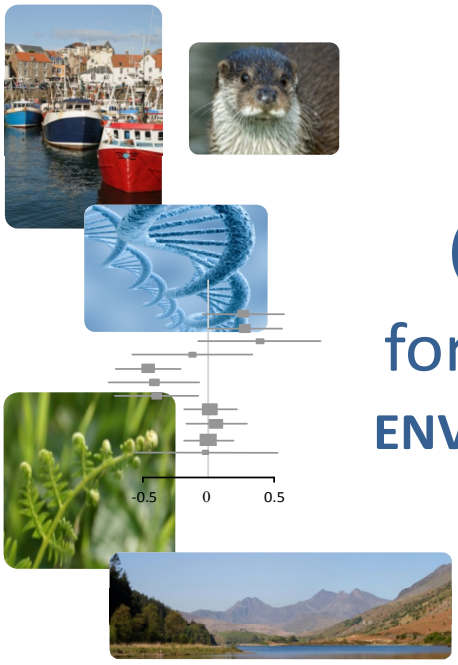
Please note that as of February 2018, the CEE Guidelines for Systematic Reviews in Environmental Management v4.2 are no longer in use.

Please visit http://www.environmentalevidence.org/information-for-authors to view the Guidelines and Standards for Evidence Synthesis in Environmental Management v5.0. The new Guidelines and any updates and corrections will now be available online only.

COLLABORATION FOR ENVIRONMENTAL EVIDENCE



GUIDELINES

for SYSTEMATIC REVIEWS in ENVIRONMENTAL MANAGEMENT

Version 4.2
March 2013

Compiled on behalf of CEE by

Centre for Evidence-Based Conservation
Bangor University, UK

Guidelines for Systematic Review and Evidence Synthesis in Environmental Management

Collaboration for Environmental Evidence

Version 4.2

March 2013

Please cite as:
Collaboration for Environmental Evidence. 2013. *Guidelines for Systematic Review and Evidence Synthesis in Environmental Management.* Version 4.2. Environmental Evidence: http://environmentalevidence.org/wp-content/uploads/2014/06/Review-guidelines-version-4.2-finalPRINT.pdf

Acknowledgements

This version of the Guidelines has developed from previous versions and we thank all those who have contributed to those versions.

We thank the many managers, policy formers and scientists who have given us constructive feedback on the review process and those who have themselves contributed to systematic reviews.

The following individuals contributed to the development of the latest version these guidelines. Claes Bernes, Monique Borgerhoff-Mulder, Adam Felton, Geoff Frampton, Markus Gusset, Neal Haddaway, Sif Johansson, Teri Knight, Magnus Land, Barbara Livoreil, Gabor Lovei, Beccy Mant, Alejandro Martinez-Abrain, Andrew Pullin, Rob Richards and Ruth Stewart.

Illustrations:
www.iStockphoto.com
The Saxifraga Foundation www.freenatureimages.eu

Aims and Scope

Systematic review (SR) and evidence synthesis methodology is now in widespread use in sectors of society where science can inform decision making and has become a recognised standard for accessing, appraising and synthesising scientific information. The need for rigour, objectivity and transparency in reaching conclusions from a body of scientific information is evident in many areas of policy and practice, from clinical medicine to social justice. Our environment and the way we manage it are no exception and there are many urgent problems for which we need a reliable source of evidence on which to base actions. Many of these actions will be controversial and/or expensive and it is vital that they are informed by the best available evidence and not simply by the assertions or beliefs of special interest groups. For SR methodology to be credible and reliable, standards need to be set and upheld. Here we present the latest guidelines for the commissioning and conduct of SR in environmental management.

The guidelines for CEE SRs have been adapted from methodologies developed and established over more than two decades in the health services sector (Higgins & Green 2009) and informed by developments in other sectors such as social sciences and education (Gough et al. 2012). Through undertaking and peer reviewing CEE SRs, researching and adapting existing methodologies, and through analysis of procedures and outcomes, we have developed specific guidelines for application to environmental management. Whilst past CEE SRs may provide some guidance, our advice is not to assume that past practices are sufficient for future CEE SRs. This document refers to examples of best practice and CEE is constantly trying to improve standards of SRs.

Although the basic ethos of SR remains unchanged, environmental methodologies are often different in nature and application from those in other fields and this is reflected in the guidelines. At first glance, many of the approaches may seem routine and common sense, but the rigour and objectivity applied at key stages, and the underlying philosophy of transparency and independence, sets them apart from the majority of traditional reviews recently published in the field of applied ecology (Roberts et al. 2006). SRs are now being commissioned by a range of organisations in the environmental sector and the need for common guidelines and collaborative development of the methodology is critical. We argue that, once more widely established, SR methodology will significantly improve the identification and provision of evidence to inform practice and policy in environmental management. For this methodology to have an impact on effectiveness of our actions, more environmental scientists and managers need to get involved in the conduct of SRs. For those intending to conduct SRs, these guidelines are provided in the spirit of collaboration and we encourage you to contribute your work to the CEE, use and improve these guidelines, and help establish an evidence-based framework for our discipline.

Who are these guidelines for?

The guidelines are primarily aimed at those teams intending to conduct a CEE SR. The structure of the document takes you through the key stages from first consideration of a CEE SR to the dissemination of the outcome. They are guidelines only and do not replace formal training in SR methodology. Review Teams should not expect that the guidelines alone will be sufficient support to conduct a SR.

We hope that these guidelines will also be of use to those considering commissioning a SR and stakeholders who may become involved in their planning.

Finally these guidelines set a standard for the conduct of SRs and are therefore intended for decision makers using evidence from SRs and wishing to understand the nature of the SR process.

For clarity of process, the guidelines are split into five sections (Figure 1). There is obviously considerable overlap between planning, conducting and reporting and we cross reference as much as possible to avoid undue repetition. We use examples of completed SRs in the CEE library to illustrate each stage of the process and to highlight key issues.

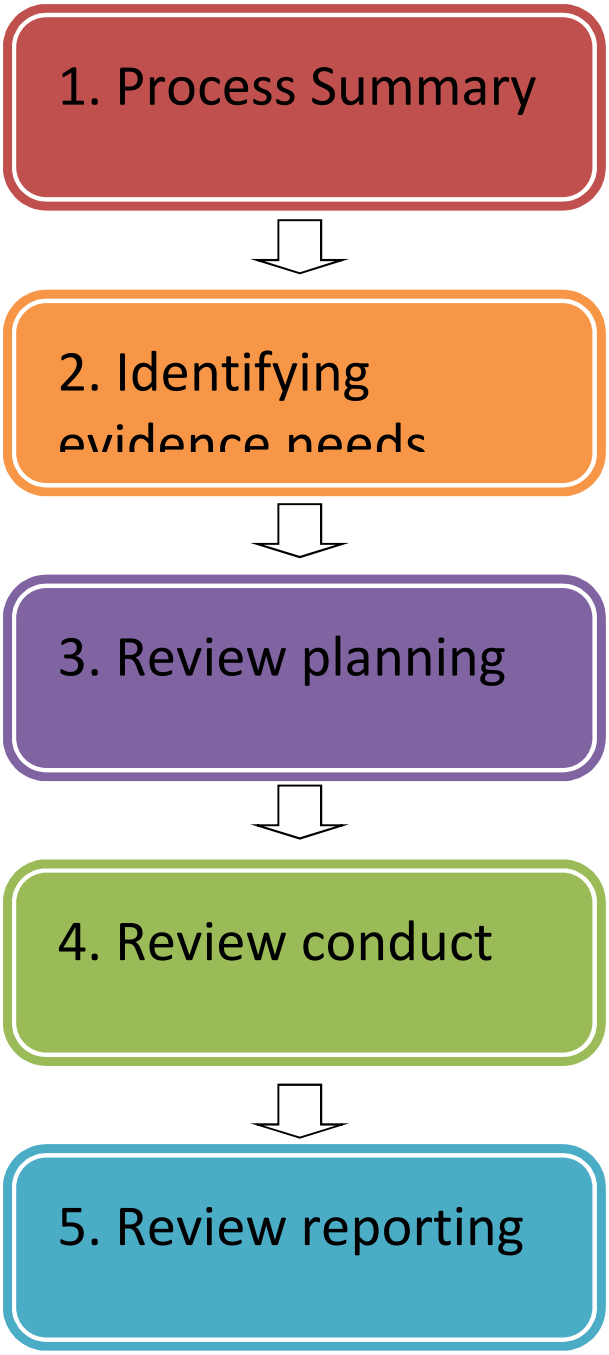


Figure 1. Basic stages of SR that form the structure of these guidelines

Contents

Section 1

Process Summary, Registration, Publication and Dissemination of a CEE Systematic Review

This section provides an summary of the steps in SR conduct and an overview how authors register their SR with CEE and of the process of submission and peer review that ensures CEE SRs are conducted to high standards.

If you are thinking of conducting a CEE systematic review for the first time we encourage you to get in touch with CEE at an early stage. We will be able to advise you on supporting materials and available training to help you develop your ideas into a viable protocol.

High standards of reporting are expected on the conduct of the review and this starts with the submission of a protocol and continues through to the provision of supplementary material such as excluded articles and data extraction spreadsheets. A template and checklist to aid report writing of SRs are available at www.environmentalevidence.org/Authors.htm

1.1 Steps in conducting an SR – A brief summary

SRs start with a question, rather like primary research, but unlike the latter SRs collect and synthesise existing data in order to attempt to answer the question. Figure 2 sets out the main stages each of which is covered in detail later in these guidelines. However, it may be useful at this point to have a basic understanding of the role of each stage.

1. Question setting: A process to derive a suitable question both in terms of evidence needs and feasibility of the SR.
2. Protocol: a plan for the conduct of the SR setting out how each stage will be conducted. The protocol is submitted to Environmental Evidence, peer reviewed and published and has a key role in maximising transparency and minimising susceptibility to bias.
3. Searching: A systematic search is conducted using a repeatable search strategy tailored to the question and likely sources of evidence.
4. Article screening: Articles retrieved from the search are examined for relevance to the review question using a-priori inclusion criteria and resulting in a collection of relevant studies.
5. Critical appraisal and data extraction: two stages that are often interlinked. In critical appraisal, studies are examined for their design and reporting standards and weighted in terms of susceptibility to bias and validity in terms of the study question. Appropriate data are extracted from each study and may be subject to further critical appraisal.
6. Data synthesis: Extracted data from individual studies are synthesised to form an overall view of the evidence. Synthesis can be narrative, quantitative, qualitative or a combination of these.
7. The SR report is written up using a specific CEE template that ensures high reporting standards for transparency and repeatability. The report is submitted to Environmental Evidence, peer reviewed and (if accepted) published in the journal and archived on the CEE website.

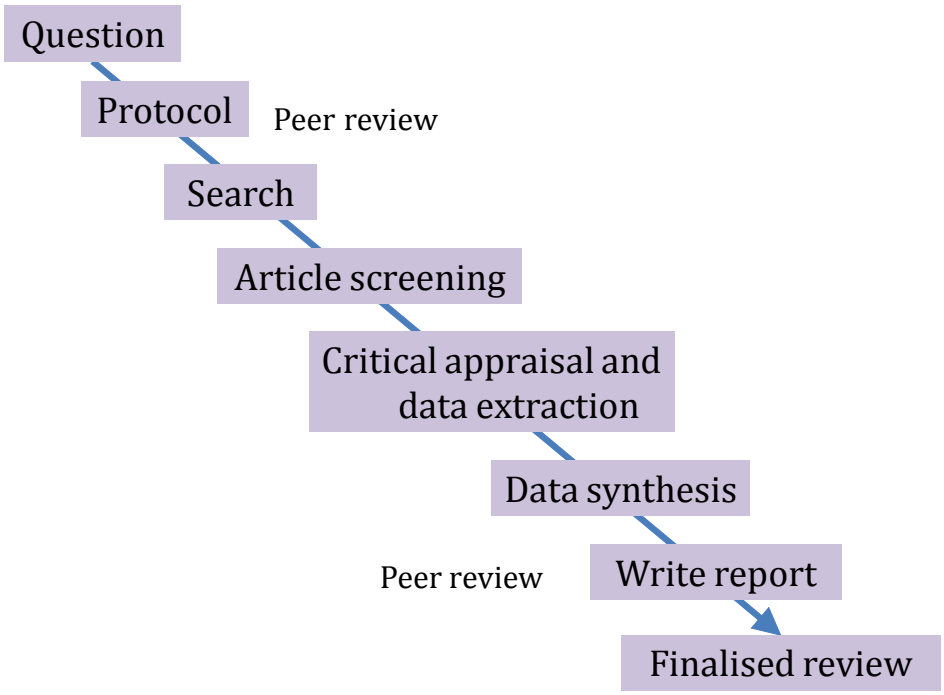


Figure 2. Basic steps in conducting a CEE systematic review

1.2 The CEE registration, submission and deposition process

CEE operates an open-access policy and all of its protocols, systematic reviews and some associated materials are published in its open-access journal 'Environmental Evidence' (www.environmentalevidencejournal.org). Article Processing Charges are payable (www.environmentalevidencejournal.org/about#apc).

Here we set out the process for publishing protocols and full reviews in 'Environmental Evidence'. Full instructions for authors on preparation of their manuscript are available at www.environmentalevidence.org/Instructionsforauthors.html

Registration and submission of a SR to Environmental Evidence is an interactive stepwise process as follows;

1. Draft protocols are submitted to Environmental Evidence through an electronic submission system. The draft protocol will be sent out for peer review. Comments will be returned to the authors and appropriate revisions may be requested to finalise the protocol.
2. The finalised protocol is published in Environmental Evidence, posted on the CEE website and the SR is then formally registered as being 'in progress'. At this point a dedicated review webpage will be created on the CEE website and can be used by the authors to post updates and news.
3. Submission of a draft review to Environmental Evidence follows the same process. If acceptable after an initial screening, the draft review will be sent out for peer

review. Comments will be returned to the authors and appropriate revisions may be requested to finalise the review.

4. The revised and completed SR (and associated supplementary material) will be published in Environmental Evidence and posted on the website in the Environmental Evidence Library as finalised.

CEE operates a very supportive policy for review teams undertaking SRs and seeks to provide help and guidance (through web-based support materials and training events) to increase the chances of SRs being successfully completed.

1.3 Supplementary materials

Beside the main text, the transparency of SRs is enhanced by the provision of a range of supplementary materials. Some can be provided as appendices whilst others may be posted as separate documents on the review webpage. For a full list see Section 5.5.

1.4 Further dissemination of findings

After all the work of searching, sifting, appraising, extracting and synthesising evidence and writing the report, it is worth considering whether the full SR format is likely to be the most effective for disseminating the key outcomes. The details of the full SR constitute an important resource and audit trail of methodology but the full technical report does not function very well as a dissemination tool. A full SR will normally include too much detail. By mutual agreement, other formats such as policy briefs, executive summaries and guidance notes can be developed and posted on the review webpage. Such documents often require some special skills in order to make the conclusions and recommendations, as well as their justification, accessible to a non-scientific audience. They can be written by the review team, but can also be designed by a specialist or during meetings with policy makers and/or practitioners and managers. For examples of these policy briefs go to www.environmentalevidence.org/Policybriefs.html

1.5 Updating a Systematic Review

SRs can only be accurate assessments of the evidence base when they are up to date. As soon as the search is completed the reliability of an SR as a synthesis of 'all available evidence' begins to decline. The rate of decline is dependent on the rate of publication of new studies and so varies from subject to subject. An outdated SR may be misleading, so they should periodically be updated. Fortunately the process of updating a SR should not be as burdensome as the original process provided accurate reporting was achieved and good records were kept of the original process. We encourage the deposition and archiving of as full a record as possible of all procedures and outcomes on the SR webpage. At the time of writing, updating a SR in environmental management is yet to be

completed; we suggest updating a review around 5 years after publication. The process for registering an update is the same as an original review and should begin with an updated protocol.

- if a review is 5 or more years out of date, the CEE editorial team will contact the authors inviting them to update the review.
- If the authors are unable to take up this invitation, the review will be marked as 'update sought' and updates will be open to any interested party.
- In the case that a new review team is formed to update a review, they will be expected to liaise closely with the original team who may also be named as authors in the updated review to reflect the intellectual input into the review as a whole.

Section 2

Identifying the need for evidence and systematic review

2.1 Addressing the need for evidence

In trying to find solutions to problems and decide among alternative interventions to achieve desired outcomes, individuals, organisations or groups may identify a need for evidence. This chapter provides guidance on the identification of evidence[1] needs to inform decision making. In doing so we provide some guidance on the initial steps that may, or may not, result in the planning and commissioning of a systematic review.

The need for evidence relating to a question of concern in policy or practice can arise in many ways from scientific curiosity of individual researchers to global policy development. Identifying and agreeing priority issues and the need for evidence to inform decisions are often iterative processes involving dialogue among many different individuals and organisations. It is not our intention here to try and describe the policy process or how management decisions are made. In the process of deciding how to spend limited resources to achieve organisational objectives, there is an opportunity for that decision to be informed by the best available evidence. However, the evidence has to be relevant and valid. Identifying exactly what evidence would help decision-making is therefore worth some thought and discussion.

The following are examples of scenarios that might generate need for evidence;

1. Assessment of a problem (e.g. is the perceived problem really a problem and, if so, how big is it?)
2. Solution scanning (e.g. what are the potential solutions to problem x?)
3. Predicting impact (e.g. what evidence exists that exposure of a population to a factor will have an impact?)
4. Relative performance (e.g. which is the best intervention, tool, mechanism for the job?)
5. Need for evidence of effect of an intervention (e.g. is the intervention we use to address a given problem working or not working?)
6. Need for evidence of influence of effect modifiers (e.g. what factors influence the effectiveness of an intervention?)
7. Need for evidence of cost effectiveness
8. What is the best combination of interventions to achieve our objectives?

Below are examples of initial concerns/questions related to evidence needs. As we shall see below, they require some discussion and reformulation to obtain questions that can be addressed by SR methodology.

[1] See the Glossary at the end of the document for definitions

Box 1. Examples of initial concerns/problems generating potential questions for SR

- Impact of roads on mammals and birds
- Responses of invasive species to climate change and impact on native species
- Management of forest by local communities to achieve better biodiversity protection
- Consequences of different levels of greenhouse gases on genetic mutation in terrestrial organisms
- Pollution of rivers and impact on the fecundity of fishes
- Resilience of different ecosystems to over-exploitation
- Is systematic conservation planning useful for biodiversity projects?
- Effectiveness of reserves as a tool to preserve migrating species
- Can we anticipate the consequences of climate change and flooding risks on urban planning?
- Improvement of agricultural practices to restore soil biodiversity
- Measurement of variation of glacial retreating
- Mitigation of the effect of climate change by urban greening
- Efficiency of reintroductions, translocation and captive breeding programmes to restore populations
- Effectiveness of science-policy communication to achieve sustainable conservation decisions
- Impact of educational programmes to protect endangered species
- How can we prevent contamination of native species by GMOs?

(more examples at the CEE Library www.environmentalevidence.org/Reviews.htm and at www.environmentalevidence.org/Reviewsinprogress.html)

Initial questions arising from discussions of evidence needs are typically open-framed, whereas questions appropriate for SR are typically close-framed. This issue is addressed in Section 2.4.1.

2.2 Getting people involved

In progressing from evidence needs to consideration and planning of a SR it is likely that several different groups will (or should) be involved. Usually, the group of people that identify a need for evidence will not be the group that undertakes a SR. There are three definable, but not mutually exclusive, groups that could be involved in the conduct of a SR from this early stage:

The User Group – policy or practice groups that identify the need for evidence and might commission a SR and/or use its findings in the context of their work.

The Review Team – the group that conducts the review; the authors of the review report.

The Stakeholder Group – all individuals and organisations that might have a stake in the findings of the review.

Normally, to avoid conflicts of interest, any individual would not be a member of more than one of these groups. **Funding** will often come from the User Group but can come from any one of these groups or be entirely independent. Funders must always be declared along with any other conflicts of interest that might arise.

The Client and the Stakeholder Groups will have a very important role in the choice of the SR question, and in its phrasing. They will also help to establish the list of sources of evidence and search terms (by providing some of them, or checking the list for completeness). Involving many people at an early stage may be particularly critical when the findings are likely to be contested (Fazey et al. 2004), such as in site selection for establishment of windfarms. However, stakeholder input needs to be carefully managed to avoid the question becoming too broad, complex or just impossible to answer (Stewart & Liabo 2012). There are further opportunities throughout the SR process for input from stakeholders, but as we shall see, identifying and obtaining initial consensus on the question is crucial to the value of the SR.

As potential questions are generated and review teams are formed there will be a process of question formulation. There is no set formal process for this but the critical elements are set out in the following section.

2.3 Why conduct a Systematic Review? Assessing the costs and benefits

SR is just one way of addressing a question. Questions are continually generated and there are limited resources for evidence synthesis and so it seems sensible to provide guidance on when SR methodology is appropriate. There are different motivations for conducting SRs. From a scientific perspective it may be motivation enough that the SR will have interesting implications for future research. This section considers in more detail the decision makers or commissioners perspective and addresses the problem of deciding whether a SR is the right option for informing their work.

The key characteristics of SR are rigour, objectivity and transparency. These characteristics serve to minimise bias and work toward consensus among stakeholders on the status of the evidence base. SRs are also readily updatable as new primary studies are made available and this serves to measure the development of the evidence base. Therefore, some examples of when an SR may be appropriate are when:

- There is a need to measure the effectiveness of an intervention or relative effectiveness of interventions.

- There is a need to measure the impact of an activity on a non-target population.
- There is a need to know how much research has been conducted on a specific question (see systematic mapping section).
- There are opposing views about effectiveness of interventions or impact of actions.
- There is a need to consider the relative effectiveness and cost of interventions.

SR may not be appropriate when the question:

- is poorly defined or too complex (but see section 2.4)
- is too simple (e.g. has species x been recorded in region y)
- does not attract stakeholder (including scientific) interest (i.e. the rigour is not necessary)
- is not judged sufficiently important for SR to be cost-effective
- very little good quality evidence exists but exposure of a knowledge gap will not be valued.

SR may not be needed when:

- a similar SR has recently been completed (but see SR updating below)
- the question can be satisfactorily answered with less rigorous and less costly forms of evidence synthesis

Funding for the review is likely to be a key factor. In our experience SRs vary in their full economic cost by as much as an order of magnitude. Some highly focused questions with few but highly accessible datasets may cost as little as US$30K whereas a broad and contested question with disparate data, much of which may be 'hidden' in the grey literature, may cost as much as US$300K and take several years to complete. Any preparatory scoping work that may help predict where on this scale a SR is likely to reside is probably time well spent (See section 3.2). Of course, the full economic cost may not be an appropriate metric and Review Teams may wish to calculate costs in other ways.

2.4 From a problem to a reviewable question: Question generation and formulation

Each SR starts with a specific question whereas evidence needs are typically much broader. For commissioners and decision makers, finding the right question to inform decisions can be a compromise (probably more so in environmental sciences than in most other disciplines) between taking a holistic approach, involving a large number of variables and increasing the number of relevant studies, and a reductionist approach that limits the review's relevance, utility and value (Pullin et al. 2009). There can be a temptation to try to squeeze too much information out of one review by including broad subject categories, multiple interventions or multiple outcome measures (this can be dealt with by first conducting a systematic map (see Section 2.4.3). Equally, there may be a tendency to eliminate variables from the question so that the utility or 'real world' credibility (external validity - see Section 4.3.2) of the review is limited.

The formulation of the question is therefore of paramount importance for many reasons. For example:

- the question must be answerable using scientific methodology (Jackson 1980; Cooper 1984; Hedges 1994), otherwise relevant primary studies are unlikely to have been conducted.
- it should be, as yet, unanswered (i.e. the Review Team should search for other related systematic reviews and specify what theirs will add)
- it should be generated by, or at least in collaboration with, relevant decision-makers (or organisations) for whom the question is real, to ensure its utility to inform.
- it may also be important for the question to be seen as neutral (unbiased) to stakeholder groups to minimise conflicts
- definitions of the structural elements of the question (see Section 2.4.2) are critical to the subsequent process because they generate the terms used in the literature search and determine relevance criteria.

The wording of the question and the definitions of question elements may be vital in establishing stakeholder consensus on the relevance of the review. Ideally, meetings should be held with key stakeholders to try to reach consensus on the nature of the question. We recommend that experts in the field be present or consulted. Ideally, a meeting would invite some of them to present the state-of-the-art on the topic of interest so that each participant (especially Review Team members that are not subject experts) could be familiarised with the context, technical jargon and challenges.

2.4.1 Open-framed and closed-framed questions

Not all types of question are suitable for SR. Many questions that might initially be posed by user groups in policy and practice are open-framed in that they lack specification or definition of some of the key components (see Section 2.4.2). Such questions are normally not answerable in a single experimental study and therefore not answerable through a synthesis of similar studies. A typical example might be '**how can we reduce the impact of alien invasive species on native biodiversity**'. This example does not specify any of the potential interventions that could be used to reduce impact (and is also poorly defined in terms of which alien invasives and what elements of biodiversity). Closed-framed questions contain all the necessary elements, although each element may need further definition. An example might be **'is poison-baiting effective at eradicating rats from islands'** (see section 2.4.2 for explanation of how this question is broken down).

Breaking down open-framed to identify closed-framed questions can be a valuable exercise in a policy context. Pullin et al. (2009) have outlined a process adapted from the health services. Essentially two stages are involved as outlined in Figure 3. The first requires that potential strategies for addressing open-framed questions are identified and the second that potential interventions are considered that would help deliver each strategy. The effectiveness of these interventions can then be the focus of a SR. The technique of systematic mapping can be used to inform the first stage and address an open-framed question (see Section 2.4.3).

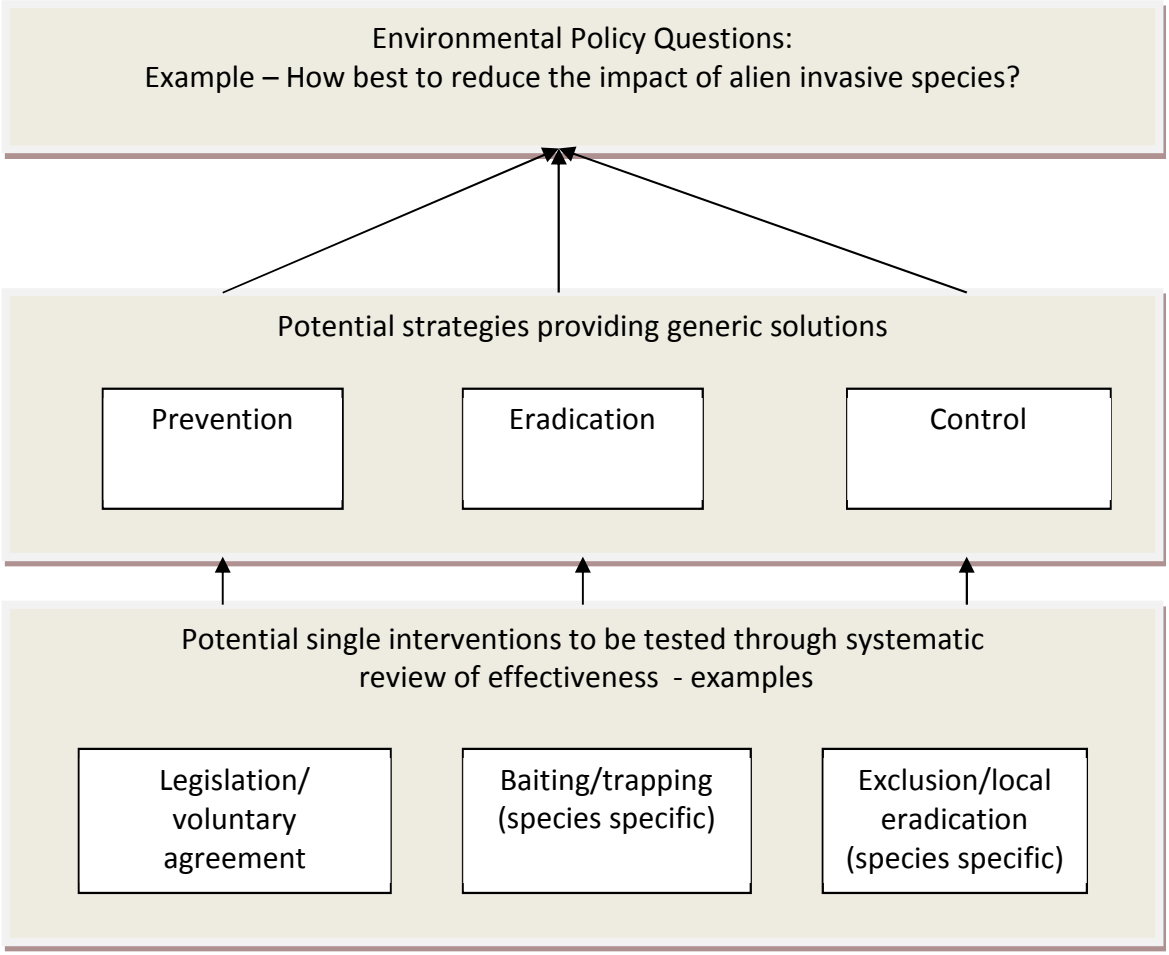


Figure 3. Relationship between a high-level open-framed policy question, potential generic solutions and individual interventions. After Pullin et al. (2009).

2.4.2 Key components of a question susceptible to SR

The most common questions for SR have four definable elements, often referred to as the PICO or PECO (**P**opulation, **I**ntervention/**E**xposure, **C**omparator, **O**utcome) elements (Table 1). Using the example in Section 2.4.1 '**is poison-baiting effective at eradicating rats from islands'**.

P = rat populations on islands,
I = poison baiting methods,
C = no baiting (or maybe other methods),
O = eradication of rat populations.

Although SR methodology was initially developed to test the effectiveness of interventions in medical practice, its use has broadened considerably and the methodology is now also used to address a range of different types of questions containing only some of the PICO/PECO elements (Table 2).

Table 1. Elements of a reviewable PICO/PECO question: normally a permutation of 'does intervention/exposure I/E applied to populations of subjects P produce outcome O?'.

Question element	Definition
Population(of subjects)	Unit of study (e.g. ecosystem, species) that should be defined in terms of the statistical populations of subject(s) to which the intervention will be applied.
Intervention/ exposure	Proposed management regime, policy, action or the environmental variable to which the subject populations are exposed.
Comparator	Either a control with no intervention/exposure or an alternative intervention or a counterfactual scenario.
Outcome	All relevant outcomes from the proposed intervention that can be reliably measured or outcome that might result from exposure to an environmental variable.

Table 2. Examples of question types answerable through SR methodology.

Question Type	Question Elements	Example elements
Effect of intervention or exposure	Population	Local human populations
	Intervention	Terrestrial protected areas/associated integrated development projects
	Comparator	Absence of PAs
	Outcome	Measures of human wellbeing
	= QUESTION "What are the human wellbeing impacts of terrestrial protected areas?" (Pullin et al. 2012)	
	Population	Vegetation in alpine/subalpine areas and arctic/subarctic tundra
	Exposure	Herbivory by reindeer/caribou
	Comparator	No/less herbivory by reindeer/caribou
	Outcome	Vegetation change (assemblage or specific groups)
	= QUESTION "What are the impacts of reindeer/caribou (*Rangifer tarandus*) on arctic and mountain vegetation?" (Bernes et al. 2013)	
Analytical accuracy (diagnostic test accuracy)	Population	Forest ecosystems
	Test being evaluated (Index test)	Estimates of carbon content
	Target Condition	Carbon release or sequestration from ecosystem change
	= QUESTION "Comparison of methods for the measurement and assessment of carbon stocks and carbon stock changes in terrestrial carbon pools? (Petrokofsky et al. 2010)	
Prevalence, occurrence, incidence	Population	Red fox populations
	Outcome	Prevalence of rabies
	= QUESTION "What is the rate of occurrence of rabies in foxes in various European countries?"	

Decision makers may often seek more than just an answer to the primary question. Secondary question elements, that follow on from the primary question, such as the cost-effectiveness of interventions; the prediction of variance in effectiveness (when or where will it work or not work?); the appropriateness and acceptability of particular interventions; and the factors which might influence the implementation of interventions 'in the real world' as opposed to the laboratory may be of equal or even greater importance. In many cases this might mean that the review essentially follows the 'effectiveness' review format but with development of synthesis strategies tailored to address a range of subquestions. Of importance is that there is discussion with funders and stakeholders at the beginning of the process to identify the type of evidence needed

– to assess whether or not an effectiveness type review is the most appropriate and if so, whether the nature of the question requires methodological variation from the standard protocol.

Other related question structures have been proposed and might be more applicable to some kinds of questions. SPICE (Setting, Perspective, Intervention, Comparator, Evaluation method) is an example that might be applicable to some questions suitable for CEE SRs (Booth 2004).

Box 2. Examples of question formulation

Concern/Problem

Protected areas (PAs) must 'at least do no harm' to human inhabitants (Vth IUCN World Parks Congress, Durban 2003), but previously some PAs have been documented to have many negative effects on humans living inside and around their borders. STAP (the Scientific and Technical Advisory Committee of the UN) wanted to know how PAs affected human wellbeing and whether impacts had changed over time and with different governance structures.

Question Development

Terrestrial protected areas were considered distinct from marine in the context of human impacts. The SR would include established and new PAs and intrinsically linked development projects. All outcomes relating to measures of human wellbeing were deemed relevant. The commissioners decided that the target populations would include all local human populations living both within and around the PA, with 'local' being defined as broadly as up to and including a national level. A cutoff of 1992 was chosen for published studies, since all PAs had to conform to IUCN category guidelines established at the CBD in Rio de Janeiro, 1992.

Final SR Question

What are the human wellbeing impacts of terrestrial protected areas?

Concern/Problem

Lowland peatland ecosystems constitute vast amounts of carbon storage relative to their geographical extent. Extraction and drainage of peat for fuel and agriculture can release greenhouse gases (GHG; CO_2, CH_4 and N_2O) and other carbon stores, contributing to global warming. Rewetting and wetland restoration aim to ameliorate these destructive practices but their effectiveness is uncertain. Whilst upland peat systems are relatively well-understood, no synthesis concerning lowland peats has been undertaken to date.

Question Development

The commissioners decided to focus the subject of a previous SR topic from all peatlands onto temperate and boreal regions, and widen the scope from water level changes to all changes in land management. Carbon fluxes and greenhouse gases were kept as relevant outcomes.

Final SR Question

How are carbon stores and greenhouse gas fluxes affected by different land management on temperate and boreal lowland peatland ecosystems?

Concern/Problem

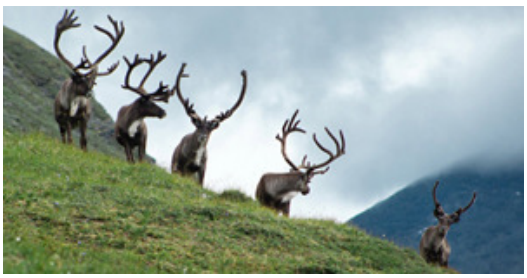
What intensity of grazing should be recommended to conserve biodiversity whilst ensuring economic sustainability of reindeer herding? An early view that reindeer were responsible for overgrazing in northern parts of Scandinavia has changed, with current opinion being that the observed overgrazing was localized and short-lived. In contrast, some are now concerned that grazing levels are insufficient to control mountain vegetation. Stakeholders identified a need to clarify a vague political dogma and goal; that the Swedish mountains should be characterised by grazing.

Question Development

Development of the review question (initially suggested by the Swedish Environmental Protection Agency, SEPA) was undertaken by a team of scientists in consultation with stakeholders. Any impact resulting from herbivory by reindeer or caribou (both *Rangifer tarandus*) from anywhere in their natural or introduced range was chosen to be included in the scope of the review. Herbivory in coniferous forests was excluded, however, since the review was to be focused on mountain and arctic regions.

Final SR Question

What is the impact of reindeer/caribou (*Rangifer tarandus*) on mountain and arctic vegetation?



2.4.3 Using a systematic mapping approach

Sometimes the evidence needs are articulated as open-framed questions and it is not feasible to derive or select a more specific question before a broader review of evidence is conducted. Initial searching for and sorting of evidence in relation to broader questions is termed systematic mapping. Thus it may be useful to undertake a two- stage review, with a systematic map of the research, followed up by SRs on subsets of research identified in the map. This permits the reviewers and users to understand the scope of current research activity in a given broad subject area before focussing on specific areas of interest.

In systematic mapping, the searching and inclusion processes are conducted with the same comprehensive method as for a full review, but the process does not extend to critical appraisal or data synthesis. Data are however extracted from included studies in order to describe important aspects of the studies using a standard template and defined keywords and coding. This approach is designed to capture information on generic variables, such as the country in which a study took place, the population focus, study design and the intervention being assessed. This standard and well-defined set of keywords and codes is essential whenever classifying and characterising studies in order for reviewers to pull out key aspects of each study in a systematic way. For an example of a systematic map see http://www.environmentalevidence.org/SR35.html. In this example, Randall et al. (2012) examined the effectiveness of integrated farm management, organic farming and agri-environment schemes for conserving biodiversity in temperate Europe. Their systematic map searched for relevant information in accordance with typical systematic review methodology. Screening was then undertaken to abstract level and a searchable database created using key wording to describe, categorise and code studies according to their focus and methodology. This searchable database is hosted on the CEE website and is freely available. Once the research has been mapped in this way it is then possible to identify pools of research which may be used to identify more narrowly defined review questions. For an example of this approach see Bowler et al. 2009. For examples within the health and social science fields see the EPPI Centre (http://eppi.ioe.ac.uk). **Systematic maps are registered and conducted according to the same procedures as CEE SRs** (See Section 1).

Section 3

Planning a CEE systematic review

3.1 Establishing a Review Team

Conducting a SR is a substantial piece of work and usually requires the input of a multidisciplinary team. Teams may consist of subject experts combined with review methodology experts, such as information specialists or statisticians. SRs are normally undertaken by a team because one person is unlikely to possess all the skills required to conduct all stages of the review, or have the appropriate combination of subject and methodological expertise, and because several stages of the review require independent testing of repeatability that requires two or more participants. The Review Team should have a designated Lead Reviewer who is experienced in the methodology and able to project manage the rest of the team. The involvement of subject experts in the team brings with it the potential for bias. Careful consideration should be given to independence of subject experts within Review Teams and conflicts of interest declared, and avoided where possible.

It is preferable that the team is constituted before or during the establishment of the review protocol (section 3.3) so that the team feels ownership and responsibility for its content. The rigorous methodology employed in SRs means substantial investment in time and it is important that careful planning of timetables and division of work is undertaken using some key predictors of the likely size and scope of the review.

3.2 Review scoping

The Review Team will need to establish a plan of how they will conduct each stage of the SR. The key stages, together with their main purpose are summarised in Section 1.1
This plan will then form the core of the SR protocol to be approved and registered by CEE. The review scoping process that aids the planning of each stage is described here followed by the structure of the protocol itself (Section 3.3).

Before the commencement of a SR, it is essential that some review 'scoping' is undertaken to guide the construction of a comprehensive and appropriate protocol, and to provide an indication of the likely form of the review and thus facilitate resource planning. In certain circumstances, it may not be efficient to commit to a full SR without some prior estimation of its value. Scoping may be undertaken by the commissioning organisation, by the Review Team itself, or a combination of the two. A thorough scope should entail:

- The development and testing of a search strategy.
- An estimate of the volume of relevant literature.
- Critical appraisal of study quality and data extraction of a small subset of relevant papers.
- An estimate of resources required based on the above.

The expected output from a scoping exercise is an estimate of the quantity and quality of evidence, and a characterisation of the likely evidence base, pertaining to the question (see Box 3 for example). The extent of investment in review scoping is a matter of

judgement and will differ with each review. We detail below the steps of a full scoping exercise.

Box 3. Example of Review Scoping

A scoping study was undertaken as part of the development of a SR protocol for the question "What is the evidence that scarcity and shocks in freshwater resources cause conflict instead of promoting collaboration?" (Johnson et al. 2011). This process involved the trialling and refining of search terms in the literature database Web of Science based on a full list of relevant exposure and outcome terms identified through an initial knowledge map and discussion with the expert review team.

Pre-scoping exposure terms	water*, riparian*, aquifer*, aqua*, dam, dams, hydrolog*, hydroelectric*, groundwater, drought*, river*, lake*, stream, streams, reservoir*, flood*, irrigat*, rain*, baseflow*, precipitation, fresh*, basin*, flow, drylands
Pre-scoping outcome terms	conflict*, dispute*, insurgen*, war*, violen*, securit*, terror*, strife, peace*, govern*, coercion, cooperat*, "co-operat*", collaborat*, collective, geopolitic*, "international relation*" allocat*, distribut*, shar*, mediat* governance, treaty, treaties, agreement*, manag*

When testing the search terms, each of 23 exposure terms above was tested individually with the outcome terms (conflict* OR cooperat*) and the first 100 articles returned were screened to assess the exposure search term's usefulness (after sorting by relevance). Once the 23 terms had been refined and finalised, these terms were searched with each outcome search term individually in a similar way to produce a final search string.

Search Terms	**Hits**	**Comments**
[all exposure terms separated by 'OR'] AND (conflict* OR cooperat*)	>100,000	Large number of unrelated articles – lacks specificity
(rain OR rains OR rainfall) AND (conflict* OR cooperat*)	604	*Rain** changed to *(rain OR rains OR rainfall)*, and resulted in relevant hits. Retained in final search string.
baseflow* AND (conflict* OR cooperat*)	3	No relevant articles; 'baseflow' excluded from final search.

Scoping searches were undertaken using three bibliographic databases: ISI Web of Knowledge, OCLC First Search and Science Direct. In addition, a web-based search for grey literature was trialled using the search engine Google. Furthermore, professional organisations were identified and additional articles were provided by the Review Team and stakeholders with a personal knowledge of the topic.

Title-level screening was based on predefined inclusion criteria, and multiple reviewer checks for consistency were used. Study quality was assessed on a sample of relevant articles and identified key areas of research, for example land subsidence and groundwater lowering, and apparent gaps in research, for example the impacts of long-term environmental change. The type of data in studies of sufficient quality was further examined to give an indication of the type of data available. Field studies with qualitative data were most common, indicating that standardised summaries by a reviewer followed by assessment by a second reviewer was the most appropriate data extraction method.

3.2.1 Developing and testing a search strategy

Systematic and comprehensive searching for relevant studies is essential to minimise publication bias (see Section 3.3) in a SR and to assess the strength of the evidence base. Enlisting an information specialist in the review team is recommended so that an efficient search strategy can be established. Aside from the review's validity, a good search strategy can make a substantial difference to the time and cost of a review.

The development of effective search terms (strings of key words and phrases) for searching should take place largely during the review scoping stage, and will most likely be an iterative process, trialling search strings using selected databases, recording numbers of hits and sampling titles for proportional relevance (the proportion of the sample that appears to be relevant to the SR question), with sensitivity improving as scoping progresses. This may include considering synonyms, alternative spellings, and non-English language terms within the search strategy. An initial list of search terms may be compiled with the help of the commissioning organisation and stakeholders. All iterations of tested terms should be recorded, along with the number of 'hits' they return (see Appendix A for an example). This should be accompanied with an assessment of proportional relevance, so that the usefulness of individual terms can be easily examined. Comparing search results when you include or exclude particular terms will allow you to identify superfluous or ineffective terms, and work out whether any should be removed from your search strategy. It is important to remember, however, that the functionality of different literature databases may vary considerably and terms that are apparently useful in one source will not always be appropriate in others: thus search strings may need to be modified to suit each one.

All scoping searches should be saved so that they may be accessed later, removing duplication of effort where possible. However, if the scoping searches are conducted well in advance of the actual review search, it would be prudent to conduct the search again in order to ensure all recent literature has been identified.

The search terms chosen will be largely influenced by the elements of the question (Section 2.4.2). For efficiency, individual terms should be combined where appropriate using Boolean operators ('AND', 'OR', 'NOT', 'SAME', etc.): particular care should be taken when employing the 'NOT' operator, to ensure that relevant papers are not inadvertently excluded. Wildcard truncation symbols to search for variant word endings are often useful (see a detailed example in Appendix A).

It is important that the search for literature and data is sufficiently rigorous and broad that as many studies as possible that are eligible for inclusion are identified. Searches must thus balance sensitivity (getting all information of relevance) and specificity (the proportion of articles that are relevant). In ecology, searches of high sensitivity often come at cost of lower specificity, which means searches are resource-intensive. This is partly because environmental science lacks the MeSH (Medical Subject Headings)-indexes and integrated databases of medicine and public health, which assign standard keywords/descriptors to articles. A high-sensitivity and low-specificity approach is often necessary to capture all or most of the relevant articles available, and reduce bias and increase repeatability in capture (see below). Typically, large numbers of articles are

therefore rejected. For example, in a review of human wellbeing impacts of terrestrial protected areas, 15,593 articles were returned from database searches. Of these, however, only 177 (1.1%) possessed all relevant inclusion criteria. Similarly, a review of the impacts of land management on carbon and greenhouse gas fluxes in lowland peats identified 18,451 articles during database searches, yet only 93 (0.5%) contained all of the required PICO structure elements.

A final step in the development of the search terms is to test the strategy with a set of known relevant articles (these may often be provided by review commissioners or subject experts or have been selected for the trial critical appraisal (see 3.2.3 below); existing meta-analyses or reviews may also provide a source of relevant studies for testing). A comprehensive set of terms with an appropriate balance of specificity and sensitivity will retrieve these relevant articles without returning an unmanageable number. Any unretrieved articles should be inspected so that the search strings can be appropriately modified to capture them.

3.2.2 Assessing the volume of literature

The volume of literature arising from scoping searches may be used as a crude predictor of the strength of the evidence base. For example, whether the review question will identify a knowledge gap (very few articles), if it is too broad and should be broken down (very many articles), or if it has the potential to provide some form of data synthesis that provides a summary answer to the question. This has implications in terms of the time and resources required to complete the review. Note, however, that the total number of returned articles is likely to reflect the specificity of the chosen search terms (and possibly searching skills of the Review Team) and thus should not be used as an accurate predictor without first sampling a random sub-set to determine proportional relevance. This can then be used to extrapolate to determine the likely quantity (but not quality) of articles relevant to the review question.

3.2.3 Trial critical appraisal, data extraction and analysis

Having developed an effective search strategy and a familiarity with the likely quantity of potentially relevant material, the next step should be an examination of a sub-set of the apparently relevant articles. These may have been provided by commissioners as literature that has formed the rationale for the review, by stakeholders or identified from scoping searches by the Review Team. This will enable the Review Team to identify whether the studies reported in the articles are likely to be of sufficient quality to allow relatively robust synthesis (see section 4.3 for a detailed discussion of the critical appraisal process) and what sorts of study designs are appropriate to include.

Having critically appraised a sub-set of relevant studies, the Review Team should attempt to perform data extraction on these studies and, in so doing, explore the form that any potential synthesis may take. This will inform the development of a suitable data extraction spreadsheet for the full review, identifying which contextual and methodological information needs to be extracted alongside the types of data to be recorded from each relevant study. Any issues with data presentation should be noted at

this point, so that they may inform review planning. For example, Review Teams may find that data are not consistently presented in a suitable form and that they may need to contact original authors for missing or raw data. This process should inform the approach to the synthesis by allowing, for example; the identification of the range of data types and methodological approaches; the determination of appropriate effect size metrics and analytical approaches (e.g. meta-analysis or qualitative synthesis); and the identification of study covariates.

3.2.4 Estimating resource requirements

Whilst the process of scoping may seem like a time-consuming one, the benefits can be considerable and this early investment may be paid back several-fold by allowing the development of a comprehensive review plan as well as improved focus and efficiency throughout the later stages of the review. Scoping should provide an estimate of the timeline of the review so that a realistic budget can be prepared or the likely costs compared with the available resources.

3.3 Developing a review protocol

The review protocol acts as an *a priori* guide and reference to the conduct of the SR that reflects views of stakeholders and that the Review Team and their commissioners agree upon. Within the CEE approach it also acts as a registration of intent by the Review Team to conduct a CEE SR. Box 4 highlights an example of the way in which a review protocol was drafted and developed.

As in any scientific endeavour, the methodology should be established in advance and made available for scrutiny and comment at an early stage. Because reviews are retrospective by nature, the protocol is essential to minimise reviewer bias (e.g. resulting from ad-hoc decisions made during the review process) and make the process as rigorous, transparent, and well-defined as possible. The background section should present some kind of 'theory of change' or conceptual model that explains how the intervention or exposure factor is thought to have an impact or cause a change in the subject population. In more complex situations a proposed causal chain, linking intervention to outcome, may be necessary. Beside a formal presentation of the question and its background (the "real world" context), a review protocol sets out (informed by the scoping process – see above) the strategy for searching for relevant studies and defines relevance criteria for article screening (Section 4.2). The question elements defined in the question-setting stage provide the *a priori* inclusion criteria important for the objectivity and transparency of the review. They should also lead to a description of the kinds of evidence (e.g. study designs) that you would consider valid to include in the review. The protocol should also detail the likely methods to be used for critical appraisal, data extraction and synthesis, and state any conflicts of interest in the review including details of funding. The following sections on the conduct of the SR also provide guidance on how to structure and describe your plans in the appropriate sections of the protocol. The format of the protocol reflects the stages of the SR as shown in Box 5.

Box 4. Example of Review Protocol Development

Following a suggestion from the Swedish Environmental Protection Agency, the MISTRA Council for Evidence-Based Environmental Management (EviEM) studied the feasibility of a SR on how mountain vegetation is affected by reindeer grazing. Review scoping was conducted, and the outcome was promising enough that the Swedish EPA remained committed to the idea. Using scoping as a basis, EviEM then drafted a first version of a review protocol. A Review Team was organised, and stakeholders (the Swedish EPA and other agencies, ministries, Sami organisations, conservationists etc.) were called to a meeting to discuss the focus of the review.

The draft protocol and the stakeholder suggestions were then discussed at a meeting of the Review Team and a CEE SR specialist. During this meeting, the experts on the Review Team confirmed their understanding of the precise scientific scope of the review question and modified the primary question, the choice of search terms and the inclusion/exclusion criteria appropriately. The SR question finally arrived at was "What are the impacts of reindeer/caribou (*Rangifer tarandus* L.) on mountain and arctic vegetation?"

Following the agreement of the experts on the draft protocol, it was uploaded to the EviEM website to allow for public scrutiny, and stakeholders, in particular, were invited to comment on it. After revision, the final draft protocol was submitted to *Environmental Evidence* for peer review.

Box 5. Protocol template for submission to Environmental Evidence

Go to www.environmentalevidence.org/Documents/Instructions_for_Authors/EE_InstructionsforAuthors_PROTOCOLS.pdf for full instructions.

Background

The need for evidence. The need for an SR and the conceptual model or theory of change that underpins the question (e.g. the theory linking an intervention to an outcome). A conceptual framework provides a description of the context within which the question is being asked, assumptions being made and the underlying logic (a logic model may be included) which links elements of the question (e.g. the intervention with the outcome measure). A theory of change is related to the conceptual framework but should explain how the variable (e.g. intervention or exposure) is thought to bring about a change in the outcome. This may involve the formation of a causal pathway (e.g. from intervention to outcome).

Question

Presentation of question, any subquestions, and definition of question elements.

Methods

Searches

Here the proposed searches should be described in sufficient detail so as to be repeatable. The following subsections are a guide to the detail required on what will be searched and how the search will be conducted.

- Search terms and languages.
- Search strings and/or combinations of searches (search strings refer to combinations of terms using Boolean characters, combinations are methods used to set-up and pool different searches run separately).
- Estimating the comprehensiveness of the search.
- Publication databases that will be searched (e.g. Web of Science).
- Internet searches conducted (e.g. Google Scholar).
- Specialist searches - searches for grey literature: contacts, searches of organisational websites, use of specific search terms or strings, filtering or limitations.
- Supplementary searches such as bibliographical searches and stakeholders (individuals or groups) who will be approached for literature.

Study inclusion criteria

Here provide explanation about the rationale you propose to include/exclude articles based on the following aspects, so that this stage is transparent and replicable by any external reader.

- Relevant subject(s)
- Relevant intervention(s)
- Relevant comparator(s) (if appropriate)
- Relevant outcomes
- Relevant types of study design
- Relevant settings / regions / countries (if appropriate)
- Any tests for consistency of decision regarding inclusion/exclusion, at title, abstract, full-text level

Potential effect modifiers and reasons for heterogeneity

Provide a list of those effect modifiers (other variables that might influence the outcome) to be considered in the review and details of how the list was compiled (including consultation of external experts).

Study quality assessment

Describe here the approach you propose to use to critically appraise and assess quality of included studies.

Data extraction strategy

Describe here how you will collect and record data from included studies.

Data synthesis and presentation

Describe here the methods you might use to synthesise the collected data and any subsequent manipulation of the data set, sub-group analysis, sensitivity analysis and tests for bias.

Since the protocol sets out what the review aims to achieve, it is useful for getting the engagement of experts who may have data to contribute. Anyone reading the protocol should clearly understand the nature of the question and what type of evidence/data will inform it. To satisfy the philosophy of transparency in undertaking a SR, the review protocol is made openly available enabling other stakeholders who have not been contacted during the development stage to provide comments on the direction of the review. Comments received can then be taken into account by the authors and, if necessary, the protocol can be updated. Publishing and posting of protocols on the CEE website also acts as a record of which reviews are in progress, enabling others to see if a review is being conducted that may be of interest to them, or to prevent the initiation of a review on a topic that is already underway. Latest guidance on developing a review protocol can be found at www.environmentalevidence.org/Instructionsforauthors.html. For examples of completed protocols, visit the Environmental Evidence Library at: www.environmentalevidence.org/Library.htm

Although changes to the protocol are best avoided, it may become necessary during the course of a review to make revisions because of deviations from the proposed methods. These changes should be clearly documented within the final review so that transparency and repeatability can be maintained.

Protocols are plans of conduct and can never be fully comprehensive. They are judged in this context during the CEE peer review process. Consequently, **the acceptance and publication by CEE of a review protocol does not guarantee acceptance of the resulting systematic review**. Problems with the latter may occur due to conduct that was not mentioned or not fully transparent in the protocol.

Section 4

Conducting a CEE systematic review

4.1 Searching for studies

This section assumes that some literature scoping has been conducted in the planning phase (if you are planning a search strategy please see Review Scoping Section 3.2). The primary sources of data are usually primary studies reported within articles (review articles should not be included but can be used as a source of primary articles). Finding these articles is usually achieved by searching databases and catalogues covering relevant subject areas. Specialist sources and web searches are also sometimes employed. Databases and catalogues vary in the manner in which they can be searched. Searches may often have to be modified for different resources as a consequence. Database help files can be useful to ascertain the search capabilities, such as the symbols for wild card terms and the use of parentheses and Boolean terms. Many of the well-known databases allow complex search strings (e.g. Web of Knowledge, Scopus). However, others only allow searching with single keywords. Obviously, resource availability will constrain the numbers of literature sources used and search term permutations applied, which will also be subject to diminishing return due to duplication. Managing the citations within a bibliographic software package can be useful to assess the amount of duplication in articles captured as the search proceeds. It is important to record the methods used in all parts of the search so that others can judge the probability that important research has been missed and so that transparency and repeatability are maintained (see Section 3.3 “Developing a review protocol”).

The literature search is normally comprised of up to six distinct actions:

1. searching online literature databases and catalogues;
2. searching websites of organisations and professional networks;
3. searching the world-wide web;
4. searching bibliographies of key articles/reviews;
5. contacting key individuals who work in the area;
6. citation searches for key papers / included papers.

4.1.1 Searching online databases and catalogues

There are a number of general scientific electronic databases that may be useful for identifying relevant articles and data sets, such as Web of Science and Scopus. Access to most of these depend on library subscriptions, and so varies between institutions and organisations. Contacting a subject librarian or information specialist to identify and discuss the resources available is recommended at an early stage of the protocol development. As well as the general scientific databases, there are also some subject-specific databases that may contain relevant information and it may be necessary to search region-specific databases if SR questions have a regional focus.

Different databases and catalogues sample different subsets of the literature, and so multiple sources should be accessed to ensure the search is comprehensive and unbiased, but avoids unnecessary duplication. To ensure the search is comprehensive yet practical, it can be useful to consider the limitations of each database (some information should be available from the database provider) to ensure that at least one resource is searched to

cover each important subset of the literature, for example, theses and dissertations, peer-reviewed and non-peer-reviewed published articles and so-called grey literature that has not been formally published.

Different Review Teams often have access to different resources, and so the list of resources searched for each review will vary, but checking bibliographies and contact with authors should help to test if relevant articles are retrieved. **To minimise the problem of publication bias, searches for both published and unpublished 'grey' literature should be conducted** (e.g. Leimu & Koricheva 2005), a standard rarely satisfied in traditional reviews. Published results tend to be positive results (statistically significant results) and positive results are usually associated with large effect sizes. Unpublished results (found in 'grey literature') on the contrary are typically associated to negative results and small effect sizes (small magnitude of the effects studied). Hence, not considering negative results leads to overestimating the overall effect size. Including unpublished studies is necessary to obtain realistic overall effect sizes. However this may increase uncertainty in the final result, shown as wider confidence intervals. But acknowledging uncertainty is a basic goal. The next two stages of the literature search help to address this issue.

The general rule with bibliographic databases is to consider all hits listed as equally valid. Thus, if you get 2000 hits from a search string you should consider the relevance of all 2000 during your screening process (please note contrast with rule for search engines in section 4.1.3). Some databases provide filters (e.g. by subject). Use of such filters should be reported and their validity may need to be tested (e.g. by examining a sample of what they have excluded to confirm no loss of relevant articles). Ranking of hits by relevance to the search string may also be available as a function. Use of such functions is not normally advised as 'relevance to search string' is not necessarily equivalent to 'relevance to review question'. If ranking is used then justification is required and some form of testing should be reported. Bibliographic databases may allow searches on title and/or abstract and keywords or on full text. Title, abstract and keyword searches are normal but you should report on which search function you used.

4.1.2 Searching specialist organisations and professional networks

Many organisations and professional networks make documents freely available through their web pages, and many more contain lists of projects, datasets and references. Often, reports referred to on a website will be provided if an organisation is contacted. Searching these organisations and networks targets the grey literature that would not be identified in a conventional bibliographic database search. The list of organisations to be searched is dependent upon both the subject of the SR and any regional focus. Stakeholders should be consulted at the planning stage and asked to suggest relevant organisations.

Many websites have a search facility but their functionality is limited. A formal, repeatable search strategy should be employed but generalised guidance is difficult to provide. If feasible, hand searching of specific sources and visits to institutions (e.g. libraries and museums) may be useful in identifying further relevant studies or datasets. Keep records and fully report on the extent of your search.

4.1.3 Web searching

The Internet can be a useful tool for identifying unpublished and ongoing studies. Careful consideration must be given to the design of the search in order to ensure that it is as focused and specific as possible (Eysenbach et al. 2001); where this is not done, searching the web can be a time-consuming task, with relatively little useful data being returned. Thus, scoping (see above) should form a key component of any web searching strategy; piloting of potential search terms is essential, as any ambiguity is likely to return spurious results. An awareness of differences in search engine functionality is also important, as these may impose inconsistencies in approach but it is reasonable to tailor the search to the search engine to maximise its usefulness.

The indexable web is now some several billion pages in size and, whilst a wide range of engines exist to enable users to search these pages, none of these individually index more than small proportion of the total web. Overlap studies (e.g. Dogpile 2007) suggest that there is relatively little cross-over between the major search engines, with the proportion of results unique to each engine as high as 88%. Therefore, to ensure maximum retrieval of the available relevant information, it is advisable that multiple engines are searched. The use of meta-engines, which simultaneously search a number of individual engines, may also offer a part-solution to the problem of patchy coverage. In general, meta-engines should be treated with caution as many of these search only the free, poorer quality engines and, in cases where the most useful engines are included, limits on the number of hits returned from each engine often mean that such searches are considerably less useful than individual searches of the single engines (University of California 2008).

In addition to discrepancies in the extent of web coverage, there are disparities in the ways in which search engines rank their results. Page position within the results is not necessarily correlated to the relevance or quality of the documents retrieved. Although a closely-guarded secret, the ranking algorithms employed by major search engines are primarily based on one or more of a set of general principles. Most use the frequency and location of keywords as a fundamental guide of relevance, with those pages containing the specified search terms most frequently and higher up the document appearing at the top of the results listing (Hock 1999). Others determine relevance from a 'popularity' scoring system, whereby pages are ranked according to the number of sites that link to them, with high rankings associated with high 'link popularity' (Introna & Nissenbaum 2000). The majority of search engine providers effectively sell search positions in one form or another: most differentiate these 'sponsored' results from 'standard' ones but it is not uncommon for the former to be embedded within the main results page and be otherwise indistinguishable from the latter. Issues with engine ranking systems will become clear during the scoping phase, and should be used to guide decisions as to engine inclusion into the final review search strategy.

Boolean logic is supported to varying degrees by the major search engines, as is truncation using wildcards. These capabilities can be checked in the engine's accompanying 'help' files when selecting engines for inclusion. Many engines lack a nesting feature (use of parentheses) that would enable the use of more complex Boolean queries (Hock 1999). Where the nature of the study necessitates multi-element search strings, it may be possible to reconstruct these searches using the advanced search features offered: the

majority of the major search engines provide a “find all the words” and “find any of the words” feature which is particularly helpful.

When searching the internet for grey literature, it might be more efficient to run searches with a restriction on the file type to be returned, on the premise that these may be more likely to contain useful data than standard web pages. For example, by limiting the search to Excel spreadsheets, raw data that would otherwise have ranked low in an unrestricted search may be captured. Most search engines provide the option of file restriction to a range of formats (.pdf, .doc, .xls, .rtf, etc.) and this is usually accessed via the engine’s “advanced search” page. A small number of engines (e.g. Scirus) allow the selection of multiple file formats per search: most do not, however, and where this is desired, visual sorting of the search results may be the only solution. Searches such as these should be recorded as part of the search strategy.

In addition to the more general search engines, the incorporation of specialised subject gateway searches into web searching strategies may be helpful. Databases such as Intute.ac.uk, ScienceResearch.com and AcademicInfo.net, contain links to hand-selected sites of relevance for a given topic or subject area and are particularly useful when searching for subject experts or pertinent organisations, helping to focus the searching process and ensure relevance.

Perhaps most importantly, remember that the hits you achieve using search engines are not to be viewed in the same way as hits achieved using literature databases (see above). Web searching may achieve very large numbers (e.g. millions) of hits and relevance may decline rapidly as you progress through the list. Specific guidance on how much searching effort is acceptable is difficult to give. In the medical literature, papers sometimes cite a “first 50 hits” approach (e.g. Smart & Burling 2001), whereby the first 50 results for each search are viewed in full. However, this appears to be an arbitrary number, and is more likely based upon the resources available to the Review Team than a reflection of the extent of searching required to effectively capture the most-relevant grey literature available. Given that the actual number of hits retrieved is review-specific, related both to the search terms used and the quantity of information available, in some instances there may be a case for modifying the recommended search limits (e.g. if there are particularly large or small numbers of relevant hits). Thus, in order to provide a consistent and practical way to limit web searching, we would recommend, at a minimum, the full viewing of each of the first 50 hits but would not advise viewing more than the first 100 (unless authors feel there is a good reason to do so). The proportion of relevant material retrieved in this subset will then provide an indication as to the potential utility of examining further hits. Review Teams must also decide the extent to which links from the original ‘hits’ to potentially relevant material will be followed, and must make sure the chased links are recorded in each instance (if a pre-determined limit is not set). It is important, both for citation purposes (should an online document be selected for inclusion in the review) and to ensure transparency and repeatability, that the dates of the web searching phase are clearly documented: the use of a simple recording form will facilitate this.

4.1.4 Searching bibliographies

It is possible to use the bibliographies of relevant articles to search for other relevant articles. This can be done manually or through databases with appropriate functionality. This approach can also involve stakeholders and experts, who may be contacted to identify other potential sources of data that may have been missed by the original search.

Where previous reviews are identified their bibliographies should be searched for relevant primary studies. These bibliographies can form a useful test of the comprehensiveness of the search strategy. If studies are found in bibliographies that were not found in the main search, you should investigate reasons for this and, if necessary, refine the search (e.g. modify search stings).

It can be useful to follow 'leads' from bibliographies in a form of chain sampling (often referred to as 'snowballing', 'citation chasing' or 'pearl growing'). This strategy may be particularly useful for checking the comprehensiveness of your database search.

4.1.5 Recording the search process

It is vital that the search strategy is transparent and repeatable. Both for the purposes of reporting and for sharing information within the Review Team, a method of recording outcomes and details of specific actions should be employed. Specifically, for each source searched a record should be made of: the dates of individual searches; the full list of search terms employed and how these were combined; any changes to the default search settings of the source used; the nature of the search (e.g. keywords, topics, or full texts) and other search options (e.g. lemmatization); the removal of duplicates if automatically carried out when downloading results; and all the results returned by each search. For an example of best practice search strategy recording, see Mant et al. (2011). Box 6 indicates the types of information that should be recorded and the levels of detail required when documenting the search process.

Box 6. Examples of the types of information that should be recorded at each of the stages in the search process (example followed by type of information)

Database searches	•e.g. Web of Knowledge; 23/02/2012, topic search, "(reindeer OR Rangifer OR caribou) AND (graz* OR herbivory OR brows* OR trampl*)" - 328 hits saved as EndNote library (/WoK search string 3.2.enl) •**Database, date, search details, search terms, hits, outputs (including any replicate removal)**
Organisational web searches	•e.g. International Centre for Reindeer Husbandry website search facility (http://icr.artcicportal.org); 05/03/12, "graz* AND vegetation" - 40 hits all checked at full text, 4 relevant articles saved as PDF/HTML files and catalogued in spreadsheet •**Web site, search type (i.e. manual search/automatic search/publications section scanned), date, search terms, result-checking method, hits, outputs**
Web search engines	•e.g. Google Scholar advanced search; 03/03/2012, reindeer AND grazing - first 100 hits checked at full text, 11 relevant articles saved to EndNote library (/Google Scholar search string 4.1.enl) •**Web site, search type, date, search terms, result-checking method, hits, outputs**
Bibliographic checking	•e.g. Suominen & Olofsson (2000). Four articles in reference list assessed as relevant at title-level and missed by searches. Two of these relevant at abstract level. One relevant at full text. •**Reference, number of titles relevant, number of relevant titles missed by searches, number of missed titles relevant at abstract, number of missed abstracts relevant at full text**
Calls for information/ expert contact	•e.g. Emilia Nordin, Swedish Environmental Protection Agency (e.nordin@sepa.se), emailed 23/04/12 requesting submission of unpublished research, responded 03/05/12 with two annual reports, email saved in templates folder and reports saved in 'Submitted Evidence' folder •**Name, affiliation, contact method, date, notes, response, outputs**

4.1.6 Managing the results of your searches

The full list of results obtained from each search and source must be recorded for transparency. This is more efficiently done through the use of reference management software, such as Endnote, Reference Manager, and Refworks, and many literature databases now allow the exporting of search results directly into such software. Other sources, such as many web engines, do not allow this however, and in such cases it is recommended that search results are initially saved into a spreadsheet or Word file. If resources allow, these may be manually entered into reference management software when all search results are combined.

When working with many different stakeholders, and when SRs are undertaken to address conflicts of evidence, it is important to ensure that the databases of search results are available for all to scrutinise. In this way, one can question why a given article is not reported in the results and can check whether it is a limitation of the search itself (the article not retrieved) or due to the study not meeting the criteria applied for screening and appraisal. As some stakeholders may not have access to the reference management software, exporting the searches into a simple database such as Excel may be a solution. This should be anticipated during the first stages of the review.

4.2 Screening articles for relevance

Once searching is complete, relevant articles must be efficiently selected without wasting resources examining irrelevant articles in too much detail. The reference management software should enable simple removal of duplicate records, which can reduce substantially the initial number of articles. Selecting only relevant articles from a potentially large body of initial literature requires the reviewer to use a-priori inclusion criteria stated in the protocol. These criteria relate directly to the elements of the question (e.g. PICO, PECO, Table 1).

Inclusion criteria can be applied at different levels of reading to impose a number of filters of increasing rigor and thus relevance assessment is normally a staged process. The exact approach to this process is a matter of preference, although it is recommended that at least two filters are applied:

1. a first reading of article titles and abstracts to efficiently remove spurious hits;

and for those passing this stage,

2. assessment of the full text.

The first stage may be split in two if desired, so that the first stage is assessment of titles, then the abstracts of those included. Whichever approach is chosen, reviewers should be conservative so as to retain articles if there is reasonable doubt as to whether all the inclusion criteria are met. For instance, on reading title and abstract, it is often difficult to assess whether a study has key elements of design such as replication or valid comparator. If such basic information is absent (or there is no abstract) then the article should be retained and the full text examined.

It is good practice at the beginning of the abstract assessment stage for two reviewers to undertake the same process on a random sub-sample of articles from the original list (the recommended sample is a minimum of 50 or 10% up to a maximum of 200 references). To check for consistency in the interpretation of the selection criteria, reviewer relevance decisions can be compared by performing a kappa analysis (Box 7), which adjusts the proportion of records for which there was agreement by the amount of agreement expected by chance alone (Cohen 1960; Edwards et al. 2002). A kappa rating of

'substantial' (0.5 or above) is recommended to pass the assessment. If comparability is not achieved, then the criteria should be further developed by redefining the scope and interpretation of the question elements. Ideally kappa analysis should be repeated on a new sample of articles, if resources allow, to check the accuracy of the redefined criteria.

Box 7. Checking for reviewer consistency and the kappa test

Example given is from the start of the abstract inclusion stage for the systematic review entitled "How effective is 'greening' of urban areas in reducing human exposure to ground level ozone concentrations, UV exposure and the 'urban heat island effect'?" (Bowler et al. 2010).

A relevance assessment of the titles of retrieved articles was conducted by a single reviewer. At the start of the abstract inclusion stage, reviewer bias was assessed by kappa analysis; two reviewers applied the inclusion criteria to 25% of the articles (n=213). The number of papers accepted or rejected by both reviewers, and the number of discrepancies was recorded as in the table below.

		Reviewer A	
		Rejected	**Accepted**
Reviewer B	**Rejected**	96	35
	Accepted	11	71

The kappa statistic was then calculated to measure the level of agreement between reviewers (see www.inside-r.org/packages/cran/fmsb/docs/Kappa.test). The kappa score was 0.57 (95% C.I: 0.46, 0.68), which indicated 'moderate' agreement between the reviewers (Landis and Koch, 1977). Discussion of the discrepancies in inclusion decisions followed, and agreements were sought to strengthen the consistency in interpretation of relevance for the remaining articles.

Remaining articles, which have not been excluded after reading their title and abstract, should be viewed in full to determine whether they contain relevant and usable data. Independent checking of a sub-sample by kappa analysis can be repeated at this stage. Obtaining the full text of all articles can be very time consuming and a realistic deadline may have to be imposed and a record kept of those articles not obtained. Shortlists of relevant articles and datasets should be made available for scrutiny by stakeholders and subject experts. All should be invited, within a set deadline, to identify relevant data sources they believe are missing from the list. Reviewers should be aware that investigators often selectively cite studies with positive results (Gotzsche 1987; Ravnskov 1992); thus, checking bibliographies and direct contacts must be used only to augment the search.

4.2.1 Recording the selection process

During the selection process the fate of each article captured during the search should be recorded. Libraries of those articles excluded at the title/abstract and full text assessments should be retained as material supplementary to the SR for future checking and for transparency of decision making. These lists can be posted alongside completed reviews in the CEE Library (currently, providing a list of articles excluded at full text assessment is mandatory). For examples of good practice in documenting the fate of reviewed articles, see Johnson et al. (2011) and Mant et al. (2011). Both reviews provide detailed diagrams of the selection process, kappa test results for reviewer agreement, and supplementary appendices of the fate of reviewed articles/studies from full text assessment onwards. A template is provided below (Figure 4).

It is important to note here the distinction between an 'article' and a 'study'. Initial searches identify articles which are screened for relevance (these include scientific papers and organisational reports). Such articles may contain more than one study or a study may be reported in more than one article. Hence the number of articles accepted at full text is frequently not the same as the number of studies from which data are extracted.

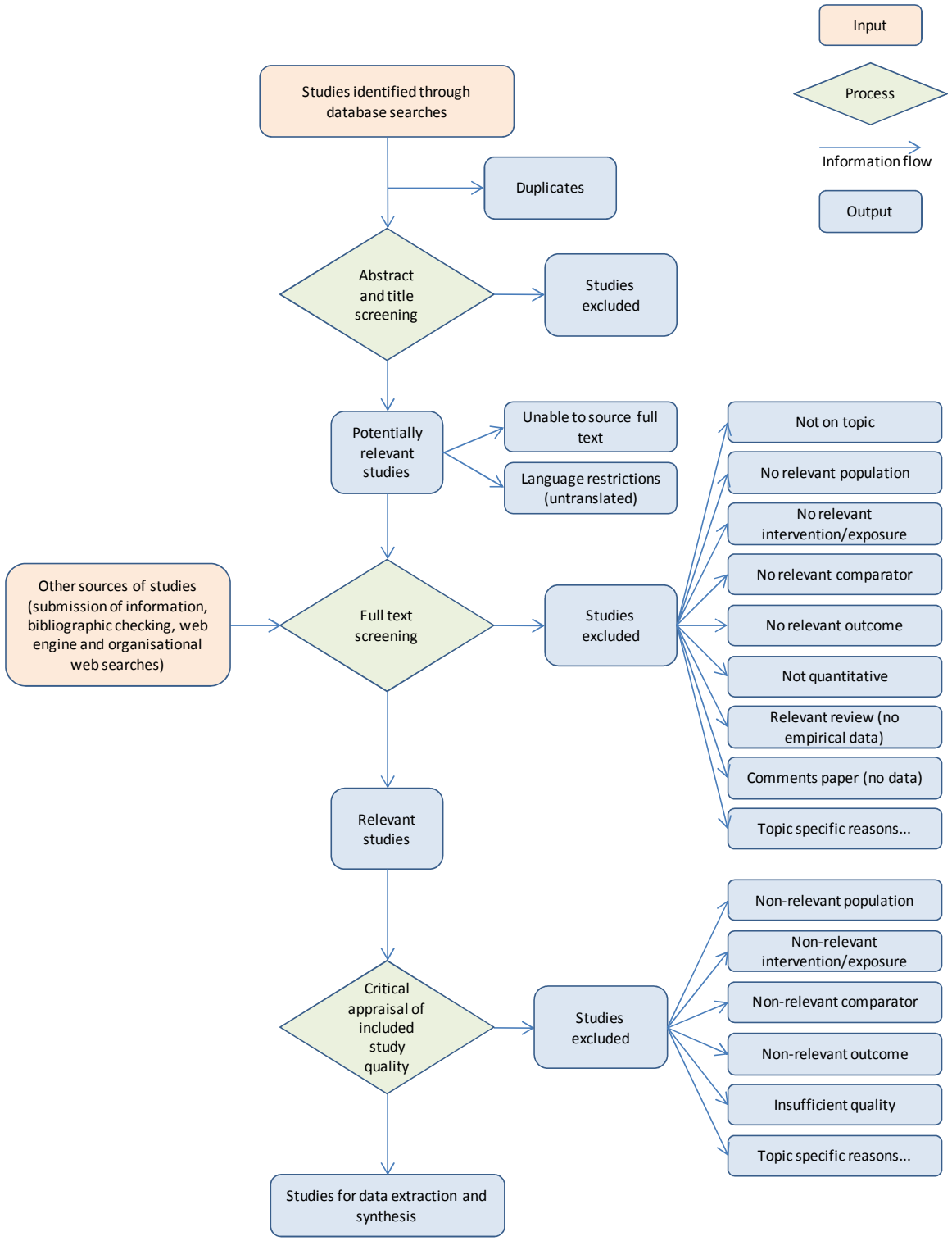


Figure 4. Template showing filtering of literature through SR stages

4.3 Critical appraisal of study quality

Some primary studies provide higher quality evidence than others. Assessing the comparative quality of the included studies (often referred to as critical appraisal) is of key importance to the resulting value of the SR (see examples in Box 8 and Table 3). It can form a basis for the differential weighting of studies in later synthesis or partitioning of studies into subgroups for separate analyses.

The precise order in which critical appraisal and data extraction (see section 4.4) are undertaken varies from SR to SR. In our experience, there is frequently an iterative relationship between the two. Hence, there is no set guideline as to which should come first.

Study quality assessment requires a number of decisions about the absolute and relative importance of different sources of bias and data quality elements common to environmental data, particularly the appropriateness of temporal and spatial scales. It is therefore vital that the assessment process be standardised and as transparent and repeatable as possible. Quality is a relative term and its measurement and scale are very dependent on the question being addressed. It may be helpful to breakdown the concept of quality into two separate units; study reliability and study relevance.

4.3.1 Study reliability

Reliability is often considered in terms of internal validity of the study methodology; the extent to which its design minimises susceptibility to bias. Four sources of systematic bias that may threaten the internal validity of a study form the basis of a methodological quality assessment (Feinstein 1985; Moher et al. 1995; Moher et al. 1996; Khan et al. 2003).

Selection bias results from the way that comparison (e.g. treatment and control) groups are assembled (Kunz 1998) and is a primary reason for randomisation in studies. This bias is common in environmental management because interventions or treatments are applied to entire sites and analogous controls often do not exist (e.g. marine protected areas). A common problem is that many studies with comparators are confounded at baseline (i.e. the treatment and control groups were not the same at the beginning of the experiment). Randomised allocation to treatments and controls is often not feasible to address this problem, but even worse, the baseline may not be measured so the extent of the problem cannot be assessed.

Performance bias refers to systematic differences in the attention given to subjects in the comparison groups and is dealt with by the experimenter being unaware of which are treatments and which controls (blinding) (Shultz 1995). It may also refer to differences in exposure or intervention received: an important consideration in some environmental contexts.

Measurement or detection bias refers to systematic differences incurred when knowledge of the intervention influences the assessment of the results in the comparison

groups and is also addressed by blinding (Shultz 1995). Blinding is generally not possible in environmental sciences and ecology and the extent of detection bias will therefore vary, depending on the rigour and objectivity of sampling methodology (e.g. when measuring abundance, percent cover assessed by eye is subject to greater potential detection bias than frequency).

The fourth, **attrition bias** (systematic differences between the comparison groups in the loss of samples) is common in population studies (e.g. individuals who die are excluded from an outcome group). This can be addressed by analysing all the data, but access to raw data may be a pre-requisite to quantify the impact of attrition bias.

In an ideal world, each data set included in a SR should be of high methodological quality, thus ensuring that the potential for error and bias is minimised and that any differences in the outcome measure between experimental groups can be attributed to the exposure or intervention of interest. To determine the level of confidence that may be placed in selected data sets, the methodology employed to generate each one must be critically appraised, using a transparent and consistent framework, to assess the extent to which it is likely to prevent systematic errors or bias (Moher et al. 1995). However, the nature of the critical appraisal and the hierarchy employed is dependent on the nature of the question and the 'theory of change' (see Section 3.3). The Review Team should justify their approach and not blindly follow an established methodology.

In the health sciences, a hierarchy of research methodology is recognised that scores the value of the data in terms of the scientific rigour; the extent to which the methodology seeks to minimise error and bias (Stevens & Milne 1997). The hierarchy of methodological design can be viewed as generic and has been translated from medicine to environmental sciences (Pullin & Knight 2003), but these generic hierarchies are crude tools and usually just a starting point and can rarely be used without modification to ensure relevance to individual review questions. Where a number of well-designed, high-quality studies are available, others with inferior methodology may be demoted from subsequent quantitative analysis to narrative tabulation, or rejected from the SR entirely. However, there are dangers in the rigid application of hierarchies as the importance of various methodological dimensions within studies will vary, depending on the study system to which an intervention is being applied. For example, a rigorous methodology, such as a randomised controlled trial (RCT), applied over inadequately short time and small spatial scales could be viewed as superior to a time series experiment providing data over longer time and larger spatial scales that were more appropriate to the question. The former has high internal validity but low external validity or generalisability in comparison to the latter. This problem carries with it the threat of misinterpretation of evidence. Potential pitfalls of this kind need to be considered at this stage and explored in covariate analyses (e.g. experimental duration or study area: see Downing et al. 1999 and Côté et al. 2001, respectively) or by judicious use of sensitivity analysis (see below).

As a consequence, authors may use existing checklists of critical appraisal tools as a basis for their specific exercise, but they should either explain why they use them as such (no modification, because not considered to be needed, and why) or adapt them to their own case-study review, in which case the decisions made must be stated and justified (see Gough et al. 2012).

We suggest that review-specific *a priori* assessment criteria for appraising the quality of methodology are included in the protocol and two or more assessors should be used. The subjective decisions may be a focus of criticism; thus, we advocate consultation with subject experts and relevant stakeholders before moving on to data extraction. Pragmatic grouping of studies into high, medium and low quality based on simple but discriminatory checklists of "desirable" study features may be necessary if sample sizes are small and do not allow investigation of all the study features individually (for example, Felton et al. 2010, and Isasi-Catalá 2010).

The scope of CEE reviews is broad and often interdisciplinary and therefore we seek to be inclusive of different forms of evidence provided their strengths and weaknesses are properly appraised and comparative study weightings are appropriate.

4.3.2 Study relevance

Relevance is often considered in terms of the external validity of the study; how transferable is it to the context of the question? As noted above, some studies can be of high internal validity (low risk of bias) but may be misleading on account of low external validity (low relevance). A simple example is a high quality study that has been conducted outside the geographical region or in a slightly different ecosystem than the one of interest.

Appraisal of study relevance can be a more subjective exercise than appraisal of study reliability. Scoring the external validity of a study may require the construction of review-specific criteria formed by fit to the question elements or similar subjective measures (see Gough et al. 2012 for examples).

Box 8. Examples of Good Practice in Critical Appraisal

The application of critical appraisal of study quality can be broken down into the following four steps in practice;

- establishment of quality assessment criteria;
- deciding on the impact of these criteria on review activities;
- enacting quality assessment and assigning criteria;
- determining the impact of these criteria on review findings.

Example 1. The Importance of Nature to Human Health
In a review of the importance of nature for health, Bowler et al. (2010) undertook critical appraisal of included studies using assessment criteria adapted from a SR of the health literature (specifically, nursing). These criteria included an assessment of: specific methodological bias (e.g. participant self-selection bias), the use of randomisation, the presence of baseline data, and the presence of other confounding variables. Five studies assessed in this review were excluded due to low quality, with the remaining study quality criteria being shown across studies in a bar chart in the results. Finally, study quality weighting was used in sensitivity analyses to compare the results of higher and lower study quality; indicating, for example, that studies with a lower quality score reported a larger effect of nature on tranquillity/calmness than those of higher quality.

Example 2. Peatland Management and Carbon Cycling/GHG Fluxes
In a SR of the impacts of land management activities in lowland peatland ecosystems on greenhouse gas fluxes and carbon cycling, three main experiment types were identified: eddy covariance towers, gas flux chambers, and extractive sampling of soil or soil pore water/air. Prior to assessing the quality of each included study, the external validity (relevance) of each article was assessed in detail. This process ensured that each aspect of the study's PICO elements was relevant to the SR question. Subsequently, critical appraisal assessed two types of methodological information. Firstly, general experimental design assessment examined matching of comparator and intervention sites, study season and length, and the time since the intervention occurred. Secondly, specific details relevant to each of the three potential experimental types were assessed; for example, the presence of mitigation measures for trampling around gas flux chambers, the height of eddy covariance towers, and the frequency of sampling. Reasons for possible concern were highlighted during the critical appraisal of each study, and a decision was made as to whether to exclude or include. A quality score based on the presence/absence of bias, appropriateness of controls, precision of methodological design and the presence of confounding variables was given to each study. This score was checked in a subset of studies by a second reviewer and modifications made where necessary. Scores were included as an explanatory variable in meta-analysis as part of a sensitivity analysis to investigate potential differences between studies of higher and lower quality.

For transparency of reporting, tables of data quality assessment should be included as an appendix or supplementary material. The data quality assessment can be incorporated in narrative synthesis tables if appropriate (see 4.5.1).

Table 3a. The data quality assessment of a study included in a SR examining impacts of land management on carbon cycling and greenhouse gas fluxes in boreal and temporal lowland peats.

Study 1

Methods	Site comparison, GHG flux measured weekly for whole year using closed chambers.
Population	Forested peatlands in Slovenia.
Intervention(s)	Drained plot (19th Century).
Comparator	Undrained plot.
Comparator-matching	Comparator plots close to intervention but distances not disclosed. Soil types moderately different (intervention=rheic hemic histosol (dystric), control=rheic fibric histosol (dystric)).
Outcomes	N_2O, CO_2, and CH_4.
Study design	CI (comparator-intervention).
Level of replication	Plot-level (1 treatment, 1 control), 3 pseudoreplicate samples per plot.
Sampling precision	Weekly measurements 60 minutes each with 3 samples per hour (regression modelling), time=zero measurement.
Confounding variables	Permanent collars account for soil disturbance, foil-covered chambers reduce temperature effects.
Conclusions	Small effective sample size, but good outcome measurement precision. High external validity to SR question. Include in review accounting for low replication.

Study 2

Methods	Site comparison, GHG flux measured once using closed chambers.
Population	Ombrotrophic fen and minerotrophic bog in Finland.
Intervention(s)	Drained plots (30 years previously).
Comparator	Undrained plots.
Comparator-matching	pH, %N and water table depth measured in all plots and appear similar.
Outcomes	CO_2 and CH_4.
Study design	CI (comparator-intervention).
Level of replication	Plot-level (one treatment, one control); two regions, one with only ombrotrophic bog, other with ombrotrophic bog and minerotrophic fen. Each site has drained and undrained

	counterparts. Each site must be treated as a separate study due to substantial differences in plot soil characteristics.
Sampling precision	One sample per plot taken between two and five times over seven month period (exact number unspecified).
Confounding variables	Drained and undrained plots actually only differ very slightly in water table depth, so stated exposure difference may have no real impact. Data extrapolated from very low degree of pseudoreplication (2 to 5 samples over 7 month period).
Conclusions	Drained and undrained plots compared in study but also shown to have minimal differences in water table depth (external validity questionable).

At the end of this stage (if not before) it should become clear what form or forms of synthesis will be possible with the available data. There are a number of different pathways from this point and therefore the following two sections become more speculative and general in terms of the guidance given. They also become more reliant on guiding the reader to more detailed information sources.

4.4 Data extraction

Alongside critical appraisal of study quality, the Review Team should extract and collate the relevant data generated by each study. Extracted data should be recorded on carefully designed spreadsheets and undertaken with the appropriate synthesis in mind (see next section).

Great care should be taken to standardise and document the process of data extraction, the details of which should be recorded in tables of included studies to increase the transparency of the process (Table 4). To some extent data extraction can be guided by *a priori* rules, but the complexity of the operation means a degree of flexibility must be maintained. Sensitivity analyses can be used to investigate the impact of extracting data in different ways when there is doubt about the optimum extraction method.

Good practice for data extraction could involve the following steps, which improve transparency, repeatability and objectivity:

- Data extractions should always present the primary data as reported in the primary study; if any corrections or transformations are needed these should be presented additionally so that all data are traceable to the primary study
- Notation of the location of data within each article and means of extraction if data are located within figures.
- Description of any pre-analysis calculations or data transformations (e.g. standard deviation calculation from standard error and sample size (e.g. Felton et al. 2010 and Smith et al. 2010), and calculation of effect sizes.

- Details of a pre-tested data extraction form.
- Data extraction in a subset of articles by multiple reviewers and checking, for example with a kappa test (for human error/consistency) (e.g. Benítez López et al. 2010, Showler et al. 2010).
- Inclusion of appendixes of extracted information (e.g. Doerr et al. 2010, Bowler et al. 2010 and Isasi-Catalá 2010).
- Contact made with authors requesting data where it is missing from relevant articles (e.g. McDonald et al. 2010 and Eycott et al. 2010).

Table 4. Example of a data extraction form from a review examining the impact of instream devices on salmonids.

Reference	**Binns & Remmick (1994)**
Location	Huff Creek, Idaho, USA
Subject	*Oncorhynchus clarki utah* (Bonneville cutthroat trout)
Intervention	Instream habitat structures (36 wooden dams, 9 rock plunges, wooden double deflector, rock deflector, 14 small rock grade controls) rock riprap, fencing of banks
Methodology	Before and after monitoring
Sources of bias	Confounding impacts concurrent with the habitat improvement are probably the most important sources of bias. Post improvement droughts occurred resulting in a likely under-estimate of effectiveness.
Outcomes Habitat quality index (HQI) Trout numbers	post intervention: n 6, m 38, sd 2; pre intervention: n 6, m 30, sd 2 post intervention: n 6, m 170, sd 59; pre intervention: n 6, m 35, sd 18
Reasons for heterogeneity	Monitoring time 11 years. Discharge is extremely variable with a mean of 6ft3/s, stream gradient (1%), proportion of cobbles in substrate (common in half of river, estimated at 25%),degree of existing modification (heavy grazing but river unmodified- low), distance from source (6km), water quality (no information), size of stream (small stream >5m), canopy cover (low >5%).
Population change/ habitat preference data Extraction	Habitat quality pre and post treatment, from text and Figure 6. Trout numbers from text and Table 2. *n* is the number of sites. Maximum time range was used for post treatment assessment (11 years). Some data are presented for individual sites which allow some separation of features. This was not extracted i) to maintain independence, ii) because no pre treatment assessments are available at a site level

Notes	HQI was evaluated for cut throat trout. Population sizes were estimated using electrofishing (Armour et al. 1983) with degree of population fluctuation assessed as in Platts & Nelson (1988). Much other data regarding both physical habitat and trout was presented but not extracted.
References	Armour, C.L., Burnham, K.P., and Platts, W.S. (1983) Field methods and statistical analysis for monitoring small salmonid streams. U.S. Fish and Wildlife Service FWS/OBS 83/33. Platts, W.S. and Nelson, R.L. (1988) Fluctuations in Trout populations and their implications for land-use evaluation. North American Journal of Fisheries Management 8. 333-345.

Note in the table above, the reporting of raw data from which effect sizes were calculated, reference to data sources and information about decisions regarding which data to extract to maintain independence.

At this stage, it may be necessary to reject articles that are seemingly relevant but do not present data in extractable format (e.g. if they do not report standard deviations for control and treatment group(s) or the information required to calculate the statistic). If possible, authors of such articles should be contacted and asked whether they can provide data in a suitable format. Contacting authors for data is not normal practice in environmental science and can be met with surprise and indignation, but it is important to develop the culture and expectation of data accessibility, particularly when the research was publicly funded.

In some cases, where the information required is not presented and cannot be obtained from authors, data can be converted into an appropriate form without problems. For example, it is relatively straightforward to substitute standard deviation for standard errors, confidence intervals, *t*-values, or a one-way *F*-ratio based on two groups (Lipsey & Wilson 2001, Deeks et al. 2005). Where missing data cannot be substituted, it can be imputed by various methods. Imputation is a generic term for filling in missing data with plausible values. These are commonly derived from average or standardised values (Deeks et al. 2005), but also from bootstrapped confidence limits (Gurevitch & Hedges 2001) or predicted values from regression models (Schafer 1997). Alternatively, data points can be deleted from some analyses, particularly where covariates of interest are missing. Such pragmatic imputation or case deletion should be accompanied by sensitivity analyses to assess its impact.

The impacts of imputation or case deletion can be serious when they comprise a high proportion of studies in an analysis. Case deletion can result in the discarding of large quantities of information and can introduce bias where incomplete data differs systematically from complete (Schafer 1997). Likewise, imputing average values or predicted values from regressions distorts covariance structure resulting in misleading p-values, standard errors and other measures of uncertainty (Schafer 1997). Where more than 10% of a data set is missing serious consideration should be given to these problems. More complex imputation techniques are available (see Schafer 1997) and should be

employed in consultation with statisticians. If this is not possible, the results should be interpreted with great caution and only presented alongside the sensitivity analysis.

It is difficult to perform formal kappa analysis on the repeatability of data extraction, but some attempt to verify repeatability should be made. A second reviewer should check a random subset (recommended sample of minimum 25%) of the included studies to ensure that the *a priori* rules have been applied or the rationale of deviations explained. This also acts as a check on data hygiene and human error (e.g. misinterpretation of a standard error as a standard deviation). Where data extraction has limited repeatability it is desirable to maintain a record of exactly how the extraction was undertaken on a study by study basis. This maintains transparency and allows authors and other interested parties to examine the decisions made during the extraction process. Particular attention should be paid to the data used to generate effect sizes. Such data extraction forms should be included in an appendix or supplementary material.

4.5 Evidence synthesis

This stage includes an overview of different forms of synthesis, narrative, quantitative and qualitative. All SRs should present some form of narrative synthesis and many will contain more than one of these approaches (e.g. Bowler et al. 2010). It is not our intention to give detailed guidelines on synthesis methods here. Detailed descriptions can be found elsewhere (e.g. Borenstein et al. 2009 for meta-analysis).

4.5.1 Narrative synthesis

A narrative synthesis is often viewed as preparatory when compared with quantitative synthesis and this may be true in terms of application of analytical rigour and statistical power but narrative synthesis has advantages when dealing with broader questions and disparate outcomes. Often narrative synthesis is the only option when faced with a pool of disparate studies of relatively high susceptibility to bias, but such syntheses can also accompany quantitative syntheses in order to provide context and background and help characterise the full evidence base. Some form of narrative synthesis should be provided in any SR, simply to present the context and overview of the evidence. A valuable guide to the conduct of narrative synthesis is provided by Popay (2006).

Narrative synthesis requires the construction of tables that provide details of the study or population characteristics, data quality, and relevant outcomes, all of which are defined *a priori*. The tendency toward simple vote counting (e.g. how many studies showed a positive versus negative outcome) at this stage should be avoided. Narrative synthesis should include an evaluation of the measured effect and the manner in which it may be influenced by study quality (including internal and external validity). Where the validity of studies varies greatly, reviewers may wish to give greater weight to some studies than others. In these instances it is vital that the studies have been subject to standardised *a priori* critical appraisal with the value judgments regarding both internal and external validity clearly stated. Ideally these will have been subject to stakeholder scrutiny prior to

application. The level of detail employed and emphasis placed on narrative synthesis will be dependent on whether other types of synthesis are also employed. An example of an entirely narrative synthesis (Davies et al. 2006) and a narrative synthesis that complements a quantitative synthesis (Bowler et al. 2010) are available in the CEE Library.

Recording of key characteristics of each study included in a narrative synthesis is vital if the SR is to be useful in summarising the evidence base. Key characteristics are normally presented in tabular form and a minimum list is given below.

Article reference
Subject population
Intervention/exposure variable
Setting/context
Outcome measures
Methodological design
Relevant reported results

It should be noted here that the interpretation of the results provided by the authors of the study is normally not summarised as this could simply compound subjective assessments or decisions.

4.5.2 Quantitative synthesis

Usually, when attempting to measure the effect of an intervention or exposure a quantitative synthesis is desirable. This provides a combined mean effect and a measure of variance within and between studies. Quantitative syntheses can be powerful in the sense of enabling the study of the impacts of effect modifiers and increasing power to predict outcomes of interventions or exposures under varying environmental conditions.

Meta-analysis is now commonly used in ecology (e.g. Arnqvist & Wooster 1995; Osenberg et al. 1999; Gurevitch & Hedges 2001; Gates 2002); consequently, we have not treated it in detail here. Meta-analysis provides summary effect sizes with each data set weighted according to some measure of its reliability (e.g. with more weight given to large studies with precise effect estimates and less to small studies with imprecise effect estimates). Generally, each study is weighted in proportion to sample size or inverse proportion to the variance of its effect. In other cases a more subjective weighting can be applied provided the methodology is transparent and repeatable.

Pooling of individual effects can be undertaken with fixed-effects or random-effects statistical models. Fixed-effects models estimate the average effect and assume there is a single true underlying effect, whereas random-effects models assume there is a distribution of effects that depend on study characteristics. Random effects models include inter-study variability (assuming a normal distribution); thus, when there is heterogeneity, a random-effects model has wider confidence intervals on its pooled effect than a fixed-effects model (NHS CRD 2001; Khan et al. 2003). Random-effects models or mixed models (containing both random and fixed effects) are often most appropriate for

the analysis of ecological data because the numerous complex interactions common in ecology are likely to result in heterogeneity between studies or sites. Exploration of heterogeneity is often more important than the overall pooling from a management perspective, as there is rarely a one-size-fits-all solution to environmental problems.

Relationships between differences in characteristics of individual studies and heterogeneity in results can be investigated as part of the meta-analysis, thus aiding the interpretation of ecological relevance of the findings. Exploration of these differences is facilitated by construction of tables that group studies with similar characteristics and outcomes together. Datasets can be stratified into subgroups based on populations, interventions, outcomes, and methodology. Important factors that could produce variation in effect size should be defined *a priori* and their relative importance considered prior to data extraction to make the most efficient use of data. Differences in subgroups of studies can then be explored.

If sufficient data exist, meta-analysis can be undertaken on subgroups and the significance of differences assessed (see Appendix B). Such analyses must be interpreted with caution because statistical power may be limited (Type I errors possible) and multiple analyses of numerous subgroups could result in spurious significance (Type II errors possible). Alternatively, a meta-regression approach can be adopted whereby linear regression models are fitted for each covariate, with studies weighted according to the precision of the estimate of treatment effect in a random-effects model (Sharp 1998).

Despite the attempt to achieve objectivity in reviewing scientific data, considerable subjective judgment is involved when undertaking meta-analyses. These judgements include decisions about choice of effect measure, how data are combined to form datasets, which data sets are relevant and which are methodologically sound enough to be included, methods of meta-analysis, and the issue of whether and how to investigate sources of heterogeneity (Thompson 1994). Reviewers should state explicitly and distinguish between the *a priori* and *post hoc* rationales behind these decisions to minimise bias and increase transparency.

If possible, a quantitative synthesis should be accompanied by a test for publication bias. Positive and/or statistically significant results are more readily available than non-significant or negative results because they are more likely published in high-impact journals and in the English language. Whilst searching methodology can reduce this bias, it is still uncertain how influential it might be. There are a number of tests for publication bias that assume a normal distribution of effects from a group of included studies. One example is the Egger test producing a funnel plot (Egger et al. 1997) (see Figure 5, below, for an example funnel plot). Another approach is to calculate the fail safe number, which is the number of null result studies that would have to be added to a meta-analysis to lower the significance of a result to a specified level (e.g. where it would be considered non-significant), but see Scargle (2000).

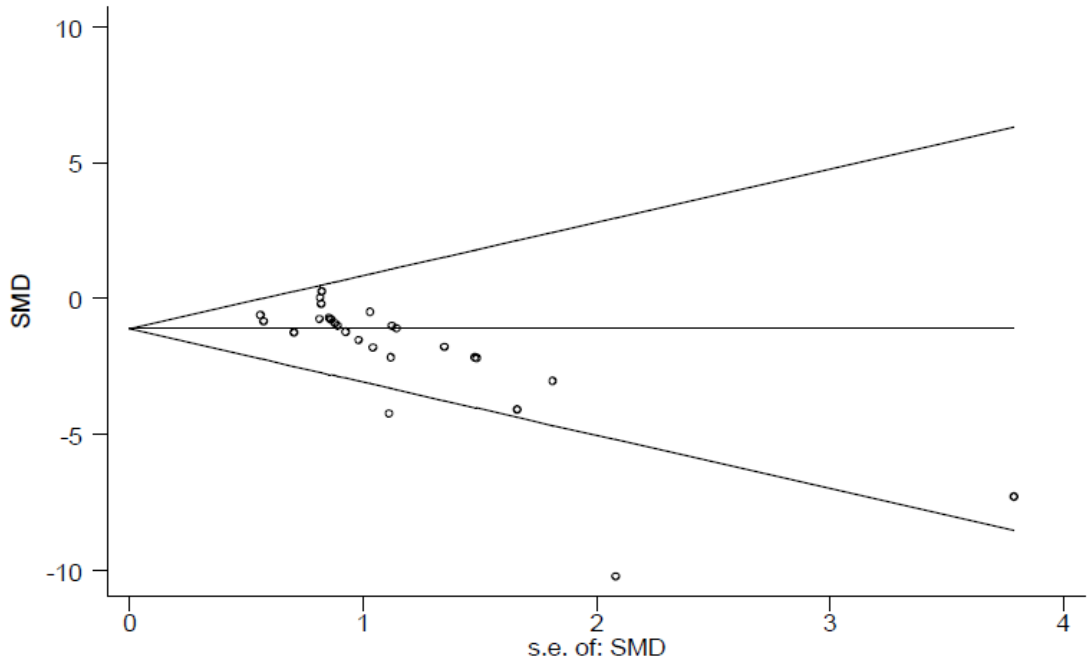


Figure 5. Begg's funnel plot for publication bias in studies investigating CH_4 emissions on drained peatlands. The standardised mean difference (SMD) is plotted against the standard error of the standardised mean difference (s.e. of SMD) to check whether studies are distributed symmetrically around the pooled effect size as expected or not. The funnel indicates 95% confidence intervals and horizontal line indicates the pooled effect size. In this case most studies reported are below the mean effect size line, suggesting the existence of publication bias as some unpublished studies may be missing from above the mean horizontal line. Taken from Bussell et al. (2010).

4.5.3 Qualitative synthesis

It is common in the social sciences to employ qualitative methods where the views of individual people are recorded in relation to a question. When open ended question are asked and complex answers received, the data are not formally quantified. In such studies the authors are often seeking to characterise the range of views or reactions to a particular question or set of questions. The role of qualitative data synthesis is therefore quite distinct and serves to increase understanding of some environmental issues and generate hypotheses that might be tested by quantitative methods. Qualitative data may also complement quantitative and contribute to a mixed method approach. Further information on these methods can be found in Gough et al. (2012) and Noyes et al. (2011).

Section 5

Reporting on the conduct and outcome of a CEE Systematic Review

5.1 The interpretation of SR evidence

SR methodology seeks to collate and synthesise data in order to present reliable evidence in relation to the review question. The strength of the evidence base and implications of the results for decision-making require careful consideration and interpretation. The discussion and conclusions may consider the implications of the evidence in relation to practical decisions, but the decision-making context may vary, leading to different decisions based on the same evidence. Authors should, where appropriate, explicitly acknowledge the variation in possible interpretation and simply present the evidence so as to inform rather than offer advice. Recommendations that depend on assumptions about resources and values should be avoided (Khan et al. 2003, Deeks et al. 2005).

Deeks et al (2005) offer the following advice that is of relevance here. Authors and end-users should be wary of the pitfalls surrounding inconclusive evidence and should beware of unwittingly introducing bias in their desire to draw conclusions rather than pointing out the limits of current knowledge. Where reviews are inconclusive because there is insufficient evidence, **it is important not to confuse 'no evidence of an effect' with 'evidence of no effect'**. The former results in no change to existing guidelines, but has an important bearing on future research, whereas the latter could have considerable ramifications for current practice or policy.

Review authors, and to a lesser extent end-users, may be tempted to reach conclusions that go beyond the evidence that is reviewed or to present only some of the results. Authors must be careful to be balanced when reporting on and interpreting results. For example, if a 'positive' but statistically non-significant trend is described as 'promising', then a 'negative' effect of the same magnitude should be described as a 'warning sign'. Other examples of unbalanced reporting include one-sided reporting of sensitivity analyses or explaining non-significant positive results but not negative ones. If the confidence interval for the estimate of difference in the effects of interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well and both should be given equal consideration in discussion of results. One-sided attempts to explain results with reference to indirect evidence external to the review should be avoided. Considering results in a blinded manner can avoid these pitfalls (Deeks *et al.* 2005). Authors should consider how the results would be presented and framed in the conclusions and discussion if the direction of the results was reversed.

Biases can occur in the SR process, which do not impair the raw data themselves (i.e. different from Section 4.3) but may affect the conclusion of the SR (through a biased selection of articles) (see review in Borenstein et al. 2009). For example:

Publication bias: statistically significant results are more prone to be published than non significant ones. Yet, there is no strict relationship between the quality of the methodology and the significance of results, and thus, their publication. A good methodology may lead to non significant results and be kept as a grey article.

Language bias: searching is generally undertaken in English because it is the most common language used in scientific writing. This may result in an over-representation of statistically significant results (Egger et al. 1997; Jüni et al. 2002) because they are more likely to be accepted in the scientific literature.

Availability bias: only the studies that are easily available are included in the analysis, whilst other significant results may be available (this can be an increasing problem as many private companies have their own research teams and publish their own journals or reports). Similarly, a **confidentiality bias** may exist in some sensitive topics (eg GMO, nuclear power) because some research results may not be available for security reasons.

Cost bias: time and resources necessary for a thorough search are not always available, which could lead to the selection of the studies only available free or at low cost.

Familiarity bias: the researcher limited the search to articles relevant to his/her own discipline.

Duplication bias: some studies with statistically significant results may be published more than once (Tramer et al. 1997).

Citation bias: Studies with significant results are more likely to be cited by other authors and thus easier to be found during the search (Gøtzsche 1997; Ravnskov 1992).

All these biases can be quantified and several methods exist to quantify their impacts on the results (Borenstein et al. 2009).

5.2 Reporting review conclusions

SRs are most often conducted to assess available evidence of effectiveness or of impact. In so doing, SRs assess the strength of a causal inference (Hill 1971). Aspects that may be reported in the conclusion section include:

1. The quality/reliability of the included studies.
2. The relevance/external validity of the included studies.
3. The size and statistical significance of the observed effects.
4. The consistency of the effects across studies or sites and the extent to which this can be explained by other variables (effect modifiers).
5. The clarity of the relationship between the intensity of the intervention and the outcome.
6. The existence of any indirect evidence that supports or refutes the inference.
7. The lack of other plausible competing explanations of the observed effects (bias or confounding).

In a review concerning the impacts of liming streams and rivers on fish and invertebrates, Mant et al. (2011) discuss all of the above points in a good example of SR conclusions. Rather than discussing the limitations of their review, the authors describe the range of quality of studies included, the size and consistency of the effect observed across studies,

the link between intervention intensity and outcome, the presence of effect modifiers, the presence of evidence in support/refute of the review findings, and the potential for other causative factors for the observed effects.

There is a range of approaches to grading the strength of evidence presented in health-related reviews, but there is no universal approach (Deeks et al. 2005). We suggest that authors of ecological reviews explicitly state weaknesses associated with each of the aspects above, but the overall impact they make on conclusions can only be considered subjectively.

5.3 Implications for policy and practice

A key objective of SR is to inform decision-makers of the implications of the best available evidence relating to a question of concern, and enable them to place this evidence in context, in order to make a decision on the best course of action. Providing evidence that increases capacity to predict the outcomes of alternative actions should lead to better decision making.

End-users must decide, either implicitly or explicitly, how applicable the evidence presented in a SR is to their particular circumstances (Deeks et al. 2005). This is particularly critical in environmental management where many factors may vary between sites and it seems likely that many interventions/actions will vary in their effectiveness/impact depending on a wide range of potential environmental variables. Authors should highlight where the evidence is likely to be applicable and equally importantly where it may not be applicable with reference to variation between studies and study characteristics.

Clearly, variation in the ecological context and geographical location of studies can limit the applicability of results. Authors should be aware of the timescale of included studies, which may be insufficiently short to make long-term predictions. Variation in application of the intervention may also be important (and difficult to predict), but authors should be aware of differences between *ex situ* and *in situ* treatments (measuring efficacy versus effectiveness respectively) where they are combined and should also consider the implications of applying the same intervention at different scales. Variation in baseline risk may also be an important consideration in determining the applicability of results, as the net benefit of any intervention depends on the risk of adverse outcomes without intervention, as well as on the effectiveness of the intervention (Deeks et al. 2005).

Where reviewers identify predictable variation in the relative effect of the intervention or exposure in relation to the specified reasons for heterogeneity, these should be highlighted. However, these relationships require cautious interpretation (because they are only correlations), particularly where sample sizes are small, data points are not fully independent and multiple confounding occurs. **When reporting implications, the emphasis should be on objective information and not on subjective advocacy.**

5.4 Implications for research

Rather like primary scientific studies, most SRs will generate more questions than they answer. Knowledge gaps will be frequent, as will areas where the quality of science conducted to date is inadequate. In conducting an SR, critically appraising the quality of existing studies and attempting to assess the available evidence in terms of its fitness for purpose, reviewers should be able to draw conclusions concerning the need for further research. This need may simply be reported in the form of knowledge gaps but may often consist of recommendations for the design of future studies that will generate data of sufficient quality to improve the evidence base and decrease the uncertainty surrounding the question.

5.5 Supplementary materials

To maximise transparency SRs should normally be supported by a number of supplementary materials. The following is a list of expected information;

1. A report of literature scoping containing combinations of search strings and the outcome of searches of different databases (this is usually as an appendix with the protocol).
2. A list of articles excluded after reading the full text, including reasons for exclusion (note: a list of articles included is expected in the main text).
3. A list of articles that could not be obtained at full text: such articles are therefore potentially relevant but not fully screened.
4. Data extraction and quality assessment tables; for example Excel files with data extracted from each included study (this may be included in the main text if a small number of studies is included).

References

Arnqvist G. and Wooster D. 1995. Meta-analysis – synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution* 10: 236–240.

Benítez López A., Alkemade R. and Verweij P. 2010. Are mammal and bird populations declining in the proximity of roads and other infrastructure? CEE Review 09-007 (SR 68). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR68.html.

Bernes C., Bråthen K.A., Forbes B.C., Hofgaard A., Moen J. and Speed J.D.M. 2013. What are the impacts of reindeer/caribou (Rangifer tarandus L.) on arctic and alpine vegetation? A systematic review protocol. *Environmental Evidence* ***2****: 6.*

Binns N. A. and Remmick R. 1994. Response of Bonneville Cutthroat Trout and Their Habitat to Drainage-Wide Habitat Management at Huff Creek, Wyoming. *North American Journal of Fisheries Management* 14(4): 669-680.

Booth A. (2004). Formulating answerable questions. In A. Booth & A. Brice (Eds.), Evidence-based practice: An information professional's handbook (pp. 61-70). London: Facet.

Borenstein M., Hedges L.V., Higgins J.P.T. and Rothstein H.R. 2009. Introduction to meta-analysis. Wiley, Chichester, UK. 421 pp.

Bowler D.E., Buyung-Ali L.M., Knight T.M., and Pullin A.S. (2009). The importance of nature for health: is there a specific benefit of contact with green space? Environmental Evidence: www.environmentalevidence.org/SR40.htm

Bowler D., Buyung-Ali L., Knigh T. and Pullin A.S. 2010. How effective is 'greening' of urban areas in reducing human exposure to ground level ozone concentrations, UV exposure and the 'urban heat island effect'? Environmental Evidence: www.environmentalevidence.org/SR41.html

Bussell J., Jones D.L., Healey J.R. and Pullin A.S. 2010. How do draining and re-wetting affect carbon stores and greenhouse gas fluxes in peatland soils? CEE review 08-012 (SR49). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR49.html.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37-46.

Cooper H.M. 1984. *Integrating Research. A Guide for Literature Reviews*. Sage Publications, Newbury Park.

Côté I.M., Mosqueira I. & Reynolds J.D. 2001. Effects of marine reserve characteristics on the protection of fish populations: a meta-analysis. *Journal of Fish Biology* 59: SA178-189.

Davies Z.G., Tyler C., Stewart G.B. & Pullin A.S. 2006. Are current management recommendations for conserving saproxylic invertebrates effective? CEE review 05-011 (SR17). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR17.html.

Deeks J.J., Higgins J.P.T. and Altman D.G. 2005. "Analysing and presenting results". In: *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5* [updated May 2005]; Section 8. (ed by J.P.T. Higgins and S. Green.) www.cochrane.org/resources/handbook/hbook.htm (accessed 12th July 2006).

Doerr V.A.J., Doerr E.D. and Davies M.J. 2010. Does structural connectivity facilitate dispersal of native species in Australia's fragmented terrestrial landscapes? CEE

Review 08-007 (SR44). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR44.html.

Dogpile. 2007. *Different Engines, Different Results: Web Searchers Not Always Finding What They're Looking for Online*. Available at: www.infospaceinc.com/onlineprod/OverlapDifferentEnginesDifferentResults.pdf Accessed: 31/03/2008.

Downing J.A., Osenberg C.W. & Sarnelle O. 1999. Meta-analysis of marine nutrient-enrichment experiements: variation in the magnitude of nutrient limitation. *Ecology* 80: 1157-1167.

Edwards P., Clarke M., DiGuiseppi C., Pratap S., Roberts I. and Wentz R. 2002. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine* 21: 1635-1640.

Egger M., Davey-Smith G., Schneider M., and Minder C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315, 629-634.

Eycott A., Watts K., Brandt G., Buyung-Ali L., Bowler D., Stewart G. and Pullin A. 2010. Do landscape matrix features affect species movement? CEE Review 08-006 (SR43). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR43.html.

Eysenbach G., Tuische J., and Diepgen T. L. 2001. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. *Medical Informatics and the Internet in Medicine*, 26(3):203-218.

Fazey I., Salisbury J.G., Lindenmayer D.B. Maindonald J. and Douglas R. 2004. Can methods applied in medicine be used to summarize and disseminate conservation research? *Environmental Conservation* 31: 190-198.

Feinstein A.R. 1985. Clinical Epidemiology: The Architecture of Clinical Research. Saunders, Philadelphia.

Felton A., Knight E., Wood J., Zammit C., and Lindenmayer D.B. 2010. A meta-analysis of fauna and flora species richness and abundance in plantations and pasture lands. CEE review 09-012 (SR73). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR73.html.

Gates S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology* 71: 547–557.

Gotzsche P. C. 1987. Reference bias in reports of drug trials. *British Medical Journal* 295: 654-656.

Gough D., Oliver S. and Thomas J. 2012. An introduction to systematic reviews. Sage. London.

Gurevitch J. and Hedges L.V. 2001. *Meta-analysis Combining the results of independent experiments*. In: "Design and Analysis of Ecological Experiments" (ed by S.M. Scheiner and J. Gurevitch) pp. 347-369. Oxford University Press, New York.

Hedges L.V. 1994. "Statistical considerations". In: *The Handbook of Research Synthesis.* (ed by H. Cooper and L.V. Hedges) pp. 30-33. Russell Sage Foundation, New York.

Higgins J.P.T. and Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.

Hill A.B. 1971 Principles of Medical Statistics. *Lancet* 9: 312-20.

Hock R. 1999. The Extreme Searcher's Guide to Web Search Engines: A Handbook for the Serious Searcher. CyberAge Books, New Jersey, USA.

Introna L. D. and Nissenbaum H. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16: 169-185.

Isasi-Catalá E. 2010. Is translocation of problematic jaguars (*Panthera onca*) an effective strategy to resolve human-predator conflicts? CEE review 08-018 (SR55). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR55.html.

Jackson G.B. 1980. Methods for integrative reviews. *Review Education Research* 50: 438-460.

Johnson V., Fitzpatrick I., Floyd R. and Simms A. 2011. What is the evidence that scarcity and shocks in freshwater resources cause conflict instead of promoting collaboration? CEE review 10-010. Collaboration for Environmental Evidence: www.environmentalevidence.org/SR10010.html.

Jüni P., Witschi A.. Bloch R. and Egger M. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of American Medical Association* 282: 1054-60.

Jüni P., Holenstein F., Sterne J., Bartlett, C. and Egger M. 2002. Direction and impact of language bias in meta-analyses of controlled-trials: empirical study. *International Journal of Epidemiology*, 31, 115-123.

Khan K.S., Kunz R., Kleijnen J. and Antes G. 2003. *Systematic reviews to support evidence-based medicine: how to apply findings of healthcare research*. Royal Society of Medicine Press Ltd, London.

Kunz R., and Oxman A.D. 1998. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised trials. *British Medical Journal* 317: 1185-1190.

Landis J.R. and Koch G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics* 33, 159—174.

Leimu R., and Koricheva J. 2005. What determines the citation frequency of ecological papers? *Trends in Ecology and Evolution* 20: 28-32.

Lipsey M.W. and Wilson D.B. 2001. Practical Meta-analysis. *Applied Social Research Methods Series*. Volume 49. Sage Publications, Thousand Oaks, California.

Mant R., Jones D., Reynolds B., Ormerod S. and Pullin A.S. 2011. What is the impact of liming of streams and rivers on the abundance and diversity of fish and invertebrates? CEE review 09-015 (SR76). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR76.html.

McDonald M., McLaren K. and Newton A. 2010. What are the mechanisms of regeneration post-disturbance in tropical dry forest? CEE Review 07-013 (SR37). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR37.html.

Moher D., Jadad A.R., Nichol G., Penman M., Tugwell P. and Walsh S. 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials* 16: 62-73.

Moher D., Jadad A.R. and Tugwell P. 1996. Assessing the quality of randomized controlled trials: current issues and future directions. *International Journal of Technology Assessment in Health Care* 12: 195-208.

NHS Centre for Reviews and Dissemination. 2001. *Undertaking systematic review of research on effectiveness*. NHS CRD, University of York.

Noyes J., Popay J., Pearson A., Hannes K. and Booth A. 2011. Chapter 20: Qualitative

research and Cochrane reviews. In: Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1. The Cochrane Collaboration. www.handbook.cochrane.org.

Osenberg C.W., Sarnelle O., Cooper S.D. and Holt R.D. 1999. Resolving ecological questions through meta-analysis: goals, metrics and models. *Ecology* 80: 1105–1117.

Petrokofsky G. et al. 2012. Comparison of methods for measuring and assessing carbon stocks and carbon stock changes in terrestrial carbon pools. How do the accuracy and precision of current methods compare? *Environmental Evidence* 1:6.

Popay J. (Ed.) 2006. Moving Beyond Effectiveness. Methodological issues in the synthesis of diverse sources of evidence. National Institute for Health and Clinical Evidence, UK.

Pullin A.S., Bangpan M., Dalrymple S., Dickson K., Healey J., Hockley N., Jones J., Knight T. and Oliver S. 2012. Human well-being impacts of terrestrial protected areas? CEE protocol 11-009. Collaboration for Environmental Evidence: www.environmentalevidence.org/SR11009.html.

Pullin A.S. and Knight T.M. 2003. Support for decision-making in conservation practice: an evidence-based approach. *Journal for Nature Conservation* 11: 83-90.

Pullin A.S., Knight T.M. and Watkinson A.R. 2009. Linking reductionist science and holistic policy using systematic reviews: unpacking environmental policy questions to construct an evidence-based framework. *Journal of Applied Ecology* 46, 970-975.

Randall N.P. and James K.L. 2012. The effectiveness of integrated farm management, organic farming and agri-environment schemes for conserving biodiversity in temperate Europe - A systematic map. *Environmental Evidence* 1:4.

Ravnskov U. 1992. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *British Medical Journal* 305: 9-15**.**

Roberts P.D., Stewart G.B. and Pullin A.S. 2006. Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. *Biological Conservation* 132, 409-423*.*

Scargle J.D. 2000. Publication Bias: The “File-Drawer” Problem in Scientific Inference. *Journal of Scientific Exploration* 14: 91–106,

Schafer J.L. (1997) Analysis of incomplete multivariate data. *Monographs on Statistics and applied Probability* 72. Chapman & Hall, Boca Raton USA.

Schulz K.F., Chalmers I., Hayes R.J. and Altman D.G. 1995. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 273: 408-412.

Sharp S. 1998. Meta-analysis regression: statistics, biostatistics, and epidemiology. *Stata Technical Bulletin* 42: 16-22.

Showler D.A., Stewart G.B., Sutherland W.J. and Pullin A.S. 2010. What is the impact of public access on the breeding success of ground-nesting and cliff-nesting birds? CEE Review 05-010 (SR16). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR16.html.

Smart J. M. and Burling D. 2001. Radiology and the Internet: A Systematic Review of Patient Information Resources. *Clinical Radiology*, 56: 867-870.

Smith R.K., Pullin A.S., Stewart G.B. and Sutherland W.J. 2010. Is predator control and effective strategy for enhancing bird populations? CEE review 08-001 (SR38).

Collaboration for Environmental Evidence: www.environmentalevidence.org/SR38.html.

Stevens A., and Milne R. 1997. “The effectiveness revolution and public health”. In: *Progress in Public Health* (ed. By G. Scally), Pp. 197-225. Royal Society of Medicine Press, London.

Stewart G.B., Pullin A.S. and Coles C.F. 2005. Effects of wind turbines on bird abundance. Environmental Evidence: www.environmentalevidence.org/SR4.htm.

Stewart R., and Liabo K. 2012. Involvement research without compromising research quality. Journal of Health Services Research and Policy 17: 248-251.

Thompson S. 1994. Systematic review: why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal 309: 1351-1355.

Tramer M.R., Reynolds D.J., Moore R.A. and McQuay H.J. 1997. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ*, 315, 635-640.

University of California 2008. *Finding information on the Internet: a Tutorial - Meta-Search Engines*. Available at: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html Accessed 31/03/2008.

APPENDICES

Appendix A. Example – Scoping: the iterative development of a database search strategy

The below example is based on pre-review scoping conducted by Bowler et al. (2010): "The Evidence Base for Community Forest Management as a Mechanism for Supplying Environmental Benefits and Improving Rural Welfare" and is presented as an illustration of the iterative nature of search term development.

Scoping searches were conducted in Web of Knowledge with the objective of testing the utility of the stakeholder-suggested search terms (see Table 5 below) and providing an idea of the potential numbers of returned hits to guide resource planning. The suggested search terms were split into three groups: the first based on the intervention of interest, the second guided by the outcome elements of the review question, and the third influenced by the types of study of interest (Table 5). Only if searches based on set one returned an unmanageable number of hits would it have been appropriate to use sets two and three.

Table 5. Original stakeholder-proposed search terms.

Set:	Search terms:
One	Community Forest Management Co-management forest Joint management forest Participatory management forest Indigenous forest reserve Decentralized Forest Governance Community engagement in forest management
Two	Biodiversity, desert*, degrad*, economic, carbon, poverty, fuel*
Three	evidence, empirical, quantitative, evaluation, assessment, measures

The results shown below in Table 6 illustrate the evolution of this set of terms, from one returning a huge number of spurious hits, to one more sensitive and manageable. On the basis of these findings, it was thus deemed appropriate to exclude the terms suggested in sets two and three, as it was felt that these may have been overly restrictive in this context.

Table 6. Search term scoping and evolution.

Search string	Number of hits (Web of Knowledge)	Change from previous
1. Topic=((community forest management) OR (co-management forest*) OR (joint management forest*) OR (participatory forest*) OR (indigenous forest* reserve*) OR (decentrali* forest*) OR (integrated conservation development pro*) OR (ICDP*))	21,464	n/a
2. Topic=("community forest management" OR "co-management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	250	Quotation marks added to improve % relevance
3. Topic=("community forest* management" OR "co-management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	256	Wildcard added to pick up alternative word endings in first phrase
4. Topic=("community forest* management" OR "co-management forest*" OR "co management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	256	De-hyphenated variant added for co-management phrase. Not useful
5. Topic=("community forest*" OR "co-management forest*" OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,008	1st phrase amended ("management" removed) to pick up alternatives such as "community forestry" or "community forests", etc.
6. Topic=("community forest*" OR "forest* co-management " OR "joint management forest*" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,019	2nd phrase amended to more probable word order
7. Topic=("community forest*" OR "forest* co-management " OR ("joint		

management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,035	Third phrase amended to pick up all variants of the term – e.g. "forest joint management" or "joint management forests/ry, etc."
8. Topic=("community forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,096	Ditto above for second phrase
9. Topic=("community forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "JFM" OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,264	"JFM" noted as a standalone term in some of the Indian literature, and thus included
10. Topic=("community forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "JFM" OR "participatory forest*" OR ("collaborative management" AND forest*) OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,279	Addition of further 'intervention' term
11. Topic=("community forest*" OR "community-based forest*" OR ("co-management " AND forest*) OR ("joint management" AND forest*) OR "JFM" OR "participatory forest*" OR ("collaborative management" AND forest*) OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*")	1,304	Ditto above
12. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" AND "social forestry")	15,195	Addition of 'social forestry'. Deemed too broad to be useful. Nothing apparently additional retrieved.
13. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*)		

OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource")	1385	"Community based natural resource" added – apparently very useful
14. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*))	1563	(community AND "natural resource management" AND forest*) added to account for alternative variants
15. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*) OR "common property")	3344	"Common property" added but broad
16. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*) OR ("common property" AND forest*))	1715	"forest*" added to common property phrase to restrict spurious hits
17. Topic=("community management" AND woodland*)	13	Not useful – all relevant papers either contained term 'forest' or other 'intervention' based terms

		e.g community-based natural resource management
18. Topic=("community management" AND tree*)	39	Ditto above
19. Topic=("community forest*" OR "community-based forest*" OR ("co-management" AND forest*) OR ("joint management" AND forest*) OR "JFM" OR ("collaborative management" AND forest*) OR "participatory forest*" OR "indigenous forest* reserve*" OR "decentrali* forest*" OR "integrated conservation development pro*" OR "ICDP*" OR "community-based natural resource" OR (community AND "natural resource management" AND forest*) OR ("common property" AND forest*))	1715	SUGGESTED TERMS (FOR DRAFT PROTOCOL)

* indicate the use of wildcards or 'truncation', to search for variant word endings. Terms in red font are those omitted or included at each stage.

Appendix B. Example of data synthesis

A SR of the impact of wind turbines on bird abundance utilised standardized mean difference meta-analysis with weighting by inverse variance to combine data from 19 globally distributed windfarms (Stewart et al. 2005). Sensitivity analyses were used to explore the effect of including data from unreplicated studies and to assess bias arising from data extraction of pseudoreplicated or aggregated data. Pooled effect sizes remained negative and statistically significant regardless of how the effect sizes were generated, indicating that the patterns in the data were robust. *A priori* and *post hoc* reasons for heterogeneity were explored with meta-regression. Of the *a priori* variables only bird taxon appeared to modify the result, with relationships between turbine number and power being too weak to have biological significance. *Post hoc* analysis revealed that the impact of windfarms became more pronounced over time, a finding not reported by any of the original research or previously assessed in the literature. This has important implications because declines in local bird abundance are more likely to have deleterious population-level impacts if they worsen over time. It also suggests that current windfarm monitoring programs are of inadequate duration to detect deleterious effects.

Box A1. Interpretation of forest plots-Example using STATA

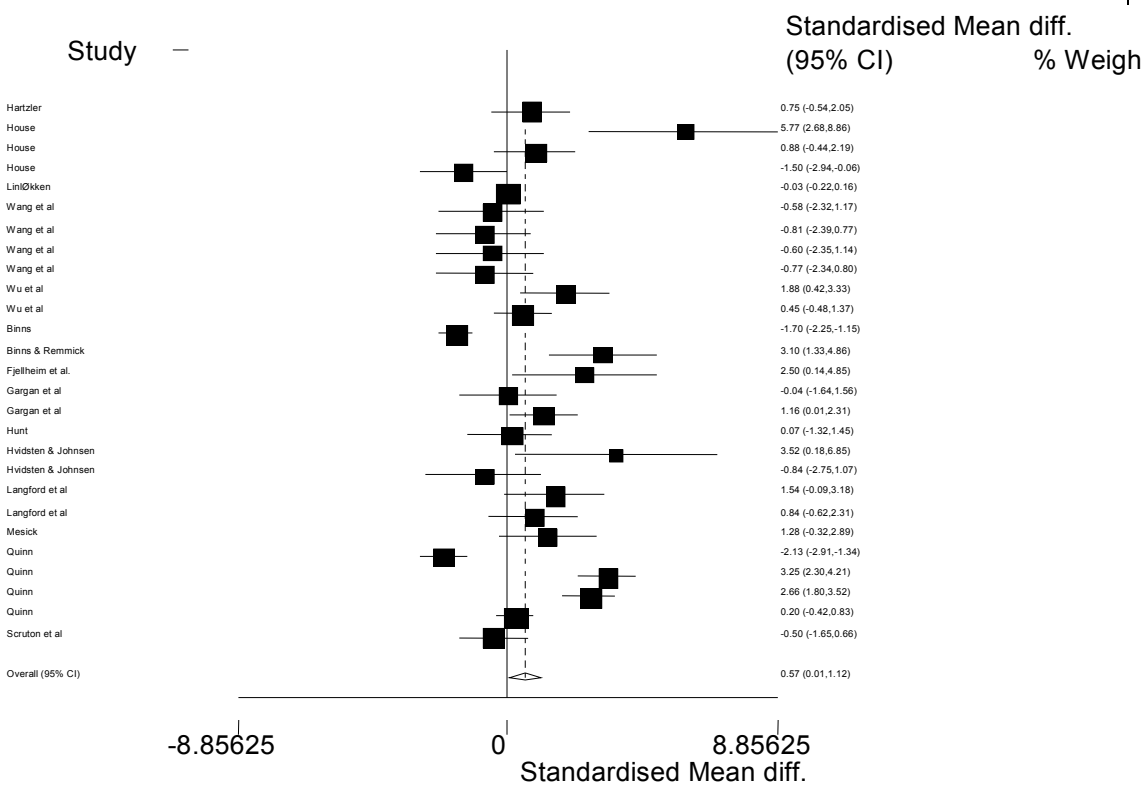


Figure A1. An example of a forest plot generated using STATA and typically included as an outcome in reviews that incorporate a meta-analysis.

The individual data points included in the meta-analysis are listed down the left side of the diagram. In this example multiple independent points have been extracted from the same references. Individual studies are typically identified by author name and year, with multiple points numbered. Full details of each study can be found in the references at the end of the SR and the tables of included studies and data extraction appendices should make it clear how multiple points were derived from individual studies.

Each data point extracted from a study is represented by a square. The size of the square represents the sample size of the study generating that data point whilst the error bar typically represents the 95% confidence interval. The position of the square on the x axis denotes the effect size (in this example Cohens D). This example also lists the effect size and confidence interval for each study to the right of the diagram, along with the weight which that study contributes to the overall synthesis (in this example weighting is by inverse variance).

Underneath the studies, there is a pooled estimate of effect represented by an open diamond. This is a graphical representation of the combined outcome for all of the included data points. The width of this diamond represents the confidence interval.

The “line of no effect” where the effect size is zero is represented by a solid vertical line, and anything that crosses this line is not statistically significant (including those studies where only the confidence interval crosses the line). Anything that falls to the left of the line of no effect has less of the outcome; whereas anything that falls to the right has more of the outcome- whether this is a positive or negative result depends on what the outcome of the meta-analysis is. Therefore a beneficial result for a negative outcome (such as habitat loss) has a significant effect size to the left of the vertical line and a beneficial result for a positive outcome (such as increase in suitable habitat) has a significant effect size to the right of the vertical line. Overall interpretation of the forest plot relies on consideration of the position and significance of individual points as well as the pooled estimate, because the pooled estimate can be misleading when heterogeneity is high (see above).

Glossary of terms

Attrition: subject units lost during the experimental/investigational period than cannot be included in the analysis (e.g. units removed due to deleterious side-effects caused by the intervention).

Bias (synonym: systematic error): the distortion of the outcome, as a result of a known or unknown variable other than intervention (i.e. the tendency to produce results that depart from the "true" result).

Confounding variable (synonym: co-variate): a variable associated with the outcome, which distorts the effect of intervention.

Critical appraisal: a formal, documented assessment of the internal and external validity of primary research.

Effectiveness: the extent to which an intervention produces a beneficial outcome under ordinary circumstances (i.e. does the intervention work?).

Effect Modifier: Any variable that modifies the impact of an intervention or exposure. Effect modifiers are one cause of heterogeneity in the outcome of interventions.

Effect size: the observed association between the intervention and outcome, where the improvement/decrement of the outcome is described in deviations from the mean.

Efficacy: the extent to which an intervention produces a beneficial outcome under ideally controlled circumstances (i.e. can the intervention work?).

Efficiency: the extent to which the effect of the intervention on the outcome represents value for money (i.e. the balance between cost and outcome).

Evidence: 1. anything that establishes a fact or gives reason for believing something. 2. statements made or objects produced as proof or to support a case.

Evidence-based health care: extends the application of the principles of evidence-based medicine to all professions associated with health care, including purchasing and management.

Evidence-based medicine (EBM): is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.

External validity: the extent to which the findings of a study can be generalised to the wider system.

Fixed effects model: a mathematical model that combines the results of studies that assume the effect of the intervention is constant in all subject populations studied. Only within-study variation is included when assessing the uncertainty of results (in contrast to a random effects model).

Forest plot: a plot illustrating individual effect sizes observed in studies included within a SR (incorporating the summary effect if meta-analysis is used).

Funnel plot: a graphical method of assessing bias; the effect size of each study is plotted against some measure of study information (e.g. sample size; if the shape of the plot resembles an inverted funnel, it can be stated that there is no evidence of publication bias within the SR).

Heterogeneity: the variability between studies in terms of key characteristics (i.e. ecological variables) quality (i.e. methodology) or effect (i.e. results). Statistical tests of heterogeneity may be used to assess whether the observed variability in effect size (i.e. study results) is greater than that expected to occur purely by chance.

Incidence = the total number of new cases occurring over a period of time, usually a year / individuals at risk

Internal validity: the degree to which a research study has attempted to minimise systematic error (bias).

Intervention: the policy or management action under scrutiny within the SR.

Mean difference: the difference between the means of two groups of measurements.

Meta-analysis: a quantitative method employing statistical techniques, to combine and summarise the results of studies that address the same question.

Meta-regression: A multivariable model investigating effect size from individual studies, generally weighted by sample size, as a function of various study characteristics (i.e. to investigate whether study characteristics are influencing effect size).

Mixed methods: research that combines qualitative and quantitative methodology to answer a given research question, and are often employed in interdisciplinary research.

Mixed effects models: a mathematical model that combines fixed and random effects.

Narrative synthesis: a textual and possibly graphical description of findings from a systematic review, i.e. not including meta-analysis.

Occurrence: a description of pattern of cases – place and time. Observation of an event.

Outcome: the effect of the intervention in a form that can be reliably measured.

Power: the ability to demonstrate an association where one exists (i.e. the larger the sample size, the greater the power and the lower the probability of the association remaining undetected).

Precision: the proportion of relevant articles identified by a search strategy as a percent of all articles found (i.e. a measure of the ability of a search strategy to exclude irrelevant articles).

Prevalence = the total number of existent cases at a specific point in time (point prevalence) or during a period of time (period prevalence –usually a year) out of the nb of individuals at risk.

Protocol: the set of steps to be followed in a SR. It describes the rationale for the review, the objective(s), and the methods that will be used to locate, select and critically appraise studies, and to collect and analyse data from the included studies.

Publication bias: the possible result of an unsystematic approach to a review (e.g. research that generates a negative result is less likely to be published than that with a positive result, and this may therefore give a misleading assessment of the impact of an intervention). Publication bias can be examined via a funnel plot.

Qualitative: a terms used for descriptive information based on a quality or characteristic rather than a quantity or metric.

Quality assessment: see critical appraisal

Random effects model: a mathematical model for combining the results of studies that allow for variation in the effect of the intervention amongst the subject populations studied. Both within-study variation and between-study variation is included when assessing the uncertainty of results (in contrast to a fixed effects model).

Review: an article that summarises a number of primary studies and discusses the effectiveness of a particular intervention. It may or may not be a SR.

Search strategy: an *a priori* description of the methodology, to be used to locate and identify research articles pertinent to a SR, as specified within the relevant protocol. It includes a list of search terms, based on the subject, intervention and outcome of the review, to be used when searching electronic databases, websites, reference lists and when engaging with personal contacts. If required, the strategy may be modified once the search has commenced.

Sensitivity: the proportion of relevant articles identified by a search strategy as a percentage of all relevant articles on a given topic (i.e. the degree of comprehensiveness of the search strategy and its ability to identify all relevant articles on a subject).

Sensitivity analysis: repetition of the analysis using different sets of assumptions (with regard to the methodology or data) in order to determine the impact of variation arising from these assumptions, or uncertain decisions, on the results of a SR.

Standardised mean difference (SMD): an effect size measure used when studies have measured the same outcome using different scales. The mean difference is divided by an estimate of the within-group variance to produce a standardised value without units.

Study quality: the degree to which a study seeks to minimise bias.

Subgroup analysis: used to determine if the effects of an intervention vary between subgroups in the SR. Subgroups may be pre-defined according to differences in subject populations, intervention, outcome and study design.

Subject: the unit of study to which the intervention is to be applied.

Summary effect size: the pooled effect size, generated by combining individual effect sizes in a meta-analysis.

Systematic review: a review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyse data from the studies that are included within the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies.

Weighted mean difference (WMD): a summary effect size measure for continuous data where studies that have measured the outcome on the same scale have been pooled.

The Collaboration for Environmental Evidence (CEE) is a partnership between scientists and managers working towards a sustainable global environment and the conservation of biodiversity. The collaboration seeks to synthesise evidence on issues of greatest concern to environmental policy and practice. CEE has formal charitable status and its Objects are: "*The protection of the environment and conservation of biodiversity through preparation, maintenance promotion and dissemination of systematic reviews of the effects and impacts of environment management interventions, for the public benefit. "*

www.environmentalevidence.org

.

With the support of:

NATURAL ENGLAND

NATURAL ENVIRONMENT RESEARCH COUNCIL