MT/WY Continuous DO – Equivalence Testing

# Introduction

Dissolved oxygen (DO) criteria in Montana and Wyoming are assessed against 30-day mean, 7-day mean, and 7-day mean minimum metrics. To optimize staff and equipment resources, it may be of interest to deploy continuous sondes for fewer than 7 and/or 30 days at a given site while still obtaining an accurate determination of the support or non-support of the DO criteria. Thus, the objective of this analysis was to identify the minimum number of days a sonde could be deployed in a stream to approximate the DO criteria metrics.

# Data

Continuous DO data from streams were used for this analysis. Data files were provided by MT, WY, and EPA. Additional files provided by EPA from another dissolved oxygen project (“Technical Support for Continuous Monitoring and Assessment Document”) were also used in this analysis. Data were observed from 112 sites across 12 years (2010 to 2021) (see Appendix Table 7). The criteria for inclusion of a site in the analysis was the availability of continuous DO data. Samples were typically taken every 15 or 30 minutes, depending on the site.

The continuous data were visually screened for erratic values, including brief (i.e., less than a day) spikes or brief dips to zero. Inconsistent and abrupt shifts in the DO time series plot could be due to equipment issues at the time of those samples. Erratic values were removed where identified and replaced with NA.

Rolling DO metrics were calculated. For data processing, the samples were filtered for the months of May through September to enable calculation of 30-day rolling mean values from June through August. First, the continuous data were aggregated to the daily timestep (both daily average and daily minimum). For each day, the DO metrics were then calculated on a rolling basis. The 30-day mean, 7-day mean, and 7-day average of daily minimums (“7-day avg min”) metrics were calculated with a centered aggregation window. Criteria were not calculated for windows with missing days.

This analysis involves comparing rolling metric values calculated using a shorter duration against values calculated using the full 7- or 30-day duration. For example, comparing a 7-day mean value vs a 5-day mean value, with both durations centered on the same date. The analysis will use paired data from a given month to assess whether the 5-day mean can be considered equivalent to the 7-day mean. To support this analysis, rolling metrics with a shortened duration of 1 to 6 days were calculated alongside the 7-day mean and 7-day avg min. Likewise, rolling metrics with a shortened duration of 1 to 21 days were calculated alongside the 30-day mean.

Data processing and analyses were performed using R (R Core Team 2023). AR1 models were run using the “MASS” package (Venables and Ripley 2002).

# Methods

The research question is “How many continuous days are sufficient to approximate a 7-day mean, 7-day avg min, and a 30-day mean?” To answer this question, the full duration metrics were compared to shortened duration metrics. Full and shortened duration metrics were both calculated for each day in the study period (June to August). As such, the data were considered *paired* because there were two values (full and shortened) for each day. Specifically, an equivalence test was run on the paired data. An equivalence test (Berger and Hsu 1996) is similar to a traditional hypothesis test, except the null and alternative hypotheses are flipped. This is necessary because the burden of proof must always be on the alternative hypothesis.

For example, a common method to compare two groups is to use a t-test, which has the hypotheses:

where Δ is the difference between the “Full” and “Shortened” datasets. If the p-value is below alpha (typically 0.05), then we reject the null hypothesis and say there is evidence to support the alternative hypothesis that the two groups are *significantly different*. *Significantly different* means that the difference between the groups is greater than zero, when comparing the confidence interval around the test statistic.

If the p-value is > 0.05, then we say that there is not sufficient evidence to reject the null hypothesis, which states that the two datasets are *not* significantly different. Note that this is **not** the same as saying there is evidence that the two groups are equivalent. To make such a statement, we must flip-flop the hypotheses as such:

Now, if there is sufficient evidence for the alternative hypothesis, we can reject the null hypothesis and say that the two datasets are *statistically equivalent*. Equivalence testing has the additional requirement that we must specify an “equivalence margin” that quantifies how close is “close enough” to be considered equivalent:

where θ is the equivalence margin. The equivalence margin can be thought of as “guard rails”. To determine that the two datasets are equivalent, the interval around our test statistic Δ must be within these guard rails.

The difference between significantly different and statistically equivalent is illustrated in Figure 1. Below are confidence intervals (CI; horizontal lines) from four separate analyses (*A* through *D*). The equivalence margin is set to ±0.5 and represented by vertical dashed lines. Analyses *A* and *C* are each statistically equivalent, because their CIs are fully within the equivalence margin (guard rails). That is, the maximum differences are within the limits defined by the equivalence margin. Analysis *B* is statistically different because the CI does not cross zero. Analysis B is also not equivalent, because its CI extends outside the equivalence margin. Analysis *D* is not statistically different (it does cross zero) and is also not statistically equivalent (it extends past the equivalence margin). Additional information on equivalence testing can be found in Walker and Nowacki (2010) and Lakens et. al. (2018).

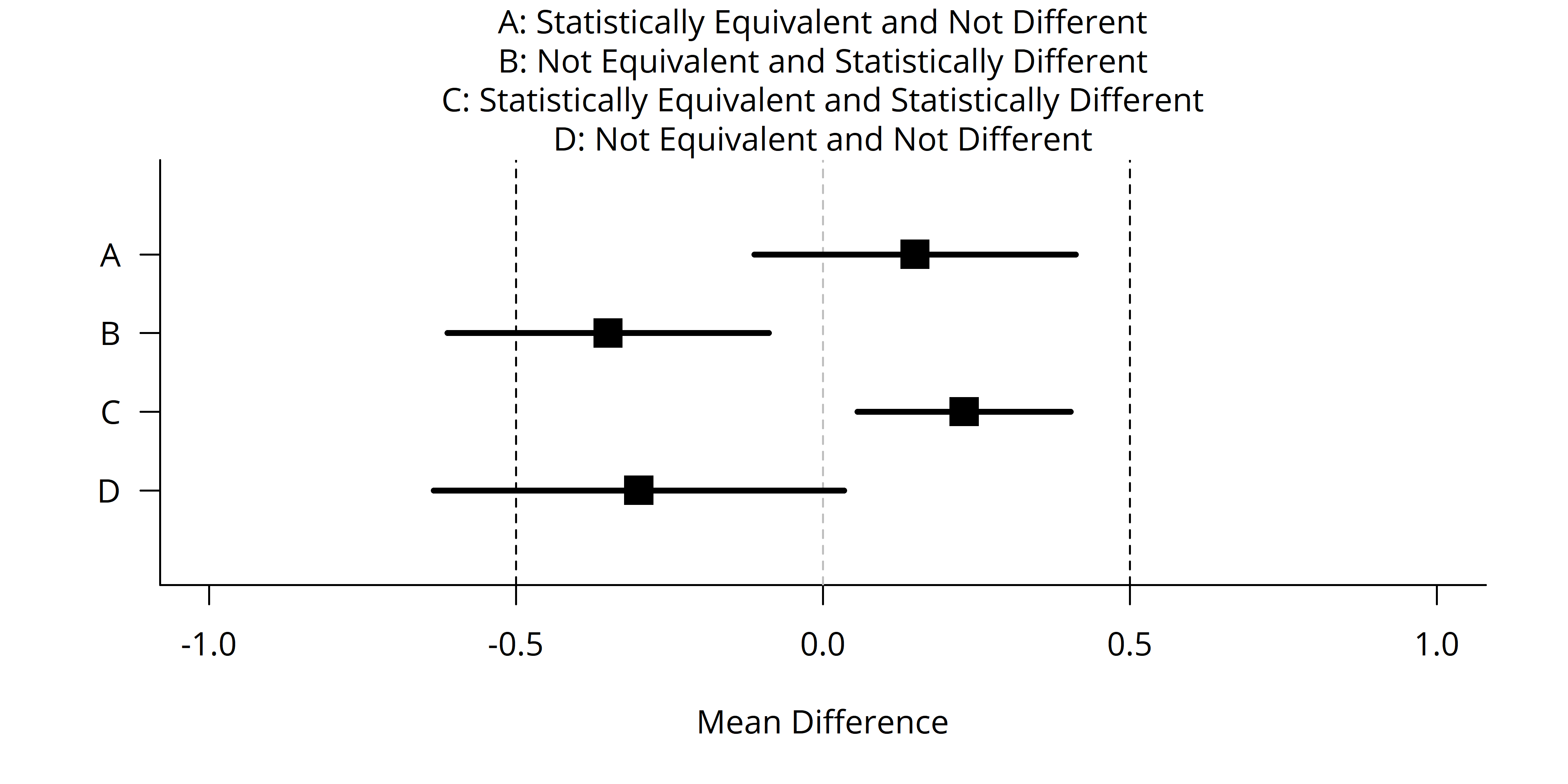


Figure . An illustration comparing standard hypothesis testing and equivalence testing. The square denotes the test statistic value. The horizontal bars represent the confidence interval. The vertical dashed lines denote the margin of equivalence (Lakens 2017). Statistical equivalence occurs if the confidence interval falls within the dashed lines; statistical difference occurs if the confidence interval does not cross 0.

In standard hypothesis testing, the Type I error rate *alpha* is used to define the level of certainty the researcher desires. In equivalence testing, the researcher sets both the alpha value (which affects the width of the CI) and the equivalence margin. Smaller p-values and smaller equivalence margins equate to an increased level of confidence that the two datasets are equivalent.

For this analysis, two different equivalence margins (θ) of ±0.2 and ±0.5 mg/L were used. The value of 0.2 mg/L represents the typical analytical precision of DO sonde equipment in use (0.1 or 0.2 mg/L). The larger value of 0.5 mg/L was chosen in consultation with the group to represent an acceptable amount of error.

The equivalence analysis was run for each Site-Year-Month combination. The typical sample size for each analysis was n = 30 (days) of paired data. The minimum paired sample size was set to 10 days for sites with missing data. (The paired sample size is not related to the number of days used to calculate the 7- and 30-day criteria.). The analysis was partitioned into months for two reasons. First, it was of interest to see if results differed between the early (June) and later (August) portion of the growing season. Second, we did not wish for any seasonal trends to influence the results. Focusing on a maximum length of one month greatly reduced the chance of the results being influenced by a seasonal pattern.

Since the data are paired, the analysis is run on the differences between the full and shortened datasets. (This is the same way a 2-sample paired t-test is analyzed.) That is, if there were 30 full values and 30 shortened values for the month of June, then the difference for each day was calculated, and the confidence interval is calculated using the 30 difference values.

For each Site-Year-Month combination, the CI was compared against both equivalence margins. A CI represents the uncertainty around the *mean* Δ of the dataset. Depending on the end goals, it may be more appropriate quantify the uncertainty around an individual, random sample from the dataset. A prediction interval quantifies this type of uncertainty. For example, if an agency wants the average of multiple (shortened) rolling averages to be equivalent to the average of multiple (full) rolling averages, then the CI will quantify that. In contrast, if an agency wishes to see if a random, single, shortened rolling average is equivalent to a full rolling average, then PI should be used. For this analysis, we used the 5th and 95th percentiles of the differences as the interval representing the uncertainty around a single random sample. This “percentile interval” PI was compared against both equivalence margins. Additional discussion on whether the CI or PI is more appropriate can be found in the Results section.

Finally, continuous datasets are autocorrelated (temporally dependent). Exploratory analyses of these data also show that aggregated continuous data can still retain a high degree of autocorrelation. Data independence is a key assumption for many statistical tests. Violating this assumption will cause CIs to be artificially narrow and p-values to be artificially smaller. This can result in increased false positives (saying datasets are equivalent when then are not), meaning that the researcher cannot trust the specified Type I error rate (alpha). To account for this, the CIs were calculated using a statistical model that assumes an autoregressive first order (AR1) correlation structure. An AR1 assumption is common for time series data and is assumed in ARIMA time series modeling. The R function “MASS::glmmPQL()” was used run the AR1 model. The PIs were calculated using percentiles and do not rely on asymptotic large sample theory like traditional CIs. Therefore, the PIs do not require any adjustment for autocorrelation.

# Results

Example output were visualized for select sites, representing one site in Wyoming (Fish Creek), one site in Montana (Little Beaver Creek), and one site in Oregon (Wildcat Creek). These example sites illustrate DO conditions that are fairly constant over time (Fish Creek), variable from day to day (Little Beaver Creek and Wildcat Creek), and have an increasing trend throughout the growing season (Wildcat Creek). The sites were chosen as illustrative rather than as statistically representative examples. shows how the rolling averages smooth the underlying daily data.

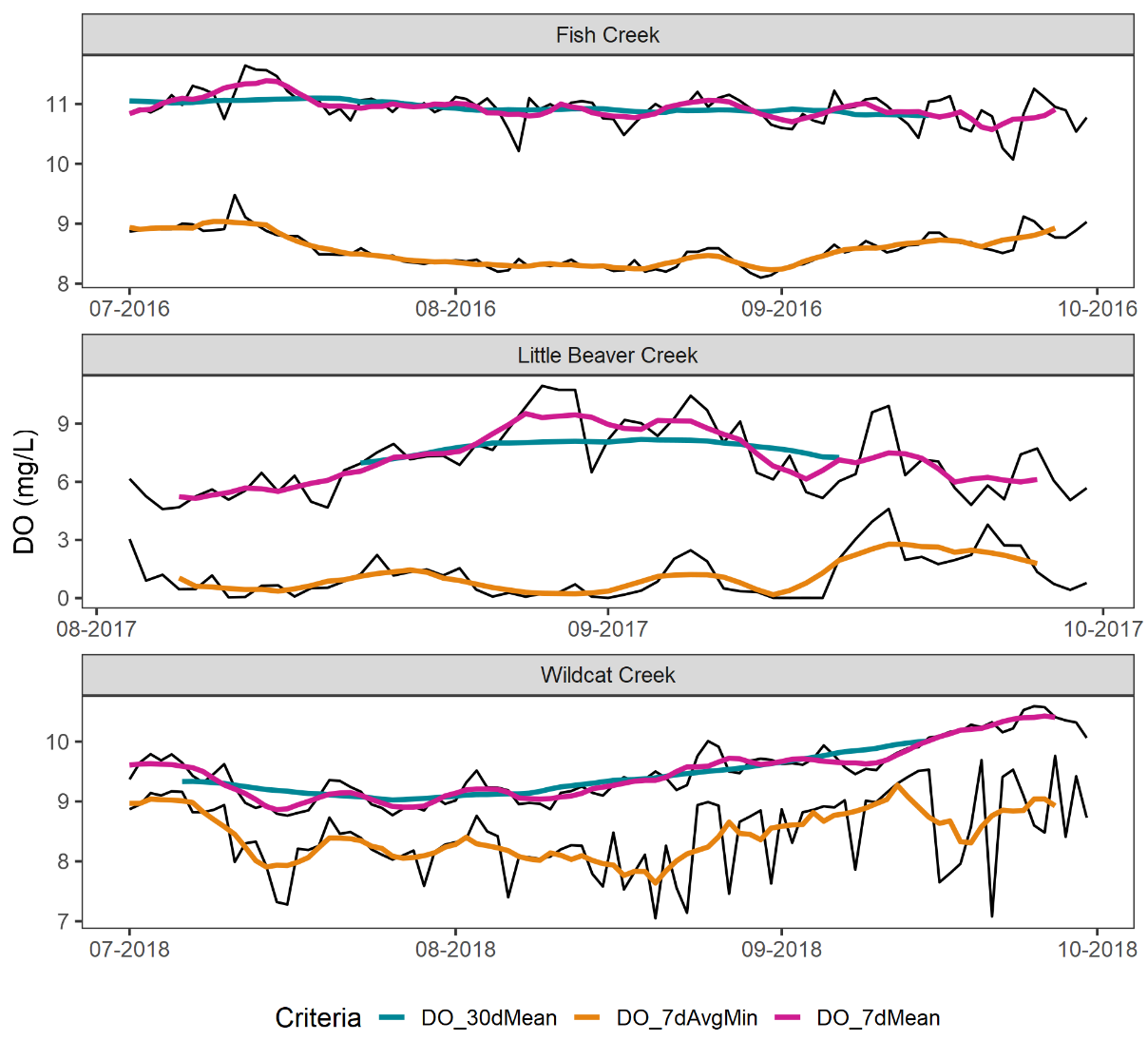
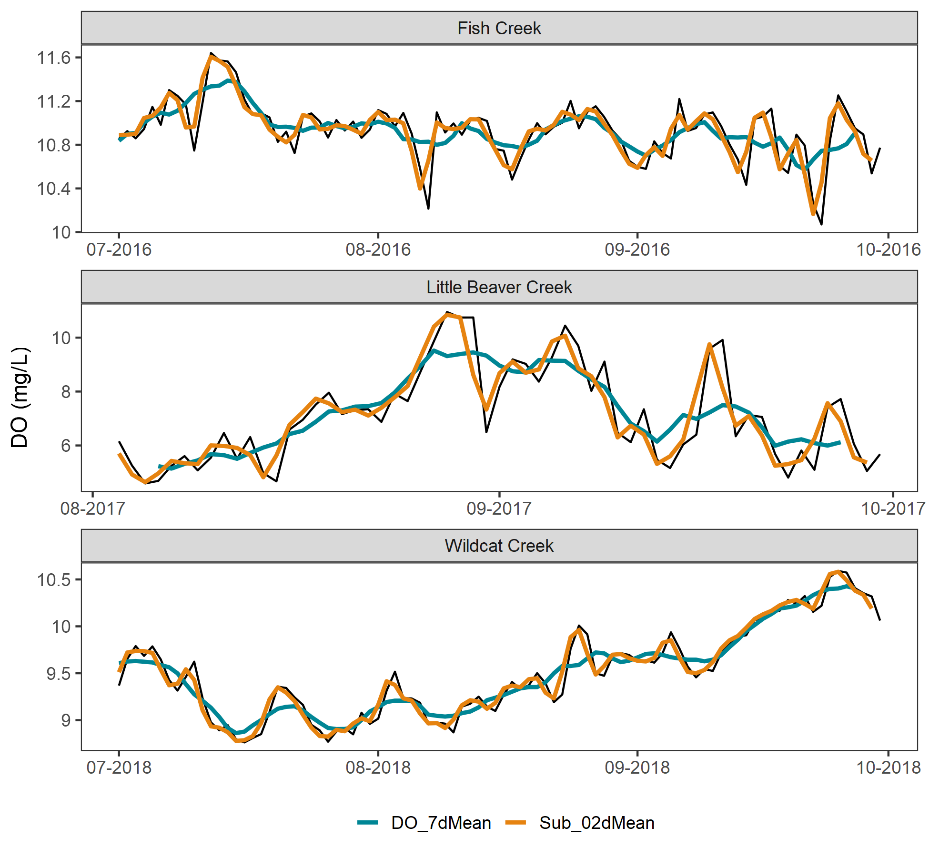


Figure . Time series plot of Fish Creek (WY), Little Beaver Creek (MT), and Wildcat Creek (OR). Daily data (mean and min) are shown in black. Rolling average data are shown in colors. These example sites illustrate the impacts of subsampling for streams that are fairly constant and invariable (Fish Creek), variable (Little Beaver Creek), and trending (Wildcat Creek).

As the duration for the shortened rolling metrics increase, they will approach the “full” rolling average metric. This is visualized in the sequence in below.



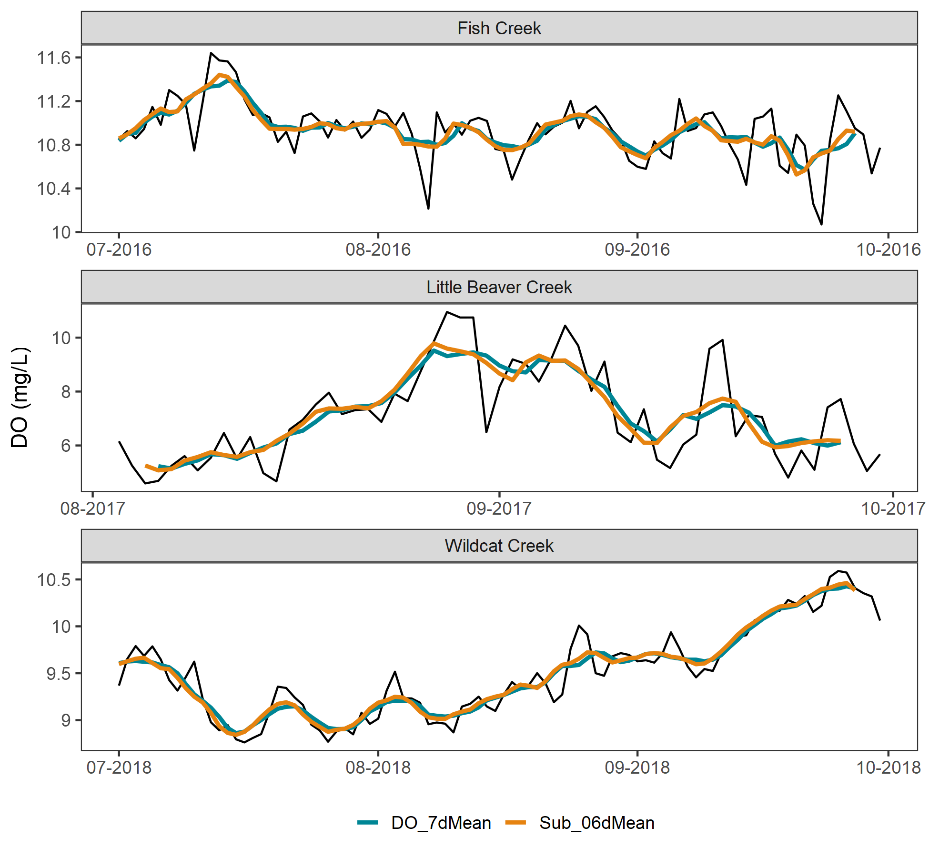


Figure . Time series plot of Fish Creek (WY), Little Beaver Creek (MT), and Wildcat Creek (OR). Black line is 1 day mean. Blue line is 7-day mean. Orange line is 2-day mean (top) and 6-day mean (bottom). The orange line approaches the curvature of the blue line as the number of days increases.

Boxplots were generated to visualize comparisons. compares June data for Little Beaver Creek (MT). The 7-day mean (left) has a smaller range than the 2-day window (right), though their medians are virtually identical.

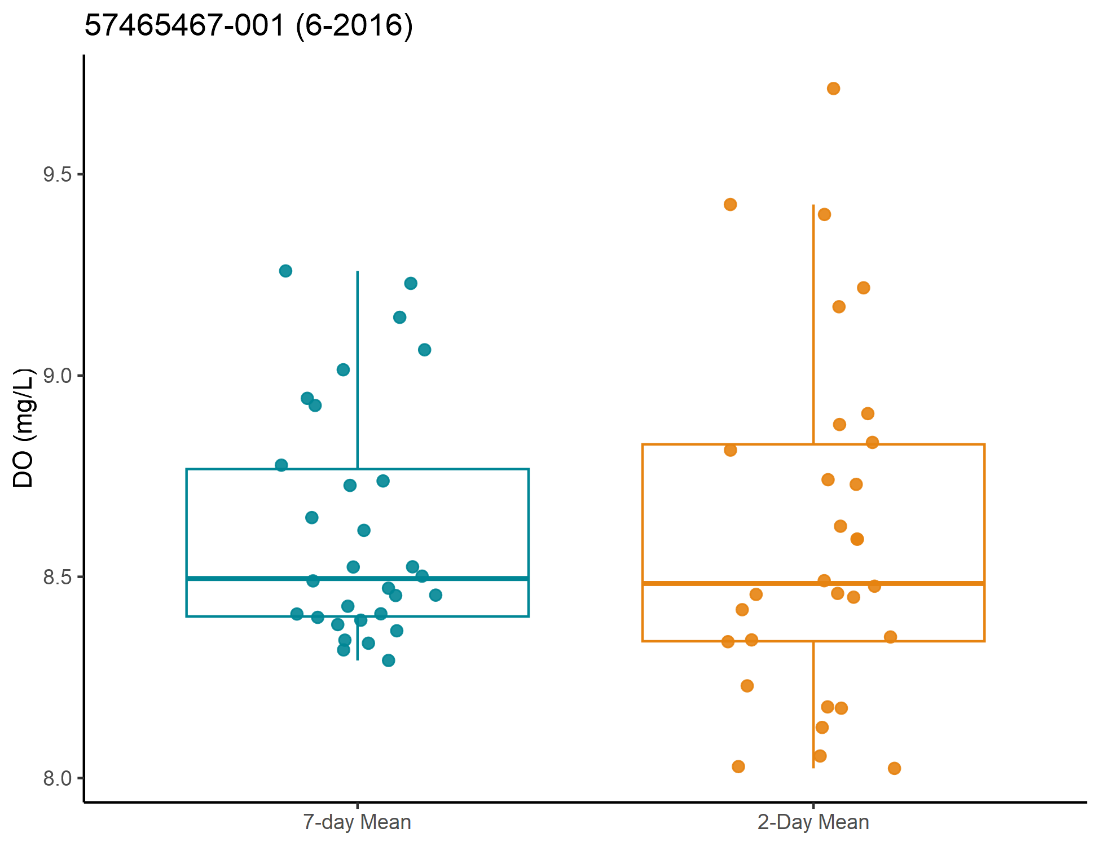


Figure . Boxplot of the 7-day mean (left) and the 2-day mean (right) for June 2016 data at Little Beaver Creek, MT.

Similarly, time series plots were also generated to visualize the differences. shows a smoother line for the 7-day avg min metric for Fish Creek (WY).

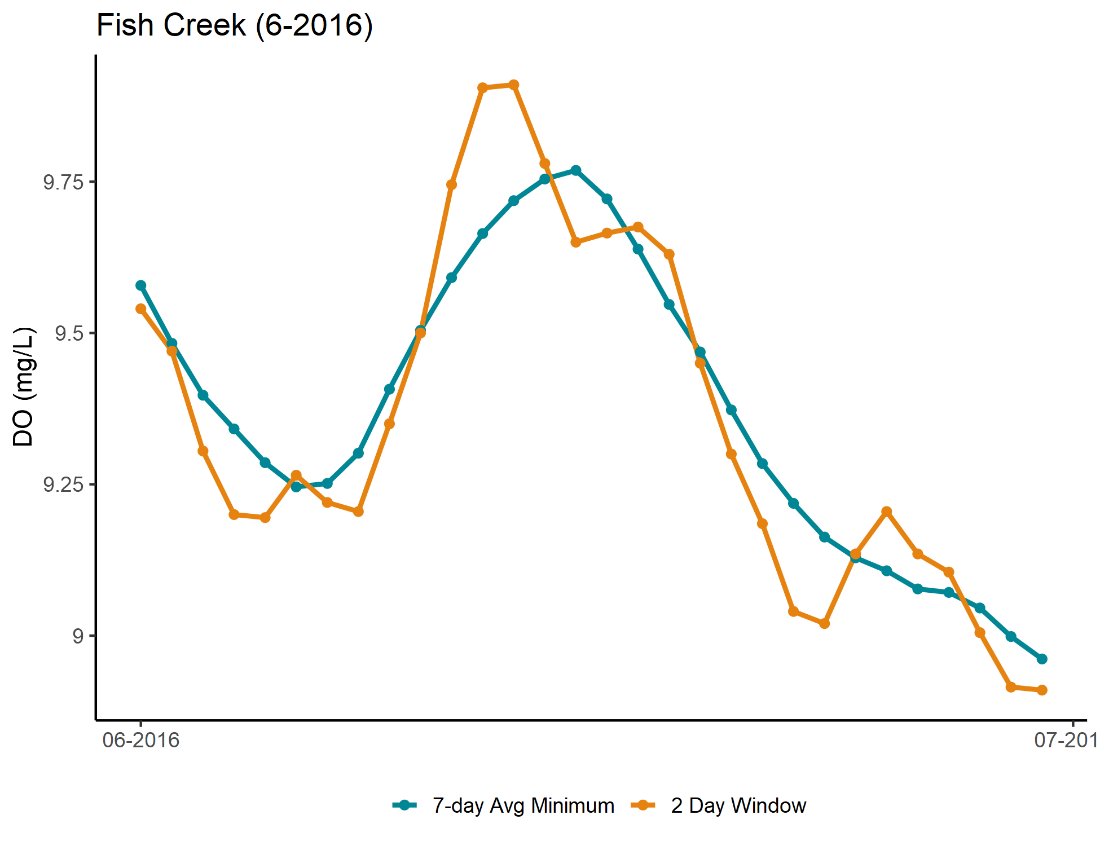


Figure . Time series plot of 7-day avg min (blue) and 2-day avg min (orange) for June 2016 data at Fish Creek, WY.

Interval plots were generated to visualize the CIs, PI, and equivalence margins. shows an example where neither the CI (thick black horizontal line) nor the PI (horizontal grey line) are equivalent at the ±0.2 mg/L margin (the lines are not fully contained by the equivalence margins), though both are equivalent at the ±0.5 mg/L margin. In both cases, the difference is not significant (the CI crosses zero).

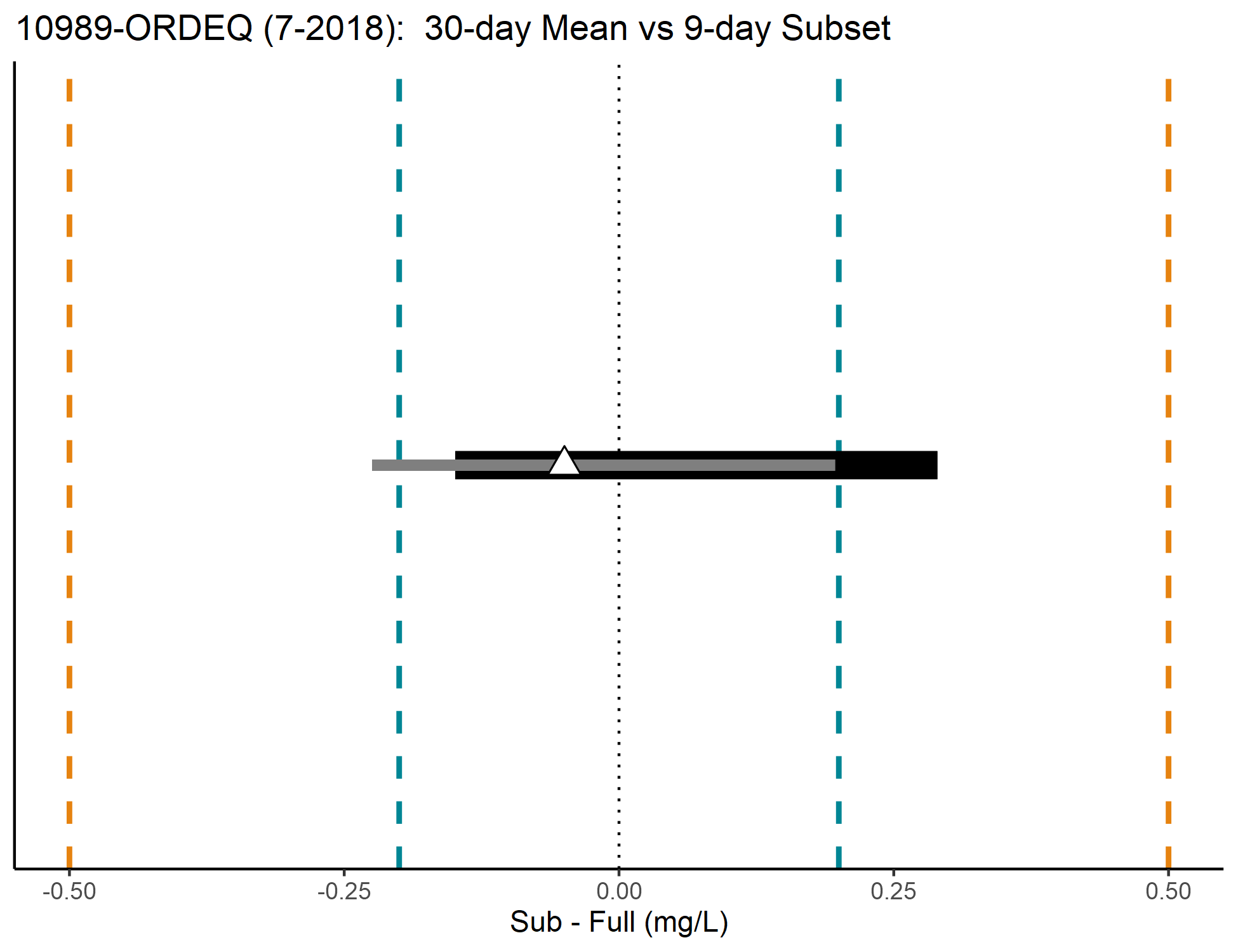


Figure . Interval plot for July 2018 data from Wildcat Creek (OR). The x-axis denotes the difference between the two datasets (full 30-day mean vs. 9-day subset). The white triangle represents the average difference. Horizontal lines are CI (thick black) and PI (grey). Vertical dashed lines are equivalence margins at ±0.2 mg/L (blue) and ±0.5 mg/L (orange). A interval plot representing the CI is overlaid for aesthetics.

shows an example from Box Elder Creek (S4603; MT). With 10 days of data, both the CI and PI in relation to the 30-day mean are within the ±0.2 mg/L equivalence region, using data from August 2017 (top). The time series plot shows how the 30-day and 10-day aggregated datasets track across the month.

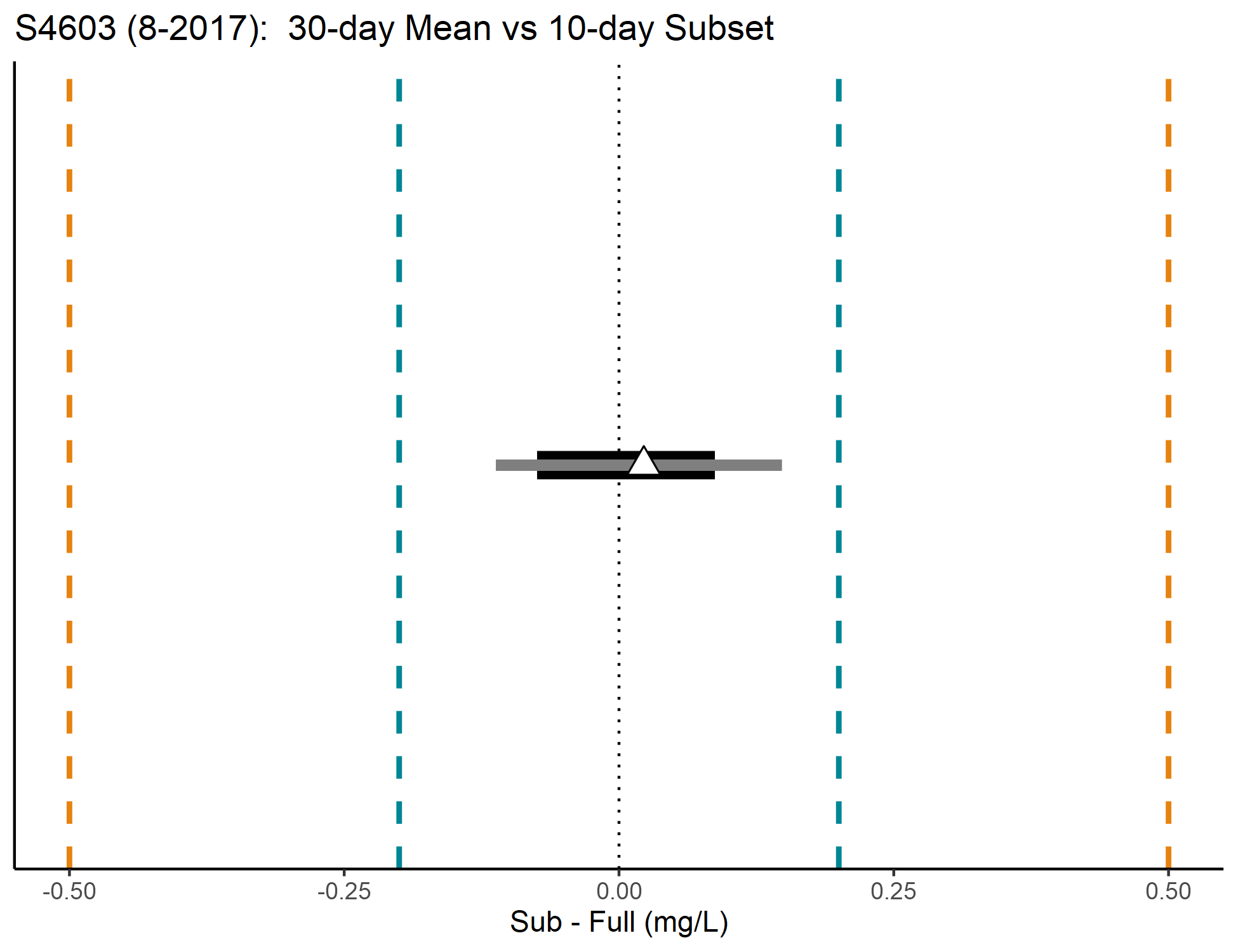
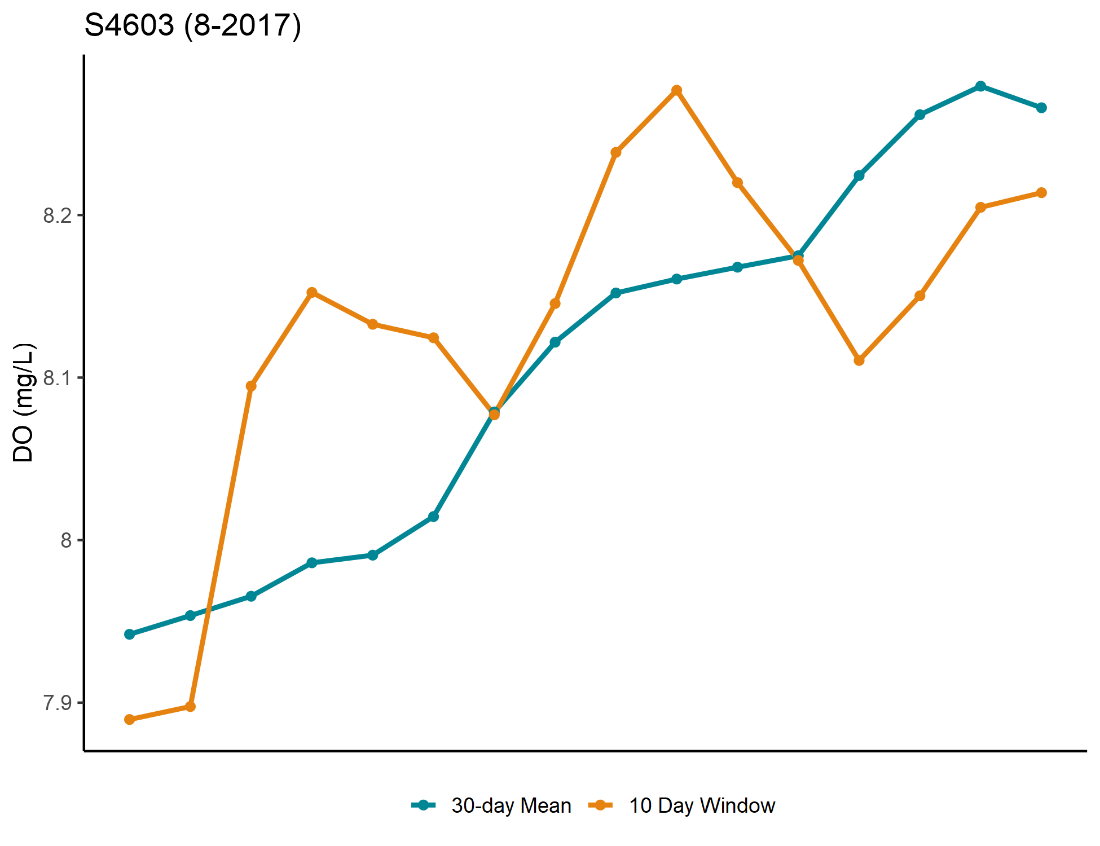
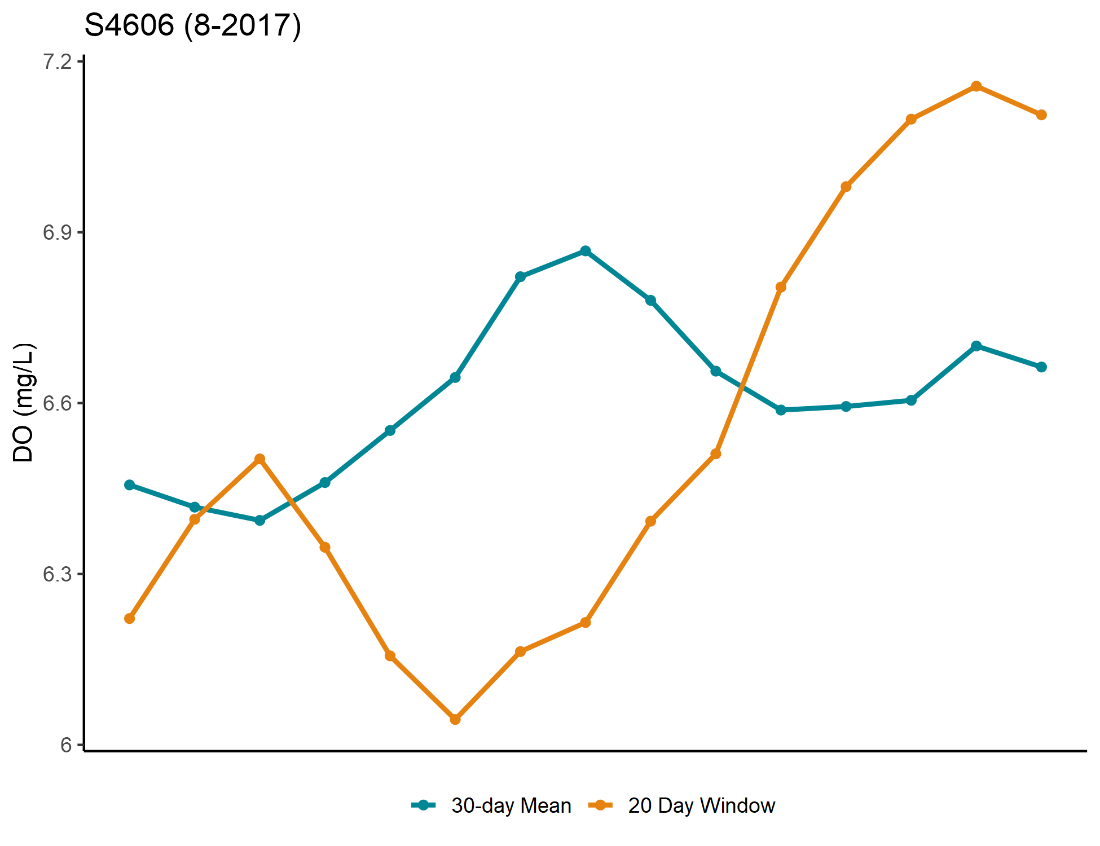
  


Figure . Interval plot (top) and time series plots (bottom) for Box Elder Creek (S4603; MT).

from Hay Creek (S4606; MT) steps through each major step of the analysis. First, 15 days of paired, rolling averaged data were calculated are plotted in the top panel. In blue, 30-day means are shown along with 20-day means in orange. The x-axis represents specific dates, so points that align vertically represent rolling averages centered on the same date. In the middle panel, the differences for each of these 15 days are plotted. The CI and PI are calculated using these 15 difference values. Finally, the bottom panel shows how the calculated CI (horizontal thick black line) and the PI (horizontal grey line) compare against the two equivalence margins. *Equivalence* is determined based on whether the intervals are fully contained within either set of equivalence margins. For this site and with 20 days of aggregation, neither the CI nor the PI are equivalent with respect to the ±0.5 mg/L equivalence region for August 2017. This finding agrees with the wide divergence in values shown in the time series plots.

  
A picture containing diagram, line, plot

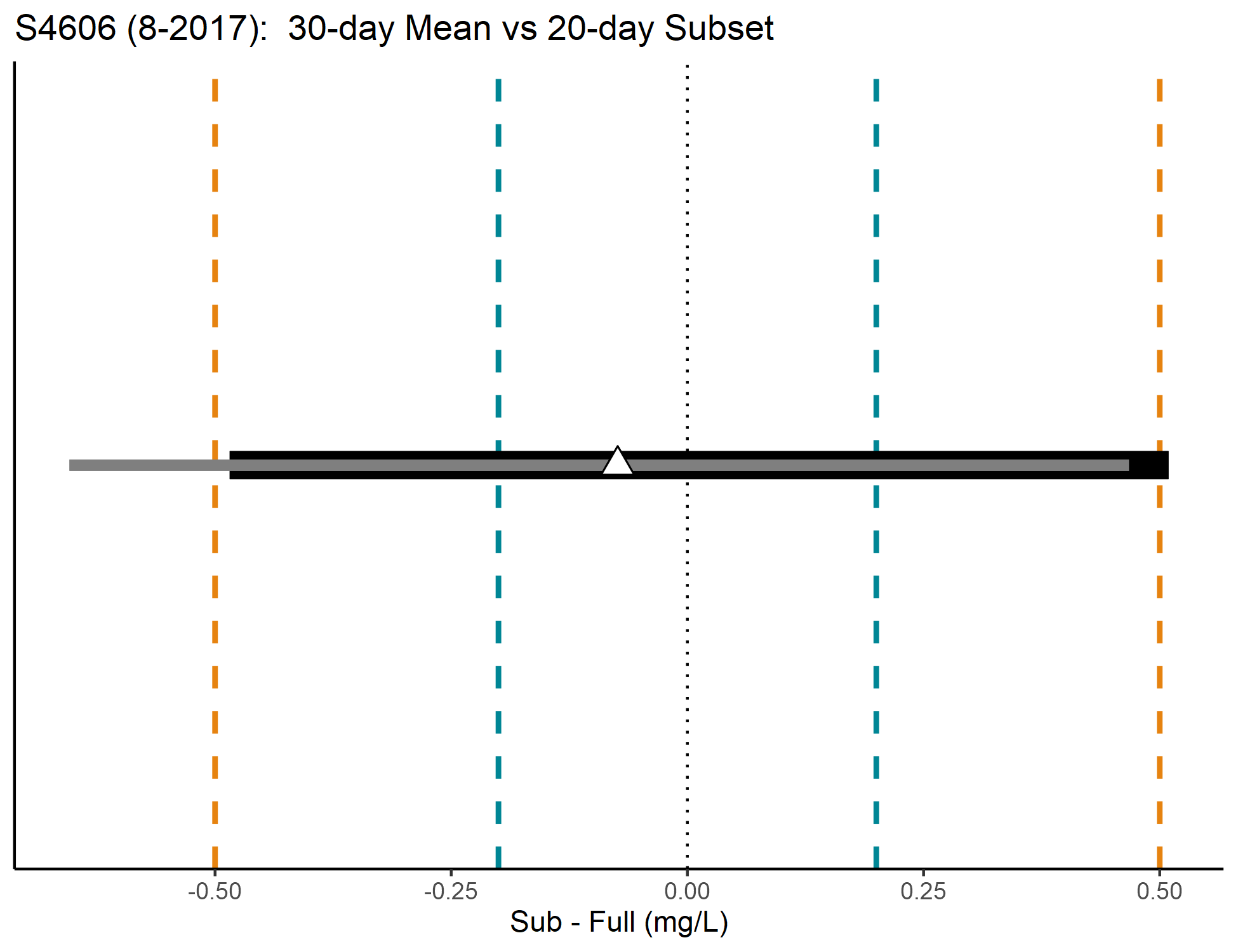
Description automatically generated  


Figure . Interval plot (top) and time series plots (bottom) for Hay Creek (S4606; MT).

The final dataset contained 13,233 Site-Year-Month-Criteria combinations from 110 sites sampled across 11 states between 2010 and 2021 (see Appendix for list of sites). Equivalence tests were run for 10,854 combinations with sufficient sample size. Boxplots of the true, full duration criteria values are shown in . To visualize over 10,000 results, plots were generated that display the proportion equivalent for each day of equivalence. If multiple tested durations were deemed equivalent, then the minimum number of days was plotted. Separate plots were generated for each combination of interval (CI and PI) and equivalence margin (±0.2 mg/L and ±0.5 mg/L).

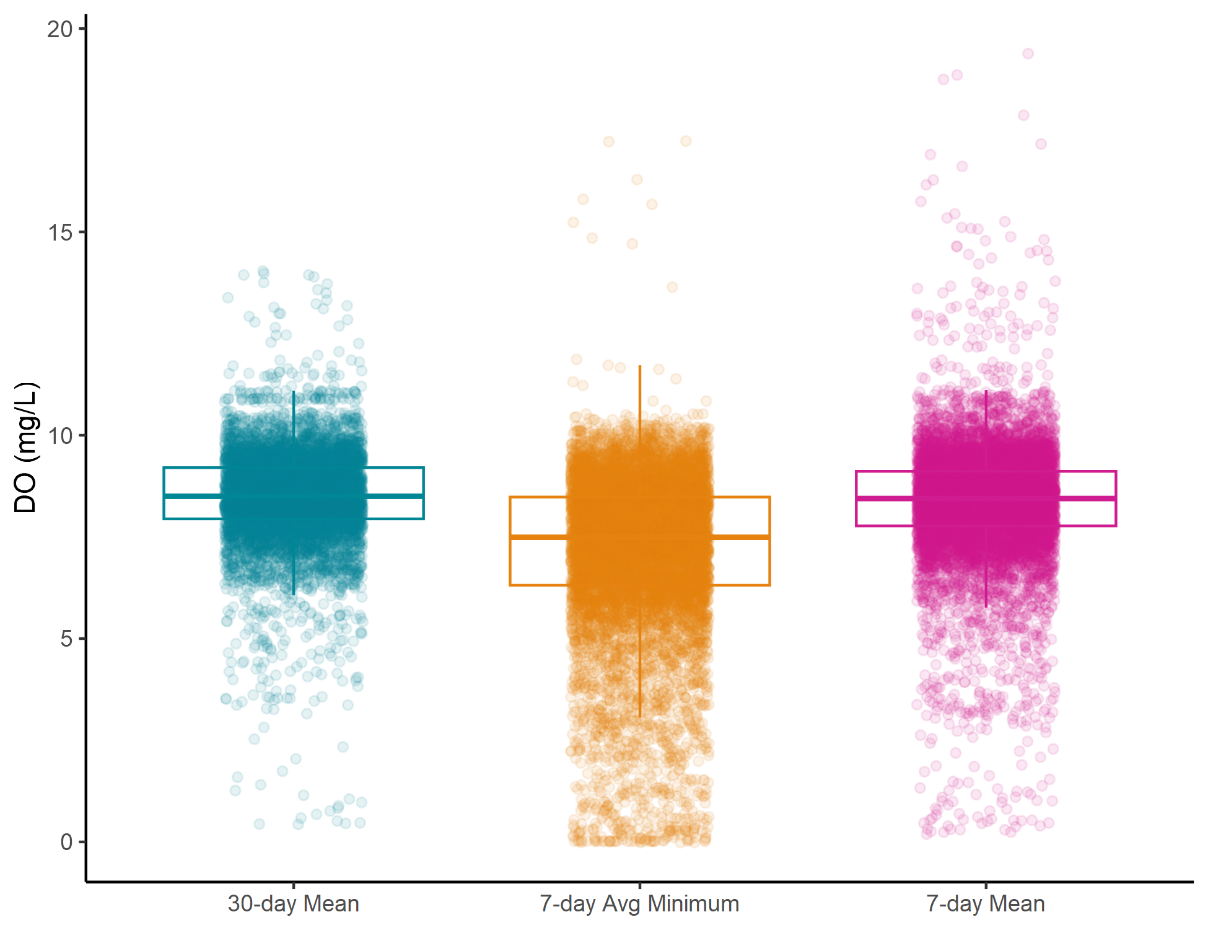


Figure . Boxplots of observed criteria values using the full duration.

## Confidence Intervale ±0.2 mg/L

Results for the CI and ±0.2 mg/L equivalence margin () suggest that 1 day of data can accurately approximate the overall 7-day mean and 7-day avg min across a month (86.5% of site-year-month samples). Tabular results are presented in Table 1. For the 30-day mean, there were many sites that were well approximated by 1 day (26.1% of site-year-month samples). However, many sites required up to 21 days (8.2%), and a larger number of sites were not well approximated when using up to 21 days (17.3%) (the “30-day” column contained the 2nd largest count of tests). The 30-day mean has a longer averaging period than the 7-day metrics. Therefore, the 30-day mean is more likely to experience increased data variation. This makes it more challenging to approximate the 30-day mean using a shorter duration, as illustrated in each of these cumulative step plots.

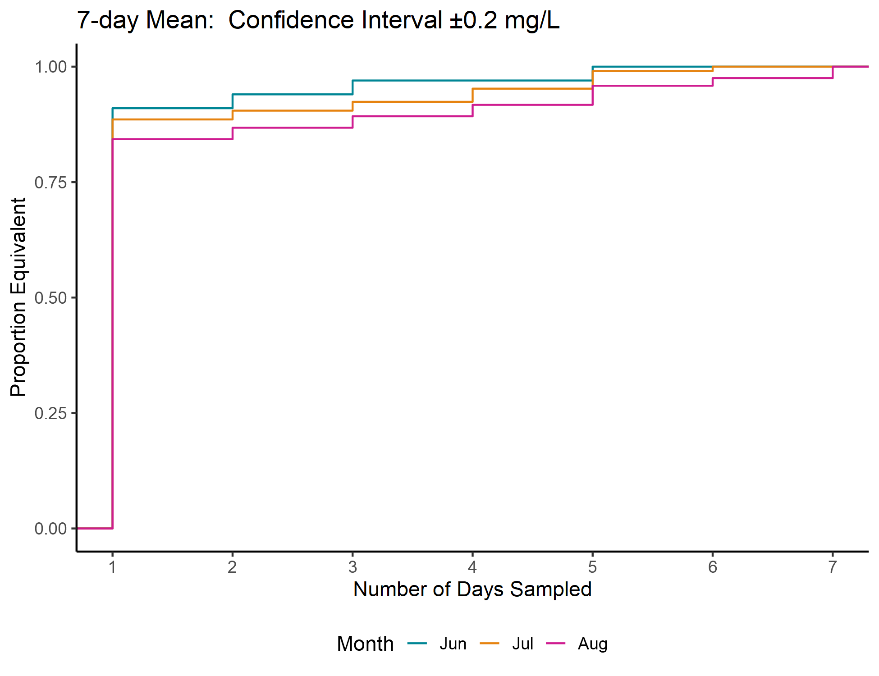
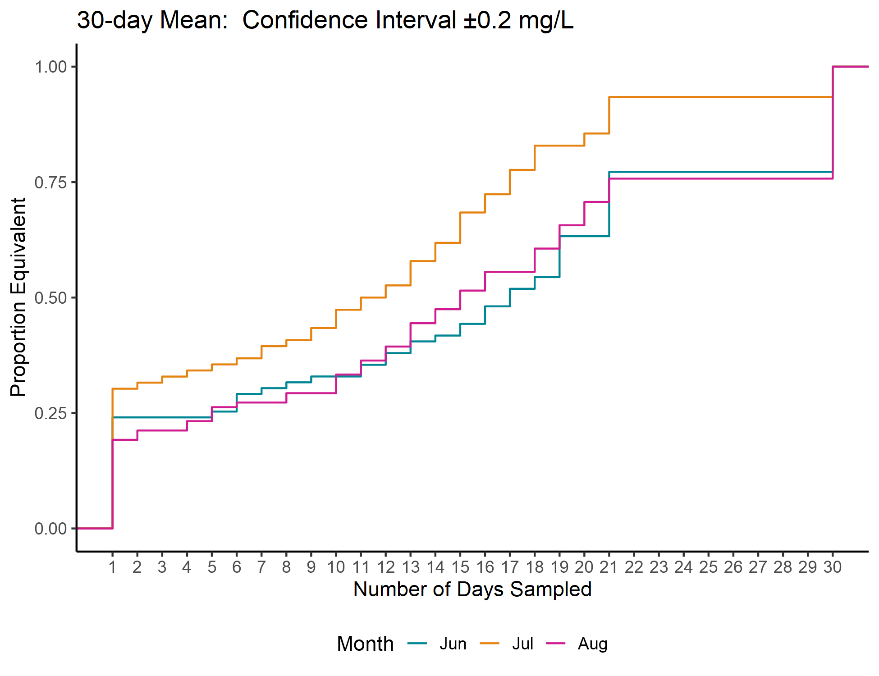
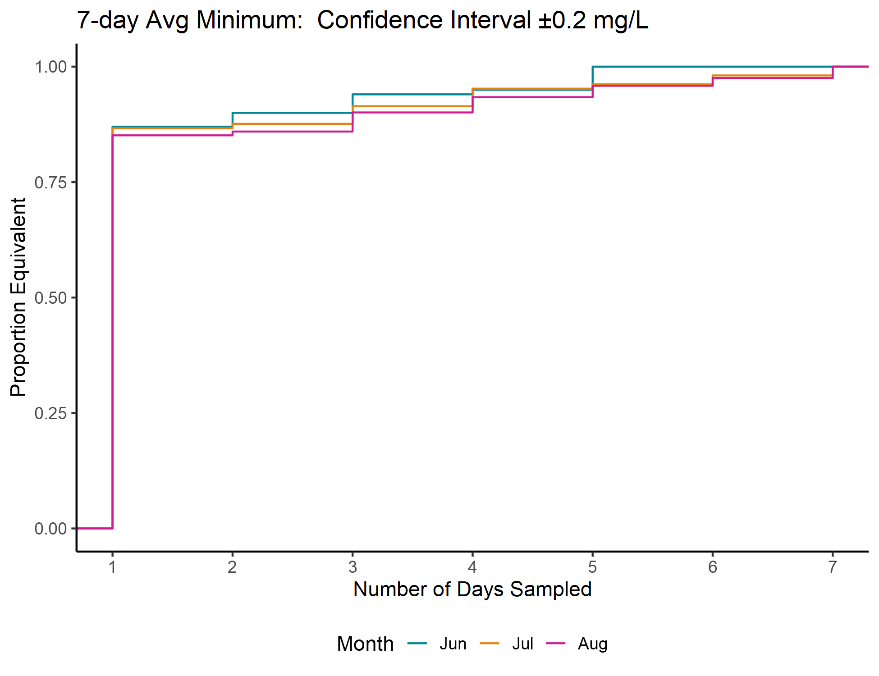
   


Figure . Cumulative step plots for CI and ±0.2 mg/L. Results are color-coded by month.

Table . Summary counts of the number of CI and ±0.2 mg/L equivalent tests for each criteria

|  |  | Number of Tests Equivalent | | | |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Criteria | Number of Days Tested | June | July | Aug | Total | Percentage | Cumulative Percentage\* |
| DO\_30dMean | 1 | 19 | 23 | 19 | 61 | 24.0 | 24.0 |
| 2 | 0 | 1 | 2 | 3 | 1.2 | 25.2 |
| 3 | 0 | 1 | 0 | 1 | 0.4 | 25.6 |
| 4 | 0 | 1 | 2 | 3 | 1.2 | 26.8 |
| 5 | 1 | 1 | 3 | 5 | 2.0 | 28.7 |
| 6 | 3 | 1 | 1 | 5 | 2.0 | 30.7 |
| 7 | 1 | 2 | 0 | 3 | 1.2 | 31.9 |
| 8 | 1 | 1 | 2 | 4 | 1.6 | 33.5 |
| 9 | 1 | 2 | 0 | 3 | 1.2 | 34.6 |
| 10 | 0 | 3 | 4 | 7 | 2.8 | 37.4 |
| 11 | 2 | 2 | 3 | 7 | 2.8 | 40.2 |
| 12 | 2 | 2 | 3 | 7 | 2.8 | 42.9 |
| 13 | 2 | 4 | 5 | 11 | 4.3 | 47.2 |
| 14 | 1 | 3 | 3 | 7 | 2.8 | 50.0 |
| 15 | 2 | 5 | 4 | 11 | 4.3 | 54.3 |
| 16 | 3 | 3 | 4 | 10 | 3.9 | 58.3 |
| 17 | 3 | 4 | 0 | 7 | 2.8 | 61.0 |
| 18 | 2 | 4 | 5 | 11 | 4.3 | 65.4 |
| 19 | 7 | 0 | 5 | 12 | 4.7 | 70.1 |
| 20 | 0 | 2 | 5 | 7 | 2.8 | 72.8 |
| 21 | 11 | 6 | 5 | 22 | 8.7 | 81.5 |
| 30 | 18 | 5 | 24 | 47 | 18.5 | 100.0 |
| DO\_7dAvgMin | 1 | 87 | 91 | 103 | 281 | 86.2 | 86.2 |
| 2 | 3 | 1 | 1 | 5 | 1.5 | 87.7 |
| 3 | 4 | 4 | 5 | 13 | 4.0 | 91.7 |
| 4 | 1 | 4 | 4 | 9 | 2.8 | 94.5 |
| 5 | 5 | 1 | 3 | 9 | 2.8 | 97.2 |
| 6 | 0 | 2 | 2 | 4 | 1.2 | 98.5 |
| 7 | 0 | 2 | 3 | 5 | 1.5 | 100.0 |
| DO\_7dMean | 1 | 91 | 93 | 102 | 286 | 87.7 | 87.7 |
| 2 | 3 | 2 | 3 | 8 | 2.5 | 90.2 |
| 3 | 3 | 2 | 3 | 8 | 2.5 | 92.6 |
| 4 | 0 | 3 | 3 | 6 | 1.8 | 94.5 |
| 5 | 3 | 4 | 5 | 12 | 3.7 | 98.2 |
| 6 | 0 | 1 | 2 | 3 | 0.9 | 99.1 |
| 7 | 0 | 0 | 3 | 3 | 0.9 | 100.0 |

\*Departures in cumulative percentage from 100% are attributed to rounding errors at the first decimal place

Results for the 7-day criteria did not appear to vary much across months (). However, results did appear to vary some across months for 30-day criteria, with July data having a larger proportion of equivalence. One hypothesis is that the more steady temperatures of July (compared to June and August) allow for a shorter minimum period of equivalence.

## Percentile Intervale ±0.2 mg/L

Results for the PI and ±0.2 mg/L equivalence margin (; Table 2) showed a range of minimum days were required to well approximate the 7-day metrics. For the 30-day mean, most sites were not well approximated by any tested duration of days. However, the counts of equivalent tests did increase over time from 1 day to 21 days. July data appear to reach 30-day equivalence somewhat more quickly than other months.

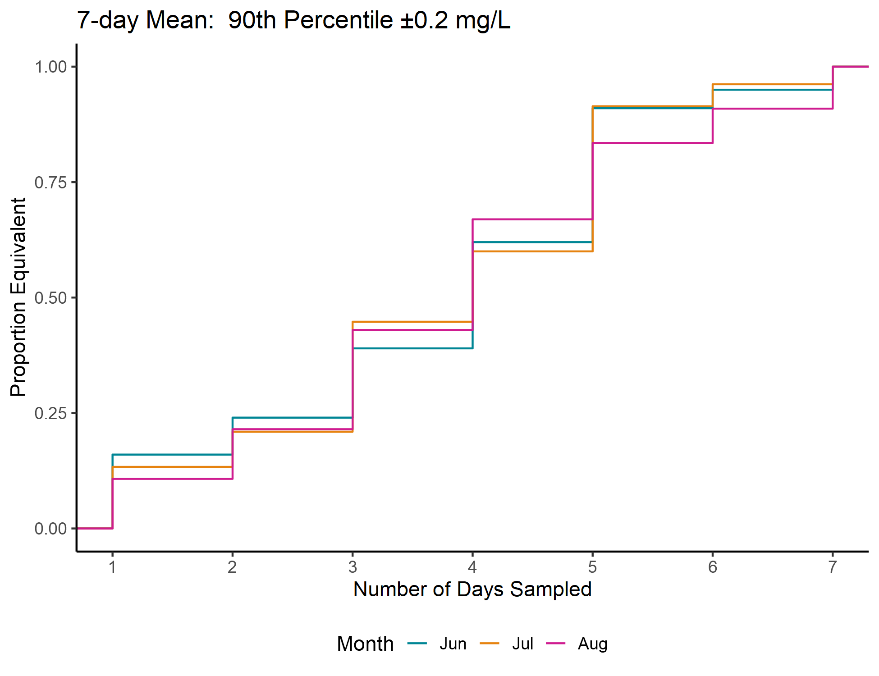
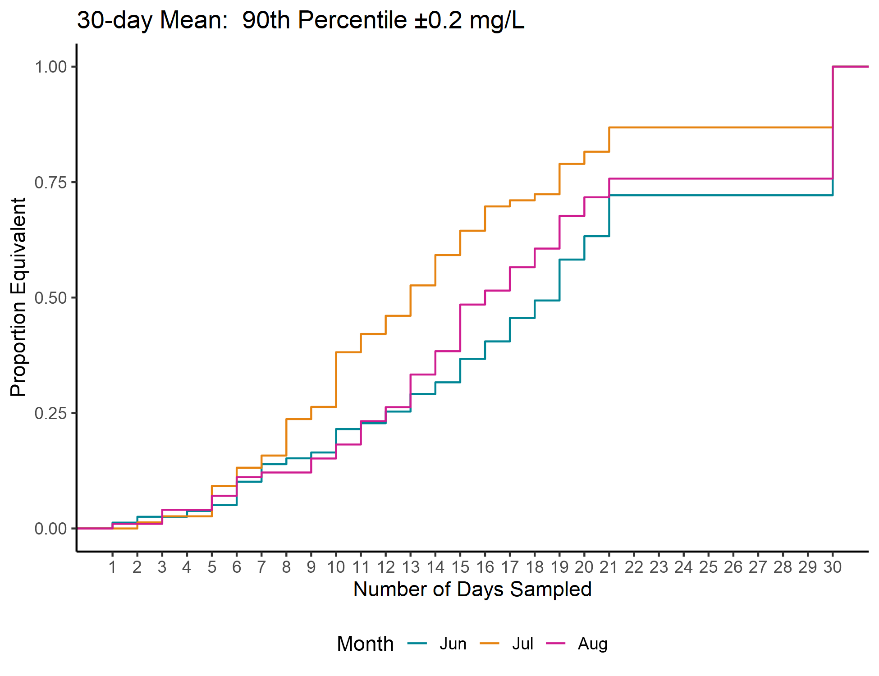
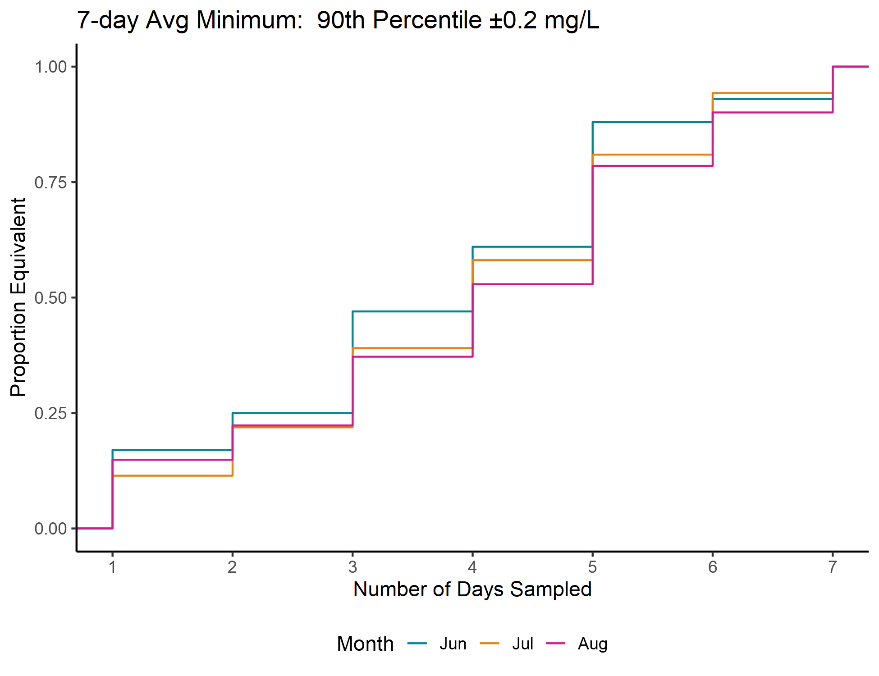
   


Figure . Cumulative step plots for PI and ±0.2 mg/L. Results are color-coded by month.

Table . Summary counts of the number of PI and ±0.2 mg/L equivalent tests for each criteria

|  |  | Number of Tests Equivalent | | | |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Criteria | Number of Days | June | July | Aug | Total | Percentage | Cumulative Percentage\* |
| DO\_30dMean | 1 | 1 | 0 | 1 | 2 | 0.8 | 0.8 |
| 2 | 1 | 1 | 0 | 2 | 0.8 | 1.6 |
| 3 | 0 | 1 | 3 | 4 | 1.6 | 3.1 |
| 4 | 1 | 0 | 0 | 1 | 0.4 | 3.5 |
| 5 | 1 | 5 | 3 | 9 | 3.5 | 7.1 |
| 6 | 4 | 3 | 4 | 11 | 4.3 | 11.4 |
| 7 | 3 | 2 | 1 | 6 | 2.4 | 13.8 |
| 8 | 1 | 6 | 0 | 7 | 2.8 | 16.5 |
| 9 | 1 | 2 | 3 | 6 | 2.4 | 18.9 |
| 10 | 4 | 9 | 3 | 16 | 6.3 | 25.2 |
| 11 | 1 | 3 | 5 | 9 | 3.5 | 28.7 |
| 12 | 2 | 3 | 3 | 8 | 3.1 | 31.9 |
| 13 | 3 | 5 | 7 | 15 | 5.9 | 37.8 |
| 14 | 2 | 5 | 5 | 12 | 4.7 | 42.5 |
| 15 | 4 | 4 | 10 | 18 | 7.1 | 49.6 |
| 16 | 3 | 4 | 3 | 10 | 3.9 | 53.5 |
| 17 | 4 | 1 | 5 | 10 | 3.9 | 57.5 |
| 18 | 3 | 1 | 4 | 8 | 3.1 | 60.6 |
| 19 | 7 | 5 | 7 | 19 | 7.5 | 68.1 |
| 20 | 4 | 2 | 4 | 10 | 3.9 | 72.0 |
| 21 | 7 | 4 | 4 | 15 | 5.9 | 78.0 |
| 30 | 22 | 10 | 24 | 56 | 22.0 | 100.0 |
| DO\_7dAvgMin | 1 | 17 | 12 | 18 | 47 | 14.4 | 14.4 |
| 2 | 8 | 11 | 9 | 28 | 8.6 | 23.0 |
| 3 | 22 | 18 | 18 | 58 | 17.8 | 40.8 |
| 4 | 14 | 20 | 19 | 53 | 16.3 | 57.1 |
| 5 | 27 | 24 | 31 | 82 | 25.2 | 82.2 |
| 6 | 5 | 14 | 14 | 33 | 10.1 | 92.3 |
| 7 | 7 | 6 | 12 | 25 | 7.7 | 100.0 |
| DO\_7dMean | 1 | 16 | 14 | 13 | 43 | 13.2 | 13.2 |
| 2 | 8 | 8 | 13 | 29 | 8.9 | 22.1 |
| 3 | 15 | 25 | 26 | 66 | 20.2 | 42.3 |
| 4 | 23 | 16 | 29 | 68 | 20.9 | 63.2 |
| 5 | 29 | 33 | 20 | 82 | 25.2 | 88.3 |
| 6 | 4 | 5 | 9 | 18 | 5.5 | 93.9 |
| 7 | 5 | 4 | 11 | 20 | 6.1 | 100.0 |

\*Departures in cumulative percentage from 100% are attributed to rounding errors at the first decimal place

## Confidence Intervale ±0.5 mg/L

Results for the CI and ±0.5 mg/L equivalence margin (; Table 3) suggest that 1 day of data can accurately approximate both the overall 7-day and 30-day metrics across a month. July data appear to reach 30-day equivalence somewhat more quickly than other months

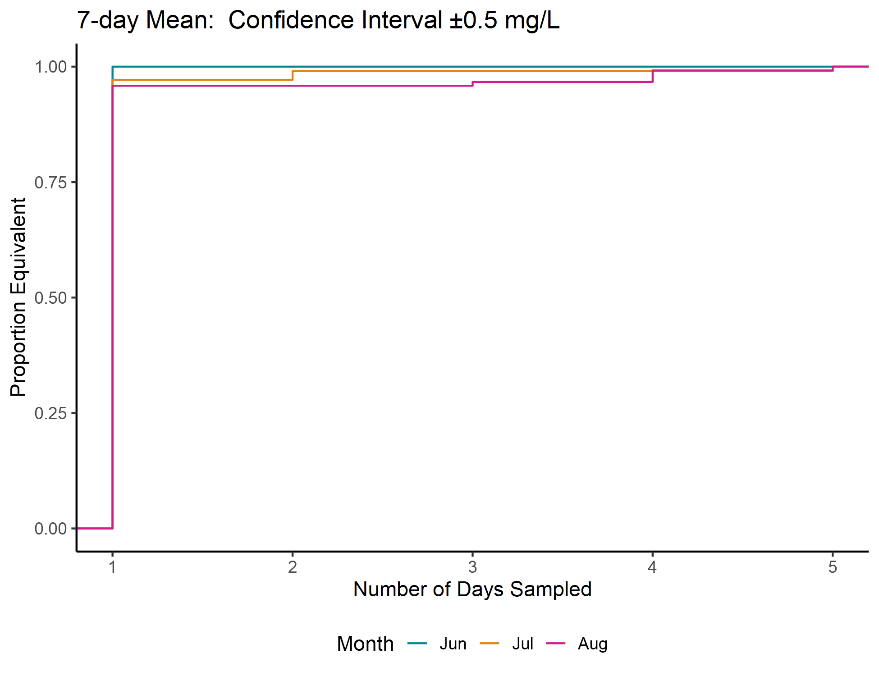
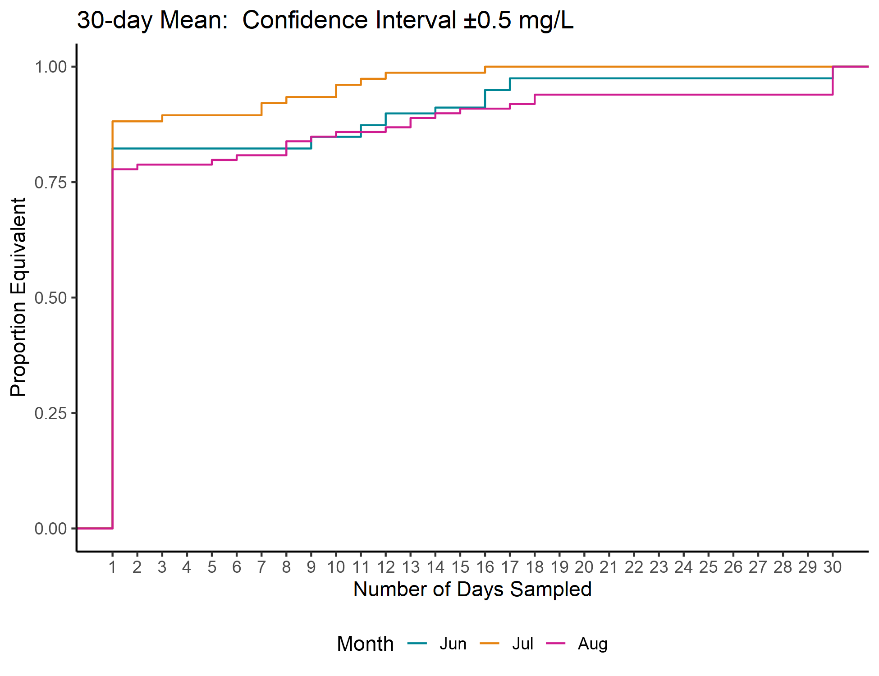
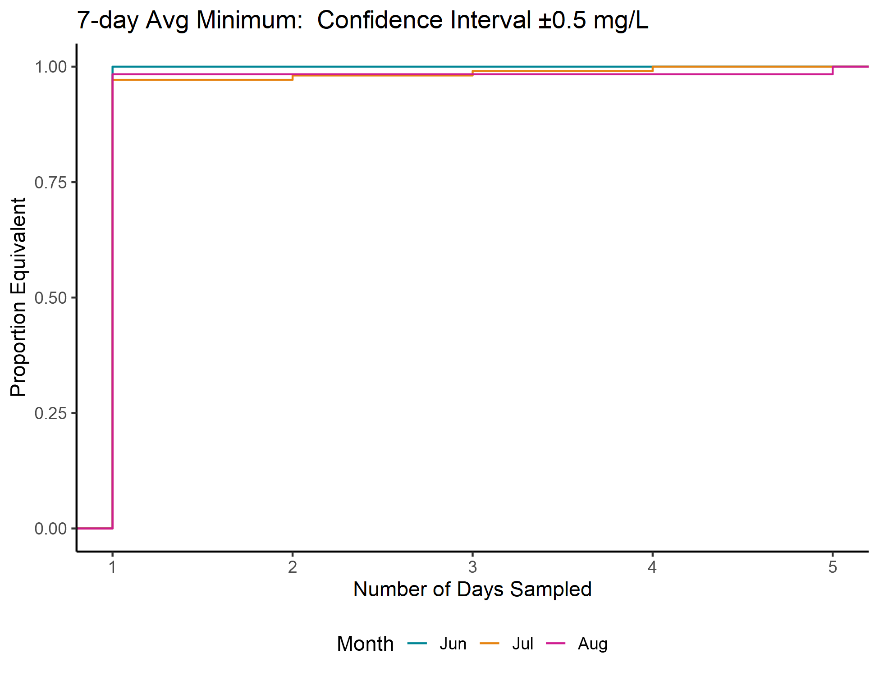
   


Figure . Cumulative step plots for CI and ±0.5 mg/L. Results are color-coded by month.

Table . Summary counts of the number of CI and ±0.5 mg/L equivalent tests for each criteria

|  |  | Number of Tests Equivalent | | | |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Criteria | Number of Days | June | July | Aug | Total | Percentage | Cumulative Percentage\* |
| DO\_30dMean | 1 | 65 | 67 | 77 | 209 | 82.3 | 82.3 |
| 2 | 0 | 0 | 1 | 1 | 0.4 | 82.7 |
| 3 | 0 | 1 | 0 | 1 | 0.4 | 83.1 |
| 4 | 0 | 0 | 1 | 1 | 0.4 | 83.5 |
| 6 | 0 | 0 | 1 | 1 | 0.4 | 83.9 |
| 7 | 0 | 2 | 0 | 2 | 0.8 | 84.6 |
| 8 | 0 | 1 | 3 | 4 | 1.6 | 86.2 |
| 9 | 2 | 0 | 1 | 3 | 1.2 | 87.4 |
| 10 | 0 | 2 | 1 | 3 | 1.2 | 88.6 |
| 11 | 2 | 1 | 0 | 3 | 1.2 | 89.8 |
| 12 | 2 | 1 | 1 | 4 | 1.6 | 91.3 |
| 13 | 0 | 0 | 2 | 2 | 0.8 | 92.1 |
| 14 | 1 | 0 | 1 | 2 | 0.8 | 92.9 |
| 15 | 0 | 0 | 1 | 1 | 0.4 | 93.3 |
| 16 | 3 | 1 | 0 | 4 | 1.6 | 94.9 |
| 17 | 2 | 0 | 1 | 3 | 1.2 | 96.1 |
| 18 | 0 | 0 | 2 | 2 | 0.8 | 96.9 |
| 19 | 2 | 0 | 6 | 8 | 3.1 | 100.0 |
| 20 | 100 | 102 | 119 | 321 | 98.5 | 98.5 |
| 21 | 0 | 1 | 0 | 1 | 0.3 | 98.8 |
| 30 | 0 | 1 | 0 | 1 | 0.3 | 99.1 |
| DO\_7dAvgMin | 1 | 0 | 1 | 0 | 1 | 0.3 | 99.4 |
| 2 | 0 | 0 | 2 | 2 | 0.6 | 100.0 |
| 3 | 100 | 102 | 116 | 318 | 97.5 | 97.5 |
| 4 | 0 | 2 | 0 | 2 | 0.6 | 98.2 |
| 5 | 0 | 0 | 1 | 1 | 0.3 | 98.5 |
| 6 | 0 | 0 | 3 | 3 | 0.9 | 99.4 |
| DO\_7dMean | 1 | 0 | 1 | 1 | 2 | 0.6 | 100.0 |
| 2 | 65 | 67 | 77 | 209 | 82.3 | 82.3 |
| 3 | 0 | 0 | 1 | 1 | 0.4 | 82.7 |
| 4 | 0 | 1 | 0 | 1 | 0.4 | 83.1 |
| 5 | 0 | 0 | 1 | 1 | 0.4 | 83.5 |
| 6 | 0 | 0 | 1 | 1 | 0.4 | 83.9 |

\*Departures in cumulative percentage from 100% are attributed to rounding errors at the first decimal place

## Percentile Intervale ±0.5 mg/L

Results for the PI and ±0.5 mg/L equivalence margin (Table 4) suggest that 1 day of data can accurately approximate most sites for the 7-day and (to a smaller extent) the 30-day metrics. July data appear to reach 30-day equivalence somewhat more quickly than other months

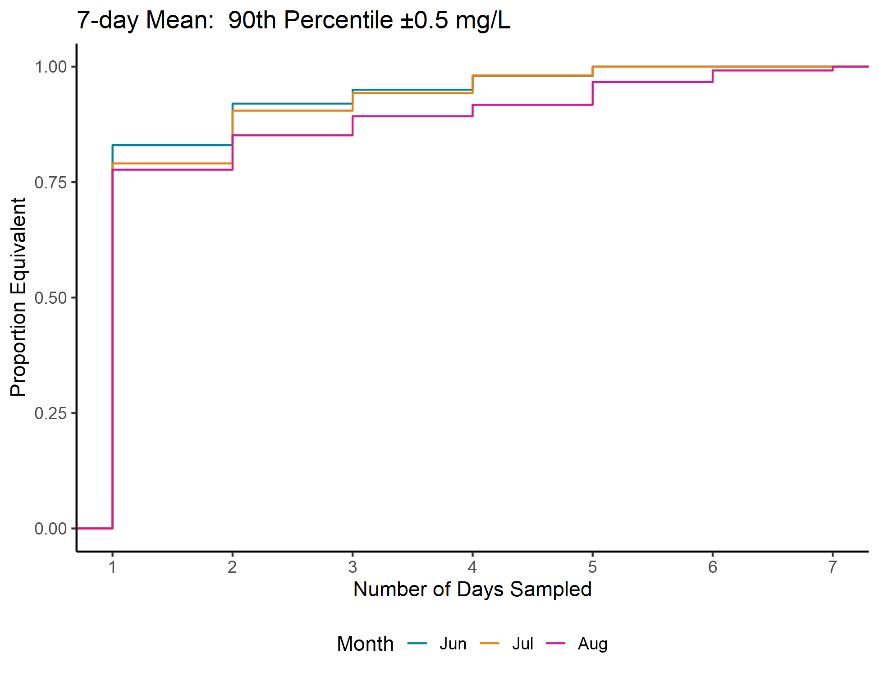
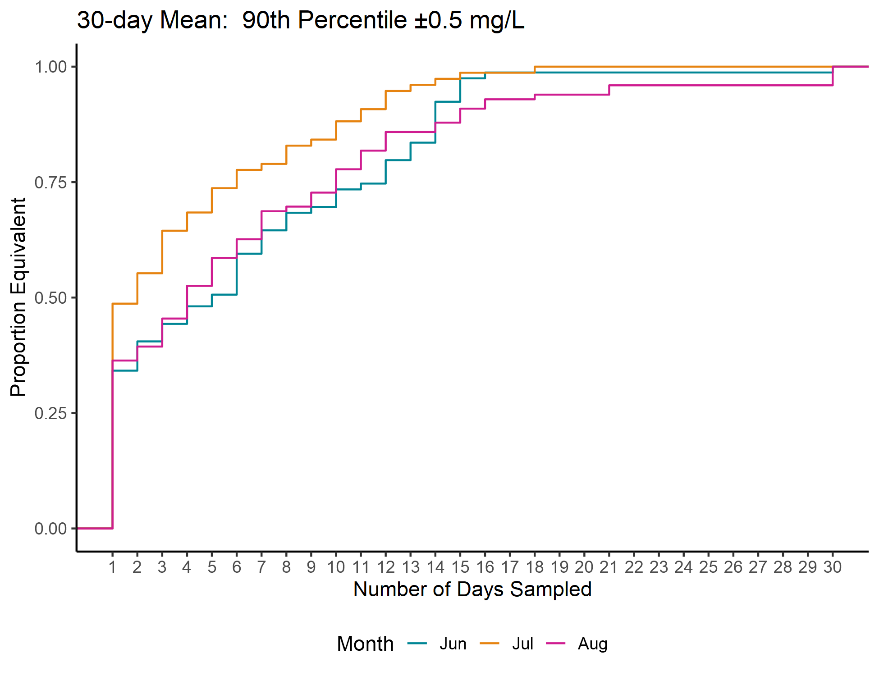
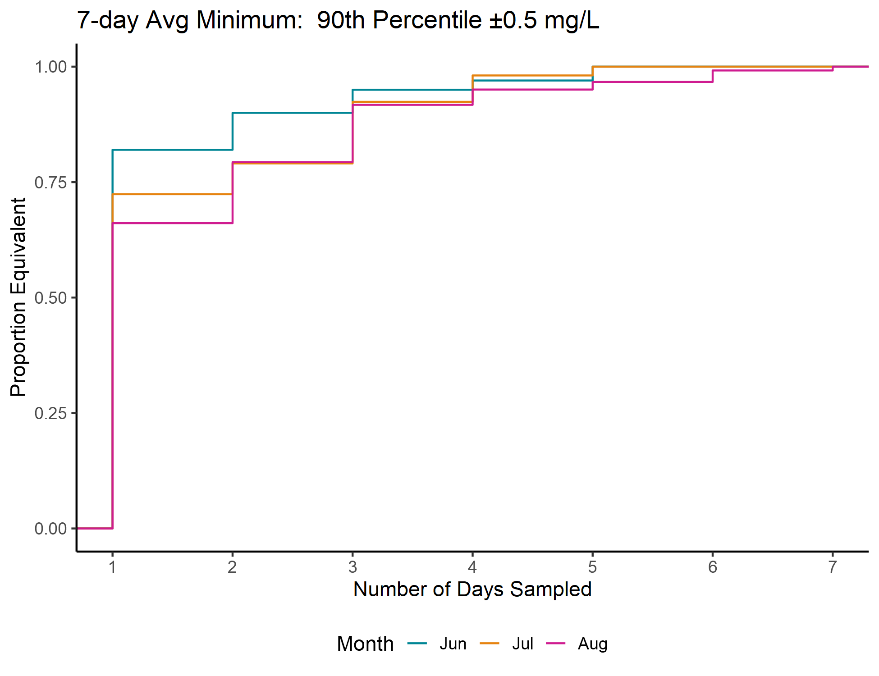
   


Figure . Cumulative step plots for PI and ±0.5 mg/L. Results are color-coded by month.

Table . Summary counts of the number of PI and ±0.5 mg/L equivalent tests for each criteria

|  |  | Number of Tests Equivalent | | | |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Criteria | Number of Days Tested | June | July | Aug | Total | Percentage | Cumulative Percentage\* |
| DO\_30dMean | 1 | 1 | 27 | 37 | 36 | 100 | 39.4 |
| 2 | 2 | 5 | 5 | 3 | 13 | 5.1 |
| 3 | 3 | 3 | 7 | 6 | 16 | 6.3 |
| 4 | 4 | 3 | 3 | 7 | 13 | 5.1 |
| 5 | 5 | 2 | 4 | 6 | 12 | 4.7 |
| 6 | 6 | 7 | 3 | 4 | 14 | 5.5 |
| 7 | 7 | 4 | 1 | 6 | 11 | 4.3 |
| 8 | 8 | 3 | 3 | 1 | 7 | 2.8 |
| 9 | 9 | 1 | 1 | 3 | 5 | 2.0 |
| 10 | 10 | 3 | 3 | 5 | 11 | 4.3 |
| 11 | 11 | 1 | 2 | 4 | 7 | 2.8 |
| 12 | 12 | 4 | 3 | 4 | 11 | 4.3 |
| 13 | 13 | 3 | 1 | 0 | 4 | 1.6 |
| 14 | 14 | 7 | 1 | 2 | 10 | 3.9 |
| 15 | 15 | 4 | 1 | 3 | 8 | 3.1 |
| 16 | 16 | 1 | 0 | 2 | 3 | 1.2 |
| 17 | 18 | 0 | 1 | 1 | 2 | 0.8 |
| 18 | 21 | 0 | 0 | 2 | 2 | 0.8 |
| 19 | 30 | 1 | 0 | 4 | 5 | 2.0 |
| 20 | 1 | 82 | 76 | 80 | 238 | 73.0 |
| 21 | 2 | 8 | 7 | 16 | 31 | 9.5 |
| 30 | 3 | 5 | 14 | 15 | 34 | 10.4 |
| DO\_7dAvgMin | 1 | 4 | 2 | 6 | 4 | 12 | 3.7 |
| 2 | 5 | 3 | 2 | 2 | 7 | 2.1 |
| 3 | 6 | 0 | 0 | 3 | 3 | 0.9 |
| 4 | 7 | 0 | 0 | 1 | 1 | 0.3 |
| 5 | 1 | 83 | 83 | 94 | 260 | 79.8 |
| 6 | 2 | 9 | 12 | 9 | 30 | 9.2 |
| 7 | 3 | 3 | 4 | 5 | 12 | 3.7 |
| DO\_7dMean | 1 | 4 | 3 | 4 | 3 | 10 | 3.1 |
| 2 | 5 | 2 | 2 | 6 | 10 | 3.1 |
| 3 | 6 | 0 | 0 | 3 | 3 | 0.9 |
| 4 | 7 | 0 | 0 | 1 | 1 | 0.3 |
| 5 | 1 | 27 | 37 | 36 | 100 | 39.4 |
| 6 | 2 | 5 | 5 | 3 | 13 | 5.1 |
| 7 | 3 | 3 | 7 | 6 | 16 | 6.3 |

\*Departures in cumulative percentage from 100% are attributed to rounding errors at the first decimal place

The results for the “PI ±0.5” appear somewhat similar to the “CI ±0.2” results. Comparing those two figures illustrates the balancing act between selecting the appropriate interval (confidence vs percentile? 90th vs another quantile?) and selecting the appropriate equivalence region. If an agency wants the average of multiple (shortened) rolling averages to be equivalent to the average of multiple (full) rolling averages, then the CI will quantify that. In contrast, if an agency wishes to see if a random, single, shortened rolling average is equivalent to a full rolling average, then PI should be used. What constitutes an appropriate equivalence region is a risk management decision that should be decided on independent of the statistical results.

The difference between the CI and PI results is profound and reflects what each interval is quantifying.. For example, refer to the middle panel of Fig8. In that example, the CI quantifies the average difference across all the points. In contrast, the PI quantifies the expected difference in a single point. Equivalence with individual samples will be harder to achieve than equivalence with monthly means. Whether to use CI or PI results depends on exactly what measurement – the average of multiple rolling average values, or a single value – the researcher wishes to test for equivalence. In general, equivalence was found more often at shorter sampling durations for CI than PI.

Recall that the research question is “How many continuous days are sufficient to approximate a 7-day mean, 7-day avg min, and a 30-day mean?” It is expected that the answer to this question will vary greatly on site-specific factors. Such factors may include discharge/watershed size, slope, percent of watershed development, and baseline DO level (e.g., typically high or typically low DO). Biological factors such as productivity, benthic and fish community composition, and available sunlight/riparian shading likely also play a part. Many of these “covariates” were not readily available for the sites and therefore were not analyzed. Future work that quantifies the relationship between these covariates and the estimated minimum number of days for equivalence will improve our understanding and provide nuance to this research question.

## Attainment Scenarios and Error Rates

DO criteria are often used to assess whether streams achieving their designated uses. Incorrect determination of whether a stream is meeting or exceeding DO criteria would lead to incorrect listing decisions, comprising situations where the stream is truly meeting the DO criteria but the assessment yields an exceedance (false positive) and situations where the stream is truly exceeding the DO criteria but the assessment indicates the stream is meeting DO criteria (false negative). To explore the type I (false positive) and II (false negative) errors associated with a shortened duration period, the decisions (exceed criteria or not exceed criteria) obtained using the full duration were compared to decisions obtained using the shortened duration. lists the DO criteria used in this analysis.

Table . DO criteria (mg/L) for cold and warm streams

|  |  |  |
| --- | --- | --- |
| DO Criteria | Cold | Warm |
| 30-day Mean | 6.5 | 5.5 |
| 7-day AvgMin | 5.0 | 4.0 |
| 7-day Mean | 6.5 | 6.0 |

For this analysis, information on whether a site is classified as a Cold or Warm stream was not available. Therefore, the dataset was analyzed twice, once assuming warm criteria and again assuming cold criteria. For each criterion, if a site had at least one exceedance using the full duration, it was labeled as *Exceeded*. Likewise, for each criterion, if a site had a least one exceedance using the shortened duration (1 day, 2 days, etc.), then the site was labeled as *Exceeded* for that duration length.

**Results – Attainment Scenario**

This setup of comparing binary outcomes between two datasets – full vs shortened – lends itself to a confusion matrix. For each class-criteria-duration combination, a confusion matrix was calculated. For example, displays the confusion matrix for Warm streams applying a 7-daymean criteria with a one day shortened window.

Table . Confusion matrix for Warm, 7-day mean criterion compared to a 1-day shortened duration.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 7-day Avg Min  (true condition) | |
|  |  | **Attain** | **Exceed** |
| 1-day Avg Min  (estimated condition) | **Attain** | 70  True negative | 0  False negative |
| **Exceed** | 17  False positive | 23  True positive |

Above, we see the sample size was 77+3+30 = 110 sites. The overall accuracy is 97.3% or (77+30)/110. The naïve accuracy (the accuracy obtained by simply guessing the majority decision) is (80+0)/110 or 72.3%. The naïve accuracy is a better baseline to compare the overall accuracy against, rather than comparing against a baseline of 50%. Type I and II error rates can also be quantified. The Type I error (false positive) rate equals 1 -77/(3+77) = 0.0375 or 3.8%. The Type II error (false negative) rate equals 1 – 30/(0+30) = 0. Type I and II errors are sometimes expressed as “1 minus specificity” and “1 minus sensitivity” results.

In total, 66 confusion matrices were calculated for each class-criteria-duration combination. We can plot the summary metrics – accuracy, type I error, etc. – from these 66 matrices to visualize how the shortened durations compared to the full duration decisions. displays the type I error rates vs the number of averaging days tested.

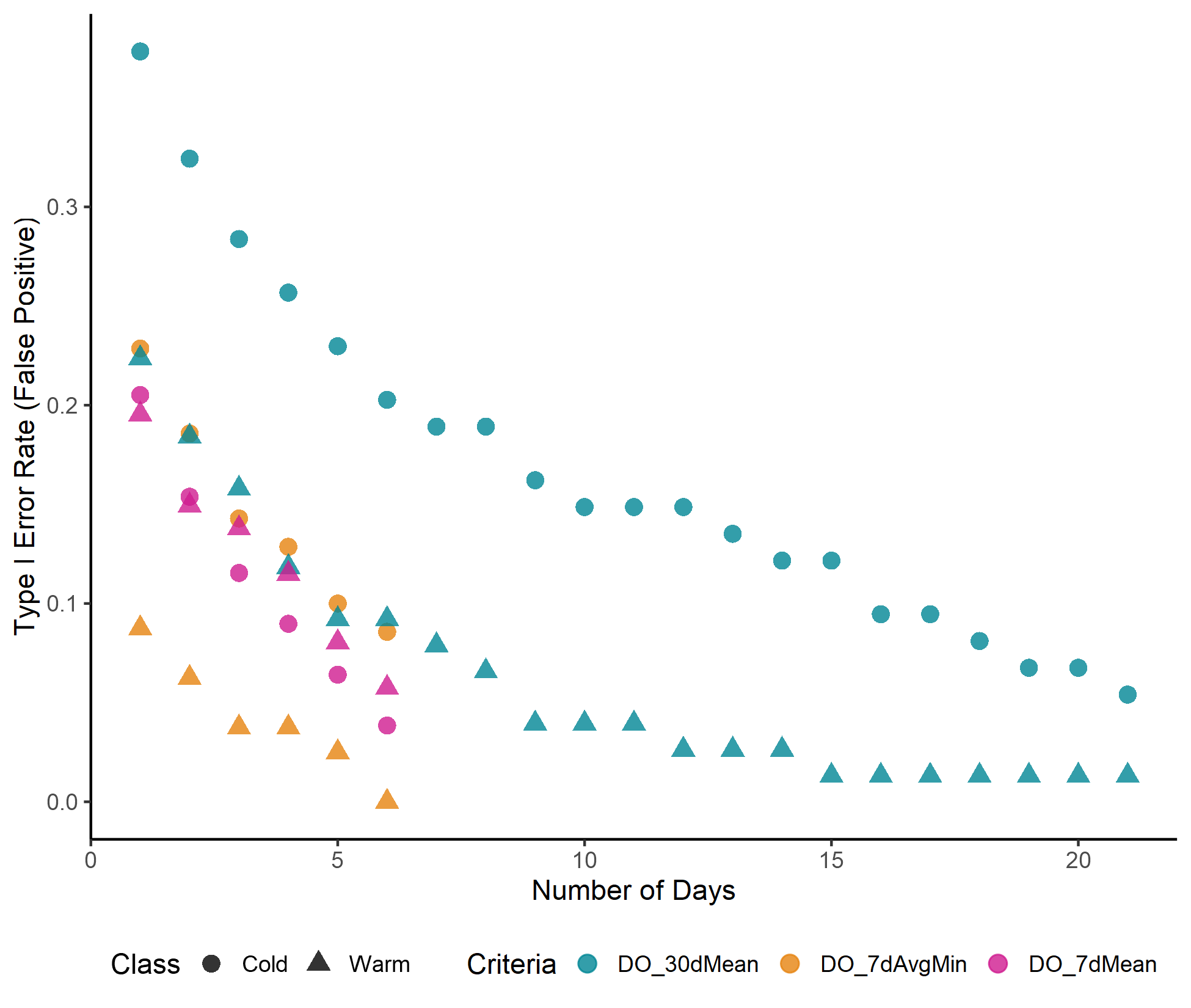


Figure . Scatterplot of false positive rate vs number of averaging days. Data from the confusion matrix results.

Unsurprisingly, the error rate declines as the averaging period approaches the full duration. False positive rates ranged from 0 to 0.38, with 30-day criteria having larger false positive rates than 7-day criteria, suggesting that it is more challenging to approximate 30-day criteria than it is to approximate 7-day criteria.

The 7-day Avg Min (Warm) showed the lowest Type I error rate. This may be because extreme statistics like the minimum are robust to perturbations in the dataset, compared to the mean. The minimum has a high breakdown point, meaning the estimate will not change unless a smaller value is added or removed to the dataset. In comparison, the estimate of the mean will fluctuate if virtually any sample is added or removed.

Overall, the Cold criteria had larger Type I error rates than the Warm criteria. Since each stream was tested twice using both criteria, this observed difference may suggest that more streams \*in this dataset\* were hovering closer to the Cold criteria than the Warm criteria. It should not be inferred that cold streams intrinsically have a larger error rate when it comes to approximating 7- and 30-day criteria.

The two analyses differ in multiple ways. One difference is that the equivalence margin has a large role in the equivalence analysis, but does not affect the error rate analysis. Another difference is that the magnitude of observed DO values does not affect the equivalence margin results, but does affect the error rate analysis. This is because the error rate analysis is sensitive to streams that are near the criteria. Achieving a low error rate is to be expected with a stream that is far away (above or below) the stated criteria. However, streams with DO values that hover near the criteria represent "edge cases" where achieving a low error rate is more challenging.

The streams selected for this analysis represent a broad selection of sites with sufficiently available continuous data. Sites were not selected based on their DO values or attainment status. That is, the dataset represents a wide range of DO values, as shown in .

# Conclusions and Recommendations

* The streams selected for this analysis represent a broad selection of sites with sufficiently available continuous data.
* Many sites were well approximated using fewer contiguous days of monitoring than the DO criteria specified:
  + 7-day mean and 7-day mean minimum DO were generally well-approximated by a single day of sampling
  + 30-day minimum DO was often not well-approximated by a single day or a small number of days of sampling.
* Approximating 30-day means was more challenging for shorter monitoring periods than the 7-day metrics, likely due to the increased chance of variation across a longer rolling average period.
* The choice of statistical interval and equivalence margin are important risk management choices that influence the statistical results. The choice of confidence interval (CI) vs percentile interval (PI) and the equivalence margin have a substantive effect on the equivalence results:
  + CI quantifies the average difference across all samples, whereas PI quantifies the expected difference for a single sample. Equivalence would be more difficult to achieve with a PI than for a CI, all else held equal.
  + Equivalence was less likely to be found with an equivalence margin of 0.2 mg/L than with an equivalence margin of 0.5 mg/L.
  + Equivalence for 7-day metrics was typically achieved with a single day of sampling for a CI and ±0.2 mg/L equivalence margin, a CI and ±0.5 mg/L equivalence margin, and a PI and ±0.5 mg/L equivalence margin. A range of days of sampling was required to well-approximate DO for a PI and ±0.2 mg/L equivalence margin.
  + Equivalence for 30-day minimum DO was achieved with a single day of sampling for 26.1% of tests for a CI and ±0.2 mg/L equivalence margin, 1.2% of tests for a PI and ±0.2 mg/L equivalence margin, 82.1% of tests for a CI and ±0.5 mg/L equivalence margin, and 39.7% of tests for a PI and ±0.5 mg/L equivalence margin. Conversely, sampling for greater than 21 days was required to achieve equivalence for 17.3% of tests for a CI and ±0.2 mg/L equivalence margin, 24.2% of tests for a PI and ±0.2 mg/L equivalence margin, 2.7% of tests for a CI and ±0.5 mg/L equivalence margin, and 2.7% of tests for a PI and ±0.5 mg/L equivalence margin.
* False negatives (i.e., situations where the stream is exceeding the DO criteria but the shortened assessment window indicates the stream meets the DO criteria) were never observed, regardless of the length of the assessment window or criteria measure.
* False positives (i.e., situations where the stream meets the DO criteria but the shortened assessment window indicates the stream exceeds the DO criteria) were observed in 0-38% of tests, occurring more frequently for the 30-day mean DO metric than for the 7-day DO metrics. False positives were more likely for shorter sampling durations. Choosing a sampling duration based on the likelihood of a false positive result may be an alternative option.
* The 7-day Avg Min showed the lowest Type I error. This may be due to the nature of the minimum statistic, which is not as easily influenced as the mean.
* The error rate analysis results are not impacted by the equivalence margin or the choice of interval. The results are impacted, however, by the magnitude of the observed DO values. That is, how close or far away the data are from the given criteria thresholds.
* Further research into the effect of covariates could improve our understanding of which factors (e.g., discharge, percent developed, etc.) affect the variability in equivalence results across sites.

# References

Berger, RL and JC Hsu. 1996. Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets. Statistical Science, 11(4), 283-302.

Lakens, D., AM Scheel, and PM Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. Advances in Methods and Practices in Psychological Science 2018, Vol. 1(2) 259–269. DOI: 10.1177/2515245918770963.

Lakens, Daniel. 2017. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Social Psychological an Personality Science. 8(4) 355–362.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Walker E., and AS Nowacki. 2010. Understanding Equivalence and Noninferiority Testing. Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA.

# Appendix

Table . List of stream sites used in the analysis

| Site Name | Site ID | State\* |
| --- | --- | --- |
| Animas River | AR 10-8 | CO |
| Animas River | AR 16-0 | CO |
| Animas River | AR 19-3 | CO |
| Animas River | AR 2-7 | CO |
| Beaver Run | 25990482-001 | PA |
| Bells Run | 57468543-001 | PA |
| Bennie Peer Creek | S4646 | MT |
| Big Elk Creek | 112189032-001 | PA |
| Birch Run | 25989452-001 | PA |
| Box Elder Creek | S4591 | MT |
| Box Elder Creek | S4603 | MT |
| Box Elder Creek | S4608 | MT |
| Browns Run | 66541519-001 | PA |
| Bull Creek | 50BullCr000.1 | NM |
| Campbells Run | 99687232-001 | PA |
| CANYONF | CANYONF | MT |
| Carley Brk | 25943266-001 | PA |
| Catatonk | Catatonk | SRB |
| Charlie Creek | S4650 | MT |
| Cherry Run | 123865227-001 | PA |
| Clark Fork | S4001 | MT |
| Clark Fork | S4004 | MT |
| Coal Bank Creek | S4619 | MT |
| Cooks Creek | 26054142-001 | PA |
| Deer Creek | S4406 | MT |
| Donegal Creek | 57464461-001 | PA |
| Dunbar Creek | 69919283-001 | PA |
| Squamscott River | SQ2 | NH |
| Squamscott River | SQ3 | NH |
| Fish Creek | Fish Creek | WY |
| Florida River (Animas) | FLR 1-2 | CO |
| Fox Creek | S4403 | MT |
| Goose Creek (Most) | 25621276-001 | PA |
| Goose Creek (Oak) | 25621300-001 | PA |
| Goose Creek (Thorn) | 25621340-001 | PA |
| Grays Run | 66910357-001 | PA |
| Groff Creek | 57462723-001 | PA |
| Hay Creek | S4606 | MT |
| Hell Run | 126216680-001 | PA |
| Horse Creek | S4621 | MT |
| Hunts | Hunts | SRB |
| Hyner Run | 61114835-001 | PA |
| Indian Creek (Berg) | 25999100-001 | PA |
| Indian Creek (Rt 63) | 25986884-001 | PA |
| Lick Run | 99408620-001 | PA |
| Little Beaver Creek | S4600 | MT |
| Little Beaver Creek | S4601 | MT |
| Little Beaver Creek | S4602 | MT |
| Little Beaver Creek | S4605 | MT |
| Little Beaver Creek | S4607 | MT |
| Little Beaver Creek (PA) | 57465467-001 | PA |
| Lone Tree Creek | S4647 | MT |
| Long Branch | 696\_15.0\_1.4 @ Jallopy Rd | MO |
| Long Branch | 696\_15.0\_1.9 @ Hwy 156 | MO |
| Long Branch | 696\_2.1 @ Kennedy St | MO |
| Long Branch | 696\_6.6 @ W Marion St | MO |
| Long Branch | 696\_9.3 @ Hwy J | MO |
| Long Grass Creek | S4679 | MT |
| Lower Flathead | C2700 | MT |
| Lower Flathead | C2777 | MT |
| Masthope Creek | 25869368-001 | PA |
| Meshoppen | Meshoppen | SRB |
| Middle Creek | 53321960-001 | PA |
| Missouri | S4018 | MT |
| MO-NUT | S4182 (MR1) | MT |
| MO-NUT | S4182 (MR2) | MT |
| MO-NUT | S4182 (MR4) | MT |
| MO-NUT | S4182 (MR5) | MT |
| MO-NUT | S4182 (MR6) | MT |
| MO-NUT | S4182 (MR7) | MT |
| MO-NUT | S4182 (MR8) | MT |
| MO-NUT | S4182 (MR9) | MT |
| MO-NUT | S4183-YSIDL | MT |
| MO-NUT | S4184-YSIDL | MT |
| MO-NUT | S4185-MDOT | MT |
| MO-NUT | S4186-YSIDL | MT |
| MO-NUT | S4187-YSIDL | MT |
| MO-NUT | S4188-YSIDL | MT |
| MO-NUT | S4189-YSIDL | MT |
| Moose | Moose | SRB |
| Moyers Mill Run | 54966869-001 | PA |
| Muddy Run | 57463365-001 | PA |
| OFallon Creek | S4644 | MT |
| Pecos River | 50PecosR540.8 | NM |
| Pecos River | 50PecosR770.0 | NM |
| Pennel Creek | S4611 | MT |
| Pennel Creek | S4613 | MT |
| Pennel Creek | S4615 | MT |
| Peters Creek | 99408584-001 | PA |
| Peters Creek | 99408690-001 | PA |
| Pickering Creek | 25980586-001 | PA |
| Pine Creek (Berks) | 25981122-001 | PA |
| Piney Fork Run | 99408674-001 | PA |
| Portage | Portage | SRB |
| Raccoon Creek | 66205927-001 | IN |
| Raritan River | BFBM000219 | NJ |
| Red Clay Creek | 26092962-001 | PA |
| Redwater Creek | S4693 | MT |
| Rife Run | 57462557-001 | PA |
| Rock Run | 66908783-001 | PA |
| Sandstone Creek | S4610 | MT |
| Sandstone Creek | S4618 | MT |
| Skippack Creek (Rt 63) | 25998866-001 | PA |
| Sterling | Sterling | SRB |
| Tinicum Creek | 26030712-001 | PA |
| Traverse Creek | 99685860-001 | PA |
| Vandenberg Drain | Y1757 | MT |
| Vandenberg Drain | Y1758 | MT |
| W Branch Brandywine Creek | 26105820-001 | PA |
| W Branch Octoraro Creek | 57469375-001 | PA |
| W Charlie Creek | S4649 | MT |
| Wildcat Creek | 10989-ORDEQ | OR |

\* SRB – Susquehanna River Basin