Water Research 282 (2025) 123769

SCCWRP #1442

Contents lists available at ScienceDirect

Water Research

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

A data-driven index for evaluating BMP water quality performance

Elizabeth A. Fassman-Beck [a,*], Edward D. Tiernan [a], Ka Lun Cheng [b], Kenneth C. Schiff [a]

[a] *Southern California Coastal Water Research Project, 3535 Harbor Blvd. Ste. 110, Costa Mesa, CA 92627, USA*
[b] *Los Angeles County Public Works, 900 S Fremont Ave, Alhambra, CA 91803, USA*

ARTICLE INFO

Keywords:
Urban runoff
NPDES, Stormwater quality, Nonpoint source pollution, Performance evaluation, Green infrastructure

ABSTRACT

A data-driven stormwater best management practice (BMP) performance index (PI) developed herein provides a simple, unbiased method to interpret the water quality treatment performance of a structural BMP, and provides actionable information for managers in planning, implementing and maintaining BMPs. The PI is derived from field-monitored influent-effluent pollutant event mean concentrations normalized by a user-specified water quality benchmark that may be adapted to reflect watershed-specific objectives. Benchmarking allows performance of any BMP to be investigated regardless of the treatment mechanisms, climate conditions, hydrologic performance, or site-specific pollutant concentrations. Quantitative monitoring data normalized by the benchmark are subsequently binned into performance categories of Success, Excess, Marginal, Insufficient, and Failure, which are indicators of the relative potential to achieve downstream receiving water goals. A single PI score per analyte ($PI_{analyte}$) is derived by compositing the categorical distribution. A simplified analytical hierarchy procedure is adapted to combine multiple $PI_{analyte}$ scores, if/when BMP selection for future applications must address a range of pollutants, e.g., where multiple water quality objectives are present. Data from the International Stormwater BMP Database are used to demonstrate PI applications such as selecting the "right" BMP to address specific water quality concerns, comparing amongst similar types of BMP to identify beneficial design features, and the clear interpretation offered by the score providing actionable information compared to a percent-removal assessment method. An index interpretation guide provides managers, maintenance teams, and designers with a feedback mechanism to refine future project designs, operations, or maintenance needs, and informs progress towards achieving successful watershed management plans.

1. Introduction

The predominant method for reducing pollutant loads in stormwater is the installation of structural best management practices (BMPs). BMPs encompass a suite of technology types, and may be referred to in the USA as stormwater control measures (SCMs), green infrastructure (GI), or low impact development (LID), while in Europe the term sustainable urban drainage systems (SUDS) may be preferred, among other terms found internationally (Fletcher et al., 2015). Each BMP type offers different pollutant removal mechanisms to treat stormwater (Clary et al. 2020), and almost all BMPs operate according to passive, gravity-flow (Davis et al. 2022). Even within one type of BMP, design and construction specificity can lead to a nearly limitless combination of treatment strategies, as evidenced by the proliferation of design guidance amongst state and local jurisdictions in the USA (Galavotti 2016; Law et al. 2008; Minnesota Pollution Control Agency 2022a; Snyder et al. 2020). Prescriptive design approaches for water quality BMPs in the USA set expectations to remove pollutants to the "maximum extent practicable" (Muthukrishnan et al. 2004), rather than to an effluent quality or system efficiency.

Calculations of "percent removal" per storm are the most common metrics for BMP performance assessment found amongst hundreds of monitoring studies (Moore et al. 2017, 2018; Rodak et al. 2019, 2020, Vogel et al. 2016), and are used in many regulatory settings (e.g., New Jersey DEP 2004, Minnesota Pollution Control Agency 2022b). Regrettably, percent removal is a strongly biased performance metric for stormwater treatment systems (Urbonas, 2003; Jones et al. 2005; Barrett 2005; Gilliom et al. 2020; Strecker et al. 2001). Percent removal calculates a change relative to an initial condition, in this case the BMP influent event mean concentration (EMC) or influent pollutant mass load. The calculation biases the outcome because significant inter-storm and inter-site variability in untreated (i.e. influent) wet weather runoff quality (Pitt et al. 2018; Simpson et al. 2022; Strecker et al. 2001) means that BMP performance is assessed against a moving target, rendering the

[*] Corresponding author.
E-mail address: elizabethfb@sccwrp.org (E.A. Fassman-Beck).

https://doi.org/10.1016/j.watres.2025.123769
Received 18 December 2024; Received in revised form 29 April 2025; Accepted 2 May 2025
Available online 2 May 2025
0043-1354/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

interpretation of the result to subjective judgement. Geosyntec et al. (2015) attributed statistically weak correlations between influent and effluent concentrations for a range of pollutants and BMP types, to the high variability in many monitoring data sets. A percent-removal metric on its own at a BMP- or site-scale cannot directly indicate whether the BMP will make a meaningful impact on protecting or restoring downstream water quality (Jones et al. 2005).

Interpreting BMP performance is also complicated by technical factors such as limited or unequal data set sizes, variations in storm characteristics, differences in BMP designs, and the wide range of BMP placements in the urban landscape (Geosyntec et al. 2015; Jones et al. 2005; McNett et al. 2010; Strecker et al. 2001). Monitoring studies may call out exceptions to overall performance assessment for the influence of extreme events (i.e. very large storms compared to the design basis or high influent concentrations). Strecker et al. (2001) suggest calculating a %-change over the sum of loads over multiple storm events, to account for the influence of storm size, but this approach is uncommon in practice likely due to the difficulty of monitoring every event.

Barrett (2005) proposed an alternative performance evaluation method based on linear regression of measured influent and effluent concentrations. Monitoring data reflecting extremely high influent concentrations or where no effluent was produced (e.g., from small storms that are completely retained) could not be included in the framework. The method was also site- and BMP-specific, without offering any means to translate to another situation. Furthermore, if additional data were obtained through monitoring, new regression equations would be necessary.

A range of statistical approaches and online tools to evaluate BMP water quality performance (notably excluding percent-removal calculations) is provided by the International Stormwater BMP Database (BMP Database, https://bmpdatabase.org/bmp-statistical-analysis-tool), along with a large-scale repository of BMP monitoring data (https://bmpdatabase.org/). Statistical tools include hypothesis testing of the differences between influent and effluent EMCs, and multiple graphical methods to investigate and evaluate data distributions. The BMP Database is a comprehensive, invaluable technical resource with statistical summaries published periodically (Clary et al., 2020); however, interpretation of the multiple types of analyses is left to the user, which often requires an in-depth understanding of complicated outputs.

Strecker et al. (2001) and Barrett (2005) suggest that BMP effluent quality should offer a more robust measure of BMP treatment compared to %-removals. Limited research has explored benchmarks for BMP effluent quality. For example, Barrett (2005) explores the influence of irreducible concentrations (originally described by Schueler [1996]) on BMP performance evaluations. The irreducible concentration concept suggests a lower limit of treatment that could be interpreted as a performance benchmark. McNett et al. (2010) proposed assessing BMP performance for total nitrogen (TN) and total phosphorus (TP) by comparing effluent EMCs against in-stream concentrations associated with benthic habitat ratings. This framework has been applied to a range of BMP monitoring studies in North Carolina (Braswell et al. 2018; Page et al. 2015; Koryto et al. 2017; Luell et al. 2021, Smolek et al., 2018, Wissler et al. 2020). The Technology Assessment Protocol:Ecology (TAPE) program for evaluating manufactured treatment devices uses %-removals that are binned according to a range of influent EMCs to benchmark performance for use applications such as pre-treatment, general use, or conditional use, among other designations (WSDE 2024). Each of these alternative approaches to BMP performance addresses at least one of the shortcomings of conventional %-removal assessments. Only the TAPE program offers a context for interpreting results, i.e., judging whether performance is "good enough" or otherwise, but there is no direct method to interpret BMP impacts to in-stream receiving water quality. Only McNett et al. (2010) provides a method to consider extreme events, as the subsequent studies cited in North Carolina used probability distributions to determine the proportion of events that met water quality criteria (Braswell et al. 2018; Page et al. 2015; Koryto et al. 2017; Luell et al. 2021, Smolek et al., 2018, Wissler et al. 2020). In practice, specific objectives for effluent quality from individual BMPs are not typically found in regional or state-issued design guidance since pollutant loadings and receiving water goals are usually site- or watershed-specific.

Ultimately, while BMPs are often considered "efficient" systems for removing pollutants from stormwater runoff, there remains no standardized or transferable method for assessing context-specific BMP effectiveness. Watershed and stormwater managers are desperate for this type of information given that the US EPA's Environmental Financial Advisory Board (2020) anticipates US$7.5 billion per year will be spent nationally for municipal separate storm sewer (MS4) permittees comply with Clean Water Act (1972) requirements for managing stormwater runoff. In Southern California, since 2018 the County of Los Angeles allocates approximately US$280M annually for stormwater quality improvement projects and studies through its Safe, Clean Water Program (https://safecleanwaterla.org/). The County of Orange (2024a & 2024b) most recent estimates exceed $413M for capital construction of structural BMP projects and approximately $13M/year in annualized operations and maintenance. BMP projects constructed on private parcels or by private ownership are not included in these allocations.

The objective of this study is to develop a comprehensive and unbiased data-driven assessment method to interpret a structural BMP's water quality treatment performance, and provide actionable information for decision-making. Criteria for what an objective framework for water quality performance evaluation should achieve are established; the quantitative and qualitative rationale of the proposed water quality performance index (PI) structure is described; and case studies using data from a range of real-world BMP monitoring studies are used to demonstrate applicability and flexibility. Interpretation of the index provides stormwater managers, maintenance teams, and designers with a feedback mechanism to refine future BMP project designs and operations. The PI is intended to replace subjective performance assessments such as %-removals with an objective, transferable, data-driven index that enables site-specific BMP performance to be interpreted in a watershed context and provides actionable information to advance the state of the practice in stormwater management.

2. Methods

2.1. Water quality performance index development criteria

At its core, any BMP performance assessment should judge performance relative to whether a BMP currently does, or has the potential to reduce or minimize pollutant transport downstream and consistently contribute to achieving receiving water quality goals. With this mindset, primary considerations for a water quality PI were identified as:

a) An index based on water quality EMCs provides insights into whether a BMP induces appropriate treatment mechanisms;
b) The index must consider influent and effluent water quality to track performance on a storm-by-storm basis, including extreme events, and over time;
c) The index must be able to incorporate site-specific water quality objectives that vary between regions and watersheds;
d) The index must be adaptable to a wide range of water quality parameters that can be assessed individually and collectively. Stormwater managers rarely design BMPs for only a single pollutant;
e) The index must be compatible with a wide variety of BMP types and designs;
f) The index must be objective, quantitative, and repeatable, while being transparent and easy to understand for technical experts and non-technical decision makers.

2.2. Performance index development approach

Index development required four steps including: (1) conceptual design, (2) interpreting index scores across multiple storm events for management implications, (3) evaluating sensitivity to benchmarks, and (4) integrating performance across multiple pollutants. The conceptual design outlines the index's foundation and identifies the quantifiable metrics. Then, the scoring criteria are developed, which translate numeric scores into management actions, which is the ultimate goal of developing an index. Sensitivity to benchmarks critically evaluates how the index responds to different water quality objectives or thresholds, which is an integral element of comparing across BMPs or pollutants. The index framework leads to an index score for a single water quality parameter, which may be aggregated to a multi-parameter index to encompass multiple water quality concerns or performance expectations.

Engaging target stakeholders and technical experts from the project's initiation was considered critical to developing an index that is useful for management application and is based on robust science while also being practical (i.e., can be applied using existing typical approaches to BMP design and monitoring). An Advisory Committee comprised of watershed managers, state and regional regulators, academia, industry, and a non-governmental organization was convened nine times over almost four years to review and inform critical decision points in the index's development.

2.3. Data sources

Selected data from the BMP Database was used to develop the PI. The data was used herein for the purposes of demonstration of the performance evaluation framework, rather than to provide a comprehensive assessment of the BMPs themselves.

Data from 18 bioretention type BMPs and 13 dry extended detention basins recommended by the Advisory Committee were used for conceptual index design and evaluating sensitivity to thresholds. Bioretention BMP data sets were selected representing multiple climates and geographies across the USA, including CO, KS, MD, NC, OH, and VA. Detention basin data sets were obtained from CA, CO, KS, MD, MN, NC, and OR. Names of specific BMPs are provided in the Supplemental Materials, Tables SI-1 and SI-2. Design approaches, as-built compliance, and maintenance conditions likely differ, thus encompassing a range of BMP conditions in the data set. Bioretention BMPs with underdrains (where treated effluent discharges to a downstream flow conduit, a.k.a. flow-through BMPs) and exfiltration (where the effluent discharges to surrounding soils, a.k.a. full capture BMPs) were included. Data analysis uses four example pollutants, total suspended solids (TSS), total copper (Cu), TP, and fecal coliform which were somewhat arbitrarily selected based on data availability and represent classes of common stormwater pollutants of concern (sediments, nutrients, heavy metals, and fecal indicator bacteria). Availability of widely distributed data points was considered more important than the specific location where data were collected in order to explore implications of benchmarks, thresholds, weighting strategies and score interpretations as the index was developed.

Paired influent and effluent EMC data were used for index development and example case studies. Data from the Shop Creek pond and wetland treatment train (Denver, CO) in the BMP Database was used to evaluate the effectiveness of the BMP design. The treatment train is comprised of an upstream retention pond discharging to a downstream wetland, i.e., the outlet of the retention pond is the inlet to a series of wetlands. It manages runoff from 550 acres of mostly single-family low-density housing in the Cherry Creek Basin near Denver, CO. Flow-weighted EMCs from monitoring data are available from 1990–1997, beginning shortly after the system was brought into service. Internal records from the Urban Drainage and Flood Control District (renamed in 2019 as the Mile High Flood District) and the Cherry Creek Basin Authority document pond re-grading and planting to introduce a vegetated littoral zone in 1994. Re-grading did not introduce additional storage capacity, but altogether represents a significant design modification that might influence performance. Monitoring data were broken down into two distinct ranges (1990–1994, 1995–1997) for analysis.

In data processing, a value of ½ the method detection limit was assigned to data reported below the detection limit, for consistency with the BMP Database protocol. The effluent EMC was assigned as a numeric zero when no measurable discharge volume was recorded, as opposed to data that were not measured or otherwise missing. Bioretention BMPs included herein are represented by at least four storm events for any analyte considered. Detention basins are represented by at least six storm events. The current effort is not intended as a meta-analysis of existing data within the BMP Database, nor a treatise on the efficacy of any type of BMP.

A publicly available web application is provided at https://sccwrp.shinyapps.io/bmp_wq_index_app/ to enable users to calculate performance index scores and generate graphics and tabular outputs from user-supplied BMP monitoring data. Additional information about the web application is provided in the Supplemental Materials.

3. Results

3.1. Conceptual index design

The PI concept is structured to answer the fundamental question of whether a BMP contributes to achieving receiving water quality goals. This management question is translated into three questions that can be answered with quantitative data (Fig. 1): (1) are pollutants being removed by the BMP? (2) what are the treated water quality conditions compared with the desired outcomes? and (3) do operating conditions impact pollutant removal? Operating conditions herein refer to whether the BMP is subjected to high or low concentration runoff. The data required to answer these study questions are paired influent-effluent EMCs sampled from a treatment BMP for multiple storm events and a user-specified threshold concentration that is used to benchmark what is "clean" or "dirty" runoff for the watershed of interest.

The index's structure is built by considering three distinct comparisons as follows:

a) The influent water quality conditions for each storm are evaluated with respect to a threshold (the normalized influent EMC),
b) The effluent water quality conditions for each storm are evaluated with respect to a threshold (the normalized effluent EMC), and
c) Influent versus effluent EMC - The influent and effluent water quality conditions are compared to each other.

The normalized EMC comparisons are dissected in Figs. 2a-c, culminating in a data visualization identifying regions that emerge from the superposition of these dividing lines (Fig. 2d). A vertical line drawn

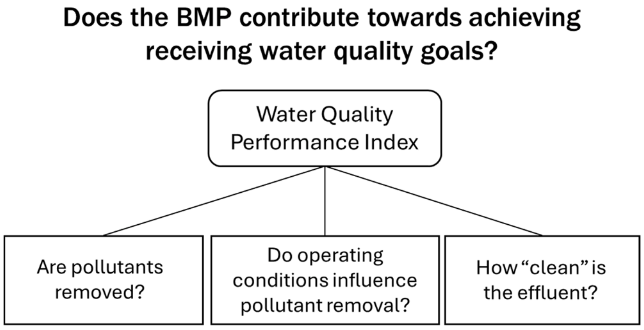

Does the BMP contribute towards achieving receiving water quality goals?
Water Quality Performance Index
Are pollutants removed?
Do operating conditions influence pollutant removal?
How "clean" is the effluent?

Fig. 1. The water quality performance index framework brings together management questions that can be answered with quantitative monitoring data.

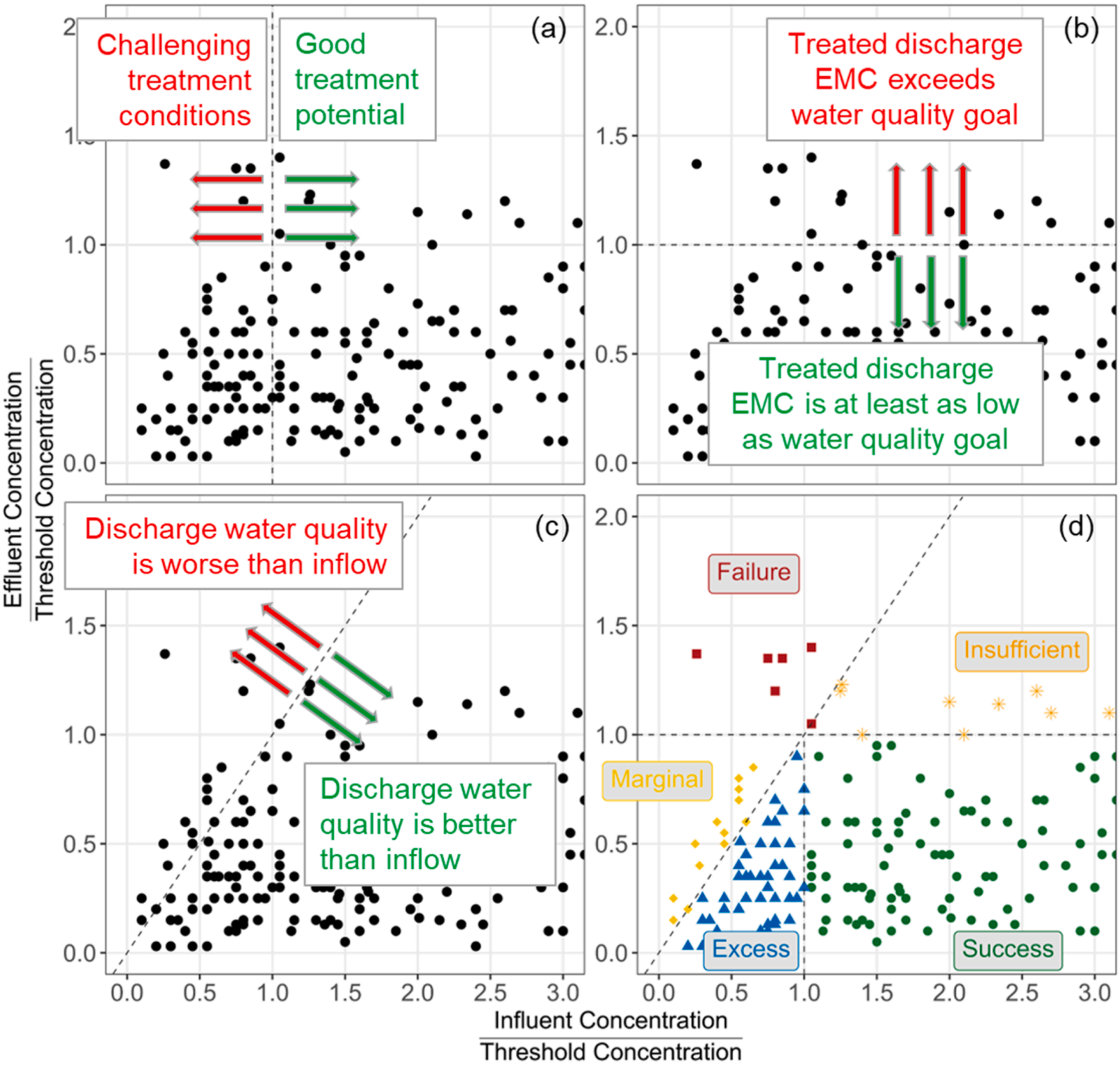


Fig. 2. BMP Performance Index framework using total suspended solids (TSS) for 18 bioretention BMPs selected from the BMP Database: (a) Does influent require treatment to meet water quality thresholds?; (b) Does effluent exceed a water quality threshold?; (c) Are pollutants removed or exported?; and (d) Translating numerical data into narrative outcomes. $N = 309$ total site-events. 69 Success data points and 10 Insufficient data points not shown to maintain figure clarity.

at a normalized influent EMC equal to 1.0 demarks measured inflows that are considered "dirty" to the right, or "clean" to the left (Fig. 2a), for the specific pollutant and receiving water quality conditions. A wide range of normalized influent EMCs indicates a BMP experiences highly variable pollutant concentrations, which is interpreted as a wide range of operating conditions, whereas a narrow range of influent EMCs indicates the pollutant concentrations entering the BMP are relatively stable and may provide limited information when accessing the performance consistency or capability. Data points to the far right might be considered extreme events in terms of influent water quality, such as the seasonal first flush typically experienced in Southern California after usually ~6 months without rainfall. Data points to the left of the vertical line demark already "clean" runoff that may not strictly have needed treatment (for that pollutant and watershed objective), or may be indicative of runoff in temperate climates subject to more frequent rainfall and pollutant wash-off. Literature suggests that passively operated BMPs may be subject to a limit of treatment, and that treatment of higher pollutant concentrations is "easier" to achieve (Barrett 2005; Jones et al. 2005; Strecker et al. 2001).

The horizontal line drawn at normalized effluent EMC = 1.0 evaluates how close the BMP performance is to achieving the desired treatment objectives (Fig. 2b). Values above the line in Fig. 1b indicate that additional treatment may be needed to meet the desired downstream water quality objectives. Values below the line suggest that the effluent water quality conditions meet downstream water quality goals.

A 1:1 line imposed over the plot indicates the paired event data disaggregated by whether any treatment is achieved at all, i.e., the effluent EMC is less or greater than the influent EMC (Fig. 1c). Unfortunately, some BMPs have been shown to export pollutants (Clary et al. 2020; Zhou et al. 2024), clearly a performance phenomenon stormwater managers seek to avoid. Pollutant export may occur because of poor design, materials specifications, construction errors, maintenance conditions, unexpected site conditions, changes in the drainage area, or a range of other factors.

The regions emerging from superposition of the dividing lines in Figs. 2a-c bin quantitative monitoring data into categorical performance interpretations (Fig. 2d). The BMP Water Quality Performance Index interprets five descriptive categories as follows in order of their managerial preference:

- Success – The most preferred outcome; influent EMC exceeds the water quality threshold, but the treated effluent EMC is below the water quality threshold.
- Excess – Influent EMC is already "cleaner" than the water quality threshold; additional treatment is obtained but not strictly needed.

- Marginal – Influent EMC is already below the water quality threshold; exported pollutants degrade quality but not enough to exceed the water quality threshold.
- Insufficient – Influent EMC exceeds the water quality threshold and some treatment occurs to reduce effluent EMC, but additional treatment is needed to meet water quality objectives.
- Failure – Regardless of influent water quality, the BMP exports pollutants at effluent concentrations greater than water quality threshold, potentially exacerbating downstream water quality conditions. Corrective action is warranted.

Data points falling on dividing lines are assigned to the more protective category. Data pairs categorized as marginal or excess may incorporate relatively more noise or uncertainty associated with detection of low concentrations.

Five typically unequal sized quintiles emerge from the dividing lines, thus converting scattered quantitative data into coherent, categorical determinations of performance benchmarked by a receiving water quality objective. The categories can be interpreted by the likelihood of a corrective action being taken; interpretations are presented in the next section. The water quality threshold is user-defined for each site or receiving water and pollutant such that many BMPs, pollutants, and storm events can be represented on the same axis and their categorical determinations compared directly, thereby satisfying the index development criteria of establishing a transferable metric.

3.2. *Interpreting index scores for management recommendations*

3.2.1. *Data consolidation strategy*

The data sets investigated herein from the BMP Database suggest that it would be unlikely for all data to fall within a single category for a single BMP. To this end, watershed managers need a way to distill the narrative behavior of a suite of monitored events into a single value that can be tracked over time or compared between BMP datasets. A weighted average ($PI_{analyte}$) was developed to consolidate the outcomes for an individual analyte into a single descriptor of performance.

The weighted average is the sum of the proportion of data in each category multiplied by the weighting factor of that category as in Eq. (1).

$$PI_{analyte} = \sum_{i=success}^{failure} w_i X_i \tag{1}$$

where $PI_{analyte}$ is the weighted average PI value for the indicated analyte, w_i is the weighting factor for each of the five categories, i, the value of which is associated with the managerial preference of that outcome. X_i is the proportion of data in each category (i.e. the number of data points in each category divided by the total number of data points). The set of category weighting factors is referred to as the weighting scheme. The weighting scheme is critical to support several of the index development criteria, including: (a) is transparent and quantitative, (b) reflects the underlying distribution of the monitoring data, (c) heavily weights the failure conditions that conclusively identify poor BMP performance, and (d) contributes to clear decision making.

A weighting scheme that satisfies the criteria and reflects the underlying managerial preference of resource allocation or follow-up action emerged establishing values for w_i as Success = 0, Excess = 1, Marginal = 3, Insufficient = 4, Failure = 10. Success is defined as a BMP that meets the performance expectations and does not warrant further action. The success of a BMP might serve as an example for design engineers of what to do in the future. Excess is defined as a BMP that outperforms the project's needs. Excess BMPs are not substantively different from success BMPs in the eyes of a manager (neither requires intervention), so they receive similar scores, but a distinction is preserved for post-hoc technical investigation. Marginal and Insufficient BMPs are acknowledged to be worse outcomes than Success or Excess; however, the Insufficient category outcomes are considered more deserving of additional resource allocation to address the elevated pollutants discharged in order to meet the downstream needs. BMPs exhibiting water quality outcomes in the Failure condition received a high penalty in the weighting scheme to call clear attention to unacceptable performance triggering a high priority for corrective action, and that likely requires additional resource allocation. The quantitative distinctions between $PI_{analyte}$ scores enabling clear interpretation for action are an important outcome of the weighting scheme that would not be discerned with assigning sequential integer values as weights (e.g., Success = 1, Excess = 2, Marginal = 3, Insufficient = 4, Failure = 5).

3.2.2. *Score interpretation*

Table 1 breaks down the interpretation of the $PI_{analyte}$ scores, and recommends follow-up actions and queries. A quick assessment for high-level distinction emerges with the left-most column. BMPs are doing what they're supposed to do (Score 0–2.0), further investigation is needed to confirm if there is an existing problem (2.0–5.0) or if the BMP needs corrective action (> 5.0). Additional lines of inquiry for technical designers, engineers, compliance officers or maintenance crews are further broken out and are discussed in more detail herein. Transition points in the performance index correspond to categorical outcome distributions that are ≥ 50 % in a less protective category.

Scores in the "no action" zone (0–2.0) indicate the set of BMPs monitored are performing well as a whole. There are no recommendations for remedial action in this range, but technical personnel may wish to investigate the PI further. A $PI_{analyte}$ score of 0.0 – 0.5 indicates a high prevalence of success and excess outcomes, suggesting the design, implementation, and maintenance condition of the BMPs are good examples for future replication. A score in the 0.5 – 1.0 range may hint that upstream runoff is not a significant source of the pollutant of concern; most of the data will be in the excess or marginal conditions. Consistently low measured runoff EMCs might suggest that better siting for the water quality BMP is warranted. Scores in the 1.0 – 2.0 range indicate some degrees of marginal or insufficient treatment; managers should watch this set of BMPs for a declining trend.

$PI_{analyte}$ scores in the "query data" zone (2.0–5.0) constitute BMP sets with ambiguous or mixed outcomes that require additional investigation to discern management implications. Scores in the range 2.0 – 3.5 should be investigated for a high prevalence of insufficient category site-events. If more than half of the dataset falls into the insufficient category, additional treatment or design modification may be needed to remove the specific pollutant from water. Additional treatment may be in the form of introducing the appropriate treatment mechanism for the pollutant type via a treatment train or constructing additional BMPs elsewhere in the watershed. Alternatively, a score of 2.0–3.5 might be achieved by mostly good effluent quality but a handful of fails, which

Table 1
Interpretation of the Water Quality Performance Index Score.

$PI_{analyte}$ Score	Suggested Actions & Data Queries	
0–2.0	No action needed.	
	0–0.5	BMP is successfully achieving treatment goals.
	0.5–1.0	Untreated runoff is not an important source for the pollutant of interest (for the storms monitored)
	1.0–2.0	Majority of data are "clean" effluent. Watch for a declining trend.
2.0–5.0	Query data: Are all/most of data of concern from a single BMP?	
	2.0–3.5	Does "insufficient" exceed 50 % of data? YES: Additional treatment needed. NO: Score is due to mostly marginal and/or a few fails. If failures are present, are there site- or storm-specific reasons?
	3.5–5.0	Additional treatment needed. Inspect maintenance condition. Investigate design/construction. ≥ 50 % of data are insufficient or significant % of failures.
5.0–10.0	Intervention needed. ≥ 50 % of data are failures.	

may not obviously point to any remedial action.

Scores in the 3.5 – 5.0 range indicate that additional treatment and perhaps BMP remediation are needed. Additional investigation into the prevalence of insufficient and failure condition BMPs is required to determine *where* the additional treatment would be most beneficial. The $PI_{analyte}$ is calculated for all site-events available in the dataset, so the easiest query is to determine if most/all of the data of concern (i.e., insufficient and failure categories) are from a single BMP. If a single BMP contains many of the poorly performing site-events, the $PI_{analyte}$ score for that slice of data will be correspondingly worse and fall into a different range of suggested actions.

$PI_{analyte}$ scores greater than 5.0 indicate a significant fraction of the data set is failures. BMPs that fall into this performance category are consistently failing to meet water quality objectives while not resolving and perhaps actively contributing to the downstream water quality problem. Corrective or restorative maintenance activities are likely necessary. Consistent failure may also identify persistent design or construction issues. As-built status and construction procedures should be reviewed.

3.2.3. *Application: comparing amongst BMPs of a similar type*

It is useful to evaluate data from multiple BMPs of the same type in order to inform future BMP selection for watershed-specific concerns. The same investigation might also help identify common design features that benefit or hinder performance. For the purposes of example, sample bioretention data were investigated for total copper and TSS performance.

The available data set for total copper comes from 11 BMPs with a total of 229 paired influent-effluent EMCs. The data yield a PI_{Cu} score of 2.35 across all 11 BMPs when an arbitrary threshold value of 15 µg/L is applied (Table 2). The threshold was selected to demonstrate the example application. The management interpretation instructs the user to query the data (Table 1), upon which breaking down the calculated results reveals that a single BMP is responsible for the majority of failure data points (anonymized as 'BMP_A' in Table 2). If the site and its information were accessible to the authors, the BMP could be investigated for sources of copper in-situ, or the media's capacity for sorption might be limited or near exhaustion. Potential contamination from components of the BMP might be revealed from design metadata or construction drawings and notes.

The remaining 10 BMPs yield a composite PI_{Cu} score of 2.02, which translates to "watch for a declining trend".

3.2.4. *Interpretating a $PI_{analyte}$ compared to a percent-removal*

Sample bioretention data were investigated for TSS performance. The average PI_{TSS} score with threshold = 20 mg/L is 1.31 across the 18 BMPs and 309 data points considered herein (Table 3), giving a clear indication that biofiltration type BMPs provide successful treatment of TSS. All but three of the BMPs considered have individual scores <2.0. To the contrary, these same BMPs yield an average EMC reduction of 62 ± 66 %, which is arguably interpreted as inconsistent performance across the BMP type. Individual BMP performance ranges 2–91 % reduction, suggesting that some bioretention systems settle and filter, while others do not. The BMP with the average 2 % percent reduction ('BMP_C', Table 4) suffers from the undue influence of a single storm where a "clean" influent (5.2 mg/L) increased to 27.4 mg/L in the effluent – a 427 % increase in concentration for an effluent quality that is reasonably close to the example threshold of 20 mg/L. In comparison, BMP_C yields a PI_{TSS} score of 1.57 at a threshold of 20 mg/L, with only 1 failing data point out of the total of 7 monitored storms. In any or all cases, the percent removal metric fails to inform managers whether any of the BMPs considered (as a group or individually) will provide meaningful advances towards minimizing TSS downstream, whereas the PI_{TSS} score suggests that they will.

Table 3
Relatively consistent PI_{TSS} scores demonstrates the success of 18 bioretention BMPs in achieving effluent EMCs ≤ 20 mg/L, compared to the widely varying %-removal metrics for those same BMPs.

Anonymized BMP Name	# EMC Pairs	Average % Removal ± St. Deviation	PI_{TSS} with threshold = 20 mg/L
BMP_A	74	85 ± 53	1.11
BMP_B	30	17 ± 52	2.67
BMP_C	7	2 ± 192	1.57
BMP_D	19	78 ± 15	0.32
BMP_E	25	72 ± 18	0.44
BMP_F	18	41 ± 61	1.17
BMP_G	18	53 ± 58	1.06
BMP_H	8	81 ± 22	0.38
BMP_I	8	65 ± 32	1.12
BMP_J	7	65 ± 74	1.0
BMP_K	10	56 ± 40	1.2
BMP_L	23	42 ± 97	1.74
BMP_M	23	57 ± 104	1.43
BMP_N	4	91 ± 10	3.0
BMP_O	14	87 ± 18	0.21
BMP_P	7	70 ± 35	1.14
BMP_Q	6	54 ± 80	2.5
BMP_R	8	66 ± 30	1.5
All	309 (Total)	62 ± 66 (Average)	1.31 (Average)

Table 4
Example BMP from the BMP Database, BMP_C, with seven monitored storm events. The TSS threshold is 20 mg/L.

Event ID	Influent EMC (mg/L)	Effluent EMC (mg/L)	Percent Removal (%)	Category
1	27.1	2.5	91	Success
2	40.2	3.3	92	Success
3	29.1	5.4	81	Success
4	44.1	5.6	87	Success
5	33.3	5.6	83	Success
6	11.2	10.2	9	Excess
7	5.2	27.4	−427	Failure
Average ± St. Dev	**27.2 ± 14.3**	**8.6 ± 8.7**	**2 ± 192 %**	**PI_{TSS} 1.57**

3.2.5. *Application: evaluating design alternatives*

The Shop Creek treatment train provides a beneficial example of how the water quality PI can be used to evaluate the impact of design alternatives. Index scores for TP and TSS are examined as TP is the primary pollutant of concern for the downstream Cherry Creek Reservoir (Urbonas et al. 1993), while TSS removal should be enhanced through the introduction of filtration in the littoral zone (Urbonas et al., 1994). Arbitrary thresholds of 15 mg/L TSS and 0.2 mg/L of TP are adopted in the absence of published watershed-specific goals, acknowledging that high quality raw water is beneficial for water supply.

Table 2
Querying the index breakdown to investigate potential sources of concern for total copper treatment with threshold = 15 µg/L.

BMP	Threshold	PI_{Cu}	# Success	# Excess	# Marginal	# Insufficient	# Failure
All Bioretention*	15 µg/L	2.35	57 (25 %)	77 (34 %)	57 (25 %)	15 (7 %)	23 (10 %)
All Bioretention except BMP_A	15 µg/L	2.02	26 (17 %)	66 (43 %)	49 (32 %)	4 (3 %)	8 (5 %)
BMP_A	15 µg/L	3.01	31 (41 %)	11 (14 %)	8 (11 %)	11 (14 %)	15 (20 %)

* 11 BMPs with a total of 229 paired influent-effluent EMCs.

The similar data distributions in Fig. 3 demonstrates that influent pollutant EMCs were comparable among the two monitoring periods, implying that the BMP was subjected to similar operating conditions and performance is justifiably compared. Most storms generated influent EMCs exceeding the water quality thresholds arbitrarily adopted for this example.

Field monitoring data between 1990 and 1994 yields $PI_{analyte}$ scores in the "query data" range for the treatment train, mostly due to data points in the insufficient category for TSS and failures for TP (Fig. 3). The treatment train was removing some TSS, but additional treatment would be necessary to meet the hypothetical water quality objectives. On the other hand, two out of three data points not meeting TP performance objectives are in the failure category, creating a warning to watch for a declining trend. TP and TSS performance index scores shift into the category of "no action needed" for the 1995–97 monitoring period after the pond was regraded, indicating the benefits of physical filtration through the vegetated zone and successful implementation of the design modification.

3.3. Sensitivity to benchmarks

Bioretention data for TSS (18 BMPs, $n = 309$ data pairs) are used to demonstrate the sensitivity of criteria that define the PI categories . A threshold of 20 mg/L (Fig. 4a) is arbitrarily selected as a reasonable benchmark for "clean" runoff as it has been identified as the lower limit of settling without the aid of a coagulant in stormwater treatment practices (Schueler 1996), although it is noted that Barrett et al. (2004) suggests a value of 25 mg/L. A more stringent threshold of 10 mg/L (Fig. 4b) is selected for comparison; it is the median effluent TSS EMC amongst 41 BMPs and 685 data points according to a statistical summary of all bioretention in the BMP database (Clary et al. 2020).

The 1:1 line emerges as the dominant reference point, as this relationship is fixed for all monitoring pairs. Changing the threshold has no impact on monitoring data relative to the 1:1 line. More stringent thresholds shift the horizontal and vertical category boundaries down and to the left along the 1:1 line, but data points shift only within categories on the same side of the 1:1 line. An increasingly stringent TSS threshold results in data moving from marginal to failing (Fig. 2a to b), or success to insufficient and from excess to success (Fig. 2b to c).

The bioretention PI_{TSS} score for the 18 BMPs considered shifts from 1.24 to 2.42 when the water quality threshold switches from 20 mg/L (Fig. 2a) to 10 mg/L (Fig. 2c). The threshold shift results from a "do nothing" score interpretation to triggering an additional query of the data. This investigation reveals that one of the 18 bioretention systems ('BMP_B') has an individual PI_{TSS} score of 5.07 at a threshold of 10 mg/L (50 % insufficient and 37 % failure), representing 30 of the 309 data points (at threshold = 20 mg/L, PI_{TSS} for this BMP is 2.67 [Table 3]). In comparison, the BMP with the largest individual data set ('BMP_A') (74/309 EMC pairs) yields a PI_{TSS} score of 1.74 for a threshold of 10 mg/L. The analysis suggests that overall, the bioretention systems considered are effective at achieving arguably quite good TSS quality. A field investigation of BMP_B is warranted for determining the appropriate corrective action, or whether there is a design concern.

3.4. Integrating performance for multiple pollutants

Water quality BMPs are typically expected to treat a plethora of contaminants of concern within stormwater. Stormwater managers may be interested in evaluating the range of pollutants that are relevant to specific local compliance targets or water quality objectives (e.g., a watershed has a TMDL for TP, but only general concerns for TSS and Cu). Up until this point, the PI analysis has focused on the performance of a BMP with regards to a single pollutant at a time, however a multi-pollutant composite index (MPI) score may also be warranted. An MPI

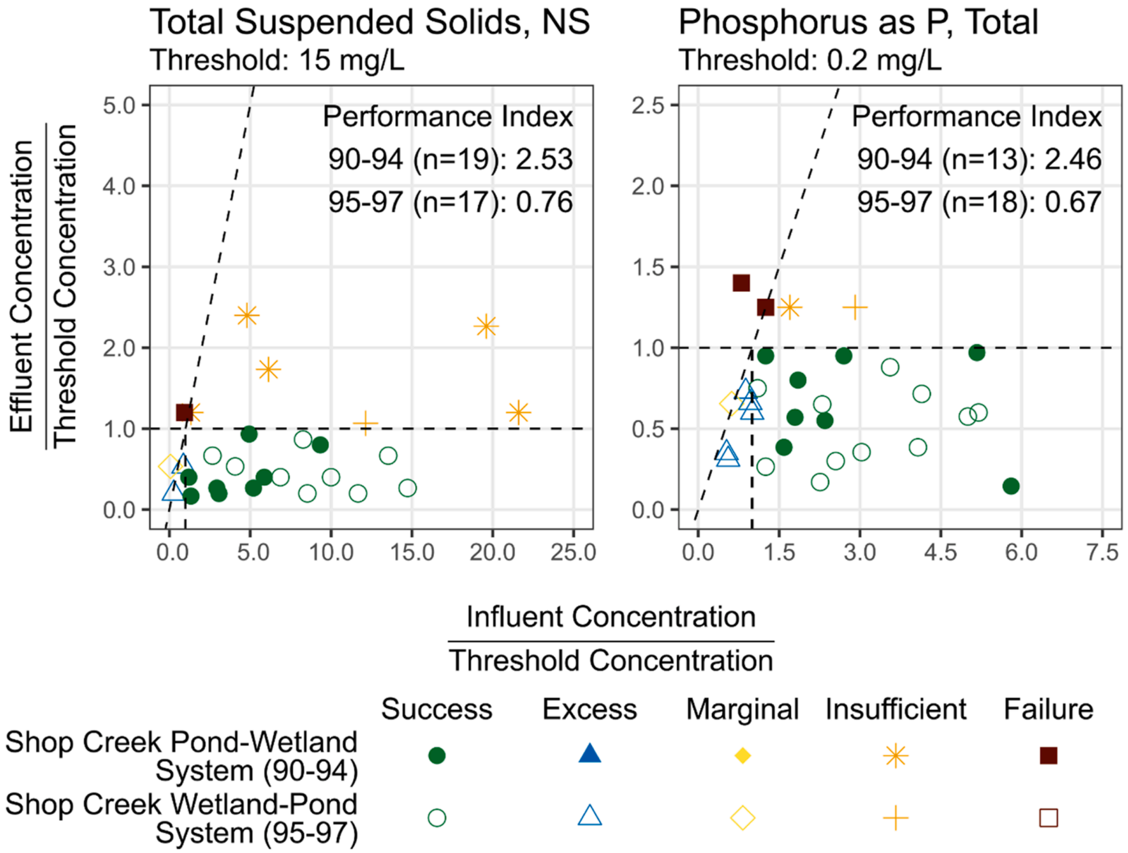


Fig. 3. Categorical performance designations for the Shop Creek pond and wetland treatment train using hypothetical water quality objectives.

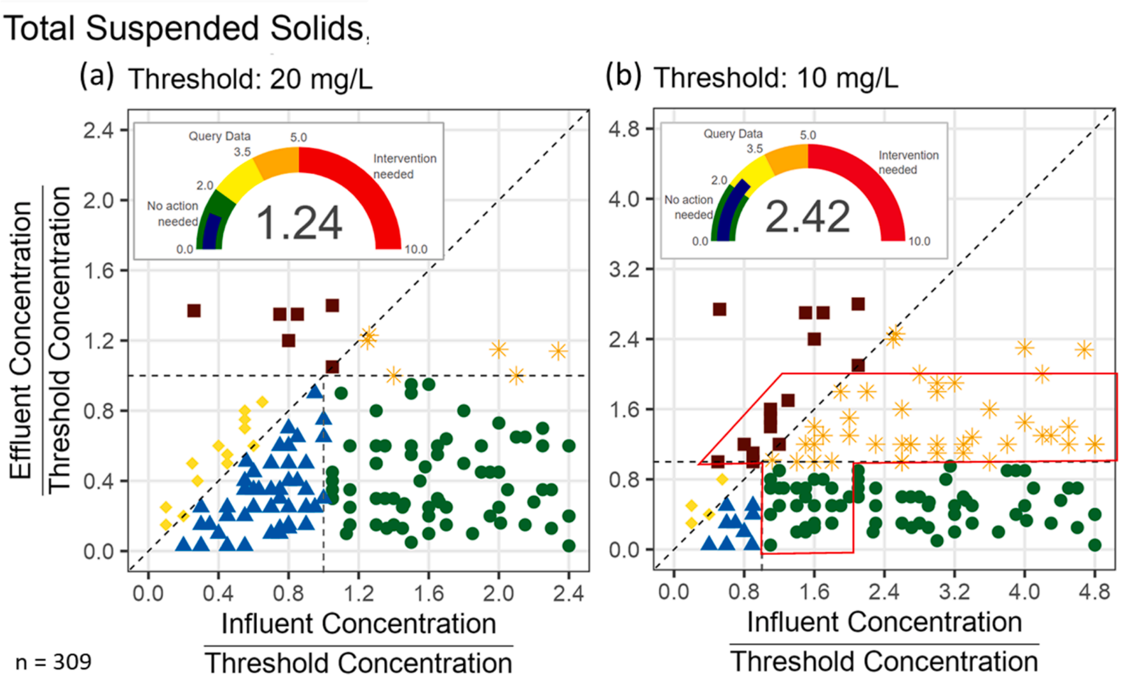


Fig. 4. Decreasing the threshold value for benchmarking performance from 20 mg/L to 10 mg/L increases the PI_{TSS} score from (a) 1.24 to (b) 2.42 for the sample bioretention BMPs representing 309 paired influent-effluent EMCs. Data points highlighted in red illustrate the majority changes in categorical outcomes. The plots have been truncated for figure clarity; additional data points are not shown to the right, but are incorporated into the performance index scores shown in the dials.

should be sensitive to user-defined preferences and priorities for water quality treatment.

To quantify the performance of a BMP across a suite of pollutants in an MPI, the analytical hierarchy process (AHP) for organizing and assessing complex options (Saaty, 2008) is adapted. The AHP method requires relational preferences between each option as inputs, e.g. is TSS mitigation from wet weather runoff a higher priority than Cu? How much more? Users define those relational preferences by placing pollutants into prioritization "bins". Herein, four rank prioritization bins are suggested, to maintain consistency of a clear overall MPI score with the interpretations for $PI_{analyte}$ scores, and to prevent "splitting hairs" amongst potentially a large range of parameters under consideration.

The prioritization bins are designated as High, Medium, Low and Not Important (i.e. ignore), with narrative interpretations in Table 5.

The AHP rank values are unique to each bin (Table 5). The AHP algorithm (Griffith, 2021) solves for the AHP bin weighting factor, r_j (not to be confused with the single parameter index category weight, w_i in Eq. (1)), based on the prioritization rank and number of pollutants considered according to Eq. (2).

Table 5
Prioritization "bins" for ranking pollutants into an MPI score using the AHP method.

Parameter Priority	Interpretation	AHP Rank
High	Parameters identified as high priority are the primary decision-makers for stormwater management solutions in the watershed. Examples of high priority might include analytes for which there is an associated TMDL, beneficial use of interest, or other public concern.	1
Medium	Parameters for which there is general concern, for example where objectives for management are identified in an MS4 permit or watershed management plan.	3
Low	Parameters of interest, but unlikely to feature as a concern in watershed management plans or public interest.	7
Not Important	Parameters for which data exists but are not currently relevant for MS4 permit requirements or watershed management plans.	∞

$$r_j = \frac{1/Rank_j}{\sum_{j=1}^{N}(1/Rank_j)} \tag{2}$$

Where the value of $Rank_j$ is as per Table 5. Eq. (2) reveals that the AHP bin weighting factor is the inverse of the priority rank normalized by the sum of the rank inverses. The sum of AHP bin weighting factors is equal to one. The value of r_j decreases with unequal increments between bins. The ratio of AHP bin weights between ranks for a prioritized parameter list is preserved according to Eq (2) and Table 5, i.e., the weight of pollutants with Rank = High will always be 3x the weight of pollutants with Rank = Medium and 7x the weight of pollutants with Rank = Low. The AHP bin weight for pollutants with Rank = Not Important is always zero.

The algorithm outputs coefficients for a weighted-average MPI score defined by Eq. (3).

$$MPI = \sum_{j=TSS}^{N} r_j * PI_j \tag{3}$$

Where the $PI_{analyte}$ score for each analyte, j, is multiplied by the AHP bin weighting factor, r_j. The result is summed for all relevant pollutants specified by the user.

The AHP method offers a dynamic approach to determining weighting factors, which enables user customization. The top pollutant (s) of concern for a watershed drive(s) the MPI score. Any number of pollutants may be identified in each bin (high, medium, low, ignore). AHP bin weighting factor, r_j, is equal for all parameters sharing bin j. As the number of parameters increases, the difference in the value of r_j between bins decreases, though the ratio of r between bins stays constant. Where all water quality concerns are equal, the AHP returns equal weighting factors and results in an arithmetic average of $PI_{analyte}$ scores for the MPI.

3.4.1. *Application: planning for TMDL objectives*

Many individual watersheds and/or waterbodies in California are subject to multiple total maximum daily loads (TMDLs) (https://www.waterboards.ca.gov/water_issues/programs/tmdl/index.html). TMDLs emerge from a program in the USA that establish wasteload allocations on a watershed basis with the intention of restoring degraded receiving waters. TMDLs are pollutant- and watershed-specific according to the designated "beneficial uses" (e.g. contact recreation, non-contact recreation, water supply, etc.). Numerical water quality standards to support beneficial uses are established typically by the US Environmental Protection Agency (US EPA). The MS4 stormwater discharge permits in the USA also often typically incorporate "general" concerns for runoff water quality such as TSS, nutrients, and heavy metals. The MPI offers a tool for selecting the "right" BMPs in support of achieving multiple water quality goals, such as a combination of watershed-specific TMDLs and general MS4 goals.

A hypothetical case of a waterbody with bacteria (Fecal coliform) and heavy metal (total zinc and copper) TMDLs is considered. Local MS4 planning (hypothetically) identifies treatment of TSS in stormwater as a priority, while nutrients (TN and TP) are a low concern. Bioretention is strongly being considered for widespread implementation partly because it fits into a range of public spaces while providing co-benefits such as urban heat island mitigation. The hypothetical permittee is more familiar with constructing detention basins, and is hesitant to adopt a new technology.

Fig. 5 illustrates the AHP bin weights, r_j, for the hypothetical case study. The weighted-average coefficients from the AHP method follow an exponential decay function where high-ranked pollutants were weighted 3x higher than medium-ranked pollutants, and 7x low-ranked pollutants, etc.

The MPI for the hypothetical prioritization using these weights is broken down in Table 6. Bioretention emerges as a better choice compared to detention basins to meet watershed water quality goals, considering only water quality treatment performance and according to the prioritization and data considered. Individual $PI_{analyte}$ scores are lower for all analytes considered. This is not surprising, given that bioretention potentially offers treatment mechanisms of settling, filtration and sorption via surface ponding followed by flow through porous media, whereas detention basins offer only settling often under dynamic flow conditions. The number of BMPs contributing to the analysis is similar, with the exception of TN, suggesting that the breadth of comparison is reasonably consistent between the BMPs. The least amount of data is available for Fecal coliform, one of the high priority analytes in this example. The lack of fecal indicator bacteria data is problematic across the industry (Clary et al., 2021), which points to an industry-wide challenge inhibiting data-informed decision making.

Table 6
MPI application comparing between different types of BMPs to best support planning for a hypothetical watershed.

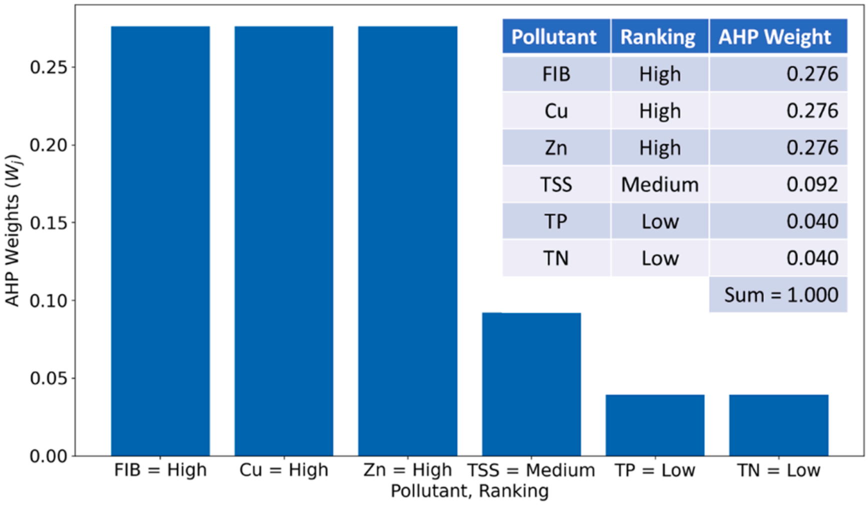
Analyte	Priority	Threshold	Detention Basins		Bioretention	
			n BMPs	$PI_{analyte}$	n BMPs	$PI_{analyte}$
Fecal coliform	High	200 cfu/100 mL	4	4.41	2	1.25
Cu (total)	High	15 μg/L	9	3.22	11	1.19
Zn (total)	High	15 μg/L	9	5.6	11	2.39
TSS	Med	20 mg/L	13	4.81	18	0.88
TP	Low	0.2 mg/L	11	3.86	18	2.59
TN	Low	1 mg/L	4	4.36	12	3.13
MPI				4.42		1.64

4. Discussion

Identifying the relative success of BMPs to treat stormwater contaminants is a critical aspect of stormwater planning, asset management, and capital improvement programs. The PI enables BMP water quality treatment effectiveness to be measured relative to site- or watershed-specific design intent or expectations, which is often a challenge when attempting to compare different types of BMPs for the same contaminant, or different contaminants for the same type of BMP.

A major advantage of the PI is the use of simple comparisons to synthesize complicated monitoring data into easily transmitted information to non-technical audiences including managers, policy decision makers, and the public. Converting quantitative monitoring data into performance categories derived from site-specific water quality goals eliminates subjective or arbitrary determinations of what might be considered acceptable performance, while consistently accounting for highly variable event-to-event performance reported in most studies, as was explored in Table 3. The ability to contextualize all monitored events into an unbiased performance assessment is a significant advancement over the historic use of %-removal metrics that subsequently require subjective judgement of "how much treatment is enough?", or any variation thereof (e.g. the TAPE program [WSDE 2024]). Likewise, the flexibility in accounting for a range of influent EMCs in the PI overcomes the limitation of the evaluation method proposed by Barrett (2005), which depends on specifying an influent concentration of interest and cannot incorporate outliers.

Selection of the water quality threshold that is used to benchmark performance is perhaps the most significant consideration for use. The PI



Pollutant	Ranking	AHP Weight
FIB	High	0.276
Cu	High	0.276
Zn	High	0.276
TSS	Medium	0.092
TP	Low	0.040
TN	Low	0.040
		Sum = 1.000

Fig. 5. Example MPI with user-defined rankings and AHP weights (from Eq. (2)).

is easily adapted to a range of water quality thresholds identified by the user. The thresholds should represent the primary water quality goals in the context or watershed where BMPs are being planned, or where they are already implemented. In most cases, the intent of BMP implementation is to protect or restore water quality at the watershed level. An "effective" water quality treatment BMP will produce effluent EMCs near or below the watershed water quality goal, because in urban watersheds, the discharge is highly likely to mix with additional untreated runoff once again elevating concentrations and loads before physically reach the receiving water (McNett et al. 2010).

Water quality thresholds are expected to vary from location to location, and potentially even at the same location between seasons. Indeed, McNett et al. (2010) acknowledge the need for regionally relevant benchmarks. The value of the threshold can be any relevant in-stream numerical water quality objective or standard, such as those associated with beneficial uses, or similar regulatory framework or authority. Elsewhere, thresholds for benchmarking effluent EMCs have been suggested for TN and TP in order to support benthic habitat quality in three ecoregions of North Carolina (McNett et al. 2010), for TSS (Barrett et al. 2004), or by comparing effluent EMCs against a reference catchment (Braswell et al. 2018).

If or where regulatory criteria or specific in-stream water quality objectives are absent, other selections might include comparison to "known" or established treatment potential, for example using a statistically representative value across a large data set, such as the example in this manuscript where threshold was set as the median effluent concentration for all of the BMPs of a given type in the BMP Database. The comparison therein can be interpreted as: "how do a single BMP or subset of BMPs perform relative to the population of BMPs (for which consistent data are available)?" Likewise, a more stringent threshold such as an irreducible concentration (Schueler, 1996a; Strecker et al. 2001) would be interpreted as "how do a single BMP or subset of BMPs perform relative to the best performing BMP?" A conservative threshold would reflect an analytical detection or reporting limit, in which case the interpretation would be: "what is the potential for the BMP to entirely remove the pollutant?" Applications where the water quality threshold varies by monitoring event such as hardness-adjusted trace metals or the biotic ligand model is feasible with the PI, and would require a straightforward data adjustment before applying the PI framework. The introduction of thresholds to benchmark performance by the authors does not imply that numerical effluent limitations should be adopted in a regulatory context. The index is intended as a tool for informing how well site- or regional-scale BMPs contribute to achieving watershed-scale receiving water goals so that appropriate action can be implemented, if needed.

A $PI_{analyte}$ or MPI score is the first step in evaluating performance (Table 1); however it is the follow-up actions from interpreting the score (e.g. Table 7) that offers the most potential for improving actual BMP performance and thereby improving receiving water quality. If a BMP fails to achieve a Success or Excess PI score, the user should not automatically assume the BMP is non-functional. Rather, the pollutant removal mechanisms offered by the BMP type should be considered when interpreting PI scores. For example, a BMP that only offers settling (e.g. a detention basin) should not be expected to achieve Success or Excess performance for dissolved contaminants. On the other hand, it would be beneficial to identify if/when a BMP or BMP type provides treatment beyond expectations. In this case, investigating the design of these exceptional BMPs might yield important information to advance the state of the practice. The components of interpretation and recommended pathway for further action are offered as significant advances over any other evaluation frameworks found in the literature. For example, in the five BMP monitoring studies in North Carolina that adopted the framework proposed by McNett et al. (2010), the evaluations are limited to determining the fraction of monitored events meeting "good" benthic water quality goals without any further interpretation. These studies fail to provide an interpretation of how often the benchmark should be met or exceeded to preserve benthic water quality. Likewise, Park and Roesner (2012) develop BMP pollutant load frequency curves with the intention to provide an uncertainty analysis in terms of the degree of TMDL compliance. The method leaves it to managers to interpret the extent of uncertainty that is acceptable, or if changes should be made to planning, design, or maintenance. Conversely, the PI enables users to examine if installing more or different BMPs in the watershed or introducing treatment trains might better achieve water quality goals.

Table 7
Suggested paths for follow-up investigations for "query data" $\mathrm{PI}_{\mathrm{analyte}}$ scores (2.0–5.0).

How many BMPs contribute to the $PI_{analyte}$?	Application	Investigation and Follow-Up for a Query Data Score
Multiple	Selecting the "right" BMP for a future installation.	If one BMP of the group is the cause of the $PI_{analyte}$ score, eliminate from group score. Opportunity to learn about a design or construction issue for that problematic BMP.
Multiple or Single	Is a newly constructed BMP "working"?	If the appropriate treatment mechanism is theoretically present in the BMP type, investigate component sources and as-built condition. • Is there excessive build-up of sediment, trash or debris in the pre-treatment area or ponding zone that needs to be removed? • Is scour occurring within the BMP? If so, introducing energy dissipation of influent and filling in scour paths may rectify the issue. • Does flow short-circuit the BMP? Short-circuiting due to overfilling of media, trash and debris build-up, or similar issues may be rectified with routine maintenance. If short-circuiting occurs because of the orientation of inlets and outlets, structural modifications may be necessary. • Have conditions changed in the contributing drainage area increasing pollutant load? This may require introducing additional pre-treatment, adding capacity to the BMP or additional downstream treatment. • Was the BMP constructed according to plan? Corrective maintenance may be required. • Were appropriate materials/components specified and supplied? For example, media composition in bioretention, unwashed gravel supporting underdrains or permeable pavement beds, have been reported to compromise performance.
Single	Maintenance monitoring	Trend analysis shows a seasonal anomaly. Likely a systematic condition, e.g., leaf accumulation in autumn. Adjust routine maintenance to target activities in advance of the season.
	Maintenance monitoring	Trend analysis shows a decline in annual $PI_{analyte}$ score. Restorative maintenance likely needed.

Tracking a PI score over time provides several insights. An improving temporal trend might indicate that a newly constructed BMP is successfully becoming established (a phenomenon particularly relevant to vegetated BMPs), while a declining trend would likely indicate an evolving maintenance concern, for example if/when filtration-type BMPs are reaching their capacity for pollutant removal. Likewise, optimum $PI_{analyte}$ scores should be restored after performing the appropriate maintenance. One-off type sampling could be used to provide a snapshot of performance against previously established expectations, if monitoring of that BMP had previously been performed or compared against a larger relevant data set (e.g. from the BMP Database or local repository), and thus give an indication of whether additional monitoring or other follow-up action is needed. The ability of the PI to track trends offers an advantage over other evaluation frameworks that rely on frequency analysis (Braswell et al. 2018; Page et al. 2015; Koryto et al. 2017; Luell et al. 2021; Park and Roesner 2012, Smolek et al., 2018, Wissler et al. 2020). In direct contrast, TAPE is a certification program that is pursued only prior to BMP installations (WSDE 2024).

A "query data" PI score (2.0–5.0) represents a more challenging range of interpretation, requiring some technical expertise for follow-up. There are a myriad of influences on BMP behavior. Influent-effluent monitoring data alone cannot provide the "answer" of what to do for a potentially problematic or declining BMP; however, the quantitative approach for combining the distribution of categorical outcomes into the $PI_{analyte}$ score was carefully designed to provide clear indicators of distinct performance differences. A few scenarios are considered in Table 7. A fundamental consideration for a poorly performing BMP or set of BMPs is whether that BMP type includes design elements promoting the appropriate pollutant removal mechanism for the pollutant of concern. If so, $PI_{analyte}$ scores exceeding 3.5 indicate field inspection (at a minimum) or possibly corrective actions are warranted. A review of BMP metadata (i.e., design data) and a visual assessment of maintenance conditions are recommended to identify causes and potential corrective actions to shift performance into the Success category.

The ability of the MPI to assess multiple pollutants according to watershed-specific priorities is essential. BMPs are rarely designed to treat a single pollutant. The MPI helps engineers to summarize BMP performance based on the device's intended use and treatment expectations. The outcomes provide engineers an in depth understanding of a treatment technology in the as-built condition and, thus, improve future project design for treating pollutants to the "maximum extent practicable". For on-going BMP monitoring programs, this type of assessment also provides value in evaluating whether a BMP implementation plan is likely to achieve long-term watershed remediation or restoration goals, or if mid-course corrections are in order. Of the other evaluation frameworks found in the literature, only McNett et al. (2010) and the TAPE program (WSDE 2024) consider multiple pollutants, but these are still more limited than the PI. McNett et al. (2010) is specific to TN and TP and interprets outcomes only in terms of benthic habitat. The TAPE protocol incorporates TSS, TP, and heavy metals in a rigid framework that does not adapt to other watershed specific concerns.

Monitoring data availability plays a critical role in performance assessment. The accuracy of the water quality PI is only as good as the quality of the monitoring data. For example, fecal indicator bacteria are the most common impairment for TMDLs across the USA; however, they are among the smallest data sets available in the BMP Database (Clary et al. 2020, 2021). The PI was developed with the intention of applying field-derived, flow-weighted, composite EMCs as these types of samples are thought to be the most representative average storm concentration and the best indicator of potential pollutant loads (Geosyntec and Wright Water Engineers 2009; Tiernan et al. 2024). Investments in additional high-quality monitoring generating data over multiple years are encouraged.

The PI structure isolates the treatment capability of a BMP by focusing on the comparison of EMCs between influent and effluent. Hydrologic performance is specifically excluded. A BMP that exfiltrates runoff may mask whether pollutant treatment actually occurs, due to the role of runoff volume in a pollutant load calculation. An exfiltrating BMP may increase effluent EMCs for a contaminant of concern (Clary et al. 2021) while simultaneously showing a decrease in the pollutant load discharging from the site. If the same BMP were configured with an underdrain and discharged runoff downstream, it would cause concern. An EMC-based index directly supports appropriate BMP selection in locations if/where exfiltration is infeasible or undesirable (e.g. soils with poor infiltration characteristics, proximity to buildings or sensitive buried infrastructure, etc.), as the treatment functions promoted by the BMP are of paramount importance.

5. Limitations and opportunities for future work

The $PI_{analyte}$ concept has been developed to evaluate BMP performance based on flow-weighted EMCs. If/where these data are not available or not able to be captured, water quality samples would not reflect a flow-weighted EMC. The scoring or interpretation of non-flow-weighted pollutant concentrations was not investigated. Flow-weighted EMCs are typically generated from either using automated sampling equipment that concurrently measures flow and collects sample aliquots at prescribed equal-flow intervals generating a single composite sample for laboratory analysis, or by collecting discrete samples (sometimes called grab samples) at multiple points along the hydrograph, and post-storm compositing (Tiernan et al. 2024). In either case, concurrent flow measurement (or modeling) and sample collection are required, which may also require substantial technical expertise. The influence of the number of data sets on the $PI_{analyte}$ score was not studied herein, but would be a useful investigation to inform resource allocation for monitoring programs.

This paper uses field monitoring data to demonstrate how the PI can be applied to compare performance across or between multiple BMPs and to evaluate BMP design aspects. It is anticipated that the index may also be used to evaluate changes in performance from the level of maintenance activity and overall device condition; insufficient data were found to examine this application in detail.

The research supporting this work designed and applied the PI using EMCs in order to isolate the pollutant treatment functions of a BMP. McNett et al. (2010) identify the importance of evaluating BMP hydrologic performance along with the water quality performance. With little effort, the PI can also be adapted to pollutant mass or hydrologic characteristics such as flow rates or volumes. As long as the measurements are made on influent and effluent, and there is an applicable threshold to benchmark "good" versus "bad", the BMP Performance Index approach can be used and categorical outcomes with narrative interpretations can be applied. The authors are continuing to work on expanding the BMP Performance Index specifically for hydrological applications.

6. Conclusions

The BMP Performance Index provides a crucial step in evaluating performance that was carefully structured to provide watershed managers, engineers, and maintenance crews clear pathways to determine how in-situ, post-construction structural stormwater control measures or BMPs are working for improving stormwater quality, and where design, construction, or maintenance improvements are needed to attain the overall watershed management goals. The $PI_{analyte}$ or MPI scores offer unbiased methods to organize and interpret the water quality treatment performance of a structural BMP and yields clear management actions based on as-built conditions. The index offers a quantitative performance evaluation method that significantly improves the utility of costly monitoring data over percent removal metrics. It can be applied to evaluate performance for a single analyte, or a suite of analytes and establishes a common basis for comparing amongst BMPs regardless of the treatment mechanism offered. Data generated by a range of real-world BMP monitoring studies are used to develop and explore the index, which benchmarks achievements towards specific water quality goals while considering storm-to-storm operating conditions. While existing public data sets were used to develop and explain the concepts of the index in this paper, watershed managers can apply the index framework to data from their own BMPs and analytes of interest using a publicly available web application (https://sccwrp.shinyapps.io/bmp_wq_index_app/). The index can be applied to a single BMP measured across one or more storms, to compare multiple BMPs of the same type, or compare across BMPs of different types. The BMP performance index enables objective data driven assessments across a range of pollutant types, water quality goals, design specifications and/or implementation across a wide variety of landscape settings. It can be applied to a long-term data set for trend analysis or used to spot-check performance.

Funding sources

The investigation presented in this manuscript was funded by Los Angeles County Public Works.

CRediT authorship contribution statement

Elizabeth A. Fassman-Beck: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Edward D. Tiernan:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Ka Lun Cheng:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Kenneth C. Schiff:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the project Advisory Committee for their dedication, critical evaluation and input to the development of the BMP Performance Index, in alphabetical order: Chris Beegan (California State Water Resources Control Board), Richard Boon (Riverside County [CA] Flood Protection & Water Conservation District), Jane Clary (Wright Water Engineers, Inc. [CO]), Bhaskar Joshi (Caltrans), Annalisa Moe (Heal the Bay [CA]), LB Nye (Los Angeles Regional Water Quality Control Board), Bridget Wadzuk, Ph.D. (Villanova University [PA]). We also extend our appreciation for technical support at SCCWRP to Paul Smith, Robert Butler, Nicholas Lombardo, Duy Nguyen, and Emily Darrin.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2025.123769.

Data availability

Data will be made available on request.

References

Barrett, M.E., Lantin, A., Austrheim-Smith, S., 2004. Stormwater pollutant removal in roadside vegetated buffer strips. Transp. Res. Rec. 1890 (1), 129–140. https://doi.org/10.3141/1890-16.

Barrett, M.E., 2005. Performance comparison of structural stormwater best management practices. Water Environ. Res. 77 (1), 78–86. https://doi.org/10.2175/106143005×41654.

Braswell, A.S., Winston, R.J., Hunt, W.F., 2018. Hydrologic and water quality performance of permeable pavement with internal water storage over a clay soil in Durham, North Carolina. J. Environ. Manage. 224 (2018), 277–287. https://doi.org/10.1016/j.jenvman.2018.07.040.

Clean Water Act (1972). 33 US.C. §§ 1251-1387. Accessed online 24 April 2025 from https://www.govinfo.gov/content/pkg/USCODE-2017-title33/html/USCODE-2017-title33-chap26.htm.

Clary, J., Jones, J., Leisenring, M., Hobson, P., Strecker, E., 2020. International stormwater BMP database: 2020 summary statistics. Rep. Water Res. Foundat., Project 4968. Accessed online 07/10/2023 from. https://www.waterrf.org/system/files/resource/2020-11/DRPT-4968_0.pdf.

Clary, J., Ervin, J., Steets, B., Olson, C., 2021. Pathogens in urban stormwater systems: where are we now? J. Sustain. Water Built Environ. 8 (1). https://doi.org/10.1061/JSWBAY.0000969.

County of Orange (2024a). Bacterial TMDL wet weather reasonable assurance demonstration for South OC. Appendix F of the South Orange County Water Quality Improvement Plan. Accessed online 24 April 2025 at https://www.southocwqip.org/pages/bacteria-tmdl.

County of Orange (2024b) Santa ana region unified annual progress report program effectiveness assessment 2023-24 reporting period. Accessed online 25 April 2025 from https://ocerws.ocpublicworks.com/service-areas/oc-environmental-resources/oc-watersheds/regional-stormwater-program/north-oc-0.

Davis, A.P., Hunt, W.F., Traver, R.G., 2022. Green Stormwater Infrastructure Fundamentals and Design. John Wiley & Sons, Inc, Hoboken, NJ.

Fletcher, T.D., Shuster, W., Hunt, W.F., Ashley, R., Butler, D., Arthur, S., Trowsdale, S., Barraud, S., Semadeni-Davies, A., Bertrand-Krajewski, J.-L., Mikkelsen, P.S., Rivard, G., Uhl, M., Dagenais, D., Viklander, M., 2015. SUDS, LID, BMPs, WSUD and more – The evolution and application of terminology surrounding urban drainage. Urban Water J 12 (7), 525–542. https://doi.org/10.1080/1573062X.2014.916314.

Galavotti, H., 2016. Summary of State Post Construction Stormwater Standards. Office of Water, Office of Wastewater Management Water Permits Division.

Geosyntec Consultants and Wright Water Engineers. 2009. Urban stormwater BMP performance monitoring. Accessed online 02/14/2025 from https://bmpdatabase.org/s/2009MonitoringManualSingleFile.pdf.

Geosyntec Consultants, Wright Water Engineers, and Veneer Consulting, 2015. Transferability of post-construction stormwater quality BMP effectiveness studies. In: Report to the American Association of State Highway and Transportation Officials. NCHRP Project 25-25, Task 92, National Cooperative Highway Research Program, Transportation Research Board. Accessed online 04/28/2025 from. https://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=3721.

Gilliom, R.L., Bell, C.D., Hogue, T.S., McCray, J.E., 2020. Adequacy of linear models for estimating stormwater best management practice treatment performance. J. Sustain. Water Built Environ. 6 (4), 04020016. https://doi.org/10.1061/jswbay.0000921.

Griffith, P., 2021. https://github.com/PhilipGriffith/AHPy.

Jones, J.E., Earles, T.A., Fassman, E.A., Herricks, E.E., Urbonas, B., Clary, C., 2005. Urban storm-water regulations—Are impervious area limits a good idea? J. Environ. Eng. 131 (2), 176–179. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:2(176).

Koryto, K.M., Hunt, W.F., Page, J.L., 2017. Hydrologic and water quality performance of regenerative stormwater conveyance installed to stabilize an eroded outfall. Ecol. Eng. 108 (2017), 263–276.

Law, N.L., Fraley-McNeal, L., Cappiella, K., 2008. Monitoring to Demonstrate Environmental Results: Guidance to Develop Local Stormwater Monitoring Studies Using Six Example Study Designs. Center for Watershed Protection.

Luell, S.K., Winston, R.J., Hunt, W.F., 2021. Monitoring the water quality benefits of a triangular swale treating a highway runoff. J. Sustain. Water Built Environ. 7 (1), 05020004, 2021.

McNett, J.K., Hunt, W.F., Osborne, J.A., 2010. Establishing storm-water BMP evaluation metrics based upon ambient water quality associated with benthic macroinvertebrate populations. J. Environ. Eng. 136 (5), 535–541. https://doi.org/10.1061/_ASCE_EE.1943-7870.0000185, 2010.

Minnesota Pollution Control Agency, 2022a. Stormwater pollutant concentrations and event mean concentrations. Minnesota Stormwater Manual. MediaWiki. Accessed January 15, 2023. https://stormwater.pca.state.mn.us/index.php?title=Stormwater_pollutant_concentrations_and_event_mean_concentrations.

Minnesota Pollution Control Agency, 2022b. Stormwater pollutant removal, stormwater credits. Minnesota Stormwater Manual. MediaWiki. Accessed Feb. 15, 2025. https://stormwater.pca.state.mn.us/index.php?title=Stormwater_pollutant_removal,_stormwater_credits.

Moore, T.L., Rodak, C.M., Vogel, J.R., 2017. Urban stormwater characterization, control, and treatment. Water Environ. Res. 89, 1876–1927. https://doi.org/10.2175/106143017×15023776270692.

Moore, T.L., Rodak, C.M., Ahmed, F., Vogel, J.R., 2018. Urban stormwater characterization, Control and Treatment. Water Environ. Res. 90, 1821–1871. https://doi.org/10.2175/106143018×15289915807452.

Muthukrishnan, B.M., Selvakumar, A., Field, R., Sullivan, D.A., 2004. The use of best management practices (BMPs) in urban watersheds. U.S. Environmental Protection Agency, Washington, DC, 600/R-04/184 (NTIS PB2007-107266), 2005Accessed Feb 15, 2025 from. https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=539707&Lab=NRMRL.

New Jersey Department of Environmental Protection (DEP). 2004. New jersey stormwater best management practices manual. Chapter 4 Stormwater Pollutant Removal Criteria. Accessed online 02/14/2024 from https://dep.nj.gov/wp-content/uploads/stormwater/bmp/nj_swbmp_4-print.pdf.M.

Page, J.L., Winston, R.J., Hunt, W.F., 2015. Soils beneath suspended pavements: an opportunity for stormwater control and treatment. Ecol Eng 82 (2015), 40–48. https://doi.org/10.1016/j.ecoleng.2015.04.060.

Park, D., Roesner, L.A., 2012. Evaluation of pollutant loads from stormwater BMPs to receiving water using load frequency curves with uncertainty analysis. Water Res 46 (20), 6881–6890. https://doi.org/10.1016/j.watres.2012.04.023.

Pitt, R., Maestre, A., Clary, J. (2018). National Stormwater Quality Database (NSQD) Summary report version 4.02. Accessed 07/10/2023 from https://bmpdatabase.org/national-stormwater-quality-database.

Rodak, C.M., Moore, T.L., David, R., Jayakaran, A.D., Vogel, J.R., 2019. Urban stormwater characterization, control, and treatment. Water Environ. Res. 91, 1034–1060. https://doi.org/10.1002/wer.1173.

Rodak, C.M., Jayakaran, A.D., Moore, T.L., David, R., Rhodes, E.R., Vogel, J.R., 2020. Urban stormwater characterization, control, and treatment. Water Environ. Res. 92, 1552–1586. https://doi.org/10.1002/wer.1403.

Saaty, T.L., 2008. Decision making with the analytic hierarchy process. Int. J. Serv. Sci. 1 (1), 83. https://doi.org/10.1504/IJSSCI.2008.017590.

Schueler, Thomas, 1996a. Irreducible Pollutant Concentrations Discharged from Stormwater Practices, 2. Center for Watershed Protection, pp. 369–372. Technical Note #74. *Watershed Protection Techniques*.

Simpson, I.M., Winston, R.J., Brooker, M.R., 2022. Effects of land use, climate, and imperviousness on urban stormwater quality: a meta-analysis. Sci. Total Environ. 809 (2022), 152206. https://doi.org/10.1016/j.scitotenv.2021.152206.

Smolek, A.P., Anderson, A.R., Hunt, W.F., 2018. Hydrologic and water-quality evaluation of a rapid-flow biofiltration device. J. Environ. Eng. 144 (2), 05017010. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001275.

Snyder, T., Whipple, R., Moneda, J., 2020. County of San Diego BMP Design Manual. County of San Diego Department of Public Works.

Schueler, T.R., 1996b. Irreducible pollutant concentrations discharged from urban BMPs. Watershed Protect. Techniq. 2 (2), 369–372.

Strecker, E.W., Quigley, M.M., Urbonas, B.R., Jones, E.J., Clary, J.K., 2001. Determining urban storm water BMP effectiveness. J. Water Resour. Plann. Manage. 144–149, 127_3_.

Tiernan, E., Fassman-Beck, E., Lombardo, N., 2024. Effects of postprocessing decisions on flow weighted event mean concentrations. J. Sustain. Water Built Environ. (10), 3. https://doi.org/10.1061/JSWBAY.SWENG-552.

Urbonas, B., J. Carlson, and B. Vang. 1994. *Joint Pond-Wetland System in Colorado, USA*. An Internal Report of the Urban Drainage and Flood Control District.

Urbonas, B., 2003. Effectiveness of urban stormwater BMPs in semi-arid climates. paper presented at. In: the regional conference on Experience with Best Management Practices in Colorado. Available at https://mhfd.org/wp-content/uploads/2019/12/Effectiveness-of-BMPs-in-Semi-Arid-Climates.pdf.

Urbonas, B., Carlson, J., Vang, B., 1993. Performance of the shop creek joint pond-wetland system. Flood Hazard News 23 (1).

US EPA Environmental Financial Advisory Board. (2020). Evaluating stormwater infrastructure funding and financing. Online report accessed 8/27/2024 from https://www.epa.gov/sites/default/files/2020-04/documents/efab-evaluating_stormwater_infrastructure_funding_and_financing.pdf.

Vogel, J.R., Moore, T.L., 2016. Urban stormwater characterization, control, and treatment. Water Environ. Res. 88, 1918–1950. https://doi.org/10.2175/106143016X14696400495938.

Washington State Department of Ecology (WSDE). 2024. Technical guidance manual for evaluating emerging stormwater treatment technologies: technology Assessment protocol - ecology (TAPE). Accessed Feb. 15, 2025 from https://ecology.wa.gov/regulations-permits/guidance-technical-assistance/stormwater-permittee-guidance-resources/emerging-stormwater-treatment-technologies.

Wissler, A.D., Hunt, W.F., McLaughlin, R.A., 2020. Hydrologic and water quality performance of two aging and unmaintained dry detention basins receiving highway stormwater runoff. J. Environ. Manage. 255 (2020), 109853. https://doi.org/10.1016/j.jenvman.2019.109853.

Zhou, B., Parsons, C., Van Cappellen, P., 2024. Urban stormwater phosphorus export control: comparing traditional and low-impact development best management practices. Environ. Sci. Technol. 58 (26), 11376–11385. https://doi.org/10.1021/acs.est.4c01705. **2024**.