Review of ESSD-2025-743 “Riverine phosphorus gain and loss across the conterminous United States” by Wang et al.

Phosphate and TP loads were estimated at 963 and 2,317 stations across the CONUS. Loads were linked to NHD to apply topology routing and estimate riverine net gains and losses. This reviewer finds immediate value in the dataset and initial gain and loss estimates. Comparison of Loadest to WRTDS is value added. This dataset should serve as a foundation for other modeling and analyses. Well done! This reviewer has only minor comments for the authors to consider to perhaps further clarify the presentation.

Response: Thank you for the positive and valuable comments. Based on these suggestions, we have revised the manuscript and provided point-by-point responses below.

General comments and questions:

What is your time period? Was it recent 2015-2020? That info is super important towards interpretation. It should be stated in the figshare dataset description too. And perhaps even included in the title.

Response: Thank you for this important comment. Because estimating spatial patterns of riverine P gain and loss requires sufficient upstream-downstream connectivity, we used available observations from all time periods to maximize spatial coverage. As a result, the compiled dataset spans multiple decades, with phosphate data from 1952 to2022 and total phosphorus data from 1958 to2023.

We agree that the temporal extent is important for interpretation. We have now clarified this information in the manuscript and added a description of the temporal coverage in the Figshare dataset. We also emphasize that the estimates represent long-term, multi-decadal conditions.

Lines 75-77: We compiled 51,394 PO_4^{3-} (USGS parameter code 00650) concentration data from 963 hydrological stations (spanning from 1952 to 2022) and 285,675 TP (USGS parameter code 00665) observations from 2,317 hydrological stations (spanning from 1958 to 2023) across the CONUS from the Water Quality Portal (Read et al., 2017).

The Skinner and Maupin, 2019 is your dataset for point sources? This is a good comprehensive dataset. So that means point source influence was represented at the annual timescale? A reader may benefit from that knowledge.

Response: Thank you for this helpful comment. Yes, point-source phosphorus inputs were obtained from Skinner and Maupin (2019), which represent annual discharges for the year 2012.

We agree that explicitly stating the temporal resolution of this dataset is important for interpretation. Accordingly, we have clarified in the Methods section that point-source inputs are represented at an annual timescale.

Lines 83-85: For TP, the "Point-Source Nutrient Loads to Streams of the Conterminous United States" dataset provides estimated annual total point-source inputs during 2012 at the HUC12 level (Skinner and Maupin, 2019).

The discussion of general factors is helpful. Is there a focus to determine or assess change or trends?

Response: Thank you for this insightful suggestion. In response, we have incorporated a trend analysis of riverine P loads based on stations with sufficiently long records.

Specifically, we estimated trends using the Sen's slope method with the Mann-Kendall test, focusing on stations with more than 30 years of load estimates to ensure robustness. This resulted in 405 TP stations and 53 PO_4^{3-} stations included in the trend analysis.

Overall, decreasing trends dominate, accounting for 72% of TP stations and 58% of PO_4^{3-} stations. In contrast, increasing trends are only observed at a subset of stations, primarily located in the Mississippi River Basin.

We have added the Sen's slope and corresponding p-values as additional fields in the "Riverine PO_4^{3-}" and "Riverine TP" datasets, and included a description of the trend analysis in the Methods and Results sections.

Lines 137-141: In addition, we evaluated long-term trends in riverine P loads using the Sen's slope estimator in combination with the Mann-Kendall test (Hamed and Rao, 1998). To ensure robust trend detection, only monitoring stations with more than 30 years of load estimates were included. This resulted in 405 TP stations and 53 PO_4^{3-} stations used for trend analysis. The Sen's slope and corresponding p-values are provided in the dataset.

Lines 192-196: We further examined temporal trends in riverine P loads at stations with long-term records (>30 years). Decreasing trends were prevalent, accounting for 72% and 58% of TP and PO_4^{3-} stations, respectively. Stations with increasing TP trends were primarily located in the Mississippi River Basin (Fig. S6), suggesting regional differences in nutrient dynamics.

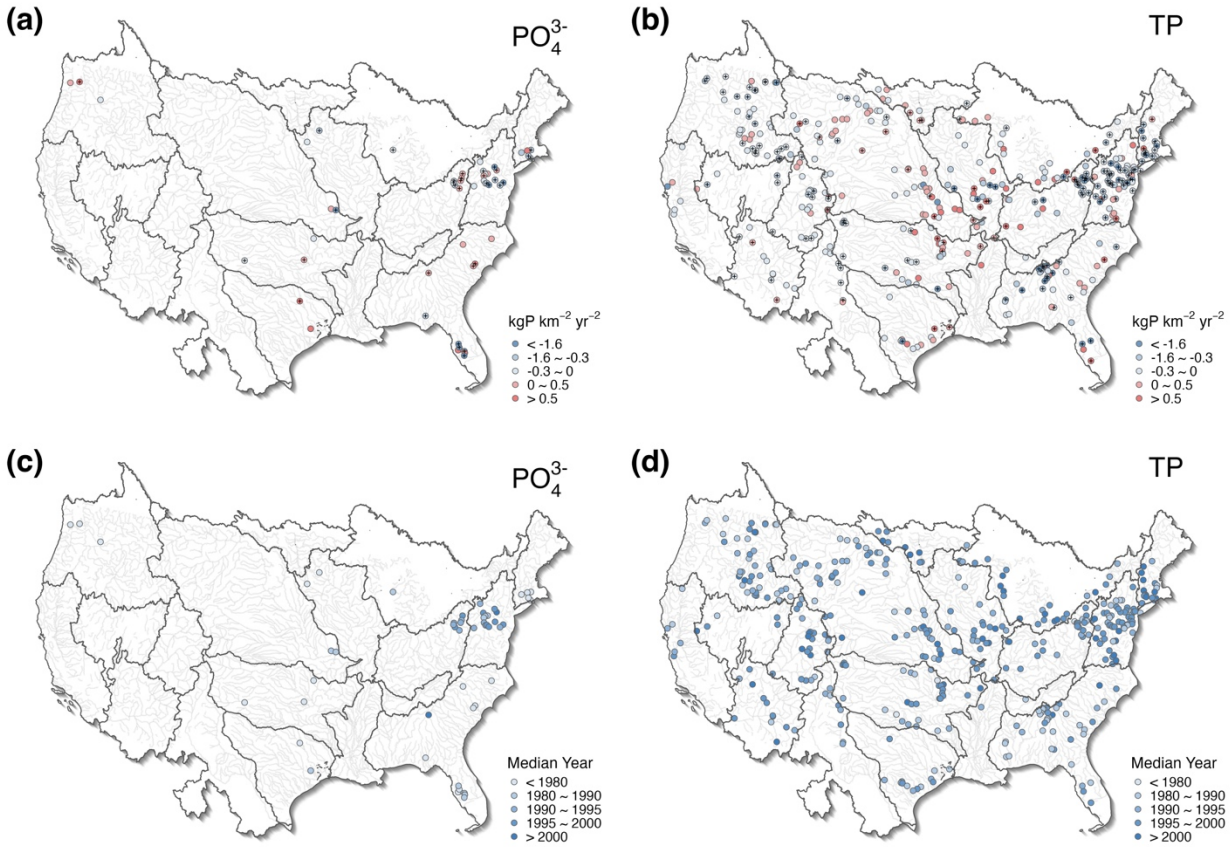

Figure S6: Spatial pattern of areal-normalized Sen's slope of riverine (a) PO_4^{3-} and (b) TP loads, and the median year of temporal coverage for (c) PO_4^{3-} and (d) TP at the monitoring stations. Crosses indicate stations with statistically significant trends ($p < 0.05$).

Results displayed by region is effective. Summarizing by HUC12 will also be immediately useful to others.

Response: Thank you for this helpful suggestion. As the current results are based on the HUC12 catchment map, we guess you mean "Summarizing by HUC2". To improve the accessibility and usability of the dataset, we have added regional summaries at the HUC2 level. Specifically, we now include area-weighted averages of TP and PO_4^{3-} gain and loss, as well as point-source and nonpoint-source contributions, for each HUC2 basin in the Supplementary Information.

In addition, we provided the spatial coverage of gain and loss estimates at the HUC2 level to help readers assess the reliability of results across regions. Corresponding descriptions have also been added to the Results section. These additions facilitate interpretation and make the dataset more readily usable for regional-scale applications.

Lines 186-187: At the HUC2 scale, the derived gain and loss estimates cover 21% to 99.9% of basin area for TP and 7% to 96.5% for PO_4^{3-}, with 72% and 44% of HUC2 basins exceeding 50% coverage, respectively (Fig. S5 and Table S2).

Lines 207-208: Notably, widespread regions in the Midwest exhibit heightened P gains (Fig. S7)

Lines 224-225: Notably, in most of the agriculturally intensive Missouri and Tennessee-Ohio river basins, total nonpoint source contribution significantly surpassed point source contributions (Fig. S8).

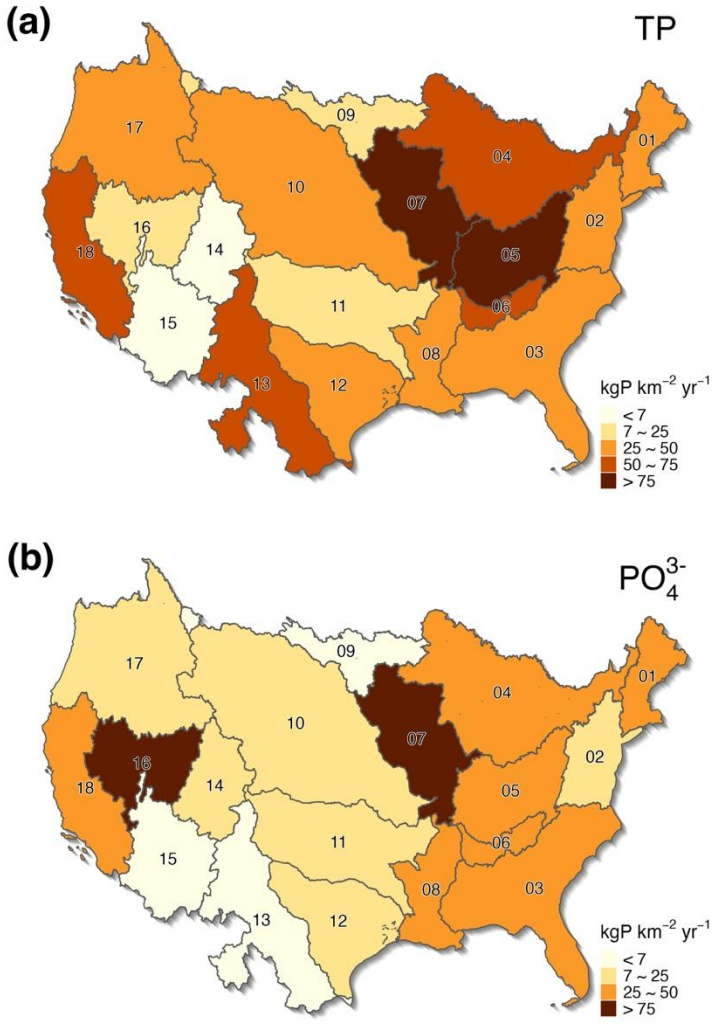

Figure S7: Area weighted average gain and loss of (a) TP and (b) PO_4^{3-} at the HUC2 level.

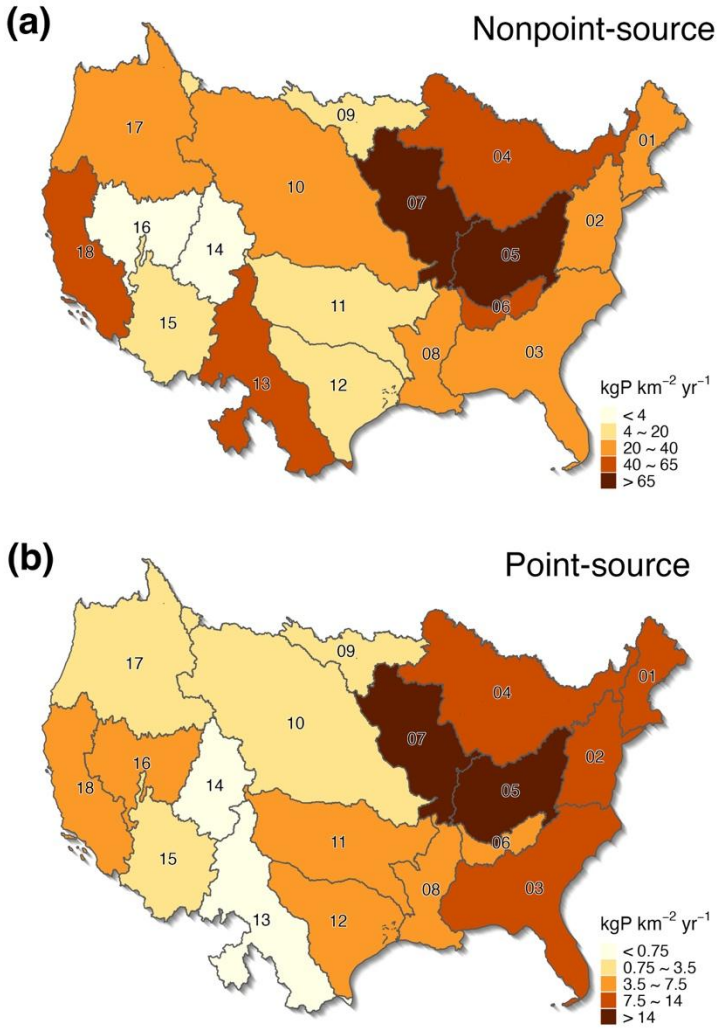

Figure S8: Area weighted average (a) non-point source and (b) point-source contributions at the HUC2 level.

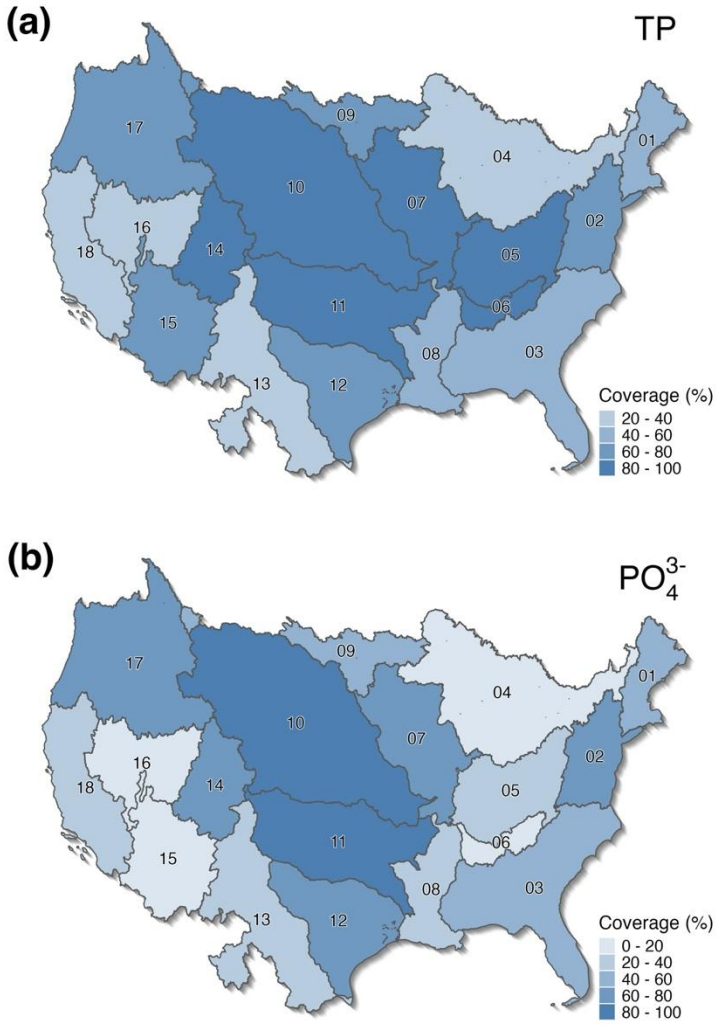

Figure S5: Spatial coverage of riverine gain and loss estimates across HUC2 basins in the CONUS for (a) TP and (b) PO_4^{3-}.

Model diagnostics seem reasonable. A statement regarding why there are possible differences between Loadest and WRTDS may benefit the reader.

Response: Thank you for this helpful suggestion. We agree that a brief explanation of the differences between LOADEST and WRTDS would benefit the reader.

Overall, the two approaches show strong agreement, as indicated by the high r^2 and relatively low RMSE. The remaining differences are primarily attributable to their distinct model structures and assumptions. LOADEST employs a set of predefined parametric regression models (Table S1) and selects the best-performing formulation based on statistical criteria (e.g., AIC), whereas

WRTDS uses a semi-parametric approach with time-varying coefficients to capture gradual changes in concentration-discharge relationships over time (eq. S1).

$$ln\ ln(c) = \beta_0 + \beta_1 t + \beta_2\ ln\ ln(Q) + \beta_3\ sin\ sin(2\pi t) + \beta_4\ cos\ cos(2\pi t) + \epsilon \qquad S1$$

where c is the concentration, β is fitted coefficient, Q is discharge, t is the time in years, and ϵ is the unexplained variation.

In addition, differences in temporal coverage and data used for model calibration may also contribute to discrepancies between the two estimates. We have added a brief explanation of these differences in Section 3.1 ("Riverine phosphorus data") to clarify this point.

Table S1. Regression models in LOADEST

ID	Regression model
1	$a_0 + a_1\ ln\ ln\ Q$
2	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ ln\ ln\ Q^2$
3	$a_0 + a_1\ ln\ ln\ Q\ + a_2 dtime$
4	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ sin\ sin\ (2\pi dtime)\ + a_3\ cos\ cos\ (2\pi dtime)$
5	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ ln\ ln\ Q^2\ + a_3 dtime$
6	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ ln\ ln\ Q^2\ + a_3\ sin\ sin\ (2\pi dtime)\ + a_4\ cos\ cos\ (2\pi dtime)$
7	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ sin\ sin\ (2\pi dtime)\ + a_3\ cos\ cos\ (2\pi dtime)\ + a_4 dtime$
8	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ ln\ ln\ Q^2\ + a_3\ sin\ sin\ (2\pi dtime)\ + a_4\ cos\ cos\ (2\pi dtime) + a_5 dtime$
9	$a_0 + a_1\ ln\ ln\ Q\ + a_2\ ln\ ln\ Q^2\ + a_3\ sin\ sin\ (2\pi dtime)\ + a_4\ cos\ cos\ (2\pi dtime) + a_5 dtime + a_6 dtime^2$

where $ln\ ln\ Q$ is the difference between $ln\ ln\ streamflow$ and $center\ of\ ln\ (streamflow)$; $dtime$ is the difference between $decimal\ time$ and $center\ of\ decimal\ time$; a_0, a_2, a_3, a_4, a_5, and a_6 are model regression coefficients.

Lines 173-174: The minor disparities observed between these two datasets are likely attributable to variations in temporal coverage of data and different regression equations.

How was "riverine removal" estimated? That is not clear to this reviewer. Is it when upstream to downstream is negative? Or did you just assume 12% loss? If yes, that would assume that

reservoirs/lakes are mostly responsible for removal. And were reservoirs/lakes considered in this loss calculation and interpretation?

Response: Thank you for this important clarification. In our framework, “riverine removal” was not explicitly estimated. Instead, we assumed that in-stream removal terms are negligible when deriving Eq. (2), such that riverine P gain and loss is defined purely as the difference between downstream and upstream loads. Therefore, negative gain values do not directly represent removal processes, but rather net decreases in load along the flow path, which may reflect a combination of retention, transformation, or data uncertainties.

As a result of this assumption, the estimated nonpoint source contribution represents a conservative (lower-bound) estimate. We did not apply a fixed removal rate (e.g., 12%) in the calculation. Instead, we used the 12% value reported in previous studies as an illustrative example in the Discussion (Section 4.1) to highlight the potential magnitude of underestimation. Incorporating such removal processes (e.g., reservoir trapping, lake retention, and wetland uptake) would likely increase the estimated nonpoint source inputs.

We have clarified this assumption and its implications in both the Methods and Discussion sections to avoid confusion.

Lines 116-120: Thus, riverine removal is not explicitly estimated in this study, and the calculated P gain and loss represent the net difference between downstream and upstream loads. Conceptually, this net difference reflects the combined effects of watershed P inputs (from both point and nonpoint sources) and in-stream processes (e.g., retention, transformation, and remobilization) occurring along the flow path, rather than a direct measure of any single process such as in-stream removal.

Lines 123-124: Accordingly, negative values of P gain and loss indicate net decreases in load along the flow path, which may reflect a combination of retention, transformation, or other processes, rather than being interpreted solely as riverine removal.

Figure 6: There is useful information here, but it seems worth suggesting something like a dynamic watershed model (e.g., SPARROW) as a possible next step for another approach to gains and losses. This reviewer agrees that this dataset is useful in supporting evaluation and diagnosis of watershed models, and because you are offering an initial cut, it may be worth making a more explicit recommendation.

Response: Thank you for this insightful suggestion. We agree that linking the dataset to dynamic watershed modeling frameworks would further enhance its utility. In response, we have revised Section 4.3 to explicitly highlight the potential application of our dataset in supporting and improving large-scale watershed models, such as SPARROW. Specifically, we now note that the

spatial patterns of riverine P gain and loss can be incorporated into such models to better constrain nutrient sources and in-stream processing across river networks. We believe this addition clarifies the broader applicability of the dataset for future modeling efforts.

Lines 312-315: They can support the evaluation and diagnosis of large-scale dynamic watershed models, the examination of environmental controls on riverine P loads, and the estimation of contributions to P gain and loss. For instance, models such as SPARROW could incorporate the spatial patterns of riverine P gain and loss to better constrain nutrient sources and in-stream processing across river networks.

Did streamflow only come from NWIS? Did you use only USGS gages? You used more than the GagesII reference gages, correct? If only reference gages, linking to agriculture is limited. And how were those paired with your discrete P data? Did you only consider stations with P and streamflow data, or did you collect P data still close to a streamflow gage? A concise statement is needed for the reader.

Response: Thank you for this important clarification. Phosphorus concentration data were obtained from the Water Quality Portal (WQP), but we only used data from USGS NWIS source.

We included all monitoring stations for which both phosphorus concentration and corresponding daily streamflow data were available. This means that only stations with co-located water quality observations and streamflow records were retained in the analysis. Phosphorus concentrations and streamflow were paired at the daily scale, using matching observation dates.

The set of stations used in this study extends beyond the USGS GAGES-II reference network. Specifically, our dataset includes an additional 177 stations for TP and 78 stations for PO_4^{3-} beyond the GAGES-II reference gages, thereby increasing spatial coverage, particularly in human-impacted basins.

We have clarified these points in the revised manuscript (Section 2.3 Data compilation).

Lines 77-81: To ensure consistency with streamflow records, only records from the U.S. Geological Survey (USGS) National Water Information System (NWIS) data source were used in this study. For each P observation, we identified co-located hydrological stations (Wang, Zhang, Zhao, et al., 2024) and downloaded and processed daily streamflow data from the USGS NWIS. Only stations with both phosphorus concentration observations and corresponding daily streamflow records were retained.

Is the WQP pull comprehensive, or are there still possible holes or gaps in that data pull?

Response: Thank you for this important question. The Water Quality Portal (WQP) is currently the most comprehensive publicly available repository for water quality data in the United States, integrating observations from multiple agencies including USGS NWIS and EPA WQX. In this study, we accessed WQP to ensure broad coverage, but retained only records from the USGS NWIS source to ensure consistency with streamflow data. This would result in the lack of observations from other sources (i.e., EPA WQX).

Despite its extensive coverage, the WQP dataset may still contain spatial and temporal gaps due to uneven monitoring efforts across regions and time periods. For example, monitoring stations are denser in the eastern U.S. than in the western regions, and historical sampling frequency varied substantially among sites. These limitations are inherent to large-scale observational datasets and are partially reflected in the spatial coverage of our derived products (Fig. S5).

Overall, while some gaps may remain, the WQP-based compilation represents the most comprehensive and consistent dataset currently available for large-scale analysis of riverine phosphorus across the CONUS.

Lines 299-301: Note that available stations with observed P concentration and streamflow data are relatively sparse in the western vs. eastern U.S., particularly for PO_4^{3-}. This led to large gaps in the spatial coverage of the datasets (Fig. 4).

Specific comments:

27: Eutrophication of what? A Wurtsbaugh citation implies lakes. Perhaps add “inland waters and estuaries” to benefit the reader. This reviewer agrees there should be more focus regarding the eutrophication of rivers and streams.

Response: Thank you for this insightful suggestion. We have clarified the scope by revising the text to “inland waters and estuaries”.

Lines 27: Eutrophication in inland waters and estuaries is a widespread water quality challenge across the globe.

34: “excessive”, why is it excessive? They are mostly unused and perhaps transported out of reach of crops. Perhaps “unused” is a better word here? This reviewer has a different

interpretation of “excessive” versus “unused.” Saying unused implies there may be some management action such as fertilizer timing or cover crops. Excessive sounds like we just put too much down without consideration.

Response: Thank you for this insightful comment. We agree that the term “excessive” may imply over application without management consideration, whereas “unused” more accurately reflects phosphorus that is not taken up by crops and remains available for transport. We have therefore revised the wording accordingly to better capture the underlying process without implying a specific management cause.

Lines 34-36: P surplus in agricultural soils due to unused fertilization and manure application can be transported to water bodies through surface runoff and groundwater pathways, and cause persistent water pollution (Stackpoole et al., 2019).

39: Agreed! This dataset is a nice contribution. Thanks.

Response: Thank you for your positive feedback. We appreciate your recognition of the value of this dataset.

53: “over 1000” is confusing. Suggest to explicitly state actual number as you do in abstract.

Response: Thank you for this suggestion. We have replaced “over 1000” with the explicit number of stations to improve clarity.

Lines 52-54: To estimate riverine P gain and loss data across the CONUS, we compiled streamflow and P concentration data (i.e., unfiltered phosphate (PO_4^{3-})) from 963 monitoring stations and total phosphorus (TP) from 2,317 stations, and calculated P loads at these stations using the Load Estimator (LOADEST) program (Runkel et al., 2004) (Fig. 1).

113: is it 12%? This reviewer considers that significant, not small.

Response: Thank you for this important comment. We agree that an in-stream removal rate on the order of ~12% should not be described as “small.” In the revised manuscript, we have removed this wording and clarified that in-stream removal can vary across systems. We now explicitly frame the assumption as a first-order simplification made to enable estimation of nonpoint source inputs, rather than implying that removal is insignificant.

Lines 114-116: Because in-stream removal terms cannot be directly constrained and may vary across systems (e.g., about 12% globally) (Maavara et al., 2015), we assume the values of these two terms to be negligible for the purpose of deriving a simplified, first-order estimate for nonpoint source inputs.

230: Maavara did not focus on rivers, 12% trapped by reservoirs. Citation does not support statement.

Response: Thank you for this helpful comment. We agree that Maavara et al. (2015) primarily focuses on phosphorus retention in reservoirs and does not directly support our statement regarding riverine processes. We have therefore replaced this citation with a more appropriate reference that specifically addresses phosphorus transport and cycling in rivers (Withers and Jarvie, 2008).

Lines 251-252: Given that the TP inputs from point and nonpoint sources are often subject to riverine removal (Withers and Jarvie, 2008)

"Withers, P. J. A. and Jarvie, H. P.: Delivery and cycling of phosphorus in rivers: A review, Science of The Total Environment, 400, 379–395, https://doi.org/10.1016/j.scitotenv.2008.08.002, 2008."

270: What is a "hydrologic station"? Do you mean a streamflow gaging station? Discrete P data paired with streamflow, this reviewer would call that a "water-quality station." Is it streamflow, WQ, or both? "Hydrologic" suggests no WQ data. This reviewer suggests picking consistent terminology.

Response: Thank you for this helpful suggestion. We agree that the term "hydrologic station" was ambiguous and could be misinterpreted as referring to streamflow-only gaging sites. In this study, we refer specifically to stations with paired phosphorus concentration and streamflow observations. To improve clarity and maintain consistent terminology, we have revised the manuscript to use "monitoring stations." We chose this term to provide a neutral and broadly applicable description, while explicitly clarifying in the text that these stations include both water-quality measurements and corresponding streamflow data. The terminology has been updated consistently throughout the manuscript.

290: Agreed!

Response: Thank you for your comment. We are glad that this point is clear and well supported.

293: What about trends? What is the period of record? If 2015-2020, this is for more recent understanding, so trends are not as priority.

309: That's a good number to have. Nice work. But what time period does it represent?

Response: Thank you for these important comments. We agree that the temporal context of the dataset is essential for interpreting both the reported values and the relevance of trends.

Our dataset does not represent a specific short-term period (e.g., 2015-2020), but instead integrates all available observations to maximize spatial coverage and upstream–downstream connectivity. As a result, the compiled data span multiple decades, with PO_4^{3-} from 1952-2022 and TP data from 1958-2023. Therefore, the reported estimates (e.g., nonpoint source contributions) represent long-term, multi-decadal average conditions rather than a specific recent time window.

To address the role of temporal change, we have additionally incorporated a trend analysis based on stations with long-term records (>30 years), using the Sen's slope and Mann-Kendall test. This analysis provides complementary insights into directional changes over time, while the main dataset focuses on spatial patterns and long-term average conditions.

We have clarified the temporal coverage and interpretation of the estimates in both the Discussion and Conclusions sections to avoid ambiguity.

Lines 315-317: The insights derived from our datasets contribute to a more comprehensive understanding of P dynamics under long-term and multi-decadal conditions, providing a foundation for improving water quality management on local, regional, and national scales.

Lines 333-334: we derived conservative estimates of the long-term average contribution of nonpoint sources to riverine TP (28.24 kgP km^{-2} yr^{-1}).