**Ecosystem Service Review: Methods for Round 1 and Round 2 Data Cleaning and Compilation, and Round 2 Review Assignment** (*plus a few other notes on our process for developing surveys and our method for review*)

**Author(s)**: Caitlin White

**Initiated**: 2020-08-23

**Last updated**: 2020-09-13

*All code and raw and cleaned datasets used in the methods below (with the exception of the Round 1 Google forms dataset) reside at github.com/caitlintwhite/kremeny\_analyses. The Round 1 Google forms dataset resides on the ES Google Drive as a Google Sheet file. Specific files and locations described below.*

*CTW used R to QA and compile all datasets in all review rounds.*

**Programming language and software used**:

R version 3.6.1 (2019-07-05). R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

RStudio 1.2.5001. RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

**R packages used**:

‘googledrive’: Lucy D'Agostino McGowan and Jennifer Bryan (2019). googledrive: An Interface to Google Drive. R package version 1.0.0. https://CRAN.R-project.org/package=googledrive

‘googlesheets4’: Jennifer Bryan (2019). googlesheets4: Access Google Sheets using the Sheets

API V4. R package version 0.1.0. https://CRAN.R-project.org/package=googlesheets4

‘lubridate’: Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

‘readxl’: Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

‘tidyverse’: Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

**Survey instruments**:

Round 1: “Abstract Review\_Exclusion Criteria for Kremen Review Framework”, Google Forms survey.

Round 2: Qualtrics software, Version Feb 2020 of Qualtrics. Copyright © 2020 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. <https://www.qualtrics.com>

(https://www.qualtrics.com/blog/citing-qualtrics/)

Metadata reference for Qualtrics fields: <https://www.qualtrics.com/support/survey-platform/data-and-analysis-module/data/download-data/understanding-your-dataset/>

**Reviewers**:

Laurel Brigham (LB), Laura Dee (LD), Nick Dragon (ND), Kathryn Grabenstein (KCG), Sierra Jech (SDJ), Claire Karban (CK), Aislyn Keyes (AK), Tim Korpita (TK), Julie Larson (JL), Travis McDevitt-Galles (TM), Anna Spiers (AIS), Grant Vagle (GV), Caitlin White (CW/CTW)

**Notes that apply to both round 1 and round 2 survey development and scope of review:**

Survey questions for both rounds were formed by class (not finalized yet, but put together) in the first half of November. KCG emailed out a preliminary version of the round 1 exclusion criteria google form, and the round 2 “coding google form” (we hadn’t yet decided to use Qualtrics) on 11/12/2019 for feedback. Round 2 questions were in part shaped/selected by what research directions people in the class were most interested in pursuing (in some class sessions we broke out into sub-groups by interest to refine survey questions). Essential research questions/aims of review, papers to allow re: EF only or explicit link to ES, topics not covered by Kremen 2005 (“gaps”, e.g. multifunctionality), and extent to which wanted to address/review scale in the review were discussed by the group on 11/19/2020 (see Google doc in Week 14 ES Google drive folder).

LB, CK and TK began developing the citation criteria used for the WOS search late October to early November. In an initial search, they found 559 papers that cited Kremen 2005 (excluding reviews and books). The group debated breadth (survey studies for extent to which they address 4 Kremen areas + new areas by using keywords in WOS search [e.g. to get boxes and arrows JL’s conceptual figure]) vs. depth (review all 559 for specific objective a la Seppelt et al. 2011) See LB’s email to group on 10/25/2019, and this doc: https://docs.google.com/document/d/1Dlv4r0HR692rWeZyU7HkpH-a-mBSC-YVhOqCAtB5vlc/edit).

LB, CK, TK, CTW, and IS beta tested reviewing the same 5 papers pulled from an initial WOS search for the group meeting on 11/5/2019 (class notes [“meta-analysis” folder], week 12). Goal was to gauge how long it took to answer exclusion questions, how easily we could link papers to JL’s conceptual figure (e.g. add +1 to a box or arrow), and how consistent we were in answering questions.

TK, KCG and AIS led developing the final WOS search syntax (class notes, Week 13 folder: <https://docs.google.com/document/d/1dy6i0J6dMhYcTcIIWfZTXiiopkKD_tYyXNDiS1xTl6U/edit>). Initially LB, CK and TK had included papers that had been published since Kremen 2005 and also cited Kremen 2005, but that generated over 52,000 papers, and some which cited her paper but didn’t respond her research agenda (class notes, week 13, <https://docs.google.com/document/d/16UjXnjULHcqLXnBYCkkACijywZ22gtjSy0K1AuU0bIQ/edit>). These are the final steps that produce 1933 returns (1932 unique articles):

e) TS =T/K/A: TS = (“ecosystem service\*” AND ecolog\*) - 10496 results

i) Filter only articles - 8690 results

ii) Only english - 8561 results

iii) Filter years 2006-2019

iv) Not Topic = econo\* + Not Title (cost, social pref, valu\*, govenmen\*, paymen\*)

v) Web of science index = check only science citation index expanded

vi) Web of Science Categories: ecology only- 1933 results

**1. Round 1**

**1.a. Initial abstract review and assignment methods.**

TK queried the following in Web of Science (Clarivate Analytics, Philadelphia, PA) to generate a starting pool of abstracts for review (see Week 13 folder): “TOPIC: ("ecosystem service\*") AND TOPIC: (ecolog\*) NOT TOPIC: (econo\*) NOT TITLE: (cost) NOT TITLE: (social pref\*) NOT TITLE: (valu\*) NOT TITLE: (govern\*) NOT TITLE: (paymen\*)”. These results were refined by “DOCUMENT TYPES: ( ARTICLE ) AND WEB OF SCIENCE CATEGORIES: ( ENVIRONMENTAL SCIENCES OR ECOLOGY ) AND LANGUAGES: ( ENGLISH ) AND WEB OF SCIENCE CATEGORIES: ( ECOLOGY ), Timespan: 2006-2019. Indexes: SCI-EXPANDED.” to generate 1933 returns, of which 1932 were unique articles. The duplicate record (article) was an online early release version of the same paper published in print the following year, and both of these records were assigned to AK (same reviewer).

Fourteen people reviewed 1932 unique abstracts from November 2019 – January 2020. Each person reviewed either 148 or 149 abstracts, with the exception of LD and CTW (LD helped CTW review about half of her abstracts as she was away). Specifically, five reviewers (one was LD/CW) were assigned 148 abstracts, and the other eight reviewers 149 abstracts.

TK emailed abstract review assignments to the group on 11/23/2019, and reviews were submitted from 11/25/2019 – 01/23/2020.

**1.b. Survey instrument**

Google form, with seven yes/no exclusion questions (a “yes” to any of these questions resulted in paper excluded from further consideration in review). Reviewers entered their name, paper title, answered the yes/no exclusion questions, and could optionally enter comments about the abstract. Exclusion questions were:

1. Is this a meta-analysis?

2. Is this a review?

3. This paper does NOT directly measure/model an EF and/or ES

4. This paper focuses ONLY on valuation or risk assessment

5. This paper describes ONLY a tool, but not does report implications for EF/ES on said tool

6. This paper only measures biodiversity/abundance but NOT as an explicit proxy for ES/EF

Regarding the 3rd question, aim was to exclude papers that only mention EF or ES in passing but don’t address or measure them directly (from class notes, “Search criteria methods”, week 13). The sixth question was added in December, after abstract screening started. Everyone agreed to go back to abstracts they had already reviewed that might screen out based on question 6, but only those abstracts (not all abstracts). Therefore, the biodiversity question was repeated in the Round 2 Qualtrics survey as a catch for any papers that should have been excluded in Round 1 based on stopping at biodiversity or abundance without connection to ecosystem function or service. Guidelines for answering Question 6 (“Q8” in the survey [numbering off]) was defined as follows (email from AK to class 12/3/2019):

Hi everyone,

Here's a quick recap of the discussion we had at the back of the table:

We grappled a lot with whether or not biodiversity/abundance needs to explicitly say it's being used as a proxy for an ES to be included. Some thoughts on this:

1. Measuring a proxy (like biodiversity) is different than measuring biodiversity and linking it to an ES/EF.

2. Many journals require that abstracts include a broader impacts section. Lots of people sprinkle ES in this section, hopefully it will be easy enough to tell whether or not they are actually measuring it or just mentioning that they're important!

We adjusted Q8 to be:

This paper only measures diversity/abundance but NOT as an explicit proxy for ES/EF

[where yes means exclude and no means keep.]

The standard moving forward:

* If biodiversity is measured as a proxy for EF/ES and EF/ES is the AIM of the study **KEEP**
* If biodiversity is measured and EF/ES ISN’T the AIM of the study **DISCARD**

**1.c. Data cleaning and compilation**

Google form survey results are read in dynamically from the ES Google Drive into R using the ‘googledrive’ and ‘googlesheets4’ packages. Because of the simplicity of the survey, data cleaning was mostly limited to correcting paper titles with typos and screening for unanswered questions or missing abstract reviews. CTW screened each paper for either all “No” answers or at least 1 “Yes” among the first six questions. CTW contacted reviewers to resubmit their survey if questions 1-5 were a combination of “No” and missing answers (looser on question 6 since it was added partway through review and was going to repeat in the round 2 survey). Similarly CTW contacted reviewers if an assigned abstract review was not in the googlesheet form (sometimes seemed to not record a submission if many people were using simultaneously). Ultimately all abstract reviews accounted for.

**1.d. Round 1 summary**

Of the 1932 unique starting abstracts, 1149 (59.5%) were excluded and 793 (40.5%) kept for Round 2 review. Reason for exclusion from most to least frequent were: 1) no direct measure of ecosystem function or service (733 papers, 63.8%), 2) review only (268 papers, 23.3%), 3) stopped at biodiversity/abundance, no connection to ecosystem service (60 papers, 5.22%), 4) meta-analysis only (38 papers, 3.31%), 5) assessment of ecosystem service valuation or risk study [social dimensions paper] (27 papers, 2.35%), 6) primary intent is to introduce new method or evaluation tool (23 papers, 2%).

**2. Round 2**

**2.a. Paper selection and assignment**

Of the 793 studies that proceeded from round 1, 392 (roughly half) were randomly selected for further in-depth systematic review. Recognizing the amount of time likely required to review a paper more in depth, the range of study topics and systems in the paper pool, and breadth of knowledge among the review group (and therefore potential for multiple interpretations/different levels of understanding), we decided to subset the dataset so two reviewers could independently review a given paper, then converge on final answers for any conflicting answers.

The same fourteen reviewers from round 1 reviewed papers for round 2 (LD and CTW reviewed papers separately). Each reviewer was assigned as a primary reviewer for 28 papers and paired as a second reviewer for two other primary reviewers and assigned 14 papers from each primary reviewer. In total, every reviewer was assigned 56 papers (28 as primary reviewer, 28 as secondary reviewer). Papers were assigned and reviewers paired with two other reviewers via random selection, with the caveat no person reviewed a paper they had already screened in round 1. We used R to randomize round 2 paper selection, reviewer pairing, and paper assignment.

Reviews were submitted from 02/05/2020 – 04/29/2020.

**2.b. Survey instrument**

As a group we designed the study survey questions and KCG created the survey in Qualtrics (Qualtrics, Provo, UT). KCG sent a finalized version of the Qualtrics survey to the group to begin reviews on Feb 5 2020. Subsequently, the following questions were added to the Qualtrics survey later (as issues arose during reviews):

* Q3b: The paper ONLY focuses on social dimensions of services (no measurement of services or processes)? (added 02/17/2020)
* Q3c: ﻿This paper is a Review, Synthesis, Meta-analysis or Methods only paper. (added 2/17/20)

While ultimately only 63 papers were double reviewed (instead of all 392 as intended), the group used a reference document to guide answering questions in consistent manner (“Survey Instructions” Google doc, in Round 2 Paper Review folder). In both review rounds, we also brought papers we were unsure of in the group to discuss with others how to answer, whether to exclude, etc.

When all reviews were complete (each reviewer had reviewed their 28 primary assignments), CTW downloaded the raw Qualtrics data as a .csv file and QA’s and compiled in the data in R.

**2.c. Data QA, cleaning, and compilation**

As the survey for round 2 was more in-depth, data quality assurance and cleaning for round 2 was more involved. Generally, data QA was as follows:

1. Match paper titles to ensure all papers assigned reviewed
   1. Correct typos in titles;
   2. Identify correct paper title for surveys submitted with incomplete or missing paper title (question required answer but value entered not a paper title)
2. Flag papers for additional review based on survey notes, methods, ecosystem; correct as needed
   1. Exclusion:
      1. Flagged: reviews that noted “exclude” or uncertainty about paper in review comments (Q24), particularly for papers where 2nd and 3rd exclusion questions added after reviewer already reviewed survey; double-reviewed papers where two reviewers inconsistent in exclude answers
      2. Several papers were added to exclusion review by GV and JL when reviewing papers for the new scale question (see below), and by CTW when reviewing papers for outstanding reviewer corrections that couldn’t be addressed by original review (e.g. people away in field, moving, non-responsive)
      3. LD, NBD, and CTW reviewed flagged papers for a final decision on whether the paper should be excluded
   2. Ecosystem
      1. Pulled all papers where reviewer entered an “Other” ecosystem; also papers studying aquatic systems (marine, freshwater, coastal) for potential reclassification as wetland (wetland was not a category available in survey)
      2. LD and IS reviewed papers, created new category of “Agricultural/Agroforestry/Rural”, reassigned any study with “Other” to a standardized system answer so no papers had an “Other” system
      3. If reviewer had entered comments in “Other” system field, CTW moved them (appended) to General Information comments (Q6) so preserved
      4. IS did additional review to reclassify studies as “wetland/riparian” if appropriate. Papers for this review were flagged by Julie and Grant during new scale question review, and by pulling any paper that had indicated an aquatic ecosystem (e.g. coastal freshwater, or something indicative in the any of the optional write-in notes questions; or matched the regular expression “basin| catchm| fen | riparian| wetland| meadow| watershed” in the abstract).
      5. To address inconsistencies in responses, CTW appended “Terrestrial” to any study that was marked as “Agricultural/Agroforestry/Rural” (some reviewers had checked both, some had not checked Terrestrial)
   3. Methods
      1. Flagged all papers that where reviewer checked “Other” methods or entered methods/general study info along with whether reviewed marked their methods as experimental, observational, or data simulation
      2. ND and AK reviewed papers; rule they determined: “If a paper selected model/data simulation but used existing data (e.g. fisheries records, LTER) then it should not select observational too. Observational should only be checked if they collected data for that study.” Also reclassified “Other” methods papers so papers in study have “Other”.
      3. If reviewed had entered text in “Other methods”, appended those comments to General Info comments (Q6) so preserved.
   4. Spatial scale
      1. Pulled any paper where reviewer entered notes about scale, and double reviewed papers to compare consistency in answers
      2. JL and GV reviewed the flagged answers and papers, decided too much inconsistency in responses that data unusable. JL and GV created new scale questions and reviewed all non-excluded papers (papers not excluded by original reviewer on Q3) to answer those (more details below in #5). CTW appended new scale question data to ES dataset, replacing old scale questions (whether study had spatial component, # of plots and # of sites).
      3. CTW retained original reviewer qualitative notes entered about spatial scale of study in the dataset.
   5. Kremen Topics
      1. As a check, pulled all records that had an environmental driver, biotic driver, had indicated study included an ESP, and any review that had answered “Yes” to temporal component, spatial component, or connectivity and answers to Kremen Topics (ESP, community structure, environment, and scale [Q14]) for comparison
      2. Flagged papers that had no Kremen topics checked to review consistent with drivers entered and responses to scale, temporal, and connectivity components
      3. Group decided if any of the above questions on drivers, ESPs, spatial-temporal scale or connectivity were affirmative, then the corresponding Kremen Topic should be checked as well. Correct via code.
   6. ESP check
      1. Pulled records based on regex match of "abund|richn|shannon|divers|biodiv|evenn|size|mass" in drivers entered or any survey notes fields to review consistency with response to Kremen community structure topic (Q13) and ESP type question (Q14)
      2. Ultimately did not apply any programmatic corrections to this in code because apparent reviewers were checked “community structure” or “ESP” as a Kremen topic addressed based on driver \*OR\* response variables entered.
3. Review and bin driver and response variables entered
   1. Extracted and wrote out all unique driver variables and response variables entered, by Ecosystem Service row (Q12), for review
   2. KCG and SDJ reviewed and binned driver and response variables into coarser groups; LD and AK gave feedback on bins created; CTW finished all assigning driver variables to bins for any SDJ did not assign (e.g. new variables added as reviewers sent corrections)
   3. CTW also re-assigned category of driver group in field “clean\_group” based on driver bin (e.g. variable that falls in “land use and land cover change” bin should fall under “Human” driver, might have been entered under Environmental or Biotic)
   4. After review and discussion of how to use dataset, group decided not to bin response variables (ultimately would not use those bins, too many response variables to bin—much more diversity in response variable type compared to driver variables since no standardized response variable options in survey like standardized driver variables)
4. Screen for missing answers or gratuitous answers
   1. Qualtrics survey set up so answers to most questions mandatory, and that certain questions could not be answered contingent on response to earlier question. That said screened for following:
      1. If “Yes” to temporal component, time internal answered; if “No”, no time answer allowed
      2. If “Yes” to connectivity, distance answered; if “No”, no distance allowed
      3. Q12: If drivers entered in ES row, at least 1 response variables entered
      4. Q12: If response variables entered in ES row, at least 1 driver variable entered
      5. Q12: If “Other” driver checked, other driver described in text field
      6. Q12: If value entered in “Other driver” text field for given driver group (Environmental, Human, or Biotic), “Other” driver checked under corresponding driver group category; if not, review
   2. Missing answers flagged, exported to a .csv, and sent to original reviewer to fill out and return to CTW for incorporation in dataset
   3. Note: questions not flagged (missing allowed):
      1. Q12: effect direction, and response variable as EF, ES or Proxy. During review and preliminary data exploration, realized those questions (because of how survey designed and/or how question set up) do not adequately capture data desired (e.g. categorical drivers that do not have any effect direction [JL example], non-linear effects [some people left answer blank, some checked “mixed”; ultimately could only give one answer for potentially many variables entered, instead of one to one, so data not useful)
5. Create and review new spatial scale question (JL and GV)
   1. JL and GV considered using answers provided for original scale questions (Q8: Does the paper consider or compare multiple spatial scales? If yes, Q9: What is the size of the unit of replication? [multiple selections allowed for multiples scales of inference. Reviewers were to select spatial extent and number of total plots and total sites, and could optionally enter notes). Were considering binning response into coarser grains, but in group discussion and JL and GV’s review determined questions were answered inconsistently (data not reliable) and so better to create new, simple scale question.
   2. New questions designed to capture: 1) spatial extent [scope of inference following Hefferman et al. 2014], defined “local”, “macro-scale”, “global”, or “NA/Other” (choose one only), and 2) spatial nestedness or multiple scales of analysis (one of: Yes, No, or Other allowed). Grant and Julie divided set of non-excluded papers and re-reviewed all, discussing ones together if answer not as apparent.
   3. More and very explicit details on criteria for answer selection in each new question found in Scale\_re-review\_trial\_18May20 (in Round 2 Paper Review/Grant&Julie\_scale folder on ES GoogleDrive).
   4. New scale data uploaded to Github repo on June 4 2020. CTW incorporated in dataset, assigned survey timestamp for scale questions matching timestamp of file put on Github respository, and changed reviewer initials in record (“Init”) to match JL, GV or JL/GV (as shown in new data file) for new scale questions only. Remaining question answers for a paper review have initials of original reviewer(s) as assigned by CTW in Feb 2020.
6. Apply reviewer corrections to individual papers
   1. Update answers for any individual corrections sent to CTW by reviewer (either Excel sheet or .csv with correction read into R and corrected answer assigned to clean\_answer in dataset)
7. Assign coarse driver bins and clean\_group for driver and direction of driver effect
   1. Once all missing response and driver variables corrected, coarser bins and clean groups attached to dataset so unique drivers and responses variables preserved but binned drivers available for analysis (every driver variable entered has a coarse bin and clean group assigned; if direction of effect answered, clean\_group assigned based on clean\_group of driver)
8. Logic checks and correction for Kremen Topics
   1. Q13, Kremen Topic 1: ESP
      1. If “Single species” checked in ESP type question (Q14), then “ESP” should be checked for Kremen Topics question (Q13)
      2. If ESP driver checked in Q12 but “ESP” or “Community Structure” not checked in Kremen Topics, add “ESP” as Kremen Topic
   2. Q13, Kremen Topic 2: Community Structure
      1. If has biodiversity driver (coarse bin, Q12) and Kremen Topic 1 or 2 not checked, add Community Structure
      2. If “across species” checked in ESP type (Q14) and did not check Kremen Topic 2, add Community Structure
   3. Note Q13 KT 1 and 2: did not ever remove KT 1 or 2 selection because reviewers also checked those based on response variables studied. Only added KT 1 or 2 if driver answers to Q12 or Q14 indicated at least one of those should be (conservative correction)
   4. Q13, Kremen Topic 3: Environment
      1. If environmental driver present (based on clean\_group = “Environment”, “Environment” should be selected for Kremen Tropics addressed
      2. If no environmental drivers entered (based on clean\_group “Environment”), “Environment” should NOT be selected for Kremen Topics addressed
   5. Q13, Kremen Topic 4: Scale
      1. If multiple spatial scales indicated in new scale data, or study has temporal component, “Scale” should be selected for Kremen Topics addressed
      2. If multiple scales not indicated in new scale data, and study does not have temporal component, “Scale” should NOT be selected for Kremen Topics addressed
9. Screen double-reviewed papers, flag as needed
   1. Screen all answers after exclusion question (Q3, addressed above) for congruency: if not the same, flag, write out for reviewers to converge on final answer
10. Apply reviewer corrections to double-reviewed papers, collapse answer into final answer
    1. Prefix reviewer initials to reviewer notes if present ([initials]: [notes]), collapse all optional notes fields, by order of reviewer (reviewer 1 notes, if any, precede reviewer 2 notes, if any)
    2. Prefix reviewer initials to original review clean\_answer, collapse both by reviewer order, and assign collapsed answer to “answer” col for the final record. If both answered agreed, that answer was used for the clean\_answer. If answer was flagged for correction/review, used corrected answer for the final record. For Q12, joined appropriate coarse driver bin and clean\_group.
    3. For final record QA note value, preserved any QA notes from the original (independent) reviews with reviewer initials prefixed to note to indicate whose review note applied to, and appended QA note for any corrections or modifications made to the final record.
11. Compile final version cleaned ES dataset
    1. Collapsed double reviewed paper responses assigned new unique ResponseID and version = “final” to distinguish from original individual reviewer responses (version = “original”). Nomenclature for new Response ID is R[#]unified, where non-excluded papers were numbered 1-34 (count of non-excluded double reviewed papers) according to paper title alphabetical order, and excluded papers were numbered 35-64 also ordered alphabetically by title. Original double-review paper responses and single review paper responses retain their Qualtrics-assigned ResponseID.
    2. Single-reviewed papers assigned version = “final”
    3. System time compiled in R was assigned as the StartDate, EndDate, and RecordedDate for final double-review paper records; otherwise original double-review paper responses and single review paper responses retain their Qualtrics-assigned timestamps.
    4. Stack all responses for non-excluded papers (original double-review responses, final double-review responses, and single-review responses) and write out clean dataset to Github repository for analysis

**2.d. Round 2 summary**

Of the 392 papers reviewed in Round 2, 119 (30.4%) were excluded. Reason for exclusion, from most to least frequent were: 1) stopped at biodiversity/abundance (49, 41.2%), 2) review/framework/synthesis/meta-analyses only (34, 28.6%), 3) social dimensions/valuation study only (31 papers, 26.1%), and 4) did not measure ecosystem function or service (should have been excluded in round 1) (5 papers, <1%).

Of the 273 studies retained, 34 were evaluated by two reviewers independently and the other papers were evaluated the assigned primary reviewer only. While our intent was for all papers to be double-reviewed, time involved in evaluating one paper necessitated single reviews and we use the 34 double reviewed papers to qualify consistency of our data. Generally, reviewers evaluated the same paper similarly for most questions. Questions that had the most frequent, albeit slight, discrepancies between reviewers concerned ecosystem studied, methods used, temporal component, which of the 4 areas Kremen (2005) outlined were addressed by the study, and type of Ecosystem Service Provider studied. These are multiple selection questions, and typically reviewers would overlap in one or two options selected and differ by one. Reviewers re-assessed questions with incongruent answers to determine on a final answer. More specifics on consistency between reviewers below.

**3. Exclusion dataset**

A cleaned dataset of the full answers for all round 1 abstracts, and cleaned dataset of all answers to all three exclusion questions for round 2 exclusion papers are saved on the Github repository in their respective round folders. Additionally, CTW compiled all excluded studies from round 1 and round 2 with a single (primary) reason for exclusion. The round 2 Qualtrics survey was designed to end as soon as a reviewer selected “Yes” to an exclusion question (multiple reasons for exclusion were not allowed), but in the Round 1 Google form, reviewers were allowed to check “Yes” to multiple questions. Not every reviewer continued answering questions once they checked “Yes”, however, so the method for compiling the complete excluded papers dataset was to use the first question with a “Yes” answer, by question order in the Round 1 survey, for the primary reason. In other words, each paper reviewed and excluded in round 1 or 2 only has 1 reason for exclusion in the master exclusion dataset. This master dataset is what is used to report summary statistics on excluded studies in the manuscript.

**4. Overall summary**

**4.a. Reasons for exclusion**

In total, of 1932 starting papers identified in WOS search, 1268 (65.6%) didn’t meet our criteria. Lumping review or framework papers, meta-analyses, and new methods or evaluation tool/approach papers into one category, across both rounds reason for exclusion from most to least frequent were: 1) no direct measure of ecosystem function or service (738 papers, 58.2%), 2) review/conceptual/synthesis/new methods only (363 papers, 28.6%), 3) study stopped at biodiversity or abundance, did not link to ecosystem service (109 papers, 8.6%), and 4) study was social dimensions/valuation paper only (58 papers, 4.6%).

(RStudio console screenshots)

Imagen que contiene pájaro

Descripción generada automáticamenteTable 1. Ungrouped reason for exclusion

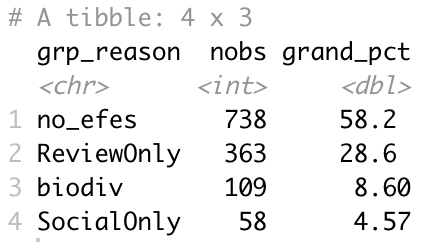


Table 2. Grouped reason for exclusion.

**4.b Double-review consistency**

Of the 392 unique papers reviewed in round 2, 63 were double reviewed. 34 of these papers were kept, and 29 excluded.

Of the 29 excluded papers that were double reviewed, reviewers agreed on the reason for exclusion for 15 papers. For the remaining 14 double-reviewed papers excluded, 6 had partial agreement (reviewers agreed on exclusion for 2 papers but based on different reasons, i.e. checked “Yes” in different sub-questions of Q3; completed the survey but both reviewers expressed doubt about whether the paper should be kept in the survey notes for 2 papers; one reviewer excluded by Q3 and the other reviewer completed the review but opined in the survey notes paper should not be kept or should be discussed for inclusion for 2 papers), and 8 did not agree (one reviewer completed survey with no hesitation in the survey notes, and the other reviewer excluded paper by Q3 or completed the survey but expressed doubts in the survey notes). Additionally, reviewers did not agree on exclusion for 3 papers ultimately kept in the study after a third party review (i.e. one reviewer excluded paper, the other reviewer kept paper, and third party opined to keep).

Stated differently, of 63 double-reviewed papers 11 (17.5%) had no agreement on exclusion, 6 had partial agreement (9.5%), and 46 (73%) had full agreement, including the reason for exclusion or to keep the paper in the study. If use the Runting et al. 2017 system of evaluation for agreement, that averages to 0.78 agreement score (where 1 = complete agreement, 0 = no agreement).

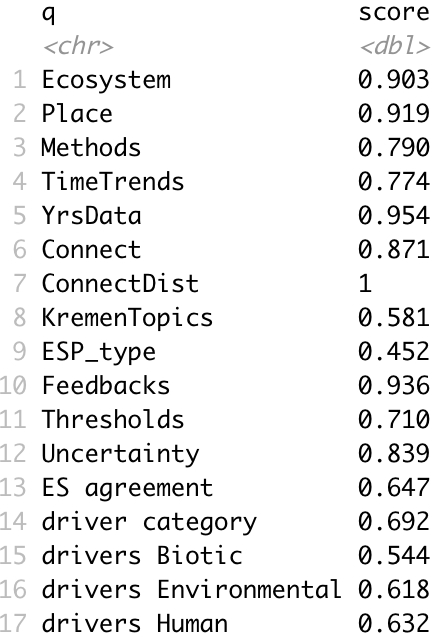
For kept papers, consistency is summarized by multiple choice non-Q12 questions and by Q12 (ES’s, responses and variables paper addresses). The overall agreement score for kept papers where both reviewers agreed paper should be kept is 0.757. Tables 3 and 4 show the breakdown by question.

Table 3. Round 2 percent agreement between reviewers for double-reviewed papers by question for papers where both reviewers agreed paper should be kept (n=31). “ES agreement” assesses agreement in Ecosystem Services selected (did reviewers fill out response and driver variables for same ES?), “driver category” assesses whether reviewers agreed on coarse driver categories in study (selected any driver or entered any other driver in same category), and the last three questions (“drivers …”) assesses whether reviewers selected same standardized driver answer and entered an other driver (ignoring verbatim other driver answer entered, just assessing whether both reviewers agreed a driver fell in that group but standardized answers available were not appropriate). For questions “YrsData” and “ConnectDist”, only considered congruence in responses when reviewers both indicated “Yes” to “TimeTrends” and “Connect” respectively since those triggered subsequent questions (i.e. not double-counting inconsistent answers when disagreed on the parent question).

Imagen que contiene texto

Descripción generada automáticamente

Table 4. Agreement score by question (0 = no agreement, 1 = complete agreement), based on procedure used for Table 3. Average agreement score for kept papers (where both reviewers agreed paper should be kept) across all questions is 0.757.



**4.c. Potential bias from publication source**

The starting pool of papers came from 128 distinct peer-review journals. After round 1 review, 98 distinct journal sources remained, and after random title selection for round 2 review, 80 distinct journal sources. Studies kept after round 2 review came from 69 distinct journals.

Table 5 shows percent contribution of study papers by each review round and used for meta-analysis results by the top 10 contributing journals in each round. Percent contribution of all journals in each round ranged from approximately 0.5-7.7% (1-149 studies of 1932 unique studies) in round 1, 0.26-7.7% (1-30 studies of 392 unique studies) in round 2, to 0.26-6.9% (1-27 studies of 273 unique studies kept) for final analysis.

Table 5. Journal contribution of studies by review round and final data analysis for the top 10 contributing journals in each review or analysis phase. Across all rounds, at most 7.7% of studies considered came from the same journal.

