
Supplementary information

A scoping review of research funding for small-scale farmers in water scarce regions

In the format provided by the
authors and unedited

Supplementary information

A scoping review of research funding for small-scale farmers in water scarce regions

In the format provided by the
authors and unedited

Supplemental Information

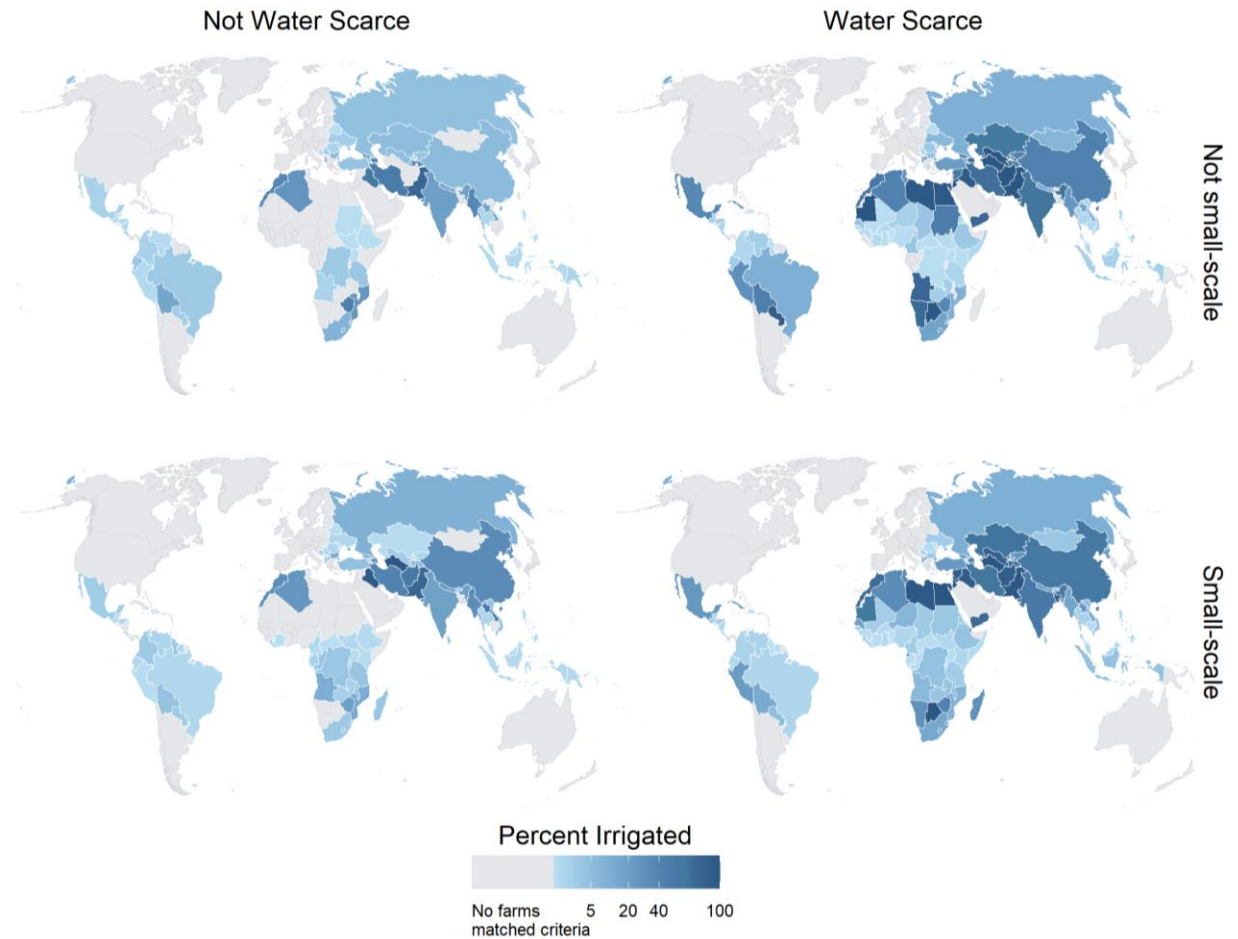
Evidence synthesis: Funding research for small-scale farmers in water scarce regions to achieve SDG 2.3

Vincent Ricciardi*, Abdrahmane Wane, Balsher Singh Sidhu, Cecile Goode, Divya Solomon, Ellen McCullough, Florian Diekmann, Jaron Porciello, Meha Jain, Nicola Randall, Zia Mehrabi

*Corresponding author

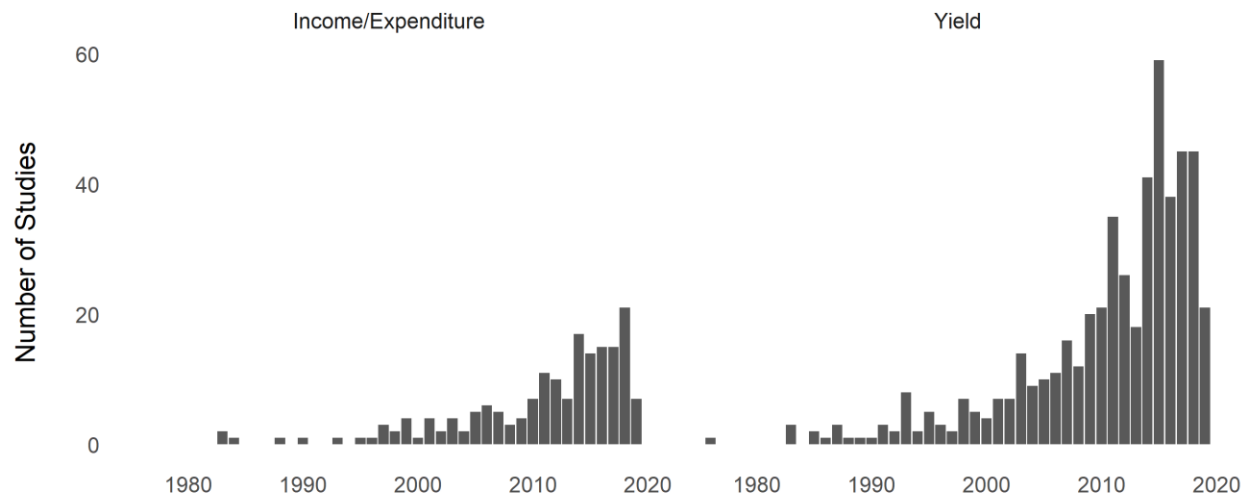
Supplementary Figures

Supplementary Figure 1: Irrigation coverage for small-scale farms in water scarce areas



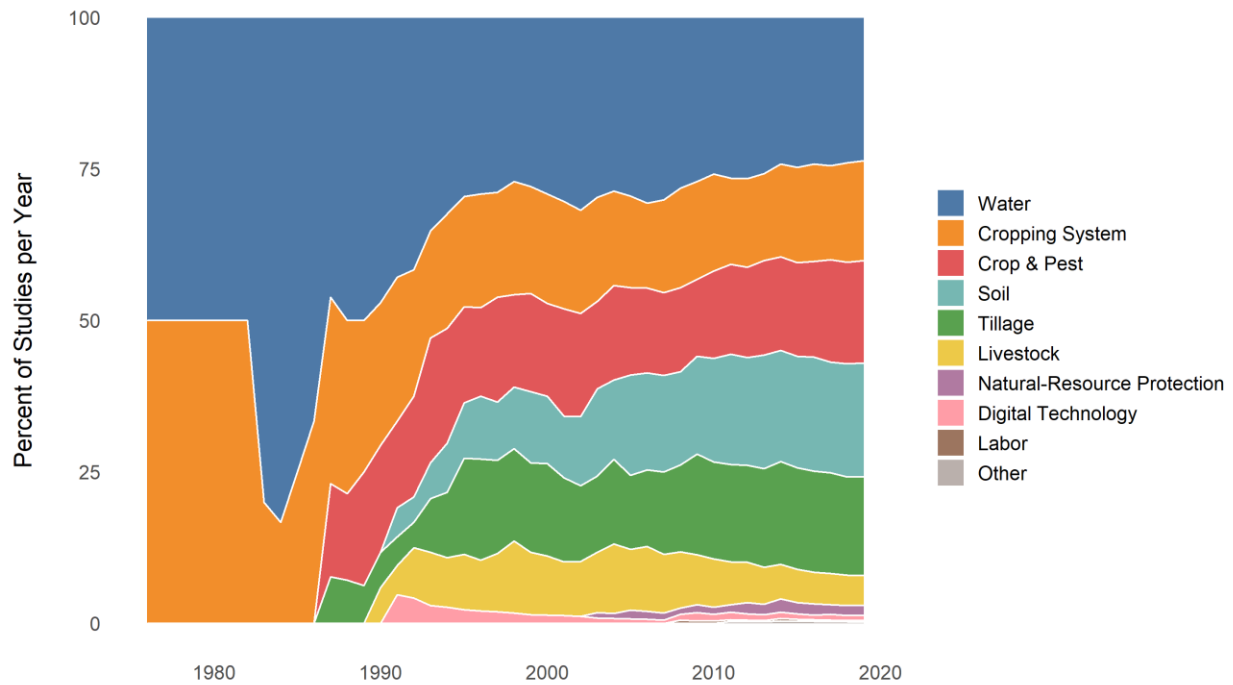
Supplementary Figure 1: Percent of farms with irrigation coverage by country in water scarce or non-water scarce regions. Small-scale and non-small-scale farms are shown. Grey indicates no farms matched our definition of small-scale farm in that country, most often due to the country being high income or having low amounts of non-water scarce regions that contained farms.

Supplementary Figure 2: Number of studies examining interventions effects of incomes and/or yields over time



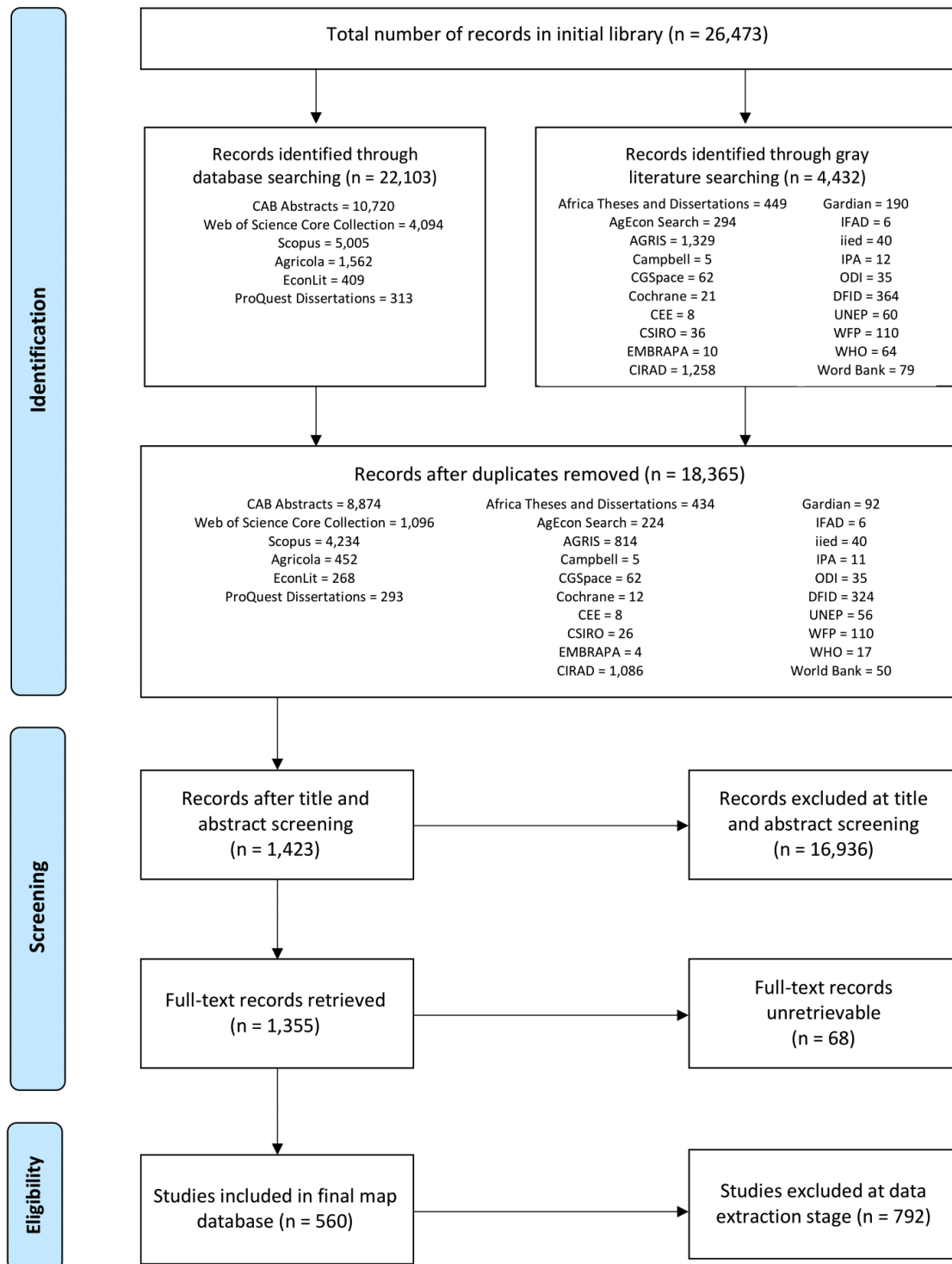
Supplementary Figure 2: Number of studies examining interventions' effects on small-scale farmers' incomes/expenditures and yields over time.

Supplementary Figure 3: Studies over time by intervention type



Supplementary Figure 3: Percent of studies per year by type of intervention.

Supplementary Figure 4: PRISMA diagram detailing the evidence review process



Supplementary Figure 4: PRISMA diagram detailing the number of studies included and excluded at each stage of the study.

Supplementary Tables

Supplementary Table 1: Examples of each intervention category.

Intervention Group	Description
Tillage	Includes all variants of tillage (e.g., conservation tillage, no-till, plough, etc.)
Natural Resource Protection	Includes protection of farm natural resources (e.g., erosion control, buffer strips, etc.)
Cropping System	Includes crop rotation, intercropping, fallow, monocropping etc.
Soil	Includes direct soil-based interventions (e.g., fertilizer, compost, liming, manure, etc.)
Crop & Pest	Includes crop species and variety-based interventions (e.g., landraces, new crop species, varietal development) and pest management as these can be overlapping interventions.
Livestock	Any intervention directly related to livestock (e.g., species feed, vaccinations, etc.)
Water	Any intervention directly related to water (e.g., irrigation and water harvesting, but not improving soil moisture content.)
Digital Technology	Any intervention dealing with digital interventions (e.g., mobile, satellite, weather advisory, automation, precision agriculture, sensors, etc.)

Supplementary Table 2: Descriptive statistics for text classification training data

Metric	Training Data
Count	1500
Mean	0.15
Std.	0.36
Min	0
Max	1

Supplementary Table 2: Descriptive statistics for the training and testing dataset for the NLP models used to assist the screening of titles and abstracts.

Supplementary Table 3: NLP model outputs to assist the screening of titles and abstracts

Model	Accuracy	Precision	Recall	F1 Score
BERT	0.877	0.792	0.877	0.830
SVM	0.850	0.875	0.850	0.836
Naive Bayes	0.831	0.969	0.831	0.832
Inter-Rater Agreement Score	0.819	-	-	-

Supplementary Table 3: For precision, recall, and the F1 score, the w8ed average between classes (i.e., include or exclude) were taken. All metrics were k-folds cross validated (k=10). All NLP models outperformed the inter-rater score, where the manual classification was performed by two researchers and their level of agreement was 81.9%.

What spectrum of farm-level interventions to alleviate water scarcity have been tested to increase small-scale farmers' yields, incomes, and productivity?

A systematic map protocol

Administrative Information

1. Identification: Protocol of a systematic map

2. Open Science Framework registration: <https://osf.io/efk7b/>

3. Authors

- Abdrahmane Wane, CIRAD-ILRI, A.Wane@cgiar.org
- Balsher Singh Sidhu, University of British Columbia, balsher.sidhu@gmail.com
- Cecile Goode, CSIRO, cecile.godde@csiro.au
- Ellen McCullough, University of Georgia, emccullough@uga.edu
- Florian Diekmann, The Ohio State University, diekmann.4@osu.edu
- Jaron Porciello, Cornell University, jat264@cornell.edu
- Meha Jain, University of Michigan, mehajain@umich.edu
- Nicola Randall, Harper Adams University, nrandall@harper-adams.ac.uk
- Tilahun Amede, ICRISAT, T.Amede@cgiar.org
- Timothy Schoepke, National Agricultural University, Timothy.Schoepke@ars.usda.gov
- Vinny Ricciardi*, University of British Columbia, vinnyricciardi@gmail.com
- Zia Mehrabi*, University of British Columbia, zia.mehrabi@ubc.ca

**Corresponding authors' physical address: Room 201, Liu Institute for Global Issues, University of British Columbia, Vancouver, BC V6T 1Z2 Canada*

4. Roles and responsibilities

Content expertise: Our team has extensive research expertise regarding small-scale farmers. Our team represents a diverse array of researchers from a variety of backgrounds, which includes: economists, ecologists, geographers, food system scientists, big data experts, synthesis researchers, and process-based modelers.

Systematic review methods: Our team includes researchers who have published meta-analyses, systematic reviews, and systematic maps on topics relating to small-scale farmers and sustainable development.

Statistical analysis: Most members of our team have substantial advanced degree training in statistical methods across a wide array of academic disciplines. We will leverage this collective knowledge to build machine learning models to assist in our title and abstract screening process as well as in the interpretation of technical literature.

Information retrieval: Our team includes two information specialists associated with the Ohio State University Food, Agricultural, and Environmental Sciences Library and the USDA National Agricultural Library.

Background

This protocol outlines the data collection, coding, and analysis for a systematic map. The systematic map will collate gray and peer-reviewed literature to identify the spectrum of smallholder interventions for alleviating water-stress and water-scarcity, the measurement of incomes, yields, and productivity for these interventions, and whether these have been linked to gender and environmental impacts.

The 2015 Sustainable Development Goals (SDG) represents a wide range of donor and government stakeholders that are committed to improve the incomes and productivity of small-scale farms. SDG target 2.3 was specifically designed to focus donor funding and programmatic efforts towards supporting small-scale farmers: *“by 2030, double the agricultural productivity and incomes of small-scale food producers, in particular women, indigenous peoples, family farmers, pastoralists and fishers, including through secure and equal access to land, other productive resources and inputs, knowledge, financial services, markets and opportunities for value addition and nonfarm employment”* (UN, 2015).

A key hurdle to achieving SDG 2.3 exists for farmers living in water-scarce regions. To improve smallholders' capacity to adapt to water-scarcity a large array of on and off-farm initiatives have been implemented across the Global South. These initiatives range from improving market access to downscaling irrigation technologies. There is a growing number of academic and non-academic programmatic assessments of these possible solutions. The rising tide of publications presents a challenge for donors and decision makers to identify and target appropriate interventions with a proven track-record of improving smallholder farmer livelihoods.

We aim to construct a systematic map to identify the range of interventions for small-scale farmers in water stressed/scarce regions. Our systematic map will attempt to synthesize the spectrum of literature that has tested the effectiveness of on-farm interventions for improving smallholders' yields, incomes, and productivity in water scarce regions. Interventions that improve smallholder welfare can range across, plot level technologies, farm-level management,

government infrastructure projects, and bi/multilateral trade agreements. We only focus on farm level interventions to generate a toolbox of actions that can be taken on the farm to tackle water scarcity directly.

Our research has three goals. First, we will identify the range of interventions that tackle water scarcity examined with respect to smallholders' incomes, productivity. Second, we will document the characteristics of the evidence, such as the experimental design, regions and time of studies. Third, we will examine if the available evidence also assessed SDG 2.3's cross cutting themes of gender and environmental impacts. These results will enable donors and decision makers to understand the geographic and socio-political contexts that on-farm interventions for tackling water scarcity have been tested, and whether they have also been assessed with respect to gender and environmental impacts. In addition, these results will help researchers identify interventions that may be suitable for an in-depth meta-analysis, systematic review, or costing exercises.

This protocol was written in references to the PRISMA and ROSES systematic review/mapping guidelines (Haddaway et al., 2018). In this protocol, we detail our review's: objectives, key definitions, research questions, eligibility criteria, information sources, search strategy, data management, and data extraction. The purpose of this protocol is to clearly define our search and coding process before conducting the review. Any amendments to this protocol will be documented for transparency.

Methods

1. Objectives

Population	- Small-scale food producers in rural areas - Farming systems in water stressed/scarce areas
Intervention	- Farming methods, technologies, or innovations that are implemented on-farm and explicitly are used to alleviate water scarcity
Comparison	- A control case (temporal, or spatial) for identifying the outcomes in the absence of the intervention (this can be from an RCT, pre-post, post-post etc), or a comparison between alternative interventions.
Outcome	- Studies that evaluate an on-farm intervention's effect on yields, productivity, and/or incomes. And concurrently study gender and environmental impacts of that intervention.

2. Definitions

- **Small-scale farmer/Smallholders:** Small-scale farmers and smallholders are synonyms. We are working with an FAO definition (Khalil et al., 2017), which we

have adapted slightly. For inclusion in the evidence synthesis, a small-scale food producer will need to meet two of the four mentioned dimensions of scale.

Definitions of “small-scale food producers” that are found in the literature are mostly based on four criteria: land size, labor input (especially of family members), market orientation and economic size. Land size is the most commonly used criterion, as the vast majority of “small-scale food producers” definitions are based on the physical size of the farm and the number of livestock heads. The second main criterion is the labor input of the farm. A third criterion is the extent of market orientation, which refers to the destination of the output of the farm, either for own-final consumption or for sale and/or barter in markets. A fourth criterion is economic size.

- **Water stress/scarcity:** Our study will take a general definition of water stress/scarcity to include a broad range of studies and interventions across a spectrum of agricultural regions. For our study, water scarcity is when there is not enough water to be used by a farmer for agricultural purposes. This includes blue water (e.g., freshwater) and green water (e.g., water retained by the soil).
- **Farming system:** We will be examining farming systems that produce crops and/or livestock. We will not explicitly examine fisheries or other forestry systems, as the range of interventions extending past our reviews’ bandwidth.
- **Interventions:** Farming methods, technologies, innovations that are implemented on farm.
- **On-farm:** It is the area operated for production by the farmer or farming household.
- **Observational Study:** Non experimental studies, e.g. a study that uses household surveys, or non-experimental plot based measurements.
- **Field trials:** A study that uses tests plots to measure different outcomes between interventions and controls.
- **Yield:** For crops, the yield is the weight of crop divided by the cultivated area. If cultivated area is not provided, we will include, in order, harvested, operated, agricultural, or owned area. For livestock, the yield is the amount of meat, milk, or egg, produced by herd/brood/drove size or Tropical Livestock Units (TLU are a standardized herd size where different types of animals can be compared).
- **Productivity:** “Land productivity is defined as the value of agricultural output per hectare of land under cultivation, land temporarily left fallow and permanent meadows, while labour productivity is defined as the value of agricultural output per worker” (Mikecz et al., 2016). This study will only be examining land productivity for crop systems and an equivalent definition for livestock systems (e.g. the value of output per herd/brood/drove susze or TLU).
- **Income:** Household income is the sum of all receipts, whether monetary or in kind (goods and services), received by the household or by individual members of the household, including remittances from members of the household that have migrated. A rural household may derive income from on-farm and off-farm activities, as well as from transfer programmes, such as a pension or child support benefit; we only include on-farm income streams. Income also includes the value of own-production (or foraging), valued at market prices. Household

income excludes windfall gains and other irregular and typically one-time receipts. Household income is measured over a fixed period of time (typically annual) (Adapted from UN 2011).

3. Research question

Primary Question

1. What spectrum of farm-level interventions to alleviate water scarcity have been tested to increase small-scale farmers' incomes, yields, and productivity?

Secondary Questions

1. Which regions have available evidence?
2. Has the evidence examined SDG 2.3's cross cutting themes of gender, and environmental impacts?
3. What is the quality of the evidence?

4. Eligibility criteria

For an article to be included in this study, it must meet all of the following inclusion criteria:

1. Studies explicitly addressing smallholder farmers.
2. Studies examining on-farm production management techniques or technologies explicitly addressing water scarcity, drought adaptation, or water efficiency adaptation.
3. Studies examining an on-farm intervention's effect on yields, productivity, or incomes.
4. Studies assessments need to either include a control case for comparison (temporal, or spatial) for identifying the outcomes in the absence of the intervention (this can be from an RCT, pre-post, post-post etc.), or a comparison between alternative interventions.

For an article to be excluded from this scoping review, it must meet one of the following exclusion criteria:

1. Studies exclusively focussing on non-smallholder farms.
2. Studies that do not examine on-farm production management techniques or technologies explicitly addressing water scarcity, drought adaptation, or water efficiency adaptation.
3. Studies not examining an on-farm intervention's effect on yields, productivity, or incomes.
4. Studies without a counterfactual, or comparison.

5. Information sources

We will conduct searches of the following electronic databases:

- CAB Abstracts (CAB Direct)
- Web of Science Core Collection (Web of Science)
- Scopus
- Agricola (EBSCOhost)
- EconLit (EBSCOhost)
- ProQuest Dissertations & Theses Global (ProQuest)

A search of the following gray literature sources will also be conducted:

- Africa Theses and Dissertations (<http://datad.aau.org/>)
- AgEcon Search (<https://ageconsearch.umn.edu/collections/>)
- AGRIS (FAO-consolidated search) (<http://agris.fao.org/agris-search/index.do>)
- Campbell Collaboration (<https://campbellcollaboration.org/>)
- Cochrane Collaboration (<https://www.cochranelibrary.com/>)
- Collaboration for Environmental Evidence (CEE) (<https://www.environmentalevidence.org/>)
- Commonwealth Scientific and Industrial Research Organisation (CSIRO) (<https://publications.csiro.au/rpr/home>)
- Brazilian Agricultural Research Corporation (EMBRAPA) (<https://www.embrapa.br/biblioteca>)
- French Agricultural Research Centre for International Development (CIRAD) (<https://agritrop.cirad.fr/cgi/search/>)
- Gardian (CGIAR) (<http://gardian.bigdata.cgiar.org/search.php#!/>)
- International Fund for Agricultural Development (IFAD) (<https://www.ifad.org/en/web/knowledge/>)
- International Institute for Environment and Development (<http://pubs.iied.org/search/>)
- JPAL/ATAI impact evaluations (IPA) (<https://www.poverty-action.org/program-area/agriculture/publications>)
- Overseas Development Institute (ODI) (<https://www.odi.org/>)
- UK Department for International Development (DFID) (<https://www.gov.uk/dfid-research-outputs?>)
- United Nations Environment Programme (UNEP) (<https://wedocs.unep.org/handle/20.500.11822/1>)
- World Food Programme (WFP) (<https://www1.wfp.org/publications>)
- World Health Organization (WHO) (<https://apps.who.int/iris/>)
- World Bank (<https://openknowledge.worldbank.org/>)

In addition, citations of included studies will be searched to capture any additional relevant studies.

6. Search strategy

A comprehensive search strategy is used to identify all relevant research pertaining to on-farm interventions that have been tested to increase small-scale farmers' incomes, yields, and productivity by addressing water-stress and water-scarcity. Search terms include variations of the key concepts in the research question: small-scale producers, water stress and scarcity, and yield, income, and productivity outcomes. Appendix A presents the master search strategy developed for CAB Abstracts (accessed via the CAB Direct platform). Full details of the search terms, search queries, the number of records generated from specific searches, and name, location, and date of the search will be provided at <https://osf.io/efk7b/>.

7. Study records

Data management: Searches will be performed across all information sources listed in Section 5 of this protocol and search results will be de-duplicated to remove redundant citations identified from multiple sources. Titles, abstracts, and keywords of all citations will be exported and stored in a csv file.

The web-based software platform Covidence (<https://www.covidence.org/>) will be used for title and article screening of the processed results. All inclusion/exclusion decisions will be documented in the Covidence system and exported to a csv file. Full article coding will be conducted using Google Forms, which will feed into a central Google Sheet document. The Google Sheet document will be explored as a csv file upon full article code completion. All analysis/figure creation will take place in R. All criteria as listed in the PRISMA and ROSES guidelines will be documented and stored (Haddaway et al., 2018; Moher et al., 2009). All data will be stored in a secure cloud server (Google Drive or in Covidence).

Selection process: The goal of the screening phase is to examine each article's title and abstract to determine if the article should be discarded or should be included in the full-text coding stage. Due to the anticipated large amount of articles to be screened, we will use a two-part title and abstract screening method. First, we will have two reviewers screen a random sample of 1500 titles and abstracts. If there is disagreement, a third independent reviewer will screen the title and abstract in question and make the final decision. Early in the screening process, each disputed inclusion/exclusion title and abstracts will be discussed with the screening team to ensure consistency. We will use this random sample to calculate a human inter-rater reliability score.

Second, these manually screened titles and abstracts will be used as training (1000 articles) and test data (500 articles) for building a binary classifier that we will train and tune to predict if articles should be included in full-text coding or excluded. If the classification algorithm exceeds the human inter-rater reliability scores, we will proceed to use it solely to screen the remainder of the titles and abstracts. If it does not, we will proceed to screen an additional 500 articles, in a stepwise fashion, at each step evaluating when the model outperforms human inter-rater reliability, at which -- a process known as active learning. An example of this dual screening

procedure can be found in Hempel et al. 2012. All articles passing screening will be used in the data extraction, but will be excluded at this stage if they fail to meet the inclusion criteria. The full text assessment of inclusion, and coding will be completed by one reviewer per article.

Data collection process: The goal of the full-text article coding phase is to extract key information from each article for creating the systematic map. One independent reviewer will code the full-text of all articles that pass our previously outlined title and abstract screening process. An initial training stage will take place to ensure all co-authors understand the coding template. We will code each article for key pieces of information: geographic location, type of farming system, intervention(s), outcome, study design, cross-cutting theme, and funding source. The coding template will include bibliographic information (e.g., pre-loaded information on title, author, DOI, etc.) and the following pieces of information:

- Georeferencing (e.g., coordinates, state(s), country(s))
- Type of study (e.g., survey, systematic review, meta-analysis, experiment, mixed survey/experiment)
 - If meta-analysis, we will develop a specific coding template to assess the quality and direction of the intervention's relationship with the outcome variable examined.
- Type of farming system (e.g., crop and/or livestock)
- Intervention
- Outcome (e.g., income, productivity, yield)
- Cross cutting theme measurements (e.g., gender, environmental impacts)
- Funding details source if provided in acknowledgements

8. Critical appraisal of individual sources of evidence

While we will not be conducting an assessment of the sign of the interventions effect on outcomes, we will be conducting quality assessments of the research, as explained below.

9. Data synthesis/charting

Anticipated Manuscript Figures and Tables

Figure 1: Evidence thematic matrix

- On-farm intervention group and subgroup by outcome and cross-cutting themes, study type, and farming system is visualized in each cell

Figure 2: Panel of geographical maps of evidence

- A-B. Overall map of evidence, color shaded by study type panel by outcome dimension

- C-D. Overall map of evidence, color shaded by cross-cutting theme panel by outcome dimension

Figure 3: Panel of regional breakdowns of Figure 1

Figure 4: Evidence type and cross cutting theme over time with regional breakdowns

Figure 5. Gap analysis of mapping the available evidence for a country (no of intervention types, quality of studies) to needs (number of farmers exposed to water scarcity)

Table 1: Category, subcategory descriptions (can add detailed inclusion in SI)

Table 2: If meta-analyses of systematic reviews are available for intervention (s), provide: citation, intervention, finding, quality score, cross-cutting themes

Table 3: Funding by frequency of studies intervention type

Supplemental information:

Figure S1: Prisma diagram

Table S1: Full list of all studies and codings

References

- Campbell Library. (n.d.). Retrieved March 6, 2019, from <https://campbellcollaboration.org/library/campbell-systematic-review-templates.html>
- Haddaway, N.R., Macura, B., Whaley, P. and Pullin, A.S. (2018). ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence*, 7(1), p.7.
- Hempel, S., Shetty, K.D., Shekelle, P.G., Rubenstein, L.V., Danz, M.S., Johnsen, B. and Dalal, S.R. (2012). Machine learning methods in systematic reviews: identifying quality improvement intervention evaluations. *United States Agency for Healthcare Research and Quality: Research White Papers*.
- Khalil CA, Conforti P, Ergin I, Gennari P (2017) Defining Small-Scale Food Producers To Monitor Target 2.3. of the 2030 Agenda for Sustainable Development. *FAO Stat Work Pap Ser* 17(12). www.fao.org/3/a-i6858e.pdf
- Mikecz, O., & Vos, R. (2016). *Can Smallholders Double their Productivity and Incomes by 2030?*. Food and Agriculture Organization of the United Nations, Agricultural Development Economics Division. <http://www.fao.org/3/a-i5959e.pdf>

- Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151(4), p.264-269.
- PRISMA-P-checklist.pdf. (n.d.). Retrieved from <http://prisma-statement.org/documents/PRISMA-P-checklist.pdf>
- Shamseer, L., Moher, D., Clarke, M., Gherzi, D., Liberati, A., Petticrew, M., ... Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, 349, p7647. <https://doi.org/10.1136/bmj.g7647>
- Sun, S. (2011). Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology*, 11(3-4), p145-163.
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, 169(7), 467. <https://doi.org/10.7326/M18-0850>
- United Nations (UN) (2015) Sustainable Development Goals (SDGs). Retrieved January 1, 2017, from: <https://sustainabledevelopment.un.org/topics/sustainabledevelopmentgoals#>.
- United Nations (UN) Economic Commission for Europe (2011). Canberra Group Handbook on Household Income Statistics. 2nd edition, United Nations: New York
- World Bank. World Bank Country and Lending Groups. Retrieved May 29, 2019, from <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519>

Appendix A

CAB Abstracts (CAB Direct)

Date of Search: 5/23/2019

Number of results: 10,720

Row #	Search string
1	(agropastoral* OR agro-pastoral* OR smallhold* OR small-hold* OR small-farm* OR microfarm* OR micro-farm OR pastoral* OR "family-run farm*" OR "family-owned farm*" OR "family-managed farm*") OR ((small-scale* OR smallscale* OR low-income* OR subsistence* OR semi-subsistence* OR poor* OR resource-limited* OR small-size* OR peasant* OR poverty*) near/3 (farm* OR mixed-farm* OR agricult* OR producer* OR grower*))
2	de:(smallholders OR "small farms" OR "marginal farms" OR "family farms")
3	1 OR 2
4	arid* OR "desert farm*" OR desert* OR drought OR "dry condition*" OR "dry* farm*" OR dry-land* OR dryland* OR evapotranspiration OR "field capacity" OR groundwater OR "integrated water management" OR precipitation OR rain* OR rainfall* OR "rain-fall*" OR rainfed OR rain-fed OR "rainwater harvest*" OR run-off OR savanna* OR "scarce water" OR "seasonal variability" OR semiarid* OR semi-arid* OR steppe* OR water* OR "water conservation" OR "water deficit" OR "water deprivation" OR water-scarc*
5	de:("arid lands" OR "arid zones" OR deserts OR drought OR "dry conditions" OR "dry farming" OR evapotranspiration OR "moisture relations" OR "plant water relations" OR rainfall OR savannas OR "semiarid grasslands" OR "semiarid zones" OR "soil moisture" OR "soil water" OR "soil water content" OR steppes OR "stress conditions" OR water OR "water availability" OR "water conservation" OR "water deficit" OR "water deprivation" OR "water flow" OR "water management" OR "water potential" OR "water resource management" OR "water resources" OR "water stress" OR "water supply" OR "water table" OR "water use" OR "water use efficiency")
6	4 OR 5
7	intensification* OR "agricult* productivity" OR "capital productivity" OR "cost benefit analysis" OR cost-benefit-analysis OR earning* OR "economic benefit*" OR "economic productivity" OR "economic return*" OR efficienc* OR extensification* OR "farm* income*" OR "household* income*" OR

	income* OR "income* earned" OR "labor productivity" OR "labour productivity" OR "land productivity" OR "net income*" OR "net return*" OR outcome* OR output* OR outturn* OR performance* OR "production cost*" OR "production economic*" OR productivity OR profit* OR return* OR revenue* OR wage* OR yield*
8	de:("capital productivity" OR "cash flow" OR "cost analysis" OR "cost benefit analysis" OR costs OR "crop yield" OR "earned income" OR "economic analysis" OR "economic evaluation" OR "economic impact" OR economics OR "extensive farming" OR "extensive livestock farming" OR "extensive production" OR "farm income" OR "farm results" OR "farmers' income" OR "financial yields" OR "gross margins" OR "household income" OR income OR "intensive cropping" OR "intensive farming" OR "intensive husbandry" OR "intensive livestock farming" OR "intensive production" OR "land productivity" OR "marginal returns" OR "maximum yield" OR "non-farm income" OR performance OR "production costs" OR productivity OR profitability OR profits OR returns OR "variable costs" OR "yield increases" OR "yield losses" OR yields)
9	7 OR 8
10	3 AND 6 AND 9
11	Refined by: Your Products: CAB ABSTRACTS OR CAB ABSTRACTS Archive OR Global Health OR Global Health Archive

Default Question Block

What is your name?

What is the study ID?

What is the study title?

Does the study meet the inclusion and exclusion criteria? Please see [this link](#) for details.

☐ Yes

☐ No

Was the study done at a global, regional, or sub/national scale?

☐ Global

☐ Regional

☐ Sub/National

Which region(s)?

☐ Northern Africa

☐ Eastern Africa

☐ Middle Africa

☐ Southern Africa

☐ Western Africa

☐ Caribbean

☐ Central America

☐ South America

☐ Northern America

☐ Central Asia

- ☐ Eastern Asia
- ☐ South-eastern Asia
- ☐ Southern Asia
- ☐ Western Asia
- ☐ Eastern Europe
- ☐ Northern Europe
- ☐ Southern Europe
- ☐ Western Europe
- ☐ Australia & New Zealand
- ☐ Melanesia
- ☐ Micronesia
- ☐ Polynesia
- ☐ Other

Which nation(s)?

Afghanistan
Albania
Algeria
Angola
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan

What was the type of farming system?

- ☐ Crop
- ☐ Livestock
- ☐ Mixed
- ☐ Other

What was the type of study? Please see this [link](#) for detailed keywords associated with each study type.

- ☐ Observational, primary data

- ☐ Experimental, primary data
- ☐ Modeling with secondary data
- ☐ Meta-analysis
- ☐ Systematic Review
- ☐ Systematic Map
- ☐ Other

Please select what interventions were examined in this study. Please see this [link](#) for detailed keywords associated with each intervention.

- ☐ Tillage
- ☐ Conservation
- ☐ Cropping system
- ☐ Soil
- ☐ Crop & pest
- ☐ Livestock
- ☐ Water
- ☐ Digital technology
- ☐ Labor
- ☐ Other

Please select which outcomes were measured in this study. Please see this [link](#) for detailed keywords associated with each intervention.

- ☐ Crop yield
- ☐ Livestock yield
- ☐ Livestock products yield
- ☐ Income/expenditure
- ☐ Crop productivity
- ☐ Livestock productivity
- ☐ Other

Were the same outcomes measured for each intervention?

- ☐ Yes
- ☐ No
- ☐ Other

Please write out which outcome(s) was examined for each intervention (please follow the format: intervention1, outcome1; intervention2, outcome2; e.g., tillage, yield; tillage, income).

Please select which environmental impacts were measured (if any) in this study. Please see this [link](#) for detailed keywords associated with each environmental impact.

- ☐ Soil health
- ☐ Water quantity
- ☐ Water quality
- ☐ Climate change
- ☐ Biodiversity
- ☐ Air quality
- ☐ Energy
- ☐ Waste
- ☐ Land cover conversion
- ☐ Environmental
- ☐ Other

Were any gender impacts measured in this study? Please see this [link](#) for detailed keywords associated with gender impacts.

- ☐ Yes
- ☐ No
- ☐ Other

Powered by Qualtrics

The Collaboration for Environmental Evidence Synthesis Appraisal Tool (CEESAT)

Version 2

The CEESAT checklist provides a point by point appraisal of the confidence that can be placed in the findings of an evidence review by assessing the rigour of the methods used by the review, the transparency with which those methods are reported and the limitations imposed on synthesis by the quantity and quality of available primary data.

Note that CEESAT does not distinguish between reviews that do not employ methodology that reduces susceptibility to bias and increases reliability of findings and reviews that may have employed such methodology but do not report it.

THE CEESAT CHECKLIST

Tick or fill one ‘O’ for each question. Start with the ‘Gold’ definition and move down the list if it does not apply. When ‘AND’ is used all statements must apply and when ‘OR’ is used one of the alternative should apply.

Please make a single decision for each question (or hypothesis). Many reviews address more than one question. This may be reflected in the review title being broad whilst subsections address more specifically defined questions. If this is the case please complete an assessment for each question and fill in multiple rows for the article on the assessment spreadsheet.

Please note that supplementary material to (and linked from) the article, including protocols, should be included in the assessment and are provided when noted by the Editorial Team. Please contact us if you have problems accessing supplementary material. Reference to methods of other original articles should not be included.

Please cite as: Collaboration for Environmental Evidence 2018. The Collaboration for Environmental Evidence Synthesis Appraisal Tool (CEESAT). Version 2.

1. THE REVIEW QUESTION

Rationale: A well-defined question (or hypothesis) is crucial for assessing the reliability of subsequent decisions on searching and screening for eligible studies, as well as forming the basis for critical appraisal of study conduct and for data extraction and synthesis.

1.1 Are the elements of the review question clear?

Tick whichever statement applies

- ☐ Gold: The review question or hypothesis is clearly stated and clearly defines key elements, (e.g. PICO, PECO, PO, PIT, etc.) correctly, such as the subject or population of interest, the intervention or exposure type, the comparator and valid measures of outcome.
- ☐ Green: The review question or hypothesis is clearly stated, and key-elements are mentioned although not formally defined in terms of PICO, PECO, PO, PIT, etc.
- ☐ Amber: The question or hypothesis is stated in broad terms but key-elements are unclear or poorly defined.
OR
Question or hypothesis not stated but problem or issue is stated such that a question can be inferred
- ☐ Red: A question, hypothesis or problem is not stated
OR
There is no stated objective to provide an answer to a question or test of a hypothesis.

2. THE METHOD/PROTOCOL

Rationale: A protocol is a document describing the methods to be used, produced prior to the commencement of an evidence synthesis. It describes the background to the synthesis, the questions, the strategy that will be used to search for primary research articles, and the criteria for deciding whether or not an article is then relevant to include in the synthesis. The protocol should also outline the approach to assessing the quality of each included study, and to extracting and synthesising data from primary research studies. Writing a protocol is therefore analogous with developing and documenting a methodology prior to conducting fieldwork or experiments and is similarly integral to producing a study that is robust against post hoc changes in methods and scope.

2.1 Is there an a-priori method/protocol document?

- Gold: The review cites a separate a-priori protocol or documented pre-defined method containing details of proposed conduct of all review and synthesis stages (e.g. question, search, eligibility screening, critical appraisal, data extraction and synthesis)
AND
It is linked from the synthesis (e.g. as supplementary material or hosted on a separate website)
AND
It was publicly accessible prior to the conduct of the review
AND
It was submitted to an independent body for peer review and publication.
- Green: The review cites a separate a-priori protocol or documented pre-defined method containing details of conduct of all review and synthesis stages (e.g. question, search, screening, critical appraisal, data extraction and synthesis)
AND
It is linked from the synthesis (e.g. as supplementary material or hosted on a separate website)
AND
It was publicly accessible prior to the conduct of the review.
- Amber: The review cites an a-priori protocol or documented pre-defined method, but this does not contain all details of conduct of all review and synthesis stages or was not publicly accessible prior to the conduct of the review
OR
The review includes a defined methods section (not a-priori) listing the synthesis stages conducted and providing sufficient detail to enable the method to be repeated.
- Red: There is no protocol and the review methods are not clearly defined in the methods section of the review or there are no methods reported.

3. SEARCHING FOR STUDIES

An optimal search for literature should possess three key properties: comprehensive (maximises the number of potentially relevant studies found), systematic (avoiding ad hoc search strategies reduces the susceptibility to bias resulting from e.g. no defined endpoint of search) and transparent (readers should be able to repeat and evaluate the search).

3.1. Is the approach to searching clearly defined, systematic and transparent?

Rationale: Search strategies should be outlined in the predefined protocol or review methods. An optimal search for literature should aim to maximise comprehensiveness (aiming to identify all relevant studies) and transparency (readers should be able to repeat and evaluate the search).

This is to avoid ‘cherry-picking’ studies or assembling a biased or unrepresentative body of evidence. Where possible, advice should be sought from an expert such as an information specialist/scientist.

- Gold: All search terms, Boolean operators (‘AND’, ‘OR’ etc.) and wildcards are clearly stated so that the exact search is repeatable by a third party
AND
There is information about the sources searched, together with dates of search and any limitations justified (e.g. language, publication date, no grey literature searches)
- Green: All search terms, Boolean operators (‘AND’, ‘OR’ etc.) and wildcards are clearly stated so that the exact search is repeatable by a third party
AND
There is information about the sources searched, together with dates of search [but no limitations justified (e.g. language, or publication date, no grey literature searches)
- Amber: The search is described but not adequately to be fully repeatable by a third party either because the specific search terms are not stated or Boolean operators/wildcards are not stated (so it is unclear how the search terms are combined).
OR
There is information about the databases searched, but dates of search not given and no limitations justified (e.g. language, or publication date, no grey literature searches)
.
- Red: No information regarding the search strategy used.

3.2. *Is the search comprehensive?*

Rationale: The resources used to find relevant literature influence the comprehensiveness and reliability of the synthesis. The principal sources for locating peer-reviewed articles are electronic databases of scientific literature and academic search engines, with a range of supplementary methods. No single database indexes all peer-reviewed articles. Moreover, these sources are unlikely to capture potentially relevant grey literature (e.g., reports by governmental and non-governmental organisations, unpublished studies) and consequently can be complemented by searching thesis repositories, websites of relevant organisations and conducting internet searches. Other supplementary search strategies include citation chasing (backwards and forwards), and contact with experts in the field.

- Gold: Sources of articles searched capture both conventionally published scientific literature and grey literature using a combination of databases, search engines and specialist websites (may also be informed by stakeholders) or limitations are fully justified.

AND

Comprehensiveness of search is demonstrated by a series of tests using independent samples of the relevant literature to demonstrate adequate sensitivity.

NB. Statements such as '*We considered only peer-reviewed material because this is more reliable than grey literature*' without evidence that the methodological quality of potentially relevant grey literature was assessed do not indicate that grey literature was objectively considered.

- Green: Sources of articles searched capture both conventionally published scientific literature and grey literature using a combination of databases, search engines and specialist websites (may also be informed by stakeholders) or limitations are fully justified.

NB. Statements such as '*We considered only peer-reviewed material because this is more reliable than grey literature*' without evidence that the methodological quality of potentially relevant grey literature was assessed do not indicate that grey literature was objectively considered.

- Amber: Resources used are stated but limited, without justification, to conventionally published scientific literature or just one or two sources.
- Red: Resources used not stated or search is not systematic (i.e. studies appear to have been selected).

4. INCLUDING STUDIES

Comprehensive searches may generate a large number of articles that vary widely in their relevance to the synthesis. Authors must then determine whether or not each article is sufficiently relevant (eligible) for inclusion in the data synthesis stage. However, the choice of eligibility criteria can influence the conclusions of the synthesis, and the application of inadequately defined criteria can be subjective and lead to biases. Decisions over which studies are relevant for inclusion should therefore be based on clearly defined criteria, and should be repeatable and transparent. Criteria 4.1-4.3 refer only to studies included/excluded on the basis of relevance to the review question – see 5 for inclusion/exclusion on the basis of methodological quality.

4.1: Are eligibility criteria clearly defined?

Rationale: Clearly stated criteria for eligibility decisions minimise the potential for subjective decisions to influence which studies are included in the review, increase the transparency of the synthesis, and allow readers to assess the validity of the criteria to the review question. In

addition to following the review question, eligibility criteria may define limits on the type of primary research to be considered in terms of (for example): geographic scope, type of data reported, type of intervention or impact, study design, date.

- Gold: Eligibility criteria are precisely defined (e.g. reliance on broad and potentially ambiguous terms should be avoided) and expressly related to each key element of the question (other criteria such as study design may also be considered)
AND
Criteria are consistent between a-priori protocol and review or differences are fully explained.
- Green: Eligibility criteria are precisely defined (e.g. reliance on broad and potentially ambiguous terms should be avoided) and all key elements of the question considered (other criteria such as study design may also be considered)
- Amber: The questions/scope/objectives of the review are stated such that the type of primary research articles/studies to be included are broadly apparent, but the review does not explicitly identify criteria to be systematically applied to all articles found during the search.
OR
Eligibility criteria are defined but no clear relationship to a review question (possibly because the question is poorly defined).
- Red: No to both amber criteria above (eligibility criteria are not stated).

4.2 Are eligibility criteria consistently applied to all potentially relevant articles and studies found during the search?

Rationale: More than one person should screen studies for inclusion to reduce the risk of human error and to ensure that the criteria are applied consistently to address the review question. If more than one person independently evaluates the relevance of the same articles, the repeatability of inclusion/exclusion decisions can be assessed. Piloting the criteria, and discussing and refining the eligibility decisions can also ensure they are consistently and meaningfully applied.

- Gold: The eligibility criteria are independently applied by more than one reviewer to all of the screened articles/studies (at title screening stage, pragmatic decisions about dual screening of subsamples is justified e.g. because large numbers of titles were screened)

AND

Repeatability of eligibility decisions was measured and reported and all disagreements between reviewers discussed so that the resolutions informed subsequent assessments.

- Green: The eligibility criteria are independently applied by more than one reviewer to a sample of justified size of the screened articles/studies at title, abstract and full text.

AND

Repeatability of eligibility decisions was measured and reported and all disagreements between reviewers discussed so that the resolutions informed subsequent assessments.

- Amber: The eligibility criteria are applied by more than one reviewer to a sample of the screened articles/studies at abstract and full text but reviewer independence is uncertain (i.e. not reported) or absent.

AND

Repeatability of eligibility decisions was measured and reported and all disagreements between reviewers discussed so that the resolutions informed subsequent assessments.

- Red: Number of reviewers not reported

OR

Only one reviewer applied criteria at abstract or full text stage,

OR

Where two reviewers, repeatability decisions not tested/reported

OR

No eligibility criteria provided (see 4.1)

4.3 Are eligibility decisions transparently reported?

Rationale: Listing all articles that were screened for eligibility and indicating whether each was included or excluded in data synthesis (usually as supplementary material), makes it clear whether potentially relevant studies have been omitted according to the eligibility criteria or were not captured by the search. Documenting the reasons for article exclusion at full text is essential for transparency.

- Gold: The number of unique articles found during the searches (after removal of duplicates) is presented

AND

The number excluded at each stage of the screening process is fully presented (e.g. in a flow diagram or table)

AND

Reasons for exclusion of each article/study considered at full-text are presented (e.g. in an appendix)

AND

A list of eligible (included) articles/studies is presented (not just included in reference list).

Green: The number of articles excluded at each stage of the screening process is reported but some aspects missing (e.g. number of unique articles or articles unobtainable)

AND

Reasons for exclusion of each article/study considered at full-text are presented (e.g. in an appendix)

AND

A list of eligible (included) articles/studies is presented (not just included in reference list).

- Amber: The number of articles excluded during the screening process is reported (or inferable) but some aspects missing (e.g. number of unique articles or articles unobtainable)

AND

A list of eligible (included) articles/studies is presented (not just included in reference list).

- Red: No to either or both of the amber criteria above

5. CRITICAL APPRAISAL

Primary research can vary widely in methodological validity (internal validity) and study context (external validity). Internal validity can influence the findings of the research, and, if not properly accounted for, the conclusions of syntheses that use it. External validity can influence the relevance/applicability of the study to users of the findings in individual contexts. Critical appraisal involves transparently evaluating the design and conduct of each included study based on methodologies can then help to objectively account for variation in study quality by placing greater emphasis on the most reliable studies.

5.1 Does the review critically appraise each study?

Rationale: Documented critical appraisal of each included study using relevant, pre-defined critical appraisal criteria allows the author(s) of the synthesis and the reader to make more objective assessments of the relative reliability (or weighting) of each study.

Some potentially relevant studies may not meet baseline methodological requirements (e.g. small sample size, pseudoreplication, spatial autocorrelation, lack of appropriate controls etc.) and so may be excluded from the synthesis. Effectively, these studies are weighted as 'zero'. Studies included in the synthesis may be treated differently according to the rigour of the sampling design, according to differences in sampling effectiveness (e.g. sample size, sampling area, study duration, etc.), or according to their generalisability for the synthesis in hand (e.g. spatial scale, study setting, etc.).

- Gold: An effort is made to identify relevant sources of bias (threats to internal and external validity)
AND
Each type of bias (threat to internal and external validity) is assessed individually for all included studies
AND
Results are reported using an a-priori defined (in protocol) critical appraisal sheet.

NB. This does not include syntheses in which the design and conduct of each study are stated but validity is not explicitly considered (no critical appraisal), or in which methodological rigour is discussed without transparent and objective assessments for each study.

- Green: An effort is made to identify relevant sources of bias (threats to internal and external validity)
AND
Each type of bias or threat to internal and external validity is assessed individually for all included studies and reported on a critical appraisal sheet.

NB. This does not include syntheses in which the design and conduct of each study are stated but validity is not explicitly considered (no critical appraisal), or in which methodological rigour is discussed without transparent and objective assessments for each study.

- Amber: Some risks of bias (threats to internal and external validity) are examined for all included studies but not reported for individual studies.
- Red: No critical appraisal conducted
OR
all critical appraisal criteria not applied to all included studies. This may include syntheses in which the methods for each study are stated but validity is not explicitly considered, or in which methodological rigour is discussed without transparent and objective assessments for each study.

5.2. During critical appraisal was an effort made to minimise subjectivity?

Rationale: Use of explicit appraisal criteria that are developed in advance of the process avoids post hoc decisions being made about study validity and changes to appraisal criteria during the assessment process. Employing more than one reviewer reduces susceptibility to bias decisions or errors in judgement.

- Gold: An effort is made to minimise subjectivity by predefining a critical appraisal process in a protocol
AND
At least two people independently critically appraised each study with disagreements and process of resolution reported.
- Green: An effort is made to minimise subjectivity by predefining critical appraisal process in a protocol
AND
At least two people critically appraised each study but not independently (e.g. second person aware of first person's decision) OR a subset of studies was appraised by at least two people independently and disagreements and process of resolution reported.
- Amber: At least two people critically appraised each study but not independently (e.g. second person aware of first person's decision) OR A subset of studies was appraised by at least two people independently? Disagreements and process of resolution MAY NOT be reported.
- Red: No to Amber above (e.g. Only one person critically appraised each study or number not reported)
OR
No critical appraisal conducted (i.e. RED for 5.1 above).

6. DATA EXTRACTION

The volume and type of data collected by primary research articles varies substantially, even when similar questions are addressed. Authors of evidence syntheses must make decisions on which data to extract and on how to extract this information. These decisions may influence the findings of the synthesis, and so to minimise bias, the approach to data extraction should be clearly stated and, wherever possible, the extracted information should be comparable and consistent between studies.

6.1. Is the method of data extraction fully documented?

Rationale: Transparently identifying a consistent set of data to extract from each study, for example into a structured data extraction sheet, allows the process to be repeated and evaluated by a third party, and reduces the potential for bias over which data are extracted from individual

studies. Typically, extracted information from each study included in the review comprises: study aims; intervention details, study design; population characteristics; comparator details and results (point estimates and measures of variance).

- Gold: The authors state in an a-priori protocol the type of data to be extracted
AND
The methods by which data from each study will be extracted so that the process can be repeated and confirm these methods were ultimately used in their report or reasons for deviation.
- Green: The authors state in the methods the type of data to be extracted
AND
The methods by which data from each study were extracted so that the process can be repeated (In some cases methods may have been partially reported in an a-priori protocol but then modified or substantially developed during the review process).
- Amber: The authors state in the methods the type of data to be extracted
AND
Although the review does not provide a fully repeatable methodology for data extraction, it is possible to infer the broad method from the reported results (e.g. a table that lists all eligible studies and data extracted might be included).
- Red: No to either part of amber above. It is not clear what data were selected for extraction and/or no consistent approach to data extraction is reported.

6.2. *Are the extracted data reported for each study?*

Rationale: Providing a summary in which the population, intervention/exposure and outcome for each study are stated makes data extraction transparent, and makes it easier for readers to locate the most relevant primary literature and conduct supplementary analyses if required (Liberati et al, 2009).

- Gold: All data selected for extraction are provided in a table or spreadsheet as set out in the a-priori protocol (this includes the data used in the synthesis for each study (e.g. outcome metrics or effect size, population, intervention etc) and meta-data)
AND
calculations or transformation of data by review authors using extracted data (e.g. effect sizes, averaging over variables summarisation of themes, coding, etc.) are reported in full and therefore repeatable.

- Green: All data selected for extraction provided in a post-hoc table or spreadsheet (this includes the data used in the synthesis for each study (e.g. outcome metrics or effect size, population, intervention etc, and meta-data)
AND
calculations or transformation of data by review authors using extracted data (e.g. effect sizes, averaging over variables summarisation of themes, coding, etc.) are reported in full and therefore repeatable.
- Amber: The review provides a table/spreadsheet that includes some of the extracted metrics data for some or all studies (e.g. Table/spreadsheet only lists e.g. 1-2 pieces of data for each study but omits other information OR Table/spreadsheet lists extracted data, but not for all studies OR A combination of these). Note: It may be unclear if all studies are included since they are not listed anywhere in the article.
- Red: No to amber above. Data extracted are not presented.

6.3. Were extracted data cross checked by more than one reviewer?

Rationale: Checking data extraction improves accuracy by ensuring the correct data are extracted for each element and reduces the risk of errors due to interpretation or transcription

- Gold: Data was extracted from each study by at least two independent reviewers.
- Green: An explanation was provided of how a sample of extracted data was cross checked between two or more reviewers?
- Amber: A statement that cross-checking between two reviewers was carried out is provided but explanation unclear or incomplete.
- Red: No report of cross checking is provided.

7. DATA SYNTHESIS

The approach to synthesising included studies varies substantially, and some approaches are more effective at ensuring objectivity and minimising potential bias than others.

7.1. Is the choice of synthesis approach appropriate?

Rationale: If appropriate, data should be pooled in a quantitative synthesis (e.g. meta-analysis, meta-regression). If substantial differences between populations, interventions, comparators or outcomes exist, meta-analysis may not be appropriate. Since meta-analysis effectively treats all individual studies part of one large study, meta-analysis is only appropriate when calculating an average effect is meaningful. If it is not appropriate to pool data across studies in meta-

analysis, a reason for this should be given, and structured approach to narrative synthesis taken, with efforts made to make sense of the whole of the data set, beyond describing results from individual studies in turn, noting differences in the weight of evidence behind statements made, and appropriate use of table and graphical presentations of results. Vote-counting (summing the studies which gave positive or negative findings) is not an appropriate synthesis method as an indication of impact or effectiveness.

- Gold: The choice of synthesis method (i.e. meta-analysis or narrative synthesis) is pre-specified and justified (e.g. in the protocol) on the basis of scoping characteristics of included studies, taking into consideration variability between studies in sample size, study design, context, etc?

OR

Justified post hoc as a result of unexpected outcome (not pre-specified) of critical appraisal and data extraction

AND

Where quantitative and statistical approach to meta-analysis is not employed, a justification for this should be given (e.g. studies too diverse or data not synthesisable).

- Green: The choice of synthesis method (i.e. meta-analysis or narrative synthesis) is (or appears) justified on the basis of characteristics of included studies, taking into consideration variability between studies in sample size, study design, context, outcomes etc.

AND

Where meta-analysis is not conducted, a justification for this should be given.

- Amber: No to either or both of the above: (e.g. no justification for not undertaking meta-analysis when it may have been appropriate)

- Red: No to green above: meta-analysis undertaken when inappropriate

OR

Vote-counting relied on as an indicator of impact or effectiveness

OR

Narrative synthesis does not include all studies or unclear if it includes all studies.

7.2 Is a statistical estimate of pooled effect (or similar) provided together with measure of variance and heterogeneity among studies?

Rationale: If a sufficient quantity and quality of data is available then the presentation and assessment of evidence can be much improved by providing statistical information. Some evidence reviews will be unable to do this because of limitations of the primary data.

- Gold: Statistical estimates of findings are presented using pre-defined (e.g. in the protocol) meta-analysis method that justifies synthesis approach including study weighting and subgroup analysis
AND
Consideration is given to study independence and bias (e.g. through sensitivity analysis).
- Green: Statistical estimates of findings are presented using meta-analysis method that justifies approach (e.g. using study weighting and subgroup analysis)
AND
Consideration is given to study independence and bias (e.g. through sensitivity analysis).
- Amber: Statistical estimates of findings presented using defined meta-analysis method but lacks justification of approach and/or consideration of study independence and bias (e.g. through sensitivity analysis).
- Red: No statistical estimate provided either because meta-analysis not conducted or not possible
OR
Statistical estimate provided but not clear what meta-analysis method was used.

7.3 Is variability in the study findings investigated and discussed?

Rationale: Studies can differ in their results (heterogeneity) which may be due to chance, but could also reflect variables other than the factor of interest that differ between studies (effect modifiers). The presence and magnitude of effect modifiers can reveal important information about a system. Investigating heterogeneity therefore indicates the degree to which effects are generalisable across taxa, regions etc., and is also necessary to evaluate the appropriateness of combining studies conducted on different populations or reporting different outcome metrics. These can be investigated statistically in meta-analysis through for example subgroup analyses, sensitivity analyses and meta-regression. In narrative syntheses differences in findings may be discussed in terms of differences in study design, context, population, focus etc.

- Gold: A strategy for investigating effect modifiers is provided in an a-priori protocol and followed (or variations explained) in the review
AND
Effect modifiers (e.g. taxa being considered, location, habitat type, study design etc.) were investigated statistically through meta-analysis (alternatively, evidence for heterogeneity between studies is tested and reported as non significant)
AND
Authors have considered results of critical appraisal (not just statistical weighting in meta-analysis) in considering findings through quantitative synthesis.

- Green: Effect modifiers (e.g. taxa being considered, location, habitat type, study design etc.) are investigated statistically through meta-analysis (alternatively, evidence for heterogeneity between studies is tested and reported as non significant) AND
Authors have considered results of critical appraisal in considering findings through quantitative synthesis.
- Amber: Effect modifiers (e.g. taxa being considered, location, habitat type, study design etc.) investigated statistically through meta-analysis (alternatively, evidence for heterogeneity between studies is tested and reported as non significant) but authors do not consider results of critical appraisal in considering findings through quantitative synthesis.
OR
Effect modifiers (e.g. taxa being considered, location, habitat type, study design etc.) were investigated descriptively through narrative synthesis.
- Red: Reasons for variability in study findings not investigated. Effect modifiers are not considered (this includes studies that use quantitative synthesis but fail to test for heterogeneity statistically).

7.4 Have the authors considered limitations in the synthesis?

Rationale. All reviews will have limitations and it is important that authors are explicit about the known limitations of the primary data and the conduct of the review process. [Here we should deal with the subjective nature of this section and make it clear to the appraiser that they must use judgement to decide if the limitations section is adequate]

- Gold: An explicit section is devoted to author's consideration of limitations of their review including limitations of the primary data (available evidence), possible sources of bias, conduct of the review process and recommendations made for future syntheses.
- Green: An explicit section or identifiable passage of text is devoted to author's consideration of limitations of their review but not containing all of the above.
- Amber: Some consideration of limitations is evident but not explicitly stated or not focus of specific section
AND
possible sources of bias, conduct of the review process and recommendations made for future syntheses not included
- Red: No evident consideration of limitations.