PNAS RESEARCH ARTICLE | ENVIRONMENTAL SCIENCES

Increasing phosphorus loss despite widespread concentration decline in US rivers

Wei Zhi[a,b,1], Hubert Baniecki[c,d], Jiangtao Liu[b], Elizabeth Boyer[e,f], Chaopeng Shen[b], Gary Shenk[g], Xiaofeng Liu[b,f], and Li Li[b,1]

Affiliations are included on p. 8.

Edited by Nils Stenseth, Universitetet i Oslo, Oslo, Norway; received January 30, 2024; accepted October 10, 2024

The loss of phosphorous (P) from the land to aquatic systems has polluted waters and threatened food production worldwide. Systematic trend analysis of P, a nonrenewable resource, has been challenging, primarily due to sparse and inconsistent historical data. Here, we leveraged intensive hydrometeorological data and the recent renaissance of deep learning approaches to fill data gaps and reconstruct temporal trends. We trained a multitask long short-term memory model for total P (TP) using data from 430 rivers across the contiguous United States (CONUS). Trend analysis of reconstructed daily records (1980–2019) shows widespread decline in concentrations, with declining, increasing, and insignificantly changing trends in 60%, 28%, and 12% of the rivers, respectively. Concentrations in urban rivers have declined the most despite rising urban population in the past decades; concentrations in agricultural rivers however have mostly increased, suggesting not-as-effective controls of nonpoint sources in agriculture lands compared to point sources in cities. TP loss, calculated as fluxes by multiplying concentration and discharge, however exhibited an overall increasing rate of 6.5% per decade at the CONUS scale over the past 40 y, largely due to increasing river discharge. Results highlight the challenge of reducing TP loss that is complicated by changing river discharge in a warming climate.

phosphorus loss | water quality | deep learning | big data | changing climate

Phosphorus (P) is essential for life on Earth. Unlike nitrogen, P is nonrenewable with limited geological deposits (1). Global analysis indicates that P shortage is possible in coming decades (2). P loss from the land to rivers depends heavily on soil erosion and hydrometeorological conditions (3), particularly precipitation and river discharge. Riverine P loss has caused eutrophication and hypoxia worldwide (4, 5), estimated to cost at least $4.3 billion annually in the United States alone (6). P loss also threatens ecosystems (7), soil productivity (8), and food production (9). Management and practices have been implemented to reduce nutrient loss since the Clean Water Act in 1972, although national-scale assessment indicates limited effectiveness (10, 11).

Total P (TP) is the sum of dissolved and particulate P. Particulate P, closely bound to soil organic matter, can be mobilized via soil erosion process during runoff events (12). The rates of P loss, quantified as fluxes (loads, quantified by multiplying concentrations and river discharge), are expected to rise with changing land use and climate that often accelerate soil erosion and sediment mobilization in rivers (13, 14). Systematic analysis of temporal trends however has remained challenging, largely due to sparse and inconsistent historical TP data across sites under diverse climate and land use conditions. The first National Water Quality Inventory (10) examined the largest 22 US rivers and concluded that TP concentrations increased in 82% of the river reaches from the mid-1960s to the early 1970s, with 57% of the rivers exceeding the limit of 0.1 mg/L. The National Water-Quality Assessment (NAWQA) Program monitored 171 streams approximately quarterly from 1993 to 2003. Results indicate minimal changes in TP concentrations in 51% of the rivers, and more increasing (33%) than decreasing (16%) trends in the remaining rivers (11). The most recent National Rivers and Streams Assessment (NRSA) sampled more than 1,800 rivers in the summer of 2013–14 and rated water quality in 58% of river miles as poor (15). Models such as SPARROW (SPAtially Referenced Regression on Watershed attributes) account for spatial variability but are limited in estimating temporal trends of TP loss (16, 17). Existing studies from regional to global scales have generally focused more on spatial variability than temporal trends and have rarely assessed temporal trends of riverine TP loss systematically (9, 14, 18).

Here, we overcome data limitation by leveraging the increasingly available Earth data (e.g., hydrometeorological data and river basin attributes) and deep learning approaches (19–21). The application of deep learning models has grown rapidly in hydrology (22)

Significance

Phosphorus (P) reserves in Earth's rocks are finite. P loss from land to rivers threatens not only food production but also aquatic ecosystem health. Long-term trend analysis of P loss has historically been challenged by sparse data. Here, we overcome this limitation by leveraging weather and earth characteristics data and building a multitask deep learning model for daily concentrations and fluxes (1980–2019) in 430 rivers at the Contiguous United States. Trend analysis shows widespread declines in concentrations, particularly in urban rivers. Concentrations in agricultural rivers, however, have mostly increased, suggesting not-as-effective controls of nonpoint sources. Despite declining concentrations, riverine P loss (fluxes) has significantly increased, driven largely by increasing streamflow in a changing climate.

Author contributions: W.Z. and L.L. designed research; W.Z. performed research; W.Z., H.B., E.B., C.S., and G.S. contributed analytic tools; W.Z., J.L., G.S., and X.L. analyzed data; and W.Z. and L.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: zhiwei@hhu.edu.cn or lili@engr.psu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2402028121/-/DCSupplemental.

Published November 18, 2024.

but is relatively nascent in water quality analysis. Here, we ask the questions: *What are the temporal trends of TP concentrations and fluxes in the past decades in contiguous United States (CONUS)? What are the most influential drivers of TP temporal trends?* We built a multitask deep learning model (long short-term memory, LSTM) to fill temporal-spatial data gaps and reconstruct continuous daily concentrations and fluxes in 430 independent, non-nested basins in CONUS from 1980 to 2019. These basins consist of 22 agricultural basins (5.1%, AG), 92 undeveloped basins (21%, UD), 102 urban basins (24%, UB), and 214 mixed (MX) basins (50%, MX). A single CONUS-scale LSTM model was trained to predict daily concentration and fluxes from 1980 to 2019 using 1) time-series hydrometeorological forcing data (e.g., discharge, air temperature, precipitation) and 2) static basin attributes including measures of topography, climate, hydrology, land use, soil, and geology. The reconstructed daily concentrations and fluxes were used to analyze temporal trends (i.e., Theil-Sen slope) and calculate TP loss under different land use conditions.

Results

Mean TP Concentrations and Fluxes Controlled by Climate and Land Use. The long-term mean concentrations (C_m) and fluxes (F_m) show different spatial patterns (Fig. 1). Mean concentrations and fluxes (daily concentration times daily discharge) were calculated as the means of all available concentration and flux data at each site. Across sites with different climate, geology, and vegetation conditions, mean concentrations are highest in arid rivers in Great Plains from North Dakota to Texas and lower along the humid coasts. In fact, mean concentrations and discharge across sites (C_m-Q_m, Fig. 1*C*) correlate negatively (R^2 = 0.063, P < 0.001, n = 430), especially in UD rivers that exhibit

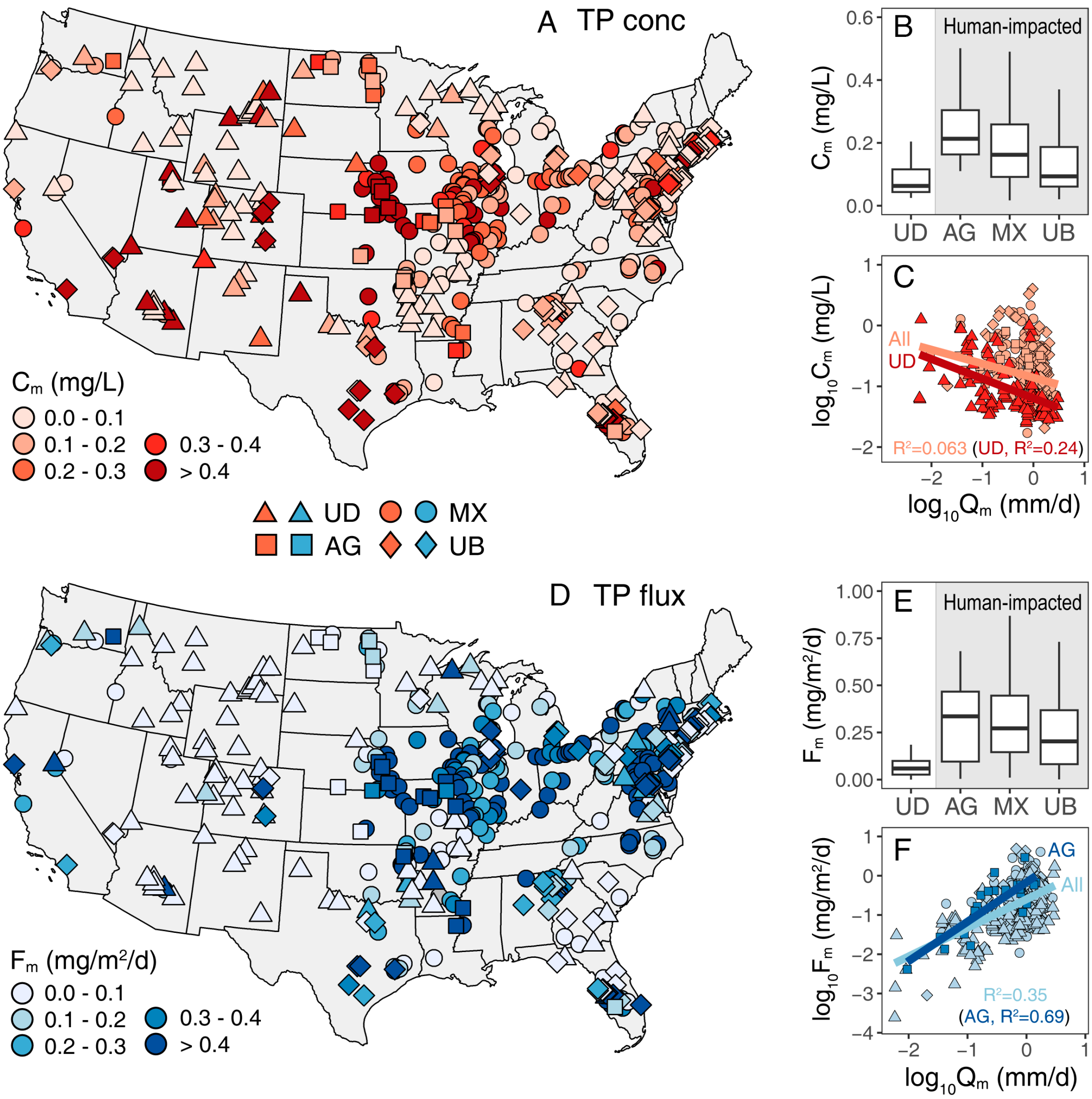

Fig. 1. Long-term mean TP concentrations (*A*, *B*, *C*) and area-normalized fluxes (*D*, *E*, *F*) and their relationships with discharges in 430 US rivers based on raw data. Mean concentrations C_m were calculated as the mean of concentrations in all years in each site; mean daily fluxes F_m were calculated as the mean of daily area-normalized fluxes (daily C times daily area-normalized Q) of all years at each site. Basin classifications of AG, UB, UD, and MX followed USGS-based land use classification: AG: >50% agricultural (planted/cultivated) lands and ≤5% UB (developed) lands; UB: >10% UB and ≤25% AG; UD: ≤25% agricultural and ≤5% UB; MX: all other combinations (details in *Materials and Methods* section). The boxplot displays median and interquartile range of mean concentrations; gray shading indicates human-impacted basins (i.e., AG, MX, and UB). In C_m-Q_m and F_m-Q_m figures (*C* and *F*), lighter lines are for all rivers; darker red and blue lines are for UD and AG rivers that have the highest R^2. The highest concentrations occur in the Midwest and the Great Plains from North Dakota to Texas. Fluxes are higher in eastern rivers and exhibit a sharp divide between the West and East.

lower mean concentrations with increasing mean discharge (R^2 = 0.24, $P < 0.001$, n = 92), possibly due to geological and land-use characteristics (e.g., limited phosphorus source). This pattern differs from the commonly observed TP mobilization patterns in individual rivers that often show high TP concentrations at high discharge and reflect enhanced TP mobilization at high discharge (23). This negative C_m-Q_m relationship of higher concentration in more arid places however has been observed for many water quality variables in large datasets at regional (24), continental (25, 26), and global scale (27). Such pattern has been explained to arise from material accumulation due to high production of materials on land relative to minimal export to rivers under arid, low discharge conditions (25).

In addition to climate, land use also drives concentration levels (Fig. 1*B*). UB rivers have point sources such as wastewaters from municipal and industrial facilities, and nonpoint sources including fertilizers from lawns, golf courses, parks, and failing septic systems (28). Agricultural lands are often dominated by nonpoint sources from fertilizers and manure (28). UD rivers here have some coverage of agricultural and developed lands, leading to slightly higher concentrations than the national background of 0.034 mg/L from pristine streams (11). UD rivers have the lowest median concentrations (0.065 mg/L, Fig. 1*B*), whereas agriculture (AG) rivers have the highest median (0.25 mg/L) with 100% rivers exceeding the maximum concentration level (MCL) of 0.1 mg/L. Rivers of MX land uses follow closely, with a median of 0.17 mg/L and 74% rivers exceeding MCL. UB rivers have a median of 0.12 mg/L and 56% exceeding MCL. Nationwide, 272 rivers (63%) exceed MCL of 0.1 mg/L (Fig. 1*A*), with exceedance occurring at an average of 80% ± 23% (mean ± SD) of the time.

TP fluxes however exhibit a clear divide between the East and West roughly along the dividing line 100°W. On average, eastern basins have 3.9 times higher fluxes than western basins, largely arising from higher river flow in the East with abundant precipitation. In fact, mean flux and discharge (F_m-Q_m) correlate robustly and positively (R^2 = 0.35, $P < 0.001$, Fig. 1*F*). This is expected, as fluxes are primarily driven by discharge. A few hotspots emerge in the flux map, including agricultural areas in the central and northeastern regions, and major metropolitan areas (e.g., New York City, NY; Philadelphia, PA) in the Northeast, indicating the influence of land use (Fig. 1*E*) (29). Other regional differences additionally influence spatial patterns. Texas is sparsely populated but has expanded UB population significantly (e.g., 30% increase in coastal counties from 1990s to 2000s), which leads to high fluxes (30). Wastewaters from hydraulic fracturing in Texas also contain phosphorous (31). The F_m-Q_m correlation is the strongest in agricultural rivers (R^2 = 0.69, $P < 0.001$), indicating TP loss is driven by discharge more in AG lands than in other land uses. This potentially arises from flow modification by AG activities such as tile drainage (32). Currently no national water quality criteria exist for TP fluxes in surface waters, although Total Maximum Daily Loads (TMDLs) exist in some areas. UD rivers have lower median normalized fluxes (0.063 mg/m^2/d) compared to human-impacted lands (0.26 to 0.38 mg/m^2/d).

Model Performance and Data-Filling Capacity. An LSTM model was trained using data from all 430 independent (nonnested) basins and predicted their daily concentrations and fluxes from 1980 to 2019 (*SI Appendix*, Fig. S1). The model achieved high performance with mean (median) Nash–Sutcliffe Efficiency (NSE)

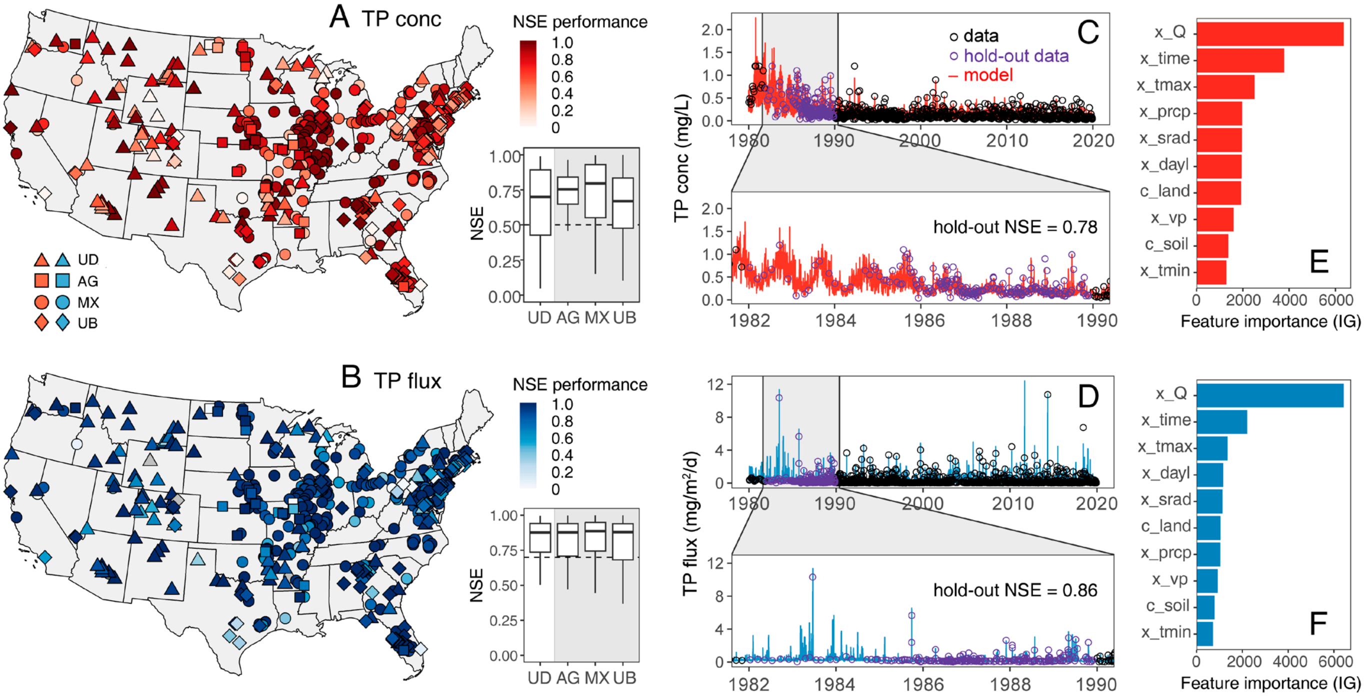

Fig. 2. Model performance, example time-series, and feature importance for TP concentrations and fluxes. (*A* and *B*) Model performance quantified by NSE. (*C* and *D*) example time-series of concentrations and fluxes. (*E* and *F*) feature importance ranking for concentrations and fluxes. NSE ranges from -∞ to 1, with 1 being the perfect match between model prediction and observation and 0 being unacceptable performance. The boxplot displays medians and interquartile range of NSE with dashed lines indicates good performance criteria of 0.5 for concentrations and 0.7 for fluxes. Reported NSE values are from the testing period. The model shows robust performance across diverse climate and land use conditions, and generally predicts fluxes better than concentrations. The time series figures (*C* and *D*) show the model ability to fill the 8-y data gaps (purple dots) where data were purposely removed from the training. The feature importance (*E* and *F*) was calculated based on IG and aggregated for all 430 basins over 40 y (details in *Materials and Methods*). Variables starting with "x_" indicate temporally varying variables, whereas those with "c_" means constant, static attributes. It shows that discharge (x_Q) as the predominant driver for both concentrations and fluxes, followed by timestamp variable (x_time), and time-series hydrometeorological forcing including daily maximum temperature (x_tmax), solar radiation (x_srad), day length (x_dayl), vapor pressure x_ (vp), and daily minimum temperature (x_tmin). Constant basin attributes such as land use (c_land) and soil properties (c_soil) were also ranked among the top 10 predictors.

of 0.62 (0.73) for concentrations and 0.75 (0.87) for fluxes in the testing period (Fig. 2 *A* and *B*), exceeding the good criteria of 0.50 for daily concentrations and 0.70 for daily fluxes (33). Agricultural and MX rivers exhibited slightly higher NSE performances for TP concentrations; for TP fluxes, the performance was relatively uniform across different land uses. The model shows robust data-filling capacities in the 8-y hold-out period (Fig. 2 *C* and *D*, hold-out NSE = 0.78 and 0.86), the period when data were excluded during the model training to test the model prediction capability. The model captured concentrations and fluxes over varying flow conditions (e.g., baseflow, high flow) across seasons in individual rivers (*SI Appendix*, Fig. S3) demonstrated robust local-scale prediction. It also reproduced the long-term data trends (i.e., decadal changing rates) in the 8 y without data, with R^2 = 0.83 and 0.54 for concentrations and fluxes (*SI Appendix*, Fig. S4), respectively.

Feature importance analysis (details in *Materials and Methods*) ranked the same three temporally varying variables (discharge, time, and maximum temperature) as the top drivers for concentrations and fluxes (Fig. 2 *E* and *F*). Notably, discharge exhibited a greater influence in fluxes than concentrations, as streamflow connects land and river P sources and thus governs P transport (28). The time variable x_time ranked as an essential driver after discharge. This variable is the timestamp used as a time-series input to facilitate dynamical learning of the input–output relationships based on the year and season along with other watershed conditions (34) (*Materials and Methods*). It serves as a latent variable representing the aggregated effects of human and management factors such as best management practices, tile drainage, and point sources. These variables change over time and are not represented by the hydrometeorological forcings; they also cannot be directly quantified or used as model inputs due to limited data availability (31). The significance of this timestamp variable indicates the impact of human activities that change over time, but their influences are not as dominant as discharge (35). Most variables in the top 10 predictors are hydrometeorology variables. Two constant attributes, land use characteristics (c_land) and soil properties (c_soil), also made the list (Fig. 2 *E* and *F*), suggesting their influences in determining TP dynamics possibly through flow paths and biogeochemical reactions (36, 37).

Widespread Decreasing Concentrations but Increasing Fluxes over Time. Most rivers (60%) see decreasing concentrations, followed by increasing (28%) and insignificant (12%) trends (Fig. 3*A*). When averaged over all rivers, the decadal rate is −1.9 ± 20% (mean ± SD) compared to their concentrations in 1980. When averaged only over rivers with a declining trend, the decadal rate is −12 ± 6.6%. Such widespread decline indicates progress in reducing TP concentrations especially in UB and MX rivers. In fact, 77% and 57% of UB and MX rivers exhibited declining trends (*SI Appendix*, Table S1), respectively, followed by 41% UD and 23% agricultural rivers. UD rivers exhibited an overall stable trend, with a median rate closest to zero (Fig. 3*A* box). However, some UD rivers exhibited significant trends, indicating that concentrations in these sites do vary under changing climate

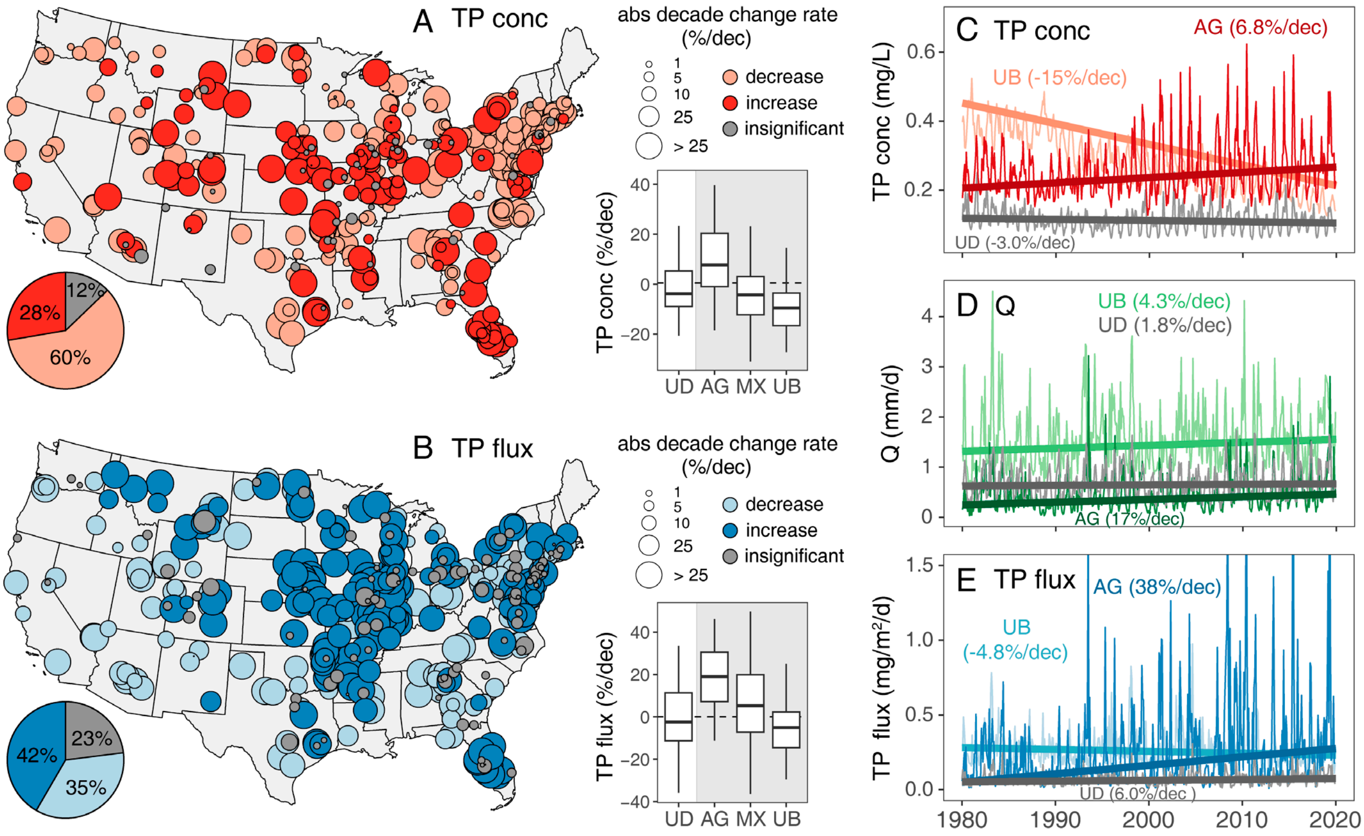

Fig. 3. Long-term trends of TP concentrations and fluxes. (*A* and *B*) long-term trends in percent change per decade (%/dec) compared to values in 1980. (*C*–*E*) time series and temporal trends of averaged concentrations, discharge, and fluxes in different land use categories. The boxplots display median and interquartile range of decadal change rates; positive and negative values indicate increasing and decreasing trends, respectively. The decline (60%) trend is more widespread in TP concentration especially in UB and MX lands than in fluxes. In (*C*–*E*), averaged concentrations, discharge, and fluxes across all UB, AG, and UD (gray) sites show different trends under different land use conditions. Increasing discharge drives the flux trends, leading to less pronounced decreasing trend of fluxes compared to concentrations in UB lands and amplifying the increasing trend of fluxes compared to concentrations in AG and MX lands. MX lies in between AG and UB and is not plotted.

conditions. Among human-impacted rivers, the average rates of AG, MX, and UB are 7.6 ± 16%, −2.1 ± 15%, and −7.4 ± 14% per decade compared to concentrations in 1980 (Fig. 3*A* box), respectively. TP concentrations in agricultural rivers have generally increased whereas those in UB areas have declined, possibly due to declining municipal wastewaters and UB runoff (38–40). MX lands often have a larger fraction of AG (47 ± 24%) than UB areas (10 ± 7.6%) but mostly followed the decreasing trend in UB sites. When averaging TP concentrations for all sites over each category (Fig. 3*C* and *SI Appendix*, Table S1), the overall trends show a similar land use pattern with +6.8% (increase) per decade in AG but −15%, −6.6%, and −3.0% per decade in UB, MX, and UD, respectively, compared to concentrations in 1980. This underscores challenges in containing and mitigating nonpoint sources in AG lands (41).

TP fluxes exhibit much less declines compared to concentrations, with decreasing (35%), increasing (42%), and insignificant (23%) trends (Fig. 3*B*). This is possibly attributed to increasing river discharge in every land use type. In UD lands, river discharge Q has increased by 1.8%/dec (Fig. 3*D*), which switched the decreasing trend of concentrations (−3.0%/dec) to an increasing trend of fluxes (6.0%/dec). River discharge in human-impacted lands increased by 4.3 to 17%/dec (Fig. 3*D*), leading to subdued decreasing trends of fluxes in UB lands (−4.8%/dec, Fig. 3*E* and *SI Appendix*, Table S1) and more pronounced increasing trends for fluxes in AG (38%/dec) and MX (6.3%/dec). This is consistent with the mean concentration and flux data analysis (Fig. 1 *C* and *F*) and feature importance analysis (Fig. 2 *E* and *F*) that indicates discharge as the most influential driver of fluxes.

TP Loss from Land to Rivers in CONUS. The trained LSTM model was applied to predict TP fluxes from HUC6 (Hydrologic Unit Code at level 6) basins to estimate total TP loss (Tg/y, teragram per year, not area-normalized) at CONUS (Fig. 4). The TP loss maps show changing patterns in 1980 and 2019 (Fig. 4 *A* and *B*), although both maps show hot spots in the eastern United States, especially in regions with heavy AG draining to the Mississippi River basin. The bottom figure (Fig. 4*C*) shows that although MX and UD occupy similar area percentages in CONUS (43 to 44%), MX basins export 3.4 times of that in UD (*SI Appendix*, Table S1). UB rivers export 15% of TP, although only drains 8.4% of the land. Total TP loss in CONUS increased from 0.43 to 0.48 Tg/y from 1980 to 2019, with a changing rate of 6.5%/dec in CONUS (Fig. 4*C*, solid trend line). These numbers are in par with TP loss reported in literature. The average TP loss in CONUS from 1980 to 2019 is 0.42 Tg/y, about half of the earlier estimation of 0.9 to 1.1 Tg/y in North America (18). Annual fluxes from the Mississippi River Basin, which drains about 41% area of CONUS, was estimated at 0.16 to 0.19 Tg/y (28, 42), consistent with 0.17 Tg/y in this study if we scale the average TP loss (0.42 Tg/y) by its drainage area fraction. The P loss from the US croplands was estimated as 0.2 Tg/y (9), accounting for about 47% of the average CONUS export from this work. This estimate is higher but close to an earlier estimate of about 38% of P loss to freshwater originated from AG (43). The overall increasing rate of 6.5%/dec in CONUS is much higher than the previously estimated 4.5%/dec in Chesapeake Bay watershed (16) based on two time snapshots of 1992 and 2012 using the SPARROW model. The upscaled TP estimates from the trained LSTM facilitate the consistent tracking of historical trends and scalable application across CONUS. However, caution will need to be exercised when using these numbers, because the upscaled estimations are subject to uncertainties of extrapolating the trained LSTM model to sites without data. Although LSTM models have been shown to reliably fill data gaps (20, 44), the reliability and accuracy of spatial data–filling hinge upon the quality and availability of data and the similarities in conditions between the sites with and without data (45).

Discussion

We trained a deep learning model to reconstruct daily TP concentrations and fluxes from 1980 to 2019, which were then used to systematically analyze their spatial patterns and temporal trends

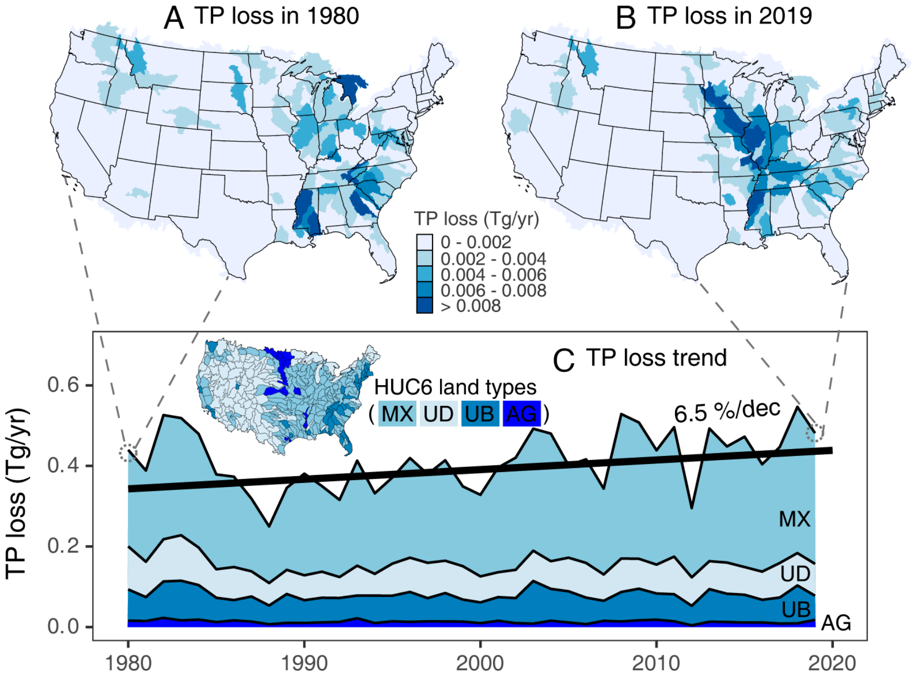

Fig. 4. The trajectory of TP loss from the CONUS with two snapshots in 1980 and 2019 (*Top* row). TP loss (Tg/y, 1 teragram = 10^{12} g) for each basin (HUC6 level) was estimated by multiplying the predicted daily TP flux (mg/m^2/d) from the trained LSTM model by its corresponding drainage area (km^2) and summing over the entire year (*A* and *B*). Total TP loss was summarized at the CONUS scale or by each land use categories (*C*). The solid line is the temporal trend of total TP loss in CONUS in the unit of 6.5 %/dec.

and upscale TP losses at the CONUS. This approach overcome data limitation and temporal bias inherent in sparse datasets such as one- or two-time snapshots, and infrequent sampling with quarterly data from annual to decades scales (10, 11, 15). TP loss from the Mississippi River Basin, for example, has been reported to exhibit inconsistency with both decreasing and increasing trends (35). Although spatial bias still exists due to inconsistent data availability across regions, this work highlights the utility of deep learning models in filling spatiotemporal data gaps and in predicting water quality in chemical-ungauged basins (45).

UB rivers have seen a pronounced decline in concentrations (−15%/dec), indicating effective practices in reducing point sources. This is particularly impressive because the U.S. UB population has increased by 64%, from 167 million in 1980 to 274 million in 2020 (https://www.macrotrends.net/countries/USA/united-states/ urban-population). Such progress however has been offset by increasing UB discharge, leading to subdued reduction in TP fluxes (−4.8%/dec) compared to concentrations. In AG-dominant MX lands, concentrations declined (−6.6%/dec) but fluxes increased (14%/dec) due to increasing discharge (6.3%/dec). TP losses in CONUS have increased at 6.5%/decade over the past 40 y, especially in the Mississippi River Basin. Such increase echoes the global observation of increasing algae blooms in lakes since 1980s (46). The increasing concentrations and fluxes in AG rivers confirm the common perception that nutrient export and water quality in AG lands have not improved (47). USEPA recently adopted a comprised goal of reducing 20% of nutrient loads in the Mississippi River Basin by 2025 after failing the original goal of reducing 45% by 2015 (48). Similarly, states that drain to the Chesapeake Bay will likely, for the third time (previous in 2000 and 2010), fail to reduce 42% of N and 64% of P by 2025 (49).

The model identified discharge as the dominant driver for the trends of both concentrations and fluxes (Fig. 2 *E* and *F*). Discharge has been known to largely drive TP export (28, 42), as discharge increases soil erosion, which often carries large quantities of sorbed and particulate P. These results highlight the importance of land-river connectivity in shaping water quality and nutrient loss in rivers and streams (50). They also underscore the challenges of controlling nonpoint sources, soil erosion, and P loss in agricultural lands, which can be further exacerbated in a warming climate, especially in more frequent climate extremes (50).

Materials and Methods

Site Selection and Riverine TP Data. Data from 430 river basins were based on the Geospatial Attributes of Gages for Evaluating Streamflow dataset version II (GAGES-II) (51), a primary database for over 9,000 basins with long-term streamflow data in the United States. Compared to streamflow data, TP data are sparse, inconsistent, and have large gaps. To ensure sufficient training data and balance the spatial coverage (i.e., number of basins) and temporal coverage (i.e., number of data points in individual basins), we used the following criteria: 1)TP concentrations have at least 100 data points (grab samples) during 1980–2019; 2) daily discharge (Q) exist for at least 50% of days during 1980–2019. Daily area-normalized fluxes were calculated by multiplying daily concentrations and daily discharge normalized by basin drainage area. To reduce spatial autocorrelation, we excluded nested watersheds, leading to the selection of 430 independent basins for model training.

The selected 430 basins vary in drainage area, hydroclimate conditions, and land uses. These basins include 71 (17%) headwater basins (1st to 3rd stream orders), 283 (65%) medium basins (4th to 6th stream orders), and 76 (18%) larger basins (≥7th stream order). The mean (median) drainage areas of headwater, medium, and larger basins are 141 (97), 3,311 (1,696), and 21,214 (18,491) km^2, respectively. Mean annual precipitation varies from 201 to 1,944 mm/y, temperature from 1.75 to 23.3 °C, and discharge from less than 5.0 to 1,202 mm/y. The corresponding means (medians) are 1,008 (1,055) mm/y, 11.3 (10.6) °C, and 346 (342) mm/y, respectively. Basin classification follows the USGS practice (11), except UB has a lower threshold. Agricultural (AG) basins are defined as having >50% agricultural land and ≤5% UB land; UD basins have ≤5% UB land and ≤25% agricultural land; UB basin has >10% UB land and ≤25% agricultural land; MX basins are all other combinations of UB, agricultural, and UD lands. Following the GAGES-II method (51), agricultural lands are defined as planted and cultivated lands, which are the sum of classes 81 and 82 from the National Land Cover Database (NLCD). UB (developed) lands are the sum of classes 21, 22, 23, and 24 from the NLCD. These basins consist of 22 AG (5.1%), 92 UD (21%), 102 UB basin (24%), and 295 (50%) MX basins. The MX basins have average (±SD) area percentages of 47 (±24%), 28 (±23%), and 10 (±7.6%) for AG, forest, and UB components, respectively. The CONUS basin classification (Fig. 4) was similarly performed on HUC6 (Hydrologic Unit Code at the level 6) using the NLCD 2006, the same data and procedure used by the GAGES-II database. NLCD temporal maps also indicate minimal changes in land use in the past decades (52).

Discharge and TP data were downloaded from the USGS National Water Information System (https://waterdata.usgs.gov/nwis) using the dataRetrieval R package (53). All retrieved data were examined for outliers and errors. Discharge data are mostly continuous and available at 93 ± 14% temporal coverage for the study period, whereas TP data only cover small temporal fractions (1.7 ± 2.1%) at the coarser resolutions of monthly or bimonthly (*SI Appendix*, Fig. S2). To address the challenge of data sparsity, we consolidated TP data from individual rivers into one training dataset, thereby improving data spatiotemporal coverage. This consolidated dataset was then used in conjunction with a comprehensive set of temporally variable hydrometeorology data and static site characteristics (detailed in the following section). This data collation enables the model to leverage auxiliary information to learn and predict TP concentrations and fluxes.

The Multitask LSTM Model. The LSTM model, a type of recurrent neural network (RNN) model, learns directly from data in a sequential manner (34, 54). LSTM solves the problem of vanishing gradients in traditional RNNs and is designed to learn and keep information for longer periods using memory cells and gates. Each memory cell has three information gates (i.e., input, forget, and output gates) and two states (i.e., cell and hidden states) to store and pass information across time steps. This structure can learn long-term dependencies in natural systems such as soil moisture (55), streamflow (56), and riverine dissolved oxygen (20). Although LSTM models have shown better performance than traditional process-based or statistical models, they are often referred to as "black boxes" due to the challenge in interpreting the relationship between data variables and model prediction. Recent advances in LSTM models such as layer-wise relevance propagation can be adapted to obtain variable attributions to inform how each value in data contributes to model's prediction (57).

Here, we develop a multitask LSTM model instead of the traditional single-task models to simultaneously predict daily TP concentrations and fluxes from 1980 to 2019 for all 430 independent basins at the CONUS-scale. A joint prediction of concentration and flux can leverage shared information between these two variables with a better capture of the underlying dynamics of the system (45). By incorporating more observational constraints, multitask learning could enhance the model's robustness across different conditions (58). The model requires two types of input data: time-series hydrometeorological forcing and TP data, and static basin attributes. The forcing data drive the model at daily resolution, including daily discharge and seven daily meteorological variables of precipitation, day length, maximum and minimum air temperature, snow water equivalent, vapor pressure, and solar radiation. These forcing data are from a gridded meteorological dataset (DAYMET, https://daymet.ornl.gov/) (59) that were basin-aggregated using delineated watershed boundaries and Google Earth Engine (60). These boundary shapefiles are from the GAGES-II database (51). We also incorporated the timestamp as a time-series input to facilitate the dynamic learning of input–output relationships based on the year and season along with other watershed conditions (61). The timestamp serves as a latent variable representing the aggregated effects of human activities such as best management practices, tile drainage, and point sources that changed over time but are not represented by the time series of hydrometeorological forcings. They also cannot be directly quantified or used as model inputs due to limited data availability (35).

The basin attributes contain essential information about intrinsic hydroclimatic, land use, vegetation, and soil characteristics. They include 37 basin characteristics of topography, climate, hydrology, land use, soil, and geology that were obtained from the Google Earth Engine using the Caravan script (https://github.com/kratzert/Caravan) (62). They include basin elevation, slope, stream gradient, annual average of air temperature, precipitation, potential and actual evapotranspiration, global aridity index, climate moisture index, snow cover extent, natural discharge, land surface runoff, land use percentages of forest, cropland, pasture, irrigated area, permafrost, and wetland, soil component percentages of sand, silt, clay, and organic carbon content, soil erosion, and lithological classes and karst area extent, among others. These dynamic and static inputs were chosen based on data availability, our domain knowledge (36, 37), and prior LSTM modeling experience (20, 44, 58). Collectively, they provide a rich context (e.g., land use conditions) for the model to learn input–output relationships, spatiotemporal TP patterns, and fill data gaps.

Model Training and Performance Evaluation. Many environmental variables, including concentration, flux, and streamflow, have highly skewed distribution that could result in biased learning processes. To address this, we followed standard data preprocessing procedures before model training (56, 63). We first transformed time-series inputs and constant basin attributes using the $\log_{10}$ equation $v^* = \log_{10}(v+0.01)$ or the bestNormalize R package to make their distributions as close to Gaussian as possible. The $\log_{10}$ transformation is known to effectively reduce the skewness of raw data (*SI Appendix*, Fig. S6) and has been used routinely in LSTM modeling (20, 64). A standardization procedure was then used to transform inputs by subtracting the CONUS-scale mean and dividing by the CONUS-scale SD (56, 63). The training and testing datasets were standardized separately using the CONUS-scale mean and SD calculated for their respective time periods. Transformation and standardization improve numerical stability and model performance and reduce training time when model inputs span different scales and ranges. After model training, we transformed the input variables back to their original scale when interpreting model results, thereby minimizing potential impacts of the transformation and standardization on interpretability. We used a flexible scheme to split concentration data into the training (75%) and testing (25%) periods for each basin based on its temporal data distribution, to ensure sufficient data coverage for model training and for model testing. Flux data inherited the same training and testing splitting as concentration to ensure synchronous multitask training. Concentrations and fluxes have equal weights in the loss function of RMS Error (RMSE) during the training process.

Nash-Sutcliffe Efficiency (NSE) was used to measure the model performance (Eq. **1**) for each of 430 basins. NSE ranges from $-\infty$ to 1, with 1 being the perfect match between observation and model prediction. NSE < 0 indicates unacceptable performance where model prediction is worse than mean observations. NSE values ≥ 0.5 and ≥ 0.7 are considered as good model performance for daily concentration and flux (33), respectively.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_{mod,i} - y_{obs,i})^2}{\sum_{i=1}^{n}(y_{obs,i} - \overline{y_{obs}})^2}, \qquad [1]$$

where $y_{mod,i}$ are the model prediction at the time of observation data $y_{obs,i}$ and $\overline{y_{obs}}$ is the observation mean, n is the total number of paired model prediction and observation in the testing period.

Long-Term Trend Analysis. We quantified the decadal change rates using the TheilSen function from the R package *openair* (65), which allows for the seasonality of average monthly data to be detrended and is robust against outliers. Theil-Sen slopes have been commonly used to determine trends of water quality (66, 67). The monthly averages of model daily outputs were used to reduce autocorrelation and the "deseason" option of the function to account for potentially important seasonal influences. The "slope.percent" option was used to express slope estimates as a percentage change per year (%/year) and then multiplied it by 10 for decadal change rate (%/decade). The slope percentage is useful for comparing slopes for different water quality indicators (e.g., TP concentration vs. flux in different units) or comparing sites with very different concentration and flux levels. The trends for TP concentration and flux were determined by the sign of the slope change and their significance at level of 0.05 (Fig. 3). Specifically, increasing and decreasing trends were assigned when the P-value ≤ 0.05 with positive and negative slope changes, respectively, while insignificant trends were assigned when P-value > 0.05.

Feature Importance Analysis. To rank the importance of different factors, we used a well-established method based on integrated gradients (IG) to interpret predominant drivers that determine model outputs (68, 69). For each basin, the LSTM model generates a 14,610-d (40-y) prediction for two target features: TP concentration and TP flux. Local feature attributions to the model's prediction were estimated for each basin at each time point (Eq. **2**).

$$IG_t(x) = \frac{x}{n}\sum_{i=0}^{50}\frac{\partial f_t\left(\frac{i}{n}\cdot x\right)}{\partial x}, \qquad [2]$$

where $\frac{\partial f_t}{\partial x}$ denotes a gradient of the model function f at time point t with respect to input x. We used the Captum Python library (70) for its open-source implementation of IG, setting the number of steps (n) in the integral approximation to 50 (default). This operation was vectorized with respect to features, i.e., $IG_t(x)$ outputs a vector of size equal to the number of features.

To assess overall feature importance (*FI*), we aggregated the above feature attributions across all basins and time points using the mean of absolute values. The resulting *FI* scores were calculated as follows:

$$FI(X) = \sum_{t=0}^{14610}\frac{1}{N}\sum_{x\in X}\left|IG_t(x)\right|, \qquad [3]$$

where X represents a set of N basins. $FI(X)$ returns a vector of size equal to the number of features. When visualized on a bar plot for each target feature, *FI* scores provide insights into the most influential features driving the model's predictions. Here 14610 is the total number of days in the 40 years.

Hold-Out Test for Reproducing TP Trend in the Presence of Large Data Gap. In addition to the base case trained by the full data, here we ran an additional hold-out case to test the model's ability to fill data gap and reproduce historical trend in the presence of large data gap. We selected 14 data-rich basins that have evenly distributed data throughout the 40 y, and randomly held out an entire 8-y period of data (e.g., 1982–1989, 1992–1999, 2002–2009) for each basin, resulting in an average (±SD) percentage of hold-out data volume as 20 ± 8%. The 8-y hold-out periods of data were excluded from the training dataset and served as ground-truth data for testing. After model retraining, model results were checked against the reserved ground-truth data in the hold-out periods (hold-out NSE, Fig. 2 *C* and *D* and *SI Appendix*, Fig. S3). Long-term trends in terms of decadal change rates (i.e., %/dec) were also compared between data and model results (*SI Appendix*, Fig. S4). Despite the challenges posed by the sparse and inconsistent TP data, the hold-out test showcased the model's capability to robustly capture historical trends and fill data gaps.

HUC6 Prediction for CONUS Estimates. To upscale TP loss at the CONUS scale, the trained LSTM was applied to estimate TP fluxes from all 336 HUC6 basins at CONUS (Fig. 4, embedded map). The number of basins at the HUC6 level is comparable to the 430 independent basins included in the training dataset. The meteorological forcing and basin attributes for these HUC6 basins were retrieved from the same datasets of Daymet and Caravan as the training inputs (62). The mean and median area of these 336 HUC6 basins are 25,513, and 21,485 km^2, respectively, which are comparable to the size of large basins (21,214 and 18,491 km^2) that constitute 18% of the training data. Additionally, the land use type distribution of these 336 HUC6 basins generally aligns with the training dataset, comprising 4.5% AG basins, 35% UD basins, 11% UB basins, and 49% MX basins. While finer resolutions (HUC8 with 2,303 subbasins or HUC10 with 18,487 watersheds) could be used for CONUS-scale TP loss estimation, we leveraged the HUC6 data due to its similarity with the training dataset, which could minimize discrepancies when upscaling with the trained LSTM model.

To accommodate the lack of long-term discharge records, we derived daily discharge data for these HUC6 basins from a CONUS-wide LSTM streamflow model (63), specifically retrained at the HUC6 level. The LSTM streamflow model was trained with time-series data of precipitation, downward shortwave radiation, surface pressure, specific humidity, and air temperature (https://www.gloh2o.org), along with basin attributes including topography (elevation, slope, roughness), land use (fraction of developed land, forest, planted/cultivated land), soil properties (depth, porosity, bulk density, percentages of clay, silt, and clay), and lithology (carbonate sedimentary rock fraction). These static data were compiled from a variety of sources, including the Global Topography (https://www.earthenv.org/topography), the National Land Cover Database (https://www.mrlc.gov/data), the Harmonized World Soil Database v1.2 (https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases), the Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers (https://doi.org/10.3334/ORNLDAAC/1304), the GLobal HYdrogeology of permeability and porosity (https://doi.org/10.1002/2014gl059856), and the Global Lithological Map (https://doi.org/10.1594/PANGAEA.788537). The streamflow model exhibited robust performance across 3,213 USGS sites (*SI Appendix*, Fig. S5), achieving a median NSE of 0.76 under all flow conditions and 0.71 under high-flow conditions (Q ≥ 50th percentile) that dominate fluxes.

The assembled hydrometeorological and basin attribute data, and modeled streamflow data were used as input for the trained LSTM model to predict daily TP fluxes in each HUC6 basin, which were then used to estimate TP losses (Tg/y) by multiplying the corresponding drainage area and summing over the entire year. Total TP loss was summarized at the CONUS scale or by each land use categories (Fig. 4).

Data, Materials, and Software Availability. The dataRetrieval R package for downloading total phosphorus and discharge data is available at https://github.com/USGS-R/dataRetrieval (53). The meteorological dataset of DAYMET is available at https://daymet.ornl.gov (59). Basin attributes were obtained from the Caravan at https://github.com/kratzert/Caravan (62). The deep learning framework is available at https://github.com/WeiZhiWater/DeepWater (71). Basin information and attributes are available at https://github.com/WeiZhiWater/Phosphorus-basin-dataset (72). The predicted HUC6 streamflow (examples in *SI Appendix*, Fig. S7) can be accessed at https://huc06-prediction-e00dcd24c887.herokuapp.com (73).

ACKNOWLEDGMENTS. W.Z. was supported by the Barry and Shirley Isett Professorship (to L.L.) at Penn State University and the High Performance Computing Platform of Hohai University. L.L. was supported by the US National Science Foundation via the Hydrologic Science program (EAR- 2346471), Critical Zone Collaborative Network (EAR-2012123 and EAR-2012669), Frontier Research in Earth Sciences (EAR-2121621), and US Department of Energy Environmental System Science (DE-SC0020146). C.S. was supported by DE-SC0016605 from the Office of Biological and Environmental Research, U.S. Department of Energy. H.B. was supported from the state budget within the Polish Ministry of Education and Science program "Pearls of Science" (PN/01/0087/2022). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author affiliations: [a]The National Key Laboratory of Water Disaster Prevention, Yangtze Institute for Conservation and Development, Key Laboratory of Hydrologic-Cycle and Hydrodynamic-System of Ministry of Water Resources, College of Hydrology and Water Resources, Hohai University, Nanjing 210024, China; [b]Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA 16802; [c]MI2.AI, University of Warsaw, Warsaw 00-927, Poland; [d]Warsaw University of Technology, Warsaw 00-661, Poland; [e]Department of Ecosystem Science and Management, The Pennsylvania State University, University Park, PA 16802; [f]Institute of Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802; and [g]Virginia and West Virginia Water Science Center, United States Geological Survey, Richmond, VA 23228

1. P. M. Vitousek, S. Porder, B. Z. Houlton, O. A. Chadwick, Terrestrial phosphorus limitation: Mechanisms, implications, and nitrogen–phosphorus interactions. *Ecol. Appl.* **20**, 5–15 (2010).
2. S. R. Carpenter, E. M. Bennett, Reconsideration of the planetary boundary for phosphorus. *Environ. Res. Lett.* **6**, 014009 (2011).
3. P. Borrelli *et al.*, An assessment of the global impact of 21st century land use change on soil erosion. *Nat. Commun.* **8**, 2013 (2017).
4. E. M. Bennett, S. R. Carpenter, N. F. Caraco, Human impact on erodable phosphorus and eutrophication: A global perspective: Increasing accumulation of phosphorus in soil threatens rivers, lakes, and coastal oceans with eutrophication. *BioScience* **51**, 227–234 (2001).
5. Z. Yuan *et al.*, Human perturbation of the global phosphorus cycle: Changes and consequences. *Environ. Sci. Technol.* **52**, 2438–2450 (2018).
6. W. K. Dodds *et al.*, Eutrophication of U.S. freshwaters: Analysis of potential economic damages. *Environ. Sci. Technol.* **43**, 12–19 (2009).
7. J. J. Elser *et al.*, Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol. Lett.* **10**, 1135–1142 (2007).
8. G. Yang *et al.*, Phosphorus rather than nitrogen regulates ecosystem carbon dynamics after permafrost thaw. *Global Change Biol.* **27**, 5818–5830 (2021).
9. T. Zou, X. Zhang, E. A. Davidson, Global trends of cropland phosphorus use and sustainability challenges. *Nature* **611**, 81–87 (2022).
10. USEPA, National water quality inventory report to congress (United States Environmental Protection Agency, Washington, D.C., 1974).
11. N. M. Dubrovsky *et al.*, The quality of our Nation's waters–Nutrients in the Nation's streams and groundwater, 1992–2004. *USGS Circ.* **1350**, 174 (2010).
12. S. Sandström *et al.*, Particulate phosphorus and suspended solids losses from small agricultural catchments: Links to stream and catchment characteristics. *Sci. Total Environ.* **711**, 134616 (2020).
13. E. Sinha, A. M. Michalak, V. Balaji, Eutrophication will increase during the 21st century as a result of precipitation changes. *Science* **357**, 405–408 (2017).
14. C. Alewell *et al.*, Global phosphorus shortage will be aggravated by soil erosion. *Nat. Commun.* **11**, 4546 (2020).
15. USEPA, National rivers and streams assessment 2013-14: A Collaborative Survey (United States Environmental Protection Agency, EPA 841-R-19-001, Washington, D.C., 2020).
16. S. W. Ator, A. M. García, G. E. Schwarz, J. D. Blomquist, A. J. Sekellick, Toward explaining nitrogen and phosphorus trends in Chesapeake Bay Tributaries, 1992–2012. *J. Am. Water Res. Assoc.* **55**, 1149–1168 (2019).
17. R. B. Alexander *et al.*, Differences in phosphorus and nitrogen delivery to the Gulf of Mexico from the Mississippi River Basin. *Environ. Sci. Technol.* **42**, 822–830 (2008).
18. A. H. W. Beusen, A. L. M. Dekkers, A. F. Bouwman, W. Ludwig, J. Harrison, Estimation of global river transport of sediments and associated particulate C, N, and P. *Global Biogeochem. Cycles* **19**, GB4S05 (2005).
19. C. Varadharajan *et al.*, Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrol. Process.* **36**, e14565 (2022).
20. W. Zhi *et al.*, From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environ. Sci. Technol.* **55**, 2357–2368 (2021).
21. K. Fang, D. Kifer, K. Lawson, D. Feng, C. Shen, The data synergy effects of time-series deep learning models in hydrology. *Water Res. Res.* **58**, e2021WR029583 (2022).
22. C. Shen, X. Chen, E. Laloy, Editorial: Broadening the use of machine learning in hydrology. *Front. Water* **3**, 681023 (2021).
23. Q. Zhang, Synthesis of nutrient and sediment export patterns in the Chesapeake Bay watershed: Complex and non-stationary concentration-discharge relationships. *Sci. Total Environ.* **618**, 1268–1283 (2018).
24. D. Kerins, L. Li, High dissolved carbon concentration in arid rocky mountain streams. *Environ. Sci. Technol.* **57**, 4656–4667 (2023).
25. L. Li *et al.*, Climate controls on river chemistry. *Earth's Future* **10**, e2021EF002603 (2022).
26. K. Sadayappan, D. Kerins, C. Shen, L. Li, Nitrate concentrations predominantly driven by human, climate, and soil properties in US rivers. *Water Res.* **226**, 119295 (2022).
27. S. E. Godsey, J. Hartmann, J. W. Kirchner, Catchment chemostasis revisited: Water quality responds differently to variations in weather and climate. *Hydrol. Process.* **33**, 3056–3069 (2019).
28. D. M. Robertson, D. A. Saad, Nitrogen and phosphorus sources and delivery from the Mississippi/Atchafalaya River Basin: An update using 2012 SPARROW models. *J. Am. Water Res. Assoc.* **57**, 406–429 (2021).
29. S. B. Bricker *et al.*, Effects of nutrient enrichment in the nation's estuaries: A decade of change. *Harmful Algae* **8**, 21–32 (2008).
30. K. Bugica, B. Sterba-Boatwright, M. S. Wetz, Water quality trends in Texas estuaries. *Mar. Pollut. Bull.* **152**, 110903 (2020).
31. S. P. Funk *et al.*, Assessment of impacts of diphenyl phosphate on groundwater and near-surface environments: Sorption and toxicity. *J. Contam. Hydrol.* **221**, 50–57 (2019).
32. D. Ren *et al.*, Modeling and assessing water and nutrient balances in a tile-drained agricultural watershed in the U.S. Corn Belt. *Water Res.* **210**, 117976 (2022).
33. D. N. Moriasi, M. W. Gitau, N. Pai, P. Daggupati, Hydrologic and water quality models: Performance measures and evaluation criteria. *T Asabe* **58**, 1763–1785 (2015).
34. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
35. S. Stackpoole, R. Sabo, J. Falcone, L. Sprague, Long-term Mississippi River trends expose shifts in the river load response to watershed nutrient balances between 1975 and 2017. *Water Res. Res.* **57**, e2021WR030318 (2021).
36. W. Zhi, L. Li, The shallow and deep hypothesis: Subsurface vertical chemical contrasts shape nitrate export patterns from different land uses. *Environ. Sci. Technol.* **54**, 11915–11928 (2020).
37. W. Zhi *et al.*, Distinct source water chemistry shapes contrasting concentration-discharge patterns. *Water Res. Res.* **55**, 4233–4251 (2019).
38. M. Lapointe, C. M. Rochman, N. Tufenkji, Sustainable strategies to treat urban runoff needed. *Nat. Sustain.* **5**, 366–369 (2022).
39. S. M. Powers *et al.*, Long-term accumulation and transport of anthropogenic phosphorus in three river basins. *Nat. Geosci.* **9**, 353–356 (2016).
40. A. Civan, F. Worrall, H. P. Jarvie, N. J. K. Howden, T. P. Burt, Forty-year trends in the flux and concentration of phosphorus in British rivers. *J. Hydrol.* **558**, 314–327 (2018).
41. A. Sharpley *et al.*, Phosphorus legacy: Overcoming the effects of past management practices to mitigate future water quality impairment. *J. Environ. Qual.* **42**, 1308–1326 (2013).
42. Z. Bian *et al.*, A century-long trajectory of phosphorus loading and export from Mississippi River Basin to the Gulf of Mexico: Contributions of multiple environmental changes. *Global Biogeochem. Cycles* **36**, e2022GB007347 (2022).
43. M. M. Mekonnen, A. Y. Hoekstra, Global anthropogenic phosphorus loads to freshwater and associated grey water footprints and water pollution levels: A high-resolution global study. *Water Res. Res.* **54**, 345–358 (2018).
44. W. Zhi, W. Ouyang, C. Shen, L. Li, Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. *Nat. Water* **1**, 249–260 (2023).
45. W. Zhi, A. P. Appling, H. E. Golden, J. Podgorski, L. Li, Deep learning for water quality. *Nat. Water* **2**, 228–241 (2024).

46. J. C. Ho, A. M. Michalak, N. Pahlevan, Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature* **574**, 667–670 (2019).

47. N. B. Basu *et al.*, Managing nitrogen legacies to accelerate water quality improvement. *Nat. Geosci.* **15**, 97–105 (2022).

48. UPEPA, States develop new strategies to reduce nutrient levels in Mississippi River, Gulf of Mexico (2015). https://www.epa.gov/archive/epa/newsreleases/states-develop-new-strategies-reduce-nutrient-levels-mississippi-river-gulf-mexico.html. Accessed 14 February 2023.

49. K. Blankenship With, "Chesapeake Bay goal out of reach, region plans for what's next" in *The Southern Maryland Chronicle*, K. Blankenship, Ed. (The Southern Maryland Chronicle, Oakland, MD, 2022).

50. L. Li *et al.*, River water quality shaped by land–river connectivity in a changing climate. *Nat. Clim. Change* **14**, 225–237 (2024).

51. J. A. Falcone, GAGES-II: Geospatial attributes of gages for evaluating streamflow (US Geological Survey, 2011).

52. X. Li, H. Tian, S. Pan, C. Lu, Four-century history of land transformation by humans in the United States: 1630–2020 (Copernicus GmbH, 2022).

53. R. M. Hirsch, L. A. De Cicco, User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data (US Geological Survey, 2015).

54. K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 2222–2232 (2016).

55. K. Fang, M. Pan, C. P. Shen, The value of SMAP for long-term soil moisture estimation with the help of deep learning. *Ieee T Geosci. Remote* **57**, 2221–2233 (2019).

56. D. Feng, K. Fang, C. Shen, Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Res. Res.* **56**, e2019WR026793 (2020), 10.1029/2019WR026793.

57. L. Arras *et al.*, "Explaining and interpreting LSTMs" in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, Eds. (Springer International Publishing, 2019), pp. 211–238, 10.1007/978-3-030-28954-6_11.

58. W. Zhi, C. Klingler, J. Liu, L. Li, Widespread deoxygenation in warming rivers. *Nat. Clim. Change* **13**, 1105–1113 (2023).

59. P. E. Thornton *et al.*, Daymet: Daily surface weather data on a 1-km Grid for North America, version 3 (ORNL Distributed Active Archive Center, 2016).

60. N. Gorelick *et al.*, Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).

61. M. H. Nour, D. W. Smith, M. G. El-Din, E. E. Prepas, The application of artificial neural networks to flow and phosphorus dynamics in small streams on the Boreal Plain, with emphasis on the role of wetlands. *Ecol. Model.* **191**, 19–32 (2006).

62. F. Kratzert *et al.*, Caravan - A global community dataset for large-sample hydrology. *Sci. Data* **10**, 61 (2023). https://doi.org/10.1038/s41597-023-01975-w.

63. W. Ouyang *et al.*, Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *J. Hydrol.* **599**, 126455 (2021).

64. F. Rahmani *et al.*, Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett.* **16**, 024025 (2021), 10.1088/1748-9326/abd501.

65. D. C. Carslaw, K. Ropkins, openair—An R package for air quality data analysis. *Environ. Model. Softw.* **27–28**, 52–61 (2012).

66. S. S. Kaushal *et al.*, Rising stream and river temperatures in the United States. *Front. Ecol. Environ.* **8**, 461–466 (2010).

67. W. Ni, M. Li, J. M. Testa, Discerning effects of warming, sea level rise and nutrient management on long-term hypoxia trends in Chesapeake Bay. *Sci. Total Environ.* **737**, 139717 (2020).

68. M. Sundararajan, A. Taly, Q. Yan, "Axiomatic attribution for deep networks" in *Proceedings of the 34th International Conference on Machine Learning*, P. Doina, T. Yee Whye, Eds. (PMLR, Proceedings of Machine Learning Research, 2017), pp. 3319–3328.

69. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, "Lecture notes in computer science" in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer Nature, 2019), vol. 11700.

70. N. Kokhlikyan *et al.*, Captum: A unified and generic model interpretability library for Pytorch. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2009.07896 (Accessed 20 March 2024).

71. W. Zhi, H. Baniecki, DeepWater. GitHub. https://github.com/WeiZhiWater/DeepWater. Accessed 16 August 2024.

72. W. Zhi, L. Li, Phosphorus basin dataset. GitHub. https://github.com/WeiZhiWater/Phosphorus-basin-dataset. Accessed 28 June 2024.

73. J. Liu, HUC06 Streamflow Prediction. https://huc06-prediction-e00dcd24c887.herokuapp.com. Accessed 22 August 2024.