Environmental Evidence

SYSTEMATIC REVIEW

Open Access

Response of chlorophyll *a* to total nitrogen and total phosphorus concentrations in lotic ecosystems: a systematic review

Micah G. Bennett[2], Sylvia S. Lee[1*], Kate A. Schofield[1], Caroline E. Ridley[1,5], Benjamin J. Washington[1,3] and David A. Gibbs[1,4]

Abstract

Background: Eutrophication of freshwater ecosystems resulting from nitrogen and phosphorus pollution is a major environmental stressor across the globe. In this systematic review, we compiled and synthesized literature on sestonic and benthic chlorophyll *a* (chl-a) responses to total nitrogen (TN) and total phosphorus (TP) concentrations in the water column in streams and rivers to provide a state-of-the-science summary of nutrient impacts on these endpoints. This review was motivated by the need for comprehensive information on stressor-response relationships for the most common nutrient and biotic response measures used by state-level environmental managers in the United States to assess eutrophication of lotic ecosystems and support environmental decision making.

Methods: Searches for peer-reviewed and non-peer-reviewed articles were conducted using bibliographic databases, specialist websites, and search engines. These returns were supplemented with citation mapping and requests for material from experts. Articles were screened for relevance using pre-determined eligibility criteria, and risk of bias was evaluated for each included article based on study type-specific criteria. Narrative summaries and meta-analysis were used to evaluate four primary stressor-response relationships: TN-benthic chl-a, TP-benthic chl-a, TN-sestonic chl-a, and TP-sestonic chl-a. Potential effects of modifying factors and study validity on review conclusions were assessed via sensitivity and sub-group analysis and meta-regression.

Results: Meta-analysis of 105 articles, representing 439 cause-effect pairs, showed that mean effect sizes of both benthic and sestonic chl-a responses to TN and TP were positive. Of the four stressor-response relationships examined, TP-sestonic chl-a had the most positive relationship, followed by TN-benthic chl-a, TN-sestonic chl-a, and TP-benthic chl-a. For individual U.S. states, mean effect sizes for the four stressor-response relationships were mostly positive, with a few exceptions. Chlorophyll measurement method had a moderately significant influence on mean effect size for TP-sestonic chl-a, with chl-a responding more strongly to TP if fluorometry versus spectrophotometry was used. Year of publication had a significant negative effect on mean effect size, as did mean nutrient concentration for both sestonic chl-a nutrient relationships. When the same study measured both TN and TP, chl-a tended to respond similarly to both nutrients. Sensitivity analysis indicated that conclusions are robust to studies with high risk of bias.

*Correspondence: lee.sylvia@epa.gov
[1] Present Address: Center for Public Health and Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, 1200 Pennsylvania Ave. NW (8623-R), Washington, DC 20460, USA
Full list of author information is available at the end of the article

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Conclusions: This systematic review confirms that nutrients consistently impact primary producer biomass in streams and rivers worldwide. It builds on previous literature syntheses evaluating chl-a responses to nutrient concentrations and confirms that benthic and sestonic chl-a respond positively to nutrients across a range of stream and river conditions, but also points to limits on these relationships (e.g., potential saturation at high nutrient concentrations). Lack of consistent reporting of contextual data limited our ability to examine how moderating factors influenced these stressor-response relationships. Overall, we provide nutrient managers responsible for protecting the quality of lotic ecosystems with a comprehensive evidence base for chl-a responses to TN and TP concentrations in the water column.

Keywords: Nutrients, Pollution, Water quality, Stressor-response, Stream, River, Primary production, Eutrophication

Background

Nutrient pollution by nitrogen (N) and phosphorus (P)—defined here as nutrient concentrations higher than background or natural levels—is a major stressor of freshwater ecosystems, both globally and across the United States (U.S.) [1–6]. Nutrient pollution and resulting stressors (e.g., eutrophication, oxygen depletion) degrade ecosystem services worth more than $2.2 billion annually in the U.S. alone [7]. Despite recognition of nutrient pollution issues by scientists and stakeholders, rigorous synthesis of scientific evidence is still needed to inform nutrient-related management decisions, especially in streams and rivers [8].

Nutrient stressor-response relationships are complicated by multiple interacting environmental factors, complex and indirect causal pathways involving diverse biotic assemblages and food web compartments, legacy (historic) nutrient sources such as agricultural sediments, and the naturally high spatiotemporal variability of lotic ecosystems [9, 10]. The complexity of stressor-response relationships and natural variability of ecosystems may result in inconsistent findings in the literature. For example, chlorophyll *a* (chl-a) is a widely used measure of eutrophication, but the response of chl-a to nutrients often depends on environmental context. The ability to place individual study results in the context of a comprehensive evidence base and increasing the accessibility of evidence may support development of environmental decisions, such as state numeric nutrient criteria. In the U.S., important water quality decisions are made at the state level, and this review aimed to increase access to and understanding of the state-relevant evidence available in the literature.

Nutrient increases can affect biota in streams and rivers through a variety of biological, chemical, and physical mechanisms [11]. Biota integrate nutrient impacts over time and may represent the ecological condition of a system more accurately than a snapshot of water chemistry measurements [12–15]. For example, primary producer biomass integrates water quality conditions over time periods of several days to months and thus provides information about conditions affecting aquatic life over longer time scales. Environmental managers often use this biological information to evaluate impacts of chronic pollution (e.g., [16]), but this can be complicated because high spatiotemporal variability, across numerous environmental factors, can mask links between nutrients and biota [17]. A synthesis of nutrient stressor-biological response relationships and how these relationships are modified by other factors could aid environmental managers in both identifying impacted systems based on biota [18] and setting targets for maintaining or reviving healthy ecosystems [19–21].

Primary producers are one of the first ecological indicators that respond to nutrient pollution in lotic systems [22]. Increases in primary producer biomass, particularly algal biomass, are also associated with many of the negative human health and ecological consequences of eutrophication, such as hypoxia, reduced drinking water quality [23, 24] and altered species composition [6]. Chl-a is a photosynthetic pigment used to measure primary producer or algal biomass [25]. In streams and rivers, researchers may sample benthic chl-a from bottom substrates or sestonic chl-a from the water column [25, 26] to determine chl-a concentrations.

In this systematic review, we compiled and synthesized literature on sestonic and benthic chl-a responses to total nitrogen (TN) and total phosphorus (TP) concentrations in the water column in streams and rivers to provide a state-of-the-science evidence base summarizing nutrient impacts on these endpoints. TN and TP were selected for both practical and ecological reasons. This review was motivated by a need for comprehensive information on stressor-response relationships to aid water quality scientists at the U.S. Environmental Protection Agency (U.S. EPA) and state environmental agencies in better understanding the effects of nutrient pollution, and TN and TP are the most common nutrient measures used by environmental managers in the United States to assess eutrophication of lotic ecosystems [19, 21, 27, 28]. Although dissolved nutrient forms may be more available for immediate uptake by biota, total nutrient forms are

often more highly correlated with chl-a [22]. Dissolved forms may undergo rapid uptake and release by primary producers, such that concentrations of dissolved nutrients in the water column may not represent true nutrient availability [29, 30]. In contrast, total nutrient forms may best represent trophic state and nutrient limitation in most lotic ecosystems because TN and TP account for N and P held within algae and sediment particles and thus represent integrated measures of biologically available nutrients [17, 30, 31].

Objective of the review

The primary question addressed by this review is: "*What is the response of chl-a to TN and TP concentrations in lotic ecosystems?*" The nutrient stressor (TN or TP concentration in the water column) and biotic response (chl-a) were chosen based on measures commonly used by U.S. state agencies to evaluate and make regulatory decisions about impairment of lotic ecosystems due to eutrophication. In several meetings held in 2016–2017, federal and state agency end users of the systematic review helped refine its scope, specific review questions and objectives, and modifying factors of interest. Based on these meetings, the primary review question consists of the following components:

Population: Lotic fresh waters, or mesocosms that mimic these systems, in any geographic location in the world.

Exposure: Concentration of TN or TP in the water column. We define TN as the sum of ammonia N, nitrate N, nitrite N, and organic nitrogen forms; we define TP as the sum of dissolved and particulate phosphorus forms.

Comparator: Control group (no added TN or TP, or low exposure to TN or TP) (for experimental studies); comparison to lower or higher TN or TP concentrations across a gradient (for observational studies).

Outcome: Chl-a concentration (sestonic, benthic, or other).

The secondary question addressed by this review is: "*How are the relationships identified in the primary question affected by other factors?*" An initial list of potential modifying factors was developed in [32] and a final list is provided below (see "Methods"—"Potential effect modifiers and reasons for heterogeneity").

Methods

This review was conducted according to an a priori protocol [32] following the Guidelines for Systematic Review of the Collaboration for Environmental Evidence (v. 4.2) [33]. This review conforms to ROSES reporting standards [34] (Additional file 1). Deviations from the protocol were made due to specific situations encountered during the review and are detailed in the next section.

Deviations from the protocol

Search for articles

- As part of a larger project, we searched for articles from databases for two other ecological endpoints (macroinvertebrates and diatoms). Articles returned in these searches were screened together with chl-a returns (see below). These additional searches are documented in Additional file 2 and followed the original chl-a protocol [32] but with endpoint-appropriate keywords replacing chl-a keywords.
- We were unable to search two websites on our original list for state-level environmental agencies (Kentucky Department of Environmental Protection, Puerto Rico Environmental Quality Board) due to broken links and lack of search capabilities.
- We conducted an initial title screen within EndNote to remove document types that were clearly ineligible (e.g., Front Matter, Meeting Programs, Abstracts, Books Reviewed). Duplicate entries were identified and removed using EndNote's algorithm followed by manual screening. Remaining articles were imported into the Rayyan software [35] (http://rayyan.qcri.org/) for title/abstract screening, and Rayyan was used to identify additional duplicates.

Article screening and eligibility criteria

- For title/abstract screening consistency, the protocol stated that a maximum of 200 articles would be screened by all reviewers; however, an error in the screening software omitted one of the selected articles from screening and, once discovered, the original process could not be replicated with the same reviewers.
- Articles obtained through snowball searches were screened similarly to database search returns using Rayyan, except for a subset of articles that were not available in Web of Science Core Collection. These articles were screened, first at the title level and then at the abstract level, directly in MS Excel and eligibility decisions were recorded separately. Separate title and abstract level screening were justified because so many of the bibliography items were easily judged ineligible at the title level (e.g., statistics textbooks).
- We screened returns from two additional endpoints together with chl-a returns at the title/abstract level. At the full text level, we removed articles that only

reported effect sizes for macroinvertebrates or diatoms, as documented in the ROSES diagram.
- We did not attempt to digitize or extract raw data to calculate effect sizes ourselves for this review, due to the number of articles and thus level of effort involved. Articles with no effect size were not advanced for validity assessment, full data extraction, or meta-analysis.

Study validity assessment

- We slightly modified the validity assessment questions for observational field studies from those listed in the protocol to distinguish between intra-site visit sample replication (e.g., triplicate water samples estimating nutrient concentrations) and repeated visits to the same site (see Table 2), both of which could reduce risk of bias as they increased.

Potential effect modifiers and reasons for heterogeneity

- Based on evaluation of highly relevant articles and consultation with stakeholders and experts, the modifying factors extracted for this review differed slightly from the list in the protocol.

Data synthesis and presentation

- We did not combine validity assessment ratings to create an overall risk of bias for each dataset, as stated in the protocol, but instead retained individual characteristic ratings in a twofold sensitivity analysis. Additionally, sample size limited our ability to conduct sensitivity analysis to only observational field studies.

Search for articles

Search terms and filters

Bibliographic databases were searched using a combination of terms representing the nutrient stressors (TN or TP), the biological response (chl-a), and habitat- or study-specific terms (e.g., terms associated with types of lotic fresh waters and experimental stream studies) (Additional file 2). For example, the following search string was used for searching Web of Science Core Collection: (benth* OR catchment OR watershed OR stream* OR creek* OR river* OR pool* OR "flood plain" OR floodplain OR riparia* OR ditch* OR lotic OR spring* OR seep* OR riffle* OR freshwater OR freshwaters OR "fresh water" OR brook OR "running water" OR headwater OR tributary OR mesocosm OR flume OR microcosm) AND ("total nitrogen" OR "total N" OR "total phosphorus" OR "total P") AND (Chlorophyll OR "chlorophyll-a" OR "chl-a" OR "chl a"). Databases varied in how they handled search strings, so searches were adapted as needed (Additional file 2). All databases requiring a subscription were accessed through the U.S. EPA Library, except SCOPUS and GreenFILE which were accessed through the University of Iowa. Books, book chapters, pamphlets, and conference abstracts were excluded from consideration unless they were submitted through calls for additional information (see "Supplemental searches" below); although non-digital library resource limitations prevented a full evaluation of these resources, they typically did not report sufficient primary data and results. No language restrictions were applied to database searches, and any other filters used for specific databases (e.g., excluding full text search to limit ineligible literature) are detailed in Additional file 2.

In addition to the main bibliographic database searches for chl-a identified in the original protocol, we conducted database searches for two other ecological endpoints (macroinvertebrates and diatoms) as part of a larger project. Some of these articles contained chl-a responses that were not captured by the original chl-a search. These searches followed the original chl-a protocol [32] but with endpoint-appropriate keywords replacing chl-a keywords. The database searches conducted for macroinvertebrates and diatoms are documented in Additional file 2.

Search limitations

All searches were conducted in English.

Databases

Sixteen bibliographic databases, representing peer-reviewed, non-peer-reviewed, and unpublished material, were searched in late 2016 and early 2017 to obtain articles for the review (Additional file 2). When databases limited the search results that could be viewed or downloaded, results were filtered by year, when possible, to obtain subsets for viewing and download. Due to limitations on batch downloading of citations, three databases (DART, National Technical Reports Library, and OpenGrey) were treated similarly to website searches and the first 50 items returned (for separate TN and TP searches) were examined (see "Specialist websites"; Additional file 3).

Specialist websites

In addition to bibliographic databases, we searched 71 specialist websites in early 2017 for eligible citations. These websites focused on environmental agencies at the federal, state and territory level and environmental non-governmental organizations; the complete list of websites searched is provided in Additional file 3. We were unable to search two websites on our original list for state-level environmental agencies (Kentucky Department of Environmental Protection, Puerto Rico Environmental Quality Board) due to broken links and lack of search capabilities.

For each website (and for the databases DART, National Technical Reports Library, and OpenGrey), the first 50 items returned, sorted by relevance, were examined for each search. For websites without a search function, "publications" sections were examined to find documents. Because many websites do not accept Boolean search strings, separate searches were conducted for TN and TP, and a smaller set of terms were used each of these searches. All website searches are documented in Additional file 3. Although the specialist website list is biased toward western countries, resource constraints limited our ability to search more broadly in non-English speaking countries. Supplemental searches (below) were used to increase capture of eligible articles from other countries.

Search engines

Searches using Google and Google Scholar were conducted in late 2018, and the first 50 search results were examined for relevance. Separate searches were conducted for TN and TP and each endpoint; search terms used for each search were documented (Additional file 3).

Supplemental searches

To supplement these searches, we sent requests for additional resources by email to colleagues with disciplinary knowledge and posted the same requests on ECOLOG-L, Twitter, and ResearchGate. We also requested resources from a list of published experts in the subject area, targeting authors from multiple regions around the globe to fill potential gaps created by U.S./Western bias in academic databases (Additional file 4). "Snowball" (citation mapping) searches were also conducted, using a "test set" of journal articles and reports that we judged as highly relevant given our systematic review questions (Additional file 4). The "test set" was created by searching the authors' personal libraries for highly relevant articles until at least 15 papers for TN-chl-a and 15 for TP-chl-a were obtained; because many articles reported relationships for both TN and TP, the "test set" included a total of 17 articles. References that cited or were cited by these articles were compiled and any novel references not found during database searches were evaluated. Web of Science Core Collection and Google Scholar were used to identify articles that cited or were cited by the highly relevant articles.

Reference management

Articles returned by the search strategy were stored in an EndNote library. In a slight deviation from the protocol, we conducted an initial title screen within EndNote to remove document types that were clearly ineligible (e.g., Front Matter, Meeting Programs, Abstracts, Books Reviewed). Duplicate entries were identified and removed using EndNote's algorithm followed by manual screening. Remaining articles were imported into the Rayyan software [35] (http://rayyan.qcri.org/) for title/abstract screening, and Rayyan was used to identify additional duplicates. Manual screening of remaining articles during full text screening also identified several duplicates (see ROSES diagram).

Assessing search comprehensiveness

Comprehensiveness of the search strategy was assessed by determining whether all articles in the predetermined "test set" of highly relevant articles (Additional file 4) were returned by at least one of the search types in the overall search strategy (i.e., database, website, search engine, or supplemental searches). Sixteen (94%) of the "test set" articles were returned using the search strategy, indicating that the search strategy was comprehensive (Additional file 4). We also conducted "snowball" searches/citation mapping of the "test set" articles to ensure that eligible citations were captured in our search (see "Supplemental searches").

Article screening and study eligibility criteria

Screening process

Before title/abstract screening all articles, consistency in applying eligibility criteria was evaluated on a subset of articles using the kappa statistic (values range from 0 to 1, with 0 indicating no agreement and 1 indicating complete agreement [36]). Five reviewers who would be involved in subsequent screening assessed the same randomly selected set of 199 articles (according to the protocol, a maximum of 200 articles would be screened by all reviewers; however, an error in the screening software omitted one of the selected articles from screening and, once discovered, the original process could not be replicated with the same reviewers). Kappa was calculated in the 'irr' package in R [37, 38], using modifications for more than two raters [39]. Kappa was 0.585, meeting the moderate or high standard set in the protocol for proceeding with screening [32, 33]. Reviewers discussed

screening decisions to resolve disagreements for the initial subset and made notes to inform future decisions. During title/abstract screening of the rest of the retrieved articles, any questions or disagreements about the eligibility of an article that could not be resolved by referring back to notes on the initial subset were discussed.

The eligibility criteria (see below) were used to identify articles that were topically relevant or contained relevant data, based on review of the title and abstract. Abstracts of non-English language articles were translated using Google Translate to assess relevance. Any article for which there was uncertainty about its relevance was included for full text screening. Articles obtained through snowball searches were screened similarly to database search returns using Rayyan, except for a subset of articles that were not available in Web of Science Core Collection. These articles were screened, first at the title level and then at the abstract level, directly in MS Excel and eligibility decisions were recorded separately. Separate title and abstract level screening were justified because so many of the bibliography items were easily judged ineligible at the title level (e.g., statistics textbooks). Articles obtained through website and other supplemental searches were screened during those searches by examining title/abstract/summary and full text when necessary, and information on the number of returns and eligible articles was recorded separately (see ROSES diagram).

We conducted title/abstract screening of three sets of database ecological endpoint searches (chl-a, macroinvertebrates, diatoms) together because of the large number of duplicate articles and because some articles contained chl-a responses that were not captured by the original chl-a search.

Following evaluation of all titles and abstracts, full text screening occurred simultaneously with data extraction and validity assessment. As full text articles were examined for data extraction and validity assessment, any article judged to be ineligible was added to a list along with the justification based on eligibility criteria (Additional file 5). Consistency during full text screening was addressed by frequently convening three reviewers to discuss the strategy and resolve any questions. This was done prior to and during initial phases of data extraction to refine data capture and improve consistency. Two additional reviewers were added partway through the full text screening and data extraction steps of the review; these additional reviewers were added to the frequent discussions noted above to ensure eligibility decisions were consistent. Overall, approximately 11% of the 2253 records that advanced to full text screening and data extraction were screened for eligibility by more than one independent reviewer.

Eligibility criteria

The criteria in Table 1 were used to identify eligible literature. We did not impose any language or date restrictions during screening.

Multiple articles using same underlying data

For cases in which multiple articles used the same or similar underlying data (e.g., a dissertation and one or more published articles from that dissertation), the following criteria (listed in order of priority) were used to select a single source: the article with the more complete underlying data, the version published as a peer-reviewed journal article, or the most recent version. The duplicative article was used to fill in gaps in methodology or contextual information where possible.

Unobtainable articles

Attempts to obtain full text of all articles not excluded during the title/abstract screening process were made using available library resources or by contacting authors. Articles for which full text was not obtainable are listed in Additional file 5. Where possible, full texts of eligible non-English language articles were also translated using Google Translate; however, some articles could not be reliably translated. All non-English articles considered eligible based on title/abstract screening but that were not able to be fully translated are also listed in Additional file 5.

Articles with no effect size

During full text screening we discovered that more articles than anticipated reported raw or averaged data on TN or TP and chl-a but did not report a calculated effect size—that is, they did not provide a quantitative measure of the strength of the relationship between TN or TP and some measure of chl-a. In a deviation from the protocol [32], we did not attempt to digitize or extract raw data to calculate these effect sizes ourselves for this review, due to the number of articles and thus level of effort involved. Thus, effect size became a de facto eligibility criterion. Articles with no effect size were not advanced for validity assessment, full data extraction, or meta-analysis but are listed in Additional file 5.

Study validity assessment

Datasets from articles that advanced to full text screening and extraction were assessed for validity and risk of bias. Validity assessments were based on a detailed guide developed before extraction began, as well as notes generated by reviewers in discussions throughout the data extraction process (Additional file 6). In

an article, there could be one dataset (e.g., generated from one mesocosm experiment) or many datasets (e.g., generated from field samples taken in five states that were analyzed separately). Aspects of validity and risk of bias from published critical appraisal frameworks in environmental science and medicine [40–42] were examined to develop a review-specific validity assessment approach [43]. For each dataset within an article, aspects of validity contributing to a "low" or "high" risk of bias were rated, based on specific criteria for three study designs: (1) observational field studies, which typically sampled chl-a along a gradient of nutrient concentrations; (2) mesocosm experiments; and (3) field experiments (e.g., Before-After-Control-Impact designs [44]) (Table 2, Additional file 7). We slightly modified the questions for observational field studies from those listed in the protocol to distinguish between intra-site visit sample replication (e.g., triplicate water samples estimating nutrient concentrations) and repeated visits to the same site (see Table 2), both of which could reduce risk of bias as they increased. We also recognize that our validity assessment approach combines aspects of random and systematic error. As with data extraction, we assessed accuracy in validity assessment by having a reviewer not involved in the initial validity assessment independently assess validity for 25% of the studies evaluated by other reviewers; reviewers then discussed and resolved any differences. In a slight deviation from the protocol, we did not combine ratings to create an overall risk of bias for each dataset, but instead retained individual characteristic ratings for the sensitivity analysis. Validity results for all datasets included in the narrative synthesis and meta-analysis are in Additional file 8.

Data coding and extraction strategy

Data were extracted from articles that were considered eligible after full text screening. A data extraction template was created in MS Excel and used by all reviewers, with some modifications made after extraction was tested on an initial subset of articles. Each reviewer read the full text of articles and manually entered information into the Excel spreadsheet based on a detailed guide developed before extraction began and updated by reviewers in frequent discussions throughout the data extraction process (Additional file 6). The data extraction strategy focused on quantitative and qualitative information about each cause-effect pair—that is, each specific reported relationship between TN or TP and benthic or sestonic chl-a—as well as environmental factors that could modify the relationship. The direction and strength

Table 1 Detailed eligibility and exclusion criteria used to determine study eligibility in the systematic review

Eligibility criteria	Exclusion criteria
Population (unit of study)[a]	
Lotic fresh waters anywhere in the world Mesocosms made to mimic lotic freshwater systems	Lentic or non-fresh waters (wetlands, lakes, reservoirs, ponds, oceans, estuaries)
Exposure (environmental variable to which population is exposed)	
Exposure to total nitrogen (TN) or total phosphorus (TP) measured as concentration (e.g., mg/L)	Exposure only to other nutrients, or nitrogen and phosphorus not reported as TN or TP
Comparators (control or alternative intervention)	
Comparison to sites or treatments with lower or higher levels of TN or TP across a gradient Comparison to control group (no or background TN or TP) or to lower or higher levels of TN or TP in experimental studies	Studies of single sites (without sampling across time) or those without comparison to lower or higher levels of TN or TP
Outcomes (eligible outcomes resulting from exposure)	
Concentration of benthic or sestonic chlorophyll *a*, measured as mass per area or volume (e.g., μg/cm^2, mg/m^2, μg/L)	Studies examining only TN or TP with no data on biological responses Studies examining other biological effects
Study type	
Experimental studies in mesocosms or field sites Field-based, observational studies	Studies examining only TN or TP with no data on biological responses Studies examining only biological effects other than chlorophyll *a*
Publications (types of sources used)	
Study must contain original data Study must contain sufficient detail on methodology to assess study validity	Articles with no original data (e.g., editorials, reviews) Articles without sufficient information to evaluate pertinent relationships (chlorophyll *a* response to TN or TP) or study validity (e.g., methodology) Retracted articles

[a] We included some search terms that may capture studies in lentic habitats related to flowing systems (e.g., floodplain, riparian) in an attempt to obtain eligible studies that might otherwise be missed. We recognize that there is some uncertainty with the lotic/lentic distinction (e.g., flowing freshwater springs) and liberally included such articles at the title/abstract screening if otherwise eligible

Table 2 Study validity assessment framework for observational field studies

Domain of bias	Characteristic	Low risk of bias	High risk of bias
Study design and sampling	Pairing of nutrient and chl-a measurements	Nutrient and chl-a measurements taken at same place and time or index period	Nutrient and chl-a measurements are not paired in time and space
	Study timeframe	Sampling includes relevant periods over multiple years	Sampling occurs only over a single season or year
	Gradient definition	Gradient based on a nutrient related variable or its causal antecedent	Gradient based on a common causal descendant of TP, TN or chl-a
	Sample size	Acceptable # sites ($\geq$ 10) across gradient	Low # sites (< 10) across gradient
	Event-based replicates (sample replicates)	Multiple samples taken during each sampling event for nutrient and chl-a	Single samples taken during each sampling event for nutrient and chl-a
	Site-based temporal replicates (within-site replicates)	Multiple samples taken at each site (over study period)	Single samples taken at each site (over study period)
	Randomization of sampling (selection bias)	Some form of randomized site selection (e.g., stratified random sampling)	No randomization of site selection
	Confounding factors	If not controlled by study design, confounding factors are measured and adjusted for in statistical analysis	Confounding factors reported and not accounted for, or are likely, and are not able to be adjusted for post hoc
Data analysis and results	Clarity and detail	Analysis methods described in detail sufficient to permit repeating	Missing information not allowing for repeatability
	Uncertainty	Some estimate of uncertainty in effect or relationships provided (e.g., confidence intervals, standard error, standard deviation, etc.)	No estimates of uncertainty provided
	Reporting bias	All variables, measurements, and statistical tests mentioned in methods are reported in results or supplementary material	Some variables, measurements, or statistical tests mentioned in methods are not reported
Other biases	Detection bias	No indication that outcomes were measured differently in high versus low exposure sites	Some indication that outcomes were measured differently in high versus low exposure sites
	Attrition bias	No differences in loss of high versus low exposure sites	Differences in loss of high versus low exposure sites

of these relationships formed the basis for meta-analysis and narrative summary of the review results.

Authors were contacted if an article indicated that an effect size was calculated but its value was not reported (e.g., for non-significant associations or effect sizes for which only direction was reported). Limited information (some contextual details about the studies but no additional effect sizes) was gained as a result of author contacts.

One to five reviewers participated in data extraction from all eligible articles. To assess accuracy in data extraction, a reviewer not involved in initial data extraction independently extracted data for 25% of studies, and any differences were resolved and used to improve extraction consistency. Extracted data from eligible articles are provided in Additional file 8 and in ScienceHub, U.S. EPA's open access data repository (https://catalog.data.gov/harvest/epa-sciencehub).

Potential effect modifiers and reasons for heterogeneity

The secondary objective of this review was to examine the apparent variability in nutrient stressor-response relationships in lotic ecosystems that could be explained by moderating factors. Factors that potentially modify stressor-response relationships were extracted from eligible articles when these factors were examined in the original article. Based on evaluation of highly relevant articles and consultation with stakeholders and experts, the modifying factors (which differ slightly from Bennett et al. [32]) included:

- ecoregion;
- latitude;
- climate;
- elevation;
- stream channel width;
- watershed area;
- spatial extent;
- temporal extent;
- stream gradient;
- water discharge;
- water velocity;
- nutrient concentration range (lowest and highest TN and/or TP);
- existing background nutrient concentrations, including NH_4, TKN, and SRP;
- dissolved organic carbon;
- dissolved oxygen;
- water temperature;
- canopy cover/light availability;
- pH;
- alkalinity;
- sediment/turbidity; and
- conductivity.

For articles with sites in the conterminous United States that did not explicitly identify ecoregion, we assigned Level III Eco-region(s) using available information in the article. Geographically broad datasets (e.g., statewide or regional sampling) without more specific maps or locational information could not be categorized by ecoregion. Other relevant modifying factors were recorded as they were encountered during screening and data extraction. Methodological modifiers, such as chlorophyll type (benthic vs. sestonic), extraction method, measurement method [25, 45], and fraction of water sample used for nutrient measurement (filtered, unfiltered), were also recorded and factored into whether and how data were analyzed (see "Review findings"). Although our data extraction strategy encompassed a large number of potential modifying variables (Additional file 8), analyses were limited based on the number of studies reporting any single variable (see "Review findings" below).

Data synthesis and presentation

To synthesize data from the systematic review, we first described the evidence base across all 151 articles. We also narratively summarized cause-effect pairs included in the meta-analysis (105 articles) and calculated summary statistics about their associated modifying factors. We then followed Assink and Wibbelink [46] to conduct a three-level meta-analysis on effect sizes for the extracted cause-effect pairs that could be converted to a standard measure (Pearson's correlation coefficient). This three-level meta-analytic model was fit separately for each of the four stressor-response combinations of TN or TP coupled with benthic or sestonic chl-a (i.e., TN-benthic chl-a, TP-benthic chl-a, TN-sestonic chl-a, and TP-sestonic chl-a).

Pearson's correlation coefficient ($-1 \leq r \leq 1$), which quantifies the strength of the linear relationship between TN or TP and chl-a, was used as the effect size of interest. Whenever possible, published equations were used to convert other reported statistical measures to Pearson's r [47]. Some reported effect sizes could not be converted to Pearson's r and therefore could not be incorporated into meta-analysis (e.g., multiple regression R^2 if standardized slope was not provided; see Additional file 9). For the few articles with experimental studies that manipulated nutrient concentrations and reported differences in chl-a concentration between control and TN and/or TP treatment groups, we calculated raw mean differences (control vs. treatment) using the escalc function in the metafor package in R and the following input variables

for the control and treatment groups: mean, standard deviation, and sample size [47, 48].

Only cause-effect pairs with sufficient information were included in the three-level meta-analytic approach, proposed by Assink and Wibbelink [46]. This approach uses a random-effects model to account for variance in three levels due to articles containing one or more effect sizes. The levels of variance are sampling variance (level one), the variance between effect sizes extracted from the same article (level two), and the variance between effect sizes regardless of the articles (level three). Unlike random-effects models, a fixed-effects meta-analysis assumes one true effect size, τ, underlies all studies within and among articles and the variation in observed effect sizes is entirely due to sampling error, which also assumes that the observed effect sizes have an approximately normally distributed sampling distribution around τ. By contrast, the proposed multilevel random-effects model allows τ to vary randomly across articles. In the proposed random-effects model, it remains possible that all studies within and among articles share a common τ, but it is also possible that the effect size will vary from article to article [49]. Put differently, instead of allowing for one τ, the random effects model allows for a series of true effect sizes, $\tau_1, \ldots, \tau_N$, where N is the total number of unique articles. Without loss of generality, the model assumes multiple effect sizes from an arbitrary study $k, 1 \leq k \leq N$, to form well-behaved sampling distributions around τ_k. Random effects models are appropriate for making unconditional inferences about a random sample of obtained datasets [48, 50, 51].

Prior to model fitting, we evaluated the relative importance of the three levels of variance to determine whether a more complex model is necessary. This approach was preferred to others (such as randomly selecting a single representative effect size from each paper) because multiple effect sizes within articles often shed light on the influence of environmental context. We examined the impacts of modifying factors and sub-groups (the four stressor-response combinations) individually using sub-group analysis and meta-regression when possible. Ideally, we would have preferred to start with a fully random model and use a systematic or stepwise process to select the most important modifying factors and their interactions [52]. However, due to the numerous factors, interactions, and missing data, the most feasible option for this study was to analyze each individual moderator as a single fixed effect. We again followed the step-by-step tutorial in [46] for examining the moderating effects of individual categorical and continuous factors. The approach in [46] combines the random-effects meta-analytic model with fixed-effects (the modifying factors themselves)—called a mixed-effect model. As one might expect, the true effect size is allowed to vary randomly from article to article, but the moderator is assumed to have a true fixed effect that remains constant across all articles. Models were fit in the R version 4.0.2 package metafor [48]. Forest plots and sub-group plots were created using the packages metafor and metaviz [48, 53].

A Fisher's z-transformation ($z = \frac{1}{2} z = [\ln(1+r) - \ln(1-r)]$) was used to improve normality and variance for the meta-analysis models [54, 55] but raw r values were used for forest plots for ease of interpretation.

Sensitivity analysis

Risk of bias characteristics (Table 2, Additional file 7) were treated as factor levels in a sensitivity analysis to explore the impact of validity on effect sizes [33]. For each risk of bias characteristic, datasets were rated as either exhibiting an unclear, low, or high risk of bias. Then, ratings for a dataset were assigned to the one or more effect sizes that were based on that dataset. In a slight deviation from the protocol, we did not combine ratings to create an overall risk of bias for each dataset, but instead retained individual characteristic ratings in a twofold sensitivity analysis. Furthermore, sample size limited our ability to conduct sensitivity analysis to only observational field studies, using the risk of bias characteristics in Table 2.

First, we visually inspected the 14 risk of bias characteristics using kernel density estimator (KDE) plots for each stressor-response relationship. For each characteristic, two estimated distributions were produced and compared using KDEs: (1) the distribution of all effect sizes, and (2) the distribution of all effect sizes excluding those that were rated high risk of bias. If these two distributions are similar to one another, this would indicate that our conclusions are robust to high risk of bias effect sizes.

Next, we fit a random forest model to each of the four stressor-response relationships to understand which characteristics best predicted effect size. We omitted the sample size characteristic because "high" risk datasets with sample sizes < 10 were not included in the meta-analysis. A random forest model is a type of bagging routine that fits a large number of minimally correlated decision trees. As is true with many bagging routines, a random forest works by fitting a tree to m bootstrapped samples of the original data [56]. In an effort to de-correlate bagged trees, nodes (or splits) are determined by exploring only a random subset (p_{sub}) of the overall number of predictors, p. Typically, $p_{sub} \approx p/3$, and when $p_{sub} = p$, we have standard bagging. By comparing variable importance extracted from each of the four random forest models (one for each stressor-response relationship), we were able to determine what risk of

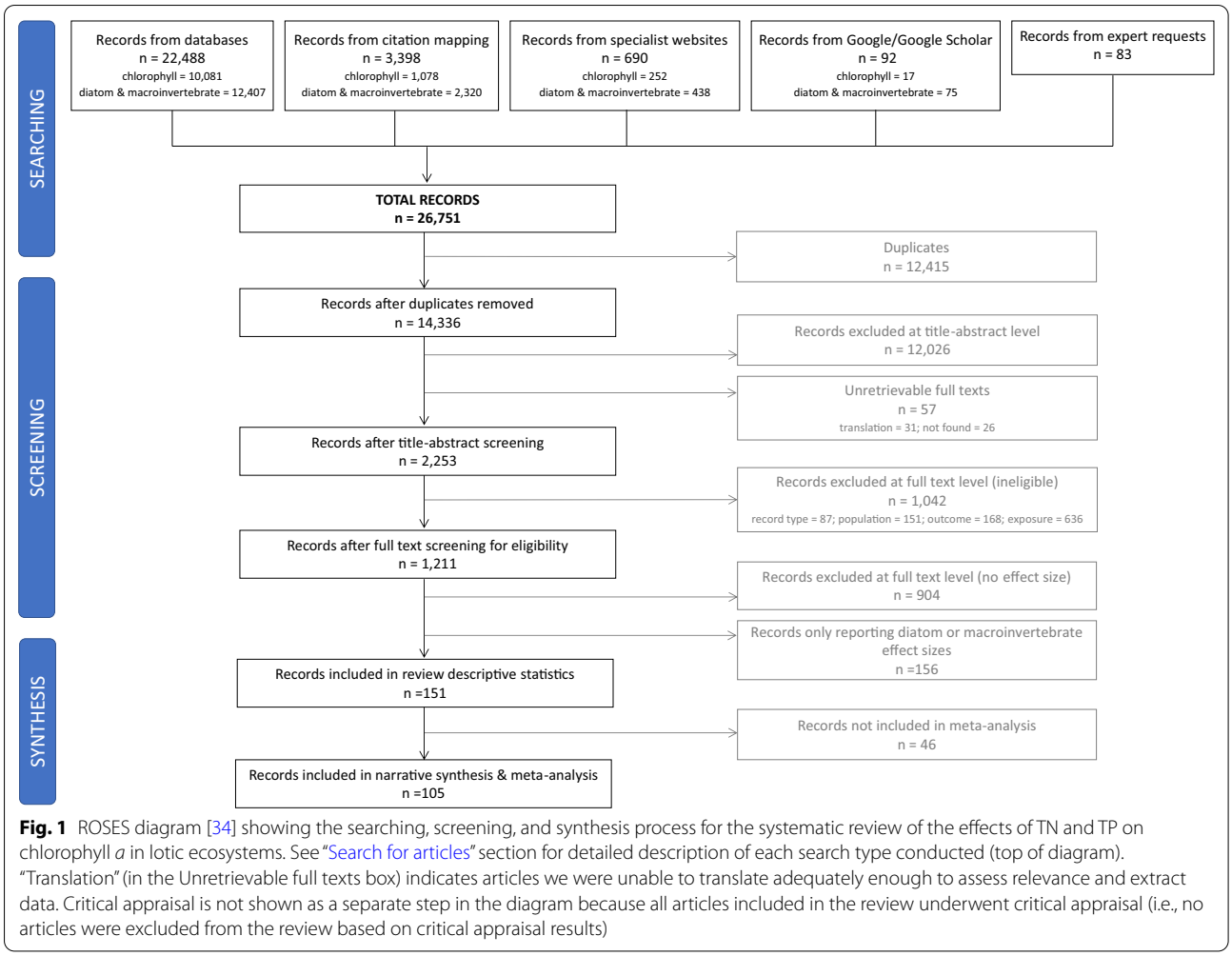

Fig. 1 ROSES diagram [34] showing the searching, screening, and synthesis process for the systematic review of the effects of TN and TP on chlorophyll *a* in lotic ecosystems. See "Search for articles" section for detailed description of each search type conducted (top of diagram). "Translation" (in the Unretrievable full texts box) indicates articles we were unable to translate adequately enough to assess relevance and extract data. Critical appraisal is not shown as a separate step in the diagram because all articles included in the review underwent critical appraisal (i.e., no articles were excluded from the review based on critical appraisal results)

bias characteristics have the largest influence on effect size. Any risk of bias characteristics that display a large influence on effect sizes need to be examined more thoroughly.

Publication bias was assessed using funnel plots comparing study effect sizes with sample size using the metafor and metaviz packages in R [48, 53].

We summarize the components of our meta-analysis using the checklist from Koricheva and Gurevitch [57] (Additional file 10).

Review findings

Review descriptive statistics

Literature searches across all three biological endpoints (chl-a, diatoms, and macroinvertebrates) returned an initial set of 26,751 articles. More than 2200 of these returns were screened at the full-text level to identify articles reporting a calculated effect size between TN or TP and at least one of the biological endpoints (Fig. 1). This review includes only the chl-a endpoint and draws from 151 articles that reported an effect size between TN or TP and chl-a (Additional file 9). An additional 1060 articles were excluded because they (1) reported cause and effect variables but did not formally analyze their correlation (904 articles) or (2) only reported an effect size for diatom or macroinvertebrate endpoints (156 articles) (Additional file 5); these articles are not considered further in this review.

Of the 151 chl-a articles, there were 873 identified cause-effect pairs. The most reported type of effect size for chl-a was R^2 from simple linear regression (n = 325), followed by Pearson's r (n = 221) and Spearman rank (n = 182). Less commonly reported measurements were Kendall's tau, partial correlation coefficient, slope coefficient, R^2 from multiple regression, and mean difference, among others (a total of 145).

Of the 873 cause-effect pairs, a total of 849 were judged for risk of bias based on observational field criteria, 15 cause-effect pairs were judged based on experimental

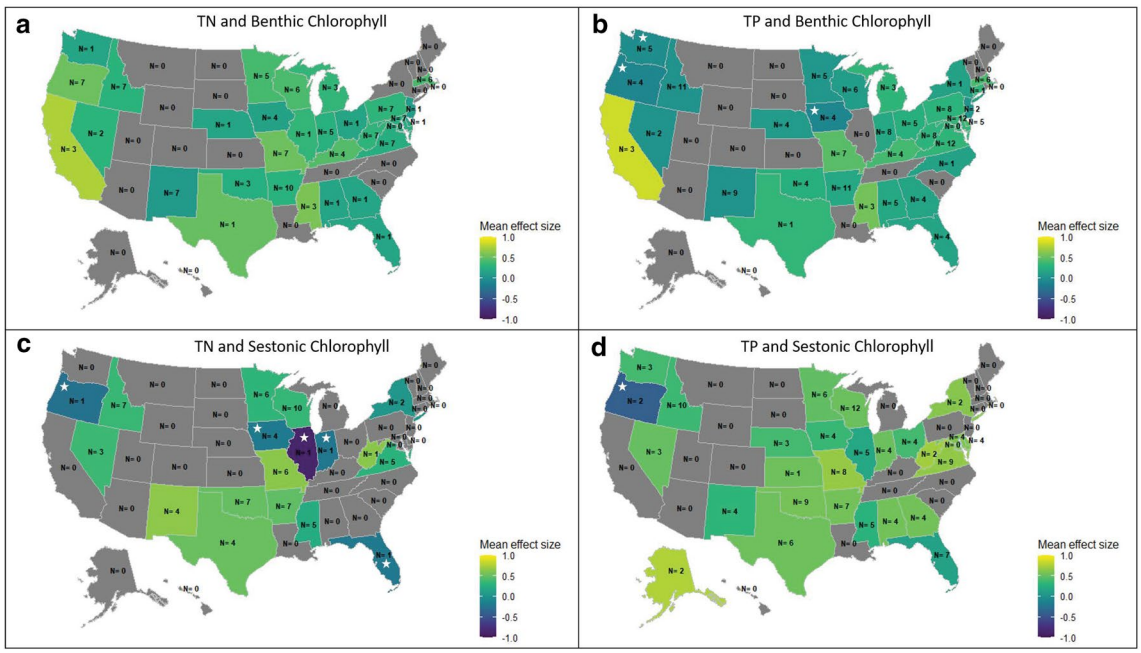

Fig. 2 Evidence availability and mean effect sizes for the U.S. (states and the District of Columbia). Maps show the number of cause-effect pairs (N) included in the narrative synthesis and meta-analysis for the stressor-response relationships: **a** TN-benthic chl-a, **b** TP-benthic chl-a, **c** TN-sestonic chl-a, and **d** TP-sestonic chl-a. White stars indicate effect sizes < 0

mesocosm criteria, and 9 cause-effect pairs were judged on field experiment criteria.

Narrative synthesis

Of the 151 articles and 873 cause-effect pairs comprising the full database, our narrative and quantitative syntheses use 105 articles and their 439 cause-effect pairs. The remaining 434 cause-effect pairs were not incorporated into the meta-analysis for the following reasons: missing results data (e.g., 355 cause-effect pairs were missing a quantitative effect size); use of analytical methods with unknown error distribution; or sample size < 10 (Additional file 9). Metadata and/or details about nutrient stressors, chl-a responses, and factors affecting the relationship between nutrients and chl-a were not reported for all cause-effect pairs, and so many of our results are based on fewer than 439 cause-effect pairs. Alternately, some cause-effect pairs could be assigned to multiple values for a metadata field (e.g., state, if the correlation was calculated based on data from 2+ U.S. states) and so totals may exceed 439.

Publication date ranged from 1980 to 2017. Actual sample collection dates ranged from 1976 to 2017. The most common temporal duration of sample collection for a cause-effect pair was at a single point in time (n = 162, 37% of cause-effect pairs), followed by months (n = 135, 31%) and then years (n = 116, 26%). However, when we resolved inconsistent extractions by multiple reviewers (see "Article screening and study eligibility criteria" section), we noticed that one area in which there was particular disagreement was the temporal duration field.

There were 144 cause-effect pairs based on samples collected outside of the U.S., while 295 of them were based on samples collected inside of the U.S. Only 34 out of 50 total U.S. states plus the District of Columbia had at least one cause-effect pair for one of the stressor-response relationships (Fig. 2). Of the states with ≥ 1 cause-effect pair, the number per state ranged from 1 (Connecticut, Kansas, North Carolina) to 35 (Arkansas, Idaho), with a median of 12.

Level III ecoregion could be assigned for 253 out of the 293 cause-effect pairs in the conterminous U.S. Cause-effect pairs existed for 63 out of 85 ecoregions. Of the ecoregions with ≥ 1 cause-effect pair, the number per ecoregion ranged from 1 (three ecoregions) to 37 (Ecoregion 39-Ozark Highlands) with a median of 12. Over half the cause-effect pairs were based on regional scale samples (n = 258, 59%). The next most common spatial scale was drainage basin (n = 137, 31%). Most cause-effect pairs were measured in a temperate climate (n = 311, 74%). The next most common climate was tropical/subtropical (n = 65, 15%).

Most cause-effect pairs were generated from field observational-type studies (n = 435, 99%), except two

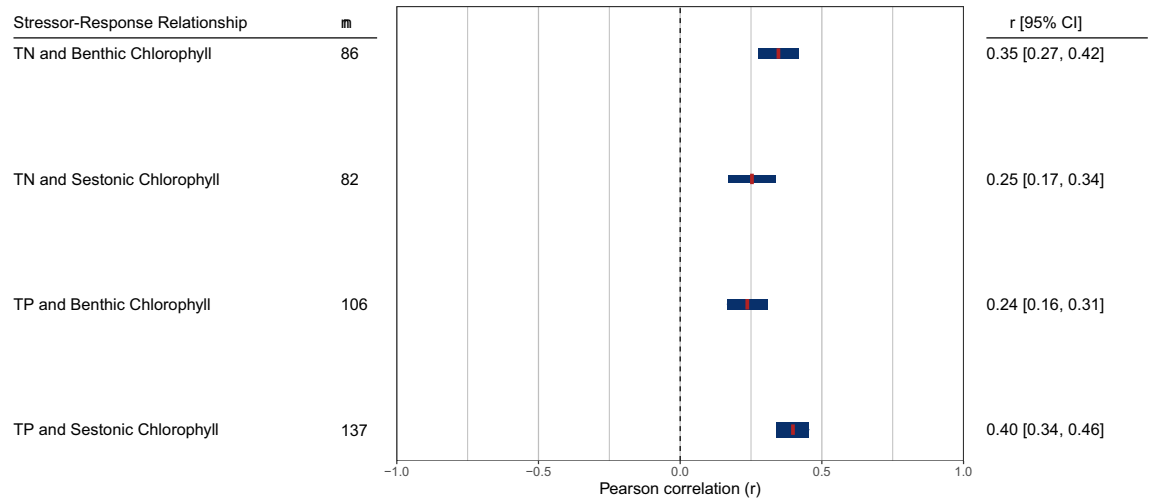

Fig. 3 Summary forest plot of the four stressor-response relationships. Forest plot shows mean effect size and 95% confidence intervals for two nutrient types (TN and TP) and two chlorophyll *a* types (benthic and sestonic) included in the meta-analysis

cause-effect pairs from a mesocosm study and two cause-effect pairs from a modeling study based on field observational data.

Of the 439 cause-effect pairs, the cause was TP for 251 (57%), TN for 173 (39%), and TN:TP for 15 (3%). The effect was sestonic chl-a for 227 (52%), benthic chl-a for 203 (46%), other chl-a for 9 (2%). "Other" chl-a included floating algae mats or a mixture containing both sestonic and benthic chl-a. Counts for the four main stressor-response relationships (excluding TN:TP and other chl-a because of low sample size) are found in Fig. 3.

The cause and effect minimum and maximum were reported often (>300 cause-effect pairs reported each). The range of mean TN concentration on which cause-effect pairs were based was 0.351–6.82 mg/L. The range for mean TP concentration was 0.008–2.2 mg/L. Mean benthic chl-a concentration varied from 6.16 to 12,879 mg/cm^2. Mean sestonic chl-a concentration varied from 0.000215 to 8.889 mg/L. Complete summary statistics for minimum, mean, median, and maximum values for cause and effect variables can be found in Additional file 11.

The most common type of effect size in the meta-analysis subset was Pearson's *r* (n = 172, 39%). The next most common effect sizes were converted to Pearson's *r* for the meta-analysis and included: R^2 from simple linear regression (n = 157, 36%), Spearman rank correlation coefficient (n = 108, 26%) and two Kendall tau measurements. The R^2 from simple linear regressions were excluded from the meta-analysis at a higher rate than the others, because sometimes the directionality of the effect could not be ascertained.

A total of 344 (78%) of cause-effect pairs indicated a positive effect between nutrients and chl-a and 93 (21%) indicated a negative effect. The meta-analysis subset appears to underrepresent relationships that were characterized as no effect (n = 2, 0.4%) compared to the full database (17%); the apparent discrepancy could be for several reasons. For instance, there could be a difference between our very strict definition of no effect (effect size = 0) for the meta-analysis subset versus a potentially more broad statistical interpretation of no effect (effect size had a calculated confidence interval that overlapped zero) reported by authors in articles in the full database. While no conclusions should be drawn about the directionality of relationships from these descriptive comparisons, we report them as evidence of potential underreporting and incomplete reporting of no effect results which is a common challenge in science [58].

In addition to nutrient (cause) and chl-a (effect) concentrations, we extracted information on instream physical and chemical modifying factors (see "Methods"—"Potential effect modifiers and reasons for heterogeneity"). We extracted minimum, mean, median, and maximum values for each factor when possible (except elevation, latitude, and longitude, for which we only extracted minimum and maximum) and present summary statistics in Additional file 11. Overall, we extracted at least one measurement for all 83 possible modifying factor values. The most reported modifying factor values were water temperature minimum, water temperature maximum, and watershed area maximum, with over 200 cause-effect pairs reporting information

associated with each. In general, the least reported modifying factor values were medians.

A total of 437, 2, and zero cause-effect pairs were judged for risk of bias based on the observational field, experimental mesocosm, and field experiment frameworks, respectively. For observational field datasets, none were judged to have a low risk of bias for all characteristics. The maximum number of low ratings was 11 (observed for only 1 cause-effect pair). Most cause-effect pairs had 5–9 low risk of bias ratings. Conversely, no observational field datasets were judged to have a high risk of bias for all characteristics. The maximum number of high ratings was 8. Most cause-effect pairs had 3–5 high risk of bias ratings.

Experimental papers

Five articles included mesocosm or field manipulation experiments that tested the effects of nutrient additions or reductions on chl-a. Outside of these five articles, we could not include many experimental studies in our analysis because the articles did not report TN and TP concentrations in the water column. Only one of these articles measured sestonic chl-a, the rest measured benthic chl-a. Most measurements of benthic chl-a were taken from artificial substrates added to mesocosms, while one study used rocks from an in-stream flow-through system. One of the experiments was conducted in New Zealand, while the remaining experiments were conducted in the U.S. We obtained 13 raw mean differences between control (no nutrient additions or reductions) and treatments (additions or reductions in N, P, or both). Sources of nutrients included experimental chemical enrichments or wastewater. Of the 13 raw mean differences, 8 were effect sizes of TP and benthic chl-a with a mean difference of 0.26 μg/cm^2 (±0.23 95% CI) chl-a between control and treatment. The mean difference for TN and benthic chl-a was 6.99 μg/cm^2 ± 1.42 95% CI, although this was based only on 3 effect sizes (Additional file 12).

Data synthesis

Our primary research question is, "*What is the response of chlorophyll* a *to total nitrogen and total phosphorus concentrations in lotic ecosystems?*" The lotic ecosystems captured by this review include streams and rivers varying in size and other characteristics that we hypothesized could moderate the response of chl-a to nutrients. These characteristics include physical (e.g., channel width, water depth, canopy cover), chemical (e.g., range of nutrient concentrations), and biological (e.g., benthic and sestonic algal taxa) components of the ecosystem. The relative importance of benthic and sestonic algae in the structure and function of small, open canopy streams likely differs from that in large, turbid rivers [59–61]. These differences likely influenced systems in which researchers measured benthic and/or sestonic algae for their studies. For example, while sestonic chl-a study sites spanned a large range of channel widths, from 60 cm to 800 m wide, the range for benthic chl-a study sites was only 30 cm to 60 m. Similarly, sestonic chl-a study sites were up to 24 m deep, while benthic chl-a study sites were less than 4 m deep. Despite this variability among sites at which benthic and sestonic chl-a were assessed, the mean effect sizes of both TN and TP on benthic and sestonic chl-a were positive (Fig. 3).

Of the four primary stressor-response relationships, TP-sestonic chl-a was the most positive ($r = 0.40$), followed by TN-benthic chl-a ($r = 0.35$), TN-sestonic chl-a ($r = 0.25$), and TP-benthic chl-a ($r = 0.24$). Pairwise comparisons showed whether the effect sizes for the four stressor-response relationships differed from each other. Holding chl-a type constant, benthic chl-a responses to TN vs. TP differed ($p = 0.01$), as did sestonic chl-a responses ($p = 0.0003$). Holding the nutrient constant, response to TP also differed between benthic vs. sestonic chl-a ($p < 0.0001$), but response to TN did not ($p = 0.42$). TN:TP had a much weaker relationship with benthic and sestonic chl-a and much greater variation around the mean effect size (Additional file 13). Because of the low sample size of TN:TP and other chl-a (e.g., floating algae mats) measurements, we excluded these cause-effect pairs from the meta-analysis.

Within the U.S., mean effect sizes were largely positive for all states and stressor-response relationships (Fig. 2). The few exceptions with slightly to moderately negative effect sizes were: Iowa, Oregon, and Washington for TP-benthic chl-a ($r = -0.11$, -0.09, -0.01, respectively); Florida, Illinois, Indiana, Iowa and Oregon for TN-sestonic chl-a ($r = -0.2$, -0.87, -0.22, -0.19, -0.26, respectively); and Oregon for TP-sestonic chl-a ($r = -0.40$).

Reasons for heterogeneity

The three-level random-effects model accounts for sampling variance (level 1), the variance between effect sizes extracted from the same study (level 2), and the variance between studies (level 3) [46]. For all four stressor-response relationships, less than 45% of the total response variance was attributable to within-study sampling variance (level 2). Despite this fact, the within-study sampling variance (level 2) was non-negligible for all stressor-response relationships excluding TN-benthic chl-a ($p = 0.6$). A larger percentage of the total response variance was attributable to differences in effect sizes between studies (level 3, 43–81%) than either the sampling variance (level 1) or the within-study variance

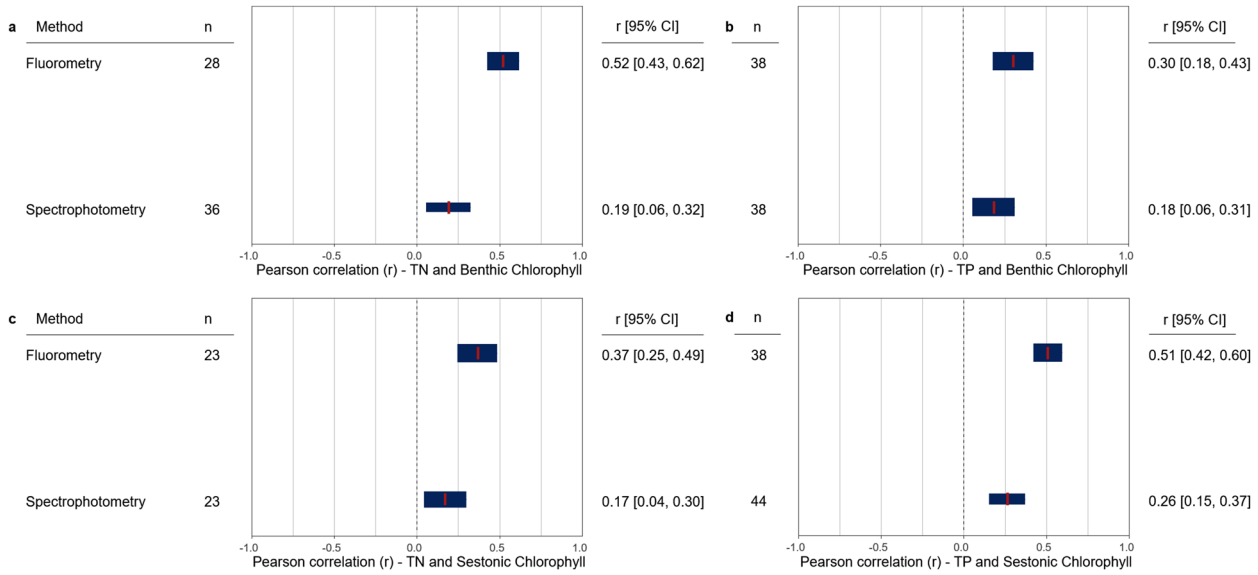

Fig. 4 Forest plots showing the influence of chlorophyll *a* measurement methods on mean effect size for the four stressor–response relationships

(level 2). The estimated variance allocations for each relationship can be found in Additional file 12. All four stressor-response relationships also displayed a significant amount of between-study variability. These results indicate that both within-study and between-study variability are non-negligible, implying that the three-level model is necessary to adequately represent the variance structure [46].

After determining that the three-level random effects model was necessary, we then proceeded to test the effect of potential moderating factors (both categorical and continuous) on mean effect size for each of the four stressor-response relationships (Additional file 13). Moderating factors were tested individually using the mixed effects model described in "Potential effect modifiers and reasons for heterogeneity" section. In total, we conducted 376 separate hypothesis tests (88 to 100 tests per stressor-response relationship). At the $\alpha = 0.05$ significance level, we would expect about five percent (~19) of these 376 tests to yield a false rejection of the null hypothesis due to multiple testing. Thus, relationships with $p < 0.05$ should be cautiously interpreted as potential ecological patterns of interest for deeper investigation.

Chlorophyll measurement method had a moderately significant effect on mean effect size for TP-sestonic chl-a ($F_{1,80} = 3.045$, $p = 0.085$): chl-a appeared to respond approximately twice as strongly to TP if fluorometry versus spectrophotometry was used, and the other stressor-response relationships consistently showed the same trend (Fig. 4). Spatial extent of sampling had a significant impact on TN-sestonic chl-a ($F_{3,78} = 2.648$, $p = 0.055$) and TP-sestonic chl-a mean effect sizes ($F_{4,134} = 2.597$, $p = 0.039$). For TN- and TP-sestonic chl-a, regional spatial extent had higher effect sizes than drainage basin. Temporal extent of sampling had a significant effect only on TP-benthic chl-a mean effect sizes ($F_{4,102} = 4.177$, $p = 0.004$). The longest temporal extent (years) had lower effect sizes than the shortest temporal extent (snapshot) (Additional file 13). The mean effect sizes of both spatial and temporal extent were consistently positive for the best represented categories (i.e., regional and drainage basin for spatial extent; snapshot, years and months for temporal extent) (Additional file 13). Climate did not have a significant effect on mean effect size for any of the four primary stressor-response relationships.

We had sufficient sample size ($n \geq 10$) to test the effect of 272 out of 420 continuous moderators. Of those, 26 (9.6%) meta-regression models (across the four stressor-response relationships analyzed) revealed a significant influence of a moderating factor on effect size at the $p < 0.05$ level. Few moderators had a significant effect on effect size for all four relationships. The complete set of meta-regression plots and test statistics not included in the main text can be found in Additional file 13.

The most notable continuous moderator is year of publication, which had a significant negative effect on mean effect size for all four primary stressor-response relationships (Fig. 5). Because these trends are consistent across all four stressor-response pairings, it is likely that these results are not merely a consequence of greater probability of significant results by chance because of multiple testing. These findings do appear to

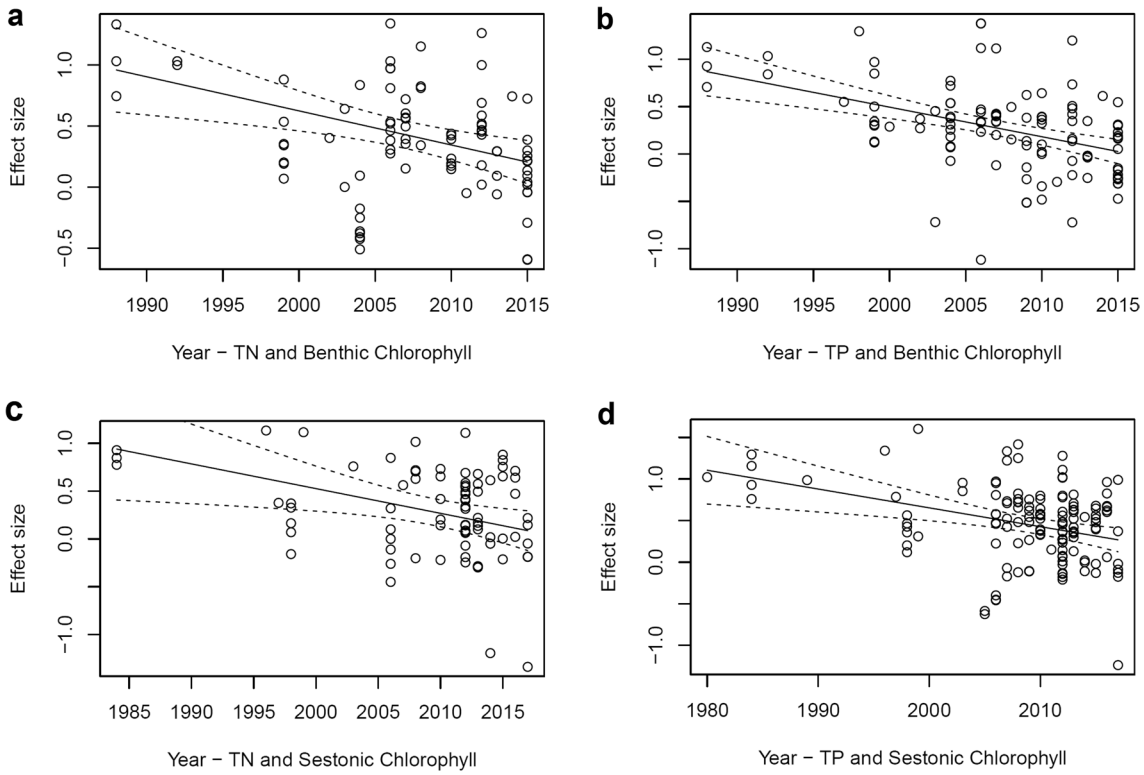

Fig. 5 Meta-regression plots showing the influence of publication year on effect size for the four stressor-response relationships. Effect size is plotted as Z-transformed Pearson correlation. Each circle is an effect size with the diameter representing total variance. Solid black line is the linear regression line and dotted black lines are 95% confidence interval

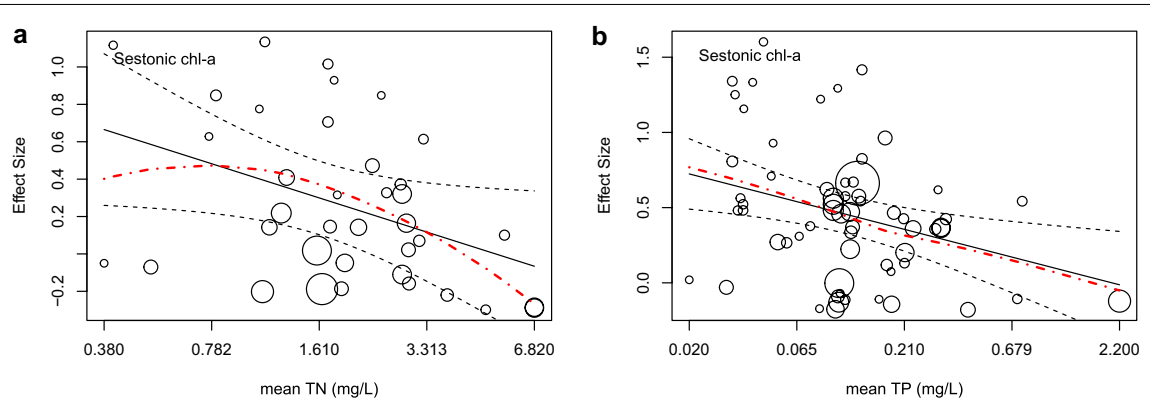

Fig. 6 Meta-regression plots showing the influence of mean nutrient concentration on effect size for the four stressor-response relationships. Effect size is plotted as Z-transformed Pearson correlation. Each circle is an effect size with the diameter representing total variance. Solid black line is the linear regression line and dotted black lines are 95% confidence interval. Red line is the LOESS curve

be driven by a few correlations reported in earlier years that tended to be large and positive. More recently, the range of correlation values has widened, and the mean effect size has decreased. When restricted to the subset of cause-effect pairs reported since 2000, the effect of year was significant *only* for TP-benthic chl-a ($F_{1,90} = 5.039$, $p = 0.027$). Other continuous moderators with $p < 0.05$ for TP-benthic chl-a included watershed area, water depth, conductivity, and turbidity; those for TN-benthic chl-a included conductivity and longitude; those for TP-sestonic chl-a included gradient and nutrient concentration; those for TN-sestonic chl-a

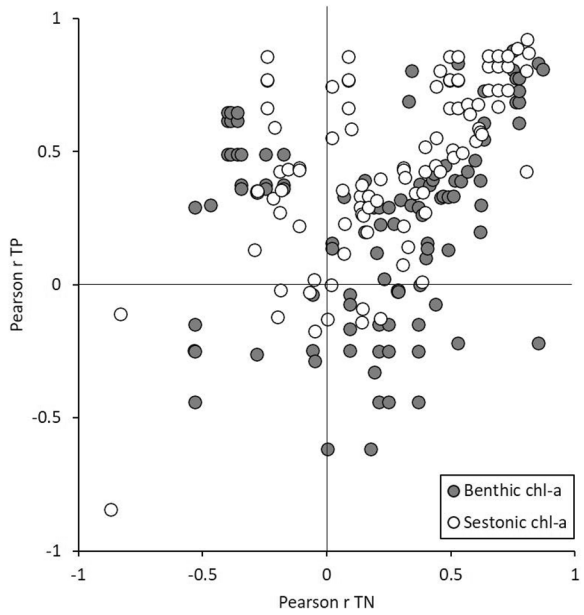

Fig. 7 Plot showing the effect size for TN and TP when measured in the same article. Effect sizes were matched within an article based on chlorophyll type (benthic, sestonic) and sample size

included watershed area, channel width, and nutrient concentration (Additional file 13).

The effect of mean nutrient concentration was negative and significant for both TN-sestonic chl-a and TP-sestonic chl-a. As the concentration of TN increased, sestonic chl-a became less correlated with TN. Similarly, as TP increased, sestonic chl-a became less correlated with TP (Fig. 6). This suggests that at some point, sestonic chl-a stops responding to additional TN and TP due to nutrient saturation or limitation by another environmental variable that is correlated with nutrients, such as light or turbidity [62, 63]. The patterns for benthic chl-a were harder to interpret. Benthic chl-a did not show the same tendency to become nutrient saturated when looking at mean nutrient concentration (although extreme values may have influenced the relationship), but did show a negative, though insignificant, relationship with median nutrient concentration (Additional file 13).

For articles reporting responses of chl-a to both TN and TP, we also examined the degree to which these responses exhibited similarity in magnitude and direction. To limit improper pairing of TN and TP effect sizes, we conservatively merged data for TN and TP based on the citation identifier, chl-a type (benthic, sestonic) and sample size. Across chl-a types, most responses to TN and TP appeared in the upper, right quadrant of the plot (Fig. 7). This shows that in places and times where chl-a responded positively to TN it also tended to respond positively to TP. We tested a few hypotheses to understand the pattern of responses, especially those with inverse relationships between TN and TP in the upper left and lower right quadrants of Fig. 7. The inverse relationships could not be explained by nutrient concentrations, turbidity, or discharge (Additional file 13, "Experimental papers" section). The most negative TN response had the largest channel width (500 m). The most negative response to both TN and TP was associated with the highest turbidity (> 785 NTU). Having no response (nearest to Pearson correlation = 0) to both TN and TP was associated with the highest discharge (> 2400 m^3/s).

We plotted Z-transformed correlation coefficients against sample size to look for evidence of publication bias. In the case of all four stressor-response relationships, we saw no evidence that either (1) small effect sizes were missing or (2) studies with small sample sizes were missing (Fig. 8).

To test the sensitivity of our conclusions to the underlying validity of data, we constructed a random forest model based on 13 risk of bias characteristics for observational field datasets for each of the four stressor-response relationships. Overall, the model for bias characteristics in TN-benthic chl-a explained the most variation in effect sizes ($R^2 = 0.485$). According to the normalized increase in mean square error, the most important characteristics in the model were "Clarity and Detail," "Study Timeframe," and "Randomization" (Table 3). Using kernel density estimation (KDE), we examined plots comparing all effect sizes (regardless of risk of bias rating) against *only* effect sizes with a low or unclear risk of bias. For the characteristic "Clarity and Detail," we found the plots to overlap almost completely (Fig. 9a). This indicates that cause-effect pairs with a high risk of bias for "Clarity and Detail" do not drive that characteristic's overall importance in the random forest model; those with an unclear risk of bias likely do. In contrast, the KDE plots for "Study Timeframe" and "Randomization" do not completely overlap. When removing high risk of bias cause-effect pairs, the central tendency of the distribution remains similar but the range contracts (Fig. 9a). For the mean effect size of TN-benthic chl-a, we conclude that our results are unlikely to be sensitive to underlying validity.

The random forest models for the three other stressor-response relationships explained less variation in effect sizes (Table 3; TN-sestonic chl-a: $R^2 = 0.245$; TP-sestonic chl-a: $R^2 = 0.186$; TP-benthic chl-a: $R^2 = 0.161$). In general, this indicates results for these three stressor-response relationships are even less sensitive to underlying validity than TN-benthic chl-a. For TN-sestonic chl-a, the most important characteristic was "Within Site Replication," for TP-sestonic chl-a it was "Pairing

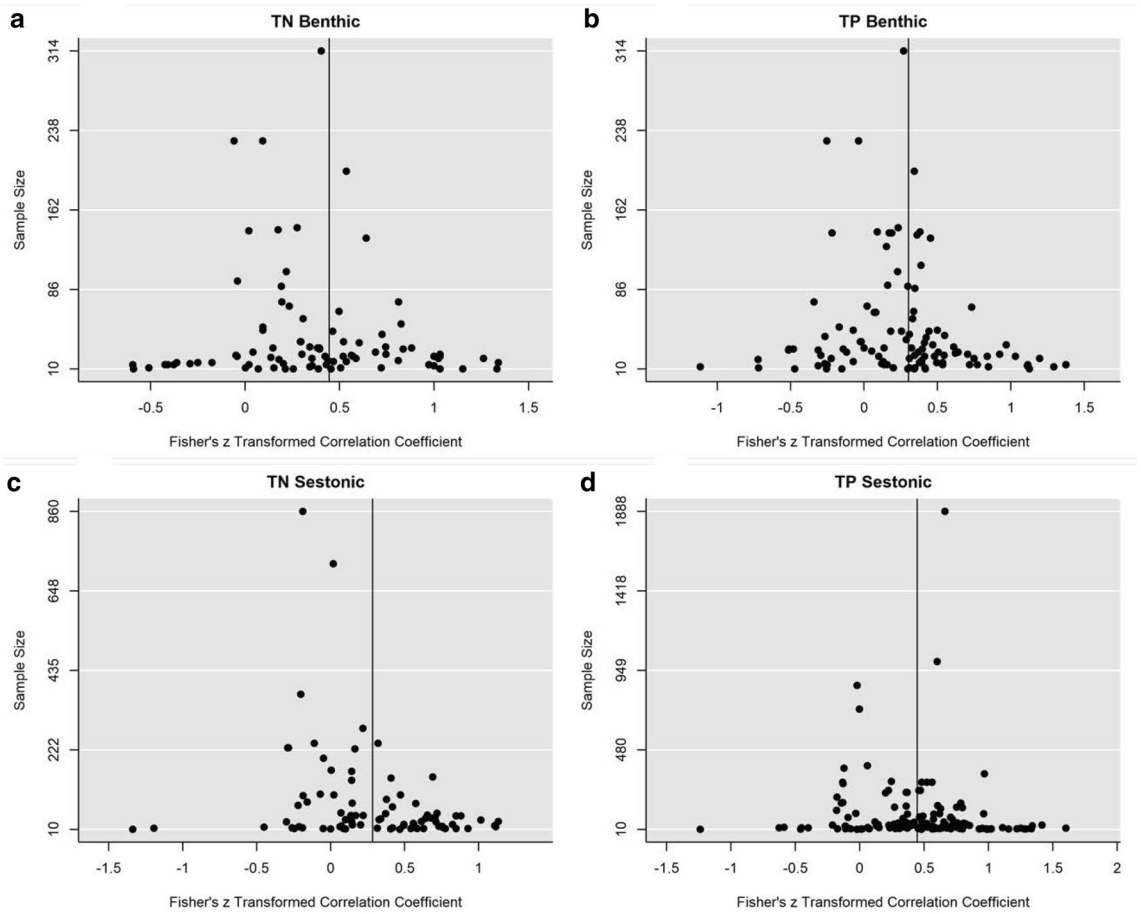

Fig. 8 Funnel plots for the four stressor-response relationships

Table 3 Results of random forest models for each of the four stressor-response relationships

Risk of bias characteristic	TN benthic R-squared: 0.485	TN sestonic R-squared: 0.245	TP benthic R-squared: 0.161	TP sestonic R-squared: 0.186	Mean
Clarity and detail	0.149	0.113	0.123	0.082	0.117
Study timeframe	0.114	0.079	0.162	0.113	0.117
Uncertainty	0.089	0.116	0.137	0.095	0.109
Gradient definition	0.087	0.090	0.085	0.113	0.094
Reporting bias	0.092	0.109	0.077	0.090	0.092
Randomization	0.099	0.056	0.117	0.072	0.086
Confounding	0.093	0.101	0.057	0.089	0.085
Pairing nutrient response	0.086	0.068	0.049	0.122	0.081
Within site replicates	0.041	0.121	0.058	0.076	0.074
Sample replicates	0.073	0.050	0.099	0.037	0.065
Attrition bias	0.069	0.050	0.032	0.050	0.050
Detection bias	0.009	0.048	0.000	0.057	0.028
Research aim consistency	0.000	0.000	0.004	0.004	0.002

Values in cells represent normalized increase in mean square error (MSE) for each risk of bias characteristic in the model. Larger numbers correspond to risk of bias characteristics having greater importance in the corresponding model. Risk of bias characteristics are ordered by mean MSE across models

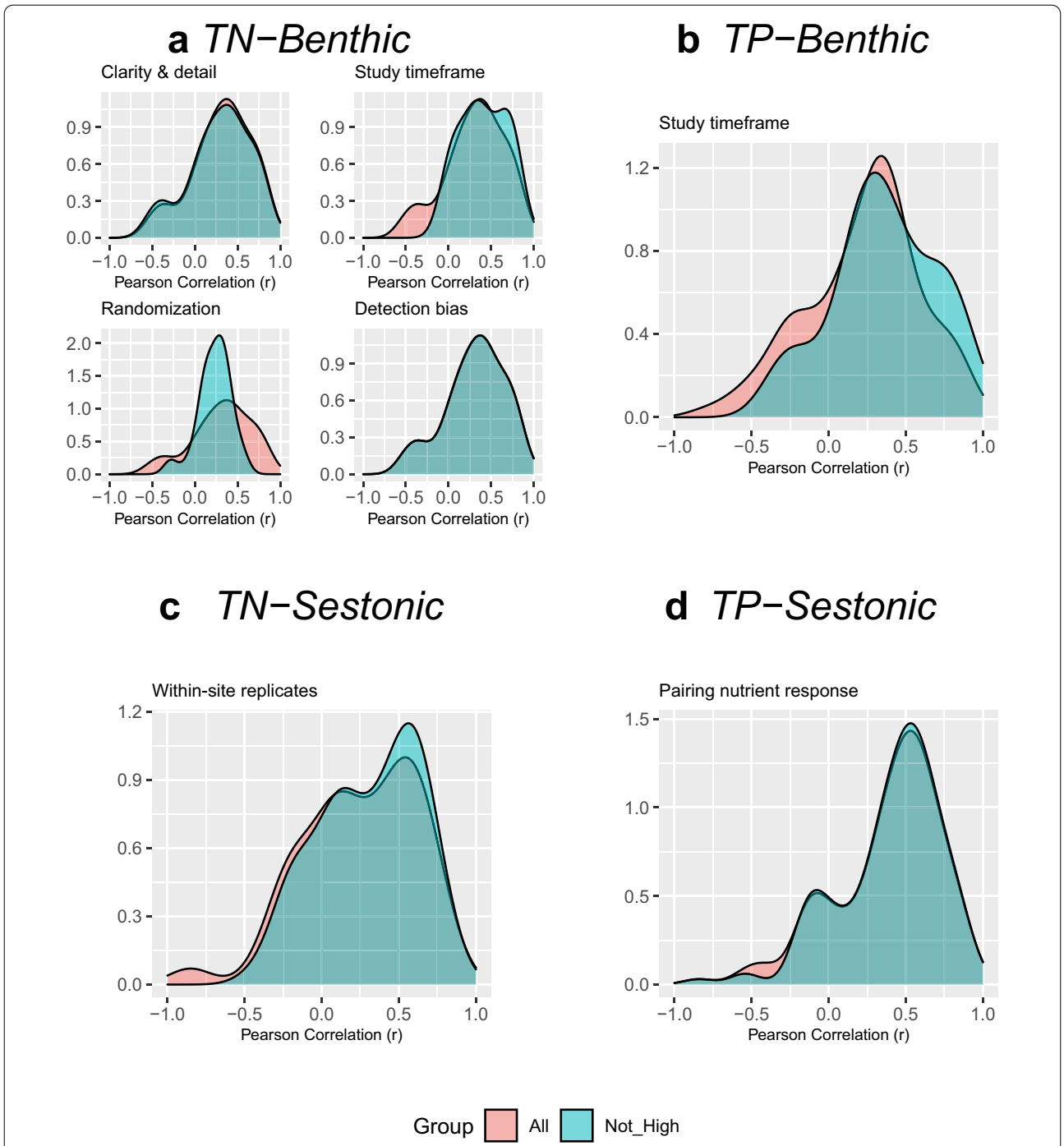

Fig. 9 Kernel density estimation (KDE) plots for selected risk of bias characteristics. Pink represents the distribution of all effect sizes. Teal represents the distribution of effect sizes with "low" or "unclear" risk of bias rating for each characteristic. KDE plots are shown for the four stressor-response relationships. Abbreviations correspond to the study validity assessment characteristics in Table 2 as follows: pairing of nutrient and chl-a measurements = pairing nutrient response; site-based temporal replicates = within-site replicates; randomization of sampling (selection bias) = randomization

Nutrient and Response," and for TP-benthic chl-a it was "Study Timeframe." Interestingly, the KDE plots for each of these most important characteristics show that removing high risk of bias cause-effect pairs contracts the distribution on the negative side (Fig. 9b–d), indicating effect sizes that are large and negative may rely on data that is of lower validity.

Review limitations

Limitations of review methodology

Our review ultimately included 151 articles across the years 1980–2017. Ideally, this review would also include more recent articles. A Web of Science Core Collection search for literature published since 2018, using the search string for chl-a described in Additional file 2 (search performed 10/16/20), returned 289 published articles—slightly over 2% of our original database search returns for chl-a. Although this cursory search provides a very rough estimate of potentially eligible, recently published articles, it does suggest that the relatively small volume of newer literature would likely have a relatively minor effect on our review conclusions.

We conducted our searches in English. This may have contributed to the geographic unevenness noted below.

The protocol for this review included a step for extracting data from articles that presented information on TN or TP and chl-a but did not statistically analyze their correlation [32]. Our intention was to extract raw or summary nutrient and chl-a concentrations and calculate correlations. Once the pool of records with no effect size was assembled, we decided there were too many to efficiently include them in this review (Fig. 1). As a result, those articles remain a potential source of stressor-response information that is untapped (Additional file 5).

Limitations of statistical methods

Generally, these meta-analytic techniques are thought of as relatively robust [46]. Using a multi-level approach, we can distinguish between sampling variance, the variance between effect sizes extracted from the same study, and the variance between effect sizes from the studies themselves. However, the moderator analyses contained herein are limited by the fact that they are only built to identify linear relationships between potential moderators and the response. Furthermore, all moderator analyses are restricted to individual variables and do not consider interactions between multiple moderating factors acting simultaneously.

Limitations of evidence base

The evidence base for this review has several notable limitations. First, evidence was not geographically comprehensive or evenly distributed. Cause-effect pairs were mostly assigned to the U.S., with just 33% assigned to the rest of the world. Within the U.S., there were gaps in evidence for states in the Plains and Intermountain West, as well as in the Northeast (Fig. 2). These gaps for states within the U.S. are important, because state governments are largely responsible for setting nutrient management goals for streams and rivers within their borders. It is possible that articles with no effect size reported (905 articles) could contain valuable data for filling these identified gaps. Regardless, this review can be used to identify regions or habitats with a particular dearth of data.

Second, our analysis of continuous moderating factors was hampered by (1) lack of reporting of this contextual information and (2) inconsistency in how this contextual information was reported when present. For the subset included in the meta-analysis, nearly all of the continuous moderating factors that we attempted to extract were reported for < 50% of cause-effect pairs, some for much less than half. The lack of and inconsistency in reporting contextual variables affected our ability to analyze the influence of all continuous moderating factors. In addition, even when the sample size was high enough, meta-regressions were sometimes influenced by the presence of one or a small number of extreme values or outliers. It is also well understood that environmental variables interact with each other, but we were unable to test the interaction of moderating factors in this review.

Finally, although field and mesocosm experiments were explicitly captured in our eligibility criteria, they made up a very small portion of our evidence base. Experiments provide a strong causal form of evidence and an important complement to observational studies. However, we found that experimental articles rarely reported measurements of total nutrients in the water column, which was an essential piece of information for this review. A recent review analyzing experimental nutrient additions more broadly, in terms of both nutrient stressors and biotic endpoints, examined 184 studies [64]. Elsewhere, we have encouraged researchers to include total nutrient measurements that can be relevant for environmental decision-making [27].

Review conclusions

Implications for policy/management

This systematic review provides nutrient managers responsible for protecting the quality of lotic ecosystems with a comprehensive evidence base of benthic and sestonic chl-a responses to total nutrient concentrations in the water column. It builds on previous

literature syntheses evaluating benthic and sestonic chl-a responses to nutrient concentrations [8, 65, 66], but focuses on nutrient measures most typically relevant to state-level nutrient managers. In the U.S., nutrient managers for individual states—the level at which many important water quality decisions are made—can use the results of this systematic review to understand and assess the evidence available in the literature that is relevant to their streams and rivers. In cases where evidence is lacking, nearby states with similar waters have applicable information.

Our review showed that the overall response of benthic chl-a and sestonic chl-a in streams and rivers to TN and TP is positive. This finding is based largely on observational studies, given the limited number of experimental studies that met our eligibility criteria. However, a similar meta-analysis of experimental nutrient additions in streams and rivers also reported positive response ratios among primary producers, as well as other trophic levels [64]. Although overall responses were positive, there were articles that showed a negative relationship for some nutrient and chl-a combinations. Although lack of data limited our ability to explore moderators that could be driving these negative relationships, we found that cause-effect pairs with "high" risk for key risk of bias characteristics across the four stressor-response relationships tended to have negative values. Further development of observational studies with both comprehensive reporting of moderating factors and low risk of bias may aid in clarifying conditions under which chl-a responses to total nutrient concentrations in the water column are more variable.

We found that the method used to measure chl-a in stream and river samples could affect the strength of association between total nutrients and chl-a. While the difference between the mean effect sizes for fluorometry and spectrophotometry-measured samples was not statistically significant, the clear and consistent pattern for all four stressor-response relationships was notable (Fig. 4). There are several plausible reasons for the difference. Fluorometry tends to have a lower detection limit, so a stronger correlation could be observed if more accurate chl-a measurements are generated by fluorometry at lower nutrient concentrations [67, 68]. Another possibility is that spectrophotometry measurements are more sensitive to sample variability (e.g., changes in algal species composition at different nutrient concentrations), resulting in a weaker correlation between nutrients and chl-a. Regardless, the observed pattern highlights the importance of reporting measurement methods when publishing chl-a concentrations. In addition, this observation should induce caution for programs that monitor or analyze chl-a. Chl-a measurement method may be an additional factor to consider when combining data from different projects or laboratories or developing standard protocols.

We found that effect size was negatively correlated with publication year (Fig. 5). Temporal instability in the body of evidence for a given ecological or evolutionary question, as well as different reasons for instability and its implications, have been identified previously [69]. A number of these reasons could apply to our systematic review, including early publication bias and a wider variety of environmental contexts included in later years. While it is possible that the overall conclusions of this systematic review could prove to be altered by literature going forward, we have some indication of temporal stability from 2000 to 2017 (see "Review findings").

In articles that report relationships between chl-a and both TN and TP, chl-a tends to show similar, positive responses (Fig. 7). Co-limitation by nitrogen and phosphorus has been demonstrated in numerous other studies, including those focused on experimental manipulations [65, 70, 71]. Our results indicate that observational studies also support the importance of controlling both nitrogen and phosphorus to limit eutrophication responses. Further examination of studies or sites in which chl-a responds positively to one nutrient but negatively to the other (e.g., studies in the upper left or lower right quadrants of Fig. 7) may help nutrient managers identify site-specific factors influencing the relative importance and specific roles of nitrogen and phosphorus.

Finally, we observed that the response of sestonic chl-a weakened as the mean concentration of both TN and TP increased (Fig. 6). Scientists and managers have long recognized that biota vary in their sensitivity to different environmental stressors [72, 73]. Taxa may also display contrasting response shapes (e.g., linear vs. S-shaped vs. parabola; [74, 75]). In agreement with primary studies, our review indicates that in the most eutrophic streams and rivers, sestonic chl-a is unlikely to be a useful biological measurement for tracking the impact of nutrients [76]. In those places, managers should consider other metrics to monitor condition. In contrast, we did not observe a clear pattern in benthic chl-a response across the range of mean TN and TP. In accordance with the Nutrient-Algal Biomass Conceptual model, we would not necessarily expect such a pattern [63, 77] due to the different environments in which benthic and sestonic algae tend to dominate (and thus be measured). Our review included responses from streams that ranged from oligotrophic to eutrophic and that encompassed many stream habitat types, all of which may have benthic algae communities that fail to respond to variation in nutrient concentration as a result of nutrient saturation

or confounding environmental factors such as canopy cover, according to the model. For both benthic and sestonic algae, responses to nutrients will be influenced by other environmental parameters (e.g., turbidity, light, discharge, grazer density) that affect algal assemblages both independently of and interactively with nutrients. Our results suggest that the usefulness of benthic chl-a as a biological indicator of nutrient impacts may be more context-dependent and influenced by study design and environmental factors.

Implications for research

Our review confirms that primary producers respond positively to nutrients across a range of stream and river conditions. Future research could explore the more nuanced patterns we observed, such as potential thresholds associated with the growth of primary producers in lotic ecosystems. For example, sestonic chl-a displays nutrient saturation or some other form of limitation at high TN and TP concentrations (Fig. 6). Our review also shows trends for other moderating factors, although these trends tend to differ across the individual stressor-response relationships. Few moderating factors showed similar influence on effect sizes in terms of significance and directionality across all four stressor-response relationships analyzed. There were some exceptions. For example, there were directionally similar, though insignificant, trends with minimum pH across all four stressor-response relationships (Additional file 13). We may attribute this general pattern to acidification acting as an overriding and fast-acting stressor on algal communities [78, 79]. Some of the limits to primary producer growth have direct application to nutrient management goals. Others would provide synthetic evidence addressing fundamental questions about aquatic ecosystems. Our review also prompts more mechanistic questions about the interactions among structural and functional components of aquatic ecosystems. Future research focused on underlying ecological processes to explain *why* we observed the patterns we did would provide additional insights that could be useful for management of nutrient pollution.

This review included cause-effect pairs from observational and experimental studies. Unfortunately, many cause-effect pairs from experimental studies had to be left out because TN or TP concentrations in the water column were not reported. Future experimental work would generate much-needed and strong evidence for how total nutrients affect growth of primary producers, as long as total nutrient concentration is reported.

The ability to conduct analysis of and see trends in moderating factors was affected by which factors were reported in the literature and in what form (e.g., range, mean, median). Future meta-analysis of these moderating factors would benefit from more complete and consistent reporting in the literature [80]. Such reporting is made possible when authors publish supplemental and/or digital files of summary statistics and raw datasets.

A potential, as yet untapped source of data and information that could address some of the research implications we have mentioned are the 904 articles that presented information on TN or TP and chl-a but did not statistically analyze their correlation. Extracting raw or summary nutrient and chl-a concentrations and calculating correlations could enhance evidence about stressor-response relationships (Additional file 5).

Chlorophyll *a* is a valid biological measurement that can be used in management of rivers and streams, but there are many other metrics for measuring biological condition. In our searching and screening process, we captured articles that reported the effects of total nutrients on diatoms and macroinvertebrates (Fig. 1). Data from these articles about diatoms and macroinvertebrates will be used in the future to conduct meta-analysis and synthesize evidence in a similar way to the current review in order to understand how these other important biological indicators respond to nutrients across a broad range of conditions in streams and rivers.

Finally, systematic reviews are typically most effective when the questions are specific and well defined (i.e., closed-framed questions) [33]. With more and better searching, screening, and data extraction tools (e.g., http://systematicreviewtools.com/), it will be possible to survey literature more broadly in a shorter period of time. The result could be more reviews like ours, with many effect sizes and diverse potential modifying factors. Advances in methods for data synthesis and analysis will likely need to accompany reviews like this. For instance, because our review included observational field data, mesocosm experiments, and field experiments, we developed sets of risk of bias characteristics that were applicable to each data source. In addition, we devised a new, efficient way to conduct sensitivity analysis of a large number of effect sizes. Future research on systematic review methods could focus on how best to draw conclusions from larger datasets of environmental evidence.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13750-021-00238-8.

Additional file 1. ROSES reporting standards.

Additional file 2. Database search strings and bibliographic databases used in review.

Additional file 3. Website searches.

Additional file 4. Other citation requests and "test set" of highly relevant citations.

Additional file 5. Articles not included in review due to irrelevance, inaccessibility, or failure to calculate an effect size.

Additional file 6. Guide to data extraction spreadsheet.

Additional file 7. Study validity assessment frameworks for experimental mesocosm and experimental field studies.

Additional file 8. Final extracted dataset.

Additional file 9. Final list of articles included in review.

Additional file 10. Meta-analysis checklist (from [57]).

Additional file 11. Summary statistics for modifying factors.

Additional file 12. Forest plots for effect sizes reported in experimental studies and variance distribution over meta-analysis models.

Additional file 13. R markdown file of the meta-analysis, showing forest plots with mean effect sizes for categorical moderating factors, results of three-level mixed effects models, and meta-regression plots for continuous moderating factors.

Acknowledgements
J. Oliver, B. Walsh, M. Pennino, T. Barnum, and two anonymous reviewers for helpful comments that improved earlier versions of the manuscript. J. Oliver, D. Thomas, B. Walsh, S. Jackson, L. Yuan, J. Alers-Garcia, G. Kaufman and many regional and state nutrient criteria coordinators provided indispensable feedback throughout the review. J. James and S. Solomon assisted with searches. M. Pennino developed Fig. 2. M. Fernandez developed Fig. 8. Disclaimer: The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. EPA. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the U.S. EPA.

Authors' contributions
This systematic review was initiated and conducted by all authors. MGB drafted the manuscript with substantial edits and improvements by KAS, SSL, CER, BJW, and DAG. All authors read and approved the final manuscript.

Funding
The U.S. EPA, through its Office of Research and Development, funded and managed the systematic review described here.

Availability of data and materials
All data associated with the review are publicly available as appendices to the publication or in ScienceHub, U.S. EPA's data repository.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests. Reviewers involved in this review that are also authors of eligible articles were not included in the decisions connected to eligibility and validity assessment of their articles.

Author details
[1]Present Address: Center for Public Health and Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, 1200 Pennsylvania Ave. NW (8623-R), Washington, DC 20460, USA. [2]Present Address: U.S. Environmental Protection Agency, Region 5, Chicago, IL, USA. [3]Present Address: Verisk Analytics, Jersey City, NJ, USA. [4]Present Address: Global Forest Watch, World Resources Institute, Washington, DC, USA. [5]Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA.

Received: 15 March 2021 Accepted: 9 September 2021
Published online: 03 October 2021

References

1. Cao D, Cao WZ, Fang J, Cai LY. Nitrogen and phosphorus losses from agricultural systems in China: a meta-analysis. Mar Pollut Bull. 2014;85(2):727–32.
2. Compton JE, Harrison JA, Dennis RL, Greaver TL, Hill BH, Jordan SJ, et al. Ecosystem services altered by human changes in the nitrogen cycle: a new perspective for US decision making. Ecol Lett. 2011;14(8):804–15.
3. Conley DJ, Paerl HW, Howarth RW, Boesch DF, Seitzinger SP, Havens KE, et al. ECOLOGY controlling eutrophication: nitrogen and phosphorus. Science. 2009;323(5917):1014–5.
4. Dubrovsky NM, Hamilton P. Nutrients in the nation's streams and groundwater: national findings and implications. In: Program NWQA, editor. United States Geological Survey; 2010. p. 1–13.
5. Jarvie HP, Sharpley AN, Spears B, Buda AR, May L, Kleinman PJA. Water quality remediation faces unprecedented challenges from "legacy phosphorus." Environ Sci Technol. 2013;47(16):8997–8.
6. Smith VH. Eutrophication of freshwater and coastal marine ecosystems—a global problem. Environ Sci Pollut Res. 2003;10(2):126–39.
7. Dodds WK, Bouska WW, Eitzmann JL, Pilger TJ, Pitts KL, Riley AJ, et al. Eutrophication of US freshwaters: analysis of potential economic damages. Environ Sci Technol. 2009;43(1):12–9.
8. Dodds WK, Smith VH, Lohman K. Nitrogen and phosphorus relationships to benthic algal biomass in temperate streams. Can J Fish Aquat Sci. 2002;59(5):865–74.
9. Pellerin BA, Bergamaschi BA, Gilliom RJ, Crawford CG, Saraceno J, Frederick CP, et al. Mississippi river nitrate loads from high frequency sensor measurements and regression-based load estimation. Environ Sci Technol. 2014;48(21):12612–9.
10. Poff NL, Ward JV. Physical habitat template of lotic systems: recovery in the context of historical pattern of spatiotemporal heterogeneity. Environ Manag. 1990;14(5):629–45.
11. Allan JD, Castillo MM. Stream ecology: structure and function of running waters. 2nd ed. Amsterdam: Springer; 2007. p. 426.
12. Barbour MT, Gerritsen J, Snyder BD, Stribling JB. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish, 2nd edition. Washington, D.C: U.S. Environmental Protection Agency, Office of Water; 1998. Report No.: EPA-841-B-99-002.
13. Hering D, Borja A, Carstensen J, Carvalho L, Elliott M, Feld CK, et al. The European water framework directive at the age of 10: a critical review of the achievements with recommendations for the future. Sci Total Environ. 2010;408(19):4007–19.
14. Karr JR. Defining and measuring river health. Freshw Biol. 1999;41(2):221–34.
15. Nichols SJ, Barmuta LA, Chessman BC, Davies PE, Dyer FJ, Harrison ET, et al. The imperative need for nationally coordinated bioassessment of rivers and streams. Mar Freshw Res. 2017;68(4):599–613.
16. Heiskary SA, Bouchard RW. Development of eutrophication criteria for Minnesota streams and rivers using multiple lines of evidence. Freshw Sci. 2015;34(2):574–92.
17. Dodds WK, Welch EB. Establishing nutrient criteria in streams. J N Am Benthol Soc. 2000;19(1):186–96.
18. Sutula M. Review of indicators for development of nutrient numeric endpoints in California Estuaries. Costa Mesa: Southern California Coastal Water Research Project; 2011.
19. U.S. Environmental Protection Agency. Nutrient criteria technical guidance manual: rivers and streams. Washington, DC: U.S. Environmental Protection Agency; 2000.
20. U.S. Environmental Protection Agency. Using stressor-response relationships to derive numeric nutrient criteria. Washington, DC: U.S. Environmental Protection Agency; 2010.

21. U.S. Environmental Protection Agency. U.S. EPA expert workshop: nutrient enrichment indicators in streams. Washington, DC: U.S. Environmental Protection Agency; 2014.
22. Dodds WK. Trophic state, eutrophication and nutrient criteria in streams. Trends Ecol Evol. 2007;22(12):669–76.
23. Otten TG, Paerl HW. Health effects of toxic cyanobacteria in US drinking and recreational waters: our current understanding and proposed direction. Curr Environ Health Rep. 2015;2:75–84.
24. Smith DR, King KW, Williams MR. What is causing the harmful algal blooms in Lake Erie? J Soil Water Conserv. 2015;70(2):27A-A29.
25. Steinman AD, Lamberti GA, Leavitt PR. Biomass and pigments of benthic algae. In: Hauer FR, Lamberti GA, editors. Methods in stream ecology. Burlington: Academic Press; 2006. p. 357–79.
26. Chambers PA, Culp JM, Roberts ES, Bowerman M. Development of environmental thresholds for streams in agricultural watersheds. J Environ Qual. 2012;41(1):1–6.
27. Bennett MG, Lee SS. Measuring lotic ecosystem responses to nutrients: a mismatch that limits the synthesis and application of experimental studies to management. Limnol Oceanogr Bull. 2019;28:26–30.
28. Poikane S, Kelly MG, Salas Herrero F, Pitt J-A, Jarvie HP, Claussen U, et al. Nutrient criteria for surface waters under the European Water Framework Directive: current state-of-the-art, challenges and future outlook. Sci Total Environ. 2019;695:133888.
29. Dodds WK. What controls levels of dissolved phosphate and ammonium in surface waters. Aquat Sci. 1993;55(2):132–42.
30. Dodds WK, Smith VH, Zander B. Developing nutrient targets to control benthic chlorophyll levels in streams: a case study of the Clark Fork River. Water Res. 1997;31(7):1738–50.
31. Dodds WK. Misuse of inorganic N and soluble reactive P concentrations to indicate nutrient status of surface waters. J N Am Benthol Soc. 2003;22(2):171–81.
32. Bennett MG, Schofield KA, Lee SS, Norton SB. Response of chlorophyll a to total nitrogen and total phosphorus concentrations in lotic ecosystems: a systematic review protocol. Environ Evid. 2017;6(1):1–13.
33. Collaboration for Environmental Evidence. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. 2013. p. 80. http://www.environmentalevidence.org/wp-content/uploads/2014/06/Review-guidelines-version-4.2-final.pdf.
34. Haddaway NR, Macura B, Whaley P, Pullin AS. 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. 2018;7(1):7.
35. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
37. Gamer M, Lemon J, Fellows I, Singh P. irr: various coefficients of interrater reliability and agreement. 0.84.1 ed2019.
38. R Core Team. R: a language and environment for statistical computing. 3.5.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2020.
39. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82.
40. Bilotta GS, Milner AM, Boyd IL. Quality assessment tools for evidence from environmental science. Environ Evid. 2014;3(1):14.
41. Mupepele AC, Walsh JC, Sutherland WJ, Dormann CF. An evidence assessment tool for ecosystem services and conservation studies. Ecol Appl. 2016;26(5):1295–301.
42. The Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions: the cochrane collaboration; 2011. www.handbook.cochrane.org.
43. Haddaway NR, Burden A, Evans CD, Healey JR, Jones DL, Dalrymple SE, et al. Evaluating effects of land management on greenhouse gas fluxes and carbon balances in boreo-temperate lowland peatland systems. Environ Evid. 2014;3(1):5.
44. Underwood AJ. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. J Exp Mar Biol Ecol. 1992;161(2):145–78.
45. Hagerthey SE, Louda JW, Mongkronsri P. Evaluation of pigment extraction methods and a recommended protocol for periphyton chlorophyll a determination and chemotaxonomic assessment. J Phycol. 2006;42(5):1125–36.
46. Assink M, Wibbelink CJ. Fitting three-level meta-analytic models in R: a step-by-step tutorial. Quant Methods Psychol. 2016;12(3):154–74.
47. Lajeunesse MJ. Power statistics for meta-analysis: tests for mean effects and homogeneity. In: Koricheva J, Gurevitch J, Mengersen K, editors. Handbook of meta-analysis in ecology and evolution. Princeton: Princeton University Press; 2013. p. 348–63.
48. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):48.
49. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1(2):97–111.
50. Gurevitch J, Hedges LV. Statistical issues in ecological meta-analyses. Ecology. 1999;80(4):1142–9.
51. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods. 1998;3(4):486–504.
52. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed effects modelling for nested data. In: Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM, editors. Mixed effects models and extensions in ecology with R. New York: Springer; 2009. p. 101–42.
53. Kossmeier M, Ulrich ST, Voracek M. metaviz: Forest plots, funnel plots, and visual funnel plot inference for meta-analysis, R package version 0.3.1. 2020.
54. Lassauce A, Paillet Y, Jactel H, Bouget C. Deadwood as a surrogate for forest biodiversity: meta-analysis of correlations between deadwood volume and species richness of saproxylic organisms. Ecol Ind. 2011;11(5):1027–39.
55. Worm B, Myers RA. Meta-analysis of cod-shrimp interactions reveals top-down control in oceanic food webs. Ecology. 2003;84(1):162–73.
56. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
57. Koricheva J, Gurevitch J. Uses and misuses of meta-analysis in plant ecology. J Ecol. 2014;102(4):828–44.
58. Jennions MD, Lortie CJ, Rosenberg MS, Rothstein HR. Publication and related biases. In: Koricheva J, Gurevitch J, Mengersen K, editors. Handbook of meta-analysis in ecology and evolution. Princeton: Princeton University Press; 2013. p. 207–36.
59. Gardner JR, Doyle MW. Sediment-water surface area along rivers: water column versus benthic. Ecosystems. 2018;21(8):1505–20.
60. Reisinger AJ, Tank JL, Rosi-Marshall EJ, Hall RO, Baker MA. The varying role of water column nutrient uptake along river continua in contrasting landscapes. Biogeochemistry. 2015;125(1):115–31.
61. Wetzel RG. Limnology: lake and river ecosystems. 3rd ed. San Diego: Academic Press; 2001. p. 1023.
62. Biggs BJF. Patterns in benthic algae of streams. In: Stevenson RJ, Bothwell ML, Lowe RL, editors. Algal ecology: freshwater benthic ecosystems. San Diego: Academic Press; 1996. p. 31–56.
63. Munn M, Frey J, Tesoriero A. The influence of nutrients and physical habitat in regulating algal biomass in agricultural streams. Environ Manag. 2010;45(3):603–15.
64. Ardón M, Zeglin LH, Utz RM, Cooper SD, Dodds WK, Bixby RJ, et al. Experimental nitrogen and phosphorus enrichment stimulates multiple trophic levels of algal and detrital-based food webs: a global meta-analysis from streams and rivers. Biol Rev. 2020. https://doi.org/10.1111/brv.12673.
65. Dodds WK, Smith VH. Nitrogen, phosphorus, and eutrophication in streams. Inland Waters. 2016;6(2):155–64.
66. Van Nieuwenhuyse EE, Jones JR. Phosphorus chlorophyll relationship in temperate streams and its variation with stream catchment area. Can J Fish Aquat Sci. 1996;53(1):99–105.
67. Arar EJ. Method 446.0: in vitro determination of chlorophylls a, b, c + c and pheopigments in 12 marine and freshwater algae by visible spectrophotometry. Cincinnati: US Environmental Protection Agency, Office of Research and Development; 1997.
68. Arar EJ, Collins GB. Method 445.0: in vitro determination of chlorophyll a and pheophytin a in marine and freshwater algae by fluorescence. Cincinnati: US Environmental Protection Agency, Office of Research and Development; 1997.
69. Koricheva J, Kulinskaya E. Temporal instability of evidence base: a threat to policy making? Trends Ecol Evol. 2019;34(10):895–902.
70. Beck WS, Rugenski AT, Poff NL. Influence of experimental, environmental, and geographic factors on nutrient-diffusing substrate experiments in running waters. Freshw Biol. 2017;62(10):1667–80.

71. Jarvie HP, Smith DR, Norton LR, Edwards FK, Bowes MJ, King SM, et al. Phosphorus and nitrogen limitation and impairment of headwater streams relative to rivers in Great Britain: a national perspective on eutrophication. Sci Total Environ. 2018;621:849–62.
72. Hering D, Johnson RK, Kramm S, Schmutz S, Szoszkiewicz K, Verdonschot PFM. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshw Biol. 2006;51(9):1757–85.
73. Segner H, Schmitt-Jansen M, Sabater S. Assessing the impact of multiple stressors on aquatic biota: the receptor's side matters. Environ Sci Technol. 2014;48(14):7690–6.
74. Sabater-Liesa L, Montemurro N, Font C, Ginebreda A, Gonzalez-Trujillo JD, Mingorance N, et al. The response patterns of stream biofilms to urban sewage change with exposure time and dilution. Sci Total Environ. 2019;674:401–11.
75. Wagenhoff A, Lange K, Townsend CR, Matthaei CD. Patterns of benthic algae and cyanobacteria along twin-stressor gradients of nutrients and fine sediment: a stream mesocosm experiment. Freshw Biol. 2013;58(9):1849–63.
76. Morgan AM, Royer TV, David MB, Gentry LE. Relationships among nutrients, chlorophyll-alpha, and dissolved oxygen in agericultural streams in Illinois. J Environ Qual. 2006;35(4):1110–7.
77. Munn MD, Frey JW, Tesoriero AJ, Black RW, Duff JH, Lee K, et al. Understand the influence of nutrients on stream ecosystems in agricultural landscapes. US Geological Survey Circular 1437. 2018:80.
78. Hirst H, Chaud F, Delabie C, Jüttner I, Ormerod SJ. Assessing the short-term response of stream diatoms to acidity using inter-basin transplantations and chemical diffusing substrates. Freshw Biol. 2004;49(8):1072–88.
79. Schneider SC, Oulehle F, Krám P, Hruška J. Recovery of benthic algal assemblages from acidification: how long does it take, and is there a link to eutrophication? Hydrobiologia. 2018;805(1):33–47.
80. Norton SB, Webb JA, Schofield KA, Nichols SJ, Ogden R, Bennett M, et al. Timely delivery of scientific knowledge for environmental management: a freshwater science initiative. Freshw Sci. 2018;37(2):205–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions