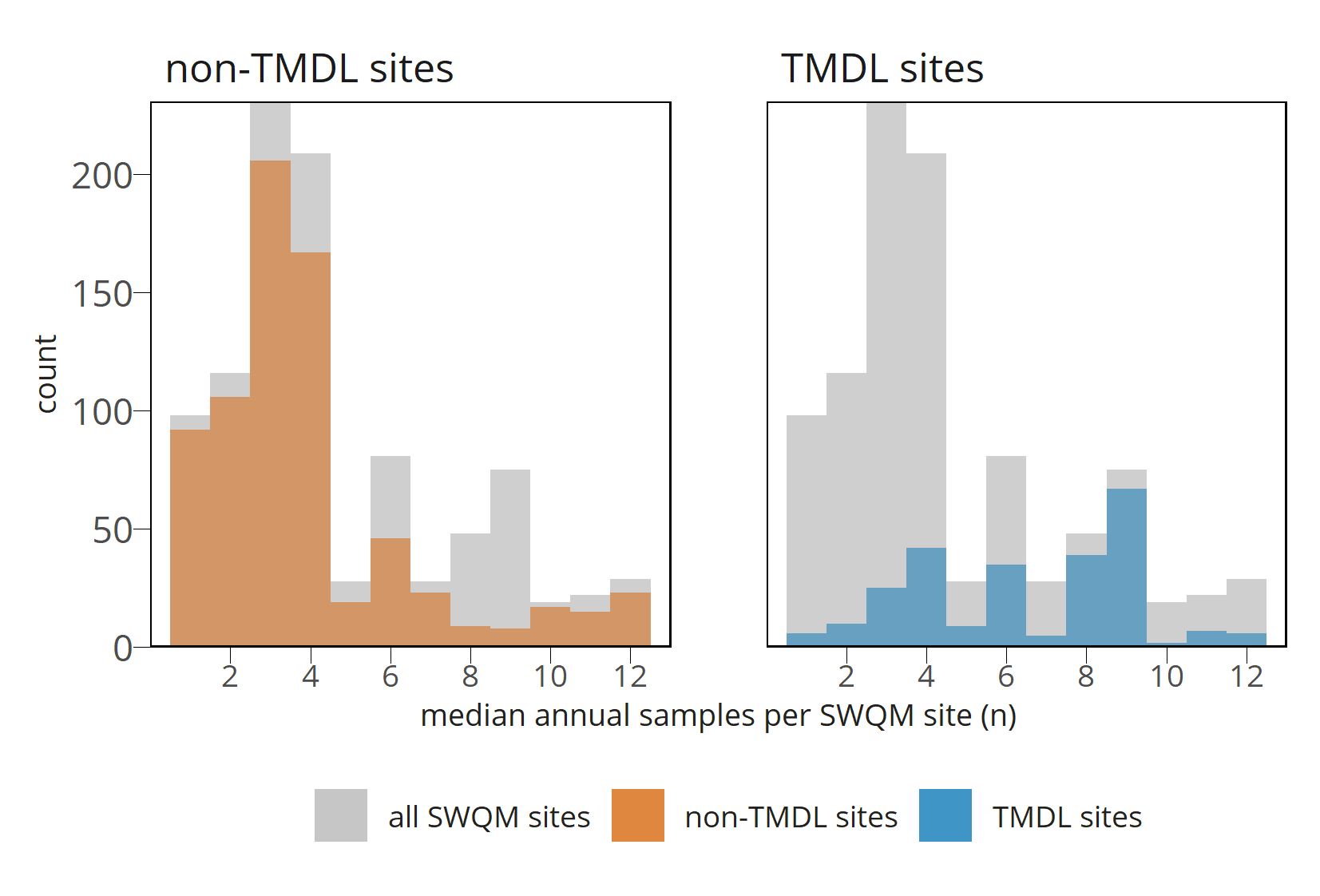


Figure : Histograms of annual *E. coli* sampling distribution for TMDL and non-TMDL SWQM sites across Texas.

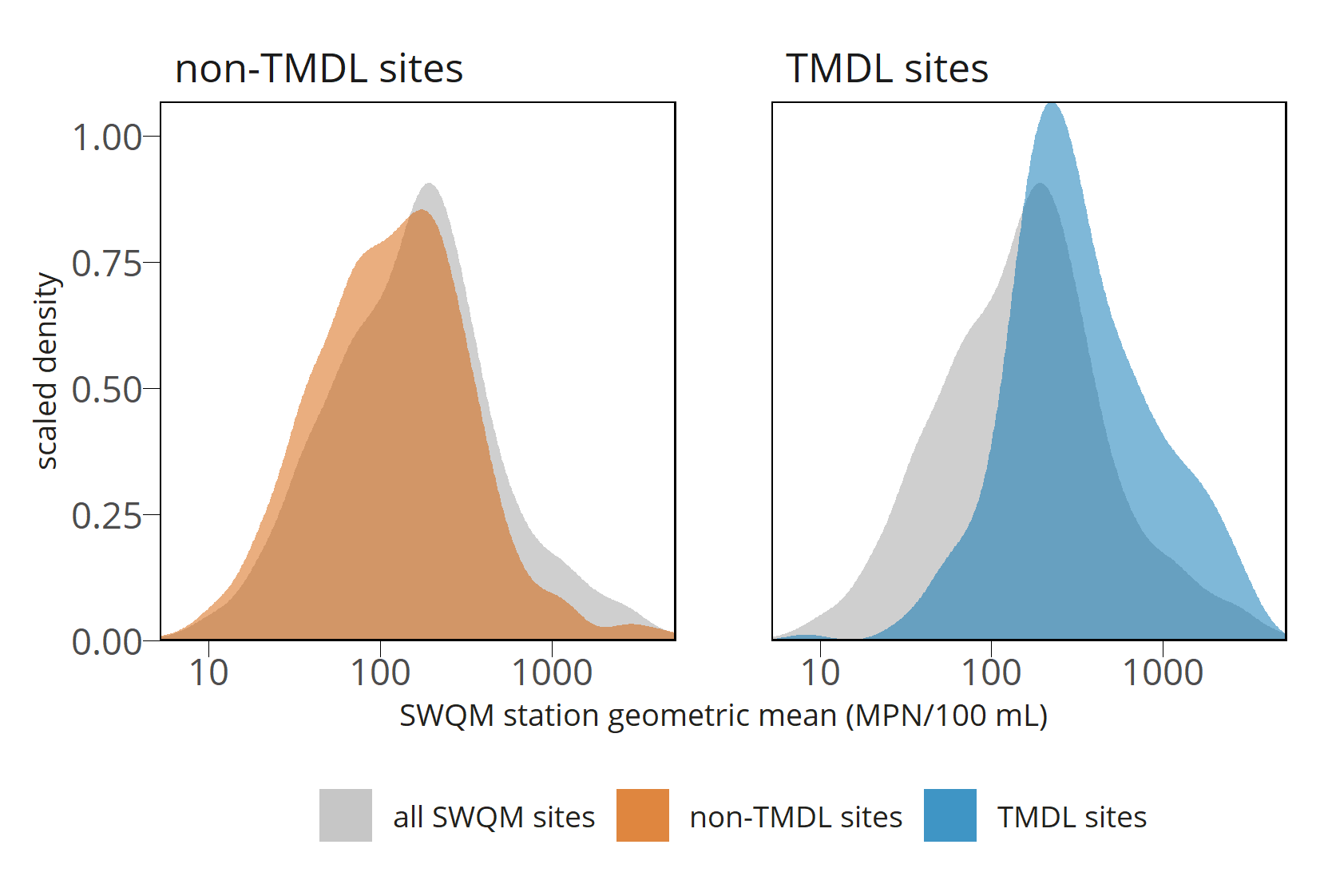


Figure : Scaled density plots of of *E. coli* geometric mean distribution for TMDL and non-TMDL SWQM sites across Texas.

## Estimated Statistical Power at SWQM Sites

At small effect sizes (5-10 percent decrease in *E. coli* concentration), all SWQM stations have relatively low statistical power (<0.30) for detecting effects using the Mann-Kendall test (Figure ). With a 40 percent decrease in *E. coli* concentrations, a large proportion of TMDL sites have sufficient statistical power for trend detection using the Mann-Kendall test. Adequate power is achieved for most sites in both groups at 90 percent effect sizes using the Mann-Kendall test.

A similiar pattern at low effect sizes is evident when using GLMs. At 5, 10, and 20 percent effect sizes SWQM sites showed low statistical power for detecting trends using GLMs (Figure ). At 40 and 80 percent decreases in *E. coli* concentration, only a small proportion of sites exhibit adequate statistical power using GLMs.

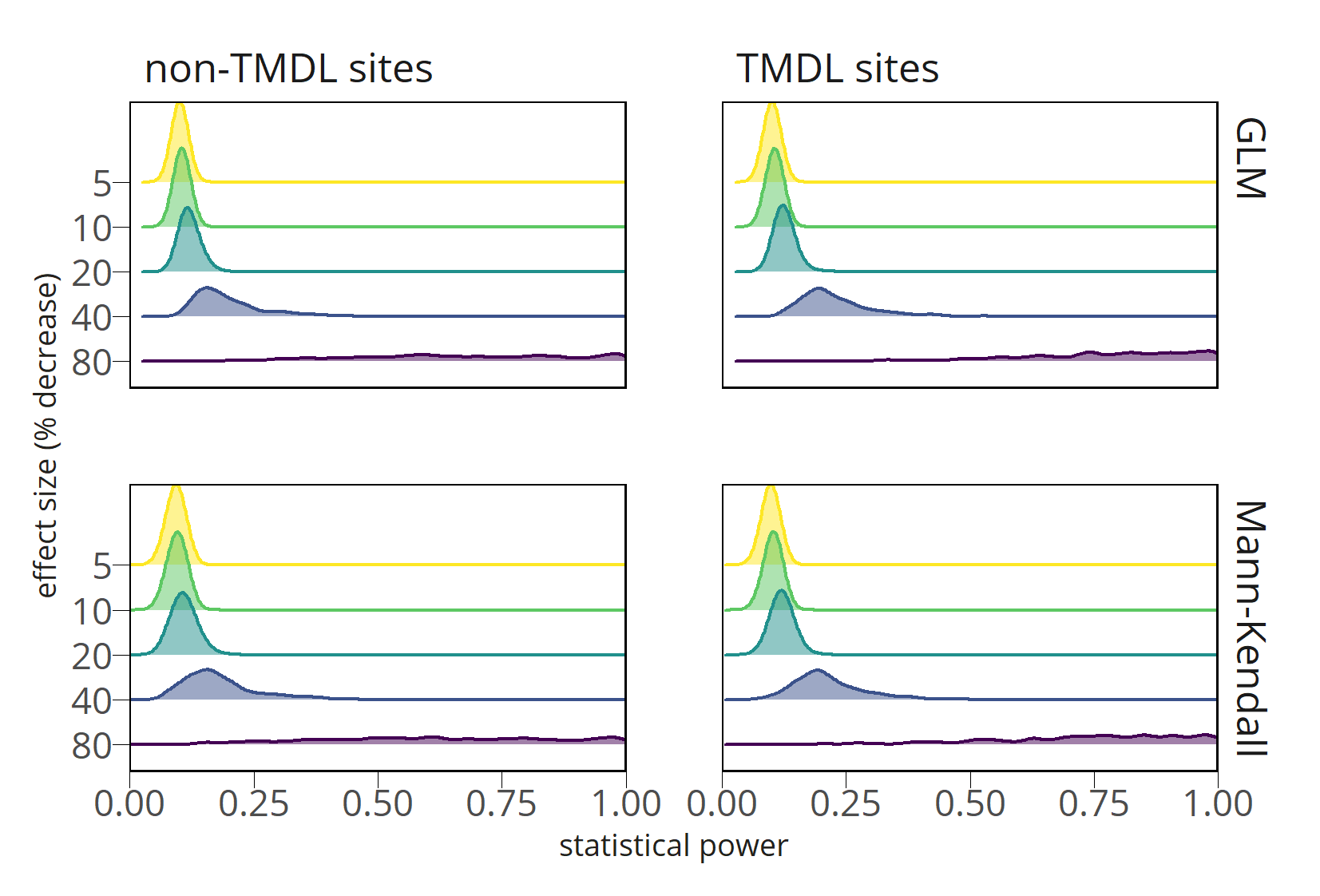


Figure : Scaled density plots of Mann-Kendall and GLM statistical power distribution for TMDL and non-TMDL SWQM sites as a function of effect size at ɑ = 0.1. Individual curves represent the scaled density estimate of statistical power values calculated for SWQM sites at a given effect size (y-axis values).

\*\* add table reporting summary of average sample size, mean (cv), statistical power \*\*

## Functional Relationships

Figure… indicates both tests are sufficiently powerful to detect trends at large effect sizes and with sufficient samples. At the median sample CV (1.94), six samples per year are required to achieve 0.8 power to detect a trend with at least an 80 percent decrease in *E. coli* concentrations over 7-years. Marginal increases in statistical power are observed with more than eight annual samples. Notably, neither method exhibit sufficient statistical power at smaller effect sizes.

# Discussion

# Bibliography

De Cicco LA, Lorenz D, Hirsch RM, Watkins W. 2018. DataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services. Reston, VA: U.S. Geological Survey; U.S. Geological Survey. <https://code.usgs.gov/water/dataRetrieval>.

Helsel D, Hirsch R. 2002. Statistical methods in water resources. U.S. Geological Survey (Techniques of water-resources investigations of the United States Geologic Survey). <http://water.usgs.gov/pubs/twri/twri4a3/>.

Novotny V. 2004. Simplified databased total maximum daily loads, or the world is log-normal. Journal of Environmental Engineering. 130(6):674–683. doi:[10.1061/(ASCE)0733-9372(2004)130:6(674)](https://doi.org/10.1061/(ASCE)0733-9372(2004)130:6(674)).

R Core Team. 2019. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Sigal MJ, Chalmers RP. 2016. Play it again: Teaching statistics with Monte Carlo simulation. Journal of Statistics Education. 24(3):136–156. doi:[10.1080/10691898.2016.1246953](https://doi.org/10.1080/10691898.2016.1246953).

TCEQ. 2019a. 2016 guidance for assessing and reporting surface water quality in Texas. <https://www.tceq.texas.gov/assets/public/waterquality/swqm/assess/16txir/2016_guidance.pdf>.

TCEQ. 2019b. Executive summary 2018 Texas integrated report for Clean Water Act Sections 305(b) and 303(d). Austin, TX: TCEQ. <https://www.tceq.texas.gov/assets/public/waterquality/swqm/assess/18txir/2018_exec_summ.pdf>.

Wilcox R. 2013. Introduction to robust estimation and hypothesis testing. Third. Academic Press Elsevier. <https://doi.org/10.1016/B978-0-12-386983-8.00015-9>.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 73(1):3–36. doi:[10.1111/j.1467-9868.2010.00749.x](https://doi.org/10.1111/j.1467-9868.2010.00749.x).

Wood SN. 2017. Generalized additive models: An introduction with R. 2nd ed. Chapman; Hall/CRC.

Yue S, Pilon P, Cavadias G. 2002. Power of the Mann–Kendall and Spearman’s rho tests for detecting monotonic trends in hydrological series. Journal of Hydrology. 259(1):254–271. doi:[10.1016/S0022-1694(01)00594-7](https://doi.org/10.1016/S0022-1694(01)00594-7).

Yue S, Wang CY. 2002. Regional streamflow trend detection with consideration of both temporal and spatial correlation. International Journal of Climatology. 22(8):933–946. doi:[10.1002/joc.781](https://doi.org/10.1002/joc.781). <http://doi.wiley.com/10.1002/joc.781>.