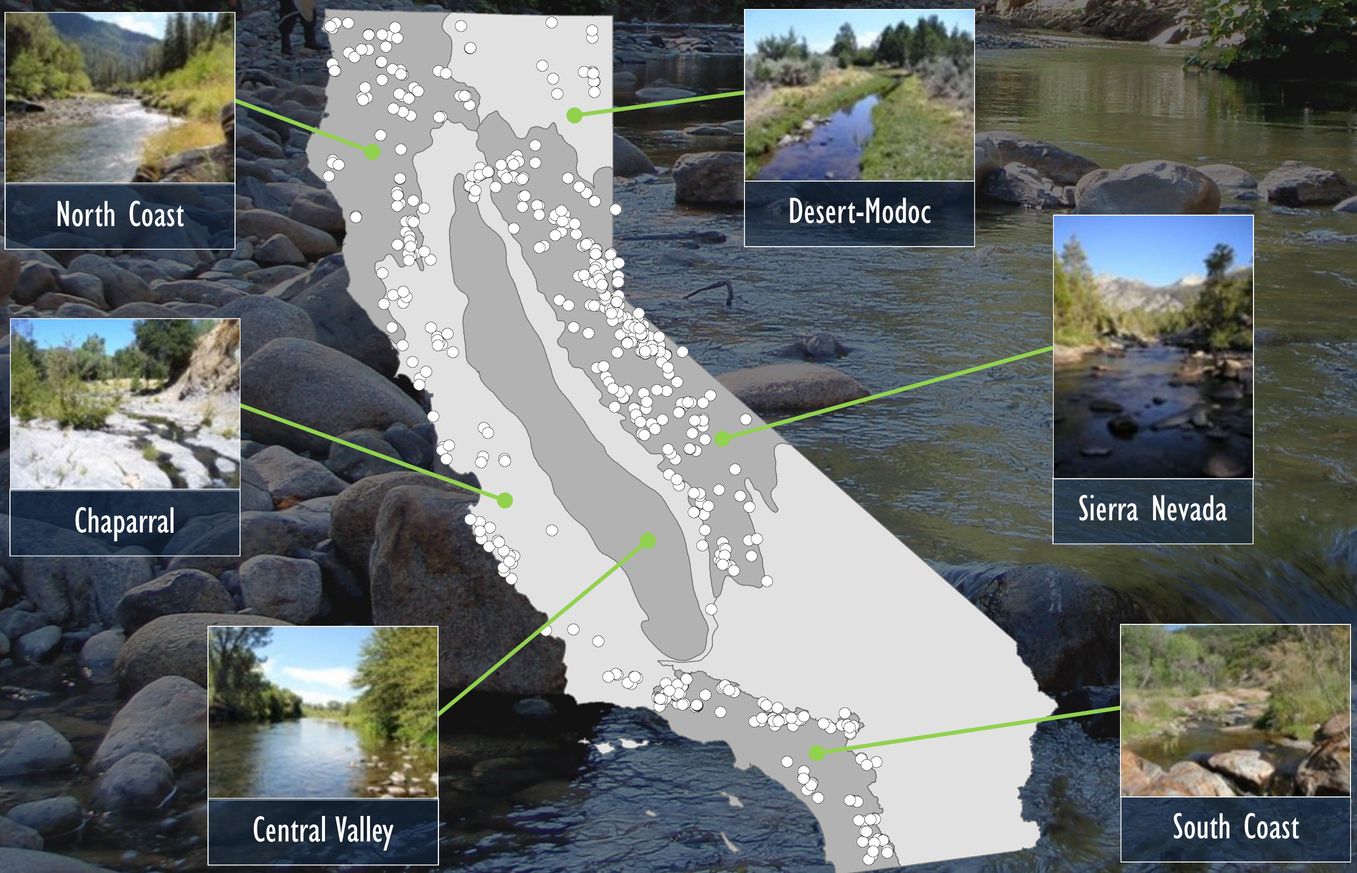
The California Stream Condition Index (CSCI)

December 2015

WHAT IS THE CALIFORNIA STREAM CONDITION INDEX?

The California Stream Condition Index (CSCI) is a biological scoring tool that helps aquatic resource managers translate complex data about benthic macroinvertebrates found living in a stream into an overall measure of stream health. The CSCI score indicates whether, and to what degree, the ecology of a stream is altered from a healthy state. Direct measures of ecosystem health like the CSCI are preferable to those based on chemical or physical measurements for many management questions. Living organisms integrate the effects of multiple stressors, such as sedimentation, nutrient enrichment and riparian disturbance, over both space and time.

Benthic macroinvertebrates are small but visible invertebrates, such as insect larvae, that live on stream bottoms. Flathead mayfly larva pictured left.



STATEWIDE REFERENCE SITES

Reference sites where human disturbance is absent or minimal are used to set benchmark expectations for healthy streams. A large set of nearly 600 reference sites (**see map**), representing the broad diversity of natural stream types found across California, was used to develop the CSCI.

CSCI vs. IBIs

Indices of biotic integrity (IBIs) were previously available for some regions of California. The CSCI is an advancement over previous indices because it is applicable statewide, accounts for a much wider range of natural variability, and provides equivalent scoring thresholds in all regions of the state. Additionally, the CSCI provides multiple lines of evidence, incorporating measures of species composition and ecological traits into a single condition score.

HOW IS THE CSCI SCORE CALCULATED?

The CSCI score is a measure of how well a site's **observed** condition matches its predicted, or **expected**, condition. Expected values of a set of **ecological indicators** are predicted using statistical models. Predictions are based on natural environmental variables resulting in a site-specific prediction for each site; greater deviations from this expectation indicate a greater likelihood of degradation. The CSCI score is calculated by comparing the expected condition with actual (observed) results. CSCI scores range from 0 (highly degraded) to greater than 1 (equivalent to reference).

EXPECTED

LOCATION

CLIMATE

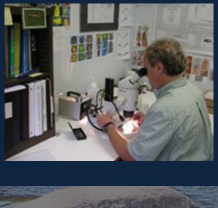
GEOLOGY

WATERSHED SIZE

Natural environmental variables are used to predict the biological composition (species and their ecological traits) at a site if it's healthy.

OBSERVED

The site is sampled and species are identified in the lab.



$$\frac{\text{Observed Species and Traits}}{\text{Expected Species and Traits}} = \text{CSCI Score}$$

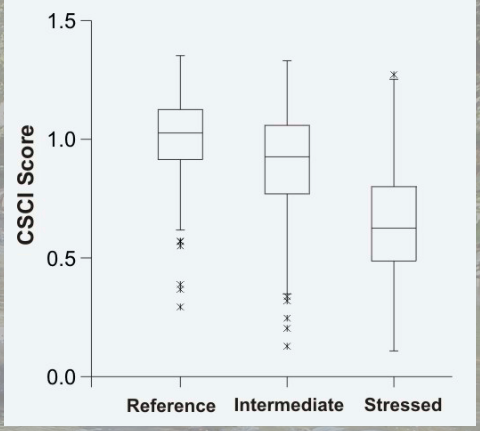
CSCI Components

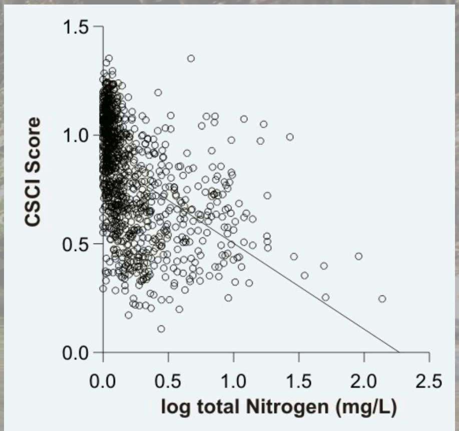
Taxonomic Completeness	Species
Measures of ecological traits (structure and function)	# Species
	# Shredders
	% Clingers
	% Coleoptera
	% EPT *
	% Intolerant

***EPT = Ephemeroptera + Plecoptera + Trichoptera**

HOW CAN THE CSCI BE USED?

The CSCI can be used to assess the status and trends of stream condition at multiple scales (sites, watersheds, regions, and statewide) and is also well-suited for compliance monitoring, evaluating the success of mitigation and restoration projects, and evaluating the success of stream protection policies and programs. It is an integral component of the state's bio-integrity plan. The CSCI is useful for measuring biological integrity in wadeable perennial and non-perennial streams throughout California. The limits of the CSCI's applicability in streams that are dry for more than 6 months each year are currently being researched by SWAMP and several regional partners.





The CSCI is responsive to human disturbance and discriminates well between reference sites and "stressed" sites, that is, sites with high levels of overall human activity in the watershed. The CSCI also responds well to individual stressor gradients such as total nitrogen, a nutrient closely associated with eutrophication in streams and rivers.

THE CSCI IS CURRENTLY BEING USED TO:

- **Assess regional and statewide stream condition**
- **Identify healthy streams and prioritize them for protection**
- **Identify impaired streams and prioritize them for restoration**
- **Evaluate effectiveness of stormwater best-management practices**
- **Assess the impacts of timber harvest activities**

For additional information on applications of the CSCI and guidance on how to calculate it, contact Lori Webber (Lori.Webber@waterboards.ca.gov) or Calvin Yang (Calvin.Yang@waterboards.ca.gov).

More information can also be found at the SWAMP Bioassessment Program website.

SWAMP Surface Water Ambient Monitoring Program

CALIFORNIA DEPARTMENT OF FISH & WILDLIFE

Aquatic Bioassessment Laboratory

SCCWRP Established 1969

The California Stream Condition Index (CSCI): A New Statewide Biological Scoring Tool for Assessing the Health of Freshwater Streams

Prepared by: Andrew C. Rehn[1], Raphael D. Mazor[2,1], Peter R. Ode[1]

[1]California Department of Fish and Wildlife
[2]Southern California Coastal Water Research Project

SWAMP Technical Memorandum
SWAMP-TM-2015-0002
September 2015

TABLE OF CONTENTS

OBJECTIVE

The objective of this technical memo is to summarize the development, features and use of SWAMP's next-generation index for monitoring stream health in California.

OVERVIEW

California's dramatic environmental diversity supports a broad array of natural stream types throughout the state. Bioassessment of freshwater stream and rivers is especially challenging in such a region because the reference condition, or the benchmark of biological condition expected when human disturbance in the environment is absent or minimal, varies greatly among natural stream types. Previous indices used by monitoring programs were developed on a regional basis to help partition the state's environmental diversity, but statewide assessments were confounded by different criteria used in different regions. The CSCI, which translates complex data about individual benthic macroinvertebrates (BMIs) found living in a stream into an overall measure of stream health, was developed specifically to address some of the shortcomings of earlier indices. First, the CSCI was developed with a much larger, more representative data set that makes it applicable statewide and that covers the broad range of environmental variability among natural stream types. Second, the CSCI sets biological benchmarks for a site based on its site-specific environmental setting. Finally, the CSCI combines two separate types of indices, each of which provides unique information about the biological condition of a stream: a multi-metric index (MMI) that measures ecological structure and function, and an observed-to-expected (O/E) index that measures taxonomic completeness. Together they provide multiple lines of evidence about the condition of a stream, providing greater confidence in results than a single index.

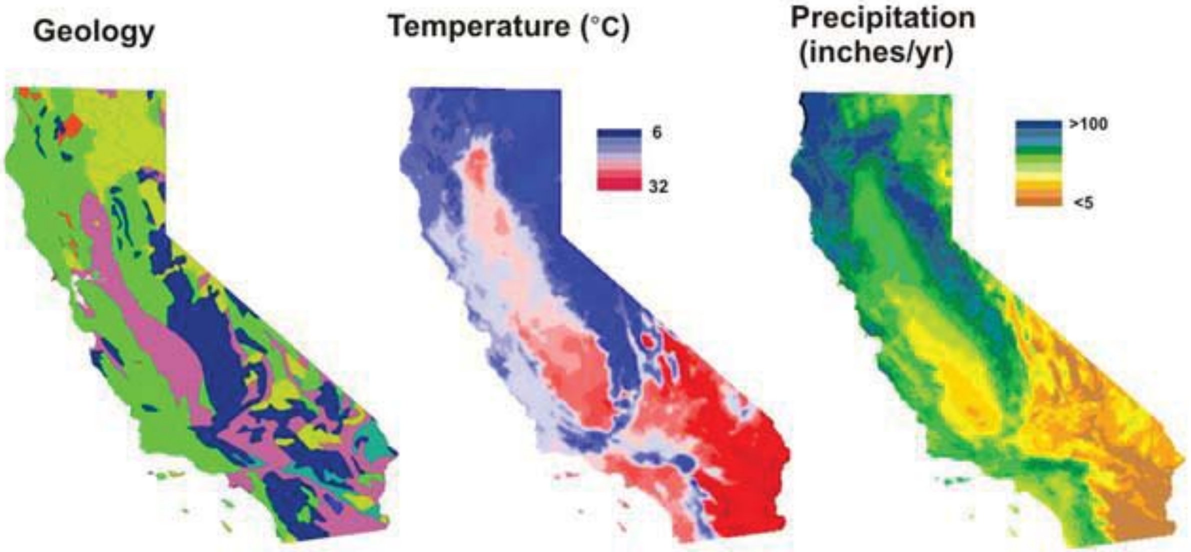


Figure 1. Extreme natural gradients in California result in a high degree of natural variation in biological expectations among stream types.

Introduction

California contains continental-scale environmental diversity within its borders, encompassing some of the most extreme gradients in elevation, climate and geology found in the United States (Figure 1). It supports temperate rainforests in the North Coast, alpine forests and meadows in the mountains, deserts in the east, and chaparral, oak woodlands, and grasslands with a Mediterranean climate in most remaining parts of the state. Such great physiographic complexity correspondingly supports a broad array of natural stream types, which in turn hosts a rich diversity of aquatic organisms. Bioassessment, which is the science of using aquatic organisms as indicators of stream health and function, is greatly complicated in such regions because the reference condition varies greatly among natural stream types (Figure 2).

Previous indices used by stream monitoring programs to rank or "score" biological condition at sampling sites relative to reference conditions were developed for specific subregions of California as a means of partitioning the state's environmental variability (e.g., Ode et al. 2005, Rehn 2009). While this approach allowed the establishment of defensible impairment thresholds within regions, comparison among regions was confounded for two closely related reasons: 1) the criteria used to define reference conditions varied among regions, and 2) each index was composed of different metrics so that deviation from the reference benchmark was not equivalently measured in all settings and did not have the same ecological meaning across the entire state.

Moreover, some portions of the state and certain stream types were unrepresented. To support the ongoing development of California's statewide Biological Integrity Plan, the State Water Board funded the development of a new index that was applicable statewide, encompassed as much natural environmental variability as possible, and allowed consistent and equivalent scoring thresholds in all regions of the state.

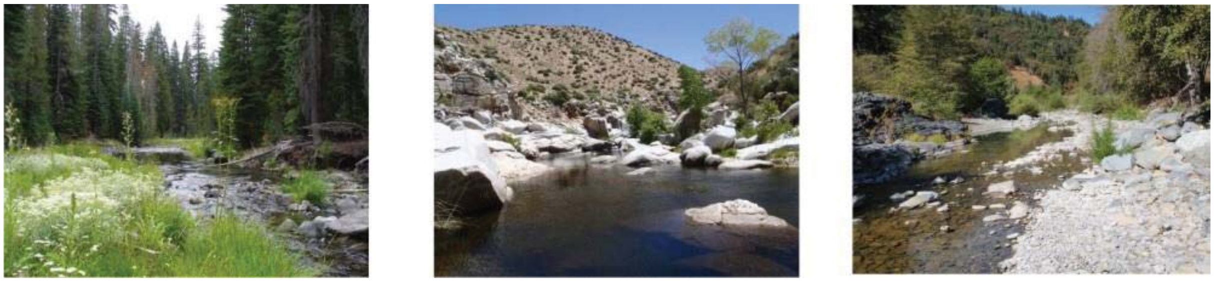


Figure 2. Bioassessment is complicated in regions with natural environmental complexity because the reference condition varies greatly among natural stream types.

The California Stream Condition Index (CSCI) is a new statewide biological scoring tool that translates complex data about benthic macroinvertebrates (BMIs) found living in a stream into an overall measure of stream health. Finalized in 2013 and recently accepted for publication (Mazor et al. in press), the CSCI represents the latest generation of biological indicators for assessing stream health in California. The CSCI combines two separate types of indices, each of which provides unique information about the biological condition at a stream: a multi-metric index (MMI) that measures ecological structure and function, and an observed-to-expected (O/E) index that measures taxonomic completeness. Unlike previous MMI or O/E indices that were applicable only on a regional basis or under-represented large portions of the state, the CSCI was built with a statewide dataset that represents the broad range of environmental conditions across California. The CSCI provides consistency and accuracy in the interpretation of biological data collected by both statewide and regional monitoring programs and will be the basis of the new statewide Biological Integrity Plan. This memo summarizes the development and key features of the CSCI, including its performance characteristics, recommended scoring thresholds, and data requirements for its use. Full details of CSCI development can be found in Mazor et al. (in press).

Compilation of Data Sets

Benthic datasets for CSCI development were compiled from more than 20 federal, state, and regional monitoring programs that sampled streams sites between 1999 and 2011. Standardization of BMI data was necessary because the level of taxonomic effort used to identify organisms and the number of specimens identified per sample varied among programs. Somewhat different data standardization approaches were used for the MMI and the O/E, but to accommodate data reduction that occurs during standardization, 600-count BMI samples identified to "Level 2a" as defined by the Southwest Association of Freshwater Invertebrate Taxonomists (SAFIT, Richards and Rogers 2011) were required[1]. BMI samples with insufficient numbers of organisms or taxonomic resolution were excluded from analyses. A final data set from 1,985 sites met all requirements and was used for development and evaluation of both the O/E and MMI indices (Figure 3).

[1] SAFIT Level 2a identifies most taxa to species and Chironomidae to subfamily.

Quantifying Natural and Anthropogenic Gradients Across Sites

Environmental data were gathered from multiple sources to characterize natural and anthropogenic factors known to affect benthic communities such as climate, elevation, geology, land cover, road density, hydrologic alteration, and mining. GIS variables that characterized natural and relatively stable environmental factors (e.g., topography, geology, climate) were used as predictors for O/E and MMI models, whereas variables related to human activity (e.g., land use, road density, etc.) were used to classify sites as reference and to evaluate responsiveness of O/E and MMI indices to human activity gradients. Most variables related to human activity were calculated at three spatial scales: within the entire upstream drainage area (watershed), within the contributing area 5 km upstream of a site, and within the contributing area 1 km upstream of a site (Appendix 1). Quantifying human activity at multiple spatial scales allowed sites to be screened for both local and catchment-scale impacts. By contrast, variables used as predictors for O/E and MMI indices were calculated at either the site (i.e., "point") scale or the watershed scale, but not at the local (1k and 5k) scales (Appendix 2).

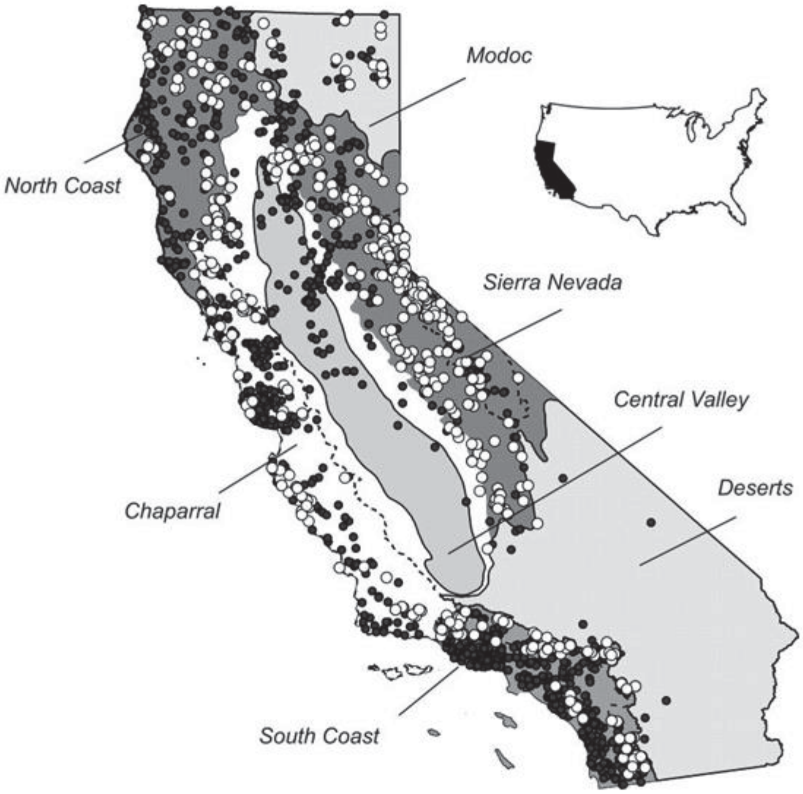


Figure 3. Distribution of 1,985 sampling sites used in development and validation of the CSCI. White circles are sites that passed reference screens (n = 590; see text) and black circles are sites that failed one or more screening criteria. Major ecological regions are those used as reporting units for the Perennial Streams Assessment (PSA).

Sites were divided into three sets for development and evaluation of indices: reference (i.e., low-activity), moderate-activity, and high-activity sites. Uniform statewide criteria for defining reference sites were recently established by Ode et al. (in press; also see Appendix 1) with an emphasis on achieving a balance between thorough environmental representativeness while still maintaining a pool of "minimally-disturbed" sites *sensu*

Stoddard et al. (2006). Nearly 600 of the 1,985 sites included in the data set for CSCI development passed reference screening criteria (Figure 3), a fairly high success rate due to an emphasis being placed on data sets likely to contain high-quality reference sites during data compilation. In addition to good geographic coverage, the final reference pool also represented several biologically important natural gradients (Figure 4). Identification of high-activity sites was necessary for MMI calibration (described below) and for performance evaluation of both MMI and O/E. High-activity sites were defined as meeting any of the following criteria: ≥50% developed land (i.e., % agricultural + % urban) at any spatial scale; ≥ 5 km/km^2 road density at any spatial scale; or riparian disturbance index (W1_HALL of Kaufmann et al. 1999) ≥ 5. Sites not identified as either reference or high-activity were designated as moderate-activity.

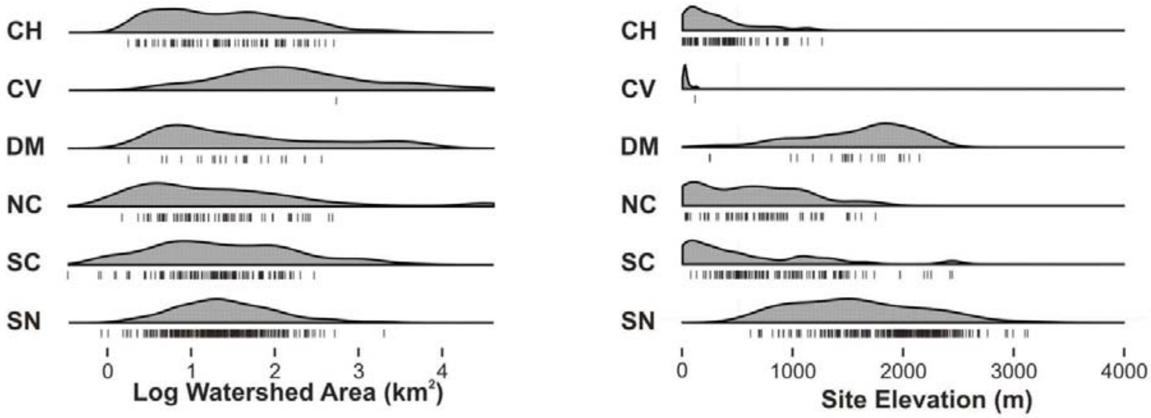


Figure 4. Examples of natural gradients that are important drivers of biological variability and are well-represented by the reference site pool. Unbiased estimates of natural gradient distributions in California were derived from probabilistic surveys conducted between 2000 and 2011 and are shown as kernel density estimates. Values of individual reference sites are shown as small vertical lines. Regions (see Figure 2) are abbreviated as follows: SN = Sierra Nevada, SC = South Coast, NC = North Coast, DM = Deserts + Modoc, CV = Central Valley, CH = Chaparral.

Building Predictive Models for O/E and MMI

The CSCI combines two different types of indices that have traditionally been used separately in stream assessments and provide unique information about the biological condition of a stream; an observed-to-expected (O/E) index that measures taxonomic completeness, and a multi-metric index (MMI) that measures ecological structure and function. Predictive modeling has been used in the development of O/E indices since their inception (Moss et al. 1987), but its use in the development of MMIs is relatively new (e.g., Pont et al. 2009). In each case, modeling improves index performance, but the process through which modeling helps achieve better performance differs somewhat between the approaches.

O/E indices assess the taxonomic completeness of a site by comparing observed and expected taxa lists. The taxa expected at a new assessment site, or a "test" site, are predicted by statistical modeling of relationships between taxonomic composition and natural environmental gradients at reference sites. Biological condition at a test site is then measured as the number of expected taxa (E) that are actually observed (O), and degradation of biological condition is quantified as loss of expected native taxa. Modeling relationships between taxonomic composition and natural environmental gradients produces indices that are more precise compared to null models where all taxa are assumed to have an equal probability of occurrence at all sites. In addition, the statistical modeling process in development of an O/E index produces site-specific expectations for each assessment site.

A multi-metric index aggregates several measures of BMI attributes, or metrics (e.g., percent predators, number of pollution tolerant species, etc.), into a single measure of biological condition. Metrics include measures of assemblage richness, composition, and diversity, and are chosen based on their responsiveness to human disturbance gradients and/or their ability to discriminate between reference and degraded condition. The challenge is that expected metric values at reference sites vary greatly among natural stream types, and natural gradients often co-vary with human disturbance gradients, thereby confounding metric response to disturbance. Previous MMIs developed for use in various subregions of California utilized regionalization approaches to control for the effects of natural variation in biological expectations, where "one size fits all" expectations were developed within large, mostly geographically defined areas (e.g., chaparral vs. mountains). Regionalization approaches have often been shown to poorly account for natural variation among sites (Hawkins et al. 2010). Therefore, models were developed to predict expected metric values at reference sites based on multiple natural environmental gradients (Appendix 2). Metric residuals (the difference between observed and expected values) were used as new metric values instead of raw metrics because they measure the range of metric variation after removing the effect of natural environmental variables. The models developed for reference sites were then used to predict expected metric values and calculate residuals at moderate- and high-activity sites. The advantages of this approach are twofold: 1) the expected metric values for any given assessment site are site-specific; and 2) metric residuals provide a more accurate evaluation of metric response to human disturbance gradients because they model out the effects of variation across natural environmental gradients. Use of modeled metrics produced an MMI with much better performance characteristics than un-modeled (null) metrics.

O/E indices do not require scoring because, as a simple ratio of observed-to-expected taxa, they are already scaled so that the mean score at reference sites is 1. Scoring is required for MMIs because individual metrics have different scales and different responses to stress, i.e., as human activity increases, some metrics decrease while others increase (Blocksom 2003). Scoring transforms metrics to a standard scale ranging from 0 (i.e., most stressed) to 1 (i.e., similar to reference sites). After scoring, final metrics[2] were chosen based on their ability to discriminate between reference and high-activity sites and by lack of bias among PSA regions (Figure 3). Scores for the final MMI at each site were then calculated by averaging the scores of the final selected metrics and rescaling (dividing) by the mean of reference calibration sites. Rescaling of MMI scores ensures that MMI and O/E are expressed in similar scales (i.e., as a ratio of observed to reference expectations), improving comparability of the two indices. A combined index (the California Stream Condition Index, CSCI) was calculated by averaging final MMI and O/E scores.

Setting Scoring Thresholds for the CSCI

The CSCI was calibrated during its development so that the mean score of reference sites is 1. Scores that approach 0 indicate great departure from reference condition and degradation of biological condition.

[2] Six metrics representing different aspects of assemblage composition (richness, trophic structure, tolerance, etc.) were chosen for inclusion in the final MMI: Taxonomic Richness, Shredder Taxa Richness, Percent Clinger Taxa, Percent Coleoptera Taxa, Percent EPT Taxa, and Percent Intolerant Individuals.

Scores > 1 can be interpreted to indicate greater taxonomic richness and more complex ecological function than predicted for a site given its natural environmental setting. In practice, CSCI scores observed from nearly 2000 study reaches sampled across California range from about 0.1 to 1.4. For the purposes of making statewide assessments, three thresholds were established based on the 30th; 10th; and 1st percentiles of CSCI scores at reference sites[3]. These three thresholds divide the CSCI scoring range into 4 categories of biological condition as follows: ≥0.92 = likely intact condition; 0.91 to 0.80 = possibly altered condition; 0.79 to 0.63 = likely altered condition; ≤0.62 = very likely altered condition.

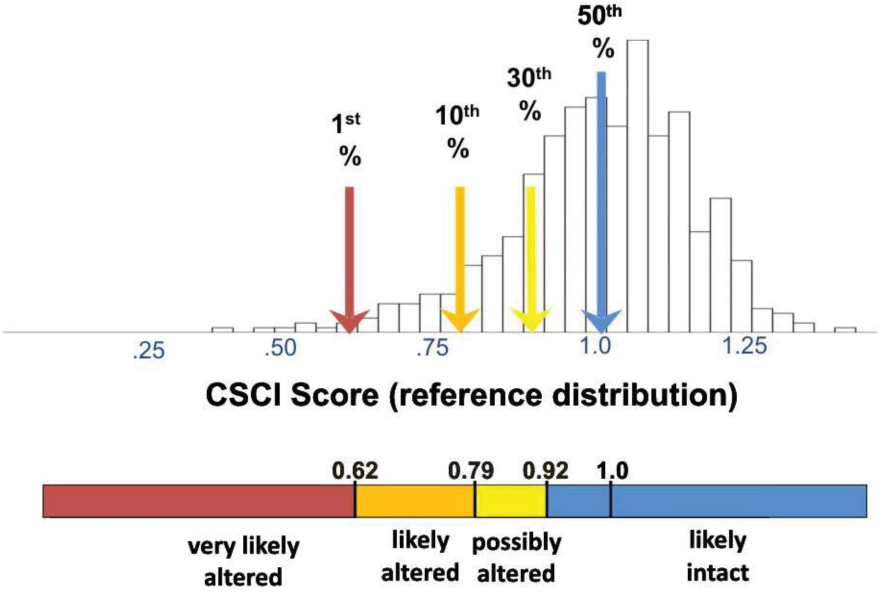


Figure 5. Distribution of CSCI scores at reference sites with thresholds and condition categories.

CSCI Performance

The CSCI had better performance than its null (un-modeled) counterpart in terms of accuracy and bias, precision, responsiveness, and sensitivity (see Appendix 3 for definitions of performance criteria). For example, mean regional differences in null CSCI scores at reference sites were large and significant, but were mostly absent in predictive CSCI scores (Figure 6a). The CSCI also was strongly responsive to human disturbance gradients, and the response was not confounded by the effects of natural gradients because those effects were modeled out by the use of metric residuals (Figure 6b).

[3] The rationale for these thresholds was to balance Type 1 errors (inferring degradation when it does not exist) and Type II errors (inferring a site is in reference condition when it is degraded). Similar thresholds have a precedent in bioassessment literature, but other methods for setting thresholds are possible, and if applied, might be equally valid.

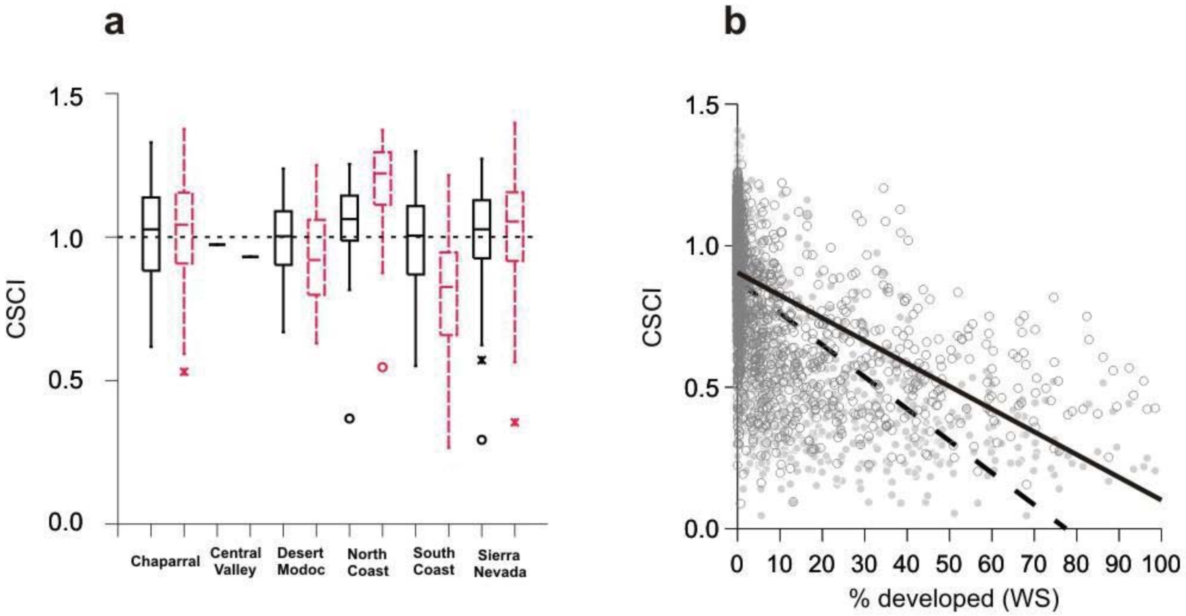


Figure 6. Examples of CSCI performance: a) Distribution of scores for predictive (black boxes) and null models (red dashed boxes) for the CSCI by geographic region. The horizontal dashed line indicates the expected value at reference sites (i.e., 1). Boxes represent the median, first, and third quartiles. Whiskers represent 1.5 times the interquartile range. Circles and X's represent outliers.
b) Relationship between predictive CSCI score (open circles and solid line) and null CSCI score (gray symbols and dashed line) and percent development in the watershed (= % urban + % ag). Note that the null CSCI has a steeper slope than the predictive CSCI because un-modeled metrics partially respond to natural gradients. By contrast, the predictive CSCI provides a more accurate response to disturbance gradients because the effects of metric variation across natural gradients have been modeled out.

Calculating the CSCI

Two types of data are required to calculate the CSCI: biological data and environmental data. Biological data are generated from benthic macroinvertebrate samples collected in accordance with standard SWAMP protocols (Ode 2007) and identified to the required taxonomic level of effort. SWAMP currently recommends a 600-count sample identified by a qualified taxonomist to *at least* SAFIT level 2a (Richards and Rogers 2015), with most taxa identified to species and Chironomidae identified to subfamily. Environmental data (e.g., watershed area, geology, precipitation) are generated by a specialist following standard geographic information system (GIS) protocols. Interim instructions (Mazor et al. 2015) that describe all steps in calculating the CSCI can be found at the SWAMP Bioassessment Program website. The first section describes the process for using GIS to delineate catchment polygons and then calculate environmental predictors (see Appendix 2 for required predictors). The second section describes the process for using the environmental predictors in conjunction with taxonomic data to calculate CSCI scores using custom libraries and scripts in the R statistical programming language. SWAMP is currently developing online tools to generate CSCI scores from user-supplied biological data and site coordinates, requiring minimal technical expertise.

More information about the SWAMP Bioassessment Program can be found at: http://www.waterboards.ca.gov/water_issues/programs/swamp/bioassessment. Those wishing to arrange training in CSCI calculation should contact Calvin Yang: calvin.yang@waterboards.ca.gov

REFERENCES

Blocksom, K. A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic highland streams. Environmental Management 42: 954-965.

Hawkins, C. P., Y. Cao, and B. Roper. 2010b. Methods of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology. 55: 1066-1085.

Kaufmann, P. R., P. Levine, E. G. Robinson, C. Seeliger, and D. V. Peck. 1999. Surface waters: Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. US EPA. Office of Research and Development. Washington, DC.

Mazor, R.D. 2015. Bioassessment of streams in southern California: A report on the first 5 years of the Stormwater Monitoring Coalition's regional stream survey. Southern California Coastal Water Research Project Technical Report 844.

R. D. Mazor, P. R. Ode, A. C. Rehn, M. Engeln, T. Boyle, E. Fintel, and S. Verbrugge. 2015. The California Stream Condition Index (CSCI): Guidance for calculating scores using GIS and R. SCCWRP Technical Report #883. SWAMP-SOP-2015-0004.

Mazor, R.D., A.C. Rehn, P.R. Ode, M. Engeln, K.C. Schiff, E.D. Stein, D. Gillett and C.P. Hawkins. In press. Bioassessment in complex environments: designing an index for consistent meaning in different settings. Freshwater Science.

Moss, D., T. Furse, J. F. Wright, and P. D. Armitage. 1987. The prediction of macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology 17: 41-52.

Ode, P.R. 2007 Standard operating procedures for collecting benthic macroinvertebrate samples and associated chemical and physical data for ambient bioassessments in California. California State Water Resources Control Board Surface Water Ambient Monitoring Program (SWAMP) Bioassessment SOP 001.

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.

Ode, P.R., T.M. Kincaid, T. Fleming and A.C. Rehn. 2011. Ecological Condition Assessments of California's Perennial Wadeable Streams: Highlights from the Surface Water Ambient Monitoring Program's Perennial Streams Assessment (PSA) (2000-2007). A collaboration between the State Water Resources Control Board's Non-Point Source Pollution Control Program (NPS Program), Surface Water Ambient Monitoring Program (SWAMP), California Department of Fish and Game Aquatic Bioassessment Laboratory, and the U.S. Environmental Protection Agency.

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. In press. Evaluating the adequacy of a reference site pool for the ecological assessment of streams in environmentally complex regions. Freshwater Science.

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. doi: 10.1029/2011WR011088

Pont, D., R. M. Hughes, T. R. Whittier, and S. Schmutz. 2009. A predictive index of biotic integrity model for aquatic-vertebrate assemblages of Western U.S. streams. Transactions of the American Fisheries Society. 138: 292-305.

Rehn, A. C. 2009. Benthic macroinvertebrates as indicators of biological condition below hydropower dams on West Slope Sierra Nevada streams, California, USA. River Research and Applications 25: 208-228.

Richards, A. B., and D. C. Rogers. 2011. List of freshwater macroinvertebrate taxa from California and adjacent states including standard taxonomic effort levels. Southwest Association of Freshwater Invertebrate Taxonomists. Chico, CA. Available from www.safit.org.

Stoddard, J. L., D. V. Peck, S. G. Paulsen, J. Van Sickle, C. P. Hawkins, A. T. Herlihy, R. M. Hughes, P.R. Kaufmann, D. P. Larsen, G. Lomnicky, A. R. Olsen, S. A. Peterson, P. L. Ringold, and T. R. Whittier. 2005. An Ecological Assessment of Western Streams and Rivers. EPA 620/R-05/005, U.S. Environmental Protection Agency, Washington, DC.

SUGGESTED CITATION

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.

APPENDIX 1. STRESSOR AND HUMAN ACTIVITY GRADIENTS USED TO IDENTIFY REFERENCE SITES AND EVALUATE INDEX PERFORMANCE.

Sites that did not exceed the listed thresholds were used as reference sites. WS: Watershed. 5 km: Watershed clipped to a 5-km buffer of the sample point. 1 km: Watershed clipped to a 1-km buffer of the sample point. Variables marked with an asterisk (*) indicate those used in the random forest evaluation of index responsiveness. W1_HALL: proximity-weighted human activity index (Kaufmann et al. 1999). Sources are as follows: A: National Landcover Data Set. B: Custom roads layer. C: National Hydrography Dataset Plus. D: National Inventory of Dams. E: Mineral Resource Data System. F: Predicted specific conductance (Olson and Hawkins 2012). G: Field-measured variables. Code 21 is a land use category that corresponds to managed vegetation, such as roadsides, lawns, cemeteries, and golf courses.

Variable		Scale	Threshold	Unit	Data source
*	% Agriculture	1 km, 5 km, WS	<3	%	A
*	% Urban	1 km, 5 km, WS	<3	%	A
*	% Ag + % Urban	1 km, 5 km, WS	<5	%	A
*	% Code 21	1 km and 5 km	<7	%	A
*		WS	<10	%	A
*	Road density	1 km, 5 km, WS	<2	km/km^2	B
*	Road crossings	1 km	<5	# crossings	B, C
*		5 km	<10	# crossings	B, C
*		WS	<50	# crossings	B, C
*	Dam distance	WS	<10	km	D
*	% Canals and pipelines	WS	<10	%	C
*	Instream gravel mines	5 km	<0.1	mines/km	C, E
*	Producer mines	5 km	0	mines	E
	Specific conductance	Site	99/1**	prediction interval	F
	W1_HALL	Reach	<1.5	NA	G
	% Sands and fines	Reach		%	G
	Slope	Reach		%	G

** The 99th and 1st percentiles of predictions were used to generate site-specific thresholds for specific conductance. Because the model was observed to under-predict at higher levels of specific conductance (data not shown), a threshold of 2000 µS/cm was used as an upper bound if the prediction interval included 1000 µS/cm.

APPENDIX 2. NATURAL GRADIENTS USED AS PREDICTORS FOR DEVELOPMENT OF O/E AND MMI INDICES.

Variable	Data Source
Site (i.e., "point")	
Latitude	
Longitude	
Elevation	A
Catchment Morphology	
Log watershed area	A
Elevation Range	A
Climate	
10-year (2000-2009) average precipitation at the sample point	B
10-year (2000-2009) average air temperature at the sample point	B
Mean June to September 1971-2000 monthly precipitation, averaged across the catchment	B
Geology	
Average bulk soil density	C
Average soil erodibility factor (k)	C
Log % phosphorus-bearing geology	C

Sources:

A. National Elevation Dataset (http://ned.usgs.gov/)

B. PRISM climate mapping system (http://www.prism.oregonstate.edu)

C: Generalized geology, mineralogy, and climate data derived for a conductivity prediction model (Olson and Hawkins 2012)

Predictors that were evaluated but not selected for any model include percent sedimentary geology, nitrogenous geology, soil hydraulic conductivity, soil permeability, sulfur-bearing geology, calcite-bearing geology, and magnesium oxide-bearing geology.

APPENDIX 3. SUMMARY OF PERFORMANCE EVALUATIONS FROM MAZOR ET AL. (IN PRESS)

Aspect	Description	Indication of good performance
Accuracy and Bias	Scores are minimally influenced by natural gradients	- Approximately 90% of validation reference sites have scores above the 10th percentile of calibration reference sites. - Landscape-scale natural gradients explain little variability in scores at reference sites, as indicated by a low pseudo-R^2 for a 500-tree random forest model. - No visual relationship evident in plots of scores at reference sites against field measurements of natural gradients.
Precision	Scores are similar when measured under similar settings	- Low standard deviation of scores among reference sites (one sample per site) - Low pooled standard deviation of scores among samples at reference sites with multiple sampling events.
Responsiveness	Scores change in response to human activity gradients	- Large t-statistic in comparison of mean scores at reference and high-activity sites. - Landscape-scale human activity gradients explain variability in scores, as indicated by a high pseudo-R^2 for a 500-tree random forest model.
Sensitivity	Scores indicate poor condition at high-activity sites	- High percentage of high-activity sites have scores below the 10th percentile of calibration reference sites.

The California Stream Condition Index (CSCI):

Interim instructions for calculating scores using GIS and R

Raphael Mazor[1,2] (raphaelm@scccwrp.org), Peter R. Ode[2],
Andrew C. Rehn[2], Mark Engeln[1], Tyler Boyle[3], Erik Fintel[3],
Calvin Yang[4] (calvin.yang@waterboards.ca.gov)

[1]Southern California Coastal Water Research Project. Costa Mesa, CA

[2]California Department of Fish and Wildlife. Rancho Cordova, CA

[3]Geographical Information Center, California State University. Chico, CA

[4]State Water Resources Control Board. Sacramento, CA

SWAMP-SOP-2015-0004

Revision Date: April 14, 2017

Table of Contents

Introduction

This document describes steps in calculating the California Stream Condition Index (CSCI), a bioassessment index that measures stream health based on benthic macroinvertebrate data. The instructions provided herein are provided as interim support for analysts requiring CSCI scores. The State Water Resources Control Board is currently developing a more automated approach to score calculation. Until that time, this document describes the only way to obtain CSCI scores.

The first section in this document describes the process for using a geographic information system (GIS) to calculate environmental predictors, such as watershed area and rainfall. The second section describes the process for using the environmental predictors, as well as taxonomic data, to calculate CSCI scores in R. A third section provides advice on interpreting scores in unusual circumstances (such as samples with poor taxonomic resolution).

The development and interpretation of the index is described in Mazor et al.(2016), which may be cited as follows:

> 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. The Society for Freshwater Science 35(1): 249-271.

A shorter summary of the index and its properties is available as a SWAMP technical memo:

> 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.

If you wish to cite this document to describe CSCI calculation (as opposed to general index properties or development), use the following citation:

> R. D. Mazor, P. R. Ode, A. C. Rehn, M. Engeln, T. Boyle, E. Fintel, and C. Yang. 2017. The California Stream Condition Index (CSCI): Interim instructions for calculating scores using GIS and R. SWAMP-SOP-2015-0004.

Computer Software Requirements:

- ArcGIS 10.2.2 or higher
    - Spatial Analyst Extension (Extension for ArcGIS)
- NHDPlusV2 Basin Delineator V2.4.0.20
- 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

Section 1: Instructions for Calculating CSCI Predictors with a Geographic Information System

The goal of this section is to guide users through the steps needed to calculate the predictors required for the California Stream Condition Index (CSCI).

These predictors are described as follows:

Predictor	Description
New_Lat	Latitude, in decimal degrees North
New_Long	Longitude, in decimal degrees East
SITE_ELEV	Site elevation in meters
ELEV_RANGE	Difference in elevation between the sample point and highest point in the catchment, in meters.
AREA_SQKM	Watershed area in square kilometers
PPT_00_09	Average precipitation (2000 to 2009) at the sample point, in hundredths of millimeters
TEMP_00_09	Average temperature(2000 to 2009) at the sample point, in hundredths of degrees Celsius
SumAve_P	Mean June to September 1971 to 2000 monthly precipitation, averaged across the entire catchment.
BDH_AVE	Average bulk soil density
KFCT_AVE	Average soil erodibility factor
P_MEAN	Average Phosphorous geology

Although the State Water Board will develop web-based tools to automate the steps described in this document, some users may be interested in calculating the CSCI on their own. We cannot guarantee the accuracy of metrics calculated using this document.

Field names and records are case-sensitive.

DOWNLOADING DATA

The necessary raster data can be downloaded from SCCWRP's FTP site.
The CSCI toolbox (top link) and the geodatabase (bottom link) can be downloaded using here:

ftp://ftp.sccwrp.org/pub/download/TMP/RaphaelMazor/CSCI_Metrics_Toolbox_10_1.zip
ftp://ftp.sccwrp.org/pub/download/TMP/RaphaelMazor/CSCI_Metric_Resources_gdb.zip

This zip files contain a geodatabase, a python script for data consolidation and export, and documentation for each step in metric calculation. The documentation is redundant with the information in this SOP. Be sure to download CSCI Metrics Toolbox for ArcGIS 10.1 and Above.

CREATING THE BASEFILES

BaseFiles are shapefiles that function as the unit of spatial analysis for calculation of CSCI predictors and other spatial metrics. The CSCI predictors are calculated with two types of BaseFiles: The site (a point representing the sample location) and a catchment (a polygon representing the contributing landuse).

All BaseFiles must contain a unique identifier of each station, which we call "StationCod" (this field name gets automatically changed to "StationCode" when data are exported for analysis in R). StationCods must be represented in all shapefiles, using the same letter case, must not contain: periods, special characters, and spaces. Each StationCod must contain no more than 18 characters.

Creating the Site BaseFile

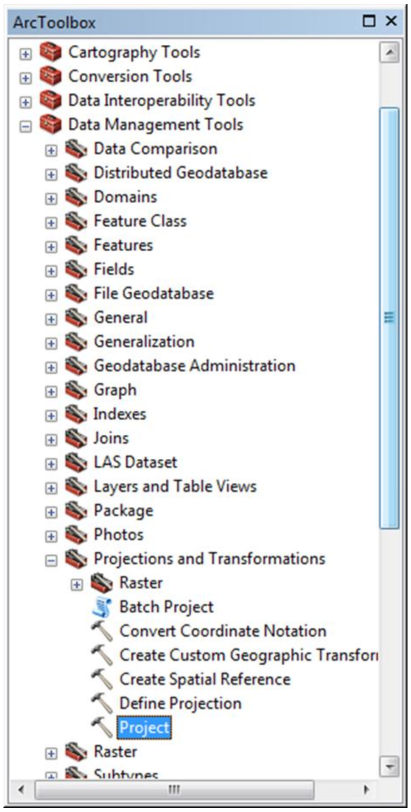
The goal of this step is to create a shapefile representing the location of sample points. Where possible, the location of sample points is adjusted ("snapped") from the actual coordinates to the nearest stream line represented in the National Hydrography Dataset Plus (NHD Plus). This "snapping" step is optional, but it is recommended because it improves the catchment delineation process, and also to help generate metrics for screening reference sites. If snapping is not desired, stop after Step 4, but be sure to give subsequent delineations, metrics, and other analytical products additional scrutiny.

Data requirements

-Spreadsheet (e.g., in .xls format) with unique site identifiers (field name: StationCode) and coordinates in decimal degrees (field names: LAT and LONG).

-NHDPlus V2 Data, including flowlines and subbasins (Hydrologic Units). Full Data Requirements can be found in the NHDPlusV2 Basin Delineator readme. Data can be downloaded from the NHDPlusV2 database here.

1. Load spreadsheet in ArcMap.
2. Right-click and display XY data. X field is the LONG, Y field is the LAT. Set the Coordinate system to WGS84.
    a. To get to WGS84:
        i. Expand Geographic Coordinate Systems
        ii. Expand World
        iii. Select WGS 1984
3. Reproject to NAD_1983_California_Teale_Albers.
    a. To reproject, open the ArcToolbox from the Geoprocessing menu or toolbar:

i. 

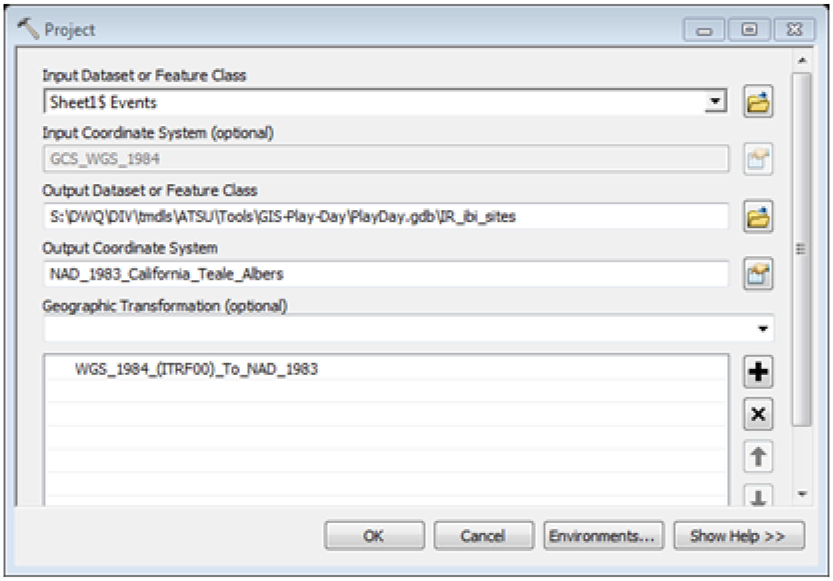
ArcToolbox

- Cartography Tools
- Conversion Tools
- Data Interoperability Tools
- Data Management Tools
  - Data Comparison
  - Distributed Geodatabase
  - Domains
  - Feature Class
  - Features
  - Fields
  - File Geodatabase
  - General
  - Generalization
  - Geodatabase Administration
  - Graph
  - Indexes
  - Joins
  - LAS Dataset
  - Layers and Table Views
  - Package
  - Photos
  - Projections and Transformations
    - Raster
    - Batch Project
    - Convert Coordinate Notation
    - Create Custom Geographic Transfor
    - Create Spatial Reference
    - Define Projection
    - Project
  - Raster
  - Subtypes

ii. Select point layer as "Input Dataset or Feature Class"

iii. Select Geodatabase you want to save to for "Output Dataset" field

iv. To fill out Output Coordinate System:

1. Expand Projected Coordinate Systems
2. Expand State Systems
3. Select NAD 1983 California (Teale) Albers (Meters)
4. 

Project

Input Dataset or Feature Class: Sheet1$ Events

Input Coordinate System (optional): GCS_WGS_1984

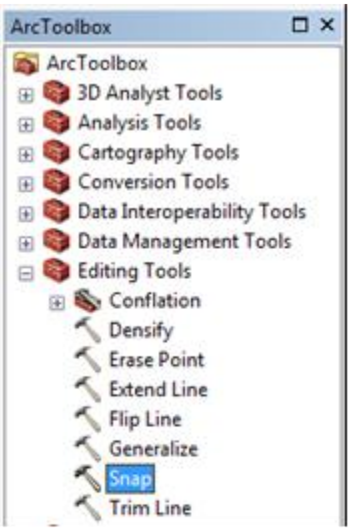
Output Dataset or Feature Class: S:\DWQ\DIV\tmdls\ATSU\Tools\GIS-Play-Day\PlayDay.gdb\IR_ibi_sites

Output Coordinate System: NAD_1983_California_Teale_Albers

Geographic Transformation (optional): WGS_1984_(ITRF00)_To_NAD_1983

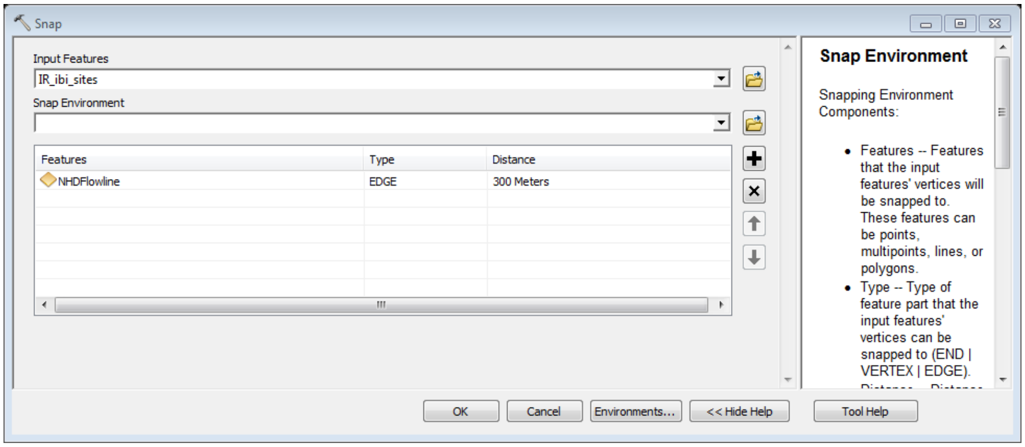
Snapping points to NHD flowlines:

4. Load the spatially corresponding NHD flowlines and the Subbasin from the USGS website (http://nhd.usgs.gov/data.html). Download MediumResolution geodatabase for CA from USGS FTP site. If all points fall within one NHD Region move on to the next steps. If not export them by region. They will need to be run separately through the delineator.
5. Snap the points to the nearest flowline in a manual edit session using "Edge Snapping". If there is no flowline near it, add a note in the attribute table. If the point is near a confluence or between two rivers look in the attribute table for clues to where it should go (see Quality Control section below).
6. To snap points using auto-snap tool from ArcToolbox:

a. 

i. To snap, open the ArcToolbox from the Geoprocessing menu or toolbar:

1. Expand Editing Tools to click on 'Snap'

b. 

Input Features: IR_ibi_sites

Snap Environment

Features	Type	Distance
NHDFlowline	EDGE	300 Meters

Snap Environment

Snapping Environment Components:

- Features -- Features that the input features' vertices will be snapped to. These features can be points, multipoints, lines, or polygons.
- Type -- Type of feature part that the input features' vertices can be snapped to (END | VERTEX | EDGE).
- Distance -- Distance

i. Select point layer as "Input Features"

ii. Select NHDFlowline as "Snap Environment"

1. Select EDGE as "Type"
2. Enter 300 Meters as "Distance"

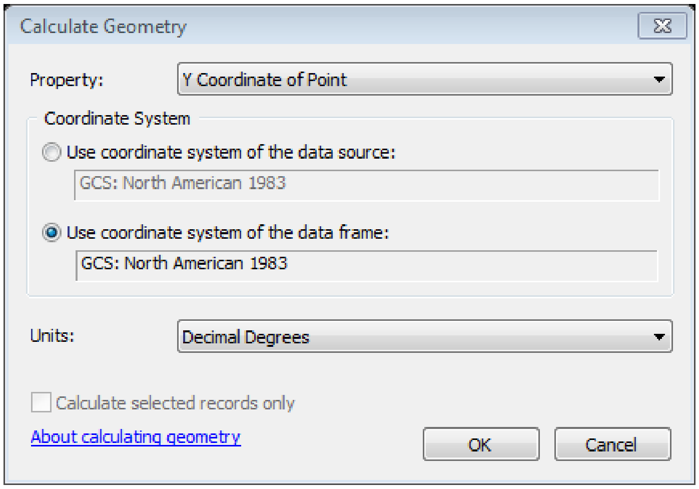
7. Once all sites are snapped, add a "New_Lat" and "New_Long" field. Calculate the latitude and longitude of the newly snapped points.

a. To Add new Fields:

i. Open Attribute Table

    ii. Select “Add Field…” under Table Options (top left icon in the table window)
    iii. Enter “New_Lat” for Name:
        1. Select “Double” under Type:
        2. Click OK
    iv. Repeat for “New_Long”

b. To Calculate New_Lat:
    i. Right-click on New_Lat field name
    ii. Select Calculate Geometry…and click Yes on warning about Calculating outside of an edit session
    iii. Fill out menu as follows and click ok:



Calculate Geometry

Property: Y Coordinate of Point

Coordinate System

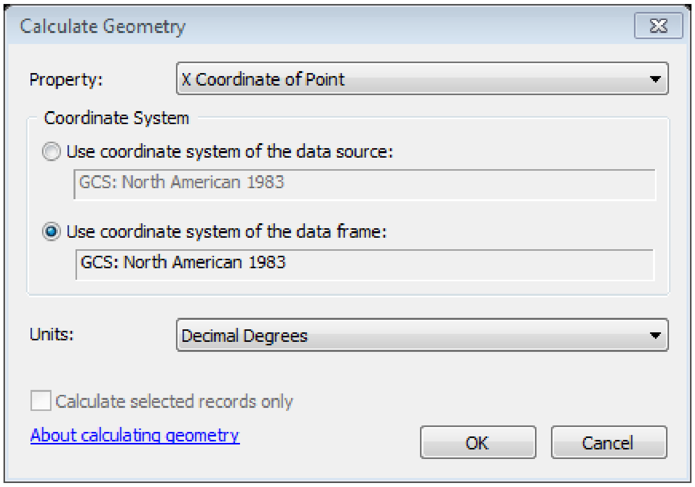
○ Use coordinate system of the data source: GCS: North American 1983

● Use coordinate system of the data frame: GCS: North American 1983

Units: Decimal Degrees

□ Calculate selected records only

        1. *Note: Latitude = Y coordinate
    iv. To Calculate New_Long:
    v. Right-click on New_Long field name
    vi. Select Calculate Geometry…and click Yes on warning about Calculating outside of an edit session
    vii. Fill out menu as follows:



Calculate Geometry

Property: X Coordinate of Point

Coordinate System

○ Use coordinate system of the data source: GCS: North American 1983

● Use coordinate system of the data frame: GCS: North American 1983

Units: Decimal Degrees

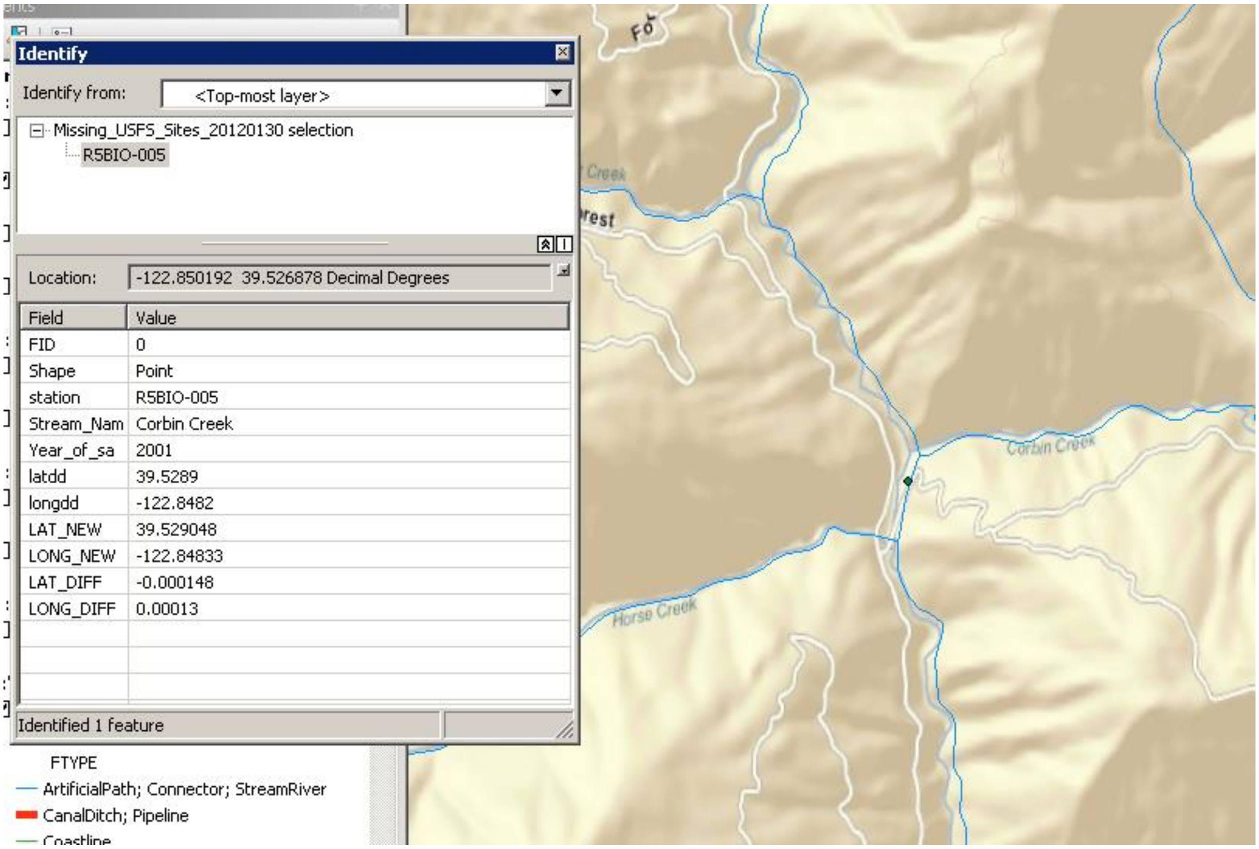
□ Calculate selected records only

    viii. Select OK and Yes on warning about calculating outside of an edit session
    *Note: Longitude = X coordinate similar to what you did for Latitude.

8. Export the points and name them appropriately to make a permanent layer. Name this shapefile XXX_Sites (where XXX is the project name).

Quality control for the Sites Basefile (for both snapped and unsnapped sites).

1. 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 large distances should be flagged, so that the catchments delineated later can receive additional review. Large snapping distances are not always problems and may have minimal impact on the catchment or the metrics calculated from the BaseFiles. 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).
2. Look for ancillary data (such as station names or descriptions, aerial imagery, USGS topographic maps) to verify sampling location. Contact sampling crews if necessary.
3. For sites close to confluences or near transitional areas, close scrutiny is required to ensure that the site is located on the correct stream segment. In the figure below, a site was sampled on Corbin Creek, near the confluence with the Eel River (as indicated by the site name). However, the point plots on the main stem of the Eel, downstream of the confluence. The coordinates needs to be manually corrected.



Identify

Identify from: <Top-most layer>

Missing_USFS_Sites_20120130 selection
R5BIO-005

Location: -122.850192 39.526878 Decimal Degrees

Field	Value
FID	0
Shape	Point
station	R5BIO-005
Stream_Nam	Corbin Creek
Year_of_sa	2001
latdd	39.5289
longdd	-122.8482
LAT_NEW	39.529048
LONG_NEW	-122.84833
LAT_DIFF	-0.000148
LONG_DIFF	0.00013

Identified 1 feature

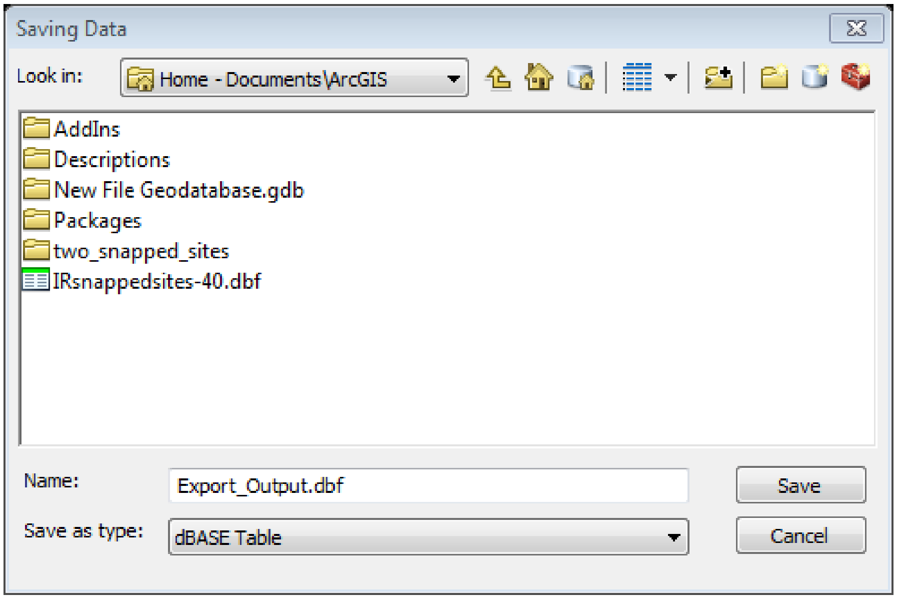
Creating the Catchments BaseFile

Below we outline the recommended approach for delineating catchments from a digital elevation model (DEM), simplified and improved by using pre-delineated watersheds in the National Hydrography Dataset Plus (NHD Plus). This approach works well for the majority of 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 ("reference") conditions. In general, dams, diversion, and inter-basin water transfers should be ignored when delineating the contributing catchment. This section describes three different ways to delineate catchments. The Basin Delineator is the preferred method followed by the ArcGIS Hydrology and Stream Stats methods. The Basin Delineator method is software that can be downloaded to user computers. The ArcGIS Hydrology method (snaps and delineates) is an ESRI online service that can only be accessed through an ArcGIS Online account. The Stream Stats method asks the user to submit sites online while returning zipped shapefiles back to the user.

Basin Delineator:

Requirements:

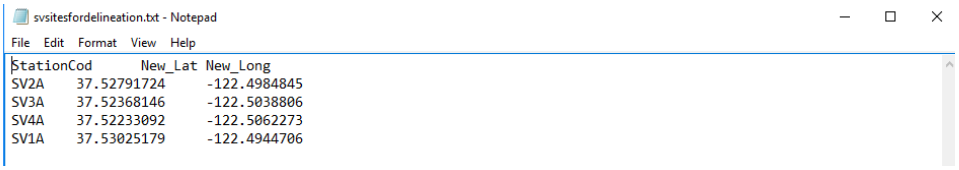
- Sites BaseFile
- 30-m DEM (link)
- NHD Plus (link)

1. Load sites BaseFile.
2. Load the NHD flowlines and Subbasins from the Hydrologic Units folder from the appropriate region; watersheds in different regions must be delineated in separate batches.
3. Save the sites attribute table as a tab-delimited text file.
   a. Export as .dbf by selecting Export… under table options
      i. 

Look in: Home - Documents\ArcGIS

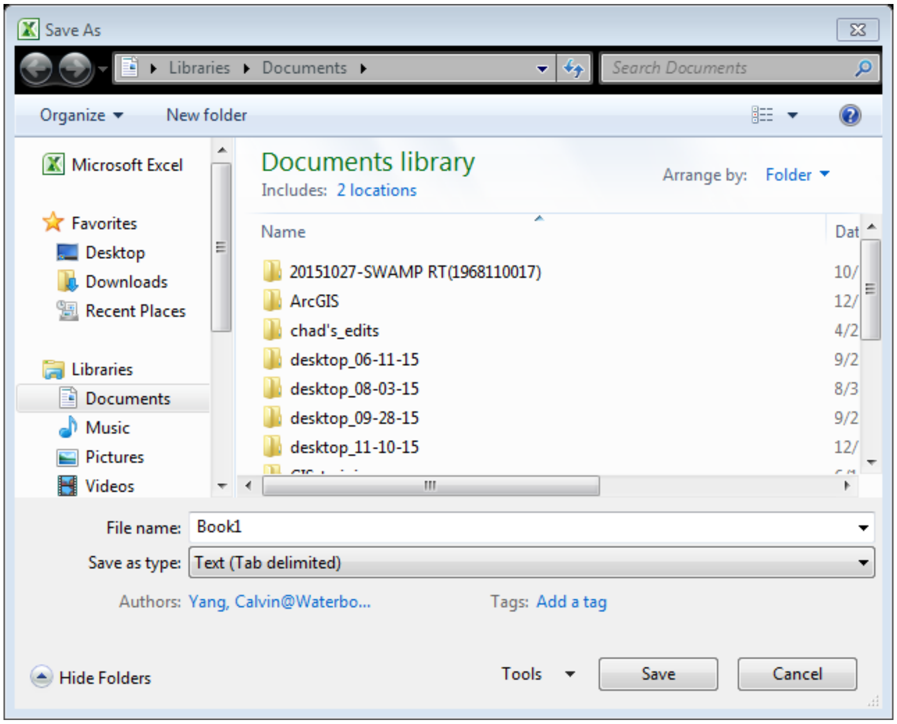
Name: Export_Output.dbf

Save as type: dBASE Table

   b. Open .dbf file in Excel
   c. Delete ObjectID column. Keep columns for StationCod, New_Lat, New_Long.
      i. 
svsitesfordelineation.txt - Notepad
File Edit Format View Help

```
StationCod	New_Lat	New_Long
SV2A	37.52791724	-122.4984845
SV3A	37.52368146	-122.5038806
SV4A	37.52233092	-122.5062273
SV1A	37.53025179	-122.4944706
```

   d. Save the file as Text (tab delimited)



File name: Book1

Save as type: Text (Tab delimited)

Authors: Yang, Calvin@Waterbo...

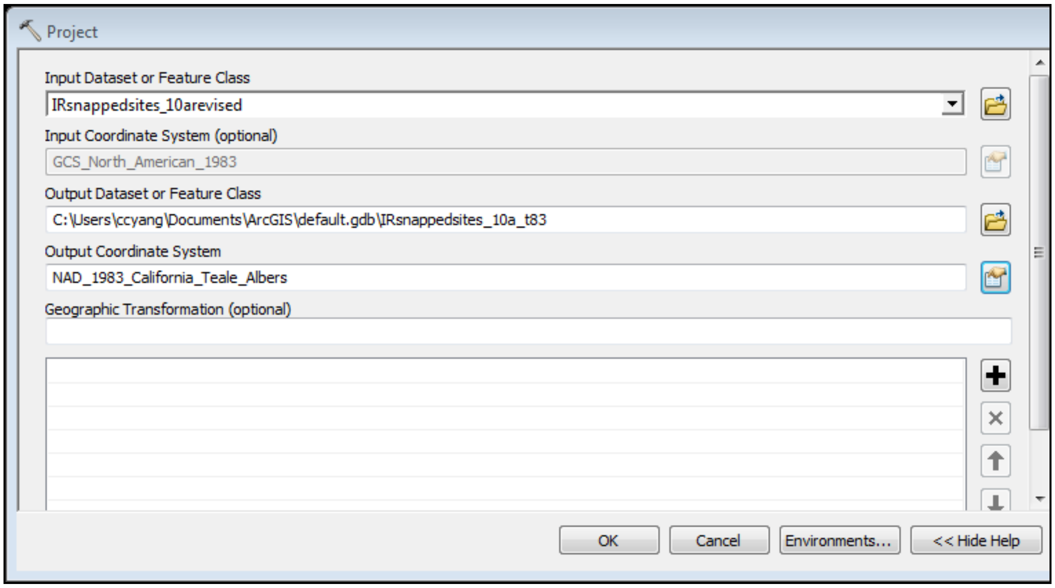
Tags: Add a tag

i.

4. Copy the text file to the NHDPlus Tools working directory (local drive) on the processing computer
5. Start the basin delineator located here (version 24020): http://www.horizon-systems.com/nhdplus/NHDPlusV2_tools.php#NHDPlusV2_BasinDelineator_Tool
    a. Before running the basin delineator, be sure that 'Split Catchments' is checked under System Setup so the delineation starts at the site rather than at the beginning of the catchment.
6. Click Run Basin Delineator.

7. The "Basin Pourpoints File" is the text file you made; browse to it.
8. Set the "Basin Shape Output File" to an appropriate directory.
9. Click Analyze, and watch for a pop-up when it completes. Depending on the number of catchments, delineation can take a long time.
10. After acknowledging the process has completed you may get a second pop-up saying that it was unable to delineate a number of catchments. You may have to delineate these manually.
11. 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.
12. Perform initial quality control checks (see section below).
13. Compare the delineated catchments to the sites BaseFile to find out which catchments have not been delineated.
14. Manually delineate those catchments that failed automatic delineation or were rejected during quality control checks in an edit session.
    a. Recalculate the New_Lat and New_Long in case there were any changes to the point locations.
15. Once all catchments have been reviewed and delineated, project the shapefile into NAD_1983_California_Teale_Albers using the project tool in the ArcToolbox

a.



Project

Input Dataset or Feature Class: IRsnappedsites_10arevised

Input Coordinate System (optional): GCS_North_American_1983

Output Dataset or Feature Class: C:\Users\ccyang\Documents\ArcGIS\default.gdb\IRsnappedsites_10a_t83

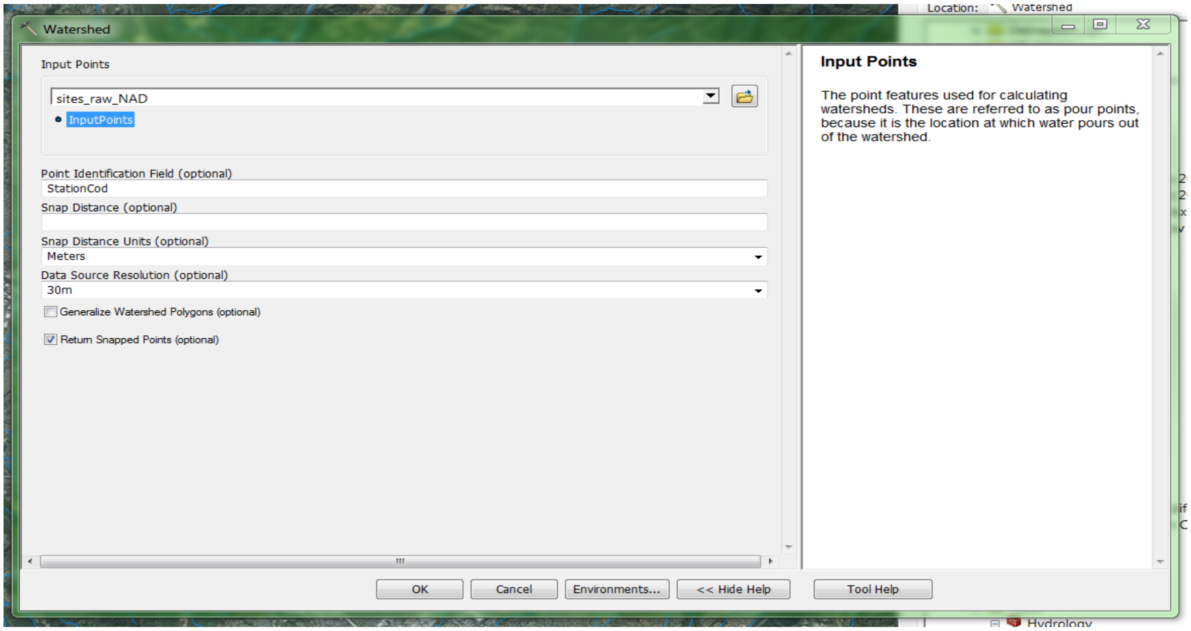
Output Coordinate System: NAD_1983_California_Teale_Albers

Geographic Transformation (optional)

16. Export the shapefile and name them appropriately to make a permanent layer. Name this shapefile XXX_WS (where XXX is the project name).

ArcGIS Hydrology

1. Input parameters:
   - Site coordinate shapefile in "Input Points"
   - Select the unique identification field (StationCod)
   - Leave snapping distance empty
   - Standard data resolution is 30m
   - Do not check the "Generalize Watershed Polygons" option
   - Check "Return Snapped Points"



Watershed

Input Points: sites_raw_NAD

● InputPoints

Point Identification Field (optional): StationCod

Snap Distance (optional):

Snap Distance Units (optional): Meters

Data Source Resolution (optional): 30m

□ Generalize Watershed Polygons (optional)

☑ Return Snapped Points (optional)

Input Points

The point features used for calculating watersheds. These are referred to as pour points, because it is the location at which water pours out of the watershed.

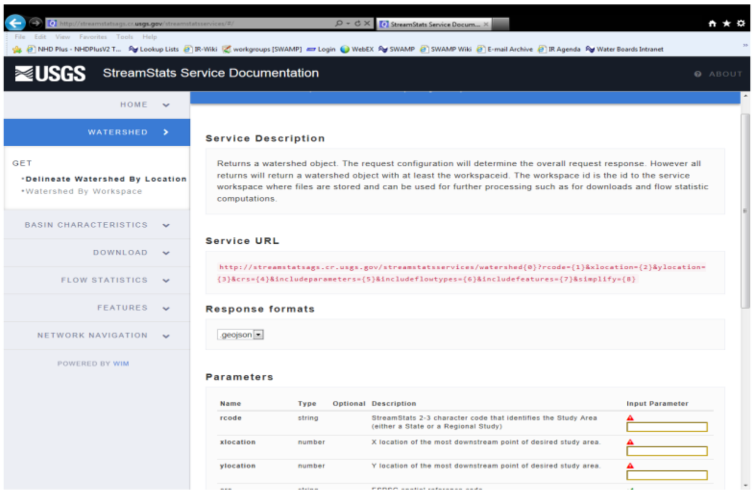
2. Export “Output Snapped Points” and “Output Watershed” outputs
3. Check the delineations for any errors/failures
    - If a snapped point is too far from the actual point, subset original point data that resulted in erroneous snapping and re-run the Watershed tool with an adjusted “Snap Distance” less than the distance to the nearest convergence. Default snapping distance is the specified resolution multiplied by 5.
    - If a delineation fails, manually snap the original points to the NHD line and re-run the Watershed tool.

4. Create new coordinate and ID fields
    - Calculate New_Lat in points file: Add new field (data type Double). Select field header - Calculate Geometry – Property = Y coordinate of point, Coordinate system = NAD 1983 (2011) California (Teale) Albers (Meters), Unit = Decimal Degrees
    - Calculate New_Long in points file: Add new field (type = Double). Select field header - Calculate Geometry – Property = X coordinate of point, Coordinate system = NAD 1983 (2011) California (Teale) Albers (Meters), Unit = Decimal Degrees
    - Update Shape_Leng: select field header - Calculate Geometry – Property = Perimeter, Unit = Meters (m)
    - Update Shape_Area: select field header - Calculate Geometry – Property = Area, Unit = Square Meters (sq m)
    - Add StationCod field: Add field – select field header – Field Calculator – Double click PourPtID in the “Fields” section

<u>Stream Stats:</u>

1. For sites not near a NHDFlowline, delineation must be done manually. We do not yet have a manual delineating process in place. Depending on the case, edits on the NHD may be required.
    a. Another option is to use USGS Stream Stats (http://streamstatsags.cr.usgs.gov/streamstatsservices/#/) may be used to delineate the watershed.
        i. To delineate a watershed using USGS Stream Stats:
            1. Fill in CA for rcode, longitude for xlocation, and latitude for ylocation. All other parameters should be left as-is.

a.



Service Description

Returns a watershed object. The request configuration will determine the overall request response. However all returns will return a watershed object with at least the workspaceid. The workspace id is the id to the service workspace where files are stored and can be used for further processing such as for downloads and flow statistic computations.

Service URL

```
http://streamstatsags.cr.usgs.gov/streamstatsservices/watershed{0}?rcode={1}&xlocation={2}&ylocation={3}&crs={4}&includeparameters={5}&includeflowtypes={6}&includefeatures={7}&simplify={8}
```

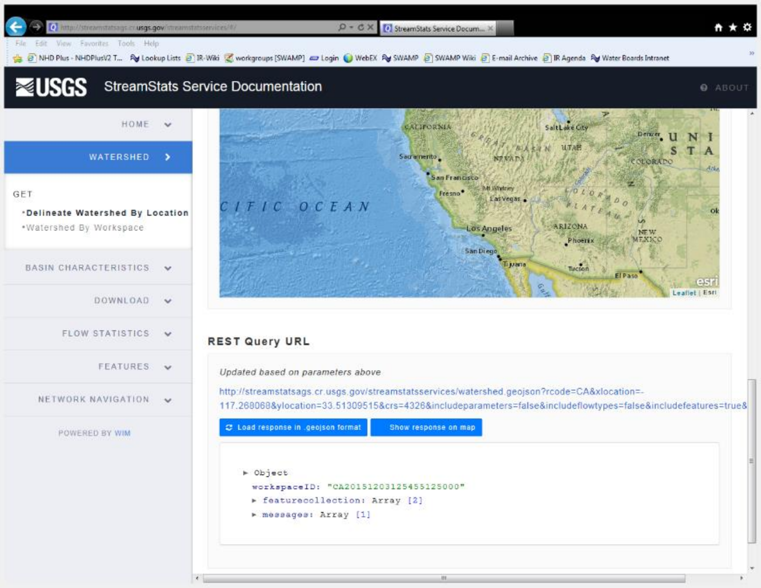
Response formats

.geojson

Parameters

Name	Type	Optional	Description	Input Parameter
rcode	string		StreamStats 2-3 character code that identifies the Study Area (either a State or a Regional Study)	
xlocation	number		X location of the most downstream point of desired study area.	
ylocation	number		Y location of the most downstream point of desired study area.	

2. Click on “Load response in .geojson format” button. Once done loading, copy the workspaceID



REST Query URL

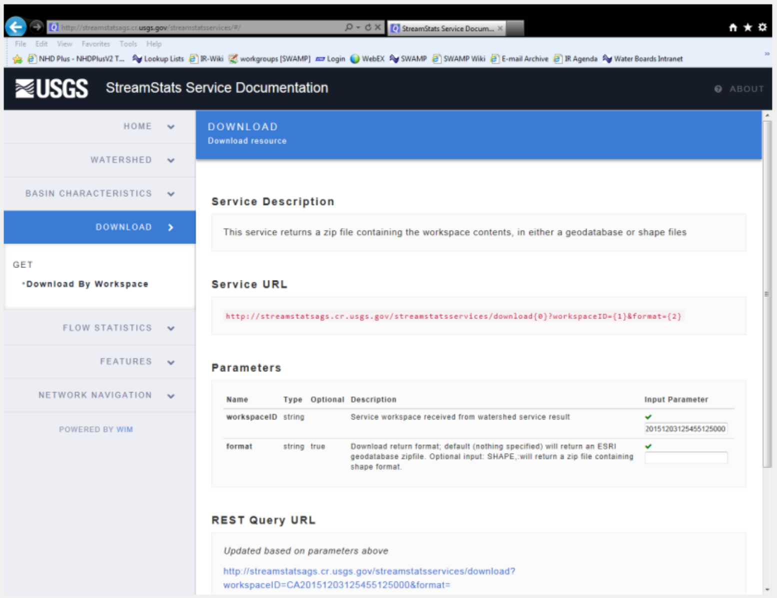
Updated based on parameters above

http://streamstatsags.cr.usgs.gov/streamstatsservices/watershed.geojson?rcode=CA&xlocation=-117.268068&ylocation=33.51309515&crs=4326&includeparameters=false&includeflowtypes=false&includefeatures=true&

```
▸ Object
  workspaceID: "CA20151203125455125000"
  ▸ featurecollection: Array [2]
  ▸ messages: Array [1]
```

a.

3. Paste unique numbers into workspaceID field and then click on hyperlink below REST Query URL



StreamStats Service Docum...
File Edit View Favorites Tools Help
NHD Plus - NHDPlusV2 T... Lookup Lists IR-Wiki workgroups [SWAMP] Login WebEX SWAMP SWAMP Wiki E-mail Archive IR Agenda Water Boards Intranet
USGS StreamStats Service Documentation
ABOUT
HOME
WATERSHED
BASIN CHARACTERISTICS
DOWNLOAD
GET
•Download By Workspace
FLOW STATISTICS
FEATURES
NETWORK NAVIGATION
POWERED BY WIM

DOWNLOAD
Download resource

Service Description

This service returns a zip file containing the workspace contents, in either a geodatabase or shape files

Service URL

```
http://streamstatsags.cr.usgs.gov/streamstatsservices/download{0}?workspaceID={1}&format={2}
```

Parameters

Name	Type	Optional	Description	Input Parameter
workspaceID	string		Service workspace received from watershed service result	✓ 20151203125455125000
format	string	true	Download return format; default (nothing specified) will return an ESRI geodatabase zipfile. Optional input: SHAPE,:will return a zip file containing shape format.	✓

REST Query URL

Updated based on parameters above

http://streamstatsags.cr.usgs.gov/streamstatsservices/download?workspaceID=CA20151203125455125000&format=

a.

4. 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.
5. Perform initial quality control checks (see Quality Control Checks for Catchment Delineations section).

Quality Control Checks for Catchment Delineations

Helpful GIS files to support QC:

- NHD Plus stream network. You may want to hide pipelines, but keep canals visible with a distinct color. **Note: the NHD (1:24K) network is often needed to resolve discrepancies between NHD+ hydrology and DEM based hydrology. If there is a conflict, the 1:24K version is usually much more accurate.**
- Elevation files, shaded relief maps or topographic maps.

1. In general, it is best to examine each catchment individually. Highlighting (or selecting) each catchment, one at a time, makes many problems obvious.
2. Look for gross irregularities, such as:
    - Holes **Fix holes** (this is a pretty rare problem). Do this by removing the polygon vertices that create the hole.
    - 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 relief.
3. Does the watershed have a "lollipop" or "frying pan" shape? This problem is most common when the site is located in a flat area with few topographic features. Unless this shape is supported by the local topography, **flag the site for further review.** Use the catchments from the corresponding regions NHDPlus as guide to fixing "lollipops". Select and merge catchments to delineated catchments where necessary to fill out catchment, or manually correct.
4. 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.
5. 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 NHD Plus 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 site with imagery from Google Earth. **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 landuse inside the delineation.
    - Watch for "divots" in the catchment. If the hydrology does not connect to the rest of the hydrologic network it will not be included in the catchment by the delineator even if they clearly feed into the catchment. Select the NHD Plus catchment and merge it into the delineated catchments in this case.

CALCULATING PREDICTOR METRICS

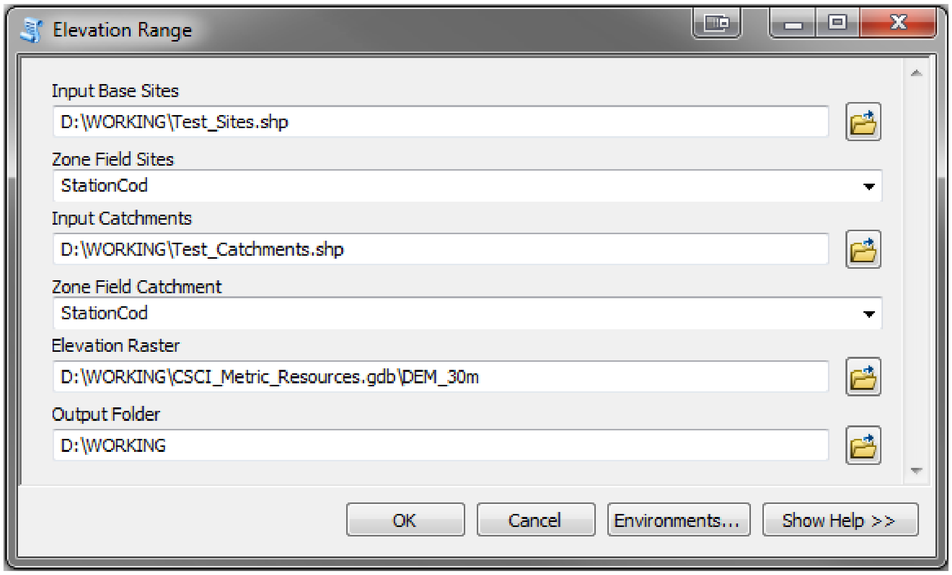
Elevation Metrics

The following describes how to process site elevation, watershed maximum elevation and the elevation change between them using the Elevation Range Python Tool in ArcGIS 10.2.2 and above. This tool requires the Spatial Analyst Extension to run.

Elevation Range Processing Tool

1. Add "CSCI_Metric_Toolbox" to your ArcToolbox
    a. right-clicking ArcToolbox
    b. select "Add Toolbox"
    c. browse to CSCI_Metric_Toolbox on your computer and click Open

2. Navigate to the "CSCI_Metric_Toolbox" and double-click the "Elevation Range" script to open its dialog box.



Elevation Range

Input Base Sites: D:\WORKING\Test_Sites.shp

Zone Field Sites: StationCod

Input Catchments: D:\WORKING\Test_Catchments.shp

Zone Field Catchment: StationCod

Elevation Raster: D:\WORKING\CSCI_Metric_Resources.gdb\DEM_30m

Output Folder: D:\WORKING

Input Base Sites: Add the site points to this input.

Zone Field Sites: Choose the field that contains the unique id for each input site point. The "Zone Field Catchments" input much have the same set of unique id values. In this example both "test_site_points.shp" and "test_catchments.shp" both contain the field "StationCod".Long

Input Catchments: Add the catchments polygons to this input. The StationCode field must correspond with the Input Base Sites for the tool to run properly.

Zone Field Catchments: Choose the field that contains the unique id for each input catchment polygon. The "Zone Field Sites" input much have the same set of unique id values. In this example both "test_site_points.shp" and "test_catchments.shp" both contain the field "StationCod".

Elevation Raster: This is the input DEM dataset. Add the “DEM_30m” raster located in the “CSCI_Metric_Resources” GDB.

Output Folder: Choose the location you wish the final results shapefiles to be saved. Intermediate files will also be saved here during processing but will be deleted upon completion.

3. Click “OK” and the tool will run. When it completes you should see two new shapefiles named “Catchments_Elevation_Ranges.shp” and “Sites_Elevation.shp”.

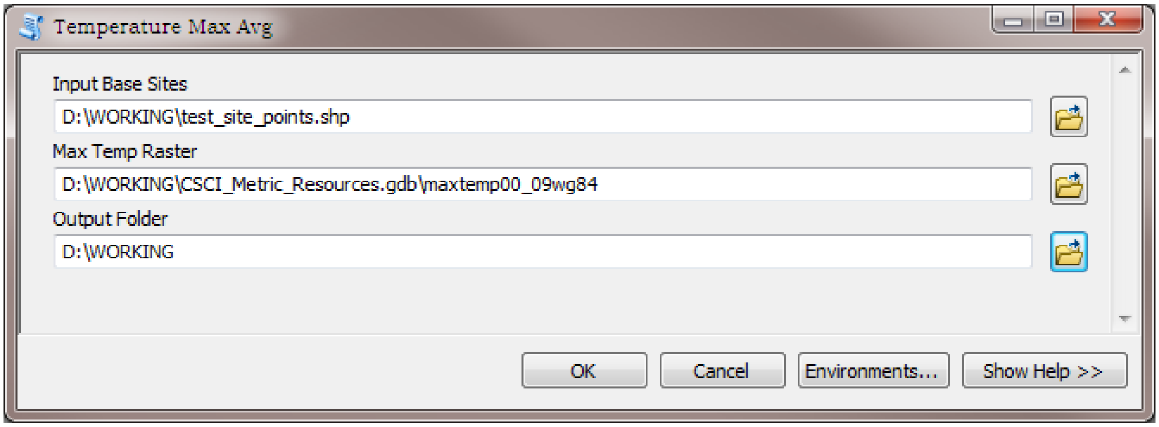
4. Add the new shapefile, “Catchment Elevation Ranges.shp” to ArcMap and open the attribute table.
    a. The follow new fields are added.
        i. SITE_ELEV – Elevation at the sample site.
        ii. MAX_ELEV – Maximum elevation of the watershed
        iii. ELEV_RANGE – Elevation range between sample site and top of watershed.
        iv. AREA_SQKM – Watershed area in square kilometers

Average Temperature

The following describes how to derive the average precipitation at a giving test site using the Temperature Avg Python Tool in ArcGIS 10.2.2 and above. This tool requires the Spatial Analyst Extension to run.

Temperature Processing Tool

1. Navigate to the “CSCI_Metric_Toolbox” and double-click the “Temperature Max Avg” script to open its dialog box.


Temperature Max Avg

Input Base Sites: D:\WORKING\test_site_points.shp

Max Temp Raster: D:\WORKING\CSCI_Metric_Resources.gdb\maxtemp00_09wg84

Output Folder: D:\WORKING

Input Base Sites: Add the site points to this input.

Max Temp Raster: This is the input max temperature dataset. Add the “maxtemp00_09wgs84” raster located in the “CSCI_Metric_Resources” GDB.

Output Folder: Choose the location you wish the final results shapefile to be saved. Intermediate files will also be saved here during processing but will be deleted upon completion.

2. Click “OK” and the tool will run. When it completes you should see a new shapefile named “TempMaxAvg_00_09wgs84.shp”. Add the new shapefile to ArcMap and open the attribute table. You will see that a new field “TEMP_00_09” has been added. It contains the maximum average temperature from 2000 to 2009 for each site.

   The temperature units are degrees Celsius multiplied by 100.

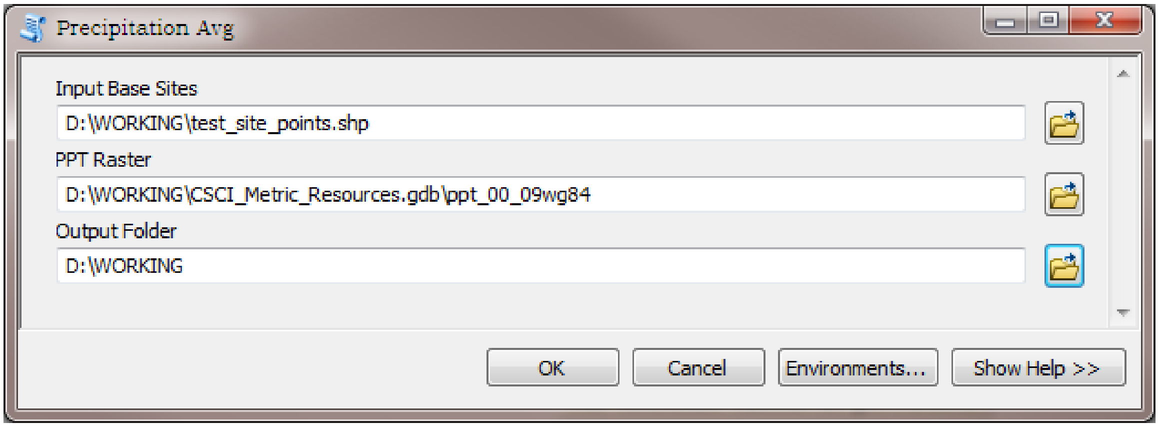
Notes: Currently to script output field assumes the standard 2000 to 2009 time frame but it can easily be modified to output a new field name for any time frame if new data is acquired.

Average Precipitation

The following describes how to derive the average precipitation at a giving test site using the Precipitation Avg Python Tool in ArcGIS 10.2.2 and above. This tool requires the Spatial Analyst Extension to run.

Precipitation Processing Tool

1. Navigate to the “CSCI_Metric_Toolbox” and double-click the “Precipitation Avg” script to open its dialog box.



Precipitation Avg

Input Base Sites
D:\WORKING\test_site_points.shp

PPT Raster
D:\WORKING\CSCI_Metric_Resources.gdb\ppt_00_09wg84

Output Folder
D:\WORKING

   Input Base Sites: Add the site points to this input.

   PPT Raster: This is the input precipitation dataset. Add the “ppt_00_09wgs84” raster located in the “CSCI_Metric_Resources” GDB.

   Output Folder: Choose the location you wish the final results shapefile to be saved.

2. Click “OK” and the tool will run. When it completes you should see a new shapefile named “PPTAvg_wgs84.shp”. Add the new shapefile to ArcMap and open the attribute table. You will see that a new field “PPT_00_09” has been added. It contains the average precipitation from 2000 to 2009 for each site.

   The precipitation units are millimeters multiplied by 100.

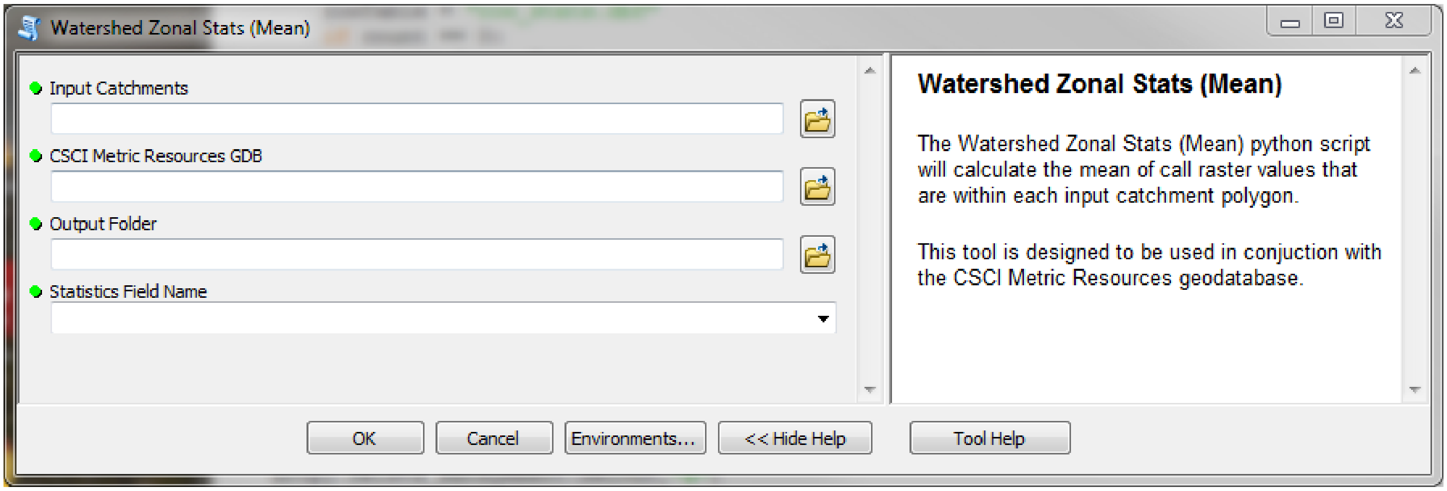
Notes: Currently to script output field assumes the standard 2000 to 2009 time frame but it can easily be modified to output a new field name for any time frame if new data is acquired.

Zonal Statistics (e.g., geology metrics)

The following section describes how to process average values of any input raster within a watershed using the Watershed Zonal Stats (Mean) Python Tool in ArcGIS 10.2.2. Currently, the tool is set up to work with only the predictors required for the CSCI (i.e., BDH_AVE, P_MEAN, SumAve_P, and KFCT_AVE), but may be expanded to other metrics in the future. This tool requires the Spatial Analyst extension to run.

Watershed Zonal Statistics Processing Tool

1. Navigate to the "CSCI_Metric_Toolbox" and double-click the "Watershed Zonal Stats (Mean)" script to open its dialog box.



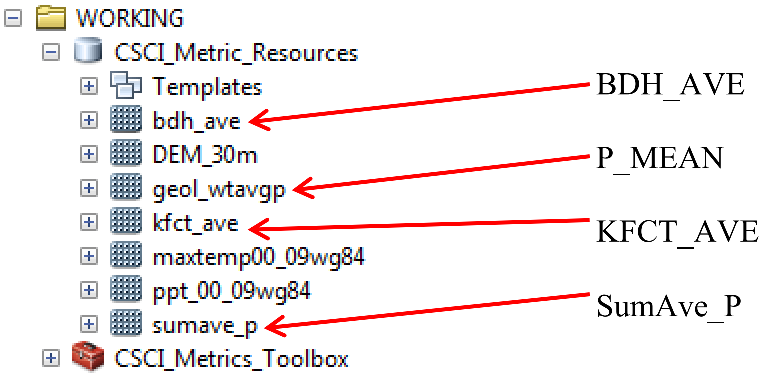
Watershed Zonal Stats (Mean)

The Watershed Zonal Stats (Mean) python script will calculate the mean of call raster values that are within each input catchment polygon.

This tool is designed to be used in conjuction with the CSCI Metric Resources geodatabase.

Input Catchments: Add the catchments polygons to this input. They must be projected in California NAD83 Teale Albers

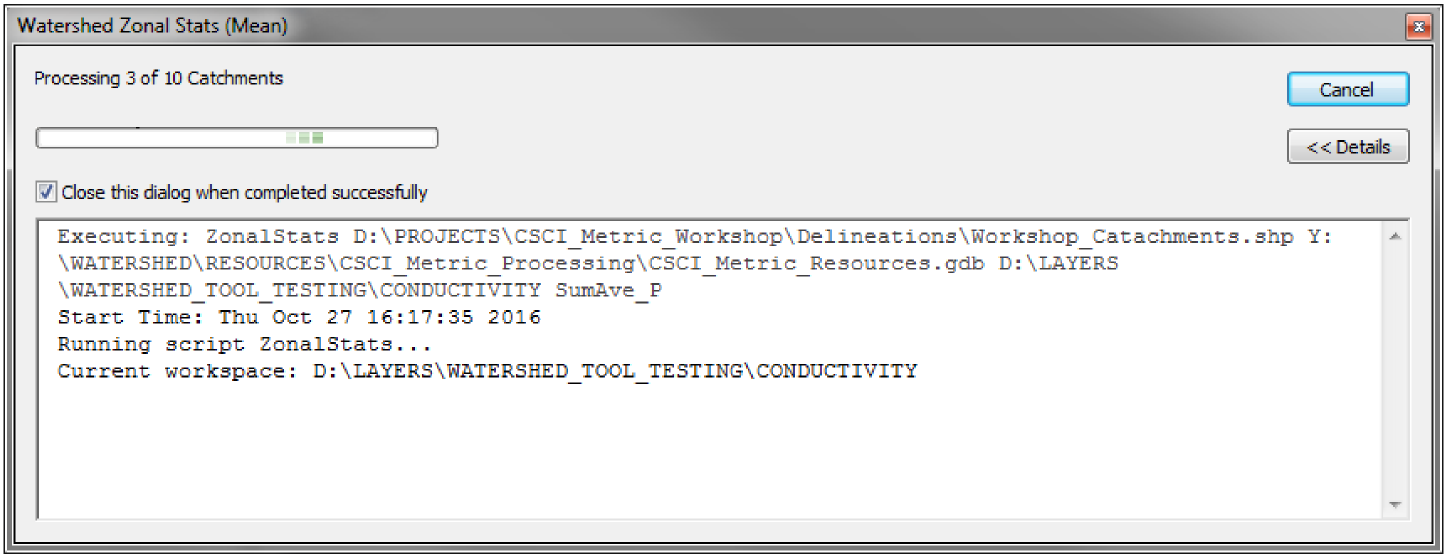
CSCI Metric Resources GDB: Select the CSCI_Metric_Resources.gdb location.



Output Folder: Choose the location you wish the final results shapefiles to be saved. Intermediate files will also be saved here during processing but will be deleted upon completion.

Statistics Field Name: From the drop down menu, choose the metric you wish to calculate. This in turn selects the correct input raster from the CSCI Metric Resources GDB and will become the name of the output field in the resultant output dataset.

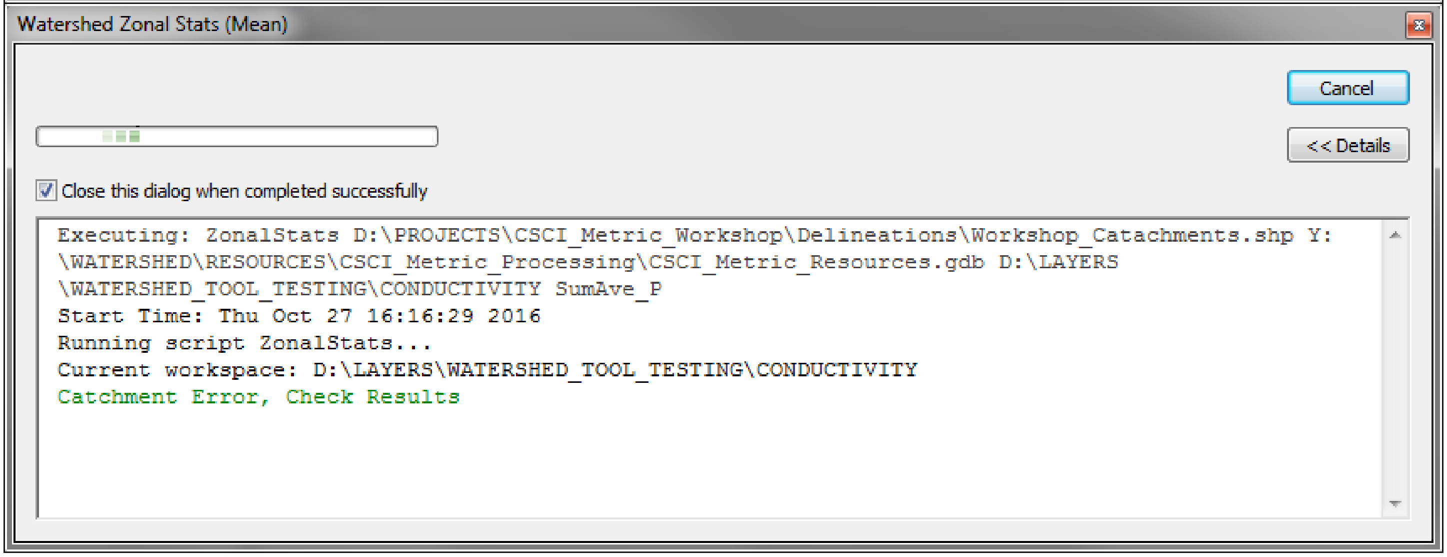
Click "OK" and the tool will run. As the tool runs, you will see a progress on how many catchments have been processed out of the total number in your input. In some cases a catchment will not overlap with the input raster, or is too small compared with the input raster cell size. In these cases, the message "Catchment Error, Check Results" is displayed.



Processing 3 of 10 Catchments

☑ Close this dialog when completed successfully

```
Executing: ZonalStats D:\PROJECTS\CSCI_Metric_Workshop\Delineations\Workshop_Catachments.shp Y:
\WATERSHED\RESOURCES\CSCI_Metric_Processing\CSCI_Metric_Resources.gdb D:\LAYERS
\WATERSHED_TOOL_TESTING\CONDUCTIVITY SumAve_P
Start Time: Thu Oct 27 16:17:35 2016
Running script ZonalStats...
Current workspace: D:\LAYERS\WATERSHED_TOOL_TESTING\CONDUCTIVITY
```



☑ Close this dialog when completed successfully

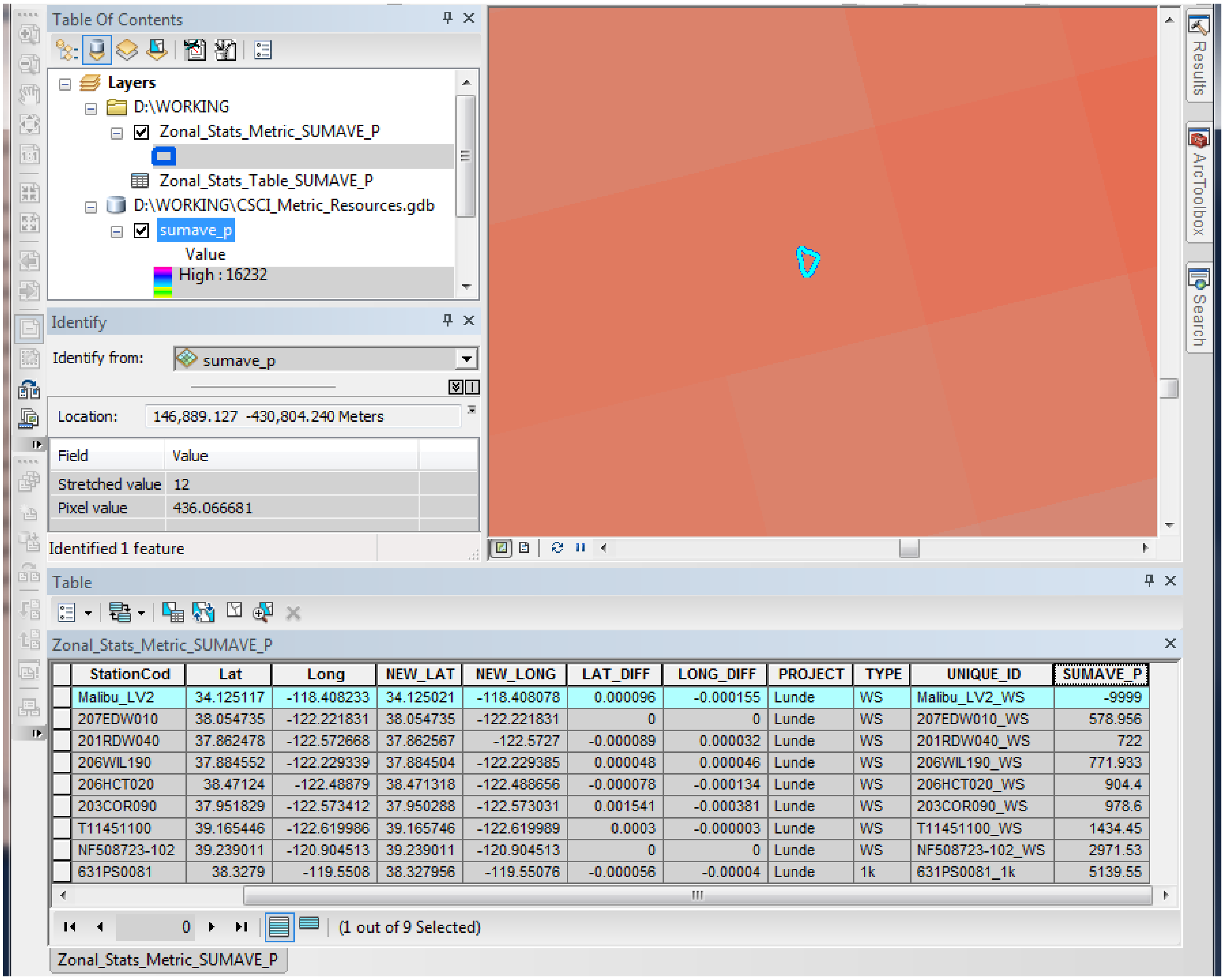
```
Executing: ZonalStats D:\PROJECTS\CSCI_Metric_Workshop\Delineations\Workshop_Catachments.shp Y:
\WATERSHED\RESOURCES\CSCI_Metric_Processing\CSCI_Metric_Resources.gdb D:\LAYERS
\WATERSHED_TOOL_TESTING\CONDUCTIVITY SumAve_P
Start Time: Thu Oct 27 16:16:29 2016
Running script ZonalStats...
Current workspace: D:\LAYERS\WATERSHED_TOOL_TESTING\CONDUCTIVITY
Catchment Error, Check Results
```

In this example, the catchment "Malibu_LV2" did not process in the model properly. It will be assigned the value -9999. Each catchment given the value -9999 will need to be manually reviewed to determine proper action. This process will be explained in more detail in steps 4 through 7.

3. When the tool completes you will have a shapefile named "Zonal_Stats_Metric_<*Statistic Field Name*>.shp. The Statistics Field Name chosen is used to name your output file. Add the file to ArcMap and open the attribute table to review.

4. Sort ascending on the Statistic field. Check for any values of -9999. If none are present then your data is complete. In the example below the catchment “Malibu_LV2” was assigned the value -9999.

5. Add the Input Stats Raster to ArcMap and zoom to “Malibu_LV2”. In the example below we can see that the catchment was too small for the Zonal Statistics operation to run properly against the “sumave_p” raster.



Identify from: sumave_p

Location: 146,889.127 -430,804.240 Meters

Field	Value
Stretched value	12
Pixel value	436.066681

Identified 1 feature

Zonal_Stats_Metric_SUMAVE_P

StationCod	Lat	Long	NEW_LAT	NEW_LONG	LAT_DIFF	LONG_DIFF	PROJECT	TYPE	UNIQUE_ID	SUMAVE_P
Malibu_LV2	34.125117	-118.408233	34.125021	-118.408078	0.000096	-0.000155	Lunde	WS	Malibu_LV2_WS	-9999
207EDW010	38.054735	-122.221831	38.054735	-122.221831	0	0	Lunde	WS	207EDW010_WS	578.956
201RDW040	37.862478	-122.572668	37.862567	-122.5727	-0.000089	0.000032	Lunde	WS	201RDW040_WS	722
206WIL190	37.884552	-122.229339	37.884504	-122.229385	0.000048	0.000046	Lunde	WS	206WIL190_WS	771.933
206HCT020	38.47124	-122.48879	38.471318	-122.488656	-0.000078	-0.000134	Lunde	WS	206HCT020_WS	904.4
203COR090	37.951829	-122.573412	37.950288	-122.573031	0.001541	-0.000381	Lunde	WS	203COR090_WS	978.6
T11451100	39.165446	-122.619986	39.165746	-122.619989	0.0003	-0.000003	Lunde	WS	T11451100_WS	1434.45
NF508723-102	39.239011	-120.904513	39.239011	-120.904513	0	0	Lunde	WS	NF508723-102_WS	2971.53
631PS0081	38.3279	-119.5508	38.327956	-119.55076	-0.000056	-0.00004	Lunde	1k	631PS0081_1k	5139.55

0 (1 out of 9 Selected)

Zonal_Stats_Metric_SUMAVE_P

6. Use the Identify tool to determine the value of the raster at “Malibu_LV2”; in this case it’s 436.066681. Use the field calculator to replace -9999 with the raster cell value.

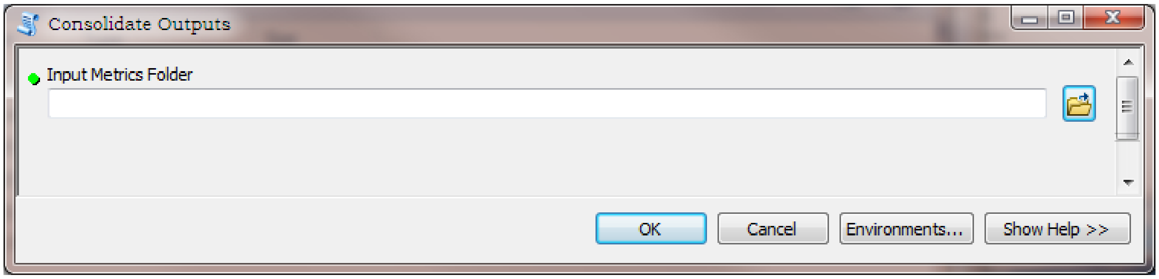
7. Repeat this process for any catchment that has a value of -9999. If the catchment does not overlap with the raster, assign the value from the raster cell closest to the catchment.

METRIC CONSOLIDATION AND DATA EXPORT

The following describes how to merge all results from other tools into a single CSV file ready for input into the R model using the Consolidate Outputs Python Tool in ArcGIS 10.2.2 and above.

Consolidate Metrics Processing Tool

1. Navigate to the "CSCI_Metric_Toolbox" and double-click the "Consolidate Outputs" script to open its dialog box.



Consolidate Outputs

● Input Metrics Folder

2. Click "OK" and the tool will run. When the tool completes a CSV file named "Final_Metrics_Consolidated.csv"

If any CSCI Metrics are missing from the chosen folder location, and error message will be given indicating which metric files are missing. Add the required files to the folder and rerun the tool.

Section 2: Instructions for Calculating CSCI Scores in R

This document assumes that the user is familiar with basic operations in the R programming language, such as data import, export, and manipulation. Although not required, we recommend using graphic interface for R, such as R-studio, which can be downloaded at http://www.rstudio.com. New users are encouraged to pursue training opportunities, such as those hosted by local R user groups. A list of such groups may be found here: http://blog.revolutionanalytics.com/local-r-groups.html. The CSCI and BMImetrics packages have a number of dependencies. Each of these packages (permute, stringr, DBl, tibble) can be installed from CRAN.

This document describes usage of CSCI package version 1.1.2.

THE SHORT VERSION

```
#Install the CSCI package the first time you run this
install.packages("devtools")#Install devtools from CRAN
install.packages("permute") #Install other required packages
install.packages("stringr")
install.packages("DBl")
install.packages("tibble")
library(devtools)
install_github("SCCWRP/BMIMetrics")
install_github("SCCWRP/CSCI")

#Load the library
library(CSCI)

#Import the bugs and stations data
bugs.df<-read.csv("bugs.csv")#                 ("") = Actual file name
stations.df<-read.csv("stations.csv")#         ("") = Actual file name

#Optional: Clean the bugs data if life stage codes are bad or missing
bugs.df<-cleanData(bugs.df)

#Calculate the CSCI
#Optional rand argument makes results repeatable
report<-CSCI(bugs.df, stations.df, rand=1)

#Export the desired reports
write.csv(report$core, "core.csv")#CSCI component scores, basic data quality info
write.csv(report$Suppl1_mmi, "Suppl1_mmi.csv")#Details about pMMI score
write.csv(report$Suppl1_grps, "Suppl1_grps.csv") #Details on ref group membership
write.csv(report$Suppl1_OE, "Suppl1_OE.csv") #Details about O/E score
write.csv(report$Suppl2_mmi, "Suppl2_mmi.csv") #Iteration-level details on pMMI score
write.csv(report$Suppl2_OE, "Suppl2_OE.csv")#Iteration-level details on O/E score
```

THE DETAILED GUIDE

Installing R-scripts

Make sure you have a good internet connection, and then run this line in the R console:

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

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 fairly large (~100 MB). You may get a warning about the file size mismatching its reported length, but this warning may be disregarded.

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 launch the CSCI library and access the help pages:

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

To receive alerts about package updates, you may join the CSCI users listserve by emailing Raphael Mazor (raphaelm@sccwrp.org). This listserve will eventually be replaced by one maintained by SWAMP.

Preparing the Input Data

Stations Data

Stations data includes all the environmental information for each station, with one row per station. The required fields are:

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
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

PPT_00_09	Long-term mean precipitation at the site
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

Field names must match spelling shown above. For the required fields, blank cells or missing values are not allowed. Please see Section 1 for information on calculating predictor data. 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 package:

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

Bugs Data

Bugs data includes all the taxonomic information for each sample, with one row per taxon (that is, flat-file format). The required fields are:

Field Name	**Description**
StationCode	Unique identifier of the site
SampleID	Unique identifier of the sample. Recommended format: A concatenation of StationCode, sample date, collection method code, and field replicate number.
FinalID	Taxonomic names. Must match values in SWAMP organism lookup lists (http://swamp.waterboards.ca.gov/SWAMP_Checker/DisplayLookUp.php?List=OrganismDetailBMILookUp). The match is not case sensitive, and a few common misspellings are recognized.
Distinct	~~Indicator of distinct taxa, provided by taxonomist. Use positive integers to indicate distinct taxa. Optional. OK to leave blank (or NA) for unknowns.~~ UPDATE: We recommend that in all cases, this field be left blank for every row of the input data.
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: http://swamp.waterboards.ca.gov/SWAMP_Checker/DisplayLookUp.php?List=OrganismDetailBMILookUp. If unknown or uncertain, you can use the cleanData() function, described below.
BAResult	Total count of the organisms

Field names must match spelling shown above. Except for Distinct and LifeStageCode, blank cells or missing values are not allowed. 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.

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

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

Getting taxonomy data from SWAMP:
If you have access to the SWAMP Reporting Module, query your benthic data as you normally would. Go to "BMI Base Queries" and export the "Benthic Taxonomy Results" report as a csv. This report should be properly formatted for calculating the CSCI.

Getting taxonomy data from CEDEN:
Benthic macroinvertebrate data are available to the general public through CEDEN (www.ceden.org). Although queries on CEDEN allow the downloading of benthic macroinvertebrate data, users will need to manually select results related to stream benthic macroinvertebrate samples (as opposed to data related to fish, algae, or plants, or non-stream macroinvertebrates). Additionally, life stage and distinct information is provided by CEDEN, but these data will require reformatting to meet the requirements of the CSCI package. The `cleanData` function and the `purge` argument (described below) may be helpful in refomatting data downloaded from CEDEN.

Calculating the CSCI

Overview:

The `CSCI` package automates all of the necessary steps to calculate CSCI scores from properly formatted input files. It uses takes 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.

The automated steps are 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 pMMI 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).

For CSCI calculation:

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

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

Caveats:

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.

Getting your score:

To calculate the CSCI, first load your bugs and stations data into the workspace, and load the CSCI library:

```
bugs.df<-read.csv("bugs.csv")
stations.df<-read.csv("stations.csv")
library(CSCI)
```

The CSCI function will 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 the following:

`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. We recommend resolving mismatches of FinalID/LifeStageCode by reviewing the data, and not by using the purge argument; however, we provide it as a shortcut for data analysis.

Interpreting the outputs:

The `CSCI()` function produces 6 reports, each as a named dataframe within a list. They can be accessed using normal R indexing (e.g., report$core, report$Suppl1_mmi, etc.). The reports are summarized as follows:

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. 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 provided below:

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.
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.
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.

Suppl1_mmi. All values are averaged across 20 iterations.

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
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

Suppl1_grps

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

Suppl1_OE

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

Suppl2_mmi

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

Suppl2_OE

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

Accessing Metadata and Reference Data

The CSCI package includes two built-in functions to give interested users access to some helpful information about the CSCI.

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 OTUs, and to assign functional feeding groups, tolerance values, and other life history information used in metric calculation.

The `loadRefData()` function generates a table containing reference data used to calibrate the CSCI. Specifically, it includes the name of each reference site, sample dates, scores, biotic group membership, and predictor values.

TROUBLESHOOTING AND FAQ

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 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:

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

Missing data

With few exceptions, missing values are not allowed.

Bad FinalIDs

Bad FinalIDs typically result from misspellings, but occasionally occur when taxonomists do not conform to SAFIT's standard taxonomic effort (available at http://safit.org/ste.html). If your data set has incorrect bug names, you may use the `purge=T` argument in the `CSCI()` function. This allows calculation of CSCI scores even if the input data has unrecognized taxa. However, it is always preferable to correct the names than to purge them, and the purge argument should only be used for preliminary analyses.

If you believe a FinalID is erroneously missing from SWAMP's lookup lists, please contact the SWAMP help desk (OIMA-Helpdesk@waterboards.ca.gov). If you believe a valid FinalID is inappropriately rejected by the scripts, contact Raphael Mazor at raphaelm@sccwrp.org.

The `loadMetaData()` function provides a 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:

```
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 work-arounds 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).

Need more help?

Join the CSCI users listserve by emailing Raphael Mazor (raphaelm@sccwrp.org). This listserve will eventually be replaced by one maintained by SWAMP.

Stations that are 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. 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 Index of Biotic Integrity (Ode et al. 2005) as a substitute index. Additionally, indices based on benthic algae (see Fetscher et al. 2014) may also be calculated in these streams.

I want to calculate the SoCal IBI/NorCal IBI, etc. Can I do that with the CSCI package?

No, but the CSCI package can make the calculations easier. There's no automatic feature to allow you to calculate any of the old IBIs (although we may add that in a future version). However, there are some functions embedded within the CSCI package that you can use to calculate the metrics. Scoring and IBI calculation could subsequently be done by hand, as per IBI requirements.

```
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)
```

Note: Users who have access to the SWAMP Reporting Module should use that tool instead.

Taxonomist over-rides of distinct taxa designations

Taxonomist over-rides of distinct taxa designations are no longer recommended for standard CSCI scoring. The CSCI calculator does not correctly score samples if the designations are at better resolution that SAFIT Level 1. That is, the calculator inlcudes 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 over-riding the automated distinct taxon designation process.

Need more help?

Join the CSCI users listserve by emailing Raphael Mazor (raphaelm@sccwrp.org). This listserve will eventually be replaced by one maintained by SWAMP.

Section 3: Cautions on Score Interpretation

Unusual Environmental Settings

Most wadeable streams can be accurately scored with the CSCI (including some nonperennial streams). However, the validity for sites from unusual environmental settings is unknown. Although indices based on predictive models typically flag sites with predictor data outside the experience of the model using a chi-square test, we do not endorse this approach, and have not included it in the `CSCI` package. Instead, we recommend a case-by-case approach to evaluating the applicability of the tool in unusual environmental settings. Data about reference sites provided by the `loadRefData()` function may help determine if a test site represents an unusual environmental setting.

Samples with Low Counts

Samples with low bug counts may have erroneously depressed CSCI scores. We have not established a minimum count of bugs for validating the CSCI, but as a rule of thumb, scores that are within 10% of the specified sample size (i.e., at least 450 individuals for the pMMI, and 360 unambiguous individuals for the O/E) may be used for most applications of the CSCI. Smaller counts may be appropriate for certain applications.

Samples with Many Ambiguous Individuals (e.g., all midges IDed to family)

Samples with many ambiguous individuals typically occur when early instars that cannot be reliably identified are abundant, or when samples were not originally taken to the desired level of taxonomic resolution (e.g., samples were identified to SAFIT Level 1). In the former case, both the O/E and pMMI may be depressed, even if the total number of individuals is very high. In the latter case, the O/E may be depressed, although the pMMI should be unaffected. Although no criteria have been established for evaluating the impacts of ambiguous organisms on the CSCI, we recommend evaluating both the Pcnt_Ambiguous_Individual and Pcnt_Ambiguous_Taxa values when interpreting scores.

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 is estimated for each sample as follows:

1. Go to Suppl1_OE, and count up the number of midge subfamilies (i.e., Chironominae, Diamesinae, Orthocladiinae, Podonominae, Prodiamesinae, and Tanypodinae) that are expected in a given sample (i.e., CaptureProb ≥ 0.5) but that are also absent (i.e., MeanObserved = 0).
2. Go to the core report, and add the number from step 1 to O for that sample. This estimates a maximum value for O.
3. Estimate the maximum O/E by dividing the estimate from step 2 by E.
4. Estimate the maximum CSCI by adding the new maximum O/E estimate from step 3 to the MMI and dividing by 2.

To automate these steps, copy and paste this function into the R console:

```
MissingMidges<-function(mylist)
{
  my.core<-mylist$core
  my.oe<-mylist$Suppl1_OE
  my.core$MissingMidges_n<-sapply(my.core$SampleID, function(x)
  {
    oe.samp<-mylist$Suppl1_OE[which(my.oe$SampleID==x),]
    length(oe.samp[which(oe.samp$OTU %in% c("Tanypodinae", "Orthocladiinae",
             "Chironominae", "Podonominae", "Diamesinae", "Telmatogetoninae",
             "Prodiamesinae") &  oe.samp$CaptureProb>=0.5 &
             oe.samp$MeanObserved==0),"OTU"])
  })
  my.core$O_MissingMidges<-my.core$Mean_O + my.core$MissingMidges_n
  my.core$OoverE_MissingMidges<-my.core$O_MissingMidges/my.core$E
  my.core$OoverE_MissingMidges_Percentile<-round(pnorm(my.core$OoverE_MissingMidges,
             mean=1, sd=0.190276), digits=2)
  my.core$CSCI_MissingMidges<-(my.core$OoverE_MissingMidges+my.core$MMI)/2
  my.core$CSCI_MissingMidges_Percentile<-round(pnorm(my.core$CSCI_MissingMidges,
             mean=1, sd=0.160299), digits=2)
  mylist$core<-my.core
  mylist
}
```

You may use this function on the outputs of the CSCI function to add new fields to the core report containing the maximum possible CSCI score (i.e., `CSCI_MissingMidges`):

```
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.

Samples Dominated by Oligochaetes

Samples dominated by taxa lacking in a certain trait information required for pMMI calculation (e.g., Oligochaeta and other non-insects) may end up failing to get scores for the pMMI and CSCI.