Science of the Total Environment 889 (2023) 164138

Contents lists available at ScienceDirect

Science of the Total Environment

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

Human and natural impacts on the U.S. freshwater salinization and alkalinization: A machine learning approach

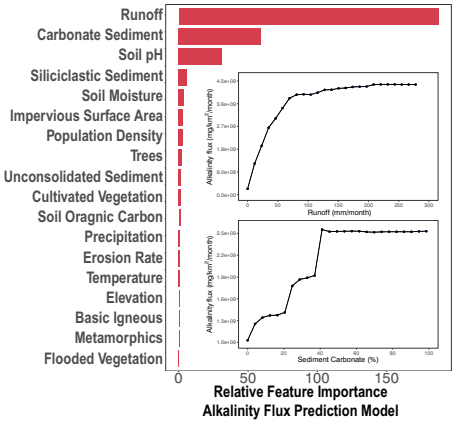
Beibei E [a], Shuang Zhang [b], Charles T. Driscoll [c], Tao Wen [a,*]

[a] Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY 13244, United States
[b] Department of Oceanography, Texas A&M University, College Station, TX 77843, United States
[c] Department of Civil and Environmental Engineering, Syracuse University, Syracuse, NY 13244, United States

HIGHLIGHTS

- Machine learning models detect causes of salinization syndrome in U.S. rivers.
- Human activities, e.g., urbanization, are the main source of U.S. river salinity.
- Alkalinization in U.S. rivers is governed mainly by natural processes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Jay Gan

Keywords:
Salinization
Alkalinization
Road salt
Rivers
Weathering
Machine learning

ABSTRACT

Ongoing salinization and alkalinization in U.S. rivers have been attributed to inputs of road salt and effects of human-accelerated weathering in previous studies. Salinization poses a severe threat to human and ecosystem health, while human derived alkalinization implies increasing uncertainty in the dynamics of terrestrial sequestration of atmospheric carbon dioxide. A mechanistic understanding of whether and how human activities accelerate weathering and contribute to the geochemical changes in U.S. rivers is lacking. To address this uncertainty, we compiled dissolved sodium (salinity proxy) and alkalinity values along with 32 watershed properties ranging from hydrology, climate, geomorphology, geology, soil chemistry, land use, and land cover for 226 river monitoring sites across the coterminous U.S. Using these data, we built two machine-learning models to predict monthly-aggregated sodium and alkalinity fluxes at these sites. The sodium-prediction model detected human activities (represented by population density and impervious surface area) as major contributors to the salinity of U.S. rivers. In contrast, the alkalinity-prediction model identified natural processes as predominantly contributing to variation in riverine alkalinity flux, including runoff, carbonate sediment or siliciclastic sediment, soil pH and soil moisture. Unlike prior studies, our analysis suggests that the alkalinization in U.S. rivers is largely governed by local climatic and hydrogeological conditions.

1. Introduction

Salinization and alkalinization in freshwaters can threaten drinking water supplies, impair freshwater biodiversity, accelerate corrosion of

* Corresponding author at: Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY 13244, United States.
E-mail address: twen08@syr.edu (T. Wen).

http://dx.doi.org/10.1016/j.scitotenv.2023.164138
Received 21 February 2023; Received in revised form 29 April 2023; Accepted 9 May 2023
Available online 13 May 2023

infrastructure, and mobilize inorganic and organic pollutants (Cañedo-Argüelles et al., 2014; DeVilbiss et al., 2021; Duan and Kaushal, 2015; Dugan et al., 2017; Hintz and Relyea, 2019; Kaushal, 2016). Over the past several decades, salinity and alkalinity in freshwaters have increased in the United States (U.S.) and worldwide (Corsi et al., 2015; Corwin, 2021; Duan and Kaushal, 2015; Findlay and Kelly, 2011; Kaushal et al., 2005, 2014, 2018; Likens and Buso, 2010; Thorslund et al., 2021). Investigators have suggested that main anthropogenic contributors to salinization and alkalinization of U.S. surface waters include road salt deicers and other industrial sources, wastewater discharge, groundwater irrigation, saltwater inundation caused by sea-level rise, and human-accelerated weathering (Bhide et al., 2021; Kaushal et al., 2013, 2017, 2018; Thorslund et al., 2021; Barnes and Raymond, 2009). These anthropogenic factors augment natural processes, such as natural weathering of rock and soil, dissolved ions in precipitation, and sea spray aerosols in coastal areas (Kaushal et al., 2013; Meybeck, 2003). Among these processes, road salt and human-accelerated weathering have been suggested as the dominant drivers of the observed salinization and alkalinization of U.S. rivers and streams over the past several decades on a continental scale (e.g., Kaushal et al., 2018). Increased urbanization and ongoing climate change have been thought to be the main driver of the ubiquitous salinization of rivers (e.g., Kaushal and Belt, 2012; Kelleher et al., 2020). In arid and semi-arid regions, salinization can be significantly accelerated during the dry season, leading to further loss of soil moisture and suppression of leaching events (Perri et al., 2022; Perri et al., 2020). Although this study focuses on surface water, it is worth noting that the increase in salinity and alkalinity in groundwater driven by human activities is not uncommon. For example, Hansen et al. (2018) suggested the increase in salinity and alkalinity in groundwater in San Joaquin Valley (California, U.S.) was primarily driven by agricultural practices.

Alkalinity in streams and rivers can be derived from both natural processes such as rock weathering and microbial processing (Kaushal et al., 2018; Regnier et al., 2022; Zhang and Planavsky, 2020), and anthropogenic processes including the interaction between human activities and watershed properties (e.g., bedrock geology). Recently, it has been suggested that human activity is driving a long-term alkalinity increase (i.e., alkalinization) in the U.S. rivers (Kaushal et al., 2018). Furthermore, previous studies indicate that freshwater salinization could lead to a further increase in river alkalinity. For example, Duan and Kaushal (2015) found that dissolved inorganic carbon (DIC; largely represents alkalinity) increased with salt inputs. In urban settings, increases in impervious surface area (Foley et al., 2005; Grimm et al., 2008) coincide with increased pH and DIC, which is thought to be linked to accelerated weathering of impervious surfaces (Kaushal et al., 2017). The few studies that have suggested that human-accelerated weathering contributes to surface water salinization and alkalinization have focused on urban areas. A broad understanding of the contribution of human-accelerated weathering to overall riverine alkalinity is limited, especially on the national or continental scale. Riverine flux of alkalinity forms a vital link in the Earth's atmosphere-land-ocean system. It acts to regulate the global atmospheric carbon cycle over multiple time scales. Therefore, a mechanistic understanding of the drivers of freshwater alkalinization will be essential to quantify atmospheric carbon dioxide sequestration and to project future changes in the geochemical composition of rivers and global carbon cycling under a changing climate.

In this study, we compiled sodium (as a salinity proxy) and alkalinity from the U.S. Geological Survey (USGS) river gauges that are used to monitor water quantity and quality in rivers and streams across the U.S. Sodium rather than chloride was selected as the salinity proxy to be consistent with the previous study (Kaushal et al., 2018). Moreover, sodium and chloride fluxes are highly correlated (Fig. S1). Therefore, we anticipate that the conclusions drawn for sodium fluxes would also be applicable to chloride fluxes. For these sites, we delineated their watershed areas and calculated and determined corresponding watershed properties. These watershed properties include features of hydrology, climate, geomorphology, geology, soil chemistry, land use, and land cover. We then developed machine learning models to predict salinity and alkalinity in U.S. rivers using selected watershed properties. In the model development phase, relative importance of watershed properties was evaluated to quantitatively assess the roles that human activities and natural processes play in regulating the salinity and alkalinity of U.S. rivers. Here we aimed to address two questions: (1) what factors are driving the spatiotemporal variation of watershed-level salinity and alkalinity in U.S. rivers on a continental scale? and (2) How does the predicted watershed-level chemical composition in U.S. rivers interact with each individual watershed property?

2. Materials and methods

2.1. Data acquisition and processing

A total of 226 USGS river sites (Fig. 1) were selected as study sites. These sites are a subset of 232 sites investigated in previous research (Kaushal et al., 2018) which concluded that human-accelerated weathering and road salts were the main drivers of salinization and alkalinization in U.S. rivers. These sites were selected because long-term continuous water quality measurements have been recorded for at least 30 years from 1942 to 2021. Six of the 232 sites were removed in this study due to three reasons: 1) three sites are located on human-built canals, 2) one site has a large uncertainty in the drainage area estimation, 3) two sites do not have paired hydrologic and water quality datasets. These 226 sites are located across six geographic zones: Northeast (n = 43), Southeast (n = 31), Midwestern (n = 62), Southwest (n = 65), Northwest (n = 21), and Pacific (n = 4) (Fig. 1). The corresponding watershed of each gauging station was delineated using the elevation, derived flow direction, and flow accumulation layers from HydroSHEDS (https://www.hydrosheds.org). Water chemistry data collected at these USGS monitoring sites were assumed to be the integrated product of solutes released from the corresponding watersheds.

Riverine sodium and alkalinity measurements were acquired from the Water Quality Portal (Read et al., 2017). To estimate the long-term impact of human activities and natural processes on surface water chemistry for each USGS site, we first aggregated daily-sodium and alkalinity measurements by year and month before calculating the annual-averaged monthly water chemistry measurements. After aggregation, we retained a total of 2685 sodium and 2691 alkalinity measurements, which were then normalized by watershed area to calculate the flux (mg/km^2/month) using Eq. (1).

$$\boldsymbol{F} = \frac{\boldsymbol{IAC} \times \boldsymbol{D} \times 28.32 \frac{liter}{ft^3} \times 3600 \frac{second}{hour} \times 24 \frac{hour}{day} \times 30 \frac{day}{month}}{\boldsymbol{WA}} \quad (1)$$

where $\boldsymbol{F}$ represents sodium or alkalinity flux (mg/km^2/month), $\boldsymbol{IAC}$ is the monthly inorganic analyte concentration (mg/L), $\boldsymbol{D}$ denotes river discharge (ft^3/s) while $\boldsymbol{WA}$ is watershed area (km^2). For alkalinity flux, it is expressed in mg $CaCO_3$/km^2/month. The numbers of 28.32, 3600, 24, and 30 are unit conversion factors.

Based on domain knowledge (see also detailed descriptions of attributes in Section 2.3), we selected and derived 32 watershed features (Table 1) that could potentially contribute to the salinization and alkalinization in U.S. rivers on the continental scale. These features include characteristics of hydrology (n = 2), climate (n = 2), geomorphology (n = 5), soil chemistry (n = 2), geology (n = 10), land use (n = 4), and land cover (n = 7). The source of each data layer is included in Table 1.

2.2. Random Forest model

Random Forest (RF) model is a machine learning technique first proposed in the 1990s (Ho, 1995) and further refined in the 2000s (Breiman, 2001). RF is based on the principle of classification and regression trees (CART) (Liaw and Wiener, 2002), which can be used for both classification and regression tasks. Compared with traditional CART models, RF has the benefit of avoiding overfitting while improving prediction accuracy. A RF model ensemble prediction results from many sub-models (i.e., decision tree) with each 'tree' model making independent predictions for the variable of interest (i.e., target variable) based on the input data (i.e., predictor variables or features). Each 'tree' model is built based on a

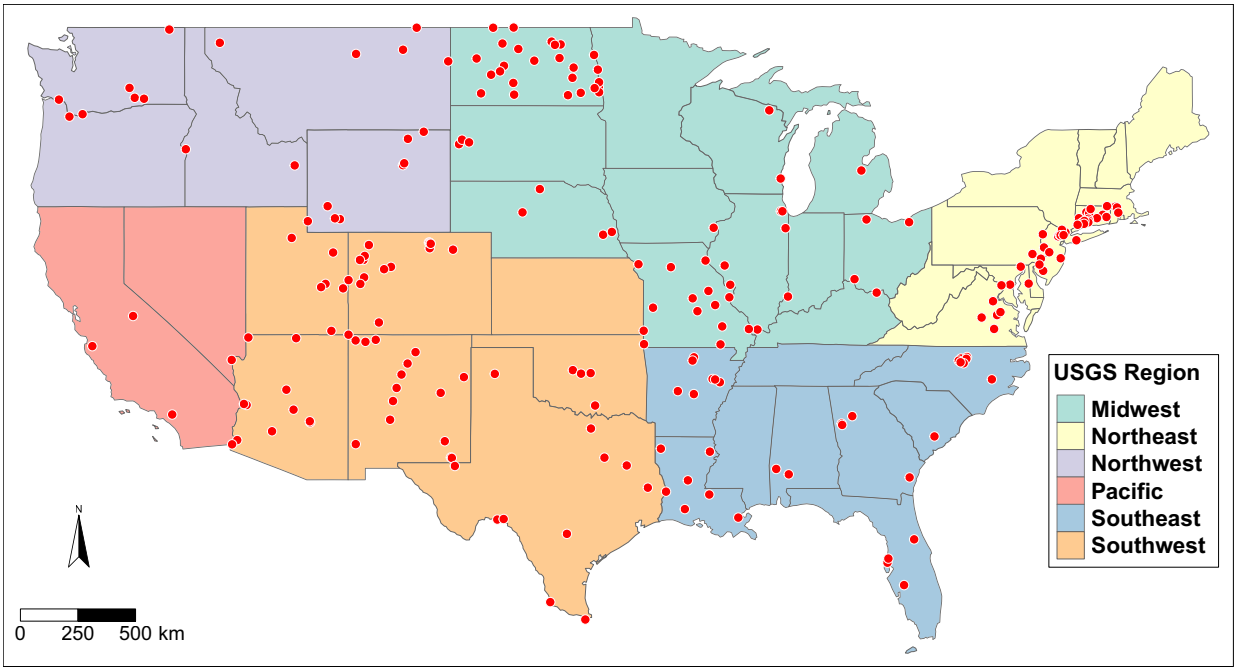


Fig. 1. Location and region of 226 USGS river monitoring sites used in this study. These sites are a subset of USGS sites previously used by Kaushal et al. (2018).

Table 1
List of all 32 watershed attributes along with their corresponding unit, feature category, and source. Features selected for the model development are marked by *.

Feature name	Unit	Category	Source
Runoff*	mm per month	Hydrology	Ghiggi et al. (2021)
Soil moisture*	m^3 per m^3	Hydrology	Wang et al. (2021)
Temperature*	Celsius	Climate	Karger et al. (2017)
Precipitation*	mm per month	Climate	
Watershed area	km^2	Geomorphology	This study
Erosion rate*	mm per year	Geomorphology	Amatulli et al. (2020); Larsen et al. (2014)
Elevation*	Meter	Geomorphology	Amatulli et al. (2020)
Slope	Degree	Geomorphology	
Aspect	Degree	Geomorphology	
Soil organic carbon*	g per kg	Soil chemistry	Poggio et al. (2021)
Soil pH*	pH unit	Soil chemistry	
Evaporite	Percentage	Geology	Hartmann and Moosdorf (2012)
Carbonate sediment*	Percentage	Geology	
Siliciclastic sediment*	Percentage	Geology	
Pyroclastic sediment	Percentage	Geology	
Mixed sediment	Percentage	Geology	
Unconsolidated sediment*	Percentage	Geology	
Igneous basic*	Percentage	Geology	
Igneous intermediate	Percentage	Geology	
Igneous acid	Percentage	Geology	
Metamorphic*	Percentage	Geology	
Impervious surface area*	Percentage	Land use	Brown de Colstoun et al., 2017
Population density*	# per km^2	Land use	Center For International Earth Science Information Network (2016)
Cultivated vegetation*	Percentage	Land use	Tuanmu and Jetz (2014)
Urban	Percentage	Land use	
Trees*	Percentage	Land cover	
Shrubs	Percentage	Land cover	
Herbaceous vegetation	Percentage	Land cover	
Flooded vegetation*	Percentage	Land cover	
Snow ice	Percentage	Land cover	
Barren	Percentage	Land cover	
Water	Percentage	Land cover	

random subset of the data. In each 'tree' model, at each split, a random set of predictor variables is selected for evaluation to divide the training data (Breiman, 2001).

Random Forest models have been increasingly used in geoscience and environmental science research. For example, Tesoriero et al. (2017) used a RF approach to predict the probability of high levels of a redox-sensitive contaminant in groundwater. Le et al. (2019) applied a RF model to forecast changes in specific conductance in surface waters that were likely to occur for the period 2070 to 2100 in response to climate change in Germany. Estévez et al. (2019) utilized a RF model to investigate the most important drivers of spatial patterns of water salinity in Spain. Also, a RF model was employed to predict satellite-derived fire severity classes from geospatial datasets of fire history, topographic setting, weather, and vegetation type (Harris and Taylor, 2017). Carranza et al. (2021) applied the RF model to estimate root zone soil moisture in data-poor regions. Other types of tree-based models (e.g., XGBoost, boosted regression tree) have also been applied to resolve hydrogeochemical questions (Choubin et al., 2018; Erickson et al., 2021a, 2021b; Ransom et al., 2022).

In this study, we developed and applied RF models to predict sodium and alkalinity fluxes in U.S. rivers using watershed properties as predictor variables. In the model training and development phase, we fine-tuned the model to find the optimal values for three model hyperparameters: num.trees (i.e., the number of sub-models), min.node.size (i.e., the minimal size of the tree branch in each sub-model), and mtry (i.e., the number of predictor variables selected at each split). In the model development process, we employed a random partitioning strategy to divide the entire dataset into 10 subsets. One of these subsets was set aside for evaluation purposes (i.e., hold-out dataset), while the remaining nine subsets were used to train the model. We randomly selected data on a sample-basis rather than on a site-basis. In addition to making predictions for river salinity and alkalinity fluxes, we also used RF model outputs to evaluate the relative importance of predictor variables in the prediction tasks by calculating and comparing the Conditional Permutation Importance (CPI) of all selected predictor variables. A higher CPI value indicates that the corresponding predictor variable plays a more important role in predicting solute flux. We also used partial dependence plots (PDP) (Le et al., 2019) to visualize the relationships between target variables (i.e., riverine sodium or alkalinity flux) and each important

individual predictor variable. The construction of the RF model and the calculation of variable importance were conducted in R (R Core Team, 2017) using the 'mlr' package (Bischl et al., 2016).

2.3. Feature selection workflow and model evaluation

The goal of the feature selection workflow is to remove predictor features which can compete for and share importance in predicting the target variable and to retain the most relevant predictor features based on either statistics or domain knowledge. To remove redundant features, we developed a workflow to assign a 'priority score' to each of the 32 predictor features. Predictor features with a 'priority score' ≥ 3 were included in the RF model, while predictor features with a 'priority score' ≤ 2 might be selected based on our research interests (Table S1). The priority score was calculated as the sum of three sub-scores. First, the predictor features with a coefficient of variation (CV) value ≥ 10 were assigned a sub-score of 1 with others rated as 0 (Table S1; 'CV' column). For a predictor feature that has a small coefficient of variation, given that the target feature is highly variable, such predictor feature will likely be deemed to be contributing little to the prediction of the target variable, i.e., this predictor feature is statistically irrelevant. The second sub-score was based on the Spearman correlation analysis results for each pair of the predictor and target variables. The variable pair with a Spearman's $r > 0.4$ and a $p < 0.05$ was defined as a statistically significantly correlated pair. Predictor features correlated with no more than two other predictor features were assigned a score of 1. Predictor features showing a correlation with at least three features were assigned a score of 1 only if they best represented the predictor features from their own feature category (e.g., land use, land cover) based on domain knowledge. When two or more such predictor features from the same feature category exist, the one that showed a higher Spearman's r ($p < 0.05$) with the target variable was assigned a score of 1. All other predictor features were assigned a score of 0 (Table S1; 'Uncorrelation' column). Finally, the ranking of predictor features in a preliminary RF model that predicted solute flux using all 32 predictor features yielded the third sub-score. The top 10 features were assigned a score of 2. Features ranked 11 to 20 and below 20 were assigned a score of 1 or 0, respectively (Table S1; 'CPI' column).

A total of 18 features were chosen and used to develop two RF models for predicting riverine sodium and alkalinity fluxes, respectively, which included features of hydrology (n = 2), climate (n = 2), geomorphology (n = 2), soil moisture (n = 2), geology (n = 5), land use (n = 3), and land cover (n = 2) (Table 1). Among these features, seven were retained based on domain knowledge regardless of their priority scores. These seven features include soil moisture, soil pH, soil organic carbon, siliciclastic sediment, erosion rate, flooded vegetation, and igneous basic rock. These features represent the earth surface processes that regulate the generation and transport of salinity and alkalinity into rivers (e.g., Brantley et al., 2008; Wen et al., 2022). To prevent the competition of feature important indices, other comparable predictor features were not selected (e.g., urban land cover). During model development, both model error and relative error were calculated for test datasets using Eqs. (2) and (3) to check spatio-temporal prediction heterogeneity in model.

$$E = PIF – OIF \tag{2}$$

$$RE(\%) = ((PIF – OIF)/OIF) * 100 \tag{3}$$

where E and RE are calculated model error and relative error, while PIF and OIF represent predicted and observed (ground truth) inorganic solute fluxes, respectively.

To explore if the fine-tuned machine learning model had different performances over space and time, we also assessed the relationship between the frequency of monthly measurement showing a large relative error and the number of samples in a state, as well as the relationship between the frequency of monthly measurement showing a large relative error and the month of the observations. Three thresholds of relative error were used: 30 %, 50 %, 100 % with a particular emphasis on the threshold of 30 %. We first performed such assessment for the sodium flux prediction model to ensure that the trained RF model yielded a great performance before analyzing the results from the alkalinity flux prediction model.

3. Results and discussion

3.1. Evaluation of sodium flux prediction model

Optimal hyperparameters for both sodium and alkalinity riverine flux prediction models are listed in the supplementary information (Table S2). Calculated mean squared error (MSE) in the training phase for both models are plotted as a function of mtry (1–18), num.trees (100–1500), and min.node.size (1–15) (Fig. S2). The trained model explained 87 % of the variation of the target variable in the hold-out dataset (Fig. S3a), indicating a satisfactory model performance in predicting riverine sodium flux for the coterminous U.S. In addition, calculated residual values showed no significant correlation with the predicted sodium flux, confirming the generalizability of the trained model in predicting sodium flux (Fig. S3b).

We also explored if the trained RF model exhibited any prediction heterogeneity over space and time. With respect to the spatial considerations, we color coded each state by the percentage of monthly data points reporting >30 % relative errors for test data of gauging stations in that state (Fig. S4a), which showed no particular geospatial patterns in the presence of highly erroneous predictions. The percentage of monthly data points showing >30 % relative errors at a site showed no statistically significant correlation with the total number of monthly measurements in each state ($p > 0.05$) (Fig. S4b). These findings suggest that the trained RF model was unlikely biased for spatial projections. As for temporal considerations, we plotted the percentage of samples in each month reporting >30 % relative errors as a function of month of the year for the conterminous U.S. (Fig. S4c). The percentage of monthly data points with over 50 % and 100 % relative errors were also plotted for reference. No month yielded significantly greater erroneous predictions of riverine sodium flux than others, suggesting that the trained RF model was unlikely to be biased across different months.

As a summary, the RF model of sodium flux manifests a consistently good performance over space and time on a continental scale. This analysis also suggests that the RF model may be applied to explore other solute fluxes in U.S. rivers.

3.2. Human activities are the main driver of salinization of U.S. rivers

The optimized riverine sodium flux prediction model also provided conditional permutation importance (CPI) values for all 18 selected predictors (Fig. 2). A predictor feature with a higher CPI value shows a larger weight

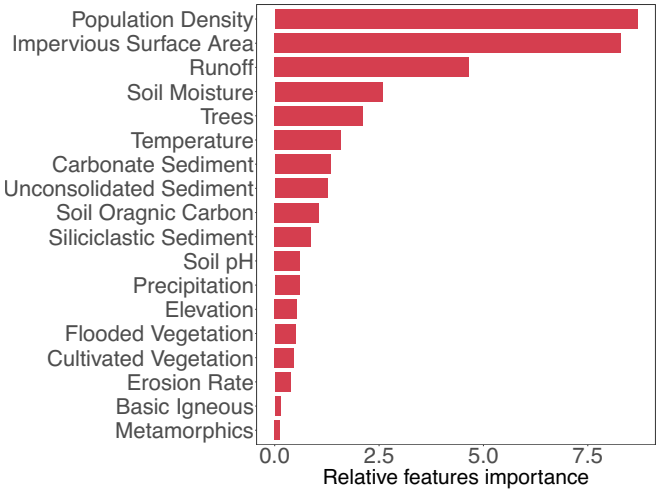


Fig. 2. Predictor feature ranking based on conditional permutation importance values from the riverine sodium flux prediction model.

(i.e., greater importance) in predicting riverine sodium flux. The five most important predictor features of the river sodium flux prediction model included two human activities related features (population density, impervious surface percentage), two hydrologic features (runoff, soil moisture), and a land cover characteristic (tree cover). The two human-related features had the highest ranking, strongly suggesting that human activities serve as the main driver of variation in sodium flux in U.S. rivers at a watershed-level. These human activities likely include road salt use, urbanization, and industrialization as suggested by previous studies (e.g., Abbott et al., 2019; ; Kaushal et al., 2013, 2018; Kaushal and Belt, 2012; Likens and Buso, 2010; Thorslund et al., 2021; Utz et al., 2022), and these activities have been thought to release large amounts of sodium into U.S. rivers, leading to marked increases in salinization over the past 80 years (Kaushal et al., 2018). In this study, we define human-related features as those that entail direct engagement in human activities, which can significantly contribute to river chemistry, such as road salting.

Partial dependence plot results show that sodium flux increases with increases in population density and percentage of impervious surface area within the watershed (i.e., top two features; Fig. 3a and b). More intensive human activities likely release more salt into surface waters, which in turn increase watershed-level sodium flux. In the initial increasing phase of partial dependence plots of impervious surface area and population density, river sodium flux experiences a marked increase before leveling off at a relatively constant rate. Such pattern might reflect the condition that as more salts are released into the environment, the transport capacity (e.g., overland runoff) can be saturated reaching a limit of sodium transport via rivers. Runoff and soil moisture were ranked as the third and fourth most important features in sodium flux prediction model. This pattern suggests a significant hydrological control on watershed-level sodium flux. With both features, sodium flux increases markedly at low (or intermediate) values of these hydrologic features before reaching a plateau (Fig. 3c and d). In the increasing

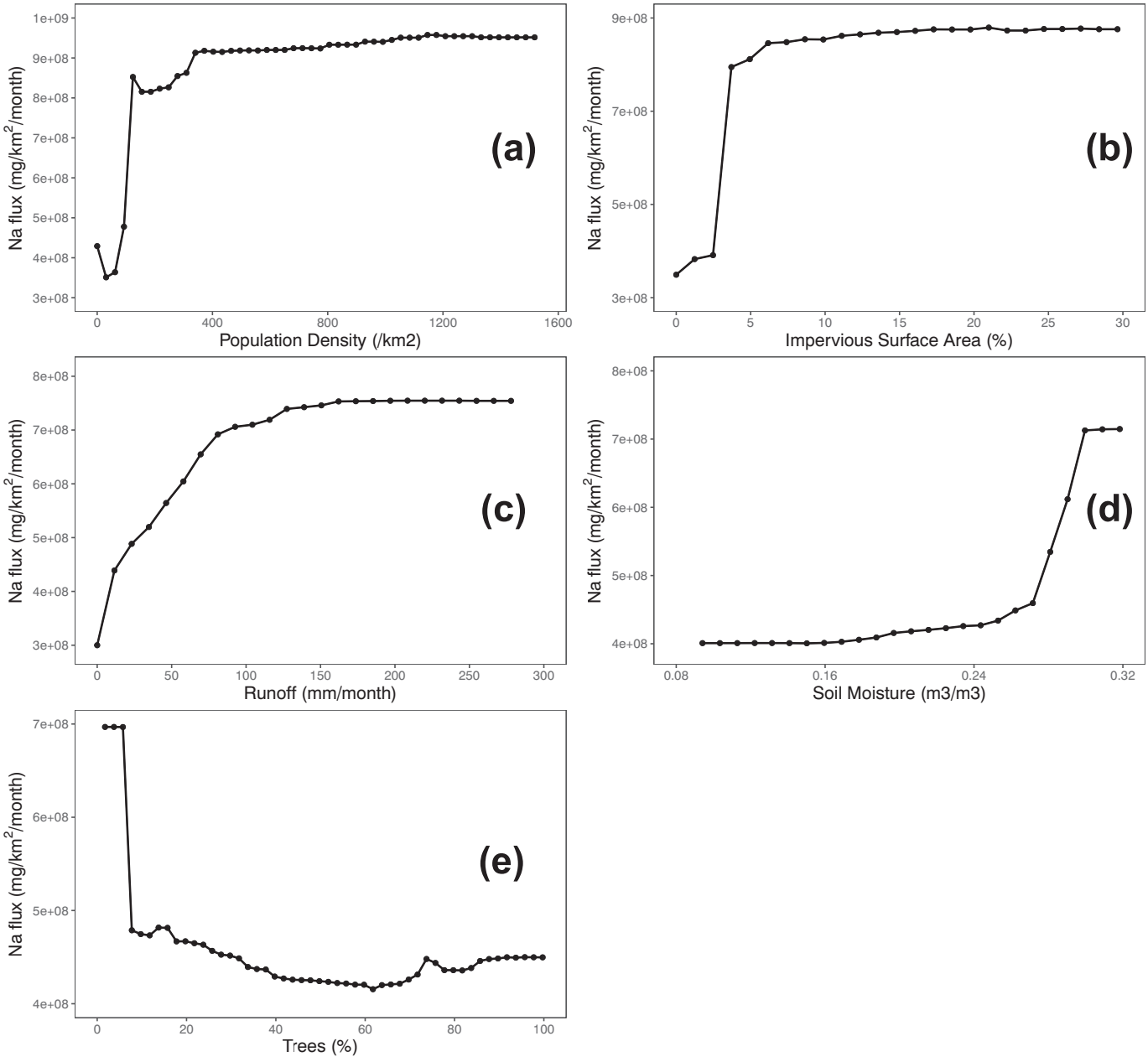


Fig. 3. Partial dependence plots showing variation in monthly riverine sodium flux varies with each individual predictor feature including (a) population density, (b) impervious surface area, (c) runoff, (d) soil moisture, and (e) percent tree cover.

phase, an increase in soil moisture and runoff provides greater capacity to transport sodium from the watershed to the corresponding river. As soil moisture and runoff further increase, the increased transport capacity might outpace the supply of available sodium within the watershed. Therefore, sodium flux might increase at a slower pace or even reach a plateau. Sodium flux generally decreases with increasing percentage of tree cover in the watershed area (Fig. 3e), which might reflect less human-contributed salt under greater tree coverage as well as water losses associated with transpiration.

The main drivers governing watershed-level sodium flux in U.S. rivers are human activities followed by secondary factors contributed by natural processes. This observation is consistent with previous research, which suggest human activities significantly contribute to the long-term salinization of rivers over past decades on the continental scale (Dugan et al., 2017; Kaushal et al., 2017, 2018; Kelting and Yerger, 2012; Thorslund et al., 2021; Utz et al., 2022).

3.3. Evaluation of alkalinity flux prediction model

Based on the same 18 features, we constructed a second RF model to predict riverine alkalinity flux in the 226 U.S. watersheds. The prediction model explained 95 % of the variation of riverine alkalinity flux in test dataset (Fig. S5a). Furthermore, computed residual values demonstrated no significant correlation with predicted alkalinity flux, indicating a well generalized alkalinity RF model (Fig. S5b).

Spatial and temporal prediction heterogeneity was also assessed for the alkalinity flux prediction model. Similar to the sodium flux model, the alkalinity flux model showed no significant spatiotemporal imbalance (Fig. S6a and b). In particular, the percentage of monthly data points with >30 % relative errors had no statistically significant correlation with the total number of monthly measurements in a state ($p > 0.05$). Moreover, no month yielded more erroneous predictions than other months (Fig. S6c).

3.4. Natural processes are the main drivers of alkalinization in U.S. rivers

Based on the calculated CPI values (Fig. 4), the five most important predictor features in regulating riverine alkalinity flux include two hydrological features (runoff, soil moisture), two geological features (carbonate and siliciclastic sediment), and one soil chemical feature (soil pH). All five features are natural factors. Although two human activity related features (population density, impervious surface percentage) were among the top 10 predictor features, their CPI values were much smaller than the top four features. These observations suggest that human factors have relatively minimal impact on watershed-level alkalinity flux compared to natural process related features at the continental scale. Although some previous research (e.g., Kaushal et al., 2017, 2018) has suggested that increasing salinity could enhance the rock weathering resulting in river alkalinization, our results suggest the impact of human activities (e.g., road salting) on watershed-level alkalinity flux might be limited in scope and only prevalent in the vicinity of urban areas. For watersheds with more diverse land use and land cover, natural sources of alkalinity flux in watershed, including those derived from silicate or carbonate weathering and microbial processes (Kaushal et al., 2018; Regnier et al., 2022; Zhang and Planavsky, 2020), appear to outpace the contribution from human-accelerated weathering. Therefore, in contrast to sodium flux, we conclude that watershed-level alkalinity fluxes are mainly controlled by natural processes on a continental scale.

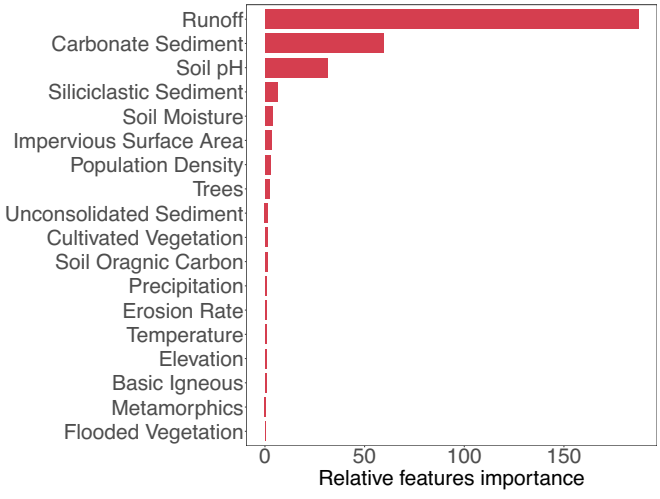


Fig. 4. Predictor feature ranking based on conditional permutation importance values from the riverine alkalinity flux prediction model.

We also evaluated the relationships between each individual important predictor feature and riverine alkalinity flux. Overall, runoff, carbonate sediment, soil pH, siliciclastic sediment, and soil moisture all display a positive correlation with alkalinity flux (Fig. 5). As runoff increases, riverine alkalinity flux rapidly increases before reaching a plateau (Fig. 5a). In the rising phase of alkalinity flux, increased runoff and soil moisture suggest increased capacity of watersheds to deliver weathering products (i.e., alkalinity) to the river. With further increase in runoff the amount of alkalinity supply is eventually outpaced by the transport capacity of the watershed, leading to a plateau in riverine alkalinity flux. In addition, higher soil moisture is likely to increase the wetted surface area of weatherable minerals which in turn should promote rock weathering and microbial activity (Berner, 2004; Kaushal et al., 2018; Regnier et al., 2022; Zhang and Planavsky, 2020). Carbonate and siliciclastic sediments (i.e., the percentage of watershed area underlain by carbonate or siliciclastic sediment) were ranked as the second and fourth most important features, respectively. Carbonate weathering provides alkalinity to rivers in a much faster pace than other rock types (Chou et al., 1989; Liu et al., 2011), which explain the positive correlation with alkalinity flux at modest levels of carbonate sediment (Fig. 5b). As the extent of carbonate deposits further increase, the transport capacity of alkalinity in watershed becomes saturated, leading to a plateau in river alkalinity flux (Fig. 5b). The observation that alkalinity flux generally positively correlate with siliciclastic sediment reflects the contribution of silicate weathering to riverine alkalinity (Fig. 5e). Soil pH, the third most important predictor feature, can be considered as a proxy of the intensity of rock weathering, i.e., the higher the soil pH, the greater the intensity of rock weathering. Given this effect, the observed positive correlation between soil pH and alkalinity flux (Fig. 5c) might be attributed to the fact that high weathering intensity contributes to elevated pH, rather than the inverse relationship.

4. Conclusions

In this study, we compiled dissolved sodium (as the salinity proxy) and alkalinity concentrations for 226 river monitoring sites across the U.S. along with a total of 32 corresponding watershed properties ranging from hydrology, climate, geomorphology, geology, soil chemistry, land use, and land cover for those river monitoring sites. We built two random forest machine learning models using 18 selected watershed features to predict monthly-aggregated sodium and alkalinity contents at these sites. Developed models yielded comparably strong performance across different months and regions. The sodium-prediction model detected population density and impervious surface percentage as the two most important features, suggesting human activities (likely road salt input) as the major sources of salinity in U.S. rivers. This finding was consistent with previous studies, providing another line of evidence that the developed machine learning model yielded reasonably accurate results. The alkalinity prediction model detected natural processes related parameters, namely runoff, percent of the area underlain by carbonate sediment or siliciclastic sediment, and soil pH or moisture, as the top five important features.

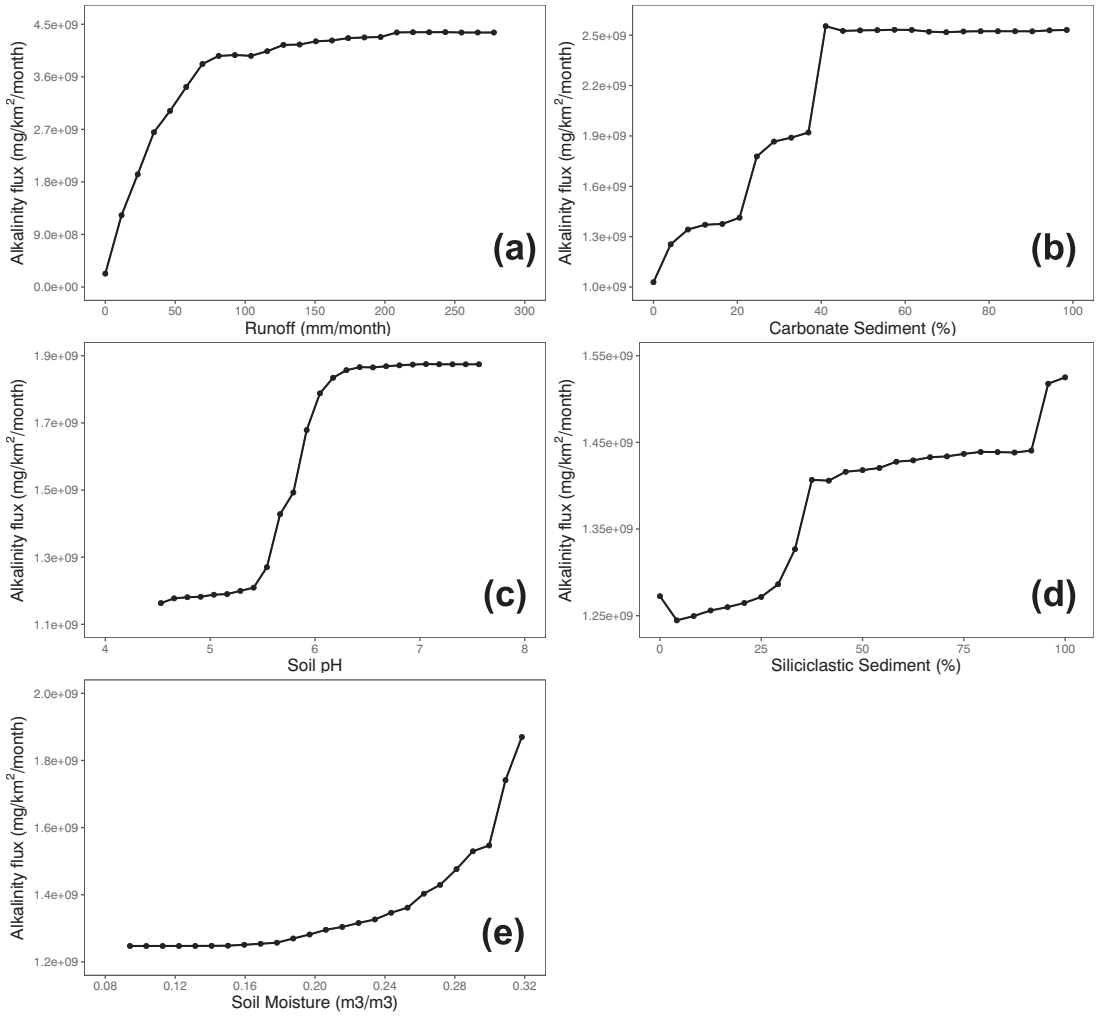


Fig. 5. Partial dependence plots showing the variation in monthly riverine alkalinity flux with each individual predictor feature including (a) runoff, (b) carbonate sediment, (c) soil pH, (d) siliciclastic sediment, and (e) soil moisture.

Therefore, unlike previous studies, we suggested that the watershed-level alkalinization in U.S. rivers might be mostly governed by climatic and hydrogeological conditions. Our finding that watershed-level alkalinity flux in the U.S. is mainly governed by water- and climate-related natural processes implies that a holistic understanding of the evolution of natural conditions of different regions is needed to implement enhanced rock weathering more effectively. Under climate change, hydrogeological conditions will shift in different regions to various extents and even in different directions. So far, many enhanced rock weathering efforts are either planned or ongoing, which would also benefit from such analysis of the relationship between natural factors and weathering flux.

CRediT authorship contribution statement

E. Beibei: Data curation, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Shuang Zhang:** Investigation, Writing – review & editing, Resources. **Charles T. Driscoll:** Investigation, Writing – review & editing. **Tao Wen:** Funding acquisition, Conceptualization, Data curation, Methodology, Investigation, Supervision, Visualization, Writing – review & editing, Resources.

Data availability

The data and codes discussed in this article are deposited in an online data repository which are publicly and freely available via this DOI: https://doi.org/10.5281/zenodo.7937820.

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.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. OAC-2209864 to TW. CTD is supported by the National Science Foundation Grant No. 3340120020186. SZ is supported by the Data Science Career Initiation Fellow Program of Texas A&M Institute of Data Science.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.164138.

References

Abbott, B.W., Bishop, K., Zarnetske, J.P., Minaudo, C., Chapin, F.S., Krause, S., Hannah, D.M., Conner, L., Ellison, D., Godsey, S.E., Plont, S., Marçais, J., Kolbe, T., Huebner, A., Frei, R.J., Hampton, T., Gu, S., Buhman, M., Sara Sayedi, S., Ursache, O., Chapin, M., Henderson, K.D., Pinay, G., 2019. Human domination of the global water cycle absent from depictions and perceptions. Nat. Geosci. 12, 533–540. https://doi.org/10.1038/s41561-019-0374-y.

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7, 162. https://doi.org/10.1038/s41597-020-0479-6.

Barnes, R.T., Raymond, P.A., 2009. The contribution of agricultural and urban activities to inorganic carbon fluxes within temperate watersheds. Chem. Geol. 266, 318–327. https://doi.org/10.1016/j.chemgeo.2009.06.018.

Berner, R.A., 2004. The Phanerozoic Carbon Cycle: CO_2 and O_2. Oxford University Press, Oxford, New York.

Bhide, S.V., Grant, S.B., Parker, E.A., Rippy, M.A., Godrej, A.N., Kaushal, S., Prelewicz, G., Saji, N., Curtis, S., Vikesland, P., Maile-Moskowitz, A., Edwards, M., Lopez, K.G., Birkland, T.A., Schenk, T., 2021. Addressing the contribution of indirect potable reuse to inland freshwater salinization. Nat. Sustain. 4, 699–707. https://doi.org/10.1038/s41893-021-00713-7.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M., 2016. mlr: machine learning in R. J. Mach. Learn. Res. 17, 1–5.. https://jmlr.org/papers/v17/15-066.html.

Brantley, S.L., Bandstra, J., Moore, J., White, A.F., 2008. Modelling chemical depletion profiles in regolith. Geoderma 145, 494–504. https://doi.org/10.1016/j.geoderma.2008.02.010.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

de Colstoun, Brown, 2017. Global Man-made Impervious Surface (GMIS) Dataset From Landsat. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4P55KKF Accessed 02/01/2019.

Cañedo-Argüelles, M., Bundschuh, M., Gutiérrez-Cánovas, C., Kefford, B.J., Prat, N., Trobajo, R., Schäfer, R.B., 2014. Effects of repeated salt pulses on ecosystem structure and functions in a stream mesocosm. Sci. Total Environ. 476–477, 634–642. https://doi.org/10.1016/j.scitotenv.2013.12.067.

Carranza, C., Nolet, C., Pezij, M., van der Ploeg, M., 2021. Root zone soil moisture estimation with Random Forest. J. Hydrol. 593, 125840. https://doi.org/10.1016/j.jhydrol.2020.125840.

Center For International Earth Science Information Network, 2016. Gridded Population of the World, Version 4 (GPWv4): Population Density. https://doi.org/10.7927/H4NP22DQ.

Chou, L., Garrels, R.M., Wollast, R., 1989. Comparative study of the kinetics and mechanisms of dissolution of carbonate minerals. Chem. Geol. 78, 269–282. https://doi.org/10.1016/0009-2541(89)90063-6.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., Kløve, B., 2018. River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. Sci. Total Environ. 615, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293.

Corsi, S.R., De Cicco, L.A., Lutz, M.A., Hirsch, R.M., 2015. River chloride trends in snow-affected urban watersheds: increasing concentrations outpace urban growth rate and are common among all seasons. Sci. Total Environ. 508, 488–497. https://doi.org/10.1016/j.scitotenv.2014.12.012.

Corwin, D.L., 2021. Climate change impacts on soil salinity in agricultural areas. Eur. J. Soil Sci. 72, 842–862. https://doi.org/10.1111/ejss.13010.

DeVilbiss, S.E., Steele, M.K., Krometis, L.-A.H., Badgley, B.D., 2021. Freshwater salinization increases survival of *Escherichia coli* and risk of bacterial impairment. Water Res. 191, 116812. https://doi.org/10.1016/j.watres.2021.116812.

Duan, S.-W., Kaushal, S.S., 2015. Salinization alters fluxes of bioreactive elements from streams and soils across land use (preprint). Biogeochemistry: Rivers & Streams https://doi.org/10.5194/bgd-12-7411-2015.

Dugan, H.A., Bartlett, S.L., Burke, S.M., Doubek, J.P., Krivak-Tetley, F.E., Skaff, N.K., Summers, J.C., Farrell, K.J., McCullough, I.M., Morales-Williams, A.M., Roberts, D.C., Ouyang, Z., Scordo, F., Hanson, P.C., Weathers, K.C., 2017. Salting our freshwater lakes. Proc. Natl. Acad. Sci. U. S. A. 114, 4453–4458. https://doi.org/10.1073/pnas.1620211114.

Erickson, M.L., Elliott, S.M., Brown, C.J., Stackelberg, P.E., Ransom, K.M., Reddy, J.E., 2021a. Machine learning predicted redox conditions in the glacial aquifer system, northern continental United States. Water Resour. Res. 57, e2020WR028207. https://doi.org/10.1029/2020WR028207.

Erickson, M.L., Elliott, S.M., Brown, C.J., Stackelberg, P.E., Ransom, K.M., Reddy, J.E., Cravotta, C.A.I., 2021b. Machine-learning predictions of high arsenic and high manganese at drinking water depths of the glacial aquifer system, northern continental United States. Environ. Sci. Technol. 55, 5791–5805. https://doi.org/10.1021/acs.est.0c06740.

Estévez, E., Rodríguez-Castillo, T., González-Ferreras, A.M., Cañedo-Argüelles, M., Barquín, J., 2019. Drivers of spatio-temporal patterns of salinity in Spanish rivers: a nationwide assessment. Philos. Trans. R. Soc. B 374, 20180022. https://doi.org/10.1098/rstb.2018.0022.

Findlay, S.E.G., Kelly, V.R., 2011. Emerging indirect and long-term road salt effects on ecosystems: Findlay & Kelly. Ann. N. Y. Acad. Sci. 1223, 58–68. https://doi.org/10.1111/j.1749-6632.2010.05942.x.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global consequences of land use. Science 309, 570–574. https://doi.org/10.1126/science.1111772.

Ghiggi, G., Humphrey, V., Seneviratne, S.I., Gudmundsson, L., 2021. G-RUN ENSEMBLE: a multi-forcing observation-based global runoff reanalysis. Water Res. 57. https://doi.org/10.1029/2020WR028787.

Grimm, N.B., Faeth, S.H., Golubiewski, N.E., Redman, C.L., Wu, J., Bai, X., Briggs, J.M., 2008. Global change and the ecology of cities. Science 319, 756–760. https://doi.org/10.1126/science.1150195.

Hansen, J.A., Jurgens, B.C., Fram, M.S., 2018. Quantifying anthropogenic contributions to century-scale groundwater salinity changes, San Joaquin Valley, California, USA. Sci. Total Environ. 642, 125–136. https://doi.org/10.1016/j.scitotenv.2018.05.333.

Harris, L., Taylor, A.H., 2017. Previous burns and topography limit and reinforce fire severity in a large wildfire. Ecosphere 8, e02019. https://doi.org/10.1002/ecs2.2019.

Hartmann, J., Moosdorf, N., 2012. The new global lithological map database GLiM: a representation of rock properties at the Earth surface. Geochem. Geophys. Geosyst. 13, 1–37. https://doi.org/10.1029/2012GC004370.

Hintz, W.D., Relyea, R.A., 2019. A review of the species, community, and ecosystem impacts of road salt salinisation in fresh waters. Freshw. Biol. 64, 1081–1097. https://doi.org/10.1111/fwb.13286.

Ho, Tin Kam, 1995. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. Presented at the 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press, Montreal, Que., Canada, pp. 278–282 https://doi.org/10.1109/ICDAR.1995.598994.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Sci. Data 4, 170122. https://doi.org/10.1038/sdata.2017.122.

Kaushal, S.S., 2016. Increased salinization decreases safe drinking water. Environ. Sci. Technol. 50, 2765–2766. https://doi.org/10.1021/acs.est.6b00679.

Kaushal, S.S., Belt, K.T., 2012. The urban watershed continuum: evolving spatial and temporal dimensions. Urban Ecosyst. 15, 409–435. https://doi.org/10.1007/s11252-012-0226-7.

Kaushal, S.S., Groffman, P.M., Likens, G.E., Belt, K.T., Stack, W.P., Kelly, V.R., Band, L.E., Fisher, G.T., 2005. Increased salinization of fresh water in the northeastern United States. Proc. Natl. Acad. Sci. U. S. A. 102, 13517–13520. https://doi.org/10.1073/pnas.0506414102.

Kaushal, S.S., Likens, G.E., Utz, R.M., Pace, M.L., Grese, M., Yepsen, M., 2013. Increased river alkalinization in the Eastern U.S. Environ. Sci. Technol. https://doi.org/10.1021/es401046s (130724203606002).

Kaushal, S.S., McDowell, W.H., Wollheim, W.M., 2014. Tracking evolution of urban biogeochemical cycles: past, present, and future. Biogeochemistry 121, 1–21. https://doi.org/10.1007/s10533-014-0014-y.

Kaushal, S.S., Duan, S., Doody, T.R., Haq, S., Smith, R.M., Newcomer Johnson, T.A., Newcomb, K.D., Gorman, J., Bowman, N., Mayer, P.M., Wood, K.L., Belt, K.T., Stack, W.P., 2017. Human-accelerated weathering increases salinization, major ions, and alkalinization in fresh water across land use. Appl. Geochem. 83, 121–135. https://doi.org/10.1016/j.apgeochem.2017.02.006.

Kaushal, S.S., Likens, G.E., Pace, M.L., Utz, R.M., Haq, S., Gorman, J., Grese, M., 2018. Freshwater salinization syndrome on a continental scale. Proc. Natl. Acad. Sci. U. S. A. 115, E574–E583. https://doi.org/10.1073/pnas.1711234115.

Kelleher, C., Golden, H.E., Burkholder, S., Shuster, W., 2020. Urban vacant lands impart hydrological benefits across city landscapes. Nat. Commun. 11, 1563. https://doi.org/10.1038/s41467-020-15376-9.

Kelting, D.L., Yerger, E.C., 2012. Regional analysis of the effect of paved roads on sodium and chloride in lakes. Water Res. 46, 2749–2758. https://doi.org/10.1016/j.watres.2012.02.032.

Larsen, I.J., Montgomery, D.R., Greenberg, H.M., 2014. The contribution of mountains to global denudation. Geology 42, 527–530. https://doi.org/10.1130/G35136.1.

Le, T.D.H., Kattwinkel, M., Schützenmeister, K., Olson, J.R., Hawkins, C.P., Schäfer, R.B., 2019. Predicting current and future background ion concentrations in German surface water under climate change. Philos. Trans. R. Soc. B 374, 20180004. https://doi.org/10.1098/rstb.2018.0004.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.

Likens, G.E., Buso, D.C., 2010. Salinization of Mirror Lake by road salt. Water Air Soil Pollut. 205, 205–214. https://doi.org/10.1007/s11270-009-0066-0.

Liu, Z., Dreybrodt, W., Liu, H., 2011. Atmospheric CO_2 sink: silicate weathering or carbonate weathering? Appl. Geochem., Ninth International Symposium on the Geochemistry of the Earth's Surface (GES-9) 26, S292–S294. https://doi.org/10.1016/j.apgeochem.2011.03.085.

Meybeck, M., 2003. Global occurrence of major elements in rivers. Treatise on Geochemistry. Elsevier, pp. 207–223 https://doi.org/10.1016/B0-08-043751-6/05164-1.

Perri, S., Molini, A., Hedin, L.O., et al., 2022. Contrasting effects of aridity and seasonality on global salinization. Nat. Geosci. 15, 375–381. https://doi.org/10.1038/s41561-022-00931-4.

Perri, S., Suweis, S., Holmes, A., Marpu, P.R., Entekhabi, D., Molini, A., 2020. River basin salinization as a form of aridity. Proc. Natl. Acad. Sci. U. S. A. 117, 17635–17642. https://doi.org/10.1073/pnas.2005925117.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7, 217–240. https://doi.org/10.5194/soil-7-217-2021.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Ransom, K.M., Nolan, B.T., Stackelberg, P.E., Belitz, K., Fram, M.S., 2022. Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. Sci. Total Environ. 807, 151065. https://doi.org/10.1016/j.scitotenv.2021.151065.

Read, E.K., Carr, L., De Cicco, L., Dugan, H.A., Hanson, P.C., Hart, J.A., Kreft, J., Read, J.S., Winslow, L.A., 2017. Water quality data for national-scale aquatic research: the water quality portal. Water Resour. Res. 53, 1735–1745. https://doi.org/10.1002/2016WR019993.

Regnier, P., Resplandy, L., Najjar, R.G., Ciais, P., 2022. The land-to-ocean loops of the global carbon cycle. Nature 603, 401–410. https://doi.org/10.1038/s41586-021-04339-9.

Tesoriero, A.J., Gronberg, J.A., Juckem, P.F., Miller, M.P., Austin, B.P., 2017. Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour. Res. 53, 7316–7331. https://doi.org/10.1002/2016WR020197.

Thorslund, J., Bierkens, M.F.P., Oude Essink, G.H.P., Sutanudjaja, E.H., van Vliet, M.T.H., 2021. Common irrigation drivers of freshwater salinisation in river basins worldwide. Nat. Commun. 12, 4232. https://doi.org/10.1038/s41467-021-24281-8.

Tuanmu, M.-N., Jetz, W., 2014. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Glob. Ecol. Biogeogr. 23, 1031–1045. https://doi.org/10.1111/geb.12182.

Utz, R., Bidlack, S., Fisher, B., Kaushal, S., 2022. Urbanization drives geographically heterogeneous freshwater salinization in the northeastern United States. J. Env. Qual. 51, 952–965. https://doi.org/10.1002/jeq2.20379.

Wang, Y., Mao, J., Jin, M., Hoffman, F.M., Shi, X., Wullschleger, S.D., Dai, Y., 2021. Development of observation-based global multilayer soil moisture products for 1970 to 2016. Earth Syst. Sci. Data 13, 4385–4405. https://doi.org/10.5194/essd-13-4385-2021.

Wen, T., Chen, C., Zheng, G., Bandstra, J., Brantley, S.L., 2022. Using a neural network – physics-based hybrid model to predict soil reaction fronts. Comput. Geosci. 167, 105200. https://doi.org/10.1016/j.cageo.2022.105200.

Zhang, S., Planavsky, N.J., 2020. Revisiting groundwater carbon fluxes to the ocean with implications for the carbon cycle. Geology 48, 67–71. https://doi.org/10.1130/G46408.1.