

Instructions for Calculating Bioassessment Indices and other Tools for Evaluating Wadeable Streams in California: The California Stream Condition Index (CSCI), Algal Stream Condition Index (ASCI) and Index of Physical Integrity (IPI)

Step-by-step Instructions for Using SWAMP Bioassessment Tools

Tyler Boyle¹, Raphael Mazor², Andrew C. Rehn³, Susanna Theroux², Marcus Beck², Marco Sigala⁴, Calvin Yang⁵, Shuka Rastegarpour⁵ and Peter R. Ode³

SWAMP-SOP-2020-0001

Revision Date: December 15, 2020

Citation:

Boyle, T., R. D. Mazor, A. C. Rehn, S. Theroux, M. Beck, M. Sigala, C. Yang, P.R. Ode. 2020. Instructions for calculating bioassessment indices and other tools for evaluating wadeable streams in California: The California Stream Condition Index (CSCI), Algal Stream Condition Index (ASCI) and Index of Physical Integrity (IPI) SWAMP-SOP-2020-0001

¹Geographical Information Center, California State University. Chico, CA

² Southern California Coastal Water Research Project, Costa Mesa, CA

³ California Department of Fish and Wildlife, Rancho Cordova, CA

⁴ Moss Landing Marine Laboratories, Moss Landing, CA

⁵ State Water Resources Control Board, Sacramento, CA

Table of Contents

Instructions for Calculating Bioassessment Indices and other Tools for Evaluating Wadeable Streams in California: The California Stream Condition Index (CSCI), Algal Stream Condition Index (ASCI) and Index of Physical Integrity (IPI).....	1
Introduction.....	3
Software and Toolbox Requirements.....	4
Watershed Metric Toolbox Download	5
Section 1: Creating Base Shapefiles for Metric Calculations	5
Creating the Sites Base File	6
Creating the Delineated Catchment Base File	11
Section 2: Calculating Bioassessment Index Predictor Variables in GIS.....	20
Calculating Index Predictor Variables in GIS.....	20
Section 3: Calculating Index Scores in R	26
The California Stream Condition Index (CSCI).....	26
The Algal Stream Condition Index (ASCI).....	41
The Index of Physical-habitat Integrity (IPI)	55
Natural (background) Specific Conductivity	65
Section 4: Calculating Reference Screening Criteria and other Site Characteristics	68
Reference Screening Processor	72
Cited literature.....	80
Appendix 1: Overview of Automated steps for Calculating CSCI	82

Introduction

This document describes steps in calculating three different ecological indices used to quantify stream conditions in California based on biological and physical data. The instructions herein are provided as support for analysts requiring scores for the California Stream Condition Index (CSCI), Algal Stream Condition Index (ASCI), and Index of Physical Integrity (IPI). Each index uses predictive modeling to compare the conditions observed at bioassessment sampling locations with those expected under natural (reference) conditions⁶. To learn more about California's Bioassessment Program, visit the [program page](#).

In addition to supporting index calculation, these instructions will help analysts calculate landscape-scale measures of human activity, which are used to determine if streams meet criteria to qualify as reference sites (Ode et al. 2016). Reference sites are stream reaches that define a benchmark of expected biological, chemical and physical attributes when human disturbance in the environment is absent or minimal, and they are the foundation of any bioassessment program.

At the beginning of each section in this document you will find background information, data and/or tools required, followed by the step-by-step instructions for that section:

[Section 1](#) describes the process for using geographic information system (GIS) to create the necessary base files for spatial analyses that underlie index calculations and reference site screening, including:

- 1) a "sites" file containing the latitude and longitude of stream sampling locations, and
- 2) a delineated catchments file containing polygons that define the upstream area that drains to a given sampling location.

[Section 2](#) describes the process for calculating environmental predictors necessary for each of the ecological indices and for the predicted conductivity model.

[Section 3](#) describes the process and data requirements for scoring sites for each ecological index (CSCI, ASCI, and IPI), calculating predicted conductivity, and troubleshooting errors when certain predictor variables fail to calculate.

⁶ Many of the predictor variables used to calculate the indices were borrowed from models developed to predict natural background specific conductivity in streams in the western United States (Olson and Hawkins 2012), which itself is a predictor for the ASCI index.

[Section 4](#) describes the process for using the Reference Screening Criteria to identify the least-disturbed reference sites in California.

The development and interpretation of the ecological indices, predicted conductivity models and the reference screening process have been documented in the following publications and technical reports:

CSCI

Mazor, R. D., P. R. Ode, A. C. Rehn, M. Engeln, K. A. Schiff, E. Stein, D. Gillett, D. Herbst, and C.P. Hawkins. 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. *Freshwater Science* 35(1): 249-271.

Rehn, A.C., R.D. Mazor and P.R. Ode. 2015. The California Stream Condition Index (CSCI): A New Statewide Biological Scoring Tool for Assessing the Health of Freshwater Streams. Swamp Technical Memorandum SWAMP-TM-2015-0002.

ASCI

Theroux, S., R.D. Mazor, M.W. Beck, P.R. Ode, E.D. Stein, and M. Sutula. 2020. Predictive biological indices for algae populations in diverse stream environments, *Ecological Indicators*, <https://doi.org/10.1016/j.ecolind.2020.106421>

IPI

Rehn, A.C., R.D. Mazor and P.R. Ode. 2017. An index to measure the quality of physical habitat in California wadeable streams. Swamp Technical Memorandum SWAMP-TM-2018-0005.

Natural (background) Specific Conductivity

Olson, J.R. and C.P. Hawkins. 2012. Predicting natural base-flow stream water chemistry in the western United States. *Water Resources Research* 48: W02504.

Reference Site Screening

Ode, P.R., A.C. Rehn, R.D. Mazor, K.C. Schiff, E.D. Stein, J.T. May, L.R. Brown, D.B. Herbst, D. Gillett, K. Lunde and C.P. Hawkins. 2016. Evaluating the adequacy of a reference site pool for the ecological assessment of streams in environmentally complex regions. *Freshwater Science* 35: 237-248.

Software and Toolbox Requirements

The following software is required to calculate indices:

- ArcGIS 10.2.2 or higher

- Be sure your license enables use of the Spatial Analyst Extension. For California State Water Board employees, do so by using “ArcMap - Advanced” instead of “ArcMap”
- You will need ArcGIS version 10.5 and above to use the Indices Tool
- Spatial Analyst Extension (extension for ArcGIS)
 - Add the extension by going to the Customize toolbar, click on “Extensions...”, check the Spatial Analyst box and press Close
- NHDPlusV2 Basin Delineator V2 2.5.0.22
- Watershed metric toolbox (download location below)
- R Studio 1.0.136 or R 3.3.2
- Microsoft .NET 4.6.10. or higher
- Microsoft SQL Server 2012 Express LocalDB 64-bit

Watershed Metric Toolbox Download

The toolbox and the geodatabase can be downloaded at the [CSU Chico Webpage](#).

The toolbox includes:

- Watershed_Metric_Resources_v*.gdb
- Watershed Metric Toolbox v*.tbx
- Watershed Metric and Index Calculation Instructions.pdf

*version number. For example: v4.6. The most up to date version will be provided at the link above.

Section 1: Creating Base Shapefiles for Metric Calculations

Base shapefiles (base files) are shapefiles that function as the unit of spatial analysis for calculation of CSCI, ASCI and IPI predictors and other spatial metrics. Index predictors are calculated with two types of base files:

- 1) a sites file containing the latitude and longitude of stream sampling locations and,
- 2) a delineated catchments file containing polygons that define the upstream area that drains to a given sampling location.

All base files must contain a unique identifier of each station, which we call “StationCod” (this field name gets automatically changed to “StationCode” when data is exported for analysis). “StationCod” fields must be represented in all shapefiles, using the same

letter case, and must not contain periods, special characters, or spaces. Each StationCod identifier must contain fewer than 18 characters.

Creating the Sites Base File

Background

The goal of creating the sites base file is to produce a shapefile representing stream sampling locations. Where possible, the location of sample points is automatically adjusted (“snapped”) from the original target coordinates to the nearest streamline represented in the National Hydrography Dataset Plus (NHDPlus). NHDPlus stream segments are based on medium resolution NHD (1:100,000) and as such sampling sites may not all have a represented stream segment in NHD. Where no NHD stream segment is present, users may need to manually digitize the corresponding upstream catchment. The “snapping” step is optional but is recommended because it improves the efficacy of the catchment delineation process and the calculation of predictor variables and metrics for screening reference sites. If snapping is not desired, you may skip the [“Snapping Sites to NHDPlus Flowlines”](#) below, but be sure to give subsequent delineations, metrics, and other analytical products additional scrutiny.

Data Requirements

Before you get started, it is recommended that a folder for all base file data is created to store site point and catchment layers. Next, create a spreadsheet with the following fields:

- 1) unique site identifier/s (field name should be StationCod)
- 2) coordinates in decimal degrees (field names should be: TargetLatitude and TargetLongitude)

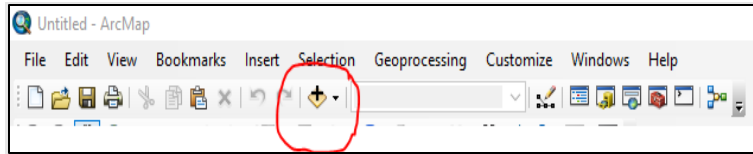
	A	B	C	D
1	StationCod	TargetLatitude	TargetLongitude	
2	509MJR099	40.28148	-121.75996	

Finally, ensure you have NHDPlus V2 Flowline data available for site snapping. The NHDPlus V2 Flowline data are provided in the Watershed metric toolbox, Watershed_Metric_Resources_v*.gdb under the “Physical_Features” dataset. Please note that data from the Watershed Metric Toolbox will need to be unzipped before uploading to ArcMap in the “Snapping Sites to NHDPlus Flowlines” step below.

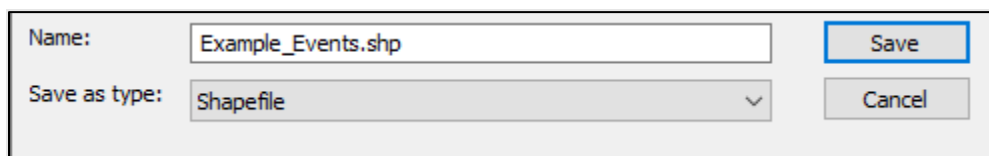
Loading Data in ArcMap

The steps below detail the process to load your data into ArcMap so that you can create the base shapefiles (sites file and delineated catchment file).

1. Load spreadsheet in ArcMap (Click the “Add Data” button)



2. Right-click the file you just loaded and click “Display XY Data...”. Fill out the “Display XY Data” window as follows:
 - a. X Field: TargetLongitude
 - b. Y Field: TargetLatitude
 - c. Z Field: <None>
 - d. Set the Coordinate system to WGS 1984.
 1. Click “Edit...” under the “Coordinate System of Input Coordinates” description box.
 2. Expand “Geographic Coordinate Systems”
 3. Expand “World”
 4. Select “WGS 1984” and click “OK”
 - A. If you receive an error about your table not having an ObjectID field, press “OK”
 5. You have now created the “Events” layer
3. Export the displayed “Events” layer as a shapefile and add to your map
 - a. Right-click on the “Events” layer
 - b. Click “Data”, then “Export Data”. Choose your desired output name and location.
 - c. Use the same coordinate system as: this layer’s source data
 - d. Name the file “XXX_Events” and save it as a shapefile

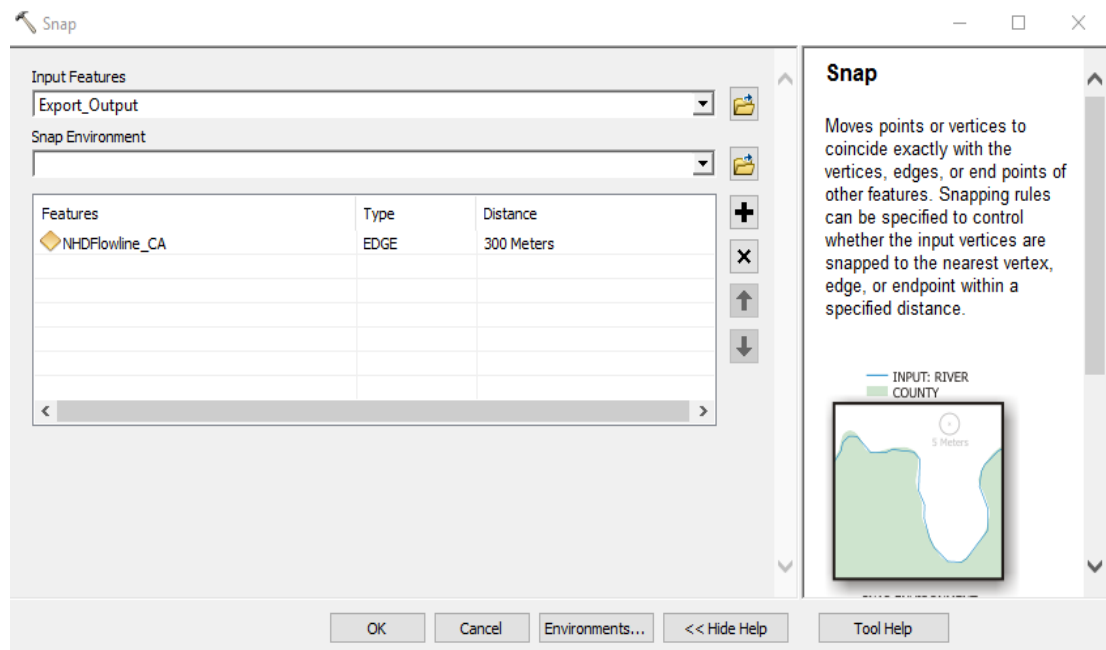


Snapping Sites to NHDPlus Flowlines

Snapping sites to NHDPlus flowlines is recommended because it improves the catchment delineation process and the calculation of predictor variables and metrics for screening reference sites. The steps below detail the snapping process.

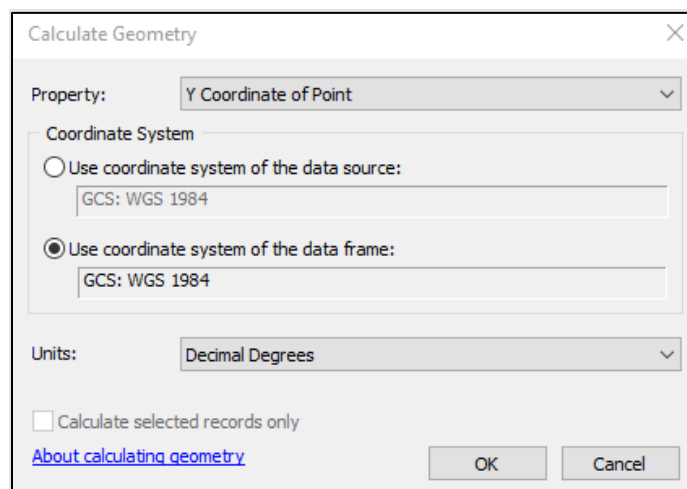
1. Load the NHDFlowline feature class from the Watershed Metric Resources geodatabase (GDB) into ArcMap
 - a. Click the “Add Data” button
 - b. Navigate to the file where you saved the Watershed_Metric_Resources_v*.gdb
 - c. Double-click “Watershed_Metric_Resources_v*.gdb” to open it

- d. Double-click the “Physical_Features” dataset
 - e. Select “NHDFlowline” and click “Add”
2. Snap the points to the nearest flowline in a manual edit session or with the “Snap” geoprocessing tool using “Edge Snapping”. To snap points:
- a. Open the ArcToolbox from the “Geoprocessing” menu or toolbar:
 - b. Expand “Editing Tools” and double-click “Snap” (see image below to what the “Snap” window looks like)



- c. Select exported sites shapefile as “Input Features”
- d. Select “NHDFlowline” as “Snap Environment”
 1. Until you complete steps “d” and “e” below, you may receive an error message via a red “x” on the right side of the feature name. Proceed with steps “d” and “e” below and the error should go away.
- e. Double click “Type” and Select “EDGE” (see image above as an example)
- f. Double click “Distance” and type: “300 Meters” (see image above as an example)
- g. Press “OK”

3. Once all sites are snapped, add a “New_Lat” and “New_Long” field to the snapped sites layer Attribute Table (see below for more guidance). Calculate the Latitude and Longitude of the newly snapped points
 - a. To add new fields, Right click the sites shapefile and open the “Attribute Table” for the snapped sites
 - b. Select “Add Field...” under “Table Options” (top-left icon in the table window)
 - c. Enter “New_Lat” for “Name”
 - d. Select “Double” under “Type”
 - e. Click “OK”
 - f. Repeat for “New_Long”
4. Calculate values “New_Lat” and “New_Long”
 - a. Right-click on field name “New_Lat”
 - b. Select “Calculate Geometry...” and click “Yes” on warning about calculating outside of an edit session
 - c. Select either X or Y Coordinate of Point (see below for specific instructions) and choose Decimal Degrees for Units.
 1. New_Lat = Y Coordinate of Point (see image below)
 2. New_Long = X Coordinate of Point (see image below)
 - d. Repeat for “New_Long”



5. Close the Attribute Table
6. Reproject the snapped sites layer to “NAD_1983_California_Teale_Albers”

- a. To reproject: select the View Tab-> Select: Data Frame Properties in the “Coordinate System” tab, expand “Projected Coordinate Systems”
 - b. Expand “State Systems”
 - c. Select “NAD 1983 California (Teale) Albers (Meters)” and click “OK”
7. Export the snapped sites layer using the coordinate system of the data frame to make the re-projection permanent
- a. Right-click on the sites shapefile and select “Data”, then select “Export Data”
 - b. Click on the yellow folder button
 - c. Rename it as “XXX_Sites” where “XXX” is the project name. Fill out the export data window with the following settings:

Name:	<input type="text" value="XXX_Sites.shp"/>	<input type="button" value="Save"/>
Save as type:	<input type="text" value="Shapefile"/>	<input type="button" value="Cancel"/>

- d. Use the same coordinate system as: “the data frame”
8. After snapping is complete, review each location per the following Quality Control checks

Quality-Control Checks for Site Base Files (for both snapped and unsnapped sites)

- Ensure that snapped locations are reasonably close to reported sampling locations (generally, less than 0.003 decimal degrees, or ~300 m on the ground). Sites that snapped larger distances should be flagged, so that the catchments delineated later can receive additional review. Larger snapping distances are not always problems and may have minimal impact on the catchment or the metrics calculated from the base files. In a few cases it can actually lead to an improvement in the position of a site (e.g., if the original coordinates plotted on a mountain side and the shift moved them down into the channel).
- Look for locations that did not snap to the NHDFlowline. These may be sites on smaller waterways not depicted in the NHDPlus data, or that were outside the snapping tolerance of 300 m.
- Look for metadata (such as station names or descriptions, aerial imagery, USGS topographic maps) to verify sampling location. Contact sampling crews if necessary.

- For sites close to confluences or near transitional areas, scrutiny is required to ensure that the site is located on the correct stream segment.

Creating the Delineated Catchment Base File

Background

Below, we outline the recommended approach for delineating catchments from a digital elevation model (DEM), simplified and improved by using pre-delineated watersheds in NHDPlus. This approach works well for most streams in California, although in certain situations, alternative delineation methods may be preferable (particularly in flat areas with minimal topographic variation). No matter what approach is used, the goal is to identify the portion of the landscape that contributes runoff to a stream under natural conditions; that is, dams, diversions, and inter-basin water transfers should be ignored when delineating the contributing catchment.

Delineation Tools

This section describes three different ways to delineate catchments: [Watershed Conversion Tool](#) web application (**preferred method**), followed by the [Basin Delineator](#), and finally the [USGS StreamStats](#) method.

Watershed Conversion Tool

This is the recommended option to delineate watersheds. The Watershed Conversion Tool was developed by the [Geographical Information Center](#) and utilizes the [USGS StreamStats Service API](#). It takes a csv file with the site ID, latitude, and longitude and outputs a polygon shapefile for spatial analyses.

Requirements:

- Sites base file (in csv format; see below instructions for guidance on how to create this file)
- Internet connection (Internet Explorer not supported)
- [Watershed Conversion Tool](#)

Steps:

1. In ArcMap, right click on your “XXX_Sites” layer
2. Click Open Attribute Table
3. Click Table Options, export All Records
4. Click yellow folder and rename it “XXX_Sites”
5. Save as text file and change .txt to .csv

6. Select ok
7. Navigate to the newly created csv and open in Excel
8. Delete all unnecessary fields, keeping only the “StationCod”, “New_Lat”, “New_Long”, in that order
9. Then delete the header row (Row 1) and ensure the file is saved as csv (comma delimited). The information should look like the example below:

	A	B	C
1	304APS	37.00187351	-121.9059523
2			

10. Navigate to the [Watershed Conversion Tool](#) website and upload the newly created csv file

WATERSHED CONVERSION TOOL

UPLOAD FILE

Choose a file or drag it here
*CSV and Excel files only

VERIFY DATA *editable*

Station Code	Latitude	Longitude
113NC0028	38.99582456	-123.6629202
113NC0104	38.71167962	-123.3991316
113NC0204	38.67217221	-123.5702619
113NC0504	39.09518828	-123.6076392
113HSCABC	38.6541368	-123.1748264
113NC0396	38.85640637	-123.5617681

CTRL + [wheel] on map to move marker
*updates data when done moving

SUBMIT DATA FOR CONVERSION

11. Inspect the editable table and ensure that all data was loaded correctly. Once verified, click “Submit Data for Conversion” on the bottom-right corner of the website. The tool will approximate the amount of time needed for conversion. An example of how the information looks when loaded correctly is below:

VERIFY DATA *editable* ADD ROW

	StationCod	Latitude	Longitude
✘	304APS	37.00187351	-121.9059523

12. Press “Submit” to begin the conversion

13. Once complete, the map window will show a preview of the watersheds. In the bottom right of the screen, name your output file and click the “Download Shapefiles” button
14. A zip file containing the delineated watershed/s will be downloaded to your computer. Find the export.zip file and extract it to the folder you are working in
15. Rename “ws_output.shp” to “XXX_Catchments”
16. Add the shapefile to your ArcMap session.
 - a. Select “Add Data”
 - b. Navigate to the folder where you extracted the delineated watershed file/s
 - c. Select the “XXX_Catchments” shapefile
17. Perform initial quality control checks (see section - [Quality Control Checks for Catchment Delineations](#))
18. Once complete, proceed to the [Finalize Delineated Catchment Base Files](#) section below (i.e. skip the “Basin Delineator” and “StreamStats” sections)

Basin Delineator

This method consists of software that can be downloaded to user computers.

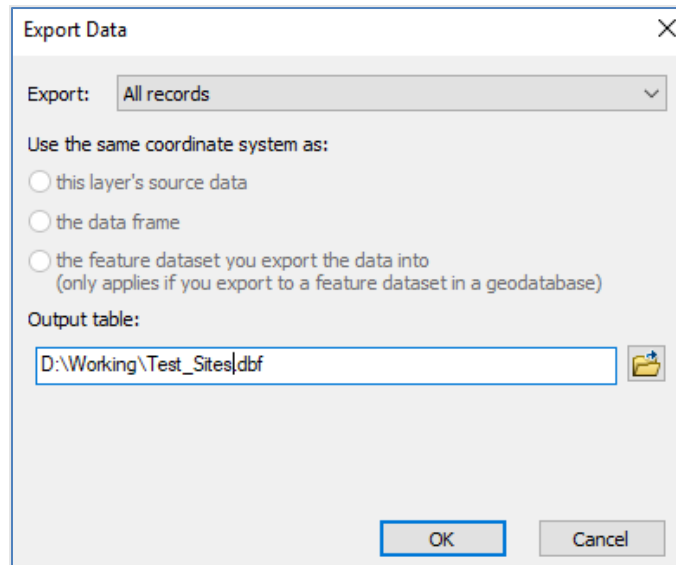
Requirements:

- Sites base file (in Tab Delimited Text format)
- [Download the NHDPlusV2 BasinDelineator Tool](#)
- ArcGIS 10.5.1 or higher and Spatial Analyst Extension
- Microsoft .NET Framework 4.6.1 or higher
- Microsoft SQL Server 2012 Express LocalDB 64-bit
- [Review the setup instructions](#)

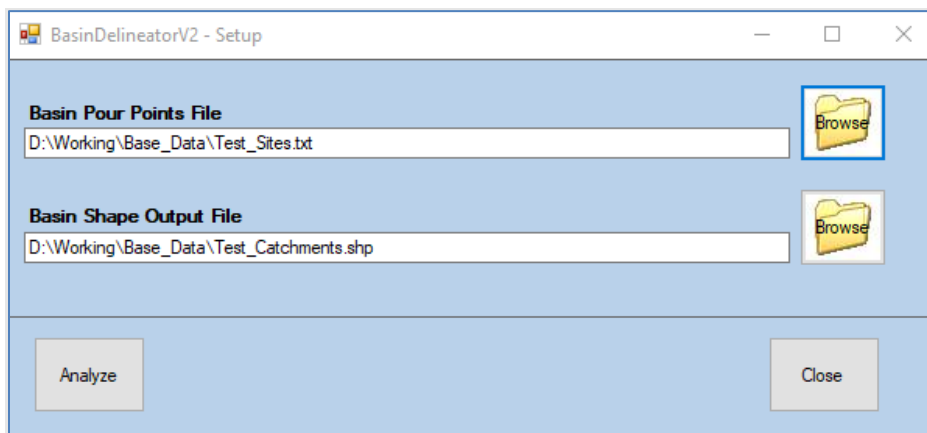
Steps:

1. Install Basin Delineator on your computer in accordance with read me instructions. Additionally, under system setup make sure to set the following parameters
 - a. The Fields in the Pour Point File are: “Basin ID, Latitude, Longitude”
 - b. Check the box for “Split Catchments”
 - c. Check the box for “Fill Interior Holes”

- d. (Optional) Check the box for “Use Simplified Polygons for the Catchments”



2. Open your sites attribute table and export as .dbf by selecting “Export...” under tables options
3. Navigate to the newly created table and open .dbf file in Excel
 - a. Delete all unnecessary fields, keeping columns for “StationCod”, “New_Lat”, “New_Long” in that order
 - b. Save the file as Text (tab delimited)
4. Start the Basin Delineator and click “Run BasinDelineatorV2”
 - a. The “Basin Pour Points File” is the tab delimited text file you made; browse to it and open
 - b. Set the “Basin Shape Output File” to an appropriate directory (e.g. your “Base_Data” folder) and save as “XXX_Catchments”, where “XXX” is the project name (example shown below):



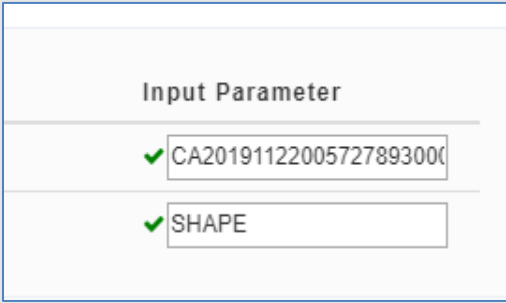
5. Click “Analyze” and watch for a pop-up when it completes. Depending on the number of catchments, delineation can take a long time
6. After acknowledging the process has completed, you may get a second pop-up saying that it was unable to delineate a number of catchments. This pop-up also tells you a new file called “SubmittedFileName_PPErrors.txt” was created. This file contains the list of sites that could not be delineated. Note that you may have to delineate these manually (See Step 9 below)
7. Load the basin shape output file in ArcMap, along with the local hydrology, HUC8 boundaries, snapped points, and a topographic base map for reference
8. Perform initial quality control checks (see section - [Quality Control Checks for Catchment Delineations](#))
9. If you have catchments that failed to delineate with Basin Delineator, try using the Watershed Conversion Tool or USGS StreamStats methods. Otherwise manually delineate failed catchments by freehand digitizing them in an edit session using a topographic base map and HUC 8, or NHDPlus boundaries as a reference for interpretation

StreamStats

The USGS StreamStats web service can be used to delineate one watershed at a time if either of the above methods fail. The [StreamStats method](#) asks the user to submit sampling locations online and returns delineated catchments as zipped shapefiles back to the user.

Steps:

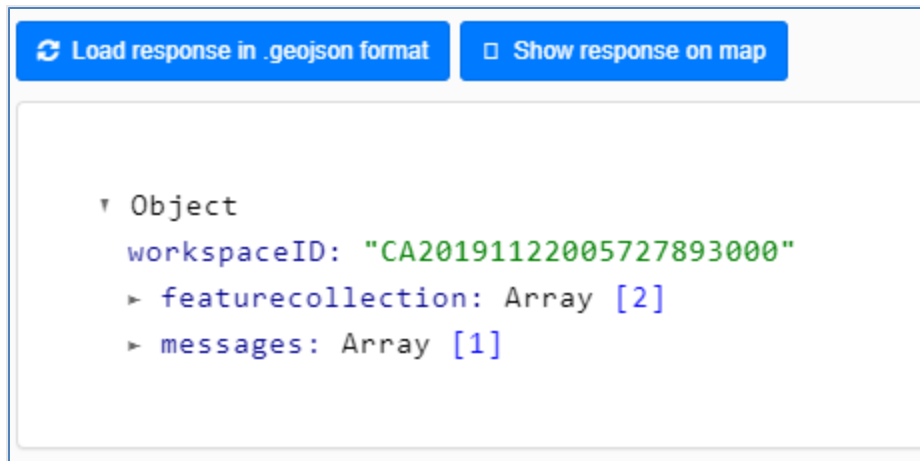
1. Navigate to “Watershed” tab and select “Delineate Watershed By Location”
2. Fill in “CA” for “rcode”, “New_Long” value for “xlocation”, and “New_Lat” value for “ylocation”. All other parameters should be left as-is



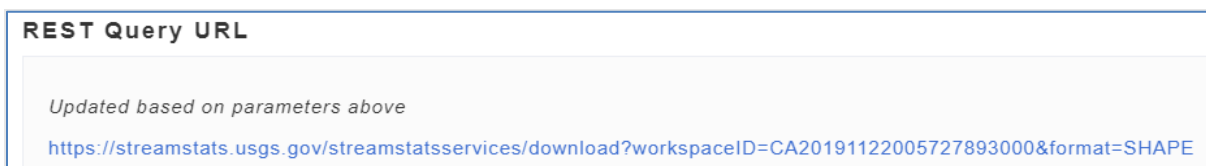
The image shows a screenshot of a web form titled "Input Parameter". It contains two input fields, each with a green checkmark to its left. The first field contains the value "CA20191122005727893000" and the second field contains the value "SHAPE".

Input Parameter	
✓	CA20191122005727893000
✓	SHAPE

3. Click on “Load response in .geojson format” button



4. Once done loading, copy the “workspaceID” value
5. Navigate to “Download” tab and select “Download By Workspace”. Paste unique numbers into “workspaceID” field and then click on hyperlink below “REST Query URL” to download the .zip file containing the delineated catchment



6. Copy the output file back to your computer and load it into ArcMap, along with the local hydrology, catchments, HUCs, snapped points, and a base map
7. Reproject the catchment output using NAD_1983_California_Teale_Albers using the same method used to reproject the base sites above
8. Perform initial quality control checks (see following section - Quality-Control Checks for Catchment Delineations)

Quality-Control Checks for Catchment Delineations

Once you have your catchment delineations, review each location per the following Quality Control checks.

- It’s helpful to reference the following datasets:
 - NHD Flow Lines: Users may want to hide pipelines but keep canals visible with a distinct color. In general, watersheds should never cut across a flow line other than pipelines or canals, except at the sample site.
 - NHD HUC 8 Boundaries: In general, watersheds should not cut across HUC 8 boundaries.

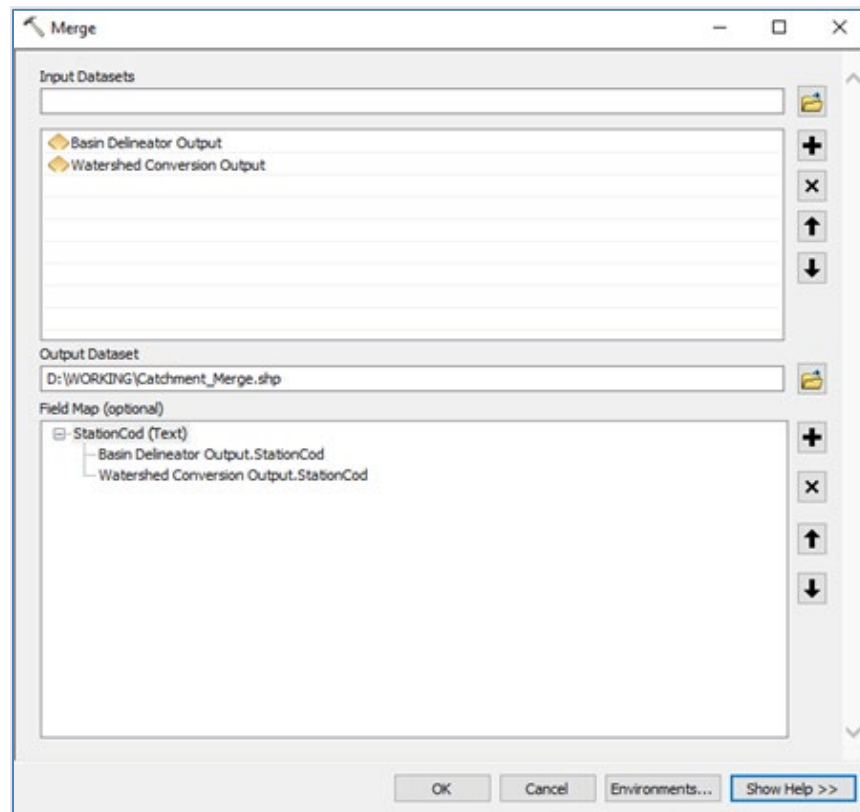
- Elevation files, shaded relief maps or topographic maps.
- In general, it is best to examine each catchment individually. Highlighting (or selecting) each catchment, one at a time, makes many problems obvious
- Look for gross irregularities, such as:
 - Holes within the catchment polygon boundary. Fix holes by removing the polygon vertices that create the hole (this is a pretty rare problem)
 - Small, nonsensical polygons that clearly don't correspond to a drainage network. These tend to occur when the coordinates plot off of a stream line and/or when the stream is in a flat area with little or no topographic relief
 - “Lollipop” or “Frying Pan”-shaped catchment polygons. This problem is most common when the site is located in a flat area with few topographic features. Unless the shape is supported by the local topography, flag the site for further review
- Use catchments from the corresponding region's NHDPlus dataset as a guide to fixing “lollipops”. Select and merge NHDPlus catchments to delineated catchments where necessary to complete and remove delineated catchment irregularities, or manually correct in ArcGIS editing session
- For sites close to confluences (within ~300 m), make sure that the “correct” catchment was delineated. The only way to verify this may be to check the original site name or description, or to check with the original field crew that sampled the site
- Follow the perimeter of the delineation around the entire watershed. Note the following potential errors:
 - Does the delineation cross any ponds, reservoirs, or lakes? If so, does the topography support inclusion in/exclusion from the watershed? Fix, or flag for further review
 - Do any NHDPlus flowlines cross the watershed border? If so, does the topography support inclusion in /exclusion from the watershed? Flowlines that represent pipelines, canals or aqueducts (or any situation where the flowline does not receive water from the immediate landscape) should be ignored. If necessary, check the site with imagery from Google Earth or other reference sources. Fix, or flag for further review
 - Most errors are small and will have negligible influence on CSCI scores or other predictors. As a rule of thumb, errors can be ignored if they would modify the total area of the catchment <5%, and do not alter the type of land-use inside the delineation
 - Watch for “divots” in the catchment perimeter. If the hydrology of peripheral drainages does not connect to the rest of the hydrologic

network, those drainages will not be included in the catchment by the delineator even if they clearly feed into the catchment. Select any affected NHDPlus catchment and merge it into the delineated catchments in this case

Finalize Delineated Catchment Base Files

Regardless of the method used to delineate catchment base files, you will need to join identifying attributes from the sites base files to the delineated catchments shapefile.

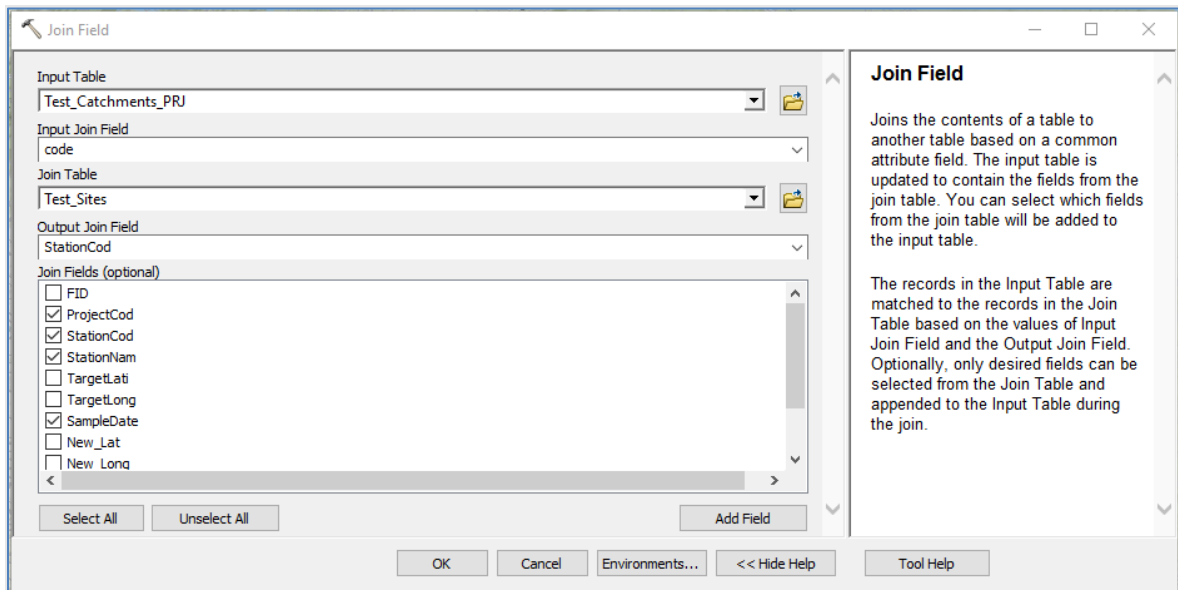
1. If multiple methods of delineation are used, merge the output catchment shapefiles into a single shapefile before proceeding
 - a. To merge shapefiles, open the ArcToolbox from the “Geoprocessing” menu or toolbar:
 - b. Expand “Data Management Tools” and “General”; double-click “Merge”
 - c. Add each catchment shapefile output you wish to merge to the Input Dataset list



- d. Select an Output Dataset location and review the Field Map. The field map controls how attributes from the inputs will be mapped and transferred to the output dataset. Be sure that your output StationCod field has the

proper corresponding fields from your inputs. You may also have other identifying attributes you wish to include in your field map

2. Once all catchments have been reviewed and delineated, project the shapefile into NAD_1983_California_Teale_Albers using the same method used to reproject the Base Sites above
 - a. Select “View”, then “Data Frame Properties”, then “Coordinate System” tab, and select “NAD_1983_California_Teale_Albers”, and click “OK”
3. Join the necessary fields from the “XXX_Sites” layer to your “XXX_Catchments” layer, as shown below
 - a. To join fields, open the ArcToolbox from the “Geoprocessing” menu or toolbar:
 - b. Expand “Data Management Tools” and “Joins”; double-click “Join Field”



1. **Input Table:** “XXX_Catchments”
2. **Input Join Field:** “code⁷”
3. **Join Table:** “XXX_Sites”
4. **Output Join Field:** “StationCod”

⁷ This example works if the GIC tool is used but not if Basin Delineator or StreamStats is used. Use the appropriate field with values that match StationCod.

5. **Join Fields:** “StationCod”, and any other identifying information you wish to include with the Catchments (e.g. “ProjectCod”, “New_Lat”, “New_Long”)
6. Click “OK”

Section 2: Calculating Bioassessment Index Predictor Variables in GIS

This section provides instruction on how to calculate index predictor variables in GIS using Indices Processor.

Calculating Index Predictor Variables in GIS

The goal of this sub-section is to guide users through the steps needed to calculate the predictors required for bioassessment indices: the California Stream Condition Index (CSCI), Algal Stream Condition Index (ASCI), Index of Physical Integrity (IPI) and predicted natural (background) specific conductivity.

These predictors are described in Table 1 below⁸:

Table 1 Index Predictor Variables

Predictor	Description
StationCode	Uniquely identifying code for the sample location
New_Lat	Latitude, in decimal degrees
New_Long	Longitude, in decimal degrees
AREA_SQKM	Watershed area in square kilometers
SITE_ELEV	Site elevation in meters.
MAX_ELEV	Elevation of the highest point of the catchment in meters.
ELEV_RANGE	Difference in elevation between the sample point and highest point in the catchment, in meters.
PPT_00_09	Average precipitation (2000 to 2009) at the sample point, in hundredths of millimeters

⁸ We cannot guarantee the accuracy of metrics calculated using this document. Field names and records are case-sensitive.

Predictor	Description
TEMP_00_09	Average temperature (2000 to 2009) at the sample point, in hundredths of degrees Celsius
PSA8	Perennial Streams Assessment Region for the Site.
PSA6	Perennial Streams Assessment Region for the Site.
AtmCa	Catchment mean of mean 1994-2006 annual ppt-weighted mean Ca concentration
AtmMg	Catchment mean of mean 1994-2006 annual ppt-weighted mean Mg concentration
AtmSO4	Catchment mean of mean 1994-2006 annual ppt-weighted mean SO4 concentration
BDH_AVE	Average bulk soil density
CaO_Mean	Average calcium oxide (quicklime) in the catchment geology
KFCT_AVE	Average soil erodibility (K) factor
LPREM_mean	Catchment mean log geometric mean hydraulic conductivity
LST32AVE	Catchment mean of mean 1961-1990 first and last day of freeze.
MAXWD_WS	Catchment mean of 1961-1990 annual max number of wet-days
MEANP_WS	Catchment mean of mean 1971-2000 annual ppt
MINP_WS	Catchment mean of mean 1971-2000 min monthly ppt
MgO_Mean	Average magnesium oxide (magnesia) in the catchment geology
PRMH_AVE	Catchment mean soil permeability
S_Mean	Catchment mean whole rock S
SumAve_P	Mean June to September 1971 to 2000 monthly precipitation, averaged across the entire catchment.
TMAX_WS	Catchment mean of mean 1971-2000 max temperature
UCS_Mean	Catchment mean unconfined Compressive Strength
XWD_WS	Catchment mean of mean 1961-1990 annual number of wet days
P_MEAN	Catchment mean whole rock P

Indices Processor Tool

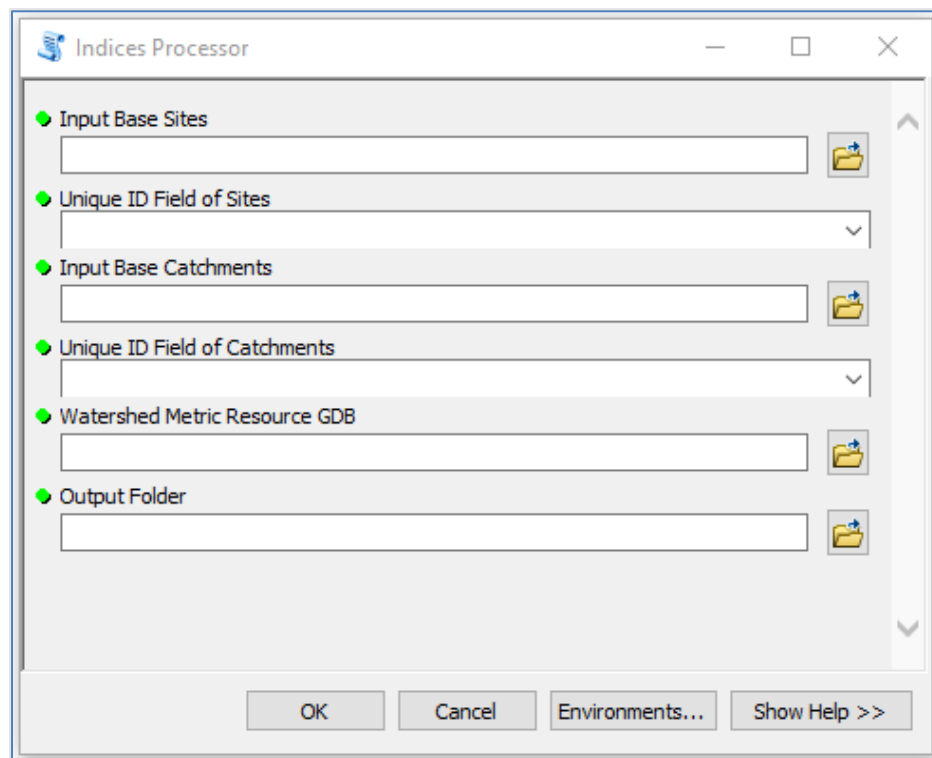
This tool is used to calculate all the predictors listed above which go into the CSCI, ASCI, and IPI indices and which are also used to calculate predicted conductivity. The following describes how to use the Indices Processor Tool in ArcGIS Desktop (version 10.5 and above).

Requirements:

- This Python Tool is designed for use with ArcGIS 10.5 and above and requires the Spatial Analyst Extension to run.

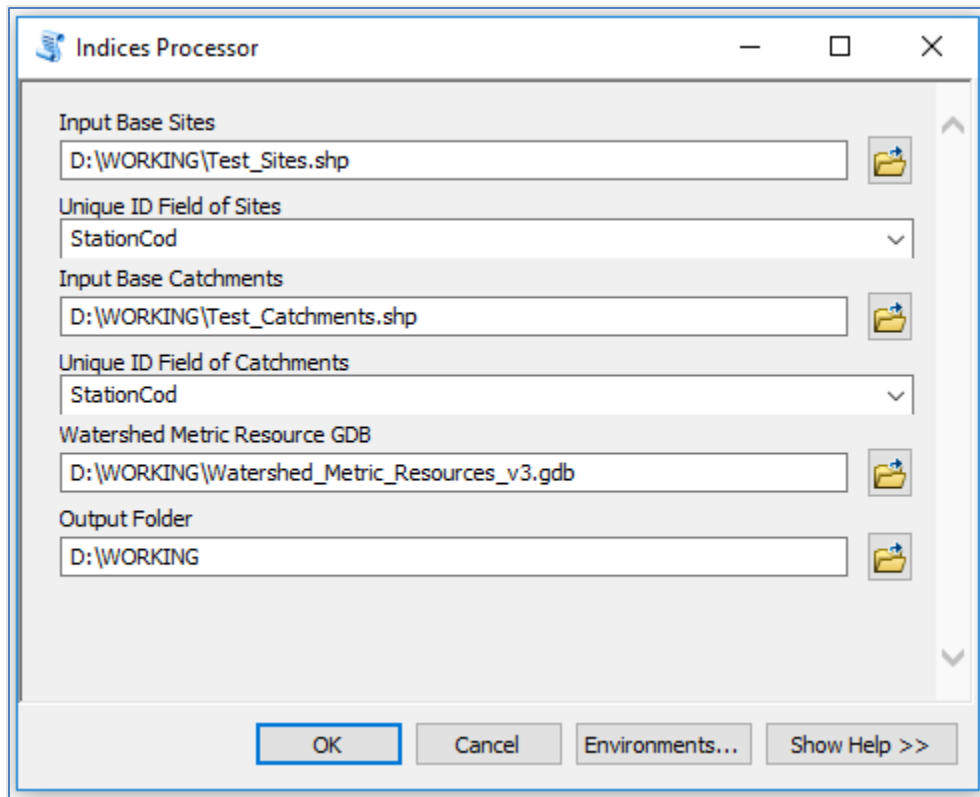
Steps:

1. Add “Watershed Metric Toolbox*.tbx” to your ArcToolbox by:
 - a. Right-clicking empty space within ArcToolbox window
 - b. Select “Add Toolbox”
 - c. Browse to “Watershed Metric Toolbox.tbx” on your computer and click “Open”
2. Navigate to the “Watershed Metric Toolbox v*” within ArcToolbox and click the “Index Toolset,” then double-click the “Indices Processor” script to open its dialog box



3. Add each of the inputs as described below. See example

- a. **Input Base Sites:** Navigate to and add the XXX_Sites point shapefile
- b. **Unique ID Field of Sites:** Choose the field that contains the unique ID for each input site point
- c. **Input Base Catchments:** Navigate to and add the XXX_Catchments polygon shapefile
- d. **Unique ID Field of Catchments:** Choose the field that contains the unique ID for each input catchment polygon. Reminder that this field must correspond with the “Input Base Sites” for the tool to run properly. All delineated catchments and corresponding site points should use the same unique ID value. In the example below, both “Test_Sites.shp” and “Test_Catchments.shp” contain the field “StationCod”. It is recommended to always use “StationCod” for the Unique ID in both the sites and catchment inputs
- e. **Watershed Metric Resource GDB:** Navigate to and add the “Watershed_Metric_Resources_v*.gdb” geodatabase containing all files for metric calculations
- f. **Output Folder:** Choose the location you wish the final results’ shapefiles to be saved. It is recommended that you create a new output folder within your working directory to store all of your metric output files. Intermediate files will also be saved here during processing but will be deleted upon completion. The output file “Metrics_Consolidated” is your stations file and can be renamed “XXX_stations” and saved as a “.csv”



4. Click “OK” and the tool will run. When it completes you should see shapefile outputs for each metric selected
 - a. Catchments_Elevation_Ranges.shp
 - b. Indices_Metrics_Consolidated.csv
 - c. PPTAvg_wgs84.shp
 - d. Sites_PSA.shp
 - e. TempMaxAvg_00_09wgs84.shp
 - f. Zonal_Stats_Metric_AtmCa.shp
 - g. Zonal_Stats_Metric_AtmMg.shp
 - h. Zonal_Stats_Metric_AtmSO4.shp
 - i. Zonal_Stats_Metric_BDH_AVE.shp
 - j. Zonal_Stats_Metric_CaO_Mean.shp
 - k. Zonal_Stats_Metric_EVI_MaxAve.shp
 - l. Zonal_Stats_Metric_KFCT_AVE.shp
 - m. Zonal_Stats_Metric_LPREM_mean.shp

- n. Zonal_Stats_Metric_LST32AVE.shp
- o. Zonal_Stats_Metric_MAXWD_WS.shp
- p. Zonal_Stats_Metric_MEANP_WS.shp
- q. Zonal_Stats_Metric_MgO_Mean.shp
- r. Zonal_Stats_Metric_MINP_WS.shp
- s. Zonal_Stats_Metric_N_MEAN.shp
- t. Zonal_Stats_Metric_P_MEAN.shp
- u. Zonal_Stats_Metric_PRMH_AVE.shp
- v. Zonal_Stats_Metric_S_Mean.shp
- w. Zonal_Stats_Metric_SumAve_P.shp
- x. Zonal_Stats_Metric_TMAX_WS.shp
- y. Zonal_Stats_Metric_UCS_Mean.shp
- z. Zonal_Stats_Metric_XWD_WS.shp

The tool may run for a long time depending on how many sites are being processed, and the size of their catchments. Processing time can range from just several minutes for 1-10 sites up to a few hours for 100+ sites. At this time, it is recommended that you process data in groups no larger than 100 sites for best performance.

Catchment Errors

When processing the predictors for specific conductivity, you will see a progress counter on how many catchments have been completed out of the total number in your input. In some cases, a catchment will not overlap properly with the input raster, or it may be too small relative to the input raster cell size. In these cases, the message “Warning: *[StationCod]* - Processing with center point selection method” is displayed. If you see this message, no action is necessary, but note that for the StationCod in question, the conductivity value was calculated using a single raster cell value, instead of the mean. You should review your catchments to make sure it is valid.

If you see the message, "Error with *[StationCod]* manual assessment required" displayed you should review your input catchment shapefile for corrupted data or geometry errors.

Section 3: Calculating Index Scores in R

Once index predictor variables are calculated following the steps described above ([Section 2](#)), they may be combined with taxonomic or habitat data to calculate bioassessment indices (CSCI, ASCI and IPI), or to estimate natural (background) levels of specific conductivity. All these steps are conducted in the R programming language (R) using packages developed for the California State Water Resources Control Board.

This document assumes that the user is familiar with basic operations in R, such as data import, export, and manipulation. Although not required, we recommend using graphic interface for R, such as RStudio. Users that are new to R or RStudio are encouraged to pursue training opportunities, such as those hosted by [local R user groups](#).

Updates to any of these packages will be announced over the SWAMP email listserve. To sign up for the listserve go to the [Water Board Email Lists website](#), scroll until you reach the category “Water Quality”, and check the box for “SWAMP Water Quality Monitoring (Surface Water Ambient Monitoring Program)”. Then scroll back up and enter your name and email address under “Signup Details” and click the Subscribe button. Users will be sent an email confirmation that they will need to accept to be added to the email list.

The California Stream Condition Index (CSCI)

The CSCI is the state’s standard bioassessment index for interpreting benthic macroinvertebrate data collected from wadeable streams using standard [SWAMP protocols](#). It replaces older regional indices of biotic integrity (IBIs), such as the Southern and Central California IBI (Ode et al. 2005), the North Coast IBI (Rehn et al. 2005), and the Central Valley IBI (Rehn et al. 2008).

The development, performance evaluation, and interpretation of this index is described in a journal article ([Mazor et al. 2016](#)), a shorter [technical memo](#), and a 2-page [fact sheet](#).

The following sub-section outlines how to install the CSCI package in R, including getting and preparing your stations and BMI taxonomy data, as well as how to calculate the CSCI. It also describes supplemental functions in the CSCI package and frequently asked questions.

Installing the CSCI package

To install the CSCI package and its dependencies, run the following lines of code:

```
install.packages("devtools") #Install devtools from CRAN
library(devtools)
install_github("SCCWRP/BMIMetrics")
install_github("SCCWRP/CSCI")
```

Once the installs have executed, users do not need to run this code unless they want to re-install these packages or if an update is issued. These lines will automatically install the `CSCI` package, as well as its dependent packages (e.g., `randomForest`, `vegan`, `stringr`, `reshape2`, `plyr`, and `data.table`). This process may take several minutes because the models and data tables required for the CSCI are large (~100 MB). Users may get a warning about the file size mismatching its reported length, but this warning may be disregarded. Depending on how the user's working environment is set up, they may need to install the `Rtools` package before they can install `devtools`.

If a user receives an error that names a package that failed to load (a "lazy loading" error), use the `install.packages()` function to load that package manually, and try again.

If installation is successful, the user should be able to load the `CSCI` library and access the help pages:

```
library(CSCI)
?CSCI
```

At the time of writing the most recent version of the CSCI package is: 1.2.2.

Stations Data

Stations data includes all the environmental information for each station, with one row per station. In the following steps, you will get the stations data from your working file and prepare it for index calculation.

Getting Stations Data

The output file "Metrics_Consolidated" from the previous Indices Processor step is your station data that you have renamed and saved as a `.csv`.

Preparing Input Data

The table below includes the required fields for your input data. Field names must match the spelling shown in the table below. For the required fields, blank cells or missing values are not allowed. Other fields of interest may be included in the stations'

data. Columns may appear in any order. Although we have implemented scripts to make the inputs case-insensitive, we recommend conforming to the capitalizations shown above.

Table 2 Station Input Data

Field Name	Description
StationCode	Unique identifier of the site
New_Lat	Latitude in decimal degrees
New_Long	Longitude in decimal degrees
SITE_ELEV	Site elevation (m)
ELEV_RANGE	Difference in elevation between the sample site and the highest point in the catchment
AREA_SQKM	Area of the catchment
TEMP_00_09	Long-term mean temperature at the site in hundredths of °C
PPT_00_09	Long-term mean precipitation at the site in hundredths of mm
SumAve_P	Mean summer precipitation across the catchment
KFCT_AVE	Average soil erodibility factor
BDH_AVE	Average soil bulk density
P_MEAN	Phosphorous content of the catchment geology

An example of properly formatted stations data is included in the package:

```
data(bugs_stations)
stations<-bugs_stations[[2]]
```

BMI (Bugs) Taxonomy Data

BMI samples should be collected with standard methods (see Ode et al. 2016b) and identified at least to the appropriate level of taxonomic resolution, SAFIT1a (i.e., most taxa to genus, with Chironomidae to subfamily). Samples identified to SAFIT2 may be scored as normal. Samples scored to SAFIT1 are generally unsuitable, but see the [Scoring SAFIT Level 1 Data Section below](#) for options on interpreting these data.

A minimum count of 500 individuals is desired to calculate CSCI scores, although smaller counts (>250) generally provide reliable information about stream condition.

Getting BMI (Bugs) Taxonomy Data

There are several sources for obtaining BMI taxonomy data: SWAMP Data Warehouse, CEDEN, and the SMC Data Portal.

SWAMP Data Warehouse

If you have access to the [SWAMP Data Warehouse](#), query your benthic data as you normally would. Go to “BMI Results” and export the report by clicking on the Download to csv button (ignore the Reporting Metrics box and download button). This report should be properly formatted for calculating the CSCI.

CEDEN

BMI data are available to the general public through the [CEDEN Advanced Query Tool](#) under the Benthic Result Category. However, these queries do not provide data in a usable format. Users will need to:

1. Filter benthic macroinvertebrate data from other organism types (filtering based on collection method is recommended)
2. Life stage and distinct information will need to be reformatted to meet the requirements of the CSCI package
3. Lastly, CEDEN data queries include header rows that must be deleted before importing into R

The `cleanData()` function and the `purge` argument may be helpful in reformatting data downloaded from CEDEN (see the [Cleaning Data with Bad or Missing Life Stage Codes Section](#) below).

SMC Data Portal

The [Stormwater Monitoring Coalition’s \(SMC’s\) data portal](#) provides access to data collected for the SMC’s stream survey, as well as under other programs in Southern California. The data portal has raw BMI taxonomy data, as well as calculated CSCI scores (where available). The “Advanced query” tool is recommended.

Preparing Input Data

BMI (bugs) data includes all the taxonomic information for each sample, with one row per taxon (i.e., flat-file format). Field names must match the spelling shown in the table below. Blank cells or missing values are not allowed, except for in the **Distinct** and **LifeStageCode** fields. All StationCodes used in the bugs file must also appear in the stations file, and vice versa. Columns may appear in any order. Although we have implemented scripts to make the inputs case-insensitive, we recommend conforming to the capitalizations shown above.

Table 3 BMI Input Data

Field Name	Description
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample. We recommend concatenating the following fields: StationCode, sample date, collection method, and field replicate number, separated by an underscore.
FinalID	Taxonomic names. Must match values in SWAMP organism lookup lists . The match is not case sensitive, and a few common misspellings are recognized.
Distinct	This field should be left blank for every row of the input data. The field is no longer used in CSCI calculations, but at this time it is still required.
LifeStageCode	Indicator of life stages: A for adult insects, L for larval insects, P for pupal insects, and X for non-insects. Not case sensitive. All combinations of FinalID and LifeStageCode must be found in SWAMP organism detail lookup lists If unknown or uncertain, you can use the cleanData() function, described below.
BAResult	Total count of organisms per FinalID

An example of properly formatted bug data is included in the package:

```
data(bugs_stations)
bugs<-bugs_stations[[1]]
```

Calculating the CSCI

Once you have obtained and prepared your stations and bugs data input files, CSCI scores can be calculated in R. The `CSCI` package automates all the necessary steps to calculate CSCI scores from properly formatted input files. It uses the predictor data in the stations input file to calculate biological expectations using random forest models. It uses the biological data in the bugs input file to calculate metrics and other biological endpoints. Additionally, it compares the endpoints to the expectations, relative to a reference distribution. We have automated many of these steps, with the goal of minimizing demands on the user. If interested to view the steps, they are available in [Appendix 1](#).

To calculate the CSCI:

First set your working directory, then load your bugs and stations data into the workspace and load the CSCI library. In this example, the bugs data is in a csv file named "bugs.csv", and the stations data is in a csv named "stations.csv":

```
# Set your Working Directory
setwd("C:/Users/FILE PATH")

# Load Data
bugs.df<-read.csv("bugs.csv")
stations.df<-read.csv("stations.csv")

# Load CSCI Library
library(CSCI)
```

Next, use the `CSCI()` function to calculate scores from the bugs and stations data:

```
Report <- CSCI(bugs=bugs.df, stations=stations.df)
```

There are only two required arguments for the `CSCI()` function: `bugs` and `stations`. Optional arguments include:

`rand`: Specify an integer to set the random seed, thereby ensuring that the subsampling procedure can be replicated on repeated runs of the script. By default, set to `sample.int(1000, 1)`.

`purge`: Automatically excludes all **FinalID/LifeStageCode** combinations that do not match associated lookup lists. If TRUE, purged taxa will be listed in the output. If FALSE (default), any unrecognized combinations will cause an error. For most applications, *we do not recommend using this feature*. Instead, we recommend resolving mismatches of **FinalID/LifeStageCode** by reviewing the data.

Accessing and Interpreting the Scores

The `CSCI()` function produces 6 reports, each as a named data frame within a list. They can be accessed using normal R indexing:

```
# Save the lists in report as separate CSVs
write.csv(report[["core"]], file = paste0("CSCI_Report_",
  Sys.Date(), "_core_.csv"))

write.csv(report[["Suppl1_mmi"]], file =
  paste0("CSCI_Report_", Sys.Date(), "_Suppl1_mmi.csv"))
```

```

write.csv(report[["Suppl2_mmi"]], file =
  paste0("CSCI_Report_", Sys.Date(), "_Suppl2_mmi.csv"))
write.csv(report[["Suppl1_grps"]], file =
  paste0("CSCI_Report_", Sys.Date(), "_Suppl1_grps.csv"))
write.csv(report[["Suppl1_OE"]], file =
  paste0("CSCI_Report_", Sys.Date(), "_Suppl1_OE.csv"))
write.csv(report[["Suppl2_OE"]], file =
  paste0("CSCI_Report_", Sys.Date(), "_Suppl2_OE.csv"))

```

Table 4 List of Produced Reports

Report Component	Description
core	A summary of the CSCI results and data quality flags, averaged across 20 iterations.
Suppl1_mmi	A detailed breakdown of the pMMI component of the CSCI. Raw, predicted, and scored metric values, averaged across 20 iterations.
Suppl1_grps	Probability of biotic group membership, with one row per SampleID.
Suppl1_OE	A detailed breakdown of the O/E component of the CSCI. Operational taxonomic unit (OTU) capture probabilities and mean abundances, averaged across 20 iterations.
Suppl2_mmi	Similar to Suppl1_mmi, except with results for each iteration provided.
Suppl2_OE	Similar to Suppl1_OE, except broken down by iteration. Iteration-wise O/E scores are also provided.

Field definitions for each report are described below:

Table 5 Core Report

Field Name	Description
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample
Count	Total number of organisms in the sample. If purge=T, the post-purge number is shown. A minimum number has not been established, but samples with low values should be evaluated with caution.

Field Name	Description
Number_of_MMI_Iterations	Number of subsamples used to calculate the pMMI. If the count is less than 500, no subsampling is performed, and this field will show 1. Otherwise, 20 subsamples are performed.
Number_of_OE_Iterations	Number of subsamples used to calculate the O/E. If the total number of unambiguous taxa is less than 500, no subsampling is performed, and this field will show 1. Otherwise, 20 subsamples are performed.
Pcnt_Ambiguous_Individuals	Percent of the total number of individuals excluded from O/E calculation. A maximum number has not been established, but samples with high values should be evaluated with caution.
Pcnt_Ambiguous_Taxa	Percent of the total number of FinalIDs excluded from O/E calculation. A maximum number has not been established, but samples with high values should be evaluated with caution.
E	The sum of all capture probabilities greater than 0.5 at a site. Interpreted as the total number of common taxa expected at a site.
Mean_O	The number of common taxa (i.e., capture probability greater than 0.5) observed at a site, averaged across iterations.
OoverE	O/E, calculated as Mean_O divided by E.
OoverE_Percentile	The percentile of the O/E score, relative to the reference distribution. A minimum threshold has not been established, but low values should be considered indicative of degradation.
MMI	The pMMI score, averaged across 20 iterations. A minimum threshold has not been established, but low values should be considered indicative of degradation.
MMI_Percentile.	The percentile of the pMMI score, relative to the reference distribution. A minimum threshold has not been established, but low values should be considered indicative of degradation.

Field Name	Description
CSCI	The CSCI score, calculated as the average of the O/E and pMMI.
CSCI_Percentile	The percentile of CSCI score, relative to the reference distribution. A minimum threshold has not been established, but low values should be considered indicative of degradation.

Table 6 Suppl1_mmi Report

Field Name	Description
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample
MMI_Score	pMMI score
Clinger_PercentTaxa	Observed percent clinger taxa
Clinger_PercentTaxa_predicted	Predicted percent clinger taxa
Clinger_PercentTaxa_score	Score for percent clinger taxa metric
Coleoptera_PercentTaxa	Observed percent Coleoptera taxa
Coleoptera_PercentTaxa_predicted	Predicted percent Coleoptera taxa
Coleoptera_PercentTaxa_score	Score for percent Coleoptera taxa metric
Taxonomic_Richness	Observed taxonomic richness
Taxonomic_Richness_predicted	Predicted taxonomic richness
Taxonomic_Richness_score	Score for taxonomic richness metric
EPT_PercentTaxa	Observed percent Ephemeroptera, Plecoptera, and Trichoptera (EPT) taxa
EPT_PercentTaxa_predicted	Predicted percent EPT taxa

Field Name	Description
EPT_PercentTaxa_score	Score for EPT percent taxa metric
Shredder_Taxa	Observed number of shredder taxa
Shredder_Taxa_predicted	Predicted number of shredder taxa
Shredder_Taxa_score	Score for shredder taxa metric
Intolerant_percent	Observed percent intolerant individuals (CTV<3)
Intolerant_percent_predicted	Predicted percent intolerant individuals
Intolerant_percent_score	Score for percent intolerant individuals metric

Note: (All values are averaged across 20 iterations)

Table 7 Suppl1_grps Report

Field Name	Description
StationCode	Unique identifier of the site
pGroupX	Probability that site is a member of group X.

Table 8 Suppl1_OE Report

Field Name	Description
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample
OTU	Operational taxonomic unit. All OTUs with capture probability greater than 0 are shown, but only those with a capture probability greater than 0.5 are used for scoring.
CaptureProb	Probability of observing the OTU at the site.
Mean Observed	Number of individuals observed in the sample, averaged across 20 iterations

Table 9 Suppl2_mmi Report

Field Name	Description
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample
metric	Name of the metric
Iteration	Unique identifier of the iteration
value	Observed metric value for each iteration
predicted_value	Predicted metric value. Same for all iterations.
score	Scored difference between predicted and observed value for each iteration of metric

Table 10 Suppl2_OE Report

Field Name	Description
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample
OTU	Operational taxonomic unit. Unlike Supplement 1, all OTUs are shown. Also, the O/E score for each iteration is shown where the OTU is “OoverE.”
CaptureProb	Probability of observing the OTU at the site.
IterationX	Number of individuals observed in Iteration X

Supplemental Functions in the CSCI Package

The `CSCI` package includes a number of functions that can facilitate analyses or provide useful ancillary information.

Accessing Metadata

The `loadMetaData()` function generates a table containing all recognized species names, including a few common misspellings. This table is used to aggregate to SAFIT Level II or to Operational taxonomic unit (OTUs), and to assign functional feeding groups, tolerance values, and other life history information used in metric calculation.

It can be particularly useful to consult this table when your input data includes unrecognized **FinalID/LifeStageCode** combinations.

Accessing Reference Data

The `loadRefData()` function generates a data frame containing reference data used to calibrate the CSCI. Predictor data, metrics, and index scores are provided. Both calibration (RefCal) and validation (RefVal) data are provided. For RefCal data, all predicted and scored values are based on out-of-bag predictions.

The `loadRefBugData()` function generates a data frame with the raw taxonomy data from reference sites.

Cleaning Data with Bad or Missing Life Stage Codes

The `cleanData()` function can help address errors caused by missing or incorrect life stage codes.

If your data are missing life stage codes, or contain values that do not match acceptable values in SWAMP, we recommend the following assumptions:

- All non-insects are X
- All *Hydraenidae* and *Hydrophilidae* are A
- All other insects are L

To automatically implement these assumptions on records that do not have acceptable life stage codes, you can use the `cleanData()` function:

```
bugs2 <- cleanData(bugs.df)
```

Although bad or missing life stage codes can be changed automatically, it may still be useful to know which records were incorrect. This information can be viewed two ways:

1. The `cleanData()` function by default returns the original input data with the addition of a new column called **fixedLifeStageCode**. This column is a T/F vector indicating which rows were corrected
2. The `cleanData()` function can be run using the `msgs=T` argument

```
bugs2 <- cleanData(bugs.df, msgs=T)
```

```
bugs2$data
```

```
bugs2$msg
```

This function will return a list object with two elements, where the first element `data` is the original data frame with the **fixedLifeStageCode** column and the second element `msg` is a string of messages indicating which rows had bad or missing life stage codes and which values were replaced.

Because this function deletes unrecognized taxa names, we strongly recommend reviewing the cleaned data for accuracy.

Scoring SAFIT Level 1 Data

In samples identified to SAFIT Level 1, *Chironomidae* (i.e., midges) are left at the family level, which is coarser than required for the O/E component of the CSCI. Thus, index scores will be depressed.

Scoring of samples identified to a SAFIT Level 1 is not recommended in most circumstances. If samples are archived, the best solution is to get midges identified to subfamily by a taxonomist who participates in SAFIT. If this is not feasible, your next best option is to calculate the range of possible CSCI scores. The lowest possible score is estimated by calculating the CSCI with all midges left at *Chironomidae*. The highest possible score can be estimated with the `MissingMidges()` function by calculating scores as though all expected *Chironomidae* subfamilies are present.

```
report <- CSCI(bugs=bugs, stations=stations)
report2 <- MissingMidges(report)
report2$core$CSCI_MissingMidges
```

In some cases, the range of possible CSCI scores may be small enough that decisions may be made with existing data (for example, if the highest possible score is below a target threshold, it may be determined that the site does not meet its objective). If the range is large enough to include an important threshold, it is recommended that samples be sent to a midge taxonomist rather than using the estimation approach described here.

Getting Metrics for Data Exploration or for Calculating Retired Indices of Biotic Integrity (IBI) Scores

The `CSCI` package does not directly calculate any of the IBI scores that predate the CSCI (e.g., the Southern and Central California IBI, Ode et al. 2005). However, it does enable calculation of the required metrics, which can then be scored by consulting the articles or reports where IBIs were originally described.

This code enables calculations of a large suite of biological metrics suitable for calculation of most of California's IBIs:

```
library(CSCI)
#Import the bugs data
bugs.df <- read.csv("bugs.csv")
#Coerce it into a "BMI" data object
bugdata <- BMI(bugs.df)
#Subsample to 500 individuals and aggregate
bugdata.samp <- sample(bugdata)
```

```
bugdata.agg <- aggregate(bugdata.samp)

#Calculate metrics at SAFIT Level 1

metrics <- BMIall(bugdata.agg, effort=1)
```

Frequently Asked Questions (FAQ) for CSCI Calculation

Most problems result from errors in data formatting, or other errors in the input data. Most errors will prevent complete execution of the `CSCI()` function. We have attempted to provide informative error messages to help guide corrections.

Bad or missing field names

All required field names must be present in input files. Please be sure to match the field names provided above. Although we have implemented scripts to make the inputs case-insensitive, we recommend conforming to the capitalizations shown in the tables above.

Bad or missing life stage codes

If your data are missing life stage codes, or contain values that do not match acceptable values in SWAMP, we recommend the following assumptions:

- All non-insects are X
- All Hydraenidae and Hydrophilidae are A
- All other insects are L

To automatically implement these assumptions on records that do not have acceptable life stage codes, you can use the `cleanData()` function described in the above [Cleaning Data with Bad or Missing Life Stage Codes Section](#).

Missing data

With few exceptions, missing values are not allowed. We recommend reviewing the data and filling in missing values as much as possible.

Bad FinalIDs

Bad FinalIDs typically result from misspellings, but occasionally occur when taxonomists do not conform to [SAFIT's standard taxonomic effort](#). If your dataset has incorrect bug names, you may use the `cleanData()` function with the `purge=T` argument:

```
bugs2 <- cleanData(bugs.df, purge=T)
```

This purged data frame can now be used with the CSCI function. However, it is *always preferable to correct the names* than to purge them, and the purge argument should only be used for preliminary analyses. To view rows in the data with incorrect entries for FinalID without purging, the `cleanData()` function can be executed with the `purge=F` and `msgs=T` arguments.

```
bugs2 <- cleanData(bugs.df, purge=F, msgs=T)
bugs2$data
bugs2$msg
```

This will return a list object with two elements, where the first element `data` is the original data frame with a new column called **problemFinalID** and the second element `msg` is a string of messages indicating which issues were encountered. The **problemFinalID** column is a T/F vector indicating which rows have incorrect **FinalID** values. The messages also indicate which observations in the input data had incorrect values for **FinalID**.

If you believe a **FinalID** is erroneously missing from SWAMP's lookup lists, please contact the [SWAMP help desk](#). If you believe a valid **FinalID** is inappropriately rejected by the scripts, contact [Raphael Mazor](#), Southern California Coastal Water Research Project.

The `loadMetaData()` function provides a table containing all recognized names, which may help identify misspellings or other problems creating errors. Please check this table before submitting a request for a modification to the script.

Importing characters as factors

R may import character vectors (like **FinalID**) as factors, which may not be interpreted correctly. We recommend importing all text fields as characters:

```
my.data.frame <- read.csv("myfile.csv", stringsAsFactors=F)
```

or coercing them into character format after they are imported:

```
my.data.frame$FinalID <- as.character(mydata.frame$FinalID)
```

Stations that are very close together

If you are scoring two stations that are so close together that the GIS data look identical, the `CSCI` function may produce an error. There are two easy workarounds you may use in this situation:

1. Remove one of the redundant rows from the stations data and treat the two samples as though they were coming from the same stations
2. Increase the precision of at least one GIS variable so they no longer appear identical (e.g., 5 or more decimal points)

Stations with catchments that include parts in Mexico

Portions of some streams include areas in Mexico. Because the geodatabases used to calculate `CSCI` predictors do not currently include this area, the `CSCI` cannot be calculated properly for these sites. In the interim, we make the following recommendations: If more than 90% of the area of a watershed is within California, treat

the state boundary as the edge of the watershed and calculate the predictors accordingly. However, you should interpret these results with caution, particularly if the portion within Mexico contains substantially different natural features. For watersheds that are less than 90% within California, we recommend using the Southern California Index of Biotic Integrity (Ode et al. 2005) as a substitute index.

Taxonomist overrides of distinct taxa designations (a.k.a., using distinct codes)

Taxonomist overrides of distinct taxa designations or pre-populating the distinct field in BMI input files are no longer recommended for standard CSCI scoring. The CSCI calculator does not correctly score samples if the designations are at better resolution than SAFIT Level 1. That is, the calculator includes taxa in richness estimates that should be aggregated to a higher taxonomic level (such as any genus, tribe, or subfamily *Chironomidae* that has been indicated as distinct). Because richness estimates appear in both the numerator and denominator of several metrics in the MMI, scores may be incorrectly inflated or deflated (although the latter is more common). We recommend leaving **Distinct** blank in all data inputs, without overriding the automated distinct taxon designation process.

The Algal Stream Condition Index (ASCI)

The ASCI refers to three individual multimetric indices based on benthic algal assemblages:

- D_ASCI, an index based on diatoms
- S_ASCI, an index based on soft-bodied (non-diatom) algae
- H_ASCI, a hybrid index based on both diatom and soft-bodied algae

These indices replace regional indices of biotic integrity (IBIs) developed for Southern California (Fetscher et al. 2014).

The manuscript describing the development, performance, and interpretation of these indices (Theroux et al., 2020) is in press and can be downloaded from the [ASCI homepage](#).

The following sub-section outlines how to install the ASCI package in R, including getting and preparing your stations and algae taxonomy data, as well as how to calculate the ASCI. It also describes supplemental functions in ASCI package and frequently asked questions.

Installing the ASCII Package

To install the `ASCII` package and its dependencies, run the following line of code:

```
install.packages("devtools") #Install devtools from CRAN  
library(devtools)  
install_github("SCCWRP/ASCII")
```

You do not need to run this code unless you want to re-install these packages or if an update is issued.

These lines will automatically install the `ASCII` package, as well as its dependent packages.

If installation is successful, you should be able to load the `ASCII` library and access the help pages:

```
library(ASCII)  
?ASCII
```

At the time of writing the most recent version of the `ASCII` package is: 2.3.1.

Stations Data

Stations data includes all the environmental information for each station, with one row per station. In the following steps, you will get the stations data from your working file and prepare it for index calculation.

Getting Stations Data

The output file “Metrics_Consolidated” from the previous Indices Processor step is your station data that you have renamed and saved as a `.csv`.

Preparing Stations Data

Stations data includes all the environmental information for each station, with one row per station. Field names must match spelling shown below. For the required fields, blank cells or missing values are not allowed. Other fields of interest may be included in the stations data. Columns may appear in any order.

Table 11 Stations Input Data

Field Name	Description
StationCode	Unique identifier of the site
AREA_SQKM	Watershed area in square kilometers
AtmCa	Atmospheric deposition of Calcium
AtmMg	Catchment mean of mean 1994-2006 annual ppt-weighted mean Mg concentration
AtmSO4	Catchment mean of mean 1994-2006 annual ppt-weighted mean SO4 concentration
BDH_AVE	Average bulk soil density
CaO_Mean	Average calcium oxide (quicklime) in the catchment geology
KFCT_AVE	Average soil erodibility (K) factor
LPREM_mean	Catchment mean log geometric mean hydraulic conductivity
LST32AVE	Catchment mean of mean 1961-1990 first and last day of freeze
MAX_ELEV	Maximum elevation in the catchment (m)
MAXWD_WS	Catchment mean of 1961-1990 annual max number of wet-days
MEANP_WS	Catchment mean of mean 1971-2000 annual ppt
MgO_Mean	Average magnesium oxide (magnesia) in the catchment geology
MINP_WS	Catchment mean of mean 1971-2000 min monthly ppt
PPT_00_09	Long-term mean precipitation at the site in hundredths of mm
PRMH_AVE	Catchment mean soil permeability
S_Mean	Catchment mean whole rock S
SITE_ELEV	Site elevation (m)
SumAve_P	Mean June to September 1971 to 2000 monthly precipitation, averaged across the entire catchment.
TMAX_WS	Catchment mean of mean 1971-2000 max temperature
UCS_Mean	Catchment mean unconfined Compressive Strength
XWD_WS	Catchment mean of mean 1961-1990 annual number of wet days

Field Name	Description
PSA6	Region of the state: NC: North Coast CH: Chaparral SC: South Coast CV: Central Valley SN: Sierra Nevada DM: Deserts and Modoc Plateau

An example of properly formatted stations data is included in the package:

```
data(demo_station)
```

The `ASCI()` function automatically calculates two additional predictors from these data if they are missing from the input data:

1. **XerMtn** is a binary field indicating if the site is located in one of the two “mountainous” PSA regions (1 if PSA6 is NC, SN; otherwise 0)
2. **CondQR50** is the predicted natural (background) conductivity, calculated as described in [Section 4](#).

These variables are ultimately used to calculate ASCI scores.

Algae Taxonomy Data

Getting Algae Taxonomy Data

Algae samples should be collected with standard methods (see Ode et al. 2016b) and identified to the appropriate level of taxonomic resolution. See section on the standardized taxonomic effort (STE) within the [Accessing Metadata section](#) below.

There are several sources for obtaining Algae taxonomy data: SWAMP Data Warehouse, CEDEN, and the SMC Data Portal.

SWAMP Data Warehouse

If you have access to the SWAMP Data Warehouse, query your benthic data as you normally would. Go to “Benthic Algae Results” and export the results (not the metrics!) as a csv. This report should be properly formatted for calculating the ASCIs.

CEDEN

Benthic algae data are available to the general public through the [CEDEN Advanced Query Tool](#) under the Benthic Result Category. However, these queries do not provide data in a usable format. Users will need to filter benthic algae data from other organism types (filtering based on collection method is recommended). Additionally, CEDEN data queries include header rows that must be deleted before importing into R.

SMC Data Portal

The [Stormwater Monitoring Coalition's \(SMC's\) data portal](#) provides access to data collected for the SMC's stream survey, as well as under other programs in Southern California. The data portal has raw algae taxonomy data, as well as calculated ASCI scores (where available). The "Advanced query" tool is recommended.

Preparing Taxonomy Data

Algae taxonomy data includes all the taxonomic information for each sample, with one row per taxon (i.e., flat-file format). Field names must match spelling shown above. No missing values are allowed except for the **Result** and **BAResult** fields (the function may execute properly, but the results may not be correct); rows where both **BAResult** and **Result** are missing will be ignored. Columns may appear in any order; non-required columns will be ignored.

Table 12 Taxonomy Input Data

Field Name	Description
StationCode	Unique identifier of the site
SampleDate	Date of sample collection
Replicate	Field replicate number
SampleTypeCode	Identifies the sample-type within a sample. Valid values are as follows: Diatoms Integrated SBA: Macroalgae Microalgae Epiphyte Qualitative Note: Qualitative components are ignored. If only a single assemblage is submitted, the H_ASCI will still be calculated however the sample will be flagged.
FinalID	Taxonomic names. Must match values in SWAMP organism lookup lists for <u>Diatoms</u> and <u>SBA</u> . The match is not case sensitive, and a few common misspellings are recognized.
BAResult	Total entity count of the organisms associated with Integrated or Epiphyte sample-types (typically used for diatoms and certain SBA taxa).
Result	Total biovolume of the organisms associated with Macroalgae and Microalgae sample-types (used for certain SBA taxa)

An example of properly formatted algae data is included in the package:

```
data(demo_algae_tax)
```

Calculating the ASCI

As with the CSCI, all steps to calculate the ASCI are automated by the ASCI package. However, we have provided a brief overview of the ASCI calculations below:

1. Final IDs are assigned to harmonized names according to the STE
2. Taxonomy data is split into diatom-only, soft-bodied only, and hybrid assemblages
3. All species observations are transformed to presence/absence data
4. Observed metric scores are calculated for all samples
5. For predictive metrics, stations data is used to predict metric values for individual samples, and residuals between observed and predicted metrics are calculated
6. Metric values are scored, averaged, and standardized by dividing by the mean from reference calibration sites

To calculate the ASCI:

First load your algae and stations data into the workspace and load the ASCI library. In this example, the algae data is in a csv file named “algae_tax.csv”, and the stations data is in a csv named “stations.csv”:

```
algae_tax.df <- read.csv("algae_tax.csv")
stations.df  <- read.csv("stations.csv")
library(ASCI)
```

The `ASCI()` function will calculate scores from the algae and stations data:

```
Report <- ASCI(taxa=algae_tax.df, stations=stations.df)
```

There are only two required arguments for the `ASCI()` function: `taxa` and `stations`.

Accessing and Interpreting the Scores

The `ASCI()` function produces a single report as a data frame, with a set of fields for each index.

The tables below summarizes the ASCI report:

Table 13 General Sample Report

Field Name	Description
SampleID	A concatenation of StationCode, SampleDate, and Replicate, delimited by an underscore.
StationCode	Unique identifier of the site
SampleDate	The date of sample collection
Replicate	The replicate number
SampleType	A concatenated string indicating which sample types were present in the input data for each sample.
D_ValveCount	Total number of diatom valves reported in the input data (i.e., sum of BAResult)
S_EntityCount	Total number of SBA entities reported in the input data (i.e., sum of BAResult)
S_Biovolume	Total SBA biovolume reported in the input data (i.e., sum of Result)
D_NumberTaxa	Number of diatom taxa
S_NumberTaxa	Number of SBA taxa
H_NumberTaxa	Number of diatom taxa + SBA taxa in hybrid ASCI
UnrecognizedTaxa	Taxa not recognized by calculator

Table 14 ASCI output

Field Name	Description
D_ASCI	Score for the Algal Stream Condition Index based on diatoms
S_ASCI	Score for the Algal Stream Condition Index based on soft-bodied algae
H_ASCI	Score for the Algal Stream Condition Index based on both diatoms and soft-bodied algae
D_cnt.spp.most.tol_pct_att	% of taxa attributed with tolerance value (QA metric)

Field Name	Description
D_cnt.spp.most.tol_pred	Count of most tolerant taxa (predicted)
D_cnt.spp.most.tol_raw	Count of most tolerant taxa (raw)
D_cnt.spp.most.tol_scr	Count of most tolerant taxa (score)
D_EpiRho.richness_pct_att	% of taxa attributed with genus (QA metric)
D_EpiRho.richness_pred	Richness of <i>Epithemia</i> and <i>Rhopalodia</i> taxa (predicted)
D_EpiRho.richness_raw	Richness of <i>Epithemia</i> and <i>Rhopalodia</i> taxa (raw)
D_EpiRho.richness_scr	Richness of <i>Epithemia</i> and <i>Rhopalodia</i> taxa (score)
D_prop.spp.IndicatorClass_TN_low_pct_att	% of taxa attributed with nitrogen indicator value (QA metric)
D_prop.spp.IndicatorClass_TN_low_pred	Proportion low total nitrogen indicator taxa (predicted)
D_prop.spp.IndicatorClass_TN_low_raw	Proportion low total nitrogen indicator taxa (raw)
D_prop.spp.IndicatorClass_TN_low_scr	Proportion low total nitrogen indicator taxa (score)
D_prop.spp.Planktonic_pct_att	% of taxa attributed with habitat value (QA metric)
D_prop.spp.Planktonic_pred	Proportion planktonic taxa (predicted)
D_prop.spp.Planktonic_raw	Proportion planktonic taxa (raw)

Field Name	Description
D_prop.spp.Planktonic_scr	Proportion planktonic taxa (score)
D_prop.spp.Trophic.E_pct_att	% of taxa attributed with trophic indicator value (QA metric)
D_prop.spp.Trophic.E_pred	Proportion eutrophic taxa (predicted)
D_prop.spp.Trophic.E_raw	Proportion eutrophic taxa (raw)
D_prop.spp.Trophic.E_scr	Proportion eutrophic taxa (score)
D_Salinity.BF.richness_pct_att	% of taxa attributed with salinity value (QA metric)
D_Salinity.BF.richness_pred	Richness of brackish/freshwater taxa (predicted)
D_Salinity.BF.richness_raw	Richness of brackish/freshwater taxa (raw)
D_Salinity.BF.richness_scr	Richness of brackish/freshwater taxa (score)
H_cnt.spp.IndicatorClass_TP_high_pct_att	% of taxa attributed phosphorus indicator value (QA metric)
H_cnt.spp.IndicatorClass_TP_high_pred	Count of high total phosphorus indicator taxa (predicted)
H_cnt.spp.IndicatorClass_TP_high_raw	Count of high total phosphorus indicator taxa (raw)
H_cnt.spp.IndicatorClass_TP_high_scr	Count of high total phosphorus indicator taxa (score)
H_cnt.spp.most_tol_pct_att	% of taxa attributed with tolerance value (QA metric)

Field Name	Description
H_cnt.spp.most.tol_pred	Count of most tolerant taxa (predicted)
H_cnt.spp.most.tol_raw	Count of most tolerant taxa (raw)
H_cnt.spp.most.tol_scr	Count of most tolerant taxa (score)
H_EpiRho.richness_pct_att	% of taxa attributed with genus (QA metric)
H_EpiRho.richness_pred	Richness of <i>Epithemia</i> and <i>Rhopalodia</i> taxa (predicted)
H_EpiRho.richness_raw	Richness of <i>Epithemia</i> and <i>Rhopalodia</i> taxa (raw)
H_EpiRho.richness_scr	Richness of <i>Epithemia</i> and <i>Rhopalodia</i> taxa (score)
H_OxyReD_DO_30.richness_pct_att	% of taxa attributed with oxygen tolerance value (QA metric)
H_OxyRed_DO_30.richness_pred	Richness of species with 30% oxygen tolerance (predicted)
H_OxyRed_DO_30.richness_raw	Richness of species with 30% oxygen tolerance (raw)
H_OxyRed_DO_30.richness_scr	Richness of species with 30% oxygen tolerance (score)
H_prop.spp.Planktonic_pct_att	% of taxa attributed with habitat value (QA metric)
H_prop.spp.Planktonic_pred	Proportion planktonic taxa (predicted)
H_prop.spp.Planktonic_raw	Proportion planktonic taxa (raw)

Field Name	Description
H_prop.spp.Planktonic_scr	Proportion planktonic taxa (score)
H_prop.spp.Trophic.E_pct_att	% of taxa attributed with trophic indicator value (QA metric)
H_prop.spp.Trophic.E_pred	Proportion eutrophic taxa (predicted)
H_prop.spp.Trophic.E_raw	Proportion eutrophic taxa (raw)
H_prop.spp.Trophic.E_scr	Proportion eutrophic taxa (score)
H_prop.spp.ZHR_pct_att	% of taxa attributed with ZHR value (QA metric)
H_prop.spp.ZHR_raw	Proportion ZHR species (raw)
H_prop.spp.ZHR_scr	Proportion ZHR species (score)
H_Salinity.BF.richness_pct_att	% of taxa attributed with salinity value (QA metric)
H_Salinity.BF.richness_pred	Richness of brackish/freshwater taxa (predicted)
H_Salinity.BF.richness_raw	Richness of brackish/freshwater taxa (raw)
H_Salinity.BF.richness_scr	Richness of brackish/freshwater taxa (score)
S_prop.spp.IndicatorClass_DOC_high_pct_att	% of taxa attributed with DOC indicator value (QA metric)
S_prop.spp.IndicatorClass_DOC_high_raw	Proportion high DOC indicator species (raw)
S_prop.spp.IndicatorClass_DOC_high_scr	Proportion high DOC indicator species (score)

Field Name	Description
S_prop.spp.IndicatorClass_NonRef_pct_att	% of taxa attributed with reference/non-reference indicator value (QA metric)
S_prop.spp.IndicatorClass_NonRef_raw	Proportion non-reference indicator species (raw)
S_prop.spp.IndicatorClass_NonRef_scr	Proportion non-reference indicator species (score)
S_prop.spp.IndicatorClass_TP_high_pct_att	% of taxa attributed phosphorus indicator value (QA metric)
S_prop.spp.IndicatorClass_TP_high_raw	Proportion high total phosphorus indicator taxa (raw)
S_prop.spp.IndicatorClass_TP_high_scr	Proportion high total phosphorus indicator taxa (score)
S_prop.spp.ZHR_pct_att	% of taxa attributed with ZHR value (QA metric)
S_prop.spp.ZHR_raw	Proportion ZHR species (raw)
S_prop.spp.ZHR_scr	Proportion ZHR species (score)
Comments	Miscellaneous flags (e.g. missing assemblages)

Supplemental Functions in the ASCI Package

The `ASCI` package includes a number of functions that can facilitate analyses or provide useful ancillary information.

Accessing Metadata

Data on traits used to calculate ASCI metrics can be accessed directly from the package:

```
data(milkup)
```

Data on standardized taxonomic effort and correct taxonomic nomenclature can be accessed directly from the package:

```
data(STE)
```

Frequently Asked Questions (FAQ) for ASCI calculation

Most problems result from errors in data formatting, or other errors in the input data. Most errors will prevent complete execution of the `ASCI ()` function. We have attempted to provide informative error messages to help guide corrections.

Accessing calibration data

Calibration data is accessible on the [ASCI homepage](#).

Bad or missing field names

All required field names must be present in input files. Please be sure to match the field names provided above. Although we have implemented scripts to make the inputs case-insensitive, we recommend conforming to the capitalizations shown above.

Missing data

With few exceptions, missing values in stations data are not allowed. The ASCI calculator will run with single assemblage (diatom or soft-bodied algae) data but it will flag these samples.

Stations with catchments that include parts in Mexico

Portions of some streams include areas in Mexico. Because the geodatabases used to calculate ASCI predictors do not currently include this area, the ASCI cannot be calculated properly for these sites. The geodatabases will be updated within the next few months. In the interim, we make the following recommendations: If more than 90% of the area of a watershed is within California, treat the state boundary as the edge of the watershed and calculate the predictors accordingly. However, you should interpret these results with caution, particularly if the portion within Mexico contains substantially different natural features. For watersheds that are less than 90% within California, we recommend using the Southern California Algal Indices of Biotic Integrity (Fetscher et al. 2014) as a substitute index.

Unrecognized taxa

Novel or misspelled species names will not be recognized by the calculator and will be output as unrecognized taxa. Users should modify these species in agreement with the SWAMP species lists and re-run the calculator.

The Index of Physical-habitat Integrity (IPI)

The IPI evaluates stream physical habitat data collected with the standard SWAMP protocol (Ode et al. 2016b) by comparing metrics to values expected under reference conditions. The development and interpretation of the index is described in a [technical memo](#).

The following sub-section outlines how to install the PHAB package in R (which will calculate the IPI scores), including getting and preparing your stations and PHAB metrics data, as well as how to calculate the IPI. It also describes supplemental functions in PHAB package and frequently asked questions.

Installing the PHAB Package

To install the `PHAB` package and its dependencies, run the following line of code:

```
install.packages("devtools") #Install devtools from CRAN
library(devtools)
install_github("SCCWRP/PHAB")
```

You do not need to run this code unless you want to re-install these packages or if an update is issued.

These lines will automatically install the `PHAB` package, as well as its dependent packages.

Depending on how your working environment is set up, you may need to install the `Rtools` package before you can install `devtools`.

If you get an error that names a package that failed to load (a “lazy loading” error), use the `install.packages()` function to load that package manually, and try again.

If installation is successful, you should be able to load the `PHAB` library and access the help pages:

```
library(PHAB)
?IPI
```

At the time of writing the most recent version of the PHAB package is: 0.1.1.

Stations Data

Stations data includes all the environmental information for each station, with one row per station. In the following steps, you will get the stations data from your working file and prepare it for index calculation.

Getting Stations Data

The output file “Metrics_Consolidated” from the previous Indices Processor step is your station data that you have renamed and saved as a “.csv”.

Preparing Stations Data

Stations data includes all the environmental information for each station, with one row per station. Field names must match spelling shown above. For the required fields, blank cells or missing values are not allowed. Other fields of interest may be included in the stations data. Columns may appear in any order. Although we have implemented scripts to make the inputs case-insensitive, we recommend conforming to the capitalizations shown below.

Table 15 Station Input Data

Field Name	Description
StationCode	Unique identifier of the site
New_Lat	Latitude in decimal degrees
New_Long	Longitude in decimal degrees
SITE_ELEV	Site elevation (m)
MAX_ELEV	Maximum elevation in the catchment (m)
ELEV_RANGE	Difference in elevation between the sample site and the highest point in the catchment
AREA_SQKM	Area of the catchment
PPT_00_09	Long-term mean precipitation at the site in hundredths of mm
KFCT_AVE	Average soil erodibility factor
MEANP_WS	Catchment mean of mean 1971-2000 annual ppt
MINP_WS	Catchment mean of mean 1971-2000 min monthly ppt

An example of properly formatted stations data is included in the package:

```
data(stations)
```

PHAB Data

Getting PHAB Data

The `IPI()` function only works with PHAB metrics, not raw PHAB data. At this time, the `PHAB` package does not support metric calculation.

There are several sources for obtaining PHAB data: SWAMP Data Warehouse, CEDEN, and the SMC Data Portal.

SWAMP Data Warehouse

If you have access to the SWAMP Data Warehouse, query your benthic data as you normally would. Go to “Habitat Results” and export the metrics by clicking on the Download Results to csv button within the Reporting Metrics box (make sure the SWAMP metrics radio button is selected). This report should be properly formatted for calculating the IPI.

CEDEN

Raw physical habitat data are available to the general public through the [CEDEN Advanced Query Tool](#) under the Habitat Category. However, these queries do not provide data in a usable format, and CEDEN data queries include header rows that must be deleted before importing into R.

At this time, CEDEN does not have a method for calculating PHAB metrics required for the IPI, as described in Rehn et al. (2017).

SMC data portal

The [Stormwater Monitoring Coalition’s \(SMC’s\) data portal](#) provides access to data collected for the SMC’s stream survey, as well as under other programs in Southern California. The data portal has raw PHAB data, as well as calculated PHAB metrics and IPI scores (where available). The “Advanced query” tool is recommended.

Preparing PHAB Metrics Data

The PHAB data includes calculated physical habitat metrics that are compiled along with the stations data to get the IPI score. These data are in long format where multiple rows correspond to physical habitat metric values for a single site.

Table 16 PHAB Metrics Data

Field Name	Description
StationCode	Unique identifier of the site
SampleDate	Date of data collection
SampleAgencyCode	Unique identifier of the agency collecting the habitat data.
Variable	The name of the PHAB metric (see below)
Result	The numeric value of the PHAB metric
Count_Calc	The number of unique observations that were used to calculate the value in Result

Values in the **Variable** column of the PHAB data indicate which PHAB metric was measured that corresponds to values in the **Result** column. The required PHAB metrics that should be provided for every unique sampling event specified by **StationCode** and **SampleDate** are in the next table.

Table 17 PHAB Metrics

Field Name	Description
XSLOPE	Mean slope of reach
XBKF_W	Mean bankfull width (m)
H_AqHab	Shannon diversity of aquatic habitat types
PCT_SAFN	Percent sand and fine (<2 mm) substrate particles
XCMG	Riparian cover sum of three layers
Ev_FlowHab	Evenness of flow habitat types
H_SubNat	Shannon diversity of natural substrate types
XC	Mean upper canopy trees and saplings
PCT_POOL	Percent pools in reach
XFC_ALG	Mean filamentous algae cover
PCT_RC	Percent concrete or asphalt

An example of properly formatted algae data is included in the package:

```
data(phab)
```

Each metric serves a specific purpose in the PHAB package: The H_AqHab, PCT_SAFN, XCMG, Ev_FlowHab, and H_SubNat metrics are used to assess habitat condition. The XSLOPE, XBKF_W, and PCT_RC metrics are used as predictors or score modifiers for different components of the IPI. Finally, the XC, PCT_POOL, XFC_ALG, and PCT_RC metrics provide information that is used for quality assurance checks for selected metrics and the overall IPI score.

All required fields for the stations and PHAB data are case-sensitive and must be spelled correctly. The order of the fields does not matter. All **StationCode** values must be shared between the datasets. The `IPI()` function automatically checks the format of the input data prior to estimating scores.

Detailed Metric Descriptions

Five of the required PHAB metrics in the input data are used directly for scoring the IPI, whereas the remainder serve a supporting role as predictors or modifiers for different parts of the complete index. Understanding what each of five core metrics describe about stream condition and how they vary with disturbance is critical for interpreting the index. Below is a detailed description of each metric.

Shannon diversity of natural instream cover (H_AqHab) measures the relative quantity and variety of natural structures in the stream, such as cobble, large and small boulders, fallen trees, logs and branches, and undercut banks available as refugia, or as sites for feeding or spawning and nursery functions of aquatic macrofauna. A wide variety and/or abundance of submerged structures in the stream provides macroinvertebrates and fish with a large number of niches, thus increasing habitat diversity. When variety and abundance of cover decreases (e.g., due to hydromodification, increased sedimentation, or active stream clearing), habitat structure becomes monotonous, diversity decreases, and the potential for recovery following disturbance decreases. Snags and submerged logs—especially old logs that have remained in-place for several years—are among the most productive habitat structure for macroinvertebrate colonization and fish refugia in low-gradient streams.

Percent sand and fine substrate (PCT_SAFN) measures the amount of small-grained sediment particles (i.e., < 2 mm) that have accumulated in the stream bottom as a result of deposition. Deposition may result from soil disturbance in the catchment, landslides, and bank erosion. Sediment deposition may cause the formation of islands or point bars, filling of runs and pools, and embeddedness of gravel, cobble, and boulders and snags, with larger substrate particles covered or sunken into the silt, sand, or mud of the stream bottom. As habitat provided by cobbles or woody debris becomes embedded, and as interstitial spaces become inundated by sand or silt, the surface area available to macroinvertebrates and fish is decreased. High levels of sediment deposition are symptoms of an unstable and continually changing environment that becomes unsuitable for many organisms. Although human activity may deplete sands and fines (e.g., by upstream dam operations), and this depletion may harm aquatic life, the IPI

treats only increases in this metric as a negative impact on habitat quality, although a post-hoc “concentrate correction” was made whereby the metric percent concrete (PCT_RC) is added to PCT_SAFN before scoring.

Shannon diversity of natural substrate types (H_SubNat) measures the diversity of natural substrate types, assessing how well multiple size classes (e.g., gravel, cobble and boulder particles) are represented. In a stream with high habitat quality for benthic macroinvertebrates, layers of cobble and gravel provide diversity of niche space. Occasional patches of fine sediment, root mats and bedrock also provide important habitat for burrowers or clingers, but do not dominate the streambed. Lack of substrate diversity, e.g., where >75% of the channel bottom is dominated by one particle size or hard-pan, or with highly compacted particles with no interstitial space, represents poor physical conditions. Riffles and runs with a diversity of particle sizes often provide the most stable habitat in many small, high-gradient streams.

Evenness of flow habitat types (Ev_FlowHab) measures the evenness of riffles, pools, and other flow microhabitat types. Optimal physical conditions include a relatively even mix of velocity/depth regimes, with regular alternation between riffles (fast-shallow), runs (fast-deep), glides (slow-shallow) and pools (slow-deep). Poor conditions occur when a single microhabitat dominates (usually glides, with pools and riffles absent). A stream that has a uniform flow regime will typically support far fewer types of organisms than a stream that has a variety of alternating flow regimes. Riffles in particular are a source of high-quality habitat and diverse fauna, and their regular occurrence along the length of a stream greatly enhances the diversity of the stream community. Pools are essential for many fish and amphibians.

Riparian vegetation cover, sum of three layers (XCMG) measures the amount of vegetative protection afforded to the stream bank and the near-stream portion of the riparian zone. The root systems of plants growing on stream banks help hold soil in place, thereby reducing the amount of erosion likely to occur. The vegetative zone also serves as a buffer to pollutants entering a stream from runoff and provides shading and habitat and nutrient input into the stream. Banks that have full, multi-layered, natural plant growth are better for fish and macroinvertebrates than are banks without vegetative protection or those shored up with concrete or riprap. Vegetative removal and reduced riparian zones occur when roads, parking lots, fields, lawns, bare soil, riprap, or buildings are near the stream bank. Residential developments, urban centers, golf courses, and high grazing pressure from livestock are the common causes of anthropogenic degradation of the riparian zone. Even in undeveloped areas, upstream hydromodification and invasion by non-native species can reduce the cover and quality of riparian zone vegetation.

Calculating the IPI

The IPI score for a site is estimated from the station and PHAB data. The score is estimated automatically by the `IPI()` function in the package following several steps.

1. Reference expectations for a site are estimated for predictive metrics using the station data
2. Observed data values are compared to reference expectations for predictive metrics and the differences between observed and predicted (i.e., metric residuals) are used for scoring. For metrics that are not predicted, raw metric values are used for scoring. Metric scores are based on the upper and lower percentiles of either metric residuals or raw metric values observed at reference and high-activity sites
3. The metric scores are then summed and standardized (i.e., divided) by the mean sum of scores at reference sites to obtain the final IPI score

To calculate the IPI:

The `IPI()` function can be used on station and PHAB data that are correctly formatted as explained above in the [Getting PHAB Data](#) section. As previously mentioned, the package includes an example of properly formatted algae data:

```
data(phab)
```

To calculate the IPI, first load the PHAB library, your stations data and your PHAB data into the workspace. In this example, the stations data is in a csv file named “phab_stations.csv”, and the phab data is in a csv file named “phab_metrics.csv”:

```
#Load the library
library(PHAB)

#Load your stations data
Stations <- read.csv("phab_stations.csv", stringsAsFactors
= F)

#Load your phab data
Phab <- read.csv("phab_metrics.csv", stringsAsFactors = F)

#Create a sample ID
phab$PHAB_SampleID <- paste(phab$StationCode,
phab$SampleDate, sep="_")
```

The `IPI()` function will calculate scores from the stations and phab data and produce a report:

```
Report <- IPI(stations = stations, phab = phab)
write.csv(report, "report.csv")
```

Accessing and Interpreting the Scores

A single data frame of IPI scores estimated at each site on each unique sample date is returned. The output data are in wide format with one row for each sample date at a site.

Table 18 IPI Output Data

Field Name	Description
StationCode	Unique identifier of the site
SampleDate	Date of data collection
SampleAgencyCode	Unique identifier of the agency collecting the habitat data.
PHAB_SampleID	Unique identifier of the sampling event. Automatically generated by concatenating site and date, separated by a vertical bar:
IPI	IPI score
IPI_percentile	Percentile of the IPI score relative to scores at reference calibration sites.
Ev_FlowHab	Evenness of flow habitat types, from the phab data input
Ev_FlowHab_score	Scored Ev_FlowHab metric
H_AqHab	Shannon diversity of natural instream cover types, from the phab data input
H_AqHab_pred	Predicted H_AqHab metric value
H_AqHab_score	Scored H_AqHab metric
H_SubNat	Shannon diversity of natural substrate types, from the phab data input
H_SubNat_score	Scored H_SubNat metric
PCT_SAFN	Percent sand and fine substrate, from the phab data input
PCT_RC	Percent concrete or asphalt, from the phab data input (used to adjust the PCT_SAFN score).
PCT_SAFN_pred	Predicted PCT_SAFN metric value
PCT_SAFN_score	Scored PCT_SAFN metric

Field Name	Description
XCMG	Riparian cover as sum of three layers, from the phab data input
XCMG_pred	Predicted XCMG metric value
XCMG_score	Scored XCMG metric
IPI_qa	Quality assurance value for the IPI score, calculated as the minimum value of all QA metrics.
Ev_FlowHab_qa	Quality assurance metric for Ev_FlowHab, calculated as the percent of expected measurements present in the phab data input.
H_AqHab_qa	Quality assurance metric for H_AqHab, calculated as the percent of expected measurements present in the phab data input.
H_SubNat_qa	Quality assurance metric for H_SubNat, calculated as the percent of expected measurements present in the phab data input.
PCT_SAFN_qa	Quality assurance metric for PCT_SAFN, calculated as the percent of expected measurements present in the phab data input.
XCMG_qa	Quality assurance metric for XCMG, calculated as the percent of expected measurements present in the phab data input.

Metrics are included in the output as observed PHAB metrics, predicted metrics (where applicable), and scored metrics. Observed PHAB metrics are returned as-is from the input data. Some PHAB metrics include a predicted column that shows the modelled metric value based on the environmental setting at a site. Scored PHAB metrics are obtained following the description above.

The last five columns include quality assurance (QA) information for the IPI score and select metrics. QA values less than one indicate less quality assurance, usually resulting from metric values being calculated from fewer measurements from a sample than specified by field protocols. The QA value for the IPI is based on the lowest score among all metrics. These columns are included by default and can be removed from the output by using the `qa = FALSE` argument:

```
report <- IPI(stations = stations, phab = phab, qa = FALSE)
```

The IPI was calibrated during its development so that the mean score of reference sites is 1; IPI scores near 1 represent locations with conditions similar to reference sites.

Scores that approach 0 indicate great departure from reference condition and degradation of physical condition. Scores > 1 can be interpreted to indicate greater physical complexity than predicted for a site given its natural environmental setting. All metric scores are weighted equally to determine the overall IPI score. For observed and scored PHAB metrics, all are expected to decrease under degraded physical conditions, except PCT_SAFN which is expected to increase.

Supplemental Functions in the PHAB Package

The `PHAB` package includes a number of functions that can facilitate analyses or provide useful ancillary information.

Accessing Calibration Data

An additional data file is available within the PHAB package that shows calibration data for scoring the IPI metrics. This file is called `refcal` and includes observed and predicted scores at reference and high-activity (or “stressed”) sites for the five PHAB metrics. Metrics are scored based on deviation from the 5th and 95th percentile of scores at reference or calibration sites. The `refcal` dataset includes observations at these sites that were used to identify percentile cutoffs for estimating metric scores.

The data may be accessed with the `data()` function:

```
data(refcal)
```

Frequently Asked Questions (FAQ) for the IPI

Most problems result from errors in data formatting, or other errors in the input data. Most errors will prevent complete execution of the `IPI()` function. We have attempted to provide informative error messages to help guide corrections.

Why can't I get IPI scores for my data?

The `IPI()` function will evaluate both the stations and phab input datasets for correct format before estimating IPI scores. IPI scores will not be calculated if any errors are encountered. The following checks are automatically made:

- No duplicate station codes in stations. That is, input data have one row per station
- All station codes in stations are in phab, and vice-versa
- All required fields are present in stations and phab (see above)
- All required PHAB metrics are present in the variable field of phab for each station and sample date (see above). An exception is made for XC, PCT_POOL, and XFC_ALG, which are not necessary for the IPI but are used for optional quality assurance checks

- No duplicate results for PHAB variables at each station and sample date. That is, one row per station, date, and metric
- All input variables for stations and phab are non-negative, excluding elevation variables in stations which may be negative if below sea level (i.e., some locations in southeast California)
- Moreover, the variables XBKF_W and Ev_FlowHab in phab must also be greater than zero

The `IPI()` function will print informative messages to the R console if any of these errors are encountered. It is the responsibility of the analyst to correct any errors in the raw data before proceeding.

Natural (background) Specific Conductivity

Olson and Hawkins (2012) developed a quantile random forest model that predicts ranges of expected levels of specific conductivity under natural conditions. The median predicted conductivity is used as a predictor in the ASCI, and the model is included in that package. However, it may be useful to access the model to predict other values than the median. For example, predictions of the upper range of natural values can help identify sites where disturbance has altered water chemistry (see section on [screening reference sites](#) below).

Preparing input data

Stations data

Stations data includes all the environmental information for each station, with one row per station. A subset of the fields required for ASCI calculation are required for this model.

Table 19 Stations Input Data

Field Name	Description
StationCode	Unique identifier of the site
KFCT_AVE	Average soil erodibility factor
LPREM_mean	Catchment mean log geometric mean hydraulic conductivity
AtmCa	Atmospheric deposition of Calcium
CaO_Mean	Average calcium oxide (quicklime) in the catchment geology
MgO_Mean	Average magnesium oxide (magnesia) in the catchment geology
S_Mean	Catchment mean whole rock S
UCS_Mean	Catchment mean unconfined Compressive Strength
AtmMg	Catchment mean of mean 1994-2006 annual ppt-weighted mean Mg concentration
AtmSO4	Catchment mean of mean 1994-2006 annual ppt-weighted mean SO4 concentration
MINP_WS	Catchment mean of mean 1971-2000 min monthly ppt
MEANP_WS	Catchment mean of mean 1971-2000 annual ppt
SumAve_P	Mean June to September 1971 to 2000 monthly precipitation, averaged across the entire catchment.
TMAX_WS	Catchment mean of mean 1971-2000 max temperature
MAXWD_WS	Catchment mean of 1961-1990 annual max number of wet-days
LST32AVE	Catchment mean of mean 1961-1990 first and last day of freeze.
BDH_AVE	Average bulk soil density
PRMH_AVE	Catchment mean soil permeability
XWD_WS	Catchment mean of mean 1961-1990 annual number of wet days

Field names must match spelling shown above. For the required fields, blank cells or missing values are not allowed. Other fields of interest may be included in the stations data. Columns may appear in any order. Although we have implemented scripts to make the inputs case-insensitive, we recommend conforming to the capitalizations shown above.

An example of properly formatted stations data is included in the [ASCII](#) package:

```
data(demo_station)
```

Calculating background conductivity

The random forest model can be accessed after loading the `ASCI` library:

```
library(ASCI)
cond.qrf <- rfmods$cond.qrf
```

This type of model is known as a quantile random forest model. Not only can it predict the most likely value, it can also predict any distribution point (i.e., any quantile), so that a prediction interval of any size may be obtained.

The `predict()` function is used to estimate these values. Due to a quirk in this function, missing values cannot be tolerated, even in fields that aren't used by model; therefore, fields with missing data need to be excluded from the demo data:

```
demo_station.2 <- demo_station %>%
  select(-CondQR50, -XerMtn)
```

This data frame can then be used with the model:

```
predict(cond.qrf,
        newdata = demo_station.2,
        what=c(0.90))
```

The `newdata` argument is used to specify the data frame containing the predictor data. The `what` argument specifies which quantiles should be predicted (in this case, the 90th percentile). Multiple quantiles can be included in this argument:

```
predict(cond.qrf,
        newdata = demo_station.2,
        what=c(0.10, 0.5, 0.90))
```

Section 4: Calculating Reference Screening Criteria and other Site Characteristics

Criteria were developed by Ode et al. (2016) to identify least-disturbed reference sites (Stoddard et al. 2006) in California. Most of the variables used in reference site screening are GIS measures that quantify human disturbance in the upstream watershed; others are measured in the field as part of physical habitat assessment. The goal of this section is to describe how to add the Watershed Metric Toolbox to ArcMap and how to use each tool within the Reference Screening Toolset. This section builds on the steps described in [Section 1](#) and assumes the user has completed delineating catchments. The GIS-based reference screening variables are in the following table (data sources for GIS variables can be found in Ode et al. 2016).

Table 20 GIS-based Reference Screening Variables

Screening variable	Description
StationCode	Uniquely identifying code for the sample location
New_Lat	Latitude, in decimal degrees
New_Long	Longitude, in decimal degrees
AG*	Sum of % row crops and % pasture (NLCD codes 81 and 82).
UR*	Sum of % low, % medium and % high intensity urban land use (NLCD codes 22-24)
AGUR*	Sum of % agriculture and % urban
CD21*	Percent developed open space (NLCD code 21)
RDRRDEN*	Density (in km/km ²) of road classes 1, 2 and 3 (i.e., sum of highway, paved and improved surface road length) plus rail (all classes)
PVD_INT*	Number (i.e., count) of intersections between paved roads and NHD flow line network (paved bridges)
NRST_DAM	Nearest upstream dam in km (value of -9999 indicates no dam in catchment)
CNL_PI_PCT	Percent of total NHD flow line length in the upstream watershed as canal or pipeline
MINES	Total count of producer mines in 5-km catchment clip

Some variables (denoted by asterisks in the table above) are calculated at three spatial scales: within 1 km upstream of the sampling site (1k), within 5 km upstream of the sampling site (5k), and within the entire watershed upstream of the sampling site (WS). In addition, the Reference Screening Toolset allows calculation of NLCD Land Cover data from different years (or all available years): 2000, 2006, 2011 and 2016. To represent each of the different scale and year combinations, the output field name is structured as follows:

[Variable + _ + Scale + _ + Year]

For example, the field name for percent AG calculated at the 5k clip scale with 2006 NLCD Land Cover data would be “AG_5k_06”.

In addition to the above screening variables, the reference screening tool also generates numerous additional variables that can be used in further analysis.

Table 21 Additional GIS-Based Reference Screening Tool Variables

Metric	Variable	Description
Basics	STATE	State name of the site.
Basics	COUNTY	County name of the site.
Basics	CALWNUM	CalWater Number of the site.
Basics	RB RN_NAME	Region Board number and name of the site.
Basics	COMID	COMID of NHDPlus flowline of the site.
Basics	RESOLUTION	NHDPlus flowline resolution
Ecoregions	LVL3CODE87 LVL3NAME87	Level III Ecoregion code and name 1987
Ecoregions	LVL3CODE10 LVL3NAME10	Level III Ecoregion code and name 2010
Ecoregions	LVL4CODE10 LVL4NAME10	Level IV Ecoregion code and name 2010
Ecoregions	LVL3CODE11 LVL3NAME11	Level III Ecoregion code and name 2011
Ecoregions	LVL4CODE11 LVL4NAME11	Level IV Ecoregion code and name 2011

Metric	Variable	Description
FCODE	ARTFCL_LEN ARTFCL_PCT	Artificial Path length in kilometers and percent of total.
FCODE	CANAL_LEN CANAL_PCT	Canal length in kilometers and percent of total.
FCODE	CNNCTR_LEN CNNCTR_PCT	Connector length in kilometers and percent of total.
FCODE	PIPELN_LEN PIPELN_PCT	Pipeline length in kilometers and percent of total.
FCODE	STREAM_LEN STREAM_PCT	Stream River length in kilometers and percent of total.
Geology	PCT_CENOZ	Percent cenozoic sediments (neogene sedimentary rocks, paleogene sedimentary rocks, and quarternary deposits)
Geology	PCT_SEDIM	Percent sedimentary geology
Geology	PCT_VOLCNC	Percent volcanic geology
Geology	PCT_QUART	Percent quarternary deposits
Geology	PCT_NOSED	Percent non-sedimentary (Gniess, Granitic, Intermediate, and Mafic_UltraMafic)
MRDS Producer Mines	GRVL_MINES	Count of gravel producing mines.
MRDS Producer Mines	MINE_DENS	Mines per square kilometer
MRDS Producer Mines	GRVL_DENS	Gravel miles per square kilometer
NID Dams	DAM_COUNT	Count of dams
NID Dams	NRML_STRG	Normal Storage capacity of all dams in Acre Feet

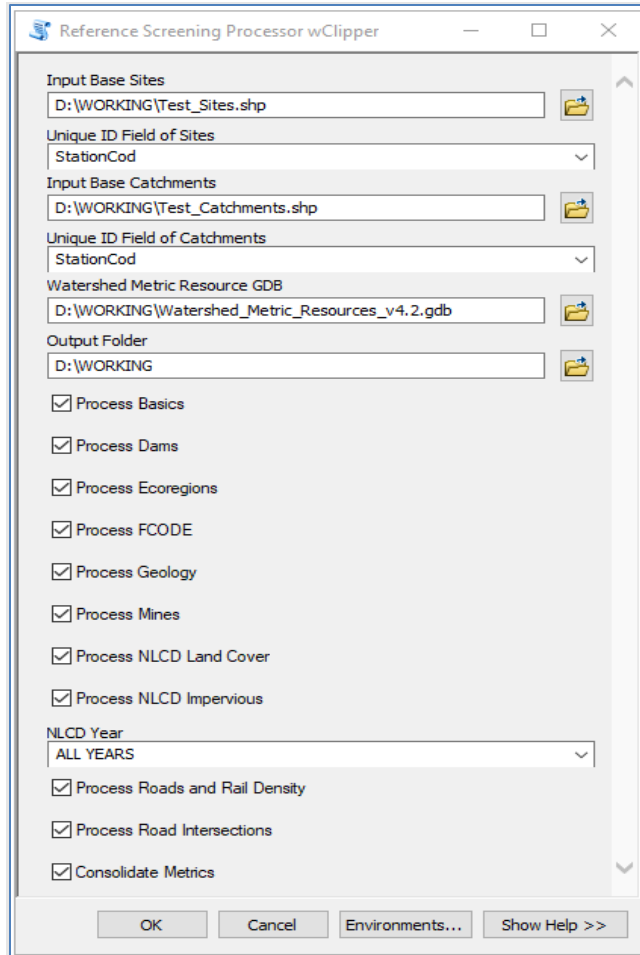
Metric	Variable	Description
NLCD Land Cover	FOREST_A FOREST	Area of forest in square meters (NLCD codes 41, 42, 43) and the percent of total.
NLCD Land Cover	WETLANDS_A WETLANDS	Area of wetlands in square meters (NLCD codes 90, 95) and the percent of total.
NLCD Land Cover	SHRUB_A SHRUB	Area of shrublands in square meters (NLCD code 52) and the percent of total.
NLCD Land Cover	NGRASSLA_A NGRASSLAND	Area of natural grasslands in square meters (NLCD code 71) and the percent of total.
NLCD Land Cover	NAT_BARR_A NAT_BARREN	Area of natural barren in square meters (NLCD code 31) and their percent of total.
NLCD Land Cover	URBAN_A	Area of urban in square meters (NLCD codes 22, 23, 24).
NLCD Land Cover	ROW_CROP_A ROW_CROPS	Area of row crops in square meters (NLCD code 82) and the percent of total.
NLCD Land Cover	PASTURE_A PASTURE	Area of pasture in square meters (NLCD code 81) and the percent of total.
NLCD Land Cover	WATER_A WATER	Area of water or ice in square meters (NLCD codes 11, 12) and the percent of total.
Road & Rail	RDLEN RDDENS	Length of all roads in kilometers and density of roads in km/sq km.
Road & Rail	RDLENC1 RDDENSC1	Length of class 1 roads in kilometers and density of roads in km/sq km.
Road & Rail	RDLENC2 RDDENSC2	Length of class 2 roads in kilometers and density of roads in km/sq km.
Road & Rail	RDLENC3 RDDENSC3	Length of class 3 roads in kilometers and density of roads in km/sq km.

Metric	Variable	Description
Road & Rail	RDLENC4 RDDENSC4	Length of class 4 roads in kilometers and density of roads in km/sq km.
Road & Rail	RDLENC5 RDDENSC5	Length of class 5 roads in kilometers and density of roads in km/sq km.
Road & Rail	RRLLEN RRDENS	Length of all railroads in kilometers and density of railroad in km/sq km.
Road & Rail	RRLENC0 RRDENSC0	Length of class 0 railroads in kilometers and density of railroad in km/sq km.
Road & Rail	RRLENC1 RRDENSC1	Length of class 1 railroads in kilometers and density of railroad in km/sq km.
Road & Rail	RRLENC2 RRDENSC2	Length of class 2 railroads in kilometers and density of railroad in km/sq km.
Road & Rail	RRLENC3 RRDENSC3	Length of class 3 railroads in kilometers and density of railroad in km/sq km.

Reference Screening Processor

The following describes how to use the Reference Screening Processor tool in ArcGIS Desktop. This tool is used to calculate all reference screening metrics listed in the table above, as well as other geospatial information about the site (e.g., ecoregion). This Python Tool is designed for use with ArcGIS 10.5 and above and requires the Spatial Analyst Extension to run.

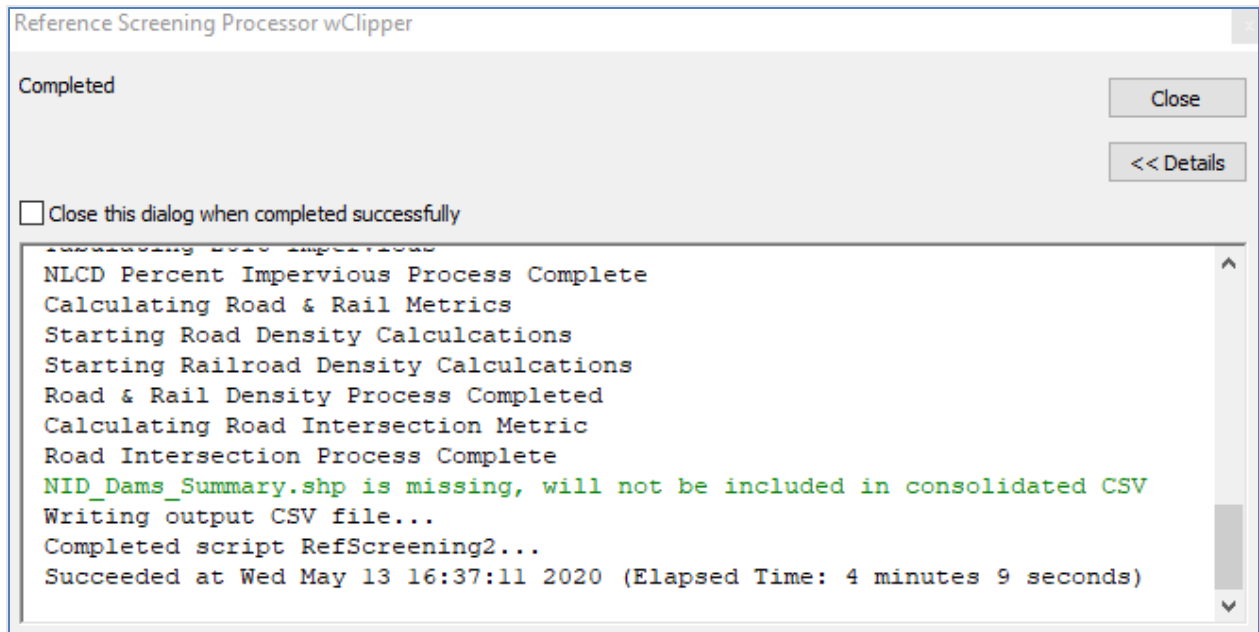
1. Add “Watershed Metric Toolbox v*.tbx” to your ArcToolbox
 - a. Right-click empty space within ArcToolbox window
 - b. Select “Add Toolbox”
 - c. Browse to “Watershed Metric Toolbox.tbx” on your computer and click “Open”
2. Navigate to the “Watershed Metric Toolbox” within ArcToolbox and double-click the “Reference Screening Processor” script to open its dialog box



3. Add each of the inputs as described below. See example:
 - a. **Input Base Sites:** Navigate to and add the site points shapefile
 - b. **Unique ID Field of Sites:** Choose the field that contains the unique ID for each input site point. All delineated catchments and corresponding site points should use the same unique ID value. In this example both “Test_Sites.shp” and “Test_Catchments_PRJ.shp” both contain the field “StationCod”
 - c. **Input Catchments:** Navigate to and add the catchments polygons shapefile. Reminder that the “StationCod” field must correspond with the “Input Base Sites” for the tool to run properly
 - d. **Unique ID Field of Catchments:** Choose the field that contains the unique ID for each input catchment polygon. All delineated catchments and corresponding site points should use the same unique ID value. In this example both “Test_Sites.shp” and “Test_Catchments_PRJ.shp” both contain the field “StationCod”

- e. **Watershed Metric Resource GDB:** Navigate to and add the “Watershed_Metric_Resources.gdb” geodatabase containing all necessary raster files for metric calculations
 - f. **Output Folder:** Choose the location you wish the final results’ shapefiles to be saved. It is recommended that you create a new output folder within your working directory to store all of your CSCI metric output files. Intermediate files will also be saved here during processing but will be deleted upon completion
 - g. **Process Basics:** Checkbox to calculate basic site information such as State, County, CalWater Number, Region Board and COMID
 - h. **Process Dams:** Checkbox to calculate Nearest Dam, Dam County, and Normal Storage metrics
 - i. **Process Ecoregions:** Checkbox to calculate ecoregion of sites at various levels
 - j. **Process FCODE:** Checkbox to calculate length of each NHD flowline feature code within catchments
 - k. **Process Geology:** Checkbox to calculate the geologic composition of catchments
 - l. **Process Mines:** Checkbox to calculate number of mines present in catchments
 - m. **Process NLCD Land Cover:** Checkbox to generate metrics on land cover classifications in catchments
 - n. **Process NLCD Impervious:** Checkbox to calculate the percent impervious surface in catchments
 - o. **NLCD Year:** Dropdown to NLCD data year for the calculations. Options available are 2001, 2006, 2011, 2016, or ALL YEARS
 - p. **Process Roads and Rail Density:** Checkbox to calculate density of roads and railroads in catchments
 - q. **Process Road Intersections:** Checkbox to calculate the number for intersections between paved roads and the NHD flowline (bridges or other crossings)
 - r. **Consolidate Outputs:** Check box to consolidate all reference screening metrics into a csv output after processing. Only the screening variables listed landscape scale screens sections below are consolidated
4. Click “OK” and the tool will run. When it completes you should see shapefile outputs for each metric selected

- a. EcoRegions.shp
 - b. FCODE_Summary.shp
 - c. Geology.shp
 - d. MRDS_Producer_Mines_Summary.shp
 - e. NID_Dams_Summary.shp
 - f. NLCD_Land_Cover_2001.shp (1k, 5k, and WS outputs)
 - g. NLCD_Land_Cover_2006.shp (1k, 5k, and WS outputs)
 - h. NLCD_Land_Cover_2011.shp (1k, 5k, and WS outputs)
 - i. NLCD_Land_Cover_2016.shp (1k, 5k, and WS outputs)
 - j. NLCD_Impervious_2001.shp
 - k. NLCD_Impervious_2006.shp
 - l. NLCD_Impervious_2011.shp
 - m. NLCD_Impervious_2016.shp
 - n. Paved_Roads_NHD_Intersect_No_Pipes.shp
 - o. Road_Rail_Density.shp (1k, 5k, and WS outputs)
 - p. Site_Basics.shp
5. The tool checkboxes allow users to omit certain metrics from being processed. This may be useful if the user is only interested in a particular metric, or if you need to reprocess a subset of data. If you omit a metric that is included with the RefScreening_Metric_Consolidated.csv output file, you will see a warning like the one below. In the example, the Dams metrics were not checked; as a result, the NRST_DAM screening variable cannot be added to the consolidation output. You will need to rerun the tool with Dams checked on if you wish to perform reference screening.



Applying reference screens

Once these variables have been calculated, they may be used to determine if a catchment meets reference criteria established in Ode et al. (2016a). These criteria are largely based on land use, using the variables described in the section above. In addition, local (reach-scale) variables measured in the field should also be used where available. Thus, the landscape-level screens described in this document are only part of the process in identifying reference sites.

Landscape-scale screens

Variables calculated by the reference screening processor should be compared to these thresholds; sites that have lower levels of development at all relevant spatial scales may be reference sites.

Table 22 Landscape-scaled Screening Variable Thresholds

Screening variable	Scale	Threshold	Unit
AG	1k, 5k, WS	3	%
UR	1k, 5k, WS	3	%
AGUR	1k, 5k, WS	5	%
CD21	1k, 5k	7	%
	WS	10	%
RDRRDEN	1k, 5k, WS	2	km/km ²
PVD_INT	1k	5	crossings
	5k	10	crossings
	WS	50	crossings
NRST_DAM	WS	10	km
CNL_PI	WS	10	%
MINES	5k	0	mines

Field-measured screens

Where available, field-measured reference screens should be used in tandem with GIS-measured screens. In Ode et al. (2016a), two field-measured screens were applied.

Table 23 Field-measured Screening Variable Thresholds

Screening variable	Threshold	Unit
W1_HALL	<1.5	-
Specific conductivity	> 99th percentile of expected background levels	uS/cm

Riparian disturbance index (W1_HALL)

The first field-measured screen is based on a proximity-weighted riparian disturbance index (“W1_HALL”) calculated from measures taken by field crews at the time of sampling. This metric was adopted from a suite of physical habitat metrics developed by the EPA (Kaufmann et al. 1999). The proximity of different types of human disturbance (roads, pipes, orchards and row crops, etc.) to the stream channel is recorded at each of 11 transects along the sampling reach; evidence of disturbance closer to (or in) the

channel is given higher weight than disturbances occurring far from the channel, and higher values of W1_HALL indicate greater levels of disturbance. Calculating W1_HALL is automated by reporting tools in the SWAMP data warehouse and the SMC data portal, but not CEDEN. However, it may also be calculated by hand as follows:

At each transect, separate observations are made on each bank, for a total of 22 observation plots per reach. Disturbances are recorded as being either on the bank or within the channel ("B"), present within the 10-m² riparian observation plots ("C"), or between 10 m and 50 m from the channel margin ("P"). Closer proximities take priority over farther proximities, e.g., if garbage is observed on the bank and in the 10-m² riparian observation plots, only the presence on the bank is recorded. Disturbances located in the channel are recorded as being on both banks.

To calculate W1_HALL, proximities are converted to weighted values: B = 1.5, C = 1, P = 0.67. Weights are summed across the sampling reach and divided by the total number of observations (n = 22).

Sites are considered to meet reference criteria if W1_HALL < 1.5.

The riparian disturbance criterion should always be applied whenever data are available. If these data are lacking, sites meeting landscape-scale reference screens should be considered provisionally reference pending further information about local conditions.

Note that other evidence of local disturbance, including site photos, may be sufficient to exclude a site from reference status, even if other reference screens are satisfied.

Specific conductivity

Specific conductivity is a measure of the ionic concentration of solutes in stream water. Elevated conductivity is associated with many human activities and may be considered an indicator of disturbance. High levels of conductivity may be toxic to certain aquatic organisms. Specific conductivity may also be high because of natural factors, such as catchment geology or climate. Natural (background) levels of specific conductivity may be estimated with a model developed by Olson and Hawkins (2012). Sites that exceed the upper range of likely levels of specific conductivity predicted by this model may be disturbed, and therefore non-reference. Thus, the model offers an additional way to screen reference sites.

In Ode et al. (2016a), sites where field-measured value that exceeded the 99th percentile of values predicted by the model were rejected from the reference pool; however, because the predicted conductivity model was observed to underpredict at higher levels of specific conductance, a threshold of 2000 $\mu\text{S}/\text{cm}$ is used as an upper bound if the prediction interval includes 1000 $\mu\text{S}/\text{cm}$.

This criterion is optional because some streams have naturally high specific conductivity due to geological factors. If a site exceeds this criterion yet meets all other reference criteria, further investigation for signs of disturbance before rejecting the site from reference status is advised. Note that the ASCIs (Theroux et al. 2020) did not apply this criterion during index calibration.

Cited literature

Fetscher, A.E., R. Stancheva, P. Kociolek, R.G. Sheath, E.D. Stein, R.D. Mazon, P.R. Ode, and L. Busse. 2014. Development and comparison of stream indices of biotic integrity using diatoms vs. non-diatom algae vs. a combination. *Journal of Applied Phycology* 26: 433-450.

Kaufmann, P. R., P. Levine, E. G. Robinson, C. Seeliger & D. V. Peck. 1999. Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. Research Ecology Branch, US Environmental Agency, Corvallis, Oregon.

Mazon, R. D., P. R. Ode, A. C. Rehn, M. Engeln, K. A. Schiff, E. Stein, D. Gillett, D. Herbst, and C.P. Hawkins. 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. *Freshwater Science* 35(1): 249-271.

Ode, P.R., A.C. Rehn, and J.T. May. 2005. A quantitative tool for assessing the integrity of southern coastal California streams. *Environmental Management* 35: 493-504. DOI 10.1007/s00267-004-0035-8.

Ode, P.R., A.C. Rehn, R.D. Mazon, K.C. Schiff, E.D. Stein, J.T. May, L.R. Brown, D.B. Herbst, D. Gillett, K. Lunde and C.P. Hawkins. 2016a. Evaluating the adequacy of a reference site pool for the ecological assessment of streams in environmentally complex regions. *Freshwater Science* 35: 237-248.

Ode, P.R., A.E. Fetscher, and L.B. Busse. 2016b. [Standard Operating Procedures \(SOP\) for the Collection of Field Data for Bioassessments of California Wadeable Streams: Benthic Macroinvertebrates, Algae, and Physical Habitat](#). California State Water Resources Control Board, Surface Water Ambient Monitoring Program (SWAMP) Bioassessment SOP 004.

Olson, J.R. and C.P. Hawkins. 2012. Predicting natural base-flow stream water chemistry in the western United States. *Water Resources Research* 48: W02504.

Rehn, A.C., J.T. May, and P.R. Ode. 2008. [An Index of Biotic Integrity \(IBI\) for Perennial Streams in California's Central Valley. SWAMP Technical Report](#). Surface Water Ambient Monitoring Program.

Rehn, A.C., R.D. Mazon and P.R. Ode. 2015. The California Stream Condition Index (CSCI): A New Statewide Biological Scoring Tool for Assessing the Health of Freshwater Streams. *Swamp Technical Memorandum SWAMP-TM-2015-0002*.

Rehn, A.C., R.D. Mazon and P.R. Ode. 2017. An index to measure the quality of physical habitat in California wadeable streams. *Swamp Technical Memorandum SWAMP-TM-2018-0005*.

Theroux, S., R.D. Mazon, M.W. Beck, P.R. Ode, E.D. Stein, and M. Sutula. 2020. [Predictive biological indices for algae populations in diverse stream environments, Ecological Indicators.](#)

Appendix 1: Overview of Automated steps for Calculating CSCI

The steps have already been automated as follows:

For O/E calculation

1. Aggregate taxa to operational taxonomic units (OTUs).
2. Exclude ambiguous taxa (e.g., taxa identified to relatively poor taxonomic resolution).
3. For samples with more than 400 remaining specimens, subsample to 400 specimens (20 iterations).*
4. Use stations data to predict group membership and calculate OTU capture probabilities.
5. Calculate O/E score for each iteration, using a minimum capture probability of 0.5.

For MMI calculation

1. Aggregate taxa to SAFIT Level 1.
2. For samples with more than 500 remaining specimens, subsample to 500 specimens (20 iterations). *
3. Calculate biological metrics.
4. Use stations data to predict metric values.
5. Calculate difference between observed and predicted metric values. Score the difference, calculate the average across metrics, and standardize by dividing by the mean from reference calibration sites (i.e., 0.628).

*Note that there are two distinct subsampling steps (i.e., for the O/E and for the MMI), and each are triggered by different criteria. The number of iterations for each subsampling step is provided in the reports.

For CSCI calculation

1. Calculate the average O/E and MMI scores, as described above.
2. Compare the CSCI, O/E, and MMI scores to the distribution of scores at reference calibration sites.

Many steps typically required of index calculation are hardwired into the scripts and are automatically handled. Specifically, FinalIDs are aggregated to the necessary taxonomic resolution, and large samples are subsampled to the required size. *We strongly*

discourage all efforts to manually aggregate or subsample your own data, and instead recommend you rely on the standardized, automated approach implemented by the provided scripts.