Causal Assessment Screening Tool (CASTool) User Guide

Table Contents

1. Background

Section 101 of the Clean Water Act establishes the goal to "restore and maintain the chemical, physical, and biological integrity of the Nation's waters." This goal is operationalized into objectives for specific waterbodies through water quality standards established by states, territories, and authorized tribes. Sections 305(b) and 303(d) of the Clean Water Act require these entities to conduct biannual assessments of the condition of their waterbodies, identify those that do not or are not expected to meet applicable water quality standards, and designate those waterbodies as impaired or threatened. For impaired and threatened waterbodies, states, territories, and authorized tribes must identify the pollutant(s) causing the impairment and establish a total maximum daily load (TMDL) of the pollutant required to achieve the applicable water quality standard. Identifying the pollutant causing impairment can be difficult when the impairment is identified through violation of a narrative or numeric expression of the qualities of a healthy organismal community—that is, a biological water quality standard.

The U.S. Environmental Protection Agency (EPA) previously developed a system for identifying the pollutants likely causing biological impairment, which it articulated in EPA's Stressor Identification Guidance Document (U.S. Environmental Protection Agency 2000) and the CADDIS (Causal Analysis/Diagnosis Decision Information System) website (U.S. Environmental Protection Agency 2010) (https://www.epa.gov/caddis). However, this process can be resource-intensive, which can create barriers to implementing both the development of biocriteria and the Stressor Identification process. As a result, states, territories, and tribes have repeatedly articulated the need for more streamlined methods of implementing EPA's Stressor Identification process. The Causal Assessment Screening Tool (CASTool) was developed to address this need in streams and rivers. It is a data-driven tool that leverages the biological, chemical, and physical monitoring data produced by assessment programs in a weight of evidence approach to screen causes of biological impairment. The first iterations of the CASTool were developed by Tetra Tech for use in

Arizona and the City of San Diego. The present version of the tool modifies this original tool and extends support for analyses across the United States.

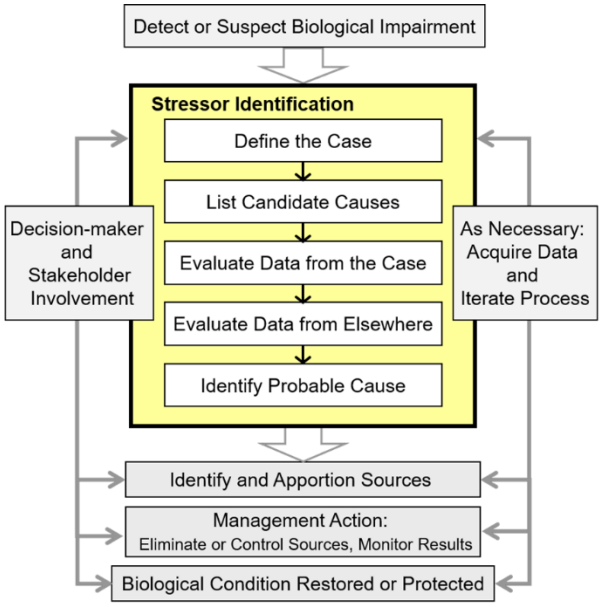


Figure 1. EPA's Stressor Identification process and its management context (reproduced from CADDIS).

2. CASTool overview

2.1 In brief

The CASTool is designed to follow the same basic steps of EPA's Stressor Identification process for a selected target site (U.S. Environmental Protection Agency 2000). The first step of the process requires defining the geographic scope of the analysis, including the sites that can be compared to the target site because they share similar characteristics and are unimpaired or differently impaired (Figure 1). The CASTool defines comparator sites more broadly than CADDIS, which advises selecting comparator sites belonging to the same system (e.g., sites upstream of the impairment) or watershed. The CASTool defines comparator sites according to the definition in (Gillett et al. 2019): "waterbodies that are ecologically similar to the biologically degraded test site (i.e., similar natural environmental setting, similar potential species recruitment), but have different levels of stressor exposure." The CASTool diverges from Gillett et al. (2019) in its approach to defining comparator sites. Gillett et al. (2019) define comparator sites based on expected biological similarity as defined by a taxonomic completeness (observed taxa divided by expected taxa, or O/E) model. Because not every region has an O/E model, the CASTool provides a

default comparator assignment algorithm based on the similarity of non-anthropogenic factors including streamflow, lithology, and climate determined using a clustering algorithm (Appendix. Additional methodological details). Using these types of environmental parameters to predict community structure is a component of developing an O/E model. Users can provide alternative clustering assignments if desired.

The second step of the Stressor Identification process requires users to select candidate causes of impairment at the selected target site (Figure 1). The CASTool initially considers all stressors that have been measured at the target site and selected by the user for evaluation as candidate causes. Because it is a data-driven tool, the CASTool cannot consider stressors that have not been measured at the target site. To help address this limitation and to provide another source of data that can be used to evaluate candidate causes, the CASTool provides a module to compare watershed stressors (e.g., land cover, dams, burned area, road density, polluted site density, and nutrient mass balances) at the target site and comparator sites (Figure 2). The CASTool generates a table of watershed stressor values elevated at the target site, defined by values greater than the median of comparator site values. Pollutants associated with watershed stressors that are elevated at a target site may merit additional consideration as causes of impairment, including additional data collection and analysis.

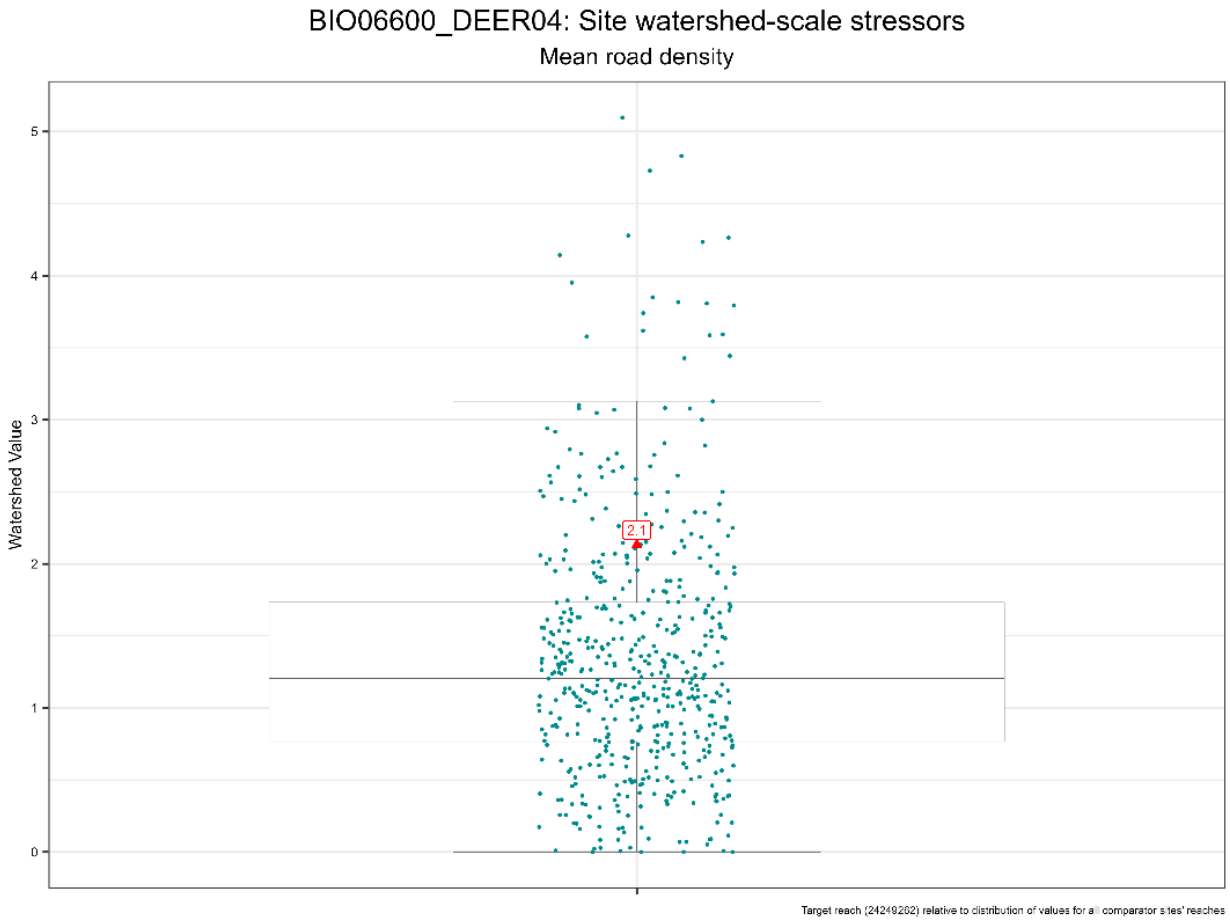


Figure 2. Example of a figure produced by the CASTool watershed stressor analysis. The blue points represent mean road density in the watersheds of comparator sites. The red triangle and label represent the watershed mean road density of the target site. The underlying boxplot summarizes the distribution of the watershed stressor values.

The third and fourth steps of the Stressor Identification process require users to evaluate stressor-response relationships using multiple types, or lines, of evidence (Figure 1, Table 1). The CASTool supports evaluation of macroinvertebrate, algae, and/or fish community responses to any stressor. To run the CASTool, the user provides formatted data in a zipped folder. To pair stressor and response observations, the user specifies a time window in terms of the number of days before and after the response observation that a stressor observation can be considered paired. If multiple stressor observations fall within the acceptable time window, the CASTool selects the stressor observation within the least number of days (before or after) of the response observation. This paired stressor-response dataset is used for all analyses. Users with sub-daily stressor observations are encouraged to calculate daily summaries to use as inputs to the CASTool.

EPA's Stressor Identification process distinguishes lines of evidence derived from comparator sites, referred to as "inside the case," and lines of evidence derived from non-comparator sites, referred to as "outside the case." In the CASTool, this distinction is operationalized using data from sites belonging to the same cluster (inside the case) versus a different cluster (outside the case). For each candidate cause (i.e., each stressor measured at the target site), a figure is generated for each line of evidence and each line of evidence is scored based on the strength of the evidence for or against each stressor as the cause of biological impairment, with -1 indicating evidence against the stressor, 0 as neutral or no evidence, and 1 indicating evidence for the stressor as the cause of impairment.

Most lines of evidence in the CASTool use a single biological community metric as the response. This metric, referred to as the "index," defines whether the associated biological community sample is considered impaired. Some CASTool analyses use metrics designed to assess impairment due to specific stressors as the response. These stressor-specific metrics can be provided to the CASTool as summary values for the entire biological community ("stressor-specific indices") or as tolerance values that apply to individual taxa ("stressor-specific tolerance values"). The CASTool can analyze other biological community metrics provided by the user, but the results of these analyses are only included in supplemental figures and are not included in the summary weight of evidence tables; these tables focus solely on the user-identified biological index and stressor-specific metrics. Each biological community sample from the target site is evaluated independently, as biological index values can change over time and site conditions.

The first line of evidence, co-occurrence, examines the target sample stressor value relative to the distribution of stressors from unimpaired comparator samples (i.e., those with index values indicative of unimpaired status). This line of evidence serves as a screening step for advancing stressors to analyses based on the other lines of evidence. For most stressors (e.g., nutrients, conductivity), stress increases with increasing stressor values. If the target sample stressor value is not elevated relative to those of the unimpaired comparator samples, the stressor is assigned a score of –1, indicating evidence against the stressor as the candidate cause. If all paired stressor samples from a target site are assigned a score of –1 for the co-occurrence line of evidence for a particular stressor, that stressor is not advanced to the other analyses as a candidate cause, as the data do not support the stressor as the cause of impairment. If stress decreases with increasing values of the stressor (e.g., pH, dissolved oxygen), the stressor is not advanced as a candidate cause if all paired sample stressor values at the target site are not low relative to unimpaired comparator samples. For pH and dissolved oxygen, the user can specify

thresholds above and below which these stressors are advanced as candidate causes of impairment regardless of the co-occurrence line of evidence score.

The final step of the Stressor Identification process is identifying the probable cause(s) of impairment using a weight of evidence approach. The CASTool displays a weight of evidence summary table showing the number of lines of evidence providing supporting, refuting, and indeterminate evidence for each candidate cause, as well as a table showing the scores for each line of evidence. As noted above, each biological community sample is scored independently and summarized in the weight of evidence summary and lines of evidence tables. However, all samples from a target site are plotted on the same line of evidence figures. Users are encouraged to view the accompanying line of evidence figures to corroborate automated scoring.

The final output from the CASTool is a zipped folder with a summary report and all accompanying figures and tables, including a table outlining the data gaps encountered. The CASTool can be run in an interactive point-and-click Shiny application or locally in a user's R console. The Shiny application is available (WEBLINK) and the R program is available (WEBLINK).

2.2 Lines of evidence

Table 1. Lines of evidence analyzed in the CASTool.

Line of evidence	**Case**	**Function**	**Evaluated when**
Co-occurrence	Inside the case	Is the target sample stressor value elevated compared to those from unimpaired sites?	Always
Sufficiency	Inside the case	Is the target sample stressor value associated with a higher probability of biological impairment?	When a candidate cause is advanced by the co-occurrence analysis
Biological gradient	Inside the case; Outside the case	Is there a linear relationship between the stressor and biological response?	When a candidate cause is advanced by the co-occurrence analysis

Time sequence	Inside the case	Does elevated stress precede the biological response?	When a candidate cause is advanced by the co-occurrence analysis
Stressor-specific tolerance values	Inside the case	Does the target sample have a lower number of individuals, percent of individuals, number of taxa, and percent of all sensitive and the most sensitive taxa compared to those from unimpaired comparator sites?	When a candidate cause is advanced by the co-occurrence analysis and stressor-specific tolerance values are provided for the candidate cause
Stressor-specific index: co-occurrence	Inside the case	Is the target sample stressor-specific index elevated compared to those from unimpaired sites?	When a candidate cause is advanced by the co-occurrence analysis and stressor specific indices are provided for the candidate cause
Stressor-specific index: sufficiency	Inside the case	Is the target sample stressor-specific index value associated with a higher probability of biological impairment?	When a candidate cause is advanced by the co-occurrence analysis and stressor specific indices are provided for the candidate cause

2.3 General caveats and limitations

- The CASTool is intended as a screening level tool. It cannot diagnose causes of biological impairment and is most effectively used to identify candidate causes with little support in the data and to inform future data collection.
- CASTool outputs are only as good as the data inputs. For example, the tool assumes that stressor data are meaningfully representative of conditions experienced by the matched biotic community sample. This assumption can be problematic for parameters with high variability, such as those sensitive to flow.
- The CASTool evaluates each stressor independently; it does not consider interactions among stressors.
- The CASTool biological gradient line of evidence evaluates the linear relationship between stressor and biological response. It may fail to adequately characterize evidence for a candidate cause when the stressor-response relationship is non-linear. For this reason, among others, users are urged to review evidence figures to corroborate automated scoring.
- The CASTool assumes that all observations with shared identifiers (e.g., observations with the same stressor parameter name) are comparable. It cannot, for example, account for changes in sampling or analytical methods.
- CASTool assumes that the user has formatted the input data and has dealt with non-detects in a manner consistent with their program.
- Biological community metrics can be more responsive by design to certain stressors than others (Jones et al. 2023), meaning that the CASTool analyses can inadvertently bias the analysis toward identifying certain stressors as candidate causes of impairment. Understanding the derivation of the index used in the CASTool to determine biological impairment and supplementing the analysis with other metrics can help combat this bias.

2.4 Definitions

candidate cause: a stressor under consideration as a cause of biological impairment.

cluster: a group of stream reaches assembled based on the expectation that they will have comparable biological responses to stressor exposure (e.g., because of their similar environmental characteristics).

comparator site: a site that can be compared to the target site because they share similar characteristics but are differently impaired.

index: the biological community response metric used to determine whether a sample is impaired.

line of evidence: an analysis used to evaluate the evidence for a candidate cause of biological impairment.

metric: a biological community response variable.

region: the geographic area (e.g., a state) that encompasses the target and comparator sites to be analyzed.

sampled site: a site included in the user's sites input file.

target site: the site with sampled stressor and biological response data that is selected to be analyzed by the CASTool.

weight of evidence: the approach of combining multiple lines of evidence to provide a more comprehensive assessment of the evidence for a candidate cause of biological impairment.

2.5 Lines of evidence figures and scoring

Every line of evidence for each stressor generates a figure. As noted above, while each target site sample is scored independently, every sample from the target site is displayed on the lines of evidence figures. The title of each figure contains the target site Sample ID and corresponding line of evidence.

2.5.1 Co-occurrence

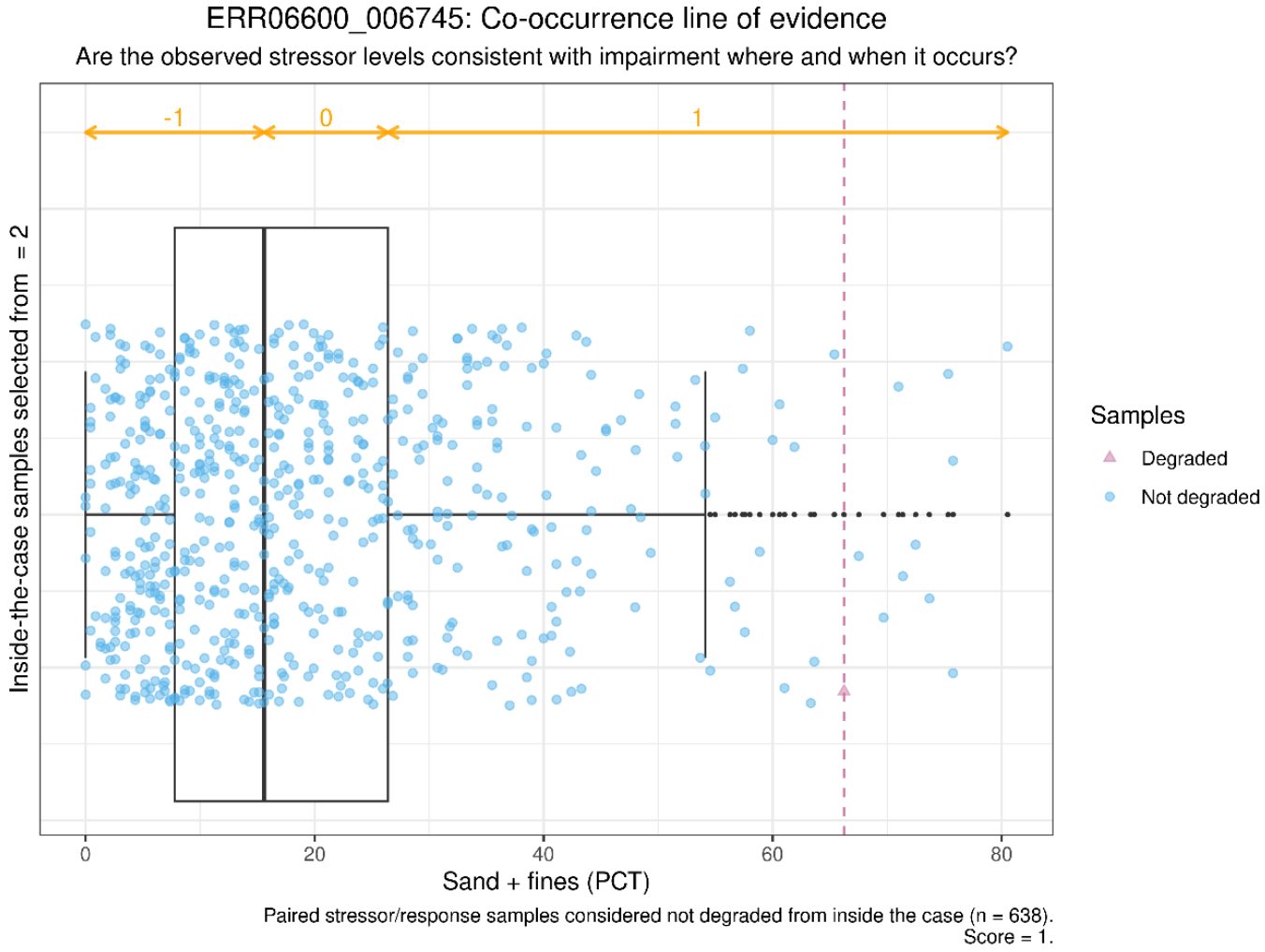


Figure 3. An example co-occurrence line of evidence boxplot illustrating the distribution of stressor values for comparator samples with biological index scores indicative of non-degraded condition. Comparator samples are depicted with blue points. The vertical dashed pink line depicts the target sample stressor value. The horizontal yellow arrows across the top of the figure show the scores that would be assigned to target sample values within their ranges. The vertical axis label indicates the cluster number of the target site and

its comparators. The text below the x axis indicates the number of non-degraded comparator samples and the assigned target sample score.

Scoring

Stress increases with higher values of the stressor:

-1 = value < 50th percentile of non-degraded comparator samples

0 = 50th percentile < value < 75th percentile of non-degraded comparator samples

1 = value > 75th percentile of non-degraded comparator samples

Stress decreases with higher values of the stressor:

-1 = value > 50th percentile of non-degraded comparator samples

0 = p25 < value < 50th percentile of non-degraded comparator samples

1 = value < 25th percentile of non-degraded comparator samples

Candidate causes scored -1 for the co-occurrence line of evidence for all target site samples are not advanced to the other analyses.

Example interpretation

The orientation of the yellow score arrows at the top of the figure illustrates that stress increases with increasing values of Sands + fines. The target sample value (depicted by the vertical pink line) is greater than the 75th percentile of non-degraded comparator samples, thus this sample receives a score of 1 (evidence for the candidate cause) and Sands + fines is advanced as a candidate cause to the other analyses.

2.5.2 Sufficiency

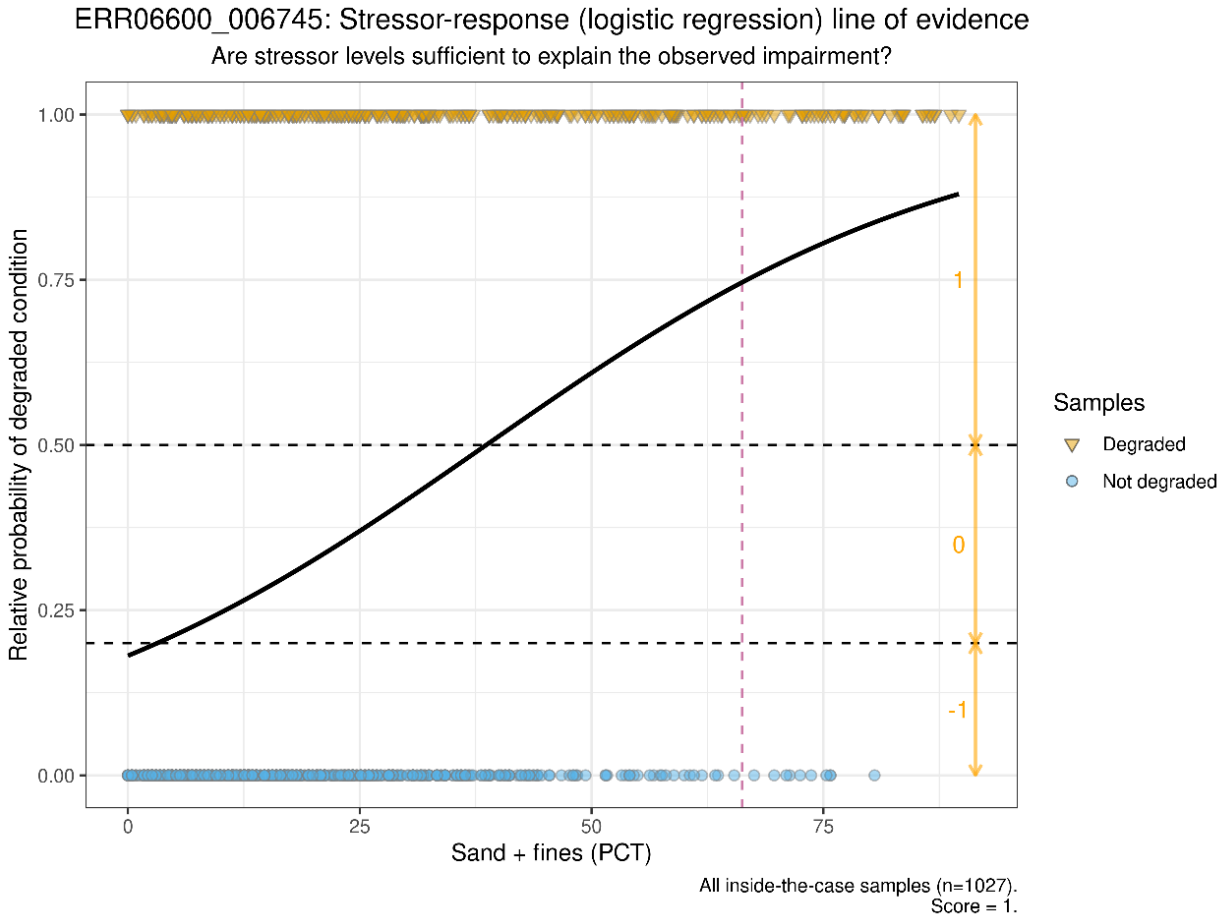


Figure 4. An example sufficiency line of evidence figure. The solid black line depicts the fit line from a logistic regression with sample stressor value as the predictor and the biological index condition (binary degraded vs. not degraded) as the response. The points along the top and bottom of the figure represent the stressor values paired with biological samples that are degraded and not degraded, respectively. The vertical, dashed, pink line depicts the target sample stressor value. The y value of the intersection between this line and the solid black regression line represents the relative probability of degraded condition based on the stressor sample value. Scores for this line of evidence are based on thresholds for the relative probability of degraded condition which are depicted by the dashed horizontal black lines. The vertical yellow arrows show the scores assigned to the relative probabilities within their ranges. The text below the x axis indicates the number of samples used to develop the regression and the target sample score.

Scoring

-1 = value corresponds to <20% probability of degraded condition

0 = value corresponds to <50% but >20% probability of degraded condition

1 = value corresponds to >50% probability of degraded condition

Note that the scoring algorithm does not account for the fit of the logistic regression, only the predicted probability of the sample stressor value.

Example interpretation

The target sample stressor value is associated with a >50% relative probability of degraded condition and therefore receives a score of 1 (evidence for the candidate cause). The vertical pink line indicating the stressor sample value intersects the regression line above the 50% horizontal threshold line, in the range of a score of 1.

2.5.3 Biological gradient (inside and outside the case)

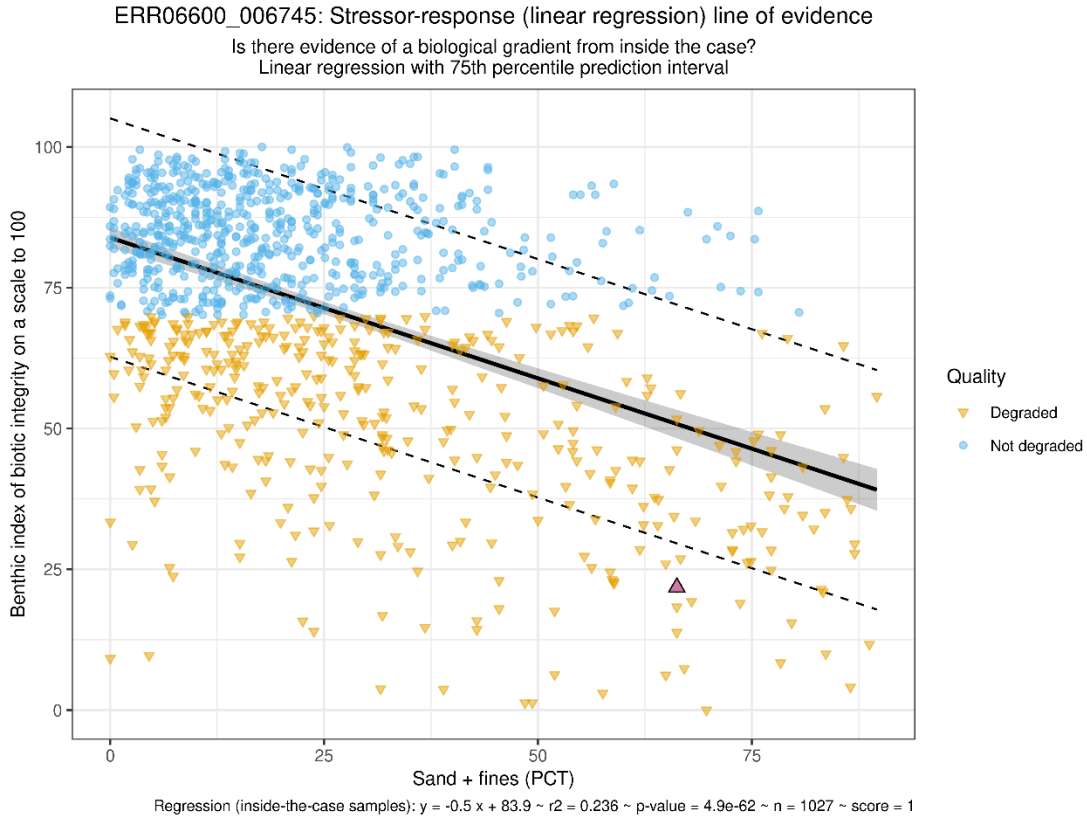


Figure 5. An example biological gradient (inside the case) line of evidence figure. The solid black line represents the linear regression between the stressor and biological response. The points represent sample values from comparator sites. The shape and color of points correspond to the binary index score (degraded vs. not degraded), and the pink triangle represents the target sample value. The dashed lines represent the 75th percent prediction interval and the shaded band represents the 95th percent confidence interval of the

regression. Regression parameters including the equation, R^2, and p-value are depicted below the x axis, along with the number of comparator samples used to develop the relationship and the line of evidence score.

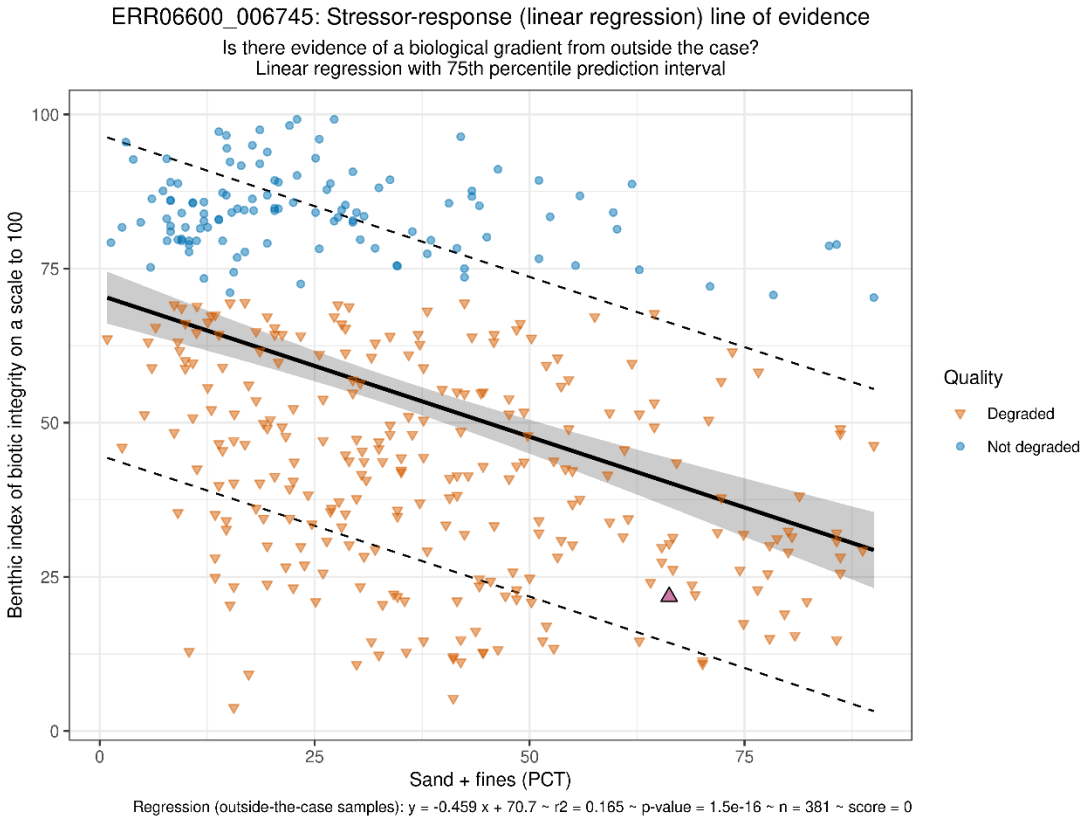


Figure 6. An example biological gradient (outside the case) line of evidence figure. The solid black line represents the linear regression between the stressor and biological response. The points represent sample values from non-comparator sites (those belonging to a different cluster than the target site). The shape and color of points correspond to the binary index score (degraded vs. not degraded), and the pink triangle represents the target sample value. The dashed lines represent the 75th percent prediction interval and the shaded band represents the 95th percent confidence interval of the regression. Regression parameters including the equation, R^2, and p-value are depicted below the x axis, along with the number of comparator samples used to develop the relationship and the line of evidence score.

Scoring

The biological gradient line of evidence is scored the same for inside and outside the case.

1 = the p-value of the slope is < a user-specified threshold, the R^2 is > a user-specified threshold, and the slope has the expected sign

0 = the p-value of the slope is not < the user-specified threshold or the R^2 is not > the user specified-threshold

-1 = the p-value of the slope is < a user-specified threshold, the R^2 is > a user-specified threshold, but the slope does not have the expected sign

Note that the biological gradient line of evidence only evaluates linear relationships. The strength of the evidence for non-linear stressor-response relationships may not be appropriately captured by the analysis.

Example interpretation

The biological gradient inside the case line of evidence receives a score of 1 (evidence for the candidate cause) because the R^2 of the regression exceeds the user-specified threshold of 0.2, the p-value is less than the user-specified threshold of 0.1, and the slope of the relationship is negative, which was the expectation for the relationship between the biological index and stressor (Sands + fines). The outside the case line of evidence receives a score of 0 (indeterminate evidence for the candidate cause), because the R^2 did not exceed the user-specified threshold.

2.5.4 Time sequence

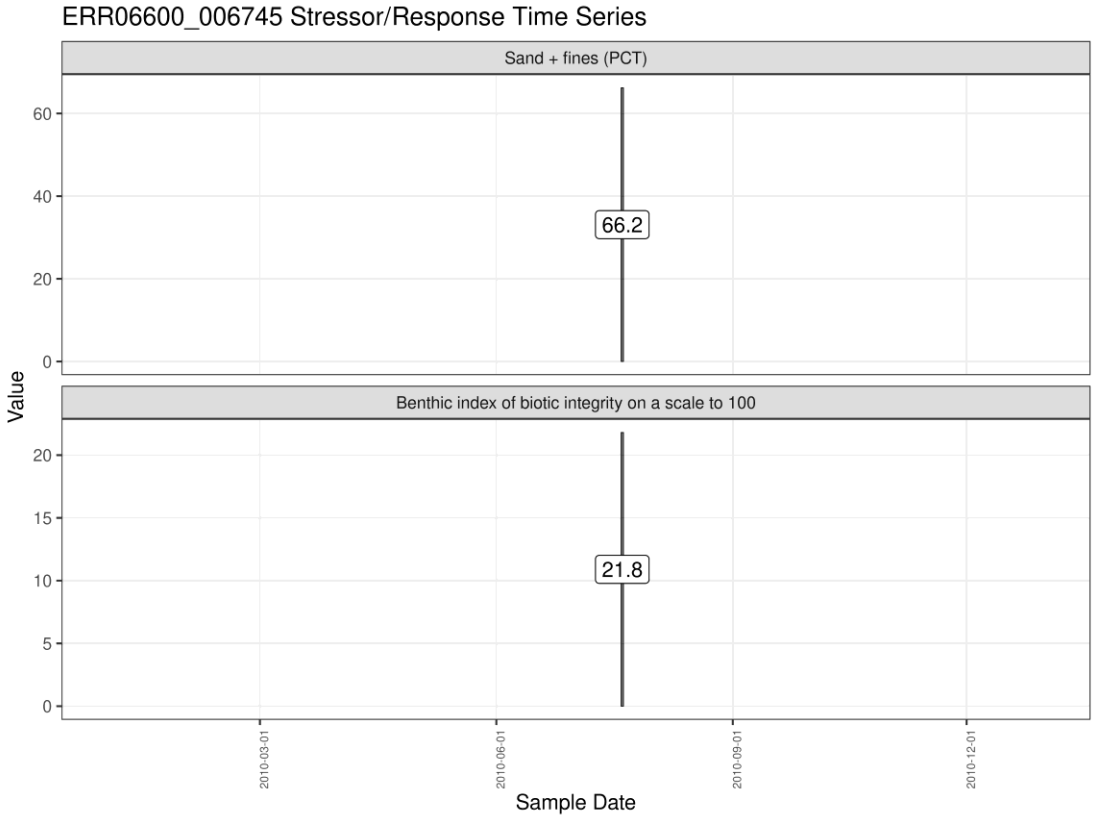


Figure 7. Example of the time sequence line of evidence figure. The top panel depicts the timing and value of stressor samples, and the bottom panel depicts the timing and value of biological response metric samples. The number displayed over the bars is the corresponding stressor or response metric value.

Scoring

The objective of the time sequence line of evidence is to determine whether the stressor precedes the biological response. Because there are infrequently sufficient samples to determine the time sequence of stressor and response, this line of evidence is not automatically scored.

Example interpretation

This example cannot be interpreted for the time sequence line of evidence because there is only one concurrent sample of stressor and response.

2.5.5 Stressor-specific tolerance values

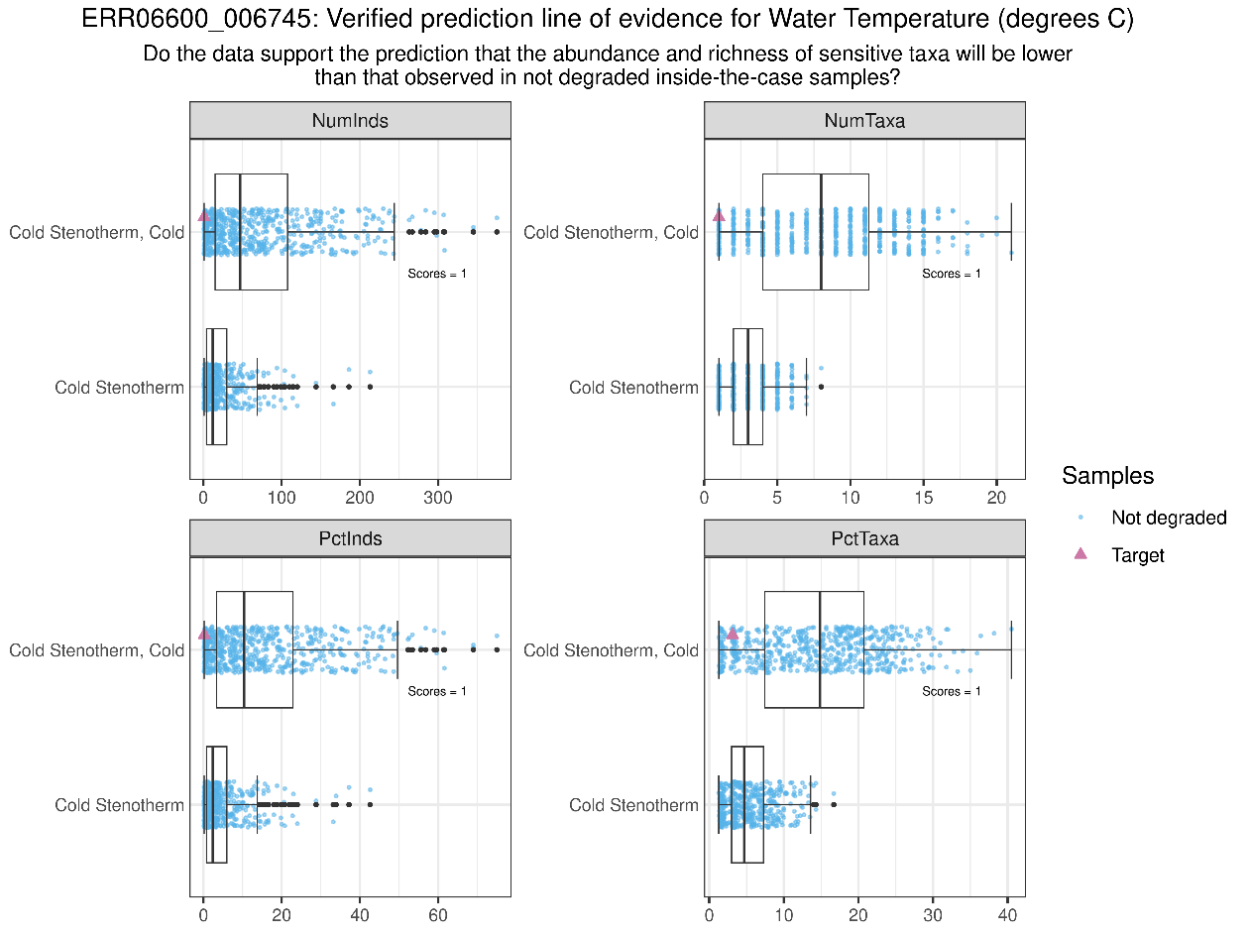


Figure 8. Example of the stressor-specific tolerance values line of evidence figure. Each panel represents one of the four component metrics (number of individuals, number of taxa, percent of individuals, and percent of taxa) that are calculated for each of two sensitivity classes (sensitive and most sensitive). The boxplots represent the distribution of component metrics for non-degraded comparator samples, which are depicted with blue points. The pink triangles represent the target sample values. Scores for each calculated component metric are displayed on the figure.

Scoring

The stressor specific tolerance value score is comprised of 4 summary statistics (number of individuals, number of taxa, percent of individuals, and percent of taxa) for 2 categories of sensitive taxa (sensitive and most sensitive), totaling 8 components, which are averaged. Assignments to sensitivity classes can be based on numeric scores or assigned categories, as in the example figure. Each component is scored according to the comparison of the target sample value and non-degraded comparator samples:

-1 = value > 50th percentile

0 = 25th percentile < value < 50th percentile

1 = value < 25th percentile

Example interpretation

In this example, sensitivity classes were determined using categorical assignments, with cold stenotherms assigned to the most sensitive class and both cold and cold stenotherms assigned to the sensitive class. The example figure does not contain scores for the most sensitive class (cold stenotherm) because no taxa in the target sample were cold stenotherms. For all four metrics, the target sample values were less than the 25th percentile of non-degraded comparator samples, thus each component metric received a score of 1 (evidence for the candidate cause). The overall stressor-specific tolerance value score is the average of the components, which in the example is 1.

2.5.6 Stressor-specific index – co-occurrence

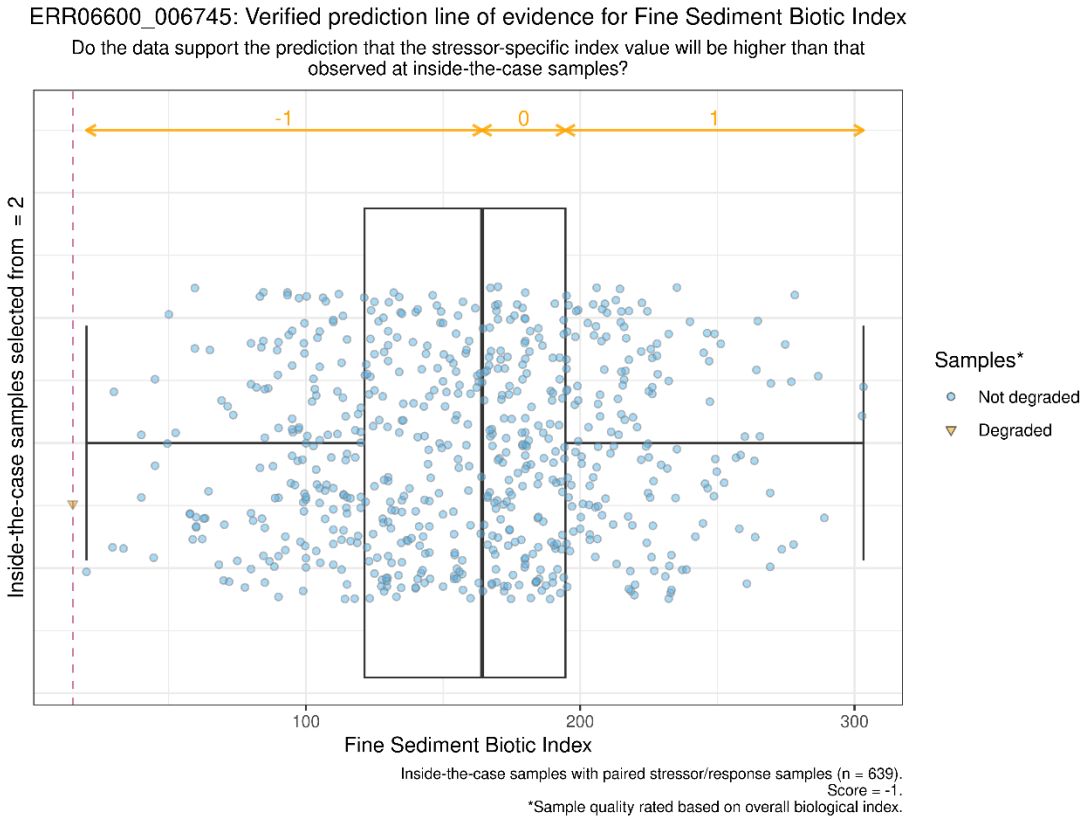


Figure 9. Example of the stressor-specific index – co-occurrence line of evidence figure. The boxplot depicts the distribution of the stressor-specific index for non-degraded comparator samples, which are displayed as the blue points. The vertical, dashed, pink line represents

the target sample value of the stressor-specific index. The horizontal yellow arrows represent the range of stressor-specific index values that would be assigned each possible score. The text below the x axis indicates the number of non-degraded comparator samples with stressor-specific index values and the target sample score.

Scoring

The stressor-specific index co-occurrence line of evidence follows the same scoring scheme as the co-occurrence line of evidence.

Stress increases with higher values of the stressor:

-1 = value < 50th percentile

0 = 50th percentile < value < 75th percentile

1 = value > 75th percentile

Stress decreases with higher values of the stressor:

-1 = value > 50th percentile

0 = p25 < value < 50th percentile

1 = value < 25th percentile

Example interpretation

The target sample stressor-specific index value (depicted by the horizontal, dashed, pink line) is less than the 50th percentile of not-degraded comparator samples. Accordingly, this sample receives a score of -1 (evidence against the candidate cause).

2.5.7 Stressor-specific index—sufficiency

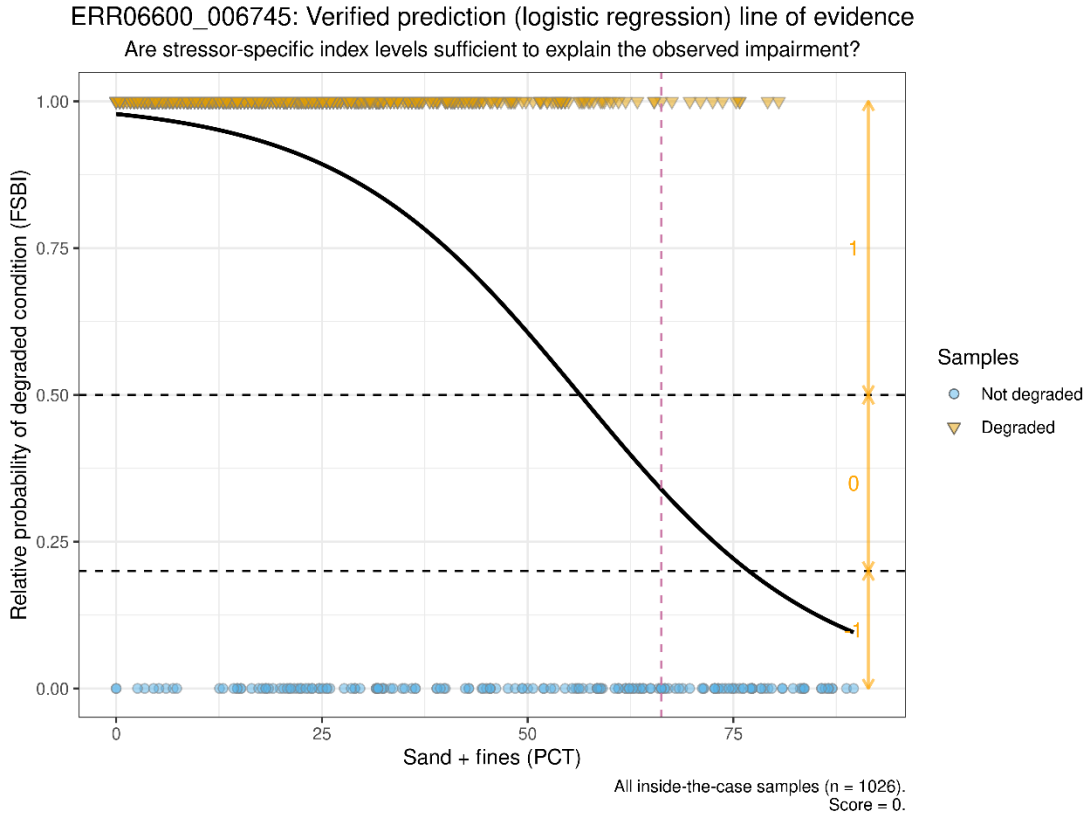


Figure 10. Example of the stressor-specific index – sufficiency line of evidence figure. The solid black line depicts the fit line from a logistic regression with sample stressor value as the predictor and the stressor-specific index condition (binary degraded vs. not degraded) as the response. The points along the top and bottom of the figure represent the stressor values paired with biological samples that are degraded and not degraded, respectively, as scored by the stressor-specific index. The vertical, dashed, pink line depicts the target sample stressor value. The y value of the intersection between this line and the solid black regression line represents the relative probability of degraded condition based on the stressor sample value. Scores for this line of evidence are based on thresholds for the relative probability of degraded condition which are depicted by the dashed horizontal black lines. The vertical yellow arrows show the scores assigned to the relative probabilities within their ranges. The text below the x axis indicates the number of samples used to develop the regression and the target sample score.

Scoring

The stressor-specific index—sufficiency line of evidence is only calculated if the user provides a threshold for the index that when exceeded indicates impairment. If such

threshold is not provided, only the stressor-specific index—co-occurrence line of evidence is calculated.

1 = value corresponds to <20% probability of impairment

0 = value corresponds to <50% but >20% probability of impairment

1 = value corresponds to >50% probability of impairment

Example interpretation

The target sample stressor-specific index value intersects the regression line between the two dashed horizontal threshold lines, indicating a probability of impairment <50% but >20%. Accordingly, the target sample receives a score of 0 (indeterminate evidence for the candidate cause).

2.6 Modifiable parameters

Table 2. A selection of parameters in the CASTool metadata modifiable for different runs of the tool with the same data.

Variable	Function	Acceptable values
exploreWSStressor	Select whether the CASTool conducts the watershed stressor analyses and generates associated figures.	TRUE/FALSE
removeOutliers	Select whether the CASTool removes outliers prior to conducting the lines of evidence analyses (Appendix. Additional methodological details).	TRUE/FALSE
samplim	Select the minimum number of samples required to analyze a candidate cause.	Positive integer
DOlim	Identify the DO limit below which DO is advanced as a candidate cause regardless of the co-occurrence analysis results.	Positive real number
pHlimLow	Identify the pH limit below which pH is advanced as a candidate cause regardless of the co-occurrence analysis results.	Positive real number
pHlimHigh	Identify the pH limit above which pH is advanced as a candidate cause regardless of the co-occurrence analysis results.	Positive real number
lagdays	Identify the number of days before and after a response sample is collected that a stressor sample can be collected and considered paired.	Positive integer
r2_cutoff	Select the R^2 cutoff value above which the biological gradient analysis will not be assigned a score of 1, indicating support for a candidate cause.	Real number between 0 and 1
p.val_cutoff	Select the p-value cutoff above which the biological gradient analysis will not be	Real number between 0 and 1

	assigned a score of 1, indicating support for a candidate cause.	
useAllCompReaches	Identify whether to use all "inside-the-case" reaches in the watershed stressor analysis including those that do not have sampled sites.	TRUE/FALSE
clusterNumber	Select the desired number of clusters for the analysis region. Only required when using cluster assignments generated by the CASTool (see Input Data Preparation). Default selects the cluster number chosen by the CASTool clustering algorithm (Appendix A.2).	default, 1, 2, 3, 4, 5

2.7 Output folder structure

The CASTool output consists of nested folders containing the summary report and associated figures and tables (Figure 11). The main output folder from the Shiny application will be named with the date and time the CASTool was run (e.g., CASTool_report_results_20260201_091216). The Region subfolder contains a status file for each run of the CASTool which describes whether a report was generated for each target site included in a CASTool run (one target site if run from the Shiny application, multiple sites possible if run from the R program). The table will include an explanation if a report was not generated for a target site.

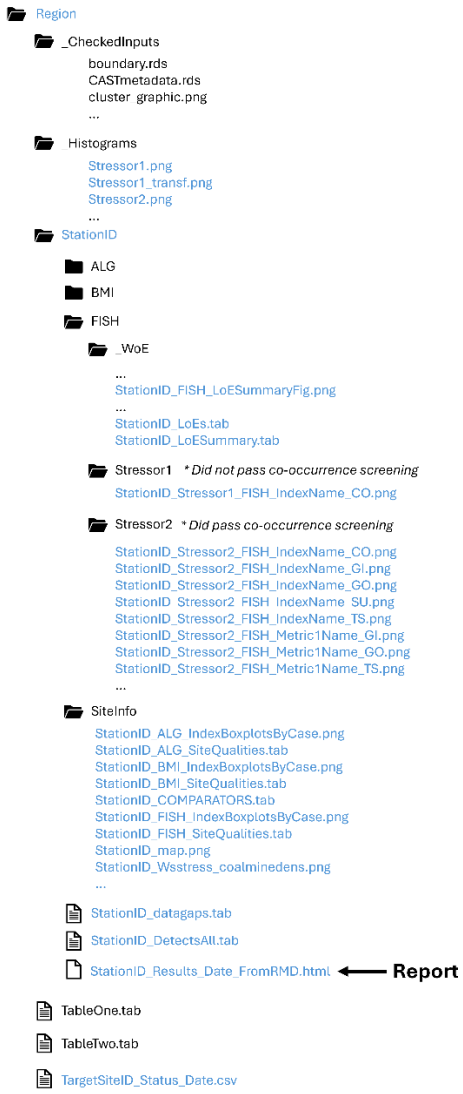


Figure 11. Structure of the Results folder output by the CASTool. Text in blue denotes folder and file names dependent on user inputs.

3. Input data preparation

Templates for CASTool input data files are available for download from the Check File Inputs tab of the Shiny application. Input files are provided to the CASTool in a zipped folder.

A key required input file is _CASTool_Metadata.xlsx, which defines many of the parameters used to set up a CASTool run, including the names of other input files. The metadata file template describes the function of each metadata variable along with its data type, acceptable values ("Domain"), whether it is required, and default value.

For some regions, some of the required geospatial and cluster input files can be loaded automatically to the CASTool from helper packages (see list of available regions). If your region is not available in the helper package, you can generate these files using the CustomBoundary package in R (Section 3.2 Geospatial and cluster file generation with a custom boundary).

The CASTool currently supports analysis of three biological communities: benthic macroinvertebrates (bmi), algae (alg), and fish. Data for at least one biological community must be provided to the CASTool.

The CASTool supports two different types of stressor data: 1) values that can be associated with a particular sampling date and site, which the tool refers to as measured data; and 2) values that are associated with only a site and not a date, which the tool refers to as modeled data. For example, in CASTool terminology, a field conductivity measurement taken on a particular date at a particular site is measured data, while a time-integrated estimate of stream flashiness at a particular site is modeled data. The user must provide the CASTool at least measured or modeled data.

3.1 Input data files

Each input data file in the templates folder has a tab for data entries and a tab for instructions on how to complete the data entries (Table 3).

If your region is not included in the CASTool helper packages, you must also include the files output from the CustomBoundary package for your region: 1) RegionName_Boundary.rda, 2) RegionName_ClusterGraphic.png, 3) RegionName_Clusters.csv, 4) RegionName_Reaches.rda. If you would like to conduct the watershed stressor analysis for a region not included in the helper packages, you must also include: 1) RegionName_WSStressor.rda and 2) RegionName_WSStressorInfo.rda, which can be output from the CustomBoundary package.

Metric and stressor names should not contain special characters like %.

Table 3. Input data files for the CASTool.

Template file name	Purpose	Required
_CASTool_Metadata.xlsx	Provides the names of the other input data files and specifies values for modifiable parameters (Table 2).	Yes
Sites.xlsx	Provides a unique identifier for each sampled site, the NHDPlusV2 flowline feature on which the site is located, coordinates, and an indication of whether a site is considered by the user to be a reference.	Yes
TargetSites.xlsx	Unique identifiers of the locations in the sites file that can be selected as a target site for analyses.	Yes
AlgaeMetrics.xlsx FishMetrics.xlsx BenthicMetrics.xlsx	Response metrics for the corresponding biological community.	At least one required.
AlgaeMetricsMetadata.xlsx FishMetricsMetadata.xlsx BenthicMetricsMetadata.xlsx	Information about the response metrics included.	Required for each biological response community with metrics data.
AlgaeCount.xlsx FishCount.xlsx BenthicCount.xlsx	Taxa count data for the corresponding biological response community*.	Required for biological response communities with stressor-specific tolerance values.
AlgaeCountMetadata.xlsx FishCountMetadata.xlsx BenthicCountMetadata.xlsx	Stressor-specific tolerance values for taxa in the corresponding count data files.	Required for each biological response

		community with count data.
MeasuredStressorData.xlsx ModeledStressorData.xlsx	Stressor data (see above for distinction between measured and modeled stressor data).	At least one required.
MeasuredStressorMetadata.xlsx ModeledStressorMetadata.xlsx	Information about the stressor data, including whether the tool should log(1 + x) transform the data before analysis.	Required for each stressor data file provided.

* Note: the CASTool does not provide a taxonomic harmonization routine, so the count files must have taxa representing the desired operational taxonomic units. Relative abundance, which is used to calculate stressor-specific tolerance values, is calculated as the number of individuals belonging to a taxon in a sample divided by the total number of individuals across all taxa present in a sample. The CASTool also does not have a subsampling routine, so the count files must already have subsampling represented in the taxa counts if desired.

3.2 Geospatial and cluster file generation with a custom boundary

Note: we are working to streamline this process, so these instructions should be considered provisional.

To generate the required geospatial and cluster files using a custom boundary (e.g., an ecoregion that crosses state boundaries or a portion of a state), you will need a geospatial file containing the region boundary, a folder to save outputs, and a region name. The CustomBoundary package uses st_read() to read in the geospatial file, which can accept a variety of common geospatial file types like shapefile and GeoJSON.

1) Install the CustomBoundary package using the code below. Uncomment the first line if you do not have the “pak” package installed.

```
# install.packages("pak")
pak::pak("laura-naslund/CustomBoundary")
```

2) Copy the code below and replace the example inputs with a) the path to your boundary file, b) the path to your desired output folder, and c) the name of your region. You can copy file paths to your clipboard by right clicking on the file name in your file explorer and selecting: Copy as path (Figure 12). If you are using a Windows machine, you will need to replace the backslashes (\) in the pasted file path with forward slashes (/). Ensure file paths and the region name are wrapped in straight quotes ("").

```
CustomBoundary::generateFiles(inputFilePath =
"C:/Users/UserName/Profile/Desktop/example_region.shp", outputFolder =
"C:/Users/UserName/Profile/Desktop/Output", region = "exampleRegion")
```

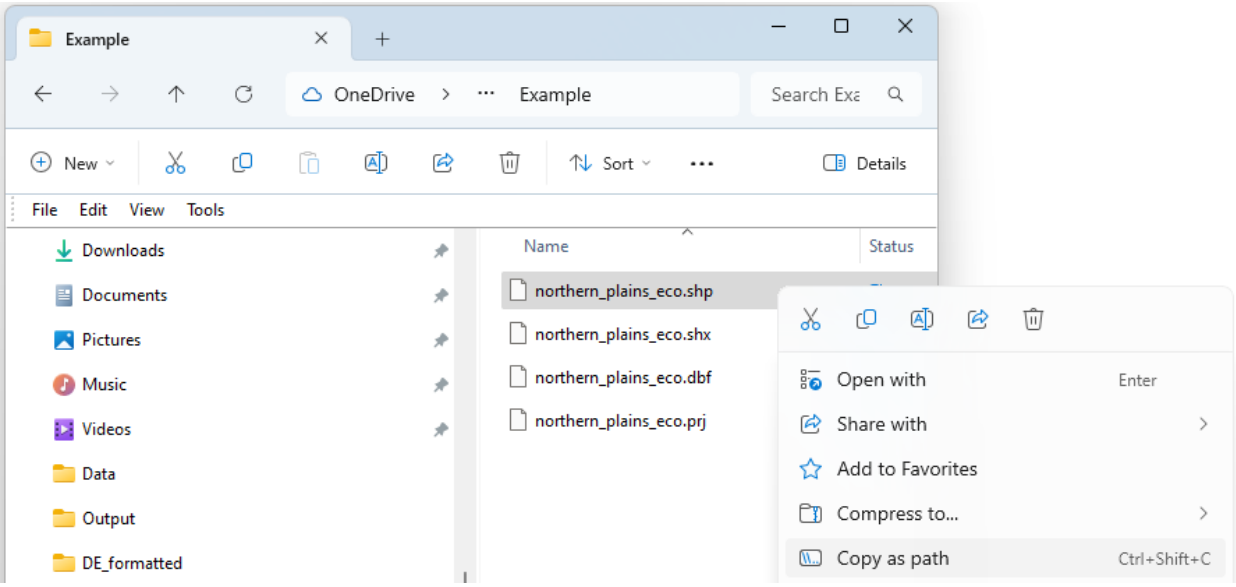


Figure 12. To copy a file path to your clipboard, right click the file name in your file explorer and select: Copy as path.

3) When generateFiles() has finished running, you will find the required input files in the specified output folder. Copy these files into the input data folder you intend to zip. Add the names of the resulting files to the corresponding parameters in the _CASTool_Metadata.xlsx (Figure 13). Note that the ClusterOutput folder contains additional files generated by the clustering algorithm that are not required as CASTool inputs. The CustomBoundary package writes the default number of clusters to the cluster graphic and cluster assignment file. If an alternative number of clusters is desired, those outputs can be found in the ClusterOutput folder.

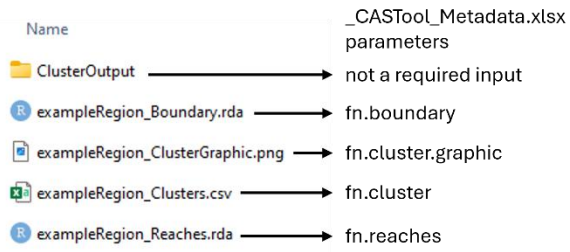


Figure 13. Output files generated by the CustomBoundary package and their corresponding metadata parameter names.

4. Shiny application

The CASTool Shiny application provides an interactive interface for users to access the CASTool in a web browser. Users do not need programming skills or specific computer programs to use the Shiny application. Unlike the R program implementation of the CASTool, the Shiny application can only run one target site at a time.

1) The landing page of the CASTool Shiny application provides a tool disclaimer and download link for this user guide and input file templates (Figure 14).

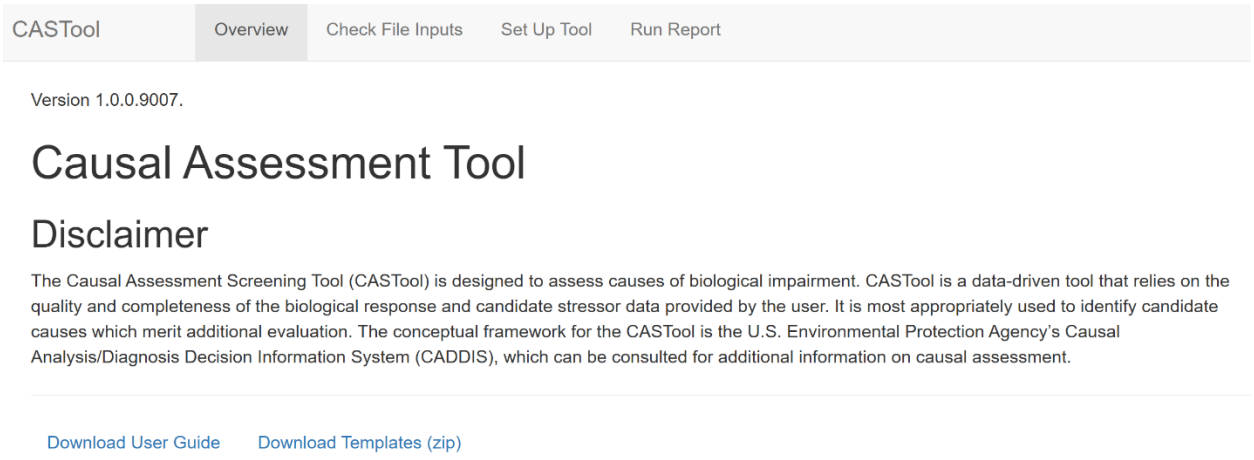


Figure 14. Landing page of the CASTool Shiny application, which includes the CASTool disclaimer and links to download this user guide and file input templates.

2) Begin your run of the CASTool on the next tab *Check File Inputs*. Upload a zipped folder (file size limit 300 MB) by selecting the Browse button and navigate to the folder in your file explorer window. If your zipped folder exceeds the file size limitation, reach out to the contact listed in the user guide for further direction (*Figure 15*).

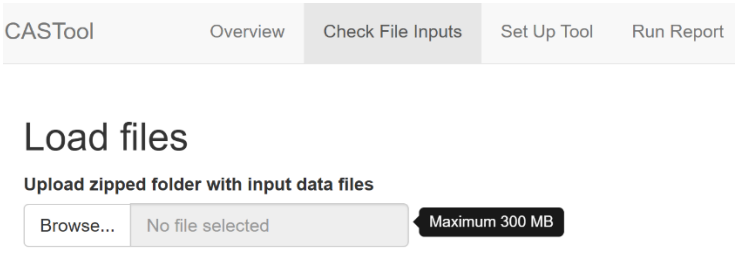


Figure 15. The upload function of the Check File Inputs tab provides the application a user's input data files in a zipped folder.

3) Ensure that all required and intended input data files are uploaded by reviewing the matching files table, missing files box, and extra files box. Rows of this table are shaded

based on their value for the “Present” column (blue = true, orange = false). Files are counted as present if they are both listed in the metadata and present in the uploaded data. Files are missing if they are listed in the metadata but absent in the uploaded data. The names of missing files will also populate in the Missing files box below. Input data files for which users have not provided a file name are not assigned a value for the Present column and the corresponding rows are not shaded (Figure 16).

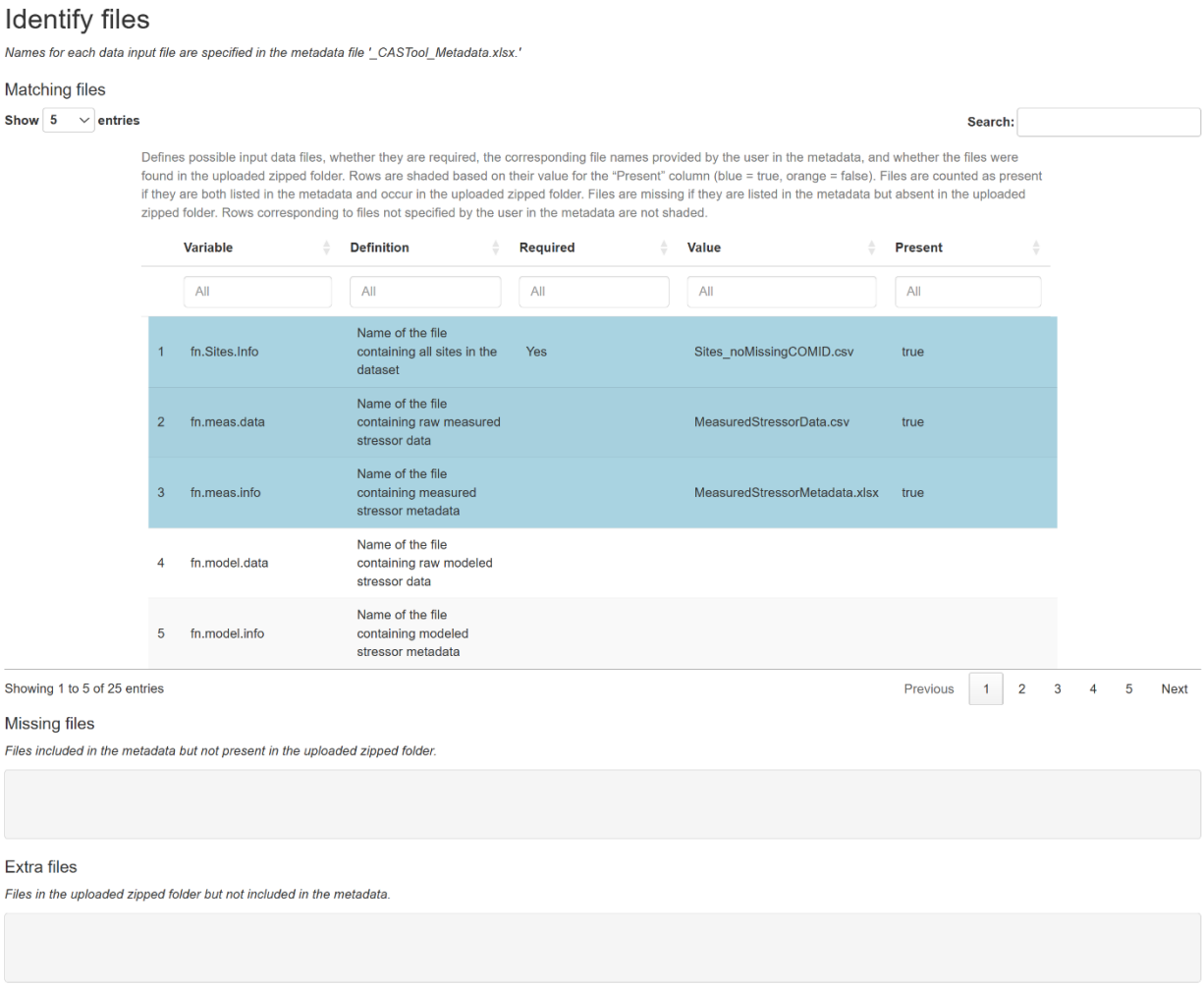


Figure 16. The Identify Files section of the Check File Inputs tab in the CASTool. The table displays the names of CASTool input files specified by the user in the metadata, whether the application found those files among those uploaded by the user, and whether an input file is required. The boxes below display the names of files included in the metadata but not present in the uploaded zipped folder and files present in the zipped folder but not included in the metadata.

4) Review the Contents of uploaded files to ensure that the tool is correctly parameterized (i.e., recognizes the available biological communities, stressor data types, stressor-tolerance values available, and outlier exclusion indicator) (Figure 17).

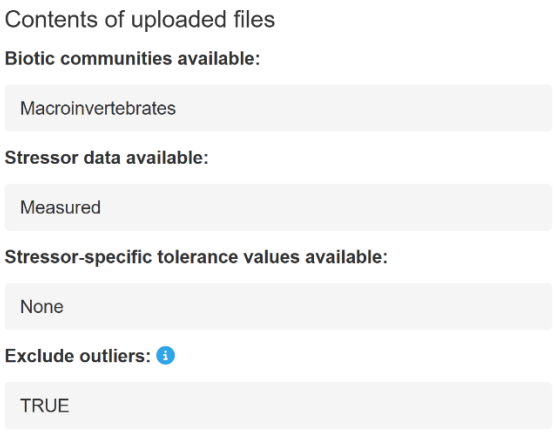


Figure 17. The Contents of uploaded files section of the Check File Inputs tab in the CASTool. The boxes display the biotic communities to be analyzed, type of stressor data available, and the biotic communities for which stressor-specific tolerance values are available, based on uploaded files. The final box displays whether to exclude outliers, based on user metadata inputs.

5) Select the Check input files button to ensure that the contents of your uploaded files meet CASTool requirements.

6) Review the Input file check table to ensure that no columns are missing, and all columns have the expected data types. Rows of the table are shaded based on their value for the "QC_Passed" column (blue = true, orange = false), which summarizes whether the input data file contains missing columns or unexpected data types (Figure 18).

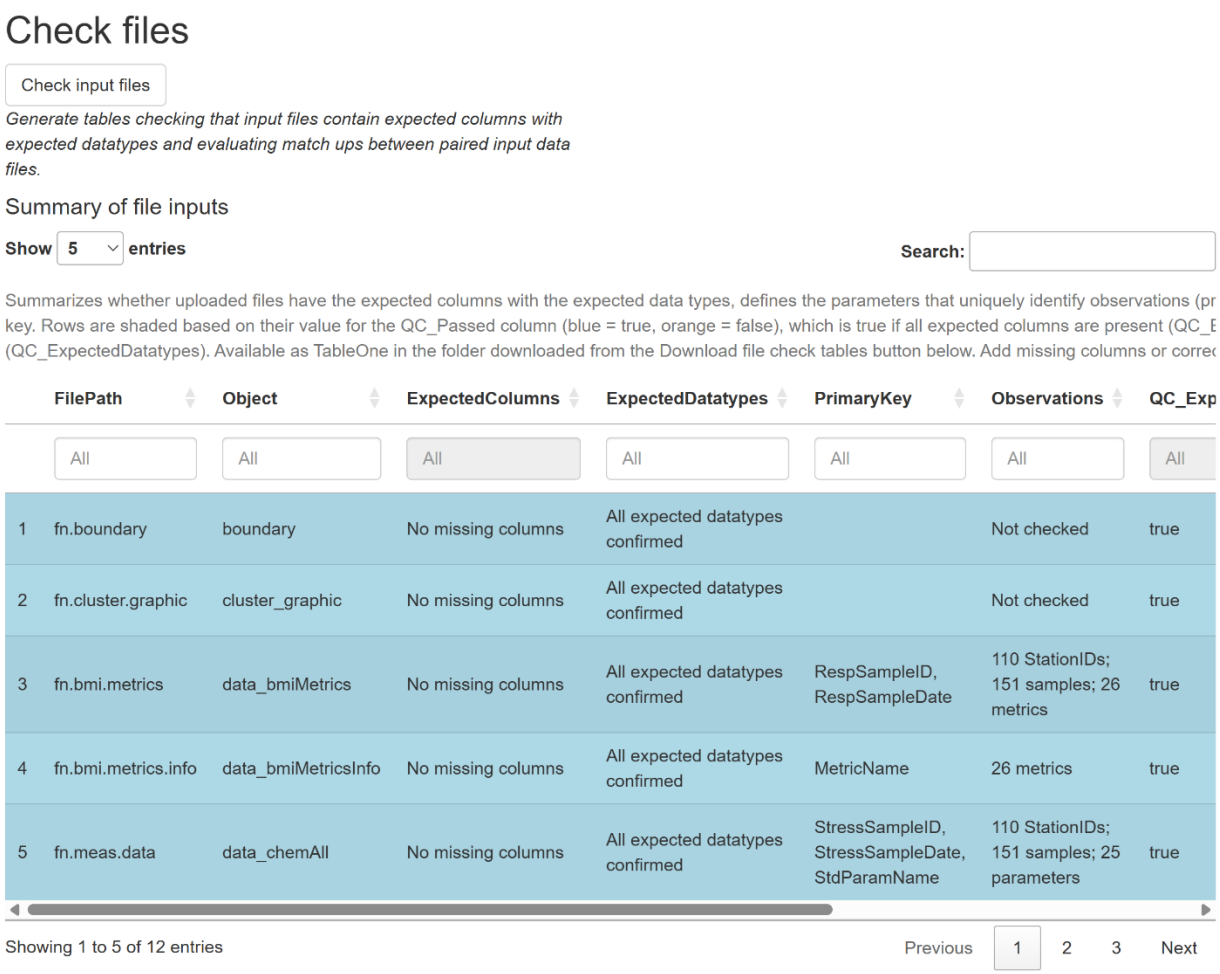
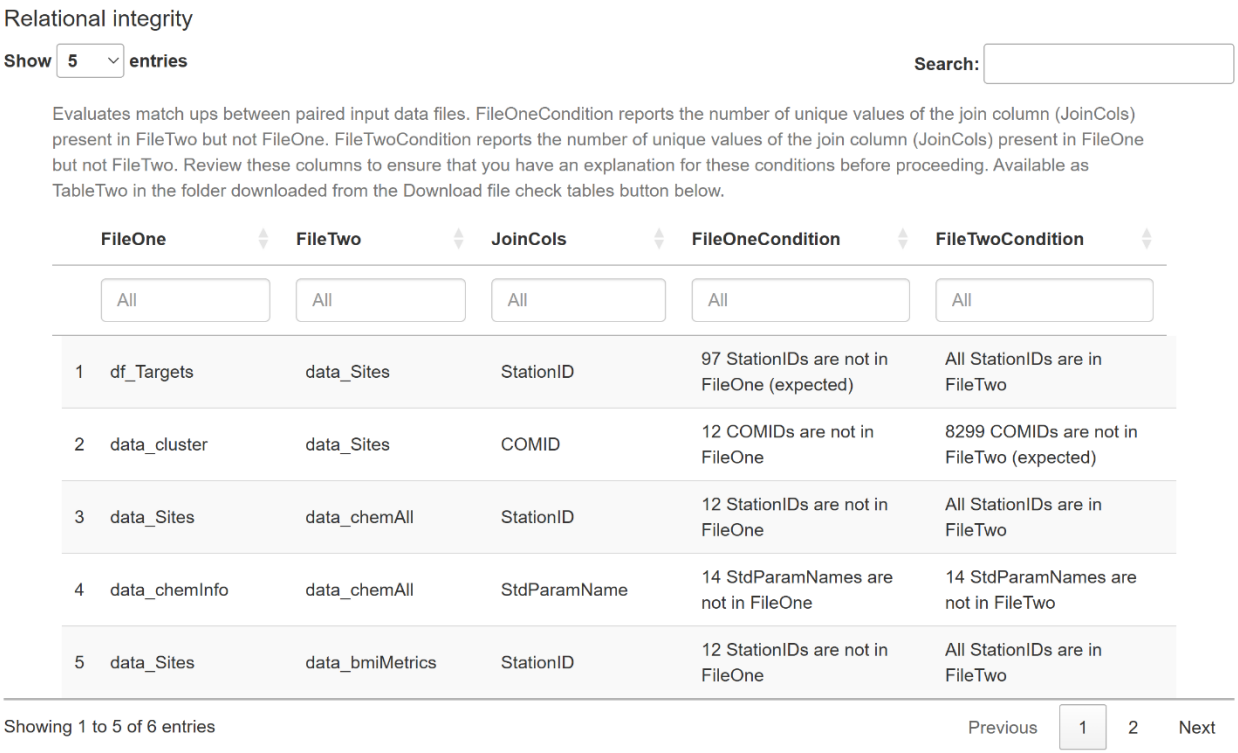


Figure 18. The Summary of file inputs table in the Check files section of the Check File Inputs Tab in the CASTool. This table describes whether input data files have the expected columns with the expected data types and summarizes unique observations of key columns.

7) Review the Relational integrity table to ensure that paired files contain the expected matchups by the joining parameter specified in JoinCols. The FileOneCondition column reports on the number of observations of the joining parameter in FileOne not shared by FileTwo. The FileTwoCondition column reports on the number of observations of the joining parameter in FileTwo not shared by FileOne (Figure 19).



Relational integrity

Show 5 entries

Search:

Evaluates match ups between paired input data files. FileOneCondition reports the number of unique values of the join column (JoinCols) present in FileTwo but not FileOne. FileTwoCondition reports the number of unique values of the join column (JoinCols) present in FileOne but not FileTwo. Review these columns to ensure that you have an explanation for these conditions before proceeding. Available as TableTwo in the folder downloaded from the Download file check tables button below.

	FileOne	FileTwo	JoinCols	FileOneCondition	FileTwoCondition
	All	All	All	All	All
1	df_Targets	data_Sites	StationID	97 StationIDs are not in FileOne (expected)	All StationIDs are in FileTwo
2	data_cluster	data_Sites	COMID	12 COMIDs are not in FileOne	8299 COMIDs are not in FileTwo (expected)
3	data_Sites	data_chemAll	StationID	12 StationIDs are not in FileOne	All StationIDs are in FileTwo
4	data_chemInfo	data_chemAll	StdParamName	14 StdParamNames are not in FileOne	14 StdParamNames are not in FileTwo
5	data_Sites	data_bmiMetrics	StationID	12 StationIDs are not in FileOne	All StationIDs are in FileTwo

Showing 1 to 5 of 6 entries

Previous 1 2 Next

Figure 19. The Relational integrity table in the Check files section of the Check File Inputs Tab in the CASTool. This table describes matchups between paired input files (i.e., unique observations of key columns that are present in one file but not the other).

The Summary of file inputs check table and Relational integrity table can be downloaded using the Download file check tables button.

8) When you have corrected any file errors identified in the Input file check and Relational integrity tables, download the zipped folder of checked files using the Download checked data button. Unless you add data to your input data files, you can skip steps 1-8 the next time you run the CASTool and proceed directly to the Set Up Tool tab using your previously downloaded checked files zipped folder.

9) Select the Set Up Tool tab. Upload the checked files zipped folder generated by the Check File Inputs tab by selecting the Browser button and navigating to the folder's location in your file explorer (Figure 20). If you saved the files to OneDrive, ensure that they have completed syncing before uploading them to the CASTool. Explore watershed stressor data will populate with the value you provided the exploreWSStressor variable in the CASTool metadata (Figure 20). The Clustering figure will populate with the cluster graphic you provided, or the figure pulled from the helper package (Figure 21).

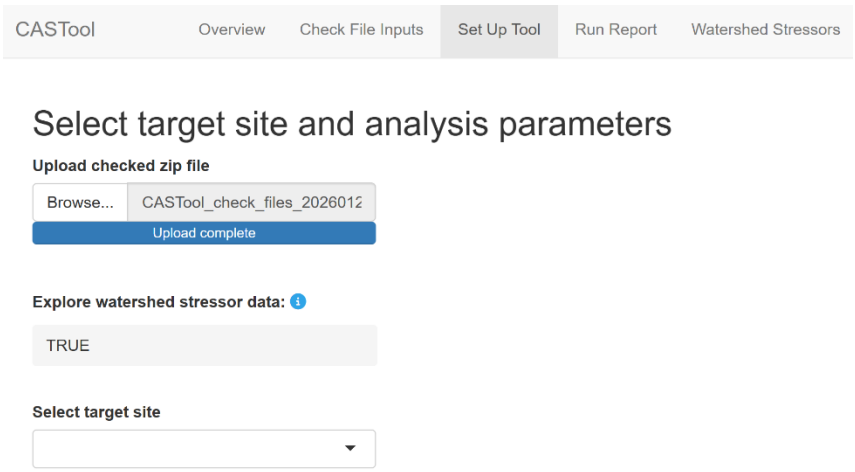


Figure 20. The Select target site and analysis parameters section of the Set Up Tool tab in the CASTool, which includes a browser window link to select the zipped folder of file inputs generated by the Check File Inputs tab, describes whether the user has selected in the metadata to explore watershed stressor data, and provides a dropdown menu to select a target site for analysis.

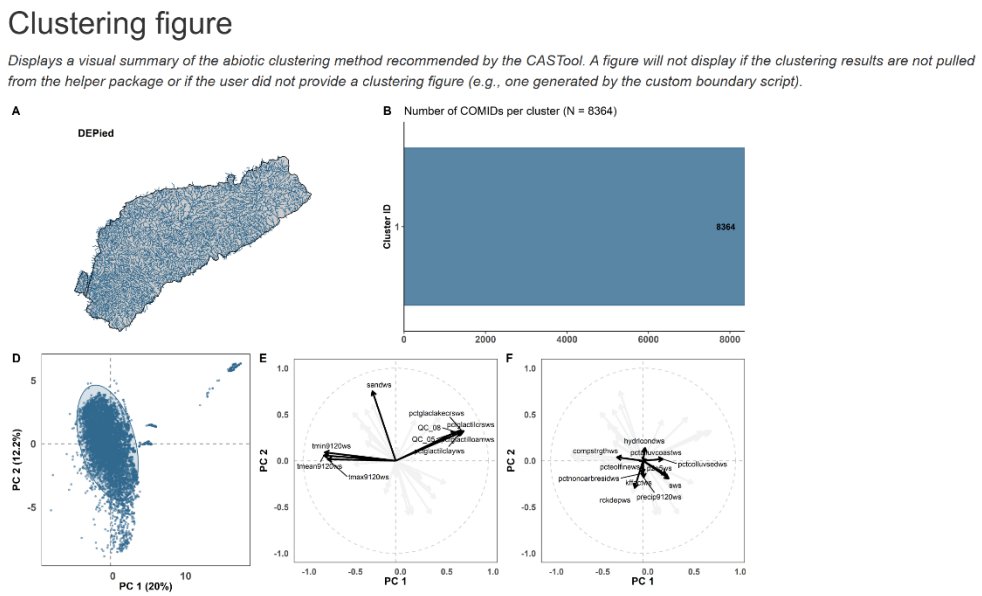


Figure 21. The clustering figure displayed on the Set Up Tool tab of the CASTool, which includes (A) a map of reach cluster assignments, (B) a bar chart of the number of reaches assigned to each cluster, (D) the distribution of reaches in the PCA, (E) the environmental variables with the greatest PCA loadings, (F) the environmental variables with the lowest PCA loadings.

10) Select your desired target site from the dropdown menu (Figure 20).

11) Select the Run Report tab and click the Generate report button. A progress bar will display in the bottom right-hand corner of the application indicating the current and remaining steps of the analysis (*Figure 22*). Note the application will time out after 15 minutes of no user clicks. If the application times out, your results will not be saved. Running the report should take less than 5 minutes per biological community analyzed, so to avoid time outs, run the report when you have time to review and download the results. A pop-up will be displayed when the report is completed (*Figure 23*).

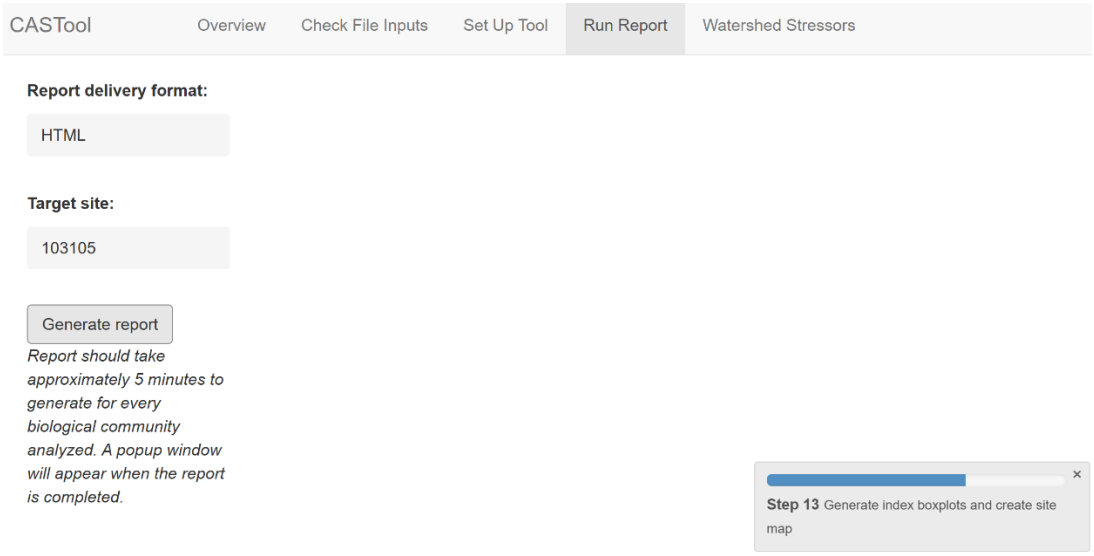


Figure 22. The Run Report tab of the CASTool, which describes the report delivery format and selected target site. After the user has selected the Generate report button, a progress bar displays in the bottom right corner.

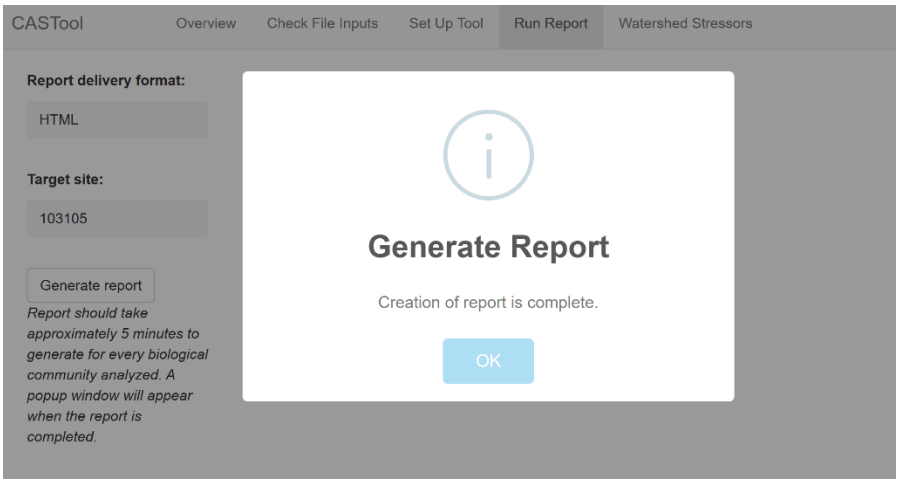


Figure 23. The pop-up displayed when the report has been generated and is available for download along with supplemental files.

12) Select the Download report and supplemental files button to download a zipped folder containing the report and accompanying files (*Figure 24*).

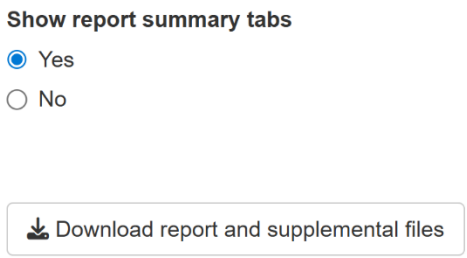


Figure 24. The radio buttons to control the display of report summary tabs and the Download report and supplemental files button on the Run Report tab of the CASTool.

13) To display tabs in the Shiny interface containing the report results select the "Yes" radio button under the "Show report summary tabs" header (*Figure 24*).

14) If Explore watershed stressor data was set to true on the Set Up Tool tab, select the Watershed Stressors tab to explore the watershed summary figures generated by the CASTool. To display a figure, select the desired watershed variable from the dropdown menu. To zoom in on a figure, use the built-in zoom functionality in your browser (*Figure 25*).

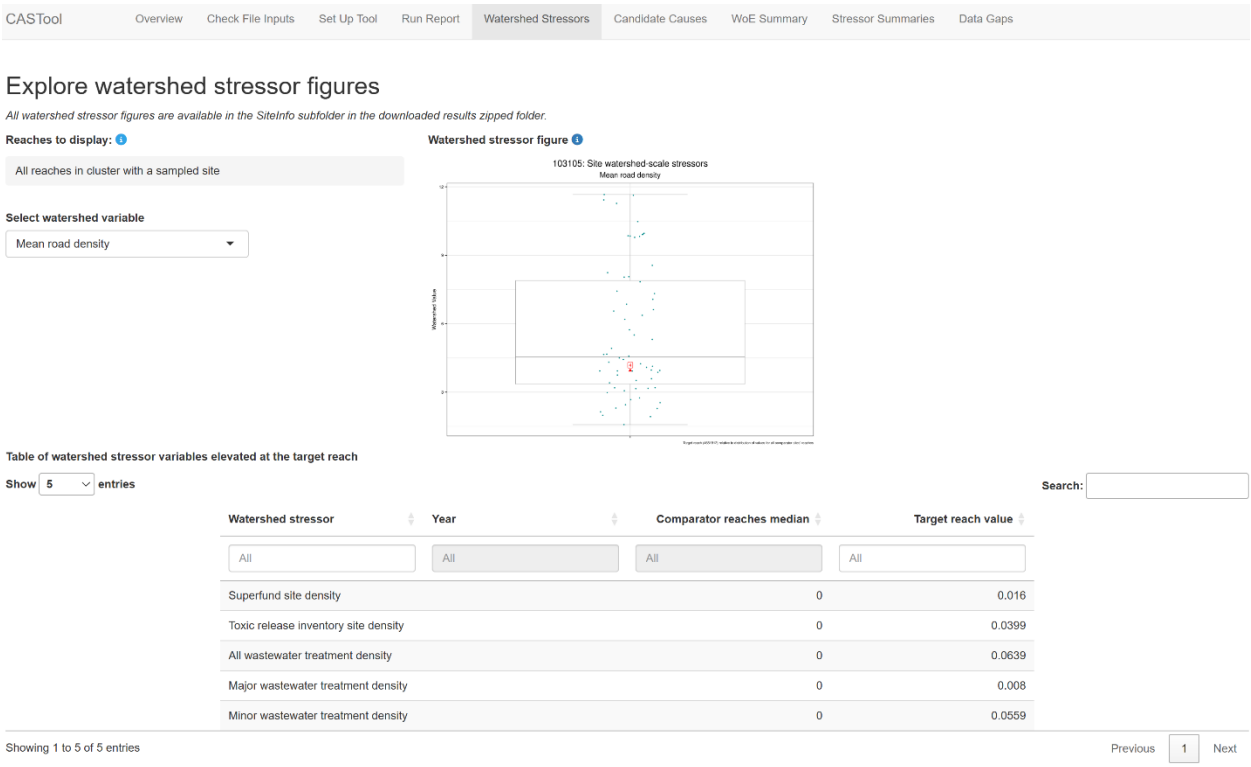


Figure 25. The Watershed Stressors tab of the CASTool, which specifies which reaches the user selected to display (all reaches in a cluster with a sampled site vs. all reaches in a cluster), includes a dropdown to display a particular watershed figure, and summarizes in a table the watershed stressor variables elevated at the target reach.

15) Select the Candidate Causes tab to view the threshold values of pH and dissolved oxygen (specified in the metadata) used to define these stressors as candidate causes regardless of the co-occurrence analysis results. This tab also lists the stressor(s) initially evaluated as candidate causes and those not advanced by the co-occurrence line of evidence (Figure 26).

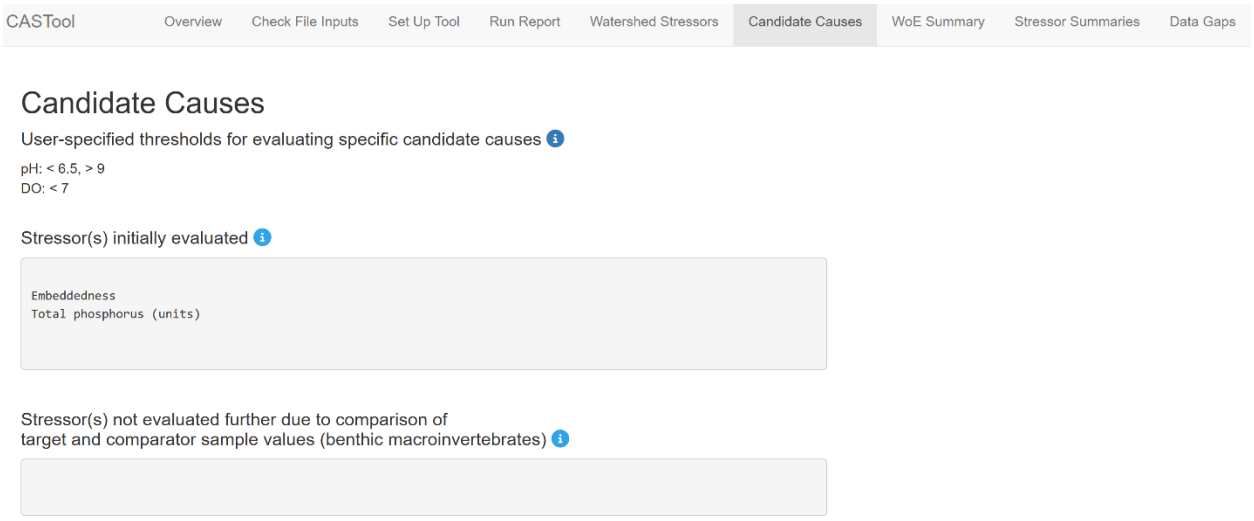


Figure 26. The Candidate Causes tab of the CASTool, which reports thresholds selected by the user in the metadata for evaluating specific candidate causes, the stressors initially evaluated, and stressors not evaluated further due to comparison of target and comparator sample values.

16) Select the WoE Summary tab to view the summary boxplots of biological community index values (Figure 27), the weight of evidence summary table (Figure 28), the lines of evidence summary table, and the summary figure.

Figure 27. The Biological index distributions boxplots available on the WoE Summary tab of the CASTool. The boxplots display the distribution of biological index values for comparator (inside the case) and non-comparator (outside the case) samples. The shape and color of points indicate whether they are from reference sites, are degraded or not degraded, and are from the target site.

Weight of evidence table

Show 10 entries

Search:

Stressor	Response Sample ID	Response Sample Date	Inside-the-Case							Outside the Case
			Co-Occurrence	Sufficiency	Biological Gradient	Time Sequence	Verified Prediction			Biological Gradient
							SSTolVals	SSI Co-Occurrence	SSI Sufficiency	
Embeddedness	103105_2011_10_24	2011-10-24	1	1	1	NE				0
Embeddedness	103105_2022_11_09	2022-11-09	-1	0	1	NE				0
Total phosphorus (units)	103105_2011_10_24	2011-10-24	1	1	0	NE				0
Total phosphorus (units)	103105_2022_11_09	2022-11-09	-1	0	0	NE				0

Showing 1 to 4 of 4 entries

Previous 1 Next

Figure 28. The Weight of evidence table available on the WoE Summary tab of the CASTool. This table summarizes the scores for each line of evidence, for each stressor, for target site sample.

Lines of evidence summary

Show 10 entries Search:

Stressor	Response Sample ID	Response Sample Date	Supporting (>0)	Refuting (<0)	Indeterminate (=0)	Not evaluated
Embeddedness	103105_2011_10_24	2011-10-24	3	0	1	1
Embeddedness	103105_2022_11_09	2022-11-09	1	1	2	1
Total phosphorus (units)	103105_2011_10_24	2011-10-24	2	0	2	1
Total phosphorus (units)	103105_2022_11_09	2022-11-09	0	1	3	1

Showing 1 to 4 of 4 entries Previous 1 Next

Figure 29. The Lines of evidence summary table available on the WoE Summary tab of the CASTool. This table summarizes the number of lines of evidence supporting, refuting, indeterminate, and not evaluated for each stressor and target site sample.

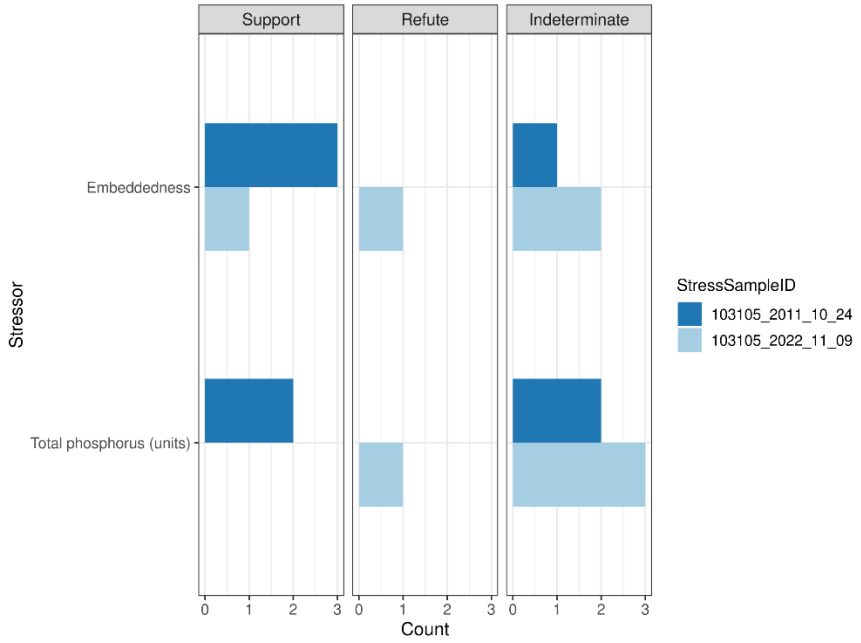


Figure 30. The Lines of evidence summary figure available on the WoE Summary tab of the CASTool. This figure summarizes the number of lines of evidence supporting, refuting, indeterminate, and not evaluated for each stressor and target site sample.

17) Select the Stressor Summaries tab to view the figures associated with the lines of evidence for each candidate cause. The contents of this tab may require up to a minute to load.

Figure 31. The Stressor Summaries tab of the CASTool, which displays the lines of evidence figures for each stressor evaluated.

18) Select the Data Gaps tab to view the data gaps table. The data gaps table is also available in the zip file downloaded on the Run Report tab.

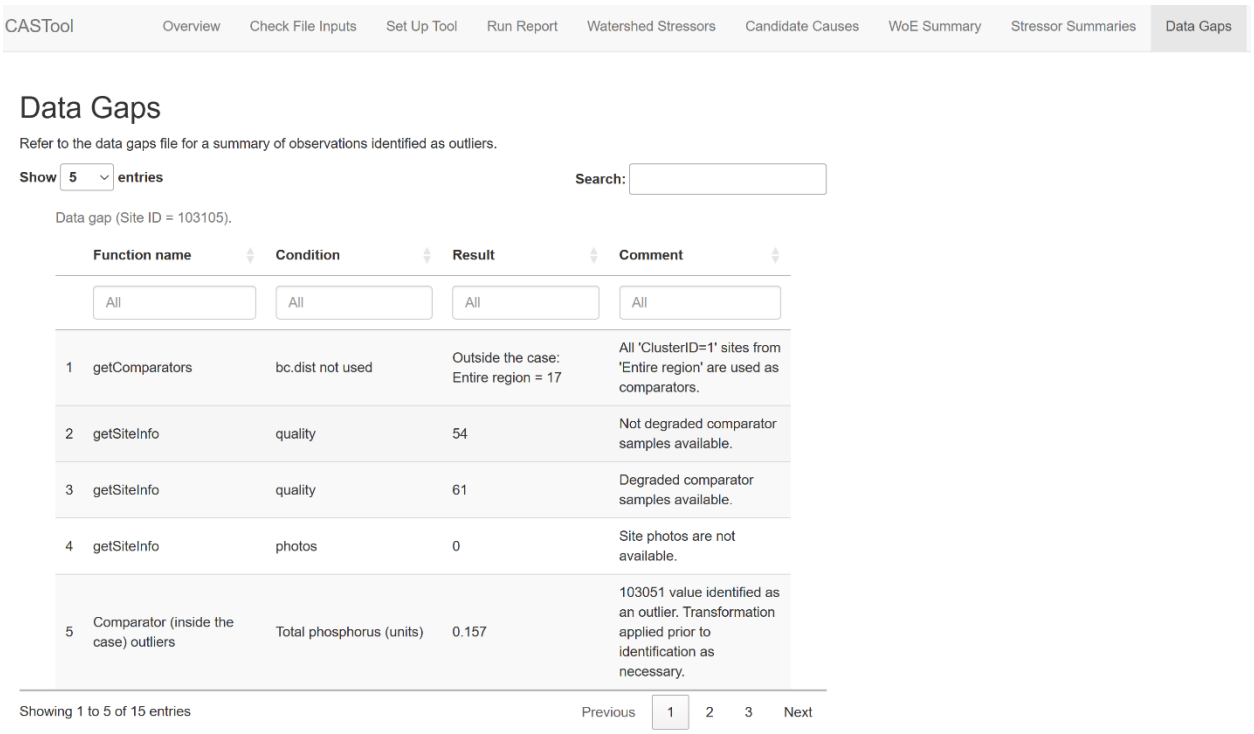


Figure 32. The Data Gaps tab of the CASTool, which displays a table of data gaps identified in the CASTool run.

5. R program

The R program implementation of the CASTool requires the user to have R installed and configured on their computer. The R program will analyze and generate a report for every site included in the target site data file.

1) To run the R program, download or clone the CASTfxn repository (WEBLINK).

2) Open the CASTool.r script (inst/shiny-examples/CASTool/CASTool.R).

3) Change the value of boo_Shiny on line 28 to FALSE, to indicate that you are running the CASTool code outside of the R Shiny application (Figure 33).

```
27  # Define global variables
28  boo_Shiny <- FALSE # Whether to run the code in Shiny mode (set to FALSE if running script outside of the app)
```

Figure 33. The definition of the boolean variable boo_Shiny in the global variable definition section of the CASTool R program. To run the R program, boo_Shiny must be set to FALSE.

4) Change the in.dir and out.dir variables to the paths for the folders containing your input data files and where you would like the CASTool to write results files, respectively (Figure 34). If you are using a Windows machine, you may need to replace the backslashes (\) in

the file path with forward slashes (/). Ensure file paths and the region name are wrapped in straight quotes (""). Change the region variable to the name of your region. Ensure that it matches the value you provided to the region variable in the CASTool metadata.

It is recommended to write files as close to the root of a local storage directory as possible to avoid issues arising from excessive file path lengths. If a target site or stressor names are particularly long, these may also need to be abbreviated in input files to avoid path length issues.

Example of the longest path: [local directory path] / [region] / Results / [TargetSiteID] / [Biocommunity abbreviation] / [Stressor name] / [TargetSiteID]_[Stressor Name]_[Biocommunity abbreviation]_[Biotic metric name]_CO.png

Check the maximum path length of your computer and compare the number of characters of the above file path parameterized with the longest TargetSiteID, biocommunity abbreviation, dtressor name, and biotic metric name.

```
in.dir <- "C:/Users/lnaslund/Documents/CASTool_Data/DataNoHelper/Data" # File path of data directory
out.dir <- "C:/Users/lnaslund/Documents/CASTool_Data/DataNoHelper/Results" # File path of results directory
region <- "DEPied" # Name of region
```

Figure 34 The definition of input and output directories and region in the global variables definition section of the CASTool R program. These values must be changed by the user prior to running the CASTool R program.

5) Highlight the entire script (e.g., by using the ctrl+A shortcut) and run the code. The first time you run the code, it will download three helper data packages. These packages may take up to 20 minutes to download but will not require downloads in subsequent runs of the script. Progress through the CASTool analyses is reported in the console. The code runs an initial check of input data files to ensure that they meet CASTool requirements. If the tool encounters an issue with the input data files, it will halt the execution of the CASTool script. Correct these issues by reviewing TableOne and TableTwo in the global environment, which contain the input file check table and relational integrity table described above (Section 4. Shiny application).

6. Contact information

For help with the CASTool or to provide feedback, please contact Laura Naslund (Naslund.Laura@epa.gov).

7. References

Gillett, D. J., R. D. Mazor, and S. B. Norton. 2019. Selecting Comparator Sites for Ecological Causal Assessment Based on Expected Biological Similarity. Freshw Sci **38**:554-565.

Jones, J. I., C. E. M. Lloyd, J. F. Murphy, A. Arnold, C. P. Duerdoth, A. Hawczak, J. L. Pretty, P. J. Johnes, J. E. Freer, M. W. Stirling, C. Richmond, and A. L. Collins. 2023. What do macroinvertebrate indices measure? Stressor-specific stream macroinvertebrate indices can be confounded by other stressors. Freshw Biol **68**:1330-1345.

U.S. Environmental Protection Agency. 2000. Stressor Identification Guidance Document. Washington, DC.

U.S. Environmental Protection Agency. 2010. Causal Analysis/Diagnosis Decision Information System (CADDIS): epa.gov/caddis

Appendix. Additional methodological details

A.1 Outlier detection algorithm

The outlier detection routine in the CASTool identifies stressor observations as outliers if they either 1) lie outside of the boxplot inner fence or 2) are more than six times the standard deviation away from the mean.

The inner fence is defined as:

25th percentile – 1.5 * interquartile range

75th percentile + 1.5 * interquartile range

A.2 Clustering algorithm

The goal of the clustering algorithm was to identify sites that we would expect to have similar biological communities to the target site in the absence of stressors; we refer to these sites as comparator sites. To identify comparator sites, we clustered stream reaches based on the similarity of their flow and slope, which we extracted from NHDPlusV2, as well as watershed summary values of geological, pedological, and topographical variables, which we extracted from EPA's StreamCat dataset. With this approach, we implicitly assume that these variables are significant drivers of stream biological communities (Poff et al. 1997, Hill et al. 2015). We generated clusters using agglomerative hierarchical clustering on principal components. We chose this method because clustering on principal components can produce more stable clusters than clustering on many input variables, and hierarchical clustering can identify clusters with different shapes and sizes (Husson et al. 2010, Saxena et al. 2017). We generated clusters by state because stream biological monitoring and assessment programs are often conducted at this scale; however, this approach can be extended beyond state boundaries if the sampled data are comparable.

We identified stream reaches within a state using level-one outlines from the Database of Global Administrative Areas (GADM 2022) with a 300 m buffer around the state border to account for simplified geometry. We extracted NHDPlusV2 flowlines and attributes within this boundary using the nhdplustools R package (Blodgett and Johnson 2023) and downloaded StreamCat watershed variables for the corresponding stream reaches using the StreamCatTools R package (Weber 2025) (Table 1). Except for percentage variables, we transformed highly skewed variables (squared skewness > 3) using a box-cox transformation. We then centered and scaled all variables and ran a PCA on complete observations using the PCA function in the FactoMineR R package (Le et al. 2008). We

determined the number of principal components required to reach a user-specified threshold of cumulative percent variation (default = 60%) and used these components to impute missing data using the regularized iterative PCA algorithm in the imputePCA function in the missMDA R package (Josse and Husson 2016). We then ran a PCA on the dataset with imputed values and passed the previously selected principal components to an agglomerative hierarchical clustering function (hclust.vector) in the fastcluster R package using Ward's method (Müllner 2013). We cut the hierarchical tree to generate the maximum number of clusters that each contained at least a user-specified minimum percentage of stream reaches (default = 20%).

Table #. Variables used to cluster stream reaches.

Variable	Definition	Source
qc_04	April mean monthly stream flow estimate	NHDPlusV2
qc_05	May mean monthly streamflow estimate	NHDPlusV2
qc_08	August mean monthly streamflow estimate	NHDPlusV2
qc_ma	Mean annual streamflow estimate	NHDPlusV2
slope	Stream slope	NHDPlusV2
al2o3	Mean % of lithological aluminum oxide (Al2O3) content in surface or near surface geology within watershed	StreamCat
bfi	Base flow is the component of streamflow that can be attributed to ground-water discharge into streams. The BFI is the ratio of base flow to total flow, expressed as a percentage, within watershed	StreamCat
cao	Mean % of lithological calcium oxide (CaO) content in surface or near surface geology within watershed	StreamCat
clay	Mean % clay content of soils (STATSGO) within watershed	StreamCat
compstrgth	Mean lithological uniaxial compressive strength (megaPascals) content in surface or near surface geology within watershed	StreamCat
elev	Mean watershed elevation (m)	StreamCat
fe2o3	Mean % of lithological ferric oxide (Fe2O3) content in surface or near surface geology within watershed	StreamCat

hydrlcond	Mean lithological hydraulic conductivity (micrometers per second) content in surface or near surface geology within watershed	StreamCat
k2o	Mean % of lithological potassium oxide (K2O) content in surface or near surface geology within watershed	StreamCat
kffact	Mean soil erodibility (Kf) factor (unitless) of soils within watershed. The Kffactor is used in the Universal Soil Loss Equation (USLE) and represents a relative index of susceptibility of bare, cultivated soil to particle detachment and transport by rainfall.	StreamCat
mgo	Mean % of lithological magnesium oxide (MgO) content in surface or near surface geology within watershed	StreamCat
na2o	Mean % of lithological sodium oxide (Na2O) content in surface or near surface geology within watershed	StreamCat
n	Mean % of lithological nitrogen (N) content in surface or near surface geology within watershed	StreamCat
om	Mean organic matter content (% by weight) of soils (STATSGO) within watershed	StreamCat
p2o5	Mean % of lithological phosphorous oxide (P2O5) content in surface or near surface geology within watershed	StreamCat
pctalluvcoast	% of watershed area classified as as lithology type: alluvium and fine-textured coastal zone sediment	StreamCat
pctnoncarbresid	% of watershed area classified as as lithology type: non-carbonate residual material	StreamCat
pctsilicic	% of watershed area classified as as lithology type: silicic residual material	StreamCat
perm	Mean permeability (cm/hour) of soils (STATSGO) within watershed	StreamCat
precip9120	PRISM climate data - 30-year normal mean precipitation (mm): Annual period: 1991-2020 within the watershed	StreamCat
rckdep	Mean depth (cm) to bedrock of soils (STATSGO) within watershed	StreamCat
runoff	Mean runoff (mm) within watershed	StreamCat

sand	Mean % sand content of soils (STATSGO) within watershed	StreamCat
sio2	Mean % of lithological silicon dioxide (SiO2) content in surface or near surface geology within watershed	StreamCat
s	Mean % of lithological sulfur (S) content in surface or near surface geology within watershed	StreamCat
tmax9120	PRISM climate data - 30-year normal maximum temperature (C°): Annual period: 1991-2020 within the watershed	StreamCat
tmean9120	PRISM climate data - 30-year normal mean temperature (C°): Annual period: 1991-2020 within the watershed	StreamCat
tmin9120	PRISM climate data - 30-year normal minimum temperature (C°): Annual period: 1991-2020 within the watershed	StreamCat
WetIndex	Mean Composite Topographic Index (CTI)[Wetness Index] within watershed	StreamCat
wtdep	Mean seasonal water table depth (cm) of soils (STATSGO) within watershed	StreamCat
pctalkintruvol	% of AOI area classified as lithology type: alkaline intrusive volcanic rock	StreamCat
pctcarbresid	% of AOI area classified as lithology type: carbonate residual material	StreamCat
pctcoastcrs	% of AOI area classified as lithology type: coastal zone sediment, coarse-textured	StreamCat
pctcolluvsed	% of AOI area classified as lithology type: colluvial sediment	StreamCat
pcteolcrs	% of AOI area classified as lithology type: eolian sediment, coarse-textured (sand dunes)	StreamCat
pcteolfine	% of AOI area classified as lithology type: eolian sediment, fine-textured (glacial loess)	StreamCat
pctextruvol	% of AOI area classified as lithology type: extrusive volcanic rock	StreamCat
pctglaclakecrs	% of AOI area classified as lithology type: glacial outwash and glacial lake sediment, coarse-textured	StreamCat

pctglaclakefine	% of AOI area classified as lithology type: glacial lake sediment, fine-textured	StreamCat
pctglactilclay	% of AOI area classified as lithology type: glacial till, clayey	StreamCat
pctglactilcrs	% of AOI area classified as lithology type: glacial till, coarse-textured	StreamCat
pctglactilloam	% of AOI area classified as lithology type: glacial till, loamy	StreamCat
pcthydric	% of AOI area classified as lithology type: hydric, peat and muck	StreamCat
pctsallake	% of AOI area classified as lithology type: saline like sediment	StreamCat

References

Blodgett, D., and M. Johnson. 2023. nhdplusTools: Tools for Accessing and Working with the NHDPlus. U.S. Geological Survey.

GADM. 2022. Database of Global Administrative Areas.

Gillett, D. J., R. D. Mazor, and S. B. Norton. 2019. Selecting Comparator Sites for Ecological Causal Assessment Based on Expected Biological Similarity. Freshw Sci **38**:554-565.

Hill, R. A., M. H. Weber, S. G. Leibowitz, A. R. Olsen, and D. J. Thornbrugh. 2015. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. JAWRA Journal of the American Water Resources Association **52**:120-128.

Husson, F., J. Josse, and J. Pagès. 2010. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? , Applied Mathematics Department.

Jones, J. I., C. E. M. Lloyd, J. F. Murphy, A. Arnold, C. P. Duerdoth, A. Hawczak, J. L. Pretty, P. J. Johnes, J. E. Freer, M. W. Stirling, C. Richmond, and A. L. Collins. 2023. What do macroinvertebrate indices measure? Stressor-specific stream macroinvertebrate indices can be confounded by other stressors. Freshw Biol **68**:1330-1345.

Josse, J., and F. Husson. 2016. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. Journal of Statistical Software **70**:1-30.

Le, S., J. Josse, and F. Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. **25**:1-18.

Müllner, D. 2013. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. Journal of Statistical Software **53**.

Poff, N. L., J. D. Allan, M. B. Bain, J. R. Karr, K. L. Prestegaard, B. D. Richter, R. E. Sparks, and J. C. Stromberg. 1997. The Natural Flow Regime. BioScience **47**:769-784.

Saxena, A., M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin. 2017. A review of clustering techniques and developments. Neurocomputing **267**:664-681.

U.S. Environmental Protection Agency. 2000. Stressor Identification Guidance Document. Washington, DC.

U.S. Environmental Protection Agency. 2010. Causal Analysis/Diagnosis Decision Information System (CADDIS).

Weber, M. 2025. StreamCatTools: Tools to work with the StreamCat API within R and access the full suite of StreamCat and LakeCat metrics.