Harnessing artificial intelligence for efficient systematic reviews: A case study in ecosystem condition indicators

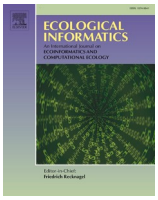
Isabel Nicholson Thomas, Philip Roche, Adrienne Grêt-Regamey

Ecological Informatics 83 (2024) 102819

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Harnessing artificial intelligence for efficient systematic reviews: A case study in ecosystem condition indicators

Isabel Nicholson Thomas [a,*], Philip Roche [b], Adrienne Grêt-Regamey [a]

[a] Planning of Landscape and Urban Systems, Institute for Spatial and Landscape Development, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland
[b] INRAE, Aix Marseille Univ, RECOVER, EMR Team, Aix-en-Provence, France

ARTICLE INFO

Keywords:
Artificial intelligence
Systematic review
Ecosystem condition
GPT

ABSTRACT

Effective evidence synthesis is important for the integration of scientific research into decision-making. However, fully depicting the vast mosaic of concepts and applications in environmental sciences and ecology often entails a substantial workload. New Artificial Intelligence (AI) tools present an attractive option for addressing this challenge but require sufficient validation to match the vigorous standards of a systematic review. This article demonstrates the use of generative AI in the selection of relevant literature as part of a systematic review on indicators of ecosystem condition. We highlight, through the development of an optimal prompt to communicate inclusion and exclusion criteria, the need to describe ecosystem condition as a multidimensional concept whilst also maintaining clarity on what does not meet the criteria of comprehensiveness. We show that, although not completely infallible, the GPT-3.5 model significantly outperforms traditional literature screening processes in terms of speed and efficiency whilst correctly selecting 83 % of relevant literature for review. Our study highlights the importance of precision in prompt design and the setting of query parameters for the AI model and opens the perspective for future work using language models to contextualize complex concepts in the environmental sciences. Future development of this methodology in tandem with the continued evolution of the accessibility and capacity of AI tools presents a great potential to improve evidence synthesis through gains in efficiency and possible scope.

1. Introduction

The importance of representative assessments of ecosystem condition, i.e. the quality of an ecosystem in terms of its abiotic and biotic characteristics (UNCEEA, 2021), is increasingly recognized in decision making (Vallecillo, 2022). However, the wide range of potential indicators used to describe ecosystem condition (Rendon et al., 2019) necessitates work to distil consistent methodologies and sets of metrics for reporting. At the same time, the number of published studies across all scientific disciplines continues to grow exponentially (Bornmann et al., 2021; Olander et al., 2017). Bridging the gap between research and implementation therefore requires first a substantial effort to compile the complex breadth of primary research into usable, relevant evidence for policy makers (Westgate et al., 2018).

Systematic reviews are well-established approaches allowing to synthesize evidence, produce an overview of the state-of-the-art in a scientific field, and identify priorities for further research. Various frameworks and guidelines exist to aid the completion of a replicable systematic review, including the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) (Moher et al., 2015) and Search, Appraisal, Synthesis, and Analysis (SALSA) (Grant and Booth, 2009) frameworks. The recognition of the need for such reviews and meta-analyses in decision-making has led to greater adoption of and advocation for these principles in the environmental science literature (Gerstner et al., 2017; Haddaway et al., 2015) and the development of the RepOrting standards for Systematic Evidence Syntheses (ROSES) framework specifically adapted for reviews in this domain (Haddaway et al., 2018). A synthesis produced by a systematic review can claim increased transparency and objectivity in its conclusions compared to a traditional literature review, on account of a strict adherence to these clear structure of methods and reporting (Mohamed Shaffril et al., 2021). Additionally, potential biases that could limit comprehensiveness are reduced through a widened selection of evidence outside of the reviewer's immediate scope of knowledge (Haddaway et al., 2015). However, as the volume of potential literature increases, the rigor of this methodology poses an increasingly resource-intensive challenge, with

* Corresponding author.
E-mail addresses: inthomas@ethz.ch (I. Nicholson Thomas), philip.roche@inrae.fr (P. Roche), gret@ethz.ch (A. Grêt-Regamey).

https://doi.org/10.1016/j.ecoinf.2024.102819
Received 1 March 2024; Received in revised form 5 September 2024; Accepted 7 September 2024
Available online 10 September 2024

the completion of a timely and comprehensive systematic review taking as long as several years and incurring substantial costs (Collaboration for Environmental Evidence (CEE), 2013).

To enhance the efficiency and accuracy of systematic review processes, a variety of software tools, both free and proprietary, are increasingly being employed. These tools leverage artificial intelligence (AI) to organize, prioritize, and preliminarily classify publications for review.(Atkinson, 2023; Gates et al., 2020; Khalil et al., 2022). Whilst studies using these methods have reported benefits with regards to efficiency, they can have varying reliability and efficacy (Blaizot et al., 2022; Khalil et al., 2022). Apart from these AI tools, systematic reviews have increasingly incorporated Machine Learning techniques, which differ by requiring the creation of a large, specialized training dataset (van Dinter et al., 2021) and significant technical understanding of the algorithms used (Ferdinands et al., 2023). Recent developments in generative AI have simplified its use for a range of users. With the release of publicly available chat-based interfaces for pre-trained Language Models such as the Generative Pretrained Transformer (GPT) series, Claude and LLAMA2 models and Application Programming Interfaces (APIs) with which to manipulate these models, the possibility to integrate AI into multiple stages of a system literature review has become more feasible (Atkinson, 2023).

Despite clear potential of AI to automate, simplify and accelerate the stages of evidence synthesis in environmental sciences, its application has so far mostly been restricted to research in the domains of Software Engineering and Medicine, where the use of systematic reviews is well-established (van Dinter et al., 2021). A systematic review aims to present an objective and reliable summary of a field of research and therefore demands a high level of transparency and accountability, including the use of appropriate, sufficiently validated and context-specific methodologies (Haddaway et al., 2015). However, key resources providing guidance on the use of systematic literature reviews in environmental science do not explicitly address how to approach the use of automation with AI in the context of a review (Collaboration for Environmental Evidence (CEE), 2013; Haddaway et al., 2015).

This paper aims to showcase the application and potential benefits of AI, particularly in automating the screening stages of systematic review processes, thereby increasing overall efficiency. The GPT models are a form of Large Language Model (LLM) trained on unlabelled text datasets with parameter sizes that would otherwise present inhibitive costs to individual researchers, and are able to generate human-like text responses (Floridi and Chiriatti, 2020). We demonstrate the utility of this approach through application to a systematic review of ecosystem condition indicators, a field characterized by diverse and sometimes conflicting terminology across various disciplines, and closely linked to a numerous concepts which incorporate the integrity and functioning of ecosystems (Rendon et al., 2019; Roche and Campagne, 2017). Previous reviews on ecosystem condition have therefore reduced the scope through a focus on specific applications of indicators (Maes et al., 2020; Smit et al., 2021), or based their analysis on the use of a limited list of synonyms (Rendon et al., 2019; Soubry et al., 2021). Ensuring the comprehensiveness of a synthesis of a body of evidence on this topic therefore requires broad search terms which can lead to a large number of potential publications for review. In the context of our review, we evaluate the performance of GPT-3.5 in terms of accuracy in classification of abstracts for review compared to expert human responses. Lastly, we offer some perspectives on the further integration of these methods into systematic reviews.

2. Material and methods

We here present the workflow used in this study. First, we identified literature of potential relevance for the review according to the guidelines of the PRISMA framework. We then developed code to query the GPT-3.5 completions API to provide a classification for papers. We compared the performance of the model to validation samples produced by expert reviewers and developed finally an optimal prompt through iterative testing.

2.1. Publications data source and search strategy

The scientific publications used for validation and testing of the approach were taken from a literature corpus compiled under the European Union Horizon project Science for Evidence-based and Sustainable Decisions about Natural Capital (SELINA). Within this project, multiple parallel reviews were planned to synthesize the state of current research in ecosystem condition, ecosystem services and ecosystem accounting. A systematic search was performed in the Web of Science and Scopus online citation databases to produce a central corpus which could be relevant for each review, thus avoiding the potential inconsistency in results retrieved from repeated querying of online citation databases (Pozsgai et al., 2021). The full details of the search and its results are described in Seguin et al. (2024). This search contained three sub-queries and covered English language entries published from 2018 to 2022. A total of 108,064 publications were retrieved across the interrogated databases.

Presently, we describe the sub-query used to identify publications used for one of these reviews, which was developed with the aim of identifying applications of spatially explicit indicators to the study of ecosystem condition and characteristics of the datasets used to develop these indicators. The review took into account the recommendations of the PRISMA statement, which provides guidelines for the reporting of systematic reviews (Moher et al., 2009).

The search logic used is visualized in Fig. 1. We filtered the central corpus using combinations of four terms with Boolean and proximity operators: an ‘ecosystem’ term (i.e. ‘ecosystem’, ‘ecological’, ‘habitat’, ‘environment’, and biological’, as well as the names of individual ecosystem types according to the Mapping and Assessment of Ecosystems and their Services (MAES) typology (Maes et al., 2013)), a ‘condition’ term (i.e. ‘condition’, ‘quality’, ‘function’, ‘state’, ‘health’), and a ‘quantification’ term (i.e., ‘map’, ‘indicator’, ‘variable’, ‘assessment’). This initial search produced 75,060 publications which were coded with a unique identifier. Following this we additionally filtered papers for a ‘spatially explicit’ term (i.e. ‘spatially explicit’, ‘map’, ‘spatial distribution’, ‘spatial modeling’, ‘spatial variability’, ‘spatial relationship’), which resulted in 5855 unique publications.

The reporting items up to title and abstract screening are shown in Fig. 2. We removed duplicated entries, retractions, and book chapters from the list of publications, and excluded an additional 1164 items published in journals targeting disciplines such as health science and medicine, mechanical and electrical engineering, and sociology, because of a low probability these papers were related to ecosystem condition. As an additional filtering step, we manually screened the full text of 4627 publications, and selected only those including a map (other than a contextual study area map) to filter out studies not using spatially explicit data. We excluded entries for which no full text was available. This first round of screening resulted in 2917 publications to be assessed for relevance.

2.2. AI implementation protocol

We screened publications based on their title and abstract using the

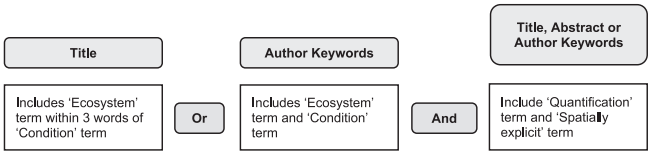


Fig. 1. Visualisation of the search logic used to identify papers in Web of Science and Scopus online databases.

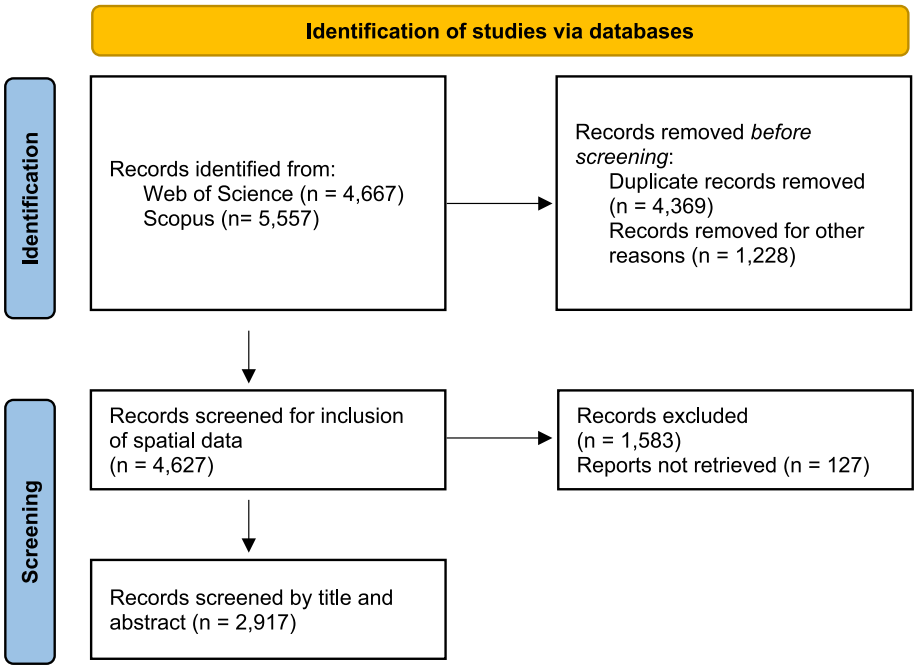


Fig. 2. Status of reporting items for systematic review at time of present study, adapted from the PRISMA statement (adapted from (Page et al., 2021), CC BY 4.0).

GPT-3.5 Turbo LLM through the completions API (OpenAI, 2023a). This approach enabled rapid processing of numerous single-turn queries, essential for efficiently screening a vast array of publications based on their titles and abstracts. Through our queries we asked the model to determine if publications should be selected or rejected for the study based on its title and abstract. We developed code to call the API and process the model responses in R (R Core Team, 2021), which is available at https://github.com/PkdRoche/Reference-Screening-ChatGPT. Furthermore, we configured the queries to specify the 'User' role, instructing the model to assume the perspective of an ecological scientist. This strategy ensured that the model's responses were framed with an understanding of relevant ecological contexts. The code used here can be easily adapted to other studies by updating the system request file holding the prompt text. To manage the API's rate limits and ensure uninterrupted operation, we incorporated a timeout strategy between queries and a backoff function. This function was designed to reattempt a query in the event of errors, such as server timeouts, thus maintaining the consistency and reliability of our data collection process.

2.3. Prompt development and refinement

We developed inclusion criteria against which to screen publications based on the aims of the review, which we translated into the initial prompt used for queries. The prompt specified that publications should be classified according to their relevance to the subject of the review (i.e. whether the publication applied indicators for the study of ecosystem condition) based uniquely on the information (title and abstract) provided. Three options for classification of the publication were offered to the model, either 'selected' or 'rejected' for the study, or 'uncertain'. The 'uncertain' option was included to mitigate against possible incorrect rejections received as a result of forcing a binary response. We assumed any publications classified as 'uncertain' should be taken forward to the full-text eligibility assessment stage of the review. The prompt began by briefly explaining the objective of the task and then described the context in which each option for classification should be chosen. In the interest of cost and efficiency the prompt specified that only the selected classification should be provided in the model response, with no additional text or justification.

We identified terms for an optimal prompt qualitatively through identifying commonalities between incorrectly classified papers, and re-running with incremental changes to the text. Additionally, interrogating the model as to why certain papers might be incorrectly classified and providing abstracts to the model along with the expected response provided insight into trends in responses. We evaluated each iteration on the basis of its rate of error in rejecting suitable publications, and its capacity to discriminate in rejecting irrelevant publications. We used 12 versions of the prompt with several "runs" during development.

We observed that due to the inherent variability of generative AI outputs, GPT-3.5 can produce inconsistent responses to the same query, leading to variability in the classification results when multiple iterations were run. However, responses tended to converge on a most common answer. To avoid relying on erroneous responses, we addressed this variability by running the query for each publication 10 times and taking the most common response as the final decision. For final implementation of screening, the number of repeats was reduced from 10 to 5 as tests showed consistent agreement across 70 % of iterations for the same publication. We re-classified as 'uncertain' any publications for which the classification was not consistent across at least 4 iterations.

GPT-3.5 offers various options of parameters which can be adjusted to shape the model's response. One key parameter with regards to the variability of observed responses is temperature. The temperature parameter controls the determinism of the model, with high values producing more random outputs and low values producing more consistent outputs. This parameter has a default value of 1 but can accept values between 0 and 2. We tested a range of values from 0 to 1 to verify the impact on the total proportion of inconsistent model responses (Fig. 3). Increasing the temperature led to higher variability in the classification result given for a publication, increasing the number of 'uncertain' results. The final prompt was therefore run with a temperature value reduced to 0.2 to minimize this variability.

2.4. Validation

We first compared early iterations of the prompt to classifications made by the authors on a subset of publications pre-selected to cover a range of degrees of relevance for the subject of interest. We then performed validation of the prompt using classifications made by a group of 10 experts. Experts were provided with a list of the titles and abstracts of

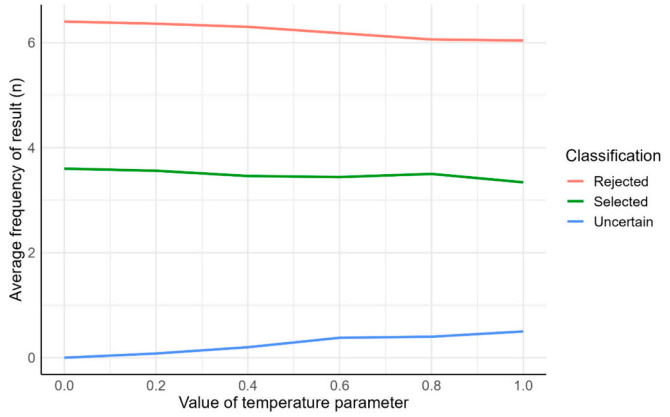


Fig. 3. Effect of adjusting the model temperature parameter on the proportion of uncertain responses.

10 publications, and instructed to respond to the same prompt text that was given to the model. We assessed performance using a first sample of expert classifications ($n = 50$) at an intermediate stage of development, and with a second sample ($n = 50$) to confirm the final version of the prompt. To assess consistency in the human response to the inclusion criteria, each entry in the validation samples was randomly assigned to 2 experts from the project team (Reviewer A and Reviewer B) who independently classified the entries according to the same instructions as given to the model. For comparison with model classifications, we re-classified instances in which expert classifications were conflicting as 'uncertain'. Additionally, we re-coded instances in which only one expert reviewer was uncertain with the second reviewer's classification.

3. Results

3.1. Definition of ecosystem condition in prompt script

The final version of the model prompt is included in Fig. 4. The resulting definition of ecosystem condition for determining selection criteria relied on the use of known synonyms of condition, including repetition of these key terms with a combination of descriptors such as 'ecosystem', 'habitat', and 'environment'. Effective prompts did not require inclusion of all possible combinations of synonyms. The definition of ecosystem condition was not changed in a significant way during prompt development, except for the addition of quality terms in later prompts. Improving prompt efficacy required more detailed description of the criteria for rejection specifying what should not be considered as ecosystem condition. Less effective prompts instructed the model to reject studies focused on 'broader' or 'general' environmental or ecological topics, while the final version detailed specific elements of ecological research to be excluded.

3.2. Performance of GPT-3.5 in classifying publications for review

Table 1 shows the results of comparing classifications made by the model in response to selected versions of the prompt with those made by expert reviewers. For brevity we present here only the most efficient and illustrative versions (v10r3, v11r4, v12r1 and v12r10). The full text of these prompts is included in appendix A.1.

Objective:

For each study with a title and abstract provided, classify them according to the following instructions. I want you to consider exclusively the provided classification criteria and not general patterns.

Classification Criteria:

'Selected': The study must explicitly discuss the creation, testing, or application of assessment tools geared towards evaluating indicators of ecosystem condition, such as ecosystem health, ecosystem state, ecological health, habitat quality, environmental quality or similar concepts. The abstract should explicitly state the use or development of quantitative methods, indicators, proxy, metrics specifically aimed at assessing indicators of ecosystem condition.

'Rejected': The study should be rejected if its primary focus is not about ecosystem condition and address issues such as landscape connectivity, habitat fragmentation, species distribution, conservation efforts, ecological conservation hotspots, pollution, human health, species or group of species or policies without a clear emphasis on the development, validation, indicators, or application of assessment tools for evaluating indicators of ecosystem condition.

'Uncertain': If the classification isn't clear or if both 'selected' and 'rejected' criteria are partially met, classify as 'uncertain'. Classify as 'uncertain' also when it is not clear if assessment tools for ecosystem condition are the main objective of the study or if the study focuses on indicators closely related to ecosystem condition without explicit mention of the term.

Instructions:

Return your classification using exclusively these terms: 'Selected', 'Rejected', or 'Uncertain'.

Fig. 4. Final version of the prompt used in queries.

Table 1
Performance of selected prompt iterations.

Prompt version	Proportion of correct selections	Proportion of correct rejections	Proportion of incorrect selections	Proportion of incorrect rejections
v10r3	0.67	0.90	0.10	0.33
v11r4	0.83	0.86	0.14	0.17
v12r1	0.42	0.90	0.10	0.58
v12r10	1.00	0.59	0.34	0.00

The first version we considered efficient enough to be compared to expert classifications, v10r3, was very effective in rejecting 90 % of papers which were also rejected by expert reviewers but showed only moderate accuracy in correctly selecting papers (67 %). Following intermediate testing of increasingly specific prompts, v11r4 was written to maintain the balanced approach taken in v10r3 to fully describe all classification options but with an increased level of detail and clarity of language used. The results of using this prompt showed an improved accuracy in correct selection (83 %) at the expense of a slightly higher level of false positives. The number of 'uncertain' classifications (2 publications) remained consistent across runs.

Multiple tests were carried out to explore the potential of improving the performance from v11r4, of which we present here two key examples as an illustration of the trade-off between accuracy and discriminatory power of the approach. Prompt v12r1 included fewer specific criteria for selection and rejection and resulted in a higher proportion of false negatives and the loss of more relevant papers. However, this version improved the proportion of correctly rejected publications compared to v11r4. In contrast, for v12r10 the opposite approach was taken to adjust the prompt by including highly specific and restrictive criteria. The model did well in selecting appropriate publications (100 %) but counter-intuitively was less restrictive, with the responses to this prompt selecting greater number of publications in total including a higher proportion of irrelevant publications.

3.3. Comparison with expert reviewers

We performed the Pearson's Chi-square test to quantify the degree of divergence between expected frequencies of classifications as provided by experts and observed frequencies produced by the model classifications (Table 2). We observed that the degree of divergence between reviewer responses was similar to the divergence between the reviewers and the model responses following classification of the first sample of publications tested (Sample 1). However following classification of Sample 2, the Chi-square test revealed no significant difference between the classifications made by Reviewer A and GPT-3.5 (Chi-square value 4.342 with 2 df, $p = 0.114$), however a significant difference was observed between the two sets of reviewers and between Reviewer B and GPT-3.5.

Human reviewers were consistently more likely to assign an 'uncertain' classification than the model, but not necessarily to the same publications. Compared to v10r3, use of the improved prompt did not have an effect on the total the number of 'uncertain' classifications assigned by the human reviewers.

Table 2
Results of Pearson's chi-square test between classifications of publication samples.

	X^2		
Publication sample	Reviewer A vs B	Reviewer A vs GPT-3.5	Reviewer B vs GPT-3.5
Sample 1	26.7 (df = 4; p = 0.000)	24.3 (df = 4, p = 0.000)	18.8 (df = 4; p = 0.000)
Sample 2	14.7 (df = 4; p = 0.000)	4.3 (df = 2, p = 0.114)	19.1 (df = 2; p = 0.000)

4. Discussion

4.1. Perspectives for defining ecosystem condition

The successive iteration of prompting scripts provided insight into how to better conceptualize ecosystem condition for the purpose of identifying relevant literature in the ecological domain. It follows that based on the terms used to appropriately select and reject items of known relevance, ecosystem condition should indeed be understood as a multidimensional concept that encompasses various attributes such as quality, state, and health. The relative ambivalence of the model towards other parts of the terminology emphasizes the importance of including these multiple attributes over broad descriptors like "ecosystem" or "environment" when defining condition. A holistic approach that highlights the quality aspects ensures a more comprehensive and consistent understanding of ecosystem condition across different ecosystem types, helping to simplify the concept while at the same time maintaining its depth.

The need to closely define what should not be considered as ecosystem condition through effective criteria for rejection however shows that the ecosystem condition concept should not be overgeneralized. The prompts were more efficient when specifically instructing the model to reject abstracts relying on strictly ecological terms such as species distribution, conservation status, connectivity and fragmentation. Such terms are associated with studies focusing on narrower concepts that only partially represent the range of relevant ecosystem characteristics, rather than the comprehensive quantification of condition through an appropriate selection of indicators. Other terms associated with human perspectives, such as human health, pollution or policies, should also be avoided when defining ecosystem condition. It is clearly important to understand the benefits to society of ecosystems in good condition. However, excluding these elements was necessary to center the conceptualization of ecosystem condition on the quantification of an ecosystem's abiotic and biotic components and landscape properties (Czúcz et al., 2021).

4.2. Accuracy and efficiency of approach

An optimal systematic review necessitates a high level of sensitivity to detect all pertinent papers, coupled with a robust specificity to ensure that the examination of full-text publications is focused exclusively on relevant papers. Our approach, utilizing GPT-3.5 for classifying publications, has demonstrated a notable efficacy in achieving this balance. Specifically, it has been effective in reducing the number of relevant papers incorrectly dismissed, while enhancing the identification of pertinent papers. The empirical results from our tests indicate that, in certain cases, GPT-3.5 aligns closely with human reviewers in terms of classification accuracy. For instance, in Sample 2, the Chi-square test suggested a comparable level of accuracy between Reviewer A and GPT-3.5. However, this was not consistently observed across all comparisons.

Overall, while GPT-3.5 does not achieve perfect accuracy, its performance in classifying publications for systematic reviews enhances efficiency when compared to the traditional review process. It is important to recognize that, akin to traditional methods which often entail a degree of error in screening (as noted by Bannach-Brown et al., 2019 and Wang et al., 2020), GPT-3.5's application is not devoid of inaccuracies. Nevertheless, its utility in streamlining the review process is evident, particularly in contexts where its classification decisions closely align with those of human reviewers. Performance of the model was improved when using a prompt that emphasizes and repeats key terms. The specificity and clarity of the classification criteria can significantly impact the accuracy of abstract classification, and we observed that it was necessary to repeat key terms within all potential

classification options (selected, rejected and uncertain). It appears that a lack of detailed specificity in the terms may lead to the model's interpretation of the criteria being too restrictive, leading to poor performance in correct selection. This is particularly relevant for the topic in question which requires the consideration of multiple potential synonyms. Striking the balance between inclusivity and exclusivity is however a key factor in model iteration.

Expert reviewers reported taking between 11 and 20 min to classify a sample of 10 publications. When validated, the model required approximately 10 % of the minimum time taken by reviewers to classify, including the additional buffer of the timeout strategy. At the quantity of papers included in this review, this presents a significant saving in researcher resources. Furthermore, participation in an extended repetitive task has been shown to result in decreased accuracy due to increasing levels of fatigue (Gonzalez et al., 2011), and we therefore mitigate the risk of error incorporated by reviewer observed with high volumes of literature for screening (Sampson et al., 2011). We therefore mitigate the risk of decreased accuracy which is observed during extended repetitive tasks. Whilst automation does entail costs in researcher time during prompt development and in API usage for querying the model, this is to an extent mitigated by removing the need to train and guide additional reviewers to a sufficient understanding of the topic.

Based on the analysis of decision mismatches between the model's decisions and those of human reviewers, several issues can be identified that affect the interpretation of abstract content relative to the prompt or instructions. Some studies included in the testing focus on concepts related to ecosystem condition (e.g., ecosystem services, habitat suitability) but do not explicitly use terms associated with the criteria defining ecosystem condition. In some cases, the abstract may not clearly articulate the methodology or its direct application to ecosystem condition assessment, leading to different possible interpretations. Studies that indirectly assess ecosystem condition (e.g., through land degradation or species distribution) may be classified differently depending on how strictly the criteria are applied. Research combining multiple aspects (e.g., ecological and human health impacts) may be classified inconsistently based on which aspect is perceived as dominant. Some studies fall into grey areas where they partially meet the criteria, leading to uncertain classifications or disagreements between reviewers. This could also be linked to the strictness in the application of criteria by human reviewers, who may be more prone to a global interpretation that is difficult for a language model like GPT-3.5 to achieve. It should be noted that when scrutinising the abstracts with conflicting decisions, the human reviewers' decision does not appear to be more reliable than that of the model. It is clearly a question of interpretation of the text content with regards to instructions. As an example of a contradictory decision, a study abstract discusses the detection of land degradation trends using remote sensing indicators like Leaf Area Index (LAI), albedo, and evapotranspiration in north-eastern Brazil. Detecting land degradation could be considered assessing ecosystem condition, which may explain why it was proposed as 'selected' by the model. However, the human reviewers, who proposed an 'uncertain' classification relied on the fact that the methodology was not clear or proven enough to definitively classify it as an ecosystem condition assessment tool.

4.3. Perspectives and limitations of using generative AI in systematic reviews

We note some key benefits compared to traditional methods employed for evidence synthesis. The systematic review process aims to increase objectivity in a review though minimising bias and interpretation from reviewers (Haddaway et al., 2015). However, previous research has shown that responses to inclusion criteria can vary between individuals and indeed evolve over the course of the screening exercise, and that individual perspectives on the quality of abstracts can lead to different decisions (Belur et al., 2021). The observed differences between our two sets of reviewers shows that subjectivity can indeed persist in the application of inclusion criteria. Automated approaches can involve a consistent interpretation of subject matter, which presents an advantage for analysing complex concepts such as ecosystem condition, where the individual understanding of expert reviewers can be placed along a continuity of definitions (Roche and Campagne, 2017).

Additionally, the iterative process of testing different prompts for use in the model queries has benefits for the development of effective inclusion criteria for a systematic review. The use of ChatGPT has been proposed for developing search strategies and inclusion criteria (Atkinson, 2023), and we found that the model's responses enabled us to clarify terminology and identify which specific phrases could lead to confusion for reviewers. Analysis of the language used in prompts offers opportunities to explore the evolution of the conceptualisation of ecosystem condition over time, which could potentially lead to the use of more understandable and continually applicable definitions. Future studies would benefit from taking the approach of the present study further and systematically assessing the impacts of small, individual additions to the definition of the ecosystem condition concept within the prompt text. In the continually evolving field of ecosystem condition and ecosystem services, there is great potential for such work to aid in improving consistency in the depiction of complex conceptual mosaics. However, it is important for researchers to bear in mind that models like GPT-3.5 are not specifically trained on scientific literature, nor do they necessarily have access to the latest published research, which may limit their capacity for analysing scientific texts. As it stands, the extraction of information on ecosystem condition from prompt development is limited by the difficulty in explaining the causality of certain changes. An AI trained specifically on scientific literature may be more effective at integrating the academic context of ecological concepts, and therefore increase the robustness of decisions.

The present analysis took place in the context of continued, rapid development of generative AI models, whose capacities and performance are constantly evolving. Using pre-trained models through an API presents new opportunities for a range of researchers as no specific capacities with using machine learning algorithms are required in order to implement the methodology. Whilst GPT-3.5's limit of 32,768 tokens restricted our queries to inclusion of the prompt text, and abstract and title of a publication, new models allow for queries of upwards of 100,000 tokens or more which permits the analysis of larger texts and opens the potential for using the GPT models to assist with the full-screening and data extraction stages of a systematic review. The increased cost of higher-performance models remains restrictive to many research contexts, with API usage of GPT-4 costing 30 times more than GPT-3.5 Turbo in 2023 (OpenAI, 2023b). This also limits access to more sophisticated LLM that could potentially better consider nuances in the queries and in the examined text. However, the creation of linked tools such as the ability to query PDF files using the same model raises the potential for their use in multiple review stages (Atkinson, 2023). The offer of LLM is rapidly increasing even including open models that could be run locally without any costs such as LLaMa2, PHI2, Mistral 7B or Mixtral. As the capability of generative AI continues to develop, the systematic review process must however continue to lean on domain-specific expertise to maintain a robust approach.

Whilst the present study included generative AI in a review on ecosystem condition indicators, the methodology can be applied to a broader field of research and would be particularly useful for ecological and environmental questions bringing together diverse perspectives from different fields. However, from our study we identify several steps that should be considered before implementation. Firstly, it should not be assumed that generative AI can replace human researchers at all stages of the review. The potential of LLMs to 'hallucinate' and give incorrect information persists, demanding effective validation strategies, based on domain-specific expertise to avoid the inclusion of erroneous information (Smit et al., 2021). The complete review process, of which only one step is described here, utilised the domain-specific

expertise of researchers, who were able to inform sufficiently descriptive prompts and filter any errors in the previous and subsequent steps respectively, guarding against the impact of hallucinations. Secondly, researchers must avoid the common pitfall of overgeneralizing (Zamfirescu-Pereira et al., 2023) and take time to identify and formulate the key details of the subject with a sufficient level of detail for a model which may not have been trained on the necessary contextual information. Finally, future implementation of this approach for screening would benefit from further in-depth testing of the effect of adjusting the model parameters. Our selection of the temperature parameter value is based on the aim of reducing the uncertainty and randomness of model responses as far as possible, whilst at the same time allowing some possibility for the model to apply the 'uncertain' classification. This is based on the assumption that more deterministic responses would be optimal for recreating the review process. Nevertheless, it could be the case that adjusting this parameter to encourage more stochastic or random responses could be useful in providing a broader range of perspectives through the variability of responses.

5. Conclusion

In summary, we present an approach for applying recent developments in generative AI to streamline the systematic review process whilst maintaining accuracy suitable for a rigorous and repeatable review. This increased efficiency opens the door for future work to derive a better understanding of the characteristics of indicators of ecosystem condition and the datasets used to calculate these indicators, through enabling the completion of a review that employs full consideration of the variety of terms employed to describe condition. In our study, we outlined the terms associated with ecosystem condition to allow a LLM to correctly separate studies presenting results on ecosystem condition indicators from other studies with different or broader scopes. Our study made clear that we cannot simply expect a LLM to have the adequate interpretation of a complex term based on its training, but there is a need to clearly express what does and what does not represent ecosystem condition as a multidimensional concept. Our experience is that the GPT-3.5 model, while not completely infallible, significantly outperformed traditional review processes in terms of speed and efficiency. This enhancement is particularly valuable given the growing volume of literature that researchers must sift through in systematic reviews. Our approach underscores the potential of AI in reducing the time and resource burden on researchers, and additionally in maintaining consistency in interpretating inclusion criteria. Moreover, our study highlighted the importance of precision in prompt design and the model's query parameters. We found that the specificity and repetition of key terms within the model's prompts greatly influenced its ability to accurately classify literature, underscoring the importance of detailed and well-considered prompt construction. The continued development and increasing availability of AI tools presents a great potential to improve evidence synthesis through gains in efficiency and the capacity to cover large volumes of data, and with sufficient validation these tools can indeed meet the standards required for systematic reviews.

CRediT authorship contribution statement

Isabel Nicholson Thomas: Writing – original draft, Methodology, Investigation, Conceptualization. **Philip Roche:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Adrienne Grêt-Regamey:** Writing – review & editing, Supervision.

Declaration of competing interest

None.

Data availability

The code used to query the model is available at https://github.com/PkdRoche/Reference-Screening-ChatGPT and data can be accessed via Zenodo at https://doi.org/10.5281/zenodo.12705805

Acknowledgements

We would like to thank our partners on the SELINA Project and colleagues at ETH Zürich who provided valuable input to this study in the form of screening and classifying publications. Additionally, we thank Christian Egger for initial technical advice and Franziska Walther for comments in the early stages of this work. Funded by the European Union under grant agreement No. 101060415, SELINA (Science for Evidence-Based and Sustainable Decisions About Natural Capital).

Appendix A. Appendix

A.1. Full text of selected prompts

Request v10r3

Objective:

For each study with a title and abstract provided, classify them according to the following instructions.

Classification Criteria:

'Selected': The study must explicitly discuss the creation, testing, or application of assessment tools geared towards evaluating indicators of ecosystem condition, such as ecosystem health, ecosystem state, ecological health, habitat quality, environmental quality or similar concepts. The abstract should explicitly state the use or development of quantitative methods, indicators, proxy, metrics specifically aimed at assessing indicators of ecosystem condition.

'Rejected': The study should be rejected if its primary focus is not about ecosystem condition and address issues such as landscape connectivity, habitat fragmentation, species distribution, conservation efforts, ecological conservation hotspots, pollution, human health, species or group of species or policies without a clear emphasis on the development, validation, or application of assessment tools for evaluating indicators of ecosystem condition.

'Uncertain': If the classification isn't clear or if both 'selected' and 'rejected' criteria are partially met, classify as 'uncertain'. Classify as 'uncertain' also when it is not clear if assessment tools for ecosystem condition are the main objective of the study or if the study focuses on indicators closely related to ecosystem condition without explicit mention of the term.

Instructions:

Returns your classification using exclusively those terms: 'Selected', 'Rejected', or 'Uncertain'.

Request v11r4

Objective:

For each study with a title and abstract provided, classify them according to the following instructions. I want you to consider exclusively the provided classification criteria and not general patterns.

Classification Criteria:

'Selected': The study must explicitly discuss the creation, testing, or application of assessment tools geared towards evaluating indicators of ecosystem condition, such as ecosystem health, ecosystem state, ecological health, habitat quality, environmental quality or similar concepts. The abstract should explicitly state the use or development of quantitative methods, indicators, proxy, metrics specifically aimed at assessing indicators of ecosystem condition.

'Rejected': The study should be rejected if its primary focus is not about ecosystem condition and address issues such as landscape connectivity, habitat fragmentation, species distribution, conservation efforts, ecological conservation hotspots, pollution, human health, species or group of species or policies without a clear emphasis on the

development, validation, indicators or application of assessment tools for evaluating indicators of ecosystem condition.

'Uncertain': If the classification isn't clear or if both 'selected' and 'rejected' criteria are partially met, classify as 'uncertain'. Classify as 'uncertain' also when it is not clear if assessment tools for ecosystem condition are the main objective of the study or if the study focuses on indicators closely related to ecosystem condition without explicit mention of the term.

Instructions:

Returns your classification using exclusively those terms: 'Selected', 'Rejected', or 'Uncertain'.

Request v12r1

Objective:

You are tasked with classifying research papers based on their titles and abstracts. Your classification should strictly follow the provided criteria. Do not rely on general patterns or external knowledge.

Classification Criteria:

- 'Selected': A study will fall into this category if its title and abstract explicitly discuss the creation, testing, or application of assessment tools designed for indicators of ecosystem condition. This encompasses evaluations of ecosystem health, ecosystem state, ecological health, habitat quality, environmental quality, or related concepts. Emphasis should be on the development or use of quantitative methods, proxies, or metrics specifically aimed at these indicators.
- 'Rejected': Classify a study as 'Rejected' if its primary emphasis, as indicated in the title and abstract, is on topics like landscape connectivity, habitat fragmentation, species distribution, conservation initiatives, ecological conservation hotspots, pollution, human health, specific species, or policy matters. These studies should not have a clear focus on the development, validation, or application of tools assessing indicators of ecosystem condition.
- 'Uncertain': Use this classification if the title and abstract do not provide a clear indication of whether the study meets the 'Selected' or 'Rejected' criteria. If there's ambiguity regarding whether the assessment tools for ecosystem condition are the core focus or if the study delves into indicators related to ecosystem condition without explicit usage of the term, classify it as 'Uncertain'.

Instructions: Return your classification for each study using only one of the following terms: 'Selected', 'Rejected', or 'Uncertain'.

Request v12r10

Objective: Your sole task is to classify each study based on the content provided in its title and abstract. Disregard any general knowledge or patterns outside the information given in these texts and this prompt.

Classification Criteria:

'Selected': A study should be classified as 'Selected' only if:

It clearly discusses the creation, testing, or direct application of tools or techniques whose primary design is for explicitly evaluating indicators of ecosystem condition.

Specific terms related to ecosystem condition — such as ecosystem health, conservation status, ecosystem state, ecological health, habitat quality, or environmental quality—are present. Indicators of impact can be accepted if associated with previously cited terms.

The focus is not merely on data integration, modeling, or general ecosystem assessment but specifically there's a clear mention of the use or development of methods, indicators, proxies, or metrics with the primary aim of assessing ecosystem condition.

'Rejected': A study should be classified as 'Rejected' if:

the title and abstract focuses on data integration, landscape connectivity, habitat fragmentation, species distribution, conservation actions, ecological conservation hotspots, pollution, human health, specific species, functional traits or general ecological or environmental considerations without direct relation to ecosystem condition indicators.

There's an absence of explicit emphasis on indicators, tools or methods specifically designed to assess ecosystem condition.

The words 'ecosystem', 'habitat', 'condition' could be present but, from the context, are not used in direct relation to ecosystem condition.

'Uncertain': If the title and abstract:

Do not provide clear-cut evidence to be classified as 'Selected' or 'Rejected'.

Instructions: Base your decision strictly on the criteria above. Returns your classification using exclusively only one of those terms: 'Selected', 'Rejected', or 'Uncertain'.

Harnessing Artificial Intelligence for efficient systematic reviews: A case study in ecosystem condition indicators.

References

Atkinson, C.F., 2023. Cheap, quick, and rigorous: artificial intelligence and the systematic literature review. Soc. Sci. Comput. Rev. https://doi.org/10.1177/08944393231196281.

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A.S.C., Ananiadou, S., Liao, J., Macleod, M.R., 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst. Rev. 8, 23. https://doi.org/10.1186/s13643-019-0942-7.

Belur, J., Tompson, L., Thornton, A., Simon, M., 2021. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociol. Methods Res. 50, 837–865. https://doi.org/10.1177/0049124118799372.

Blaizot, A., Veettil, S.K., Saidoung, P., Moreno-Garcia, C.F., Wiratunga, N., Aceves-Martins, M., Lai, N.M., Chaiyakunapruk, N., 2022. Using artificial intelligence methods for systematic review in health sciences: a systematic review. Res. Synth. Methods 13, 353–362. https://doi.org/10.1002/jrsm.1553.

Bornmann, L., Haunschild, R., Mutz, R., 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit. Soc. Sci. Commun. 8, 224. https://doi.org/10.1057/s41599-021-00903-w.

Collaboration for Environmental Evidence (CEE), 2013. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2.

Czúcz, B., Keith, H., Driver, A., Jackson, B., Nicholson, E., Maes, J., 2021. A common typology for ecosystem characteristics and ecosystem condition variables. One Ecosyst. 6, e58218 https://doi.org/10.3897/oneeco.6.e58218.

Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D.L., Tummers, L., Teijema, J.J., van de Schoot, R., 2023. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the average time to discover relevant records. Syst. Rev. 12, 100. https://doi.org/10.1186/s13643-023-02257-7.

Floridi, L., Chiriatti, M., 2020. GPT-3: its nature, scope, limits, and consequences. Minds Mach. 681–694.

Gates, A., Gates, M., Sebastianski, M., Guitard, S., Elliott, S.A., Hartling, L., 2020. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. BMC Med. Res. Methodol. 20, 139. https://doi.org/10.1186/s12874-020-01031-w.

Gerstner, K., Moreno-Mateos, D., Gurevitch, J., Beckmann, M., Kambach, S., Jones, H.P., Seppelt, R., 2017. Will your paper be used in ameta-analysis?Make the reach of your research broader and longer lasting. Methods Ecol. Evol. 777–784. https://doi.org/10.1111/2041-210X.12758.

Gonzalez, C., Best, B., Healy, A.F., Kole, J.A., Bourne, L.E., 2011. A cognitive modeling account of simultaneous learning and fatigue effects. Cogn. Syst. Res. 12, 19–32. https://doi.org/10.1016/j.cogsys.2010.06.004.

Grant, M.J., Booth, A., 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info. Libr. J. 26, 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.

Haddaway, N.R., Woodcock, P., Macura, B., Collins, A., 2015. Making literature reviews more reliable through application of lessons from systematic reviews: making literature reviews more reliable. Conserv. Biol. 29, 1596–1605. https://doi.org/10.1111/cobi.12541.

Haddaway, N.R., Macura, B., Whaley, P., Pullin, A.S., 2018. ROSES RepOrting standards for systematic evidence syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environ. Evid. 7, 7. https://doi.org/10.1186/s13750-018-0121-7.

Khalil, H., Ameen, D., Zarnegar, A., 2022. Tools to support the automation of systematic reviews: a scoping review. J. Clin. Epidemiol. 144, 22–42. https://doi.org/10.1016/j.jclinepi.2021.12.005.

Maes, J., Teller, A., Erhard, M., Liquete, C., 2013. Mapping and assessment of ecosystems and their services :an analytical framework for ecosystem assessments under action 5 of the EU biodiversity strategy to 2020 : discussion paper – final, April 2013. Publications Office, LU.

Maes, J., Driver, A., Czúcz, B., Keith, H., Jackson, B., Nicholson, E., Dasoo, M., 2020. A review of ecosystem condition accounts: lessons learned and options for further development. One Ecosyst. 5, e53485 https://doi.org/10.3897/oneeco.5.e53485.

Mohamed Shaffril, H.A., Samsuddin, S.F., Abu Samah, A., 2021. The ABC of systematic literature review: the basic methodological guidance for beginners. Qual. Quant. 55, 1319–1346. https://doi.org/10.1007/s11135-020-01059-6.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., For the PRISMA Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339, b2535. https://doi.org/10.1136/bmj.b2535.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., PRISMA-P Group, 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst. Rev. 4, 1. https://doi.org/10.1186/2046-4053-4-1.

Olander, L., Polasky, S., Kagan, J.S., Johnston, R.J., Wainger, L., Saah, D., Maguire, L., Boyd, J., Yoskowitz, D., 2017. So you want your research to be relevant? Building the bridge between ecosystem services research and practice. Ecosyst. Serv. 26, 170–182. https://doi.org/10.1016/j.ecoser.2017.06.003.

OpenAI, 2023a. API Reference. URL. https://platform.openai.com/docs/api-reference/ (accessed 12.4.23).

OpenAI, 2023b. Pricing. URL. https://openai.com/pricing (accessed 11.10.23).

Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., McKenzie, J.E., 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ n160. https://doi.org/10.1136/bmj.n160.

Pozsgai, G., Lövei, G.L., Vasseur, L., Gurr, G., Batáry, P., Korponai, J., Littlewood, N.A., Liu, J., Móra, A., Obrycki, J., Reynolds, O., Stockan, J.A., VanVolkenburg, H., Zhang, J., Zhou, W., You, M., 2021. Irreproducibility in searches of scientific literature: a comparative analysis. Ecol. Evol. 11, 14658–14668. https://doi.org/10.1002/ece3.8154.

R Core Team, 2021. R: A language and environment for statistical computing.

Rendon, P., Erhard, M., Maes, J., Burkhard, B., 2019. Analysis of trends in mapping and assessment of ecosystem condition in Europe. Ecosyst. People 15, 156–172. https://doi.org/10.1080/26395916.2019.1609581.

Roche, P.K., Campagne, C.S., 2017. From ecosystem integrity to ecosystem condition: a continuity of concepts supporting different aspects of ecosystem sustainability. Curr. Opin. Environ. Sustain. 29, 63–68. https://doi.org/10.1016/j.cosust.2017.12.009.

Sampson, M., Tetzlaff, J., Urquhart, C., 2011. Precision of healthcare systematic review searches in a cross-sectional sample. Res. Synth. Methods 2, 119–125. https://doi.org/10.1002/jrsm.42.

Seguin, J., Lange, S., Barton, D., Grêt-Regamey, A., Immerzeel, B., Rendón, P., Roche, P., Nicholson Thomas, I., 2024. SELINA report 02: development of the SELINA super-query. One Ecosyst. (Manuscript in preparation).

Smit, K.P., Bernard, A.T.F., Lombard, A.T., Sink, K.J., 2021. Assessing marine ecosystem condition: a review to support indicator choice and framework development. Ecol. Indic. 121, 107148 https://doi.org/10.1016/j.ecolind.2020.107148.

Soubry, I., Doan, T., Chu, T., Guo, X., 2021. A systematic review on the integration of remote sensing and GIS to forest and grassland ecosystem health attributes, indicators, and measures. Remote Sens. (Basel) 13, 3262. https://doi.org/10.3390/rs13163262.

UNCEEA, 2021. System of Environmental-Economic Accounting - Ecosystem Accounting, Background Document Submitted to United Nations Statistical Commission, February 2021. United Nations, New York.

Vallecillo, et al., 2022. EU-Wide Methodology to Map and Assess Ecosystem Condition: Towards a Common Approach Consistent with a Global Statistical Standard. Publications Office, LU.

van Dinter, R., Tekinerdogan, B., Catal, C., 2021. Automation of systematic literature reviews: a systematic literature review. Inf. Softw. Technol. 136, 106589 https://doi.org/10.1016/j.infsof.2021.106589.

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., Murad, M.H., 2020. Error rates of human reviewers during abstract screening in systematic reviews. PloS One 15, e0227742. https://doi.org/10.1371/journal.pone.0227742.

Westgate, M.J., Haddaway, N.R., Cheng, S.H., McIntosh, E.J., Marshall, C., Lindenmayer, D.B., 2018. Software support for environmental evidence synthesis. Nat. Ecol. Evol. 2, 588–590. https://doi.org/10.1038/s41559-018-0502-x.

Zamfirescu-Pereira, J.D., Wong, R.Y., Hartmann, B., Yang, Q., 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 437, pp. 1–21. doi:10.1145/3544548.3581388.