Heriot-Watt University
Research Gateway

Your evidence or mine? Systematic evaluation of reviews of marine protected area effectiveness

Citation for published version:
Woodcock, P, O'Leary, BC, Kaiser, MJ & Pullin, AS 2017, 'Your evidence or mine? Systematic evaluation of reviews of marine protected area effectiveness', *Fish and Fisheries*, vol. 18, no. 4, pp. 668-681. https://doi.org/10.1111/faf.12196

Digital Object Identifier (DOI):
10.1111/faf.12196

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
Fish and Fisheries

Publisher Rights Statement:
This is the peer reviewed version of the following article: Woodcock, P., O'Leary, B. C., Kaiser, M. J. and Pullin, A. S. (2017), Your evidence or mine? Systematic evaluation of reviews of marine protected area effectiveness. Fish Fish, 18: 668-681, which has been published in final form at https://doi.org/10.1111/faf.12196
This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

General rights
Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 27. May. 2022

Title (option 1): Your evidence or mine? Systematic evaluation of reviews of marine protected area effectiveness

Title (option 2): Your evidence or mine? Systematic evaluation of the scope and reliability of reviews of marine protected area effectiveness

Running Title: Evaluating marine protected area reviews

Paul Woodcock[1], Bethan C. O'Leary[1], Michel J. Kaiser[2], Andrew S. Pullin[1]*,

[1] Centre for Evidence-Based Conservation, School of Environment, Natural Resources and Geography, Bangor University, Gwynedd LL57 2UW, UK

[2] School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey, UK, LL59 5AB

* Corresponding author email: a.s.pullin@bangor.ac.uk, fax: 01248 354997, tel: 01248 382444

Abstract

Marine Protected Areas (MPAs) are a key strategy for mitigating the impacts of fisheries, but their designation can be controversial, and there is uncertainty surrounding when and where MPAs are most effective. Evidence synthesis that collates primary research on MPA effectiveness can provide a crucial bridge between research, policy, and practice. However, reviews vary in scope and rigour, meaning decision-makers face the challenge of identifying appropriate reviews. Documenting differences amongst reviews can therefore support non-specialists in locating the most relevant and rigorous reviews, and can also assist researchers in targeting evidence gaps. We addressed these priorities by systematically searching for reviews examining effectiveness of MPAs for biodiversity, critically appraising methods used, and categorising review scope. The 27 reviews assessed overlapped in scope (suggesting some redundancy) and differed substantially in reliability. Key strengths related to the effects of MPAs on fish abundance and the influence of MPA size and age on effectiveness. However, several gaps were noted, with some questions not addressed and others lacking highly reliable syntheses – importantly, the latter may create the perception that particular questions have been adequately addressed, potentially deterring new syntheses. Our findings indicate key aspects of review conduct that could be improved (e.g. documenting critical appraisal of primary research, evaluating potential publication bias), and can facilitate evidence-based policy by guiding non-specialists to the most reliable and relevant reviews. Lastly, we suggest that future reviews with broader taxonomic coverage and considering the influence of a wider range of MPA characteristics on effectiveness would be beneficial.

Keywords: biodiversity conservation, CEESAT, evidence review, evidence synthesis, evidence-base, review evaluation.

Table of contents:

55 **Introduction**

56 Fisheries exert one the most widespread anthropogenic impacts on marine ecosystems, and can threaten the
57 populations and processes that underpin vital ecosystem services (Butchart et al., 2010, Ramirez-Llodra et
58 al., 2011, Halpern et al., 2012). Establishing marine protected areas (MPAs), in which fishing is restricted
59 to varying degrees, is one of the principal tools for mitigating these impacts (Gaines et al., 2010, Halpern et
60 al., 2010, OSPAR, 2010, Lascelles et al., 2012). Accordingly, the extent of the marine environment with
61 some level of protection from fisheries (and other human activities) has increased steadily from around 0.9%
62 in 2000 to an estimated 3.5% in 2015 (Thomas et al. 2014; Lubchenco & Grorud-Colvert 2015), and is set
63 to rise further in line with the Convention on Biological Diversity target of 10% coverage by 2020 (CBD,
64 2010).

65 There are a range of options for expanding the MPA network in terms of design, placement, and
66 management. Different taxa may also benefit from protection to varying degrees, and across different
67 timescales (Fox et al., 2012, Hays and Scott, 2013). Given the importance that strategies to mitigate fisheries
68 impacts place on MPAs, the increasing promotion of MPAs as a fisheries management tool, and the potential
69 socio-economic and political challenges associated with establishing new reserves, it is essential that
70 scientific evidence is used to identify and communicate the factors that influence effectiveness – thereby
71 allowing new MPAs to be optimally designed and the predicted benefits to be understood. While primary
72 research forms the basis of this evidence, increasing publication rates (Pautasso, 2012, Larsen and von Ins,
73 2010, Li and Zhao, 2015) and the variable quality of primary studies (Willis et al., 2003, Caveen et al., 2012)
74 creates problems for decision-makers in: (1) keeping up-to-date with emerging research; (2) evaluating the
75 appropriateness of methods, data analysis and interpretation in each study; and (3) obtaining an accurate
76 representation of the overall evidence base on MPA effectiveness.

77 Evidence syntheses can assist decision-makers by summarising primary literature on MPAs, with reviews
78 providing a crucial bridge linking primary research with policy and practice. The number of reviews
79 examining MPAs is increasing rapidly (Caveen et al., 2012). However, reviews that do not follow rigorous
80 methods to maximise objectivity and comprehensiveness in searching for, appraising, and synthesising
81 primary research may unintentionally misinform or misrepresent the evidence base. For example,
82 Huntington (2011) argued that the majority of meta-analyses that examined the effectiveness of MPAs did

not address possible publication bias (the tendency to publish positive or hypothesis-affirming results rather than null or controversial findings; Møller and Jennions, 2001) and so may have provided an incomplete picture of the available primary research. Decision-makers and other non-specialists may lack the resources or expertise to systematically collate and appraise all reviews prior to use, and are therefore faced with a similar challenge as for primary literature: identifying the most relevant and rigorous reviews and appreciating the strengths and limitations of the reviews used. Furthermore, where reviews overlap in scope, apparently conflicting interpretations of evidence can reflect variation in review reliability or subtle differences in emphasis amongst reviews. This leads to a perception amongst policymakers that the science is inconclusive, resulting in no decisions being made, unnecessary delays, or selective use of evidence. The existence of a review on a particular topic could also give the impression that the topic has already been investigated and so does not require further exploration, even if the review is potentially less reliable. Future high quality syntheses might thus be deterred, resulting in what could be termed 'cryptic' evidence gaps.

To address the above issues, we evaluated the scope and the methods used by reviews that examine the effectiveness of MPAs as a tool for mitigating the impacts of fisheries on biodiversity. We carried out a systematic search for relevant reviews and categorised the scope of each review according to: (i) the geographic region(s) explored (global, temperate, tropical, polar), (ii) the taxa considered (fish, invertebrates, algae, birds, mammals, reptiles), (iii) the characteristics of MPAs investigated (size, age, level of protection, size of buffer zone, connectivity), and (iv) the measures used to evaluate MPA effectiveness (abundance, biomass, species richness, size distribution of individuals within or amongst species). We then assessed the reliability (objectivity, transparency and comprehensiveness) of each review using a standardised, published protocol (Woodcock et al., 2014), and identified general strengths and weaknesses in the review literature. Finally, we combined the categorisation of review scope with the assessment of review rigour to describe the review landscape on MPA effectiveness.

The principal objectives of our study are therefore to:

1) Assist decision-makers in quickly identifying the most relevant and rigorous reviews on topics of interest, and any limitations in the evidence used.
2) Assist decision-makers and researchers in targeting gaps in the review literature and avoiding duplication of previous reviews.

3) Identify strengths and weaknesses in the methods used by reviews, to assist researchers in maintaining and improving the rigour of future evidence syntheses.

We focused on reviews that synthesised empirical research on MPA effectiveness. Empirical data represent a large and growing volume of evidence, and reviews of this research have clear potential to support decision-making if results are provided on the outcomes of implementing MPAs, or on the characteristics that influence MPA effectiveness (e.g. Lester et al. 2009; Sciberras et al. 2013). We stress however, that the findings from such reviews should be considered in conjunction with insights from the extensive body of theoretical work on MPA effectiveness (e.g. Gaines et al. 2003; White et al. 2011), as well as site-specific considerations relating to stakeholder priorities and the objectives of individual MPAs.

Materials and methods

Review searching and screening

We compiled a database of review articles that examined MPA effectiveness through searches of peer-reviewed and grey literature using multiple databases (Web of Science, Scopus, Aquatic Sciences and Fisheries Abstracts, ScienceDirect, Centre for Agriculture and Bioscience International, Directory of Open Access Journals, and Index to Theses online), www.googlescholar.com, and websites of a range of organisations (Table A1). We used search terms adapted from a recent systematic review that evaluated the effectiveness of fully and partially protected MPAs (Sciberras et al., 2013, Sciberras et al., 2015). Search strings were modified according to the database used, but included the terms 'marine protected area', 'marine reserve', 'marine sanctuary', 'no-take area', 'partially protected area', 'fishery reserve', 'marine area closure', 'gear restriction zone' and 'buffer zone' to identify research related to MPAs. To narrow the focus to review articles we combined these terms with 'review, 'meta-analysis' and 'synthesis' For example, the search string used for locating studies in Web of Science, AFSA, CABI, ScienceDirect and Scopus was:

("marine reserve"* OR *"marine sanctuary"* OR *(marine* AND *"no-take zone")* OR *(marine* AND *harvest refug*)* OR *(marine* AND *"buffer zone")* OR *(marine* AND *partial** AND *protect*)* OR *(marine* AND *"closed area")* OR *(marine* AND *"area closure")* OR *(fisher** AND *(reserve* OR *closure))* OR *("fishing gear restriction*")* OR *("recreational fishing"* AND *protection)* OR *"marine protected area*")*

AND

(*review* OR *"meta-analy*"* OR *synthes**)

We only considered reviews published in the year 2000 or later to restrict our assessment to recent literature. Searches took place from 14-21 May 2014 and so our study encompasses the period from 2000 up to this point. Full details of the search strategy including search strings with Boolean operators and search dates are given in Table A1.

All studies found by the search were assessed for relevance and retained if the following inclusion criteria were met: Type of Article: Relevant reviews should be focused on synthesising primary research that collects field data to compare MPAs (fully or partially protected) with unprotected areas. This excludes articles clearly marked as opinions, perspectives, technical reports/management documents that are not explicitly presented as syntheses, modelling studies in which parameters are estimated through literature review, and studies that analyse long-term survey data (including articles that apply meta-analytical techniques – e.g. Ojeda-Martinez et al., 2007). Reviews primarily focused on synthesising the results from models, or on methodological aspects of MPA monitoring and evaluation were also excluded, as were reviews that focused on the ecological principles of MPA design (rather than synthesising empirical research on MPA effectiveness). Whilst each of these pieces of evidence are potentially valuable, it would not be appropriate to evaluate such studies using a tool designed for assessing reviews of primary research. For example, analyses of long-term survey data would not necessarily be expected to follow all of the methods required to produce a rigorous review of primary research (e.g. comprehensive and transparent search for relevant literature, assessment of publication bias etc.). Population: Reviews can consider any taxa, in any region. Intervention: Reviews must primarily examine the effects of fully and/or partially protected MPAs. Outcome: Reviews must clearly examine the effectiveness of MPAs with respect to at least one of: abundance, species richness, biomass, organism size. Because our emphasis was primarily on the direct implications of MPAs for biodiversity conservation and mitigating the impacts of fisheries, reviews that focused principally on ecosystem properties (e.g. nutrient cycling) or ecological processes (e.g. competition, trophic interactions) were not considered. Questions relating only to socio-economic effects also fall outside the scope of our study.

We screened all articles returned by the search for relevance, first based on the title with retained articles then assessed based on the abstract. Decisions on article inclusion can be subjective and so 10% of articles

166 screened at the abstract stage were also independently evaluated for relevance by a second person. Following
167 conventional practice for systematic reviews (CEE 2013), kappa values were used to evaluate agreement on
168 article relevance (Cohen 1960, Landis & Koch 1977). Kappa values account for the agreement expected by
169 chance, and are calculated as:

170 κ = (observed agreement-expected agreement) / (1-expected agreement)

171 'Observed agreement' is the proportion of decisions in which there is agreement (i.e. both assessors regard
172 an article as relevant, or both assessors regard an article as non-relevant). 'Expected agreement' is calculated
173 as: [(proportion of articles accepted as relevant by A1 * proportion of articles accepted by A2) + (proportion
174 of articles rejected by A1 * proportion of articles rejected by A2)], where A1 and A2 are the two assessors.
175 Kappa scores of 0.6-0.8 tend to be regarded as indicating good agreement: we obtained a kappa score of
176 0.75, indicating that decisions over article relevance were sufficiently repeatable (CEE, 2013). Where there
177 was disagreement on relevance during Abstract screening, articles were retained. All articles retained after
178 the abstract screening stage were then read in full and assessed for relevance. Articles in which the relevance
179 was uncertain at a particular stage were retained for the subsequent stage. Lastly, bibliographies of all
180 reviews retained after full-text screening were searched for additional references – this approach increases
181 the comprehensiveness of our search by capturing relevant reviews that may have omitted our search terms
182 from the Abstract. Any potentially relevant studies located in this way were screened using the same
183 title>abstract>full-text process.

184 *Assessing review scope*

185 We compiled 153 questions related to the effectiveness of MPA for biodiversity conservation and mitigating
186 the impacts of fisheries. The parameters of these questions are provided in Table 1 and consider region (e.g.
187 global, tropical etc.), taxa (fish, invertebrates etc.), MPA characteristics (e.g. size, age etc.) and outcome
188 measures (e.g. effects on abundance, biomass etc.). Questions therefore take the broad form: 'What are the
189 effects of MPAs on [*fish*]?', 'How does MPA [*size*] influence effectiveness?' etc. At this level, there are 19
190 distinct questions, representing each element in Table 1. We then considered each possible two-way
191 combination of Taxa, Region, MPA Characteristic, and Outcome Measure to assess specific questions, e.g.
192 'What are the effects of MPA [*size*] on [*fish*]?', What are the effects of [*tropical*] MPAs on [*species*

193 *richness*]?' At this level, there are 134 distinct questions. Note that constructing questions by systematically
194 combining terms in this way results in some questions that are likely to be more relevant than others.
195 However, our intention is for the evaluation of review scope and rigour to be valuable to policymakers and
196 researchers with a diverse range of priorities. In the case of the MPA literature, much of the research focus
197 has been on harvested species, whereas policy questions are increasingly broad, addressing a wide range of
198 taxa (e.g. EU Birds Directive, EU Habitats Directive). For this type of exercise, we therefore view a
199 systematic approach as preferable to identifying questions in a more *ad hoc* manner based on perceived
200 importance.

201 Reviews were categorised according to the question(s) addressed and the type of synthesis undertaken
202 (narrative synthesis [reviews that use prose to summarise and draw conclusions from primary research] or
203 meta-analysis). For the purposes of this study, we did not focus on any specific element of MPA connectivity
204 – reviews examining how any aspect of MPA connectivity influences effectiveness were therefore
205 considered to address this question. Defining the questions addressed by narrative syntheses proved
206 challenging because such syntheses are often broad-ranging with no clear boundaries to objectively decide
207 whether or not a particular question has been addressed in sufficient detail. As such, narrative syntheses were
208 categorised according to the broad focus (biodiversity conservation or fisheries) and the region, type of
209 protection (highly protection MPA [no-take] or all forms of protection [MPA]) and MPA characteristic(s)
210 explored. Each meta-analysis was categorised according to all 153 questions outlined in the preceding
211 paragraph. We categorised a meta-analytical review as addressing a particular question if effect sizes were
212 quoted directly (e.g. response ratio comparing fish density inside vs outside MPA), presented graphically or
213 used in statistical tests of relationships (e.g. relationship between effect size and MPA size).

214 In calculating effect sizes for one property (e.g. the influence of MPA size), meta-analyses could include
215 other properties as potential confounding variables (e.g. MPA age), without directly calculating effect sizes
216 for these confounding variables. From a policy perspective, it would therefore not be possible to use such an
217 analysis to fully understand the relationship between MPA age and effectiveness. Reviews that included
218 relevant terms (from Table 1) as potentially confounding variables without directly reporting effect sizes for
219 these terms were therefore noted (Supplementary Information) but not considered to directly address
220 questions relating to the confounding variables. Finally, there may be instances in which meta-analyses are

based on a small number of primary studies and so the generality of findings would be less certain. To be included as addressing a particular question, we set an arbitrary minimum threshold that meta-analyses should contain at least 10 primary research studies addressing that question. Where meta-analyses addressed a particular question but contained less than 10 studies, we noted this intended focus, as well as instances in which reviews indicated an intention to investigate a question but expressly stated that insufficient studies were available. If recent, such reviews might suggest the need for further primary research rather than additional reviews.

Critical appraisal of review reliability

We used a standardised protocol designed to assess the reliability of environmental evidence reviews (the Collaboration for Environmental Evidence Synthesis Assessment Tool [CEESAT], Woodcock et al., 2014) to critically appraise the methods of each relevant review. CEESAT assesses reviews based on 13 criteria (Table 2) for which a review can receive 3 points, 1 point or 0 points (maximum 39). The higher the score, the greater the confidence that the review is robust. Whilst CEESAT does have important limitations (e.g. does not account for methodological or interpretation errors or fraud, or include a detailed evaluation of the appropriateness of any statistical techniques used) it considers each key step of the review process and so provides a good overall picture of the likelihood that the review uses transparent methods to produce an objective, rigorous, and comprehensive synthesis of all available primary research.

All reviews were independently appraised by two assessors using CEESAT. Disagreements in scoring were then discussed and scores were amended if relevant information was overlooked by an assessor. When disagreements reflected uncertainty between assessors over whether or not a criterion was met, the midpoint score of the two assessors was used. We evaluated agreement in scoring by using a Spearman's rank test to compare the overall scores for each review between assessors, and by examining repeatability in scoring for individual criteria using (i) % agreement and (ii) kappa test as described above, but extended to the three scoring categories of 0, 1, or 3. However, larger disagreements in the scores awarded for a criteria are more important than smaller disagreements (e.g. if one scorer awards a 0 for a given criterion, it would be more concerning if the second scorer awarded a 3 than a 1). As such, we also assessed agreement in scoring for each criterion using weighted kappa (Cohen 1968; Landis & Koch 1977; Viera & Garrett 2005; Shea et al. 2007). Matrices of the observed scores awarded by the two assessors were produced for each individual

criterion, giving 13 separate matrices, each containing nine cells indicating the number of reviews awarded 0, 1, or 3 points by each assessor. Similar matrices of expected scores for each criterion were calculated as for a chi-squared test. A matrix containing nine cells representing the magnitude of disagreement between assessors was then constructed, e.g. a 1-0 disagreement is ranked as magnitude 1, whereas a 3-0 disagreement is ranked as magnitude 3 (Viera and Garrett, 2005, Shea et al., 2007). For an individual criterion, each cell in the observed matrix is then multiplied by the corresponding weight (e.g. cells where there is a 3-1 disagreement or a 1-3 disagreement are multiplied by 2). The observed weighted disagreement for that criterion is the sum of these values, with the expected weighted disagreement calculated in the same manner. The weighted kappa score for a criterion (which reflects *agreement*, and is interpreted in the same way as the unweighted kappa) is then:

κ = 1-(observed weighted disagreement/expected weighted disagreement)

Lastly, we divided the total CEESAT scores into three categories: 0-13, 13.5-26 and 26.5+ (reflecting an average score across the 13 criteria of 0-1, 1-2 and 2-3) to represent low, intermediate/moderate and high reliability (although see Woodcock et al., 2014 for further discussion regarding the interpretation of scores). Each review was assigned to one of these reliability categories based on the overall CEESAT score.

Evaluating the review landscape

Using our critical appraisal and assessment of review scope we then visually represented reviews examining the effectiveness of MPAs for biodiversity conservation and mitigating the impacts of fisheries in two matrices, one covering meta-analytical reviews and one summarising narrative syntheses. These matrices were designed to guide decision-makers to the most relevant and reliable reviews, and to enable easy visualisation of gaps and redundancy (multiple reviews on closely related topics) to target future reviews. Detailed information indicating which reviews address each specific question is given in a series of supporting tables. Strengths of MPA reviews and aspects of review methods that could be improved were explored and evidence gaps and redundancy were identified.

Results

Searching and screening

Searches (Table A1) returned 2,485 results; these were refined to 287 after screening at title stage, 98 after abstract screening, and finally reduced to 24 relevant reviews following full-text examination. The bibliographies of relevant reviews were then hand-searched for additional references, giving a final total of 27 included reviews. To maintain transparency, a complete list of all included and excluded articles (at full text) together with reasons for exclusion is provided in Table A2-A3.

Review scores

Review scores ranged from 0-34 (mean = 12.3 $\pm$ 1.8 standard error [SE]), median = 13.5, Fig. 1a): note that because scores are the average across the two assessors, non-integer values are possible. Although no review achieved the maximum score of 39, the maximum possible points (3) were awarded for each criterion at least once. The majority of reviews (93%) achieved low ($\leq$13, N=13) or intermediate (13-26, N=12) scores. Criteria 3.1, 5.1 and 6.1 represented particular strengths (see Table 2 for explanation of criteria), whilst criteria 1, 3.2, 3.3 and 6.3 were consistent weaknesses in review conduct (Fig. 1b). Of the 27 reviews assessed, 18 contained meta-analysis and nine conducted a narrative synthesis. As would be expected, the mean score for meta-analyses was higher than for all reviews combined (mean = 17.3 $\pm$ 1.6 SE, median = 16), although a substantial range of scores was still evident (9.5-34).

Repeatability of scoring

The total scores awarded to each review were highly correlated between assessors (Spearman's ρ=0.96, $p<0.001$) and the mean absolute difference in scores was small (1.7 $\pm$ 0.3). Scoring for individual criteria was also generally consistent: for 11 of 13 criteria, agreement was greater than 70% and weighted kappa scores were around 0.7 or higher (Table 3; substantial agreement, Landis and Koch, 1977). The latter indicates that most disagreements were relatively minor (e.g. 0 vs 1, rather than 0 vs 3).

Scope of meta-analytical reviews

Most of the broad question elements in Table 1 were examined to some degree by meta-analyses but a clear focus was apparent with respect to the taxa (fish), MPA characteristics (size and age), and outcome measures (abundance) considered (Fig. 2, Table A4). For example, the vast majority of meta-analyses examined if/how MPA size influences the effectiveness of the protected area, with fewer investigations into the importance

of other characteristics such as the level of protection (N=3), buffer zone size (N=2) or connectivity (N=1). No meta-analyses were detected that examined the effectiveness of MPAs in polar regions, or the effects of MPAs on birds, mammals or reptiles. At least one high-scoring review (≥26 points) was available for 11 of the broad questions, although these questions were also the subject of low-moderate scoring reviews. Two broad questions ('effects of MPAs on algae' and 'influence of connectivity on MPA effectiveness') were each only addressed by one moderate scoring review, representing cryptic evidence gaps that might not be readily apparent.

Gaps across review questions became more pronounced when specific questions were considered (Fig. 2, Tables A5-A10). Several more specific questions were the subject of multiple reviews with at least one high scoring review (e.g. 'fish' and 'abundance') demonstrating that there is some duplication in the review literature. However there are also examples of cryptic evidence gaps in which reviews were present but none were high-scoring (e.g. 'tropics' and 'species richness'). Furthermore, in addition to an absence of reviews considering polar regions, birds, mammals and reptiles, no reviews were identified for a further 15 specific questions, and an additional seven questions were either not addressed due to the low number of primary studies available, or were only addressed through a synthesis of <10 studies.

Scope of narrative syntheses

Narrative syntheses were generally of broader scope than meta-analyses (Fig. 3, Tables A11-A17). The majority (7 out of 9 narrative reviews) discussed the effects of MPAs globally rather than focusing on specific regions. The focus was split evenly between MPA effects on biodiversity and fisheries and most reviews considered MPAs as a whole rather than just highly protected (no-take) MPAs. No moderate- or high-scoring narrative reviews were identified (range in narrative review scores 0-12). Gaps and/or redundancy were noted in the majority of questions.

Discussion

The increasing importance of MPAs in global conservation strategies has stimulated extensive primary research examining the effectiveness of MPAs for mitigating the impacts of fisheries (Lester et al., 2009, Caveen et al., 2012). Reliably incorporating this research into policy requires syntheses that use systematic, with objective methodologies to address key questions. However, our findings highlight substantial variation

328 in scope and rigour amongst reviews that examine the effectiveness of MPAs for biodiversity conservation
329 (Figure 1). This variation illustrates the need to ensure non-specialists can locate the most rigorous reviews
330 on questions of interest, and parallels that found in other fields of ecology and environmental management
331 (Philibert et al. 2012; Koricheva et al. 2014; O'Leary et al. 2016). Furthermore, we found that reviews
332 strongly favoured particular questions – e.g. exploring if the size and age of an MPA influenced the effects
333 on fish. Whilst these questions are vital for both biodiversity conservation and fisheries management, there
334 is a danger that findings from such syntheses could be extrapolated to other taxonomic groups (e.g. birds,
335 mammals) and that rigorous syntheses exploring the importance of other MPA design characteristics (e.g.
336 connectivity) will not be undertaken. The summary of review scope and rigour provided here can assist
337 future reviews in ensuring that the intended questions complement rather than duplicate the existing review
338 literature. Note that our study encompasses the time period from 2000-2014. As with reviews of primary
339 literature, the value of this information as a means to explore review rigour and scope will thus be maximised
340 if updates are conducted after a suitable time period: by providing detailed methods and transparent
341 descriptions we hope to facilitate such updates.

342 *General strengths and weaknesses in the conduct of reviews: implications for policy and research*

343 We found reviews to be of variable reliability with often overlapping scope (Figs. 2 and 3, Tables A4-A17).
344 Reviews regularly applied several approaches (e.g. meta-analytical techniques and transparent reporting of
345 inclusion criteria) that are important for rigorous synthesis (Fig. 1a and b). However, certain aspects of MPA
346 review conduct could be improved, such as ensuring that decisions over which articles are relevant to include
347 in the review are repeatable and transparent (by conducting kappa tests and listing all articles read at full-
348 text but excluded from the synthesis) and that critical appraisal of the methods of included studies is
349 undertaken and clearly reported (Fig. 1b). Narrative reviews were all assessed as being of low reliability
350 (N=9), partly reflecting the lack of quantitative synthesis. Nonetheless, there is no inherent reason that
351 narrative reviews cannot, for example, provide clear information on search strategies and scope, and
352 document the extracted data. Given that narrative reviews can still contain valuable insights (and do
353 influence policy) we argue that such reviews could benefit considerably from adopting such practices –
354 indeed, one narrative review (Peppin et al. 2011) assessed during the initial evaluation of CEESAT achieved
355 a score of 20 (Woodcock et al. 2014), which is similar to many of the meta-analyses considered here.

356 Scoring cannot distinguish between reviews undertaken using less rigorous methods and those that do not
357 document rigorous methods where used. Just as with primary research, transparent reporting of review
358 methods is vital, because it allows the review to be verified and updated. We therefore highlight the
359 importance of effective reporting, and suggest that this represents a relatively straightforward means by
360 which many reviews (narrative and meta-analyses) could be improved. More generally, we stress that in our
361 view, limitations in reviews in environmental science reflect a lack of awareness of relevant systematic
362 review methods, rather than a deliberate intention to mislead.

363 From a policy perspective, the large number of reviews with low-intermediate scores represents a potential
364 problem. In low-intermediate scoring reviews, steps that are important for producing a comprehensive,
365 objective, and transparent evidence synthesis are either absent or incomplete. Such limitations reduce the
366 likelihood that the review provides an accurate picture of all available primary research. Although the effects
367 of omitting certain steps on review reliability and findings are context-specific, in the absence of clear
368 mechanisms to communicate the rigour of review methods to non-specialists, there is a risk that decision-
369 makers will not take into account potential limitations in the conduct of the review(s) consulted.

370 *Redundancy in the review literature*

371 We identified substantial redundancy in the review literature (multiple reviews asking the same question)
372 which could create difficulties for decision-makers looking to base decisions on the most robust synthesis
373 available. In some instances, redundancy is a consequence of reviews providing effect sizes for broader
374 questions and then exploring a range of more specific questions, or updating a particular question. Although
375 such analyses are valuable for completeness and comparisons, decision-makers often lack the resources to
376 locate and evaluate all relevant reviews. These situations therefore risk leading to policy and practice that is
377 not based on the most rigorous available evidence. As such, we hope that the results from studies such as
378 ours can assist decision-makers in rapidly locating the reviews most likely to accurately synthesise all
379 relevant evidence on the specific questions of interest. These outputs may also inform future research
380 direction. For example, questions relating to fish abundance and MPA size have been the subject of reliable
381 meta-analyses and so in the absence of substantial new research, attention might be better focused on
382 synthesising evidence on other questions. Such investigations might include consideration of a broader range

of taxa and MPA characteristics, as well as more specific factors that influence the effects of MPAs on fish abundance in order to inform on the degree to which findings are generalisable.

Gaps in the review literature

Gaps in the review literature are to be expected to an extent as a result of differences in public interest, policy relevance, availability of primary research (potentially influenced by e.g. logistical constraints in sampling fauna or flora), and question validity. However, some evidence gaps are in areas of high policy relevance. For example, the protection of seabirds and marine mammals is an important driver of MPA designation under European Directives, and MPAs are globally important tools in the conservation of a range of taxa (Hooker & Gerber 2004; Christianen et al. 2014). Furthermore, designation (or non-designation) can be controversial, and greater confidence in decisions would likely arise if robust evidence syntheses on the effectiveness of MPAs for multiple taxa were available. Our study also suggests some differences in the availability of reviews on tropical versus temperate MPAs. Relatively few meta-analyses quantify the effectiveness of MPAs in the former, particularly for less well-studied taxa and certain aspects of MPA design (note that global-scale reviews incorporating primary research from the tropics do not necessarily specifically evaluate the effectiveness of tropical MPAs). Decision-makers in the tropics would therefore be reliant on moderately reliable syntheses from this region and/or global syntheses that combine data from temperate and tropical MPAs. This contrasts with temperate regions, for which the effects of MPA size and age are specifically quantified by several reviews (Figure 2-3 and Supporting Information). There are also some instances in which reviews have been conducted but a highly rigorous synthesis is lacking (Figure 2-3). These could represent cryptic evidence gaps, in which the presence of existing reviews may create the perception that the question has been considered, and potentially deter new synthesis or primary research for several years.

Identification of gaps in the review literature highlights the need for new, more reliable syntheses (or primary research) to be conducted, providing a more solid basis for policy. Importantly, gaps become more frequent as questions become more specific, indicating that users should consider how applicable more general reviews are to particular contexts. It is important to stress that our primary emphasis was on properties relevant to the effectiveness of MPAs as a conservation tool for mitigating the impacts of fisheries on biodiversity. Valuable extensions of our study could therefore more specifically consider the review

literature examining the extent to which MPAs provide fisheries benefits, as well as possible gaps in terms of the effects of MPAs on ecosystem functioning (e.g. productivity, nutrient cycling, food web structure), more sophisticated outcome metrics relating to conservation effectiveness (e.g. IUCN threat status), and the socioeconomic consequences of MPAs.

Conclusions

MPAs are a key component of global conservation strategies, but there is considerable uncertainty surrounding when and where reserves are most effective. Evidence reviews examining the effectiveness of MPAs are therefore likely to directly influence decision-making and future research. However, the overlapping scope and variation in reliability we identified amongst reviews presents a potentially important problem from the perspective of decision-makers seeking to make evidence-informed decisions. Our evaluation of reviews is intended to support decision-making by guiding non-specialists to the most reliable and relevant reviews. Findings from such reviews should be considered alongside other key pieces of evidence, in particular the extensive body of theoretical work on MPA effectiveness (e.g. Gaines et al. 2003; White et al. 2011) and more context-specific information relating to individual MPAs. Our findings can also assist researchers in identifying and targeting key knowledge gaps for review or new data collection including (but not limited to) ensuring broader taxonomic coverage, consideration of a wider range of MPA characteristics and examination of more specific questions for which we have identified evidence gaps.

Acknowledgements

We thank our potential end-users Ally Dingwall (Sainsbury's), Tom Pickerell (Seafish, Seafood Watch), Jon Harman (Seafish), Mike Mitchell, David Parker (Young's Seafood), David Jarrad (Shellfish Association of Great Britain) for their contribution to discussions regarding review reliability. This project was supported by a UK Natural Environmental Research Council Knowledge Exchange Grant NE/J006386/1.

References

Butchart, S.H.M., Walpole, M., Collen, B.*, et al.* (2010) Global Biodiversity: Indicators of Recent Declines. *Science* **328,** 1164-1168.

436 Caveen, A.J., Sweeting, C.J., Willis, T.J., Polunin, N.V.C. (2012) Are the scientific foundations of temperate
437 marine reserves too warm and hard? *Environmental Conservation* **39,** 199-203.
438 CBD (2010) COP Decision X/2. Strategic plan for biodiversity 2011–2020. Available at:
439 http://www.cbd.int/decision/cop/?id=12268. Accessed 6 April 2015.
440 CEE (2013) Guidelines for systematic review and evidence synthesis in environmental management.
441 Christianen, M.J.A., Merman, P.M.J., Bouma, T.J., *et al*. (2014) Habitat collapse due to overgrazing
442 threatens turtle conservation in marine protected areas. *Proceedings of the Royal Society B* **281**
443 20132890, doi:10.1098/rspb.2013.2890.
444 Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological*
445 *Measurement* **20**, 37-46
446 Cohen, J. (1968) Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological*
447 *Bulletin* **70** 213-220
448 Fox, H.E., Soltanoff, C.S., Mascia, M.B., *et al.* (2012) Explaining global patterns and trends in marine
449 protected area (MPA) development. *Marine Policy* **36,** 1131-1138.
450 Gaines, S.D., Gaylord, B., Largier, J.L. (2003) Avoiding current oversights in marine reserve design.
451 *Ecological Applications*, **13**, S32-46.
452 Gaines, S.D., Lester, S.E., Grorud-Colvert, K. et al. (2010) Evolving science of marine reserves: New
453 developments and emerging frontiers. *Proceedings of the National Academy of Sciences* **107**,
454 18251-18255.
455 Grorud-Colvert, K., Claudet, J., Tissot, B.N., *et al.* (2014) Marine Protected Area networks: assessing
456 whether the whole is greater than the sum of its parts. *PLoS ONE* **9,** e102298.
457 Halpern, B.S., Lester, S.E., McLeod, K.L. (2010) Placing marine protected areas onto the ecosystem-based
458 management seascape. *Proceedings of the National Academy of Sciences* **107**, 18312-18317.
459 Halpern, B.S., Longo, C., Hardy, D., *et al.* (2012) An index to assess the health and benefits of the global
460 ocean. *Nature* **488,** 615-620.
461 Hays, G.C., Scott, R. (2013) Global patterns for upper ceilings on migration distance in sea turtles and
462 comparisons with fish, birds and mammals. *Functional Ecology* **27,** 748-756.
463 Hooker, S.K., Gerber, L.R. (2004) Marine reserves as a tool for ecosystem-based management: the potential
464 importance of megafauna. *BioScience* **54**, 27-39.

Huntington, B.E. (2011) Confronting publication bias in marine reserve meta-analyses. *Frontiers in Ecology and Environment* **9,** 375-376.

Koricheva, J. & Gurevitch (2014) Uses and misuese of meta-analysis in plant ecology. *Journal of Ecology* **102**, 828-844.

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

Larsen, P.O., von Ins, M. (2010) The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **84,** 575-603.

Lascelles, B.G., Langham, G.M., Ronconi, R.A., Reid, J.B. (2012) From hotspots to site protection: identifying marine protected areas for seabirds around the globe. *Biological Conservation* **156,** 5-14.

Lester, S.E., Halpern, B.S., Grorud-Colvert, K.*, et al.* (2009) Biological effects within no-take marine reserves: a global synthesis. *Marine Ecology Progress Series* **384,** 33-46.

Lubchenco, J., Grorud-Colvert, K. (2015) Making waves: The science and politics of ocean protection. *Science* **23**, 382-383.

Li, W., Zhao, Y. (2015) Bibliometric analysis of global environmental assessment research in a 20-year period. *Environmental Impact Assessment Review* **50,** 158-166.

Marine Conservation Institute (2015) MPAtlas. Available from: www.mpatlas.org. Accessed 27/10/2015.

Marinesque, S., Kaplan, D.M., Rodwell, L.D. (2012) Global implementation of marine protected areas: is the developing world being left behind? *Marine Policy* **36,** 727-737.

Møller, A.P., Jennions, M.D. (2001) Testing and adjusting for publication bias. *Trends in Ecology & Evolution* **16,** 580-586.

Ojeda-Martinez, C., Bayle-Sempere, J.T., Sanchez-Jerez, P., Forcada, A., Valle, C. (2007) Detecting conservation benefits in spatially protected fish populations with meta-analysis of long term monitoring data. *Marine Biology* **151,** 1153-1161.

O'Leary, B.C., Kvist, K., Bayliss, H.R., Derroire, G., Healey, J.R., Hughes, K., Kleinschroth, F., Sciberras, M., Woodcock, P., Pullin AS (2016) The reliability of evidence review methodology in environmental science and conservation. *Environmental Science and Policy* **64**, 75-82.

493 OSPAR (2010) The North-East Atlantic environment strategy of the OSPAR Commission for the protection
494 of the marine environment of the North-East Atlantic 2010-2020 (OSPAR Agreement 2010-2013).
495 Pautasso, M. (2012) Publication growth in biological sub-fields: patterns, predictability and sustainability.
496 *Sustainability* **4,** 3234-3247.
497 Peppin, D., Fule, P., Beyers, J. et al. (2011) Does seeding after severe forest firest in western USA mitigate
498 impacts on soils and plant communities. CEE review 08-023 (SR60). Collaboration for
499 Environmental Evidence. www.environmentalevidence.org/SR60.html.
500
501 Philibert, A., Loyce, C., Makowski, D. (2012) Assessment of the quality of meta-analyses in agronomy.
502 *Agriculture, Ecosystems & Environment* **148**, 72-82.
503 Ramirez-Llodra, E., Tyler, P.A., Baker, M.C*., et al.* (2011) Man and the last great wilderness: human impact
504 on the deep sea. *PLoS ONE* **6,** e:22588.
505 Sciberras, M., Jenkins, S.R., Kaiser, M.J., Hawkins, S.J., Pullin, A.S. (2013) Evaluating the biological
506 effectiveness of fully and partially protected marine areas. *Environmental Evidence* **2:4**.
507 Sciberras, M., Jenkins, S.R., Mant, R., Kaiser, M.J., Hawkins, S.J., Pullin, A.S. (2015) Evaluating the relative
508 conservation value of fully and partially protected marine areas. *Fish and Fisheries* **16,** 58-77.
509 Shea, B.J., Bouter, L.M., Peterson, J*., et al.* (2007) External validation of a measurement tool to assess
510 systematic reviews. *PLoS ONE* **2,** e:1350.
511 Thomas, H.L., Macsharry, B., Morgan, L. et al. (2014) Evaluating official marine protected area coverage
512 for Aichi Target 11: appraising the data and methods that define our progress. *Aquatic*
513 *Conservation: Marine and Freshwater Ecosystems* **24**, 8-23.
514 Viera, A.J., Garrett, J.M. (2005) Understanding interobserver agreement: the kappa statistic. *Family*
515 *Medicine* **37,** 360-363.
516 White, J.W., Botsford, L.W., Baskett, M.L., *et al.* (2011) Linking models with monitoring data for assessing
517 performance of no-take marine reserves. *Frontiers in Ecology and the Environment*, **9**, 390-399.
518 Willis, T.J., Millar, R.B., Babcock, R.C., Tolimiera, N. (2003) Burdens of evidencs and the benefits of
519 marine reserves: putting Descartes bes des horse? *Environmental Conservation* **30,** 97-103.

520 Woodcock, P., Pullin, A.S., Kaiser, M.J. (2014) Evaluating and improving the reliability of evidence
521 syntheses in conservation and environmental science: a methodology. *Biological Conservation* **176,**
522 54-62.
523

524 **Supporting Information**

525 Additional Supporting Information may be found in the online version of this article: **Tables A1-A17**.

526 **Table 1**: Key aspects of Marine Protected Areas (MPAs) that may influence effectiveness (geographic
527 region, taxon of interest, design characteristic), and outcome measures frequently used to assess MPA
528 effectiveness.

Region	**Taxa**	**MPA Characteristic**	**Outcome Measure**
Global	Fish	MPA Size	Abundance
Tropical	Invertebrate	MPA Age	Biomass
Temperate	Algae	MPA Connectivity	Species Richness
Polar	Mammal	MPA Buffer Zone Size	Organism Size
	Bird	MPA Protection Level	
	Reptile		

529

530 **Table 2:** Criteria and brief rationale for the Collaboration for Environmental Evidence Synthesis Assessment

531 Tool (CEESAT). See Woodcock et al. (2014) for details.

	Criteria	Rationale
1.1	Was an *a-priori* protocol available for comment before the synthesis was conducted?	Increases robustness of review against *post hoc* changes in methods and scope.
2.1	Does the search for literature use a comprehensive range of resources?	Increases likelihood that all potentially relevant articles are captured by search.
2.2	Are the search strings clearly defined?	Allows search to be repeated and evaluated. Avoids open-ended searches.
3.1	Does the review apply clearly documented inclusion criteria to all potentially relevant studies found during the search?	Increases transparency. Reduces risk of subjective decisions influencing the studies included in the review.
3.2	Does the review demonstrate that inclusion decisions are repeatable?	Demonstrates repeatability of review, and that subjective decisions have not overly influenced the articles included.
3.3	Are inclusion/exclusion decisions transparent?	Ensures that the process of including and excluding studies can be externally verified.
4.1	Does the review report critical appraisals of the methods of each study?	Makes quality of the evidence-base for the synthesis clear.
4.2	Are studies objectively weighted according to methodological quality?	Gives greater emphasis to more robust studies.
5.1	Is data extraction documented, repeatable and consistent?	Reduces potential for bias in the extraction of metrics from individual studies.
5.2	Are the extracted data reported for each study?	Ensures that the extracted data can be verified and analysed by readers.
6.1	Is a quantitative synthesis conducted?	Reduces potential for subjectivity to influence data synthesis.
6.2	Is heterogeneity in the impact of the intervention investigated statistically?	Indicates the degree to which results are generalisable and the appropriateness of combining studies.
6.3	Does the review consider possible publication bias?	Reduces potential for bias arising from non-publication of non-significant or controversial results.

532

Table 3: Agreement in scoring between reviewers. Data shown for each criterion are % of reviews for which the two reviewers awarded the same score, kappa test, and kappa test weighted by magnitude of disagreement. Kappa score of 1 = perfect agreement, kappa score of 0 = agreement no different from that expected by chance).

	Criteria	Agreement (%)	Kappa	Weighted Kappa
1.1	Protocol	100	1.00	1.00
2.1	Search resources	85	0.70	0.80
2.2	Search string stated	41	0.15	0.36
3.1	Documented inclusion criteria	74	0.51	0.72
3.2	Evidence that inclusion decisions repeatable	100	1.00	1.00
3.3	Documented exclusion decisions	85	0.53	0.67
4.1	Critical appraisal of methods	81	0.65	0.71
4.2	Objective weighting	78	0.62	0.70
5.1	Data extraction documented	78	0.59	0.70
5.2	Extracted data reported	59	0.35	0.49
6.1	Quantitative synthesis	96	0.92	0.97
6.2	Heterogeneity investigated	81	0.65	0.66
6.3	Publication bias considered	93	0.78	0.79

Fig. 1. **CEESAT scores for reviews examining the effectiveness of MPAs. (a)** total review scores, and **(b)** mean score ± S.E. for each criterion. Scores are white (mean score per criterion of <1), grey (mean score from 1-2), and black (mean score per criterion of >2). Higher scores indicate that the review demonstrates greater objectivity, transparency, and comprehensiveness, and is therefore more likely to provide an accurate reflection of the primary literature.

Fig. 2: Matrix summarising the reliability and scope of meta-analytical reviews that examine MPA effectiveness for biodiversity conservation. Matrix overview of the 19 broad and 134 specific questions we considered in our evaluation. Doughnut pie charts indicate the proportion of review achieving low (0-13; white), moderate (13-26; grey), or high (>26; black) CEESAT scores. Total number of reviews considering each question is in the centre of each chart. The matrix should be read using combinations from the top and left headings to form the question of interest; relevant reviews can then be found in Tables A4-10. For example, to explore the effect of MPA size on fish, locate MPA size under MPA Characteristics in the top set of headings and read down to fish under Taxa on the left; consult Table A6 for details of reviews. Stars indicate reviews that considered the question but with <10 primary studies, or stated that the question could not be investigated due to low number of primary studies. White areas indicate questions that are not applicable, e.g. Global/Temperate question combinations. Abbreviations in headings refer to: Outcome Measures - Abund=abundance and Sp.Rich=species richness; MPA Characteristics - Conn=connectivity, Buff=buffer zone size, Prot=level of protection; and Taxa – Invert=invertebrates.

Fig. 3: Matrix summarising the reliability and scope of narrative syntheses that examine MPA effectiveness for biodiversity conservation. Matrix should be read using combinations from the top and left headings to form the question of interest; full details of reviews can then be found in Tables A11-17. For consistency, shading of doughnut pie charts are as for Figure 2. In practice, all narrative reviews we assessed scored from 0-13, and so are coloured white. Blank areas indicate questions that are not applicable, e.g. Global/Temperate question combinations. Abbreviation 'Conn.' in MPA Characteristics refers to connectivity.