Ecological Applications, 26(5), 2016, pp. 1295–1301
© 2016 by the Ecological Society of America

An evidence assessment tool for ecosystem services and conservation studies

Anne-Christine Mupepele,[1,3] Jessica C. Walsh,[2] William J. Sutherland,[2] and Carsten F. Dormann[1]

[1]*Department of Biometry and Environmental System Analysis, University of Freiburg, Tennenbacherstr. 4, 79106 Freiburg, Germany*
[2]*Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, United Kingdom*

Abstract. Reliability of scientific findings is important, especially if they directly impact decision making, such as in environmental management. In the 1990s, assessments of reliability in the medical field resulted in the development of evidence-based practice. Ten years later, evidence-based practice was translated into conservation, but so far no guidelines exist on how to assess the evidence of individual studies. Assessing the evidence of individual studies is essential to appropriately identify and synthesize the confidence in research findings. We develop a tool to assess the strength of evidence of ecosystem services and conservation studies. This tool consists of (1) a hierarchy of evidence, based on the experimental design of studies and (2) a critical-appraisal checklist that identifies the quality of research implementation. The application is illustrated with 13 examples and we suggest further steps to move towards more evidence-based environmental management.

Key words: governance; quality checklist; quantification; rigour; valuation.

Introduction

Conservation and ecosystem services studies are important scientific sources for decision-makers seeking advice on environmental management (Daily and Matson 2008, Kareiva and Marvier 2012). Their results potentially influence actions and it is therefore crucial to assess transparently the reliability of current research and its recommendations (Pullin and Knight 2003, Boyd 2013).

Evidence-based practice was introduced in the medical field aiming to assess the reliability of scientific statements and identify the best available information to answer a question of interest (Sackett et al. 1996, GRADE Working Group 2004, OCEBM Levels of Evidence Working Group 2011). In conservation, evidence-based practice was first mentioned 15 yr ago (Sutherland 2000, Pullin and Knight 2001). Today, the Collaboration for Environmental Evidence fosters the creation of systematic reviews to collate the strongest possible evidence (Petrokofsky et al. 2011, Collaboration for Environmental Evidence 2013; see also Journal for Environmental Evidence), together with Conservation Evidence (Hopkins et al. 2015), which focuses on the development of summaries and guidelines, and the communication of evidence to practitioners (Sutherland et al. 2012, Dicks et al. 2014). Summaries, contrary to systematic reviews, do not focus on a specific question but bring together information from a much broader topic, e.g., from a whole animal group, such as bees (Dicks , et al. 2010, 2014, Walsh et al. 2015).

Systematic reviews and summaries compile individual studies and therefore require the evaluation of the evidence at the level of the individual study. In systematic reviews this is typically mentioned as one step of the critical appraisal. However, to date, such critical appraisal is often implicit, based on criteria varying for every systematic review (Collaboration for Environmental Evidence 2013, Carroll and Booth 2015, Stewart and Schmid 2015). We therefore introduce an evidence assessment tool providing a clear appraisal guideline to score the reliability of individual studies.

Definitions and Terminology

A well-defined terminology is essential for effective communication between practitioners and scientists. Evidence is the "ground for belief" or "the available body of information indicating whether a belief or proposition is true or valid" (Howick 2011). Evidence describes the knowledge behind a statement and expresses how solid our recommendations are (see also Higgs and Jones 2000:311; Rychetnik et al. 2001, Lohr 2004, Binkley and Menyailo 2005, Pullin and Knight 2005). The strength of evidence reflects the reliability of information and we can identify whether a statement is based on strong or weak evidence, i.e., very reliable or hardly reliable. Hence evidence-based practice means to identify the reliability of current knowledge, based on research integrated with

Manuscript received 14 April 2015; revised 6 November 2015; accepted 23 November 2015. Corresponding Editor: D. Schimel.
[3]E-mail: anne-christine.mupepele@biom.uni-freiburg.de

expertise, and to act according to this best available knowledge. The collation and appraisal of the best available evidence follow strict criteria to ensure transparency and to reduce bias. A goal of evidence-based practice is to act on best available evidence while being aware of the strength of inference this evidence permits (Howick 2011:15).

Setting Question and Context

The formulation of a clear research question and the purpose of investigation is highly emphasized throughout the evidence literature (Higgins and Green 2011, Collaboration for Environmental Evidence 2013:20–23). Questions should specify which ecosystem service, species or aspect of biodiversity will be investigated in which system, as this will help to determine the external validity of the answer provided in a study.

We further recommend to determine the focus of the question as either quantification, valuation, management, or governance. Quantification studies measure the amount of an ecosystem service, species abundance, biodiversity, or other conservation targets. Measures can be taken in absolute units or relative to another system. Valuation studies assess the societal value of ecosystem services. The most common way is monetary valuation. Management is the treatment designed to improve or benefit specific ecosystem services, target species, or other conservation aspects. For example, leaving dead wood in forests to increase biodiversity or reducing agricultural fertilizer to decrease nearby lake eutrophication. Governance is seen as the strategy or policy to steer a management intervention, such as REDD (Reducing Emissions from Deforestation and Forest Degradation), which aims to encourage forest protection and reforestation (Kenward et al. 2011). The strategies used by policy makers include incentives (subsidiaries) or penalties (law/tax; see also Bevir 2012). When the effectiveness of management and governance strategies is determined, evidence-based quantification or valuation is required to measure the outcome of the management or governance intervention. Acuña et al. (2013), for example, used valuation methods to determine success or failure of a management strategy, while Walsh et al. (2012) quantified malleefowl abundance through monitoring survey data to assess the management impact of fox baiting. The distinction of four different foci is essential to assess the whole range of environmental management.

We have described how to set the context of questions that can be useful in environmental management. Once the question has been determined, and the investigation carried out, the strength of the resulting evidence should be assessed (Fig. 1).

Evidence Assessment

The reliability of a study is characterized by its study design and the quality of its implementation. Both are evaluated in the evidence assessment.

Evidence hierarchy

The study design refers to the set-up of the investigation, e.g., controlled or observational design (GRADE Working Group 2004). These study designs are not equally compelling with respect to inferring causality. Differences in study designs typically translate into weak or strong evidence. To identify the reliability of a study, study designs can be ranked hierarchically according to a level-of-evidence scale, henceforth, the evidence hierarchy (Fig. 2).

Systematic reviews (LoE1a) are at the top of the evidence hierarchy (Fig. 2) and provide the most reliable

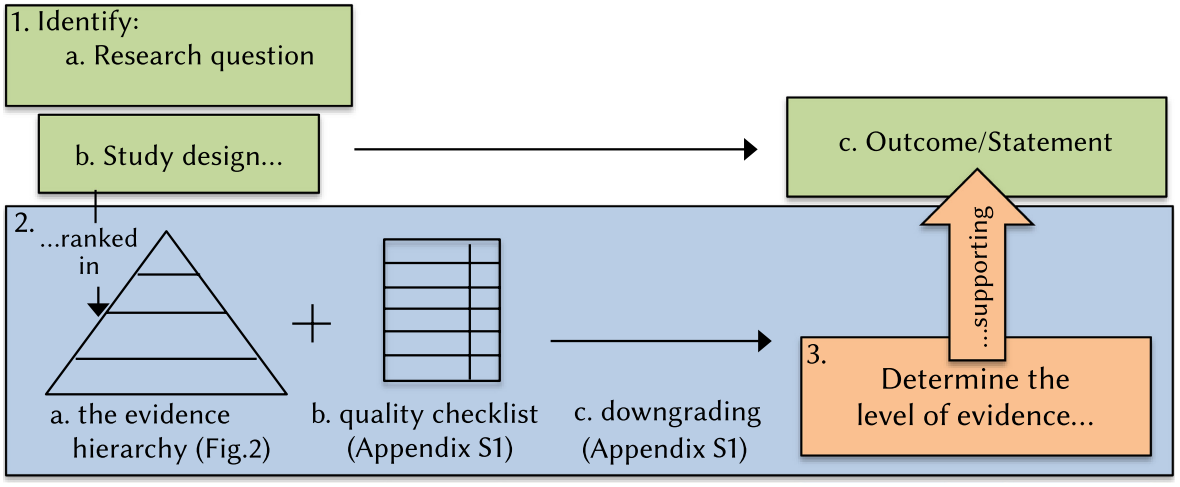

Fig. 1. Schematic representation of the evidence assessment tool. (1) Identification of study question, design, and outcome. (2) Assessing a level of evidence based on the underlying study design and calculating a quality score based on the quality checklist. (3) Determining the final level of evidence supporting the outcome by downgrading the originally assigned level of evidence according to the quality score.

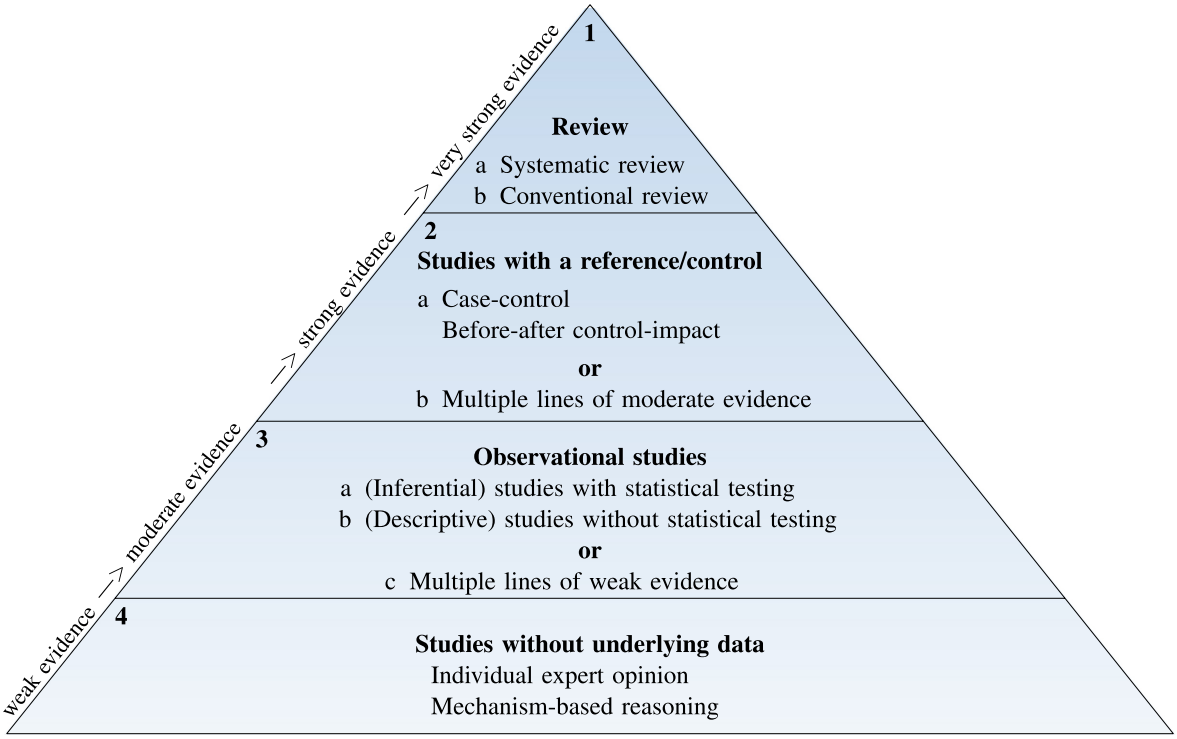

FIG. 2. Level-of-evidence (LoE) hierarchy ranking study designs according to their evidence. Very strong evidence (LoE1) to weak evidence (LoE4) with internally ranked sublevels a, b, and c.

information. They summarize all information collated in several individual studies, have an a priori protocol on design and procedure, and are conducted according to strict guidelines (e.g., Collaboration for Environmental Evidence 2013). If possible, they ideally include quantitative measures, i.e., a meta-analysis (see Koricheva et al. 2013, Vetter et al. 2013). All other, non-systematic and more conventional reviews (LoE1b) may also include quantitative analysis or are purely qualitative. Both types of review summarize the findings of several studies, but systematic reviews assess the completeness and reproducibility more carefully and strive to reduce bias by having transparent, thorough, pre-defined methods (Freeman et al. 2006, Higgins and Green 2011, Collaboration for Environmental Evidence 2013, Haddaway and Bayliss 2015, Haddaway and Bilotta 2015).

The necessary condition for any review is that appropriate individual studies are available. The most reliable individual study design is a study with a reference/control (LoE2). Typically, these are case-control or before–after control–impact studies (LoE2a; Smith et al. 2014). Investigations that cannot follow such a controlled design may alternatively seek to gain strong evidence through multiple lines of moderate evidence (LoE2b). Multiple lines of evidence require at least two unrelated and consistent arguments to confirm the study conclusions, thereby forming a non-contradicting picture (see also Smith et al. 2002). Illustrative examples are the valuation of ecosystem services (e.g., Mogas et al. 2006) or long-term environmental processes that are difficult to control (e.g., Dorman et al. 2015). Multiple lines of evidence can be collected in individual studies using different approaches within one study context (LoE2b, LoE3c) or in reviews (LoE1) including evidence from different studies.

Observational studies (LoE3) are individual studies without a control. These include studies employing inferential and correlative statistics (LoE3a), e.g., testing for the influence of environmental variables on the quantity of an ecosystem service. Descriptive studies (LoE3b) imply data collection and representation without statistical testing (e.g., data summaries, ordinations, histograms, surveys). Multiple lines of weak evidence (LoE3c) can increase the evidence of LoE4 investigations; elicitation of independent expert opinions is a well-known example (Sutherland et al. 2013, Morgan 2014, Smith et al. 2015, Sutherland and Burgman 2015; see also Appendix S1).

The lowest level of evidence are statements without underlying data (LoE4). These are usually individual expert opinions, often not distinguishable from randomness (Tetlock 2005, Drolet et al. 2015). Other statements without underlying data are reasoning based on mechanism. Mechanism-based reasoning involves an inferential chain linking an intervention to the outcome (Howick et al. 2010, Howick 2011). If this chain of mechanisms is not supported by data, there is no possibility to assess whether all relevant mechanisms linking the

intervention to the outcome have been included. Mechanism-based reasoning without corroborative data provides only weak evidence. On the other hand, mechanism-based reasoning can result in a model that is validated and tested on real world data. With such a data validation, the model could reach moderate evidence or strong evidence, depending on the underlying study design.

It is important to note that method and design should not be confused. Methods are the means used to collect or analyze data, e.g., remote sensing, questionnaires, or ordination techniques. Design reflects how the study was planned and conducted, e.g., a case-control or observational design (GRADE Working Group 2004). The same methods can be employed for different underlying designs. Remote sensing, for example, can be done purely descriptively (LoE3b) or with a reference such as ground-truthing or in a before-and-after design (LoE2a). Analogously, models can represent theories without supporting data (LoE4), involve data input to determine parameters (LoE3b), or be tested and validated (LoE3a). To achieve strong evidence, model predictions have to be confirmed by several unrelated data sets forming a non-contradicting picture (LoE2b) or should be built on information derived from controlled studies unequivocally identifying the underlying causal mechanism (LoE2a; Kirchner 2006).

Critical appraisal

Study design alone is an inadequate marker of the strength of evidence (Rychetnik et al. 2001). A study with a strong-evidence design may be poorly conducted. The critical appraisal assesses the implementation of the study design, specifically the methodological quality, the actual realization of the study design, and its reporting (Higgins and Green 2011). It identifies the study quality and may lead to a downgrading in the evidence hierarchy. Quality, in this context, is the extent to which all aspects of conducting a study can be shown to protect against bias and inferential error (Lohr 2004). Quality checklists can be used to detect bias and inferential error. Combining 30 published quality checklists, we provide the first quality checklist for conservation and ecosystem services (Appendix S1: Table S1), that can be used to comprehensively assess the internal validity of a study, covering questions on data collection, analysis, and the presentation of results. The checklist consists of 43 questions, of which some apply only to a specific context, e.g., for reviews or studies focusing on valuation. All questions answered with yes receive one point. In the case of non-reported issues, we advise the answer no to indicate a deficient reporting quality. The percentage of points received can help to decide whether to downgrade the level of evidence (Appendix S1: Table S2).

Reviews provide information at the highest level of evidence and their critical appraisal is different from other designs because they are based on studies with weaker evidence (see Appendix S1: Table S1, Review). Every single study included in the review can be assessed for its level of evidence using the evidence hierarchy and the checklist for quality criteria. If only studies based on weak evidence were included, then the review should be downgraded, regardless of other quality criteria. In addition, a review can be assessed for other quality shortcomings using again the quality checklist.

The checklist should make the assessment more transparent, but we are aware that the process may not always be straightforward. Questions in the checklist can be subjective and depend on the judgment of the assessor. Cohen's kappa test was used to test the agreement in 13 exemplary studies between two different assessors (Appendix S1: Table S3). It ranges from 0 to 1, representing random to perfect agreement. Our result revealed a moderate agreement (unweighted Cohen's kappa $= 0.49$; P-value <0.001; Cohen 1960, Landis and Koch 1977, Gamer et al. 2015). Depending on the context, the assessor may decide to give more weight to particular questions or add questions to the checklist. Although the procedure cannot be fully standardized, we are not aware of a better alternative, and we encourage the use of the checklist as a baseline that can be adapted for specific studies.

The combination of study design (Fig. 2) and quality criteria (Appendix S1: Table S1) is the last step and identifies the strength of evidence supporting the study result (schematic representation in Fig. 1). The level of evidence derived by the study design should be downgraded depending on the quality score calculated from the quality checklist (Appendix S1: Table S2).

Application of the Evidence Assessment Tool

The suggested method was applied to assess the evidence of 13 studies (Appendix S1: Table S3). They were selected to serve as examples and illustrate the applicability of the tool to the whole range of study designs and foci. The first example was a management-related systematic review of Mant et al. (2013), conducted according to the guidelines of the Collaboration for Environmental Evidence (2013). They investigated the effect of liming rivers or lakes on fish and invertebrate populations. They found that liming increased fish abundances and acid-sensitive invertebrates, but may have a negative impact on the abundance of all invertebrate taxa combined. According to the critical appraisal, the study achieved 21 out of 24 points (88%) and it therefore remained at the originally assigned LoE1a, the highest level of evidence (Appendix S1: Table S3).

A second example tackles the question: How does adding dead wood to rivers influence the provision of ecosystem services? (Acuña et al. 2013). The authors investigated two ecosystem services (fishing and retention of organic and inorganic matter) in a river-forest ecosystem in Spain and Portugal and studied the effect of this management intervention. Their study design followed a before–after control–impact approach, equivalent to LoE2a. The critical appraisal revealed shortcomings, e.g., no blinding, no randomization, and no probability sampling: only 17 out of 25 points (68%) were

achieved. The level of evidence was downgraded by one level to LoE3a. We therefore conclude that the statement made by Acuña et al. (2013): "restoration of natural wood loading in streams increases the ecosystem service provision" is based on moderate evidence (LoE3a).

We provide further examples in the Appendix (Appendix S1: Tables S3 and S4). All but one study revealed quality shortcomings and had to be downgraded. Most were scored as LoE3 or LoE4.

Relevance for Different User Groups

In the previous section it was elaborated how to assess the strength of evidence for individual studies and reviews. Now we provide a few notes on *who* should use the evidence assessment tool.

1. Scientists conducting their own studies have to be aware of how to achieve strong evidence, particularly during the planning phase. Choosing a study design that provides strong evidence and respects the quality criteria will substantially increase the potential contribution to our knowledge.
2. Scientists advising decision-makers should be explicit about the strength of evidence of information they include in their recommendations. Weighting all scientific information equally, or subjectively, runs the risk of overconfidence and bias.
3. Decision-makers receiving information from scientists should demand a level-of-evidence statement for the information provided. Alternatively, they can assess the strength of evidence themselves. However, this may be difficult as it takes time and requires some scientific training to identify the study design and evaluate the quality questions.
4. We further encourage consortia, international panels and learned societies, such as the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES), the Ecological Societies (ESA, BES, GFÖ, and others), the Society for Conservation Biology (SCB), and the Ecosystem Services Partnership (ESP) to support the development of guidelines that include an evidence assessment (Graham et al. 2011, Sutherland et al. 2015). These best-practice guides are based on the collection of scientific evidence synthesized and judged by a group of experts. They provide recommendations on how to best quantify, value, manage, or govern a desired ecosystem service or conservation target, giving decision-makers transparent advice with an emphasis on the strength of the evidence available (Graham et al. 2011).

Discussion

We have outlined an evidence assessment tool for ecosystem services and conservation studies, encompassing a hierarchy to judge the available evidence based on study design and a quality checklist to facilitate critical appraisal. We have further illustrated with examples how to apply the tool (see also Appendix S1: Tables S3 and S4).

Evidence-based practice seeks to complement existing management frameworks by emphasizing the importance of systematically collating the existing scientific evidence and assessing it for its reliability and relevance. The IPCC report, for example, uses a combined measure of evidence and level of agreement (Mastrandrea et al. 2010, Spiegelhalter and Riesch 2011). Our suggested approach is more detailed, describing how one can actually assess the evidence.

Evidence-based practice has faced criticism of its evidence hierarchies, claiming that controlled trials are not always more reliable than observational studies. A main argument against hierarchies is that they are rigid and only consider the study design to assign a level of evidence (Petticrew and Roberts 2003, Adams and Sandbrook 2013, Stegenga 2014). With our quality checklist, we emphasize the critical appraisal to check for an appropriate implementation and methodological quality of study designs. The proposed assessment therefore does not overestimate the results of deficiently implemented meta-analyses and controlled studies. Some science sectors have to rely on observational studies because their study units cannot be controlled. This usually applies to environmental governance, conservation biology of rare species, or global theories that lack a second earth as a control. Multiple lines of evidence can lead to strong evidence using only observational study designs (Fig. 2, LoE2b). However, a central task of natural science is to determine causal relationships, and observational studies do not have the same strength to determine causal relationships as replicated and randomized case-control studies (Holland 1986, Grimes and Schulz 2002, Illari et al. 2011). We should acknowledge that in some areas of science causality cannot be established, and hence the reliability achieved remains lower than in areas where it can.

Other criticism has been directed toward the fact that every system is unique and the external validity of studies is low. We are aware that generalizability of results is problematic in ecosystems, where many different drivers take influence at the same time and hence, the general evidence may not apply due to particular circumstances. At this point the judgment of experts on the external validity of the currently best available evidence is irreplacable (Karanicolas et al. 2008, Howick 2011). Evidence-based practice means integrating individual expertise with the best available evidence from systematic research (Sackett et al. 1996, Straus et al. 2010). More reflection and responses to criticism of evidence-based practice can be found in Mullen and Streiner (2004), Sutherland et al. (2004, 2005), and Haddaway and Pullin (2013).

Despite the criticism raised against evidence-based practice the benefits are clear (Gilbert et al. 2005, Howick 2011, Walsh et al. 2015). Rating the strength of evidence matters as it clarifies the reliability of research results and,

thus, the strength of conclusions, decisions, or recommendations drawn from that research (Lohr 2004).

Reliable scientific evidence in environmental management is pivotal, and its use (or misuse) can have immense impacts on environmental outcomes and the society. It is essential that scientists and decision makers consider the strength of evidence when conducting studies, provding advice, and taking decisions. In the interest of responsible use of environmental resources and processes, we strongly encourage embracing evidence-based practice as a paradigm for all research contributing to environmental management.

Acknowledgments

We thank Andrew Pullin, Sven Lautenbach, and Ian Bateman for valuable comments on earlier versions of the manuscript. This work was supported by the 7th framework programme of the European Commission in the project OPERAs (grant number 308393, www.operas-project.eu).

Literature Cited

Acuna, V., J. R. Diez, L. Flores, M. Meleason, and A. Elosegi. 2013. Does it make economic sense to restore rivers for their ecosystem services? Journal of Applied Ecology 50:988–997.

Adams, W. M., and C. Sandbrook. 2013. Conservation, evidence and policy. Oryx 47:329–335.

Bevir, M. 2012. Governance: a very short introduction. Oxford University Press, Oxford, UK.

Binkley, D., and O. Menyailo. 2005. Gaining insights on the effects of tree species on soils. Pages 1–16 *in* editor. Tree species effects on soils: implications for global change, chapter 1. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Boyd, I. 2013. A standard for policy-relevant science. Nature 501:159–160.

Carroll, C., and A. Booth. 2015. Quality assessment of qualitative evidence for systematic review and synthesis: is it meaningful, and if so, how should it be performed? Research Synthesis Methods 6:149–154.

Cohen, J. 1960. A coefficient of agreement for nomial scales. Educational and Psychological Measurement 20:37–46.

Collaboration for Environmental Evidence. 2013. Guidelines for Systematic Review and Evidence Synthesis in Environmental Management. Version 4.2. Environmental Evidence. URL www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf.

Daily, G. C., and P. A. Matson. 2008. Ecosystem services: from theory to implementation. Proceedings of the National Academy of Sciences 105:9455–9456.

Dicks, L. V., D. A. Showler, and W. J. Sutherland. 2010. Bee conservation: evidence for the effects of interventions. Pelagic Publishing, Exeter, UK.

Dicks, L. V., J. C. Walsh, and W. J. Sutherland. 2014. Organising evidence for environmental management decisions: a '4S' hierarchy. Trends in Ecology & Evolution 29:1–7.

Dorman, M., T. Svoray, A. Perevolotsky, Y. Moshe, and D. Sarris. 2015. What determines tree mortality in dry environments? a multi-perspective approach. Ecological Applications 25:1054–1071.

Drolet, D., A. Locke, M. A. Lewis, and J. Davidson. 2015. Evidence-based tool surpasses expert opinion in predicting probability of eradication of aquatic nonindigenous species. Ecological Applications 25:441–450.

Freeman, S. R., H. C. Williams, and R. P. Dellavalle. 2006. The increasing importance of systematic reviews in clinical dermatology research and publication. Journal of Investigative Dermatology 126:2357–2360.

Gamer, M., J. Lemon, I. Fellows, and P. Singh. 2015. irr: various coefficients of interrater reliability and agreement. R-package version 0.84 https://cran.r-project.org/web/packages/irr/irr.pdf.

Gilbert, R., G. Salanti, M. Harden, and S. See. 2005. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. International Journal of Epidemiology 34:874–887.

GRADE Working Group. 2004. Grading quality of evidence and strength of recommendations. BMJ 328:1–8.

Graham, R., M. Mancher, D. M. Wolmann, S. Greenfield, and E. Steinberg. 2011. Clinical practice guidelines we can trust. The National Academies Press, Washington, D.C., USA.

Grimes, D. A., and K. F. Schulz. 2002. Descriptive studies: what they can and cannot do. Lancet 359:145–149.

Haddaway, N. R., and H. R. Bayliss. 2015. Clarification on the applicability of systematic reviews. Frontiers in Ecology and the Environment 13:129.

Haddaway, N. R., and G. S. Bilotta. 2015. Systematic reviews: separating fact from fiction. Environment International (in press).

Haddaway, N. R., and A. S. Pullin. 2013. Evidence-based conservation and evidence-informed policy: a response to Adams & Sandbrook. Oryx 47:336–338.

Higgins, J. P. T., and S. Green. 2011. Cochrane handbook for systematic reviews of interventions, Version 5.1.0. [updated March 2011]. The Cochrane Collaboration.

Higgs, J., and M. Jones. 2000. Will evidence-based practice take the reasoning out of practice? Pages 307–315 *in* J. Higgs, and M. Jones, editors. Clinical reasoning in the health professionals. Two edition. Butterworth Heineman, Oxford, UK.

Holland, P. W. 1986. Statistics and causal inference. Journal of the American Statistical Association 81:945–960.

Hopkins, J., N. Ockendon, and W. J. Sutherland. 2015. Our mission to transform conservation practice. Conservation Evidence 12:1. http://www.conservationevidence.com/individual-study/5495

Howick, J. 2011. The philosophy of evidence-based medicine. Wiley-Blackwell, Oxford, UK.

Howick, J., P. Glasziou, and J. K. Aronson. 2010. Evidence-based mechanistic reasoning. Journal of the Royal Society of Medicine 103:433–441.

Illari, P. M., F. Russo, and J. Williamson. 2011. Causality in the sciences. Oxford University Press, Oxford, UK.

Karanicolas, P. J., R. Kunz, and G. H. Guyatt. 2008. Evidence-based medicine has a sound scientific base. Chest 133:1067.

Kareiva, P., and M. Marvier. 2012. What is conservation science? BioScience 62:962–969.

Kenward, R. E., et al. 2011. Identifying governance strategies that effectively support ecosystem services, resource sustainability, and biodiversity. Proceedings of the National Academy of Sciences 108:5308–5312.

Kirchner, J. W. 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research 42:1–5.

Koricheva, J., J. Gurevitch, and K. Mengersen. 2013. Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, New Jersey, USA.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Lohr, K. N. 2004. Rating the strength of scientific evidence: relevance for quality improvement programs. International Journal for Quality in Health Care 16:9–18.

Mant, R. C., D. L. Jones, B. Reynolds, S. J. Ormerod, and A. S. Pullin. 2013. A systematic review of the effectiveness of liming to mitigate impacts of river acidification on fish and macro-invertebrates. Environmental Pollution 179: 285–293.

Mastrandrea, M., et al. 2010. Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change (IPCC). Available at http://www.ipcc.ch.

Mogas, J., P. Riera, and J. Bennett. 2006. A comparison of contingent valuation and choice modelling with second-order interactions. Journal of Forest Economics 12:5–30.

Morgan, M. G. 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National Academy of Sciences 111:7176–7184.

Mullen, E. J., and D. L. Streiner. 2004. The evidence for and against evidence-based practice. Brief Treatment and Crisis Intervention 4:111–121.

OCEBM Levels of Evidence Working Group. 2011. The Oxford Levels of Evidence 1. URL http://www.cebm.net/index.aspx?o=5653.

Petrokofsky, G., P. Holmgren, and N. D. Brown. 2011. Reliable forest carbon monitoring-systematic reviews as a tool for validating the knowledge base. International Forestry Review 13:56–66.

Petticrew, M., and H. Roberts. 2003. Evidence, hierarchies, and typologies: horses for courses. Theory and Methods 57:527–529.

Pullin, A. S., and T. M. Knight. 2001. Effectiveness in conservation practice: pointers from medicine and public health. Conservation Biology 15:50–54.

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

Pullin, A. S., and T. M. Knight. 2005. Assessing conservation management's evidence base: a survey of management-plan compilers in the United Kingdom and Australia. Conservation Biology 19:1989–1996.

Rychetnik, L., M. Frommer, P. Hawe, and A. Shiell. 2001. Criteria for evaluation evidence on public health interventions. Journal of Epidemiology and Community Health 56:119–127.

Sackett, D. L., W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. 1996. Evidence based medicine: what it is and what it isn't. Clinical Orthopaedics and Related Research 455:3–5.

Smith, E. P., I. Lipkovich, and K. Ye. 2002. Weight-of-Evidence (WOE): quantitative estimation of probability of impairment for individual and multiple lines of evidence. Human and Ecological Risk Assessment: An International Journal 8:1585–1596.

Smith, R. K., L. V. Dicks, R. Mitchell, and W. J. Sutherland. 2014. Comparative effectiveness research: the missing link in conservation. Conservation Evidence 11:2–6.

Smith, S. D. P., et al. 2015. Rating impacts in a multi-stressor world: a quantitative assessment of 50 stressors affecting the Great Lakes. Ecological Applications 25:717–728.

Spiegelhalter, D. J., and H. Riesch. 2011. Don't know, can't know: embracing deeper uncertainties when analysing risks. Philosophical Transactions of the Royal Society A 369:4730–4750.

Stegenga, J. 2014. Down with the hierarchies. Topoi 33:313–322.

Stewart, G. B., and C. H. Schmid. 2015. Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies. Research Synthesis Methods 6:109–110.

Straus, S. E., P. Glasziou, W. S. Richardson, and R. B. Haynes. 2010. Evidence-based medicine: how to practice and teach it, 4e (Straus evidence-based medicine). Churchill Livingstone.

Sutherland, W. J. 2000. The conservation handbook: research, management and policy. Blackwell Science Ltd.

Sutherland, W. J., and M. A. Burgman. 2015. Use experts wisely. Nature 526:317.

Sutherland, W. J., A. S. Pullin, P. M. Dolman, and T. M. Knight. 2004. Response to Griffiths. Mismatches between conservation science and practice. Trends in Ecology & Evolution 19:565–566.

Sutherland, W. J., A. S. Pullin, P. M. Dolman, and T. M. Knight. 2005. Response to Mathevet and Mauchamp: evidence-based conservation: dealing with social issues. Trends in Ecology & Evolution 20:424–425.

Sutherland, W. J., R. Mitchell, and S. V. Prior. 2012. The role of 'Conservation Evidence' in improving conservation management. Conservation Evidence 9:1–2.

Sutherland, W. J., T. A. Gardner, L. J. Haider, and L. V. Dicks. 2013. How can local and traditional knowledge be effectively incorporated into international assessments? Oryx 48:1–2.

Sutherland, W. J., L. V. Dicks, and R. K. Smith. 2015. What works in conservation? Lessons from conservation evidence. OpenBooks, Cambridge, UK.

Tetlock, P. E. 2005. Expert political judgment: how good is it? How can we know? Princeton University Press, Princeton, New Jersey, USA.

Vetter, D., G. Riicker, and I. Storch. 2013. Meta-analysis: a need for well-defined usage in ecology and conservation biology. Ecosphere 4(6):74.

Walsh, J. C., K. A. Wilson, J. Benshemesh, and H. P. Possingham. 2012. Integrating research, monitoring and management into an adaptive management framework to achieve effective conservation outcomes. Animal Conservation 15:334–336.

Walsh, J. C., L. V. Dicks, and W. J. Sutherland. 2015. The effect of scientific evidence on conservation practitioners management decisions. Conservation Biology 29:88–98.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1890/15-0595/suppinfo

Appendix S1: Details and examples for the application of the ev-idence assessment tool

An excel file of all tables is available on GitHub:
https://github.com/biometry/EvidenceAssessmentTool/blob/master/Examples.xlsx

Quality checklist

The quality checklist was composed of 30 published quality checklists. They were retrieved from intiatives renowned for evidence-based practice, namely the Cochrane Collaboration, Oxford Centre for Evidence-based Medicine, publications from the McMaster University in Canada, the 'cradle of evidence-based medicine', and the Collaboration for Environmental Evidence. From these starting points, cross-referenced quality checklists were evaluated. Some of these linked citations summarized several checklists, such as Higgins *et al.* (2011) or Bilotta *et al.* (2014). Lists and quality aspects from other disciplines, that we have come across while developing the paper, were also considered. All references are given in Table S1 in this Appendix. Not all questions of all checklists were included. Often they were related to medical issues, such as diagnosis- or therapy-related quality aspects, which did not fit to our quality checklist for environmental management. Many checklists covered the same questions and a level of saturation with no new questions arising was reached for medical checklists. There were only three environmentally related references in the quality checklist. Söderqvist and Soutukorva (2006) provide a very detailed list on economic valuation and we extracted general aspects from their list. The guidelines from the Collaboration for Environmental Evidence (2013) are not a checklist, but discuss quality issues in their section about critical appraisal. The quality assessment from Bilotta *et al.* (2014) is less comprehensive than the checklist we present. Because publications about general quality aspects in environmental management are scarce, we did not only use existing checklists to compile the presented quality checklist, but also reflected about important questions based on our own experience. This resulted in eight additional questions, that are mentioned without citations in Table 1. The quality checklist should be understood as a mean to raise attention on common problems in study quality. We do not claim the checklist to be comprehensive.

The checklist is structured into categories and not all aspects apply to all questions. We are fully aware, that the checklist does not cover detailed quality questions for all methodological approaches, considering for example the extensive literature that exists on expert elicitation (e.g. Burgman *et al.*, 2011; Bolger and Wright, 2011; Nowack *et al.*, 2011; Morgan, 2014; Mukherjee *et al.*, 2015; Sutherland and Burgman, 2015), modeling (e.g. Seibert, 2003; Ajami *et al.*, 2007; Reichert and Mieleitner, 2009; Jackson *et al.*, 2011; Kirchner, 2006), or statistical methods such as GLMs or Bayesian statistics. It is clearly beyond the scope of a quality checklist with such a broad range to provide detailed questions on each of these aspects. The questions listed in the section 'Data collection' and 'Analysis' (Question 3 to 12) implicitly include many detailed aspects depending on the methods used in the study.

Table S1: Quality checklist questions with references. Each question answered with 'yes' will receive one point. If a question is not appropriate, it may be left out.

	Quality checklist question	Source of quality checklist criterion
INTERNAL VALIDITY		
Research aim		
1	Does the study address a clearly focused question?	Spencer *et al.* 2003, Lohr 2004, SIGN 2006, CEBM 2010, Collaboration for Environmental Evidence 2013
2	Does the question match the answer?	
Data collection		
3	Was the population/area of interest defined in space, time and size?	Spencer *et al.* 2003, Lohr 2004, Söderqvist & Soutukorva 2006, Brouwers *et al.* 2010, Santaguida *et al.* 2012, AHRQ 2014
4	Selection bias: Was the sample area representative for the population defined?	National Health and Medical Research Council 2000, Söderqvist & Soutukorva 2006, Tong *et al.* 2007, Moher *et al.* 2010, Santaguida *et al.* 2012
5	Was the sample size appropriate?	Jadad *et al.* 1996, Ah-See & Molony 1998, Verhagen *et al.* 1998, Söderqvist & Soutukorva 2006, Tong *et al.* 2007, Moher *et al.* 2012, AHRQ 2014
6	Was probability/random sampling used for constructing the sample?	Söderqvist & Soutukorva 2006
7	If secondary data were used, did an evaluation of the original data take place?	Söderqvist & Soutukorva 2006
8	If data collection took place in form of a questionnaire, was it pre-tested/piloted?	Söderqvist & Soutukorva 2006, Rattray & Jones 2007, Tong *et al.* 2007
9	Were the data collection methods described in sufficient detail to permit replication?	Brouwers *et al.* 2010, CEBM 2010, Moher *et al.* 2010
Analysis		
10	Were the statistical/analytical methods described in sufficient detail to permit replication?	Lohr 2004, Brouwers *et al.* 2010, CEBM 2010, Moher *et al.* 2010
11	Is the choice of statistical/analytical methods appropriate and/or justified?	Jadad *et al.* 1996, Ah-See & Molony 1998, Söderqvist & Soutukorva 2006
12	Was uncertainty assessed and reported?	Ah-See & Molony 1998, Söderqvist & Soutukorva 2006, Bastuji-Garin *et al.* 2013
Results and Conclusions		
13	Do the data support the outcome?	Jadad *et al.* 1996, Ah-See & Molony 1998
14	Magnitude of effect: Is the effect large, significant and/or without large uncertainty?	Jadad *et al.* 1996, Rychetnik *et al.* 2001, SIGN 2006, CEBM 2010, Singh *et al.* 2012
15	Are all variables and statistical measures reported?	CEBM 2010, Higgins *et al.* 2011, Bilotta *et al.* 2014
16	Attrition bias: Are non-response/drop-outs given and is their impact discussed?	Jadad *et al.* 1996, Ah-See & Molony 1998, SIGN 2006, Söderqvist & Soutukorva 2006, Tong *et al.* 2007, Bilotta *et al.* 2014
DESIGN-SPECIFIC ASPECTS		
Review		
17	Is there a low probability of publication bias?	National Health and Medical Research Council 2000, SIGN 2006, Shea *et al.* 2007, CEBM 2010, AHRQ 2014
18	Is the review based on several strong-evidence individual studies?	SIGN 2006
19	Do the studies included respond to the same question?	AHRQ 2014
20	Are results between individual studies consistent and homogeneous?	Rychetnik *et al.* 2001, SIGN 2006, CEBM 2010
21	Was the literature searched in a systematic and comprehensive way?	SIGN 2006, Shea *et al.* 2007, Brouwers *et al.* 2010
22	Was a meta-analysis included?	
23	Were appropriate a priori study inclusion/exclusion criteria defined?	Jadad *et al.* 1996, Ah-See & Molony 1998, Verhagen *et al.* 1998, Lohr 2004, Shea *et al.* 2007, CEBM 2010, Tong *et al.* 2012, Moher *et al.* 2014
24	Did at least two people select studies and extract data?	SIGN 2006, Shea *et al.* 2007, CEBM 2010
Study with a reference/control		
25	Allocation bias: Was the assignment of case-control groups randomized?	Jadad *et al.* 1996, Ah-See & Molony 1998, Verhagen *et al.* 1998, National Health and Medical Research Council 2000, Lohr 2004, SIGN 2006, CEBM 2010, Moher *et al.* 2010, Higgins *et al.* 2011
26	Were groups designed equally, aside from the investigated point of interest?	Lohr 2004, SIGN 2006, CEBM 2010
27	Performance bias: Was the sampling blinded?	Jadad *et al.* 1996, Ah-See & Molony 1998, Verhagen *et al.* 1998, Rychetnik *et al.* 2001, Lohr 2004, SIGN 2006, CEBM 2010, Moher *et al.* 2010, Higgins *et al.* 2011, Bilotta *et al.* 2014
28	Were there sufficient replicates of treatment and reference groups?	SIGN 2006
29	Detection bias: Were outcomes equally measured and determined between groups?	Bilotta *et al.* 2014
Observational studies		
30	Were confounding factors identified and strategies to deal with them stated?	Joanna Briggs Institute 2014

Table S1: Quality checklist questions with references (continued from previous page)

	Quality checklist question	Source of quality checklist criterion
FOCUS-SPECIFIC ASPECTS		
Quantification		
31	Is the unit of the quantification measurement appropriate?	
32	Was temporal change (e.g. annual or long-term) of quantities measured (e.g. species abundance or an ecosystem service) discussed?	
Valuation		
33	If discounting of future costs and outcomes is necessary, was it performed correctly?	SIGN 2006, Söderqvist & Soutukorva 2006
34	If aggregate economic values for a population were estimated, was this estimation consistent with the sampling and the definition of the population?	Defra 2007, de Groot *et al.* 2012
Management		
35	Was the aim of the management intervention clearly defined?	
36	Were side effects and trade offs on other non-target species, ecosystem services or stakeholders considered?	
37	Were both long-term and short-term effects discussed?	AHRQ 2014
38	Did monitoring take place for an appropriate time period?	Jadad *et al.* 1996, CEBM 2010
39	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	Jadad *et al.* 1996, SIGN 2006
Governance		
40	Were long-term effects assessed?	Biermann & Pattberg 2012, ARHQ 2014
41	Was the policy instrument that was used described?	
42	Was the influence of the applied policy instrument (incentive/law) on the society discussed?	
43	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	Jadad *et al.* 1996, SIGN 2006

Table S2: Downgrading the level of evidence (LoE) according to the percentage of quality points reached

Percentage of quality points	Downgrading of the LoE
> 87%	–> none
75 - 87%	–> by half a level (e.g. LoE1a to LoE1b)
62 - 74%	–> by one level (e.g. LoE1a to LoE2a)
50 - 61%	–> by one and a half levels (e.g. LoE1a to LoE2b)
37 - 49%	–> by two levels
25 - 36%	–> by two and a half levels
< 24%	–> by three levels

Table S3: Evidence assessment tool applied to 13 case studies.

	Reference		Mant *et al.* 2013	Lindhjem 2007	Liu *et al.* 2008	Acuna *et al.* 2013
1. Question, outcome and the context	**Context**	**Subject/Ecosystem services; Ecosystem(s); Location**	Fish and aquatic invertebrates; freshwater; global	Non-timber benefits (mainly recreation); boreal forests; Norway, Sweden, Finland	Timber, soil erosion, carbon sequestration, recreation; forests; China	Fish, recreation, erosion control; stream; Spain, Portugal
	Focus	**Quantification/Valuation/Management/Governance**	Management	Valuation	Governance	Management
	Question/Purpose		What is the impact of 'liming' (adding Calcium carbonate) of streams and rivers on the abundance and diversity of fish and invertebrate populations?	How to explain systematic variation in Willingness-to-Pay (WTP) for the value of non-timber benefits from forests in Fennoscandia?	What is the socioeconomical and ecological impact of two payments-for-ecosystem-services programs in China?	Does adding dead wood to streams affect the value of selected ecosystem services and is it cost-effective?
	Outcome		Liming increased fish abundances and acid sensitive invertebrates, but effects were variable and for all invertebrate taxa combined liming may decrease abundance.	WTP is insensitive to the size of the forest and tends to be higher if individuals are asked instead of households.	Socioeconomical impact: income increased, but revenues declined for local governments. Ecological impact: Timber harvest decreased locally but import increased. Carbon sequestration increased and soil erosion declined.	Restoration of natural wood loading in streams increases the ecosystem service provision. The cost–benefit analysis reveals differences between stream orders in the net benefit of the restoration.
	2a. Study design		Systematic Review	Review	Review	BACI
	Level of evidence		LoE1a	LoE1b	LoE1b	LoE2a
	2b. QUALITY CHECKLIST FOR THE CRITICAL APPRAISAL		Quality checklist	Quality checklist	Quality checklist	Quality checklist
	INTERNAL VALIDITY	**Description/Example**	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"
	Research aim					
1	Does the study address a clearly focused question?	See main text, section 'setting question and the context'	yes	yes	no	yes
2	Does the question match the answer?	Answers may not directly correspond to the originally formulated question, e.g. 'Does hunting lead to genetic changes in the moose population of North America?' is answered by: 'hunting reduces the size of calves'. The missing match is obvious when question and answer are written next to each other, but in publications with much text in between it may be more difficult to identify. The result of reduced calf size may be interesting, but special care should be taken while assessing the evidence base.	yes	yes	yes	yes
	Data collection					
3	Was the population/area of interest defined in space, time and size?	'Population/area' is the target, we aim to say something about; e.g. North America's moose population.	no	yes	yes	yes
4	Selection bias: Was the sample area representative for the population defined?	Usually samples are not taken from the whole population/area; e.g. only several North American forests were selected to measure moose. Were the selected forests representative? Did they cover the north, south, east and western part of North America?	/	yes	yes	yes
5	Was the sample size appropriate?	Were the criteria used to determine the sample size (e.g. power calculation) reasonable?	yes	yes	/	yes
6	Was probability/random sampling used for constructing the sample?	Probability sampling means random sampling with known selection probabilities for all objects in the population, while nonprobability sampling does not involve random selection (Trochim, 2014; Söderqvist and Soutukorva, 2009). Most often *equal* probability sampling is used: e.g. all forests in North America have the same chance of being randomly selected. *Unequal* probability sampling can be used to ensure representativeness of result, e.g. if a forest in the south of the area is selected, the selection of the next forest far away from the first will be favored. Unequal probability sampling can also mean that forests easy to access obtain a higher selection probability. Probability sampling is important in addition to representative sampling (question 4).	/	/	/	no
7	If secondary data were used, did an evaluation of the original data take place?	Secondary data, such as used in cost-benefit transfer for example, need to be evaluated to make sure that the data used are not prone to bias.	yes	no	no	no
8	If data collection took place in form of a questionnaire, was it pre-tested/piloted?	Questionnaires need to be professionally designed to ensure that they measure what they intend to measure. Therefore a questionnaire should be pre-tested/piloted on a smaller sample size to test its performance (see Rattray & Jones 2007).	/	/	/	/
9	Were the data collection methods described in sufficient detail to permit replication?		yes	no	no	yes
	Analysis					
10	Were the statistical/analytical methods described in sufficient detail to permit replication?		yes	yes	/	yes
11	Is the choice of statistical/analytical methods appropriate and/or justified?		yes	yes	/	yes
12	Was uncertainty assessed and reported?		yes	yes	no	yes
	Results and Conclusions					
13	Do the data support the outcome?	Are the conclusions drawn of the analytical results valid?	yes	yes	/	yes
14	Magnitude of effect: Is the effect large, significant and/or without large uncertainty?	This question aims to identify the magnitude and precision of results. Precise results are usually characterized by low uncertainty (CEBM 2010) and in combination with a large effect the appropriate statistical analysis (question 11) will lead to a significant result. Not all studies allow the judgment of all three aspect and we therefore combine them in one question and recommend context specific decisions.	yes	yes	/	no
15	Are all variables and statistical measures reported?		yes	yes	/	yes
16	Attrition bias: Are non-response/drop-outs given and is their impact discussed?		/	yes	/	/

Table S3: Evidence assessment tool applied to 13 case studies (continued from previous page)

		Reference		Mant *et al.* 2013	Lindhjem 2007	Liu *et al.* 2008	Acuna *et al.* 2013
2. Evidence Assessment	**DESIGN-SPECIFIC ASPECTS**						
	Review						
	17	Is there a low probability of publication bias?	An assessment of publication bias should include a combination of graphical aids (e.g. funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test, Hedges-Olken) (CEBM 2010). If no quantitative analysis is included, discussion of possible publication bias can be sufficient.	no	yes	no	/
	18	Is the review based on several strong-evidence individual studies?	Most ideally every included study should be assessed for its level of evidence. Several strong evidence individual studies should be included to achieve strong evidence in the review. See main text for further details.	yes	yes	no	/
	19	Do the studies included respond to the same question?		yes	yes	/	/
	20	Are results between individual studies consistent and homogeneous?		yes	no	yes	/
	21	Was the literature searched in a systematic and comprehensive way?		yes	yes	no	/
	22	Was a meta-analysis included?	The term 'meta-analysis' has been vaguely defined in ecology and conservation (Vetter et al. 2013). In this context we do not talk about any summary analysis (e.g. vote counting), but an explicit meta-analysis as defined by Vetter et al. 2013 or Koricheva et al. 2013	yes	yes	no	/
	23	Were appropriate a priori study inclusion/exclusion criteria defined?		yes	no	no	/
	24	Did at least two people select studies and extract data?	At least two people should select papers and extract data. There should be a consensus procedure to resolve any differences (CEBM 2010). In most cases it is too costly to extract data from every paper twice. It might be sufficient to follow the consensus procedure for the first few studies.	yes	no	no	/
	Study with a reference/control						
	25	Allocation bias: Was the assignment of case-control groups randomized?		/	/	/	no
	26	Were groups designed equally, aside from the investigated point of interest?		/	/	/	yes
	27	Performance bias: Was the sampling blinded?	Blinding means that e.g. researchers taking samples of a specific area wouldn't know the differences between these areas.	/	/	/	no
	28	Were there sufficient replicates of treatment and reference groups?		/	/	/	yes
	29	Detection bias: Were outcomes equally measured and determined between groups?	Beside the importance to *design* groups equally (Question 26), the outcome has to be *measured* equally. This is necessary to avoid a bias due to the measurement method.	/	/	/	yes
	Observational studies						
	30	Were confounding factors identified and strategies to deal with them stated?	Controlled studies have equally designed groups (Question 26). Observational studies can not be so easily controlled for potential confounders. It is therefore particularly important to identify them and discuss strategies to avoid biasing results.	/	/	/	/
	FOCUS-SPECIFIC ASPECTS						
	Quantification						
	31	Is the unit of the quantification measurement appropriate?		/	/	/	/
	32	Was temporal change (e.g. annual or long-term) of quantities measured (e.g. species abundance or an ecosystem service) discussed?		/	/	/	/
	Valuation						
	33	If discounting of future costs and outcomes is necessary, was it performed correctly?	Discounting ecosystem services is less straightforward than discounting purely economic values. Nevertheless, it has to be considered when talking about future values (TEEB 2010, ch.6)	/	no	/	no
	34	If aggregate economic values for a population were estimated, was this estimation consistent with the sampling and the definition of the population?	Individual values are summed up to total economic values (TEV), for example in cost-benefit analysis. This should be done thoroughly (e.g. avoiding double counting, considering system boundaries...)	/	/	/	
	Management						
	35	Was the aim of the management intervention clearly defined?		yes	/	/	yes
	36	Were side effects and trade offs on other non-target species, ecosystem services or stakeholders considered?		no	/	/	no
	37	Were both long-term and short-term effects discussed?		yes	/	/	no
	38	Did monitoring take place for an appropriate time period?		/	/	/	yes
	39	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	Ideally the outcome, e.g. increase in biodiversity, is measured according to an evidence-based quantification or valuation tool.	yes		/	yes
	Governance						
	40	Were long-term effects assessed?		/	/	yes	/
	41	Was the policy instrument that was used described?		/	/	yes	/
	42	Was the influence of the applied policy instrument (incentive/law) on the society discussed?		/	/	yes	/
	43	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	Ideally the outcome, e.g. increase in biodiversity, is measured according to an evidence-based quantification or valuation tool.	/	/	yes	/
	2b. Quality points			21	17	8	17
	Possible points		**- depending on the number of questions answered**	24	23	18	25
	Quality score			87.50	73.91	44.44	68.00
	Downgrading			no downgrading	one level	two levels	one level
	Level of evidence			LoE1a	LoE2a	LoE3b	LoE3a

Table S3: Evidence assessment tool applied to 13 case studies (continued from previous page)

	Reference	Kleijn *et al.* 2006	Millar *et al.* 2010	Dorman *et al.* 2015	Goulson *et al.* 2002	Rundio and Olson 2007
1. Question, outcome and the context	**Context**	Biodiversity (vascular plants, birds, bees, grasshoppers, crickets, spiders); farmland; Europe	Soil; grassland; USA	*Pinus halepensis*; forests; Israel	*Bombus terrestris*; farmland, suburban area; UK	Salamanders; forests; USA
	Focus	Governance	Quantification	Quantification	Management	Management
	Question/Purpose	Do agri-environment schemes have an effect on biodiversity and endangered species?	Does commercial sod soil production result in net soil loss? Is there a way to measure the natural occurring soil that is lost with each harvest?	What determines tree mortality in dry environments?	Do measures to promote farmland biodiversity have an influence on nest growth of *Bombus terrestris*?	What are the short-term effects of forest thinning on terrestrial salamanders in managed headwater forests? Can down wood or riparian buffers influence these effects?
	Outcome	Agri-environmental schemes had marginal to moderately positive effects on biodiversity, but endangered species rarely benefit.	Yes. There is a net soil loss of around 100 Mg per year, which is considerably higher than the tolerable soil loss.	Mortality risk was higher in older-aged sparse stands, on southern aspects, and on deeper soils. Association of mortality with lower tree densities was found.	Schemes deployed to enhance farmland biodiversity appear to have little measurable impact on nest growth of this bumblebee species.	Forest thinning decreases salamander abundance in forests that have a low down-wood volume. In stands with little down wood, riparian buffer width would need consideration and may help minimize negative effects of thinning on salamanders.
	2a. Study design	Case control	Case control	Multiple lines of evidence	Case control	BACI
	Level of evidence	LoE2a	LoE2a	LoE2b	LoE2a	LoE2a
	2 b. Q U A L I T Y C H E C K I	Quality checklist	Quality checklist	Quality checklist	Quality checklist	Quality checklist
	INTERNAL VALIDITY	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"
	Research aim					
	1 Does the study address a clearly focused question?	yes	yes	yes	no	yes
	2 Does the question match the answer?	yes	yes	yes	yes	yes
	Data collection					
	3 Was the population/area of interest defined in space, time and size?	yes	yes	yes	yes	yes
	4 Selection bias: Was the sample area representative for the population defined?	yes	no	yes	no	yes
	5 Was the sample size appropriate?	yes	yes	yes	yes	no
	6 Was probability/random sampling used for constructing the sample?	no	no	yes	yes	no
	7 If secondary data were used, did an evaluation of the original data take place?	/	/	/	/	/
	8 If data collection took place in form of a questionnaire, was it pre-tested/piloted?	/	/	/	no	/
	9 Were the data collection methods described in sufficient detail to permit replication?	yes	yes	yes	yes	no
	Analysis					
	10 Were the statistical/analytical methods described in sufficient detail to permit replication?	yes	yes	yes	yes	no
	11 Is the choice of statistical/analytical methods appropriate and/or justified?	yes	yes	yes	yes	no
	12 Was uncertainty assessed and reported?	yes	no	no	yes	no
	Results and Conclusions					
	13 Do the data support the outcome?	yes	yes	yes	no	yes
	14 Magnitude of effect: Is the effect large, significant and/or without large uncertainty?	no	yes	no	no	no
	15 Are all variables and statistical measures reported?	yes	yes	yes	yes	no
	16 Attrition bias: Are non-response/drop-outs given and is their impact discussed?	yes	/	/	yes	/

Table S3: Evidence assessment tool applied to 13 case studies (continued from previous page)

		Reference	Kleijn *et al.* 2006	Millar *et al.* 2010	Dorman *et al.* 2015	Goulson *et al.* 2002	Rundio and Olson 2007
2. Evidence Assessment		**DESIGN-SPECIFIC ASPECTS**					
		Review					
	17	Is there a low probability of publication bias?	/	/	/	/	/
	18	Is the review based on several strong-evidence individual studies?	/	/	/	/	/
	19	Do the studies included respond to the same question?	/	/	/	/	/
	20	Are results between individual studies consistent and homogeneous?	/	/	/	/	/
	21	Was the literature searched in a systematic and comprehensive way?	/	/	/	/	/
	22	Was a meta-analysis included?	/	/	/	/	/
	23	Were appropriate a priori study inclusion/exclusion criteria defined?	/	/	/	/	/
	24	Did at least two people select studies and extract data?	/	/	/	/	/
		Study with a reference/control					
	25	Allocation bias: Was the assignment of case-control groups randomized?	no	no	/	no	no
	26	Were groups designed equally, aside from the investigated point of interest?	no	yes	/	no	no
	27	Performance bias: Was the sampling blinded?	no	no	/	no	no
	28	Were there sufficient replicates of treatment and reference groups?	yes	yes	/	yes	no
	29	Detection bias: Were outcomes equally measured and determined between groups?	yes	yes	/	yes	yes
		Observational studies					
	30	Were confounding factors identified and strategies to deal with them stated?	/	/	no	/	/
		FOCUS-SPECIFIC ASPECTS					
		Quantification					
	31	Is the unit of the quantification measurement appropriate?	yes	yes	yes	yes	yes
	32	Was temporal change (e.g. annual or long-term) of quantities measured (e.g. species abundance or an ecosystem service) discussed?	no	yes	yes	no	yes
		Valuation					
	33	If discounting of future costs and outcomes is necessary, was it performed correctly?	/	/	/	/	/
	34	If aggregate economic values for a population were estimated, was this estimation consistent with the sampling and the definition of the population?	/	/	/	/	/
		Management					
	35	Was the aim of the management intervention clearly defined?	/	/	/	no	yes
	36	Were side effects and trade offs on other non-target species, ecosystem services or stakeholders considered?	/	/	/	no	no
	37	Were both long-term and short-term effects discussed?	/	/	/	no	yes
	38	Did monitoring take place for an appropriate time period?	/	/	/	no	yes
	39	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	/	/	/	no	no
		Governance					
	40	Were long-term effects assessed?	no	/		/	/
	41	Was the policy instrument that was used described?	yes	/	/	/	/
	42	Was the influence of the applied policy instrument (incentive/law) on the society discussed?	no	/	/	/	/
	43	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	yes	/	/	/	/
		2b. Quality points	17	15	13	13	11
		Possible points	25	20	16	27	25
		Quality score	68.00	75.00	81.25	48.15	44.00
		Downgrading	one level	half a level	half a level	two levels	two levels
		Level of evidence	LoE3a	LoE2b	LoE3a	LoE4	LoE4

Table S3: Evidence assessment tool applied to 13 case studies (continued from previous page)

	Reference	Entenmann and Schmitt 2013	Karimzadegan *et al.* 2007	Xie *et al.* 2011	Desanker 2005
1. Question, outcome and the context	**Context**	Biodiversity; forests; Peru	Gas regulation, pollination, pest control and other ecosystem services; forests; Iran	Air quality; urban area; China	Global climate regulation (C-sequestration); tropical forest; Africa
	Focus	Governance	Valuation	Quantification	Governance
	Question/Purpose	How do actors involved in REDD+ processes relate REDD+ implementation to biodiversity conservation? What aspects of biodiversity do they regard as especially important (biodiversity conservation values)?	What is the economic value of ecosystem services provided by Iran's forests and rangelands?	The air quality indicators: CO_2, O_2, SO_2, transpiration cooling and dust interception were quantified (and valuated) for sixteen plant species.	How can the Clean Development Mechanism be better engaged in Africa?
	Outcome	Biodiversity is not a major issue for actors, but direct synergies between REDD+ and biodiversity conservation were assumed by most actors. Values most often mentioned were direct or indirect use values. Option values for future benefits and resilience were rarely mentioned.	The economic value of nonmarket ecosystem services of forests and rangelands' is US$ 53441 million annually. This is equivalent to 43% of Iran's GDP.	Plants with high leaf area indices and photosynthetic rates resulted in an increased transpiration cooling. Species with rough leaf surfaces are efficient in capturing dust and those with thick sclerophyllous leaves best remove SO_2.	Projects should be developed by locals. Carbon money alone may not be enough. Values from the services should be factored into the economic analysis of the country.
	2a. Study design	Descriptive	Descriptive	Descriptive	Expert opinion
	Level of evidence	LoE3b	LoE3b	LoE3b	LoE4
	2 b. Q U A L I T Y C H E C K I	Quality checklist	Quality checklist	Quality checklist	Quality checklist
	INTERNAL VALIDITY	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"	Answer: "Yes/No"
	Research aim				
	1 Does the study address a clearly focused question?	yes	yes	yes	not required - already on lowest level of evidence
	2 Does the question match the answer?	yes	yes	yes	
	Data collection				
	3 Was the population/area of interest defined in space, time and size?	yes	yes	yes	
	4 Selection bias: Was the sample area representative for the population defined?	no	/	yes	
	5 Was the sample size appropriate?	yes	/	yes	
	6 Was probability/random sampling used for constructing the sample?	no	/	no	
	7 If secondary data were used, did an evaluation of the original data take place?	/	no	/	
	8 If data collection took place in form of a questionnaire, was it pre-tested/piloted?	no	/	/	
	9 Were the data collection methods described in sufficient detail to permit replication?	yes	no	yes	
	Analysis				
	10 Were the statistical/analytical methods described in sufficient detail to permit replication?	yes	yes	yes	
	11 Is the choice of statistical/analytical methods appropriate and/or justified?	yes	yes	yes	
	12 Was uncertainty assessed and reported?	no	no	no	
	Results and Conclusions				
	13 Do the data support the outcome?	yes	yes	yes	
	14 Magnitude of effect: Is the effect large, significant and/or without large uncertainty?	/	/	no	
	15 Are all variables and statistical measures reported?	no	yes	yes	
	16 Attrition bias: Are non-response/drop-outs given and is their impact discussed?	/	/	/	

Table S3: Evidence assessment tool applied to 13 case studies (continued from previous page)

		Reference	Entenmann and Schmitt 2013	Karimzadegan *et al.* 2007	Xie *et al.* 2011	Desanker 2005
2. Evidence Assessment	**DESIGN-SPECIFIC ASPECTS**					
	Review					
	17	Is there a low probability of publication bias?	/	/	/	
	18	Is the review based on several strong-evidence individual studies?	/	/	/	
	19	Do the studies included respond to the same question?	/	/	/	
	20	Are results between individual studies consistent and homogeneous?	/	/	/	
	21	Was the literature searched in a systematic and comprehensive way?	/	/	/	
	22	Was a meta-analysis included?	/	/	/	
	23	Were appropriate a priori study inclusion/exclusion criteria defined?	/	/	/	
	24	Did at least two people select studies and extract data?	/	/	/	
	Study with a reference/control			/		
	25	Allocation bias: Was the assignment of case-control groups randomized?	/	/	/	
	26	Were groups designed equally, aside from the investigated point of interest?	/	/	/	
	27	Performance bias: Was the sampling blinded?	/	/	/	
	28	Were there sufficient replicates of treatment and reference groups?	/	/	/	
	29	Detection bias: Were outcomes equally measured and determined between groups?	/	/	/	
	Observational studies					
	30	Were confounding factors identified and strategies to deal with them stated?	no	no	no	
	FOCUS-SPECIFIC ASPECTS					
	Quantification					
	31	Is the unit of the quantification measurement appropriate?	/	/	yes	
	32	Was temporal change (e.g. annual or long-term) of quantities measured (e.g. species abundance or an ecosystem service) discussed?	/	/	no	
	Valuation					
	33	If discounting of future costs and outcomes is necessary, was it performed correctly?	/	no	/	
	34	If aggregate economic values for a population were estimated, was this estimation consistent with the sampling and the definition of the population?	/	no	/	
	Management					
	35	Was the aim of the management intervention clearly defined?	/	/	/	
	36	Were side effects and trade offs on other non-target species, ecosystem services or stakeholders considered?	/	/	/	
	37	Were both long-term and short-term effects discussed?	/	/	/	
	38	Did monitoring take place for an appropriate time period?	/	/	/	
	39	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	/	/	/	
	Governance			/		
	40	Were long-term effects assessed?	yes	/	/	
	41	Was the policy instrument that was used described?	no	/	/	
	42	Was the influence of the applied policy instrument (incentive/law) on the society discussed?	yes	/	/	
	43	Appropriate outcome measures: Are all relevant outcomes measured in a reliable way?	yes	/	/	
	2b. Quality points		11	7	11	
	Possible points		18	13	16	
	Quality score		61.11	53.85	68.75	
	Downgrading		one level	one and a half levels	one level	
	Level of evidence		LoE4	LoE4	LoE4	LoE4

Table S4: Studies on carbon sequestration (CS) in forests. Examples are given for each focus (quantification, valuation, management, governance) and all levels of evidence. No critical appraisal was performed, but this example highlights the use of the evidence hierarchy and the range of foci from quantification to governance. Carbon sequestration was a prominent topic over the previous years (Oren *et al.*, 2001; Fernández-Martínez *et al.*, 2014) and we found studies about carbon sequestration following different study designs. The studies vary in their geographical region and purpose of investigation. They may also investigate a broader range, e.g. the value of all ecosystem services, and we extracted only the question related to carbon sequestration.

	Quantification	**Valuation**	**Management**	**Governance**
Question:	**How much carbon can be captured and stored by a forest?**	**What is the value of carbon sequestration in a forest?**	**How can we manage a forest to maximize carbon sequestration?**	**What are the best governance measures to manage a forest to maximize carbon sequestration?**
Review (LoE1 if there are no qualtiy shortcomings)	Does nutrient availability determine CS in forests? (Fernandez-Martinez et al. 2014)	What is the monetary value of CS provided by urban trees in Lisbon? (Roy, Byrne & Pickering 2012)	What is the effect of forest management on CS in soils? (Jandl et al. 2007)	How can we overcome critical challenges to scale up carbon investments in carbon sequestration projects in Africa? (Jindal, Swallow & Kerr 2008)
Referenced study (LoE2 if there are no quality shortcomings)	Does CS in forests depend on soil fertility? (Oren et al. 2001)	What is the non-market value from an afforested area in Spain? - Comparing results with contingent valuation and choice modelling (Mogas, Riera, Bennett 2006)	Impact of prescribed fire and small clear-cut tree harvesting on carbon dynamics in a mixed-conifer forest in Sierra Nevada? (Stephens et al. 2013)	What are barriers in implementing forest carbon trading? A comparison between the Clean Development Mechanism and a State-run carbon forestry program. (Corbera & Brown 2008)
Observational study (LoE3 if there are no quality shortcomings)	What is the reason for an increased CS in boreal deciduous forests in Canada between 1994 and 1998? (Black et al. 2000)	What is the value of CS provided by Canberra's urban forests? (Brack 2002)	Does carbon fixation increase with different forest managment strategies (e.g. fertilization, thinning)? (Hoen 1994)	What are the effects of carbon taxes and subsidies on the supply of carbon services in West-Canada? (Van Kooten, Binkley & Delcourt 1995)
Based on no data (LoE4)	No study	No study	Does proper design and management of agroforestry result in effective carbon sinks? (Montagnini & Nair 2012)	What governance conditions have to be met to succesfully put in practice small-scale forest carbon projects? (Boyd, Gutierrez & Chang 2007)

References

Acuña V, Díez JR, Flores L, *et al.* 2013. Does it make economic sense to restore rivers for their ecosystem services? *Journal of Applied Ecology* **50**: 988–997.

Ah-See KW and Molony NC. 1998. A qualitative assessment of randomized controlled trials in otolaryngology. *The Journal of Laryngology & Otology* **112**: 460–463.

AHRQ - Agency for Healthcare Research and Quality. 2014. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville: MD: Agency for Healthcare Research and Quality.

Ajami NK, Duan Q, and Sorooshian S. 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research* **43**: n/a–n/a.

Bastuji-Garin S, Sbidian E, Gaudy-Marqueste C, *et al.* 2013. Impact of STROBE statement publication on quality of observational study reporting: interrupted time series versus before-after analysis. *PLoS ONE* **8**: 2–9.

Bilotta GS, Milner AM, and Boyd IL. 2014. Quality assessment tools for evidence from environmental science. *Environmental Evidence* **3**: 14.

Black TA, Chen WJ, Barr AG, *et al.* 2000. Increased carbon sequestration by a boreal deciduous forest in years with a warm spring. *Geophysical Research Letters* **27**: 1271–1274.

Bolger F and Wright G. 2011. Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting and Social Change* **78**: 1500–1513.

Boyd E, Gutierrez M, and Chang M. 2007. Small-scale forest carbon projects: Adapting CDM to low-income communities. *Global Environmental Change* **17**: 250–259.

Brack CL. 2002. Pollution mitigation and carbon sequestration by an urban forest. *Environmental Pollution* **116**: 195–200.

Brouwers M, Kho M, Browman GP, *et al.* 2010. AGREE II: Advancing guideline, development, reporting and evaluation in healthcare. *Canadian Medical Association Journal* **182**: 839–842.

Burgman Ma, Carr A, Godden L, *et al.* 2011. Redefining expertise and improving ecological judgment. *Conservation Letters* **4**: 81–87.

Collaboration for Environmental Evidence. 2013. Guidelines for Systematic Review and Evidence Synthesis in Environmental Management. Version 4.2. Environmental Evidence. URL www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf.

Corbera E and Brown K. 2008. Building institutions to trade ecosystem services: marketing forest carbon in Mexico. *World Development* **36**: 1956–1979.

de Groot RS, Brander L, van der Ploeg S, *et al.* 2012. Global estimates of the value of ecosystems and their services in monetary units. *Ecosystem Services* **1**: 50–61.

DEFRA - UK Department for Environment Food and Rural Affairs. 2007. An introductory guide to valuing ecosystem services. URL https://www.gov.uk/government/publications/an-introductory-guide-to-valuing-ecosystem-services.

Desanker PV. 2005. The Kyoto protocol and the CDM in Africa: a good idea but .... *Unasylva* **56**: 24–26.

Dorman M, Svoray T, Perevolotsky A, *et al.* 2015. What determines tree mortality in dry environments? a multi-perspective approach. *Ecological Applications* **25**: 1054–1071.

Entenmann SK and Schmitt CB. 2013. Actors' perceptions of forest biodiversity values and policy issues related to REDD plus implementation in Peru. *Biodiversity and Conservation* **22**: 1229–1254.

Fernández-Martínez M, Vicca S, Janssens IA, *et al.* 2014. Nutrient availability as the key regulator of global forest carbon balance. *Nature Climate Change* **4**: 471–476.

Goulson D, Hughes W, Derwent L, and Stout J. 2002. Colony growth of the bumblebee, Bombus terrestris, in improved and conventional agricultural and suburban habitats. *Oecologia* **130**: 267–273.

Gowdy J, Howarth RB, and Tisdell C. 2010. Discounting, ethics, and options for maintaining biodiversity and ecosystem integrity. In: The economics of ecosystems and biodiversity: The ecological and economic foundations, chapter 6.

Higgins JPT, Altman DG, Gøtzsche PC, *et al.* 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ (Clinical research ed)* **343**: d5928.

Hoen HF. 1994. Potential and economic efficiency of carbon sequestration in forest biomass through silvicultural management. *Forest Science* **40**: 429–451.

Jackson CH, Bojke L, Thompson SG, *et al.* 2011. A Framework for Addressing Structural Uncertainty in Decision Models. *Medical Decision Making* **31**: 662–674.

Jadad AR, Moore RA, Carroll D, *et al.* 1996. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled clinical trials* **17**: 1–12.

Jandl R, Lindner M, Vesterdal L, *et al.* 2007. How strongly can forest management influence soil carbon sequestration? *Geoderma* **137**: 253–268.

Jindal R, Swallow B, and Kerr J. 2008. Forestry-based carbon sequestration projects in Africa: potential benefits and challenges. *Natural Resources Forum* **32**: 116–130.

Joanna Briggs Institute. 2014. Checklist for critical appraisal of descriptive studies. URL http://joannabriggs.org/assets/docs/jbc/operations/criticalAppraisalForms/JBC_Form_CritAp_DescCase.pdf.

Karimzadegan H, Rahmatian H, Dehghani Salmasi M, *et al.* 2007. Valuing forests and rangelands-ecosystem services. *International Journal of Environmental Research* **1**: 368–377.

Kirchner JW. 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research* **42**: 1–5.

Kleijn D, Baquero RA, Clough Y, *et al.* 2006. Mixed biodiversity benefits of agri-environment schemes in five European countries. *Ecology Letters* **9**: 243–54.

Koricheva J, Gurevitch J, and Mengersen K. 2013. Handbook of Meta-analysis in Ecology and Evolution. Princeton University Press.

Lindhjem H. 2007. 20 years of stated preference valuation of non-timber benefits from Fennoscandian forests: a meta-analysis. *Journal of Forest Economics* **12**: 251–277.

Liu J, Li S, and Ouyang Z. 2008. Ecological and socioeconomic effects of China's policies for ecosystem services. *Proceedings of the National Academy of Sciences* **105**: 9477–9482.

Lohr KN. 2004. Rating the strength of scientific evidence: relevance for quality improvement programs. *International Journal for Quality in Health Care* **16**: 9–18.

Mant RC, Jones DL, Reynolds B, *et al.* 2013. A systematic review of the effectiveness of liming to mitigate impacts of river acidification on fish and macro-invertebrates. *Environmental Pollution* **179**: 285–293.

Millar D, Stolt M, and Amador JA. 2010. Quantification and implications of soil losses from commercial sod production. *Soil Science Society of America Journal* **74**: 892–897.

Mogas J, Riera P, and Bennett J. 2006. A comparison of contingent valuation and choice modelling with second-order interactions. *Journal of Forest Economics* **12**: 5–30.

Moher D, Hopewell S, Schulz KF, *et al.* 2010. CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical research ed)* **340**: 1–2.

Moher D, Liberati A, Tetzlaff J, and Altman DG. 2014. Preferred reporting items for systematic reviews and meta-analyses. *Annals of Internal Medicine* **151**: 264–269.

Montagnini F and Nair PKR. 2012. Carbon sequestration: an underexploited environmental benefit of agroforestry systems. *Agroforestry Systems* 281–295.

Morgan MG. 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences* **111**: 7176–7184.

Mukherjee N, Huge J, Sutherland WJ, *et al.* 2015. The Delphi technique in ecology and biological conservation: applications and guidelines. *Methods in Ecology and Evolution* .

National Health and Medical Research Council. 2000. How to review the evidence: Systematic identification and review of the scientific literature. Government Info Bookshops.

Nowack M, Endrikat J, and Guenther E. 2011. Review of Delphi-based scenario studies: Quality and design considerations. *Technological Forecasting and Social Change* **78**: 1603–1615.

OCEBM. 2010. Oxford Centre for Evidence-Based Medicine - Critical Appraisal Worksheets. URL http://www.cebm.net/critical-appraisal/.

OCEBM Levels of Evidence Working Group. 2011. The Oxford Levels of Evidence 1. URL http://www.cebm.net/index.aspx?o=5653.

Oren R, Ellsworth DS, Johnsen KH, *et al.* 2001. Soil fertility limits carbon sequestration by forest ecosystems in a CO2-enriched atmosphere. *Nature* **411**: 469–72.

Rattray J and Jones MC. 2007. Essential elements of questionnaire design and development. *Journal of Clinical Nursing* **16**: 234–243.

Reichert P and Mieleitner J. 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Research* **45**: 1–19.

Roy S, Byrne J, and Pickering C. 2012. A systematic quantitative review of urban tree benefits, costs, and assessment methods across cities in different climatic zones. *Urban Forestry & Urban Greening* **11**: 351–363.

Rundio DE and Olson DH. 2007. Influence of headwater site conditions and riparian buffers on terrestrial salamander response to forest thinning. *Forest Science* **53**: 320–330.

Rychetnik L, Frommer M, Hawe P, and Shiell A. 2001. Criteria for evaluation evidence on public health interventions. *Journal of Epidemiology and Community Health* **56**: 119–127.

Santaguida PL, Riley CM, and Matchar DB. 2012. Assessing risk of bias as a domain of quality in medical test studies. In: Methods Guide for Medical Test Reviews, chapter 5. Rockville: MD: Agency for Healthcare Research and Quality.

Seibert J. 2003. Reliability of Model Predictions Outside Calibration Conditions. *Nordic Hydrology* **34**: 477–492.

Shea BJ, Grimshaw JM, Wells Ga, *et al.* 2007. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC medical research methodology* **7**.

SIGN -Scottish Intercollegiate Guidelines Network. 2006. Reporting and quality checklists. URL http://www.sign.ac.uk/methodology/tutorials.html.

Singh S, Chang S, Matchar DB, and Bass EB. 2012. Grading a body of evidence on diagnostic tests. In: Methods Guide for Medical Test Reviews, chapter 7, 1–15. Rockville: MD: Agency for Healthcare Research and Quality.

Söderqvist T and Soutukorva Å. 2006. An instrument for assessing the quality of environmental valuation studies. Swedish Environmental Protection Agency (Naturvårdsverket). Stockholm.

Spencer L, Ritchie J, Lewis J, and Dillon L. 2003. Quality in Qualitative Evaluation: A framework for assessing research evidence. Government Chief Social Researcher's Office, UK.

Stephens S, Collins BM, Dore S, *et al.* 2013. Climate change, carbon sequestration, and wildfire management in sierran mixed conifer forests. JFSP Research Project Reports.

Sutherland WJ and Burgman MA. 2015. Use experts wisely. *Nature* **526**: 317.

Tong A, Flemming K, McInnes E, *et al.* 2012. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC medical research methodology* **12**: 181.

Tong A, Sainsbury P, and Craig J. 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International journal for quality in health care* **19**: 349–57.

Trochim WMK, Donnelly JP, and Arora K. 2016. Research Methods Knowledge Base. College Bookstore, 2 edition.

van Kooten GC, Binkley CS, and Delcourt G. 1995. Effect of carbon taxes and subsidies on optimal forest rotation age and supply of carbon services. *American Journal of Agricultural Economics* **77**: 365.

Verhagen AP, de Vet HCW, de Bie RA, *et al.* 1998. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of clinical epidemiology* **51**: 1235–41.

Vetter D, Rücker G, and Storch I. 2013. Meta-analysis: A need for well-defined usage in ecology and conservation biology. *Ecosphere* **4**.

Xie QJ, Zhou ZX, and Chen F. 2011. Quantifying the beneficial effect of different plant species on air quality improvement. *Environmental Engineering and Management Journal* **10**: 959–963.