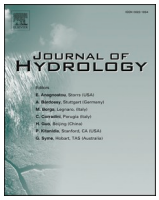
Journal of Hydrology 660 (2025) 133319

Contents lists available at ScienceDirect

Journal of Hydrology

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



Research papers

Critical zone controls on stream chemistry: Lessons from multiple machine learning methods and irregular data across large watersheds

G.M. Goldrich-Middaugh [a], K. Johnson [a], L. Ma [b], M.A. Engle [b], S.W. Fleming [a], J.W. Ricketts [b], P.L. Sullivan [a,*]

[a] Oregon State University, College of Earth, Ocean, and Atmospheric Sciences, USA
[b] University of Texas at El Paso, Department of Earth, Environmental and Resource Sciences, USA

ARTICLE INFO

Keywords:
Critical zone
Machine learning
Self-organizing map
Water quality
Regional hydrology
Clustering

ABSTRACT

Understanding surface water quality and managing river systems that span a variety of climatic, lithology, and land cover types requires regional data analysis approaches. At this scale, spatial and temporal variations and irregularities in publicly available stream chemical data limit the application of many traditional techniques for evaluating controls on water quality. This work uses publicly available measurements of major ion concentrations (>31,000 solute measurements from 670 sites; 1944–2018) from the Colorado, Brazos, Red, and Pecos Rivers in Texas to parse out controls on stream water chemical variations across space. We used an emergent self-organizing map (ESOM) to identify structures in the data. K-means clustering was then performed on the ESOM structure (e.g., proportional node weights for each solute) and six clusters were most optimal for characterizing distinct spatial patterns in stream chemistry. For example, a unique chemical signature marked by elevated concentration of magnesium relative to calcium and bicarbonate was observed over the Balcones fault zone and elevated silicon concentrations were observed in the wetter, more weathered eastern portion of Texas where forest cover is more dominant. A random forest classification model was used to predict cluster membership from sub-basin characteristics and had an overall accuracy rate of 78.9 %. Mean annual precipitation was found to be the most important variable for distinguishing between clusters. The melding of ESOM, clustering, and random forest machine learning approaches reveals complex hydrogeochemical processes, informs regional watershed management, and pinpoints areas needing further study.

Plain Language Summary: Water quality measurements are geographically widespread across the United States but are irregularly distributed in space and time, which can make it difficult to use them with traditional data analysis approaches for understanding watershed processes. Here, we used machine learning to examine data from four large watersheds in Texas to understand the influence of different land use, geologic, and climate factors on water quality. The four rivers cross a range of rock types, land uses, and precipitation regimes and have different chemical compositions. We found tens of thousands of observations from this large area could be summarized using only six chemical groups and that there were strong relationships between watershed characteristics and water chemistry. Overall, these techniques were effective in capturing major spatial water quality trends across these watersheds and could be used to target watershed management and monitoring efforts.

1. Introduction

River chemistry in regional scale watersheds is of crucial importance for managing environmental and anthropogenic water uses as well as understanding global biogeochemical cycling (Dupré et al., 2003; Gaillardet et al., 1999). At the regional scale, a critical zone approach that examines the integrated processes from the top of the canopy to the depths of circulating groundwater (National Research Council, 2012) can help elucidate the network of complex and heterogeneous factors that control stream water chemistry. In the United States (US) and Europe large bodies of publicly available data from various agencies are widely available but are challenging to use in scientific investigations.

* Corresponding author.
E-mail addresses: goldricg@oregonstate.edu (G.M. Goldrich-Middaugh), johnkeir@oregonstate.edu (K. Johnson), lma@utep.edu (L. Ma), maengle@utep.edu (M.A. Engle), fleminse@oregonstate.edu (S.W. Fleming), jricketts@utep.edu (J.W. Ricketts), pamela.sullivan@oregonstate.edu (P.L. Sullivan).

https://doi.org/10.1016/j.jhydrol.2025.133319
Received 6 August 2024; Received in revised form 20 March 2025; Accepted 12 April 2025
Available online 13 April 2025

Their irregularity in space and time limits applications of many techniques such as the assessment of concentration-discharge relationships, or application of physically-based hydrologic models that are commonly applied at smaller scales with more detailed sampling (e.g., Zhi et al., 2024; Stewart et al., 2022). However, these large-scale datasets contain information that can improve our understanding of how critical zone processes influence the dynamics of large river systems (e.g., Jankowski et al., 2023; Johnson et al., 2024a; Zhi et al., 2023) and can be used to improve further sampling efforts and water quality management (Wai et al., 2022).

Stream chemical composition can be used to understand structure, function, and processes occurring within watersheds draining various environments (Li et al., 2021; Singha et al., 2024). However, in large and complex watersheds, even where there are extensive bodies of publicly available data and hypothesis-driven knowledge, such as in the Mississippi and Chesapeake Bay river watersheds in the US (Giri, 2021), subsurface, land cover, and climate processes that influence stream water composition are not fully understood (Li et al., 2024). Lithology and flow regime have been identified as two of the primary drivers of stream chemical composition at the regional scale (Baronas et al., 2017; Gaillardet et al., 1999; Godsey et al., 2009; Kirchner, 2009; Torres et al., 2017). Often land use and climate also emerge as important influences across a range of scales (Allan, 2004; Shi et al., 2017). Particularly, climate patterns control the degree of chemical weathering, recharge, and therefore solute transport at local to regional scales (Brantley et al., 2017). Land use and land cover (LULC) characteristics also influence many hydrologic factors including infiltration rates, timing of peak discharge, erosion, and surface runoff volumes (Brooks et al., 2012; Keen et al., 2023; Sadayappan et al., 2023). Additionally, LULC can act as point and non-point source inputs of solutes, collectively these factors impact water quality (e.g., fertilizers, deicers; Li et al., 2024).

Multivariate techniques, including machine learning methods, are necessary to understand and depict complex hydrochemical relationships within the critical zone. Applications of machine learning to critical zone science have facilitated greater understanding of the influence of watershed characteristics on flow metrics across space and through time (Addor et al., 2018; Hammond et al., 2021; Wlostowski et al., 2021; Xu Fei & Harman, 2020). Applying these techniques to publicly available, noisy and spatiotemporally complex data could illuminate known and additional relationships between stream water quality and critical zone factors (Fleming et al., 2019; Melo et al., 2019; Nearing et al., 2021; Zhi et al., 2023). Studies of variations in major ion chemistry have revealed aquifer structure, contaminant flow paths, and groundwater solute sources using machine learning techniques including self-organizing maps (SOMs), also known as a Kohonen neural network (Chen et al., 2018; Haselbeck et al., 2019; Sahour et al., 2020). SOMs are an unsupervised neural network algorithm that converts non-linear relationships in large, high-dimensional datasets into a two-dimensional grid of nodes to facilitate visual assessment or further quantitative analysis, and can augment, improve, or outperform other techniques such as principal components analysis and hierarchical cluster analysis (Chavoshi et al., 2012; Kohonen, 2001; Wehrens & Buydens, 2007). They are useful in reducing dimensionality while preserving key characteristics, or topology, of the input datasets (Haselbeck et al., 2019; Melo et al., 2019; Sinha et al., 2010). SOMs have advantages over other multivariate techniques in that they are highly flexible (can capture nonlinear relationships) and rely on fewer assumptions of the input data (i.e., non-standard distribution). Emergent Self-Organizing Maps (ESOMs) are SOMs constructed using many nodes ($n > 4{,}000$) such that emergent structures in high-dimensional data are visible (Thrun et al., 2016; Ultsch, 1999). While almost identical to SOMs, ESOMs have the ability to organize intricate datasets and delineate clusters with complex geometries due to their capacity to accommodate hundreds to thousands of nodes (Ultsch, 2007). When observations are matched to nodes on the ESOM, each node can be viewed as a proto-cluster, where each observation belongs to the cluster (node) with the highest degree of similarity (Vesanto et al., 2000). This proto-clustering can be applied as the first step in a two-step clustering process where the ESOM nodes are then clustered using an additional method such as k-means. This two-step clustering process reduces the influence of outlying data points and simplifies structures in the data to highlight overarching patterns (Vesanto et al., 2000). Thus, the application of a two-step clustering process holds the promise of providing the flexibility and adaptability to elucidate patterns in water quality data collected irregularly in space and time.

Supervised machine learning algorithms such as random forests and extreme gradient boosting (XGBoost) models can then be applied to these emergent water quality patterns or clusters generated from ESOMs to understand the possible factors that control a particular chemical signature (Bolotin et al., 2023; Nasir et al., 2022; Sadayappan et al., 2022; Yang & Olivera, 2023). The random forest algorithm consists of an ensemble of decision trees generated using bootstrapped subsamples of the dataset. Random forests are a flexible multivariate technique that is appropriate for analyzing multiple predictor variables and nonlinear relationships and has been widely applied in hydrologic sciences (Addor et al., 2018; Brown et al., 2014; Fleming et al., 2021a; Hammond et al., 2021; Konapala & Mishra, 2020; Oppel & Schumann, 2020; Singh et al., 2019). Random forest classification can be used to predict cluster membership using watershed factors (e.g., lithology, land cover), and assess the importance of each watershed factor in differentiating between clusters. The success in cluster membership prediction indicates the overall quality of model training, uniqueness of the cluster attributes, and the strength of the connection between stream chemistry and the landscape factors of interest. The importance of each watershed factor within the random forest algorithm gives an indication of the strength of the relationship between that attribute and stream chemistry in the watersheds of interest.

Detailed analyses of water quality data using multivariate and machine learning approaches could provide a pathway for modeling critical zone structure and function at a regional scale and be used to evaluate and target large-scale sampling campaigns to areas where hydrochemical behaviors are not well understood. Large bodies of water quality data have been collected over the preceding decades and are now easily accessible. These data have been collected for myriad purposes and are therefore spatially and temporally heterogeneous and may not be well-suited for traditional hydrochemical analyses. However, patterns and relationships can be identified from large scale, publicly available datasets to understand water quality responses to critical zone characteristics. Therefore, we investigate hydrochemical data at a regional scale to address the research question: To what degree can patterns in surface water chemistry across large regions be identified based on critical zone characteristics using emergent self-organizing maps (ESOMs), k-means clustering, and random forest classification, and can these tools shed light on water quality management strategies?

In particular, we investigate the relationships between major ion chemistry of four major rivers in Texas; the Colorado, Brazos, Red, and Pecos rivers. These four rivers run roughly in parallel across low-relief environments which cross precipitation, lithology, and LULC gradients. Previous analysis suggests that much of the variation in stream chemical composition in the Colorado river, Texas, was controlled by interactions between underlying lithology and processes impacting flow paths and residence times (including climate and LULC factors; Goldrich-Middaugh et al., 2022). Studies in the region have also shown that urbanization, especially in large metropolitan areas within these watersheds, has impacted surface and groundwater quality (Aitkenhead-Peterson et al., 2011). Other land uses in the region include widespread irrigated agriculture, pastures and grazing, and oil and gas production in the Midland and Eagle Ford basins (Jiang et al., 2022) and along the Colorado river (Goldrich-Middaugh et al., 2022), which may impact water quality and quantity. Additionally, the region is sensitive to the impacts of a changing climate, with water resources projected to become scarcer and decline in quality as the population continues to grow and

water demand increases (Harwell et al., 2020; Jager et al., 2015). Thus, the large, publicly available water chemistry datasets from these four watersheds may help us understand the critical zone controls on stream chemistry.

2. Site Description

The Brazos, Colorado, Pecos, and Red are four adjacent rivers that run roughly parallel across Texas with watershed areas extending into New Mexico, Oklahoma, and Louisiana (Fig. 1). Each river has a

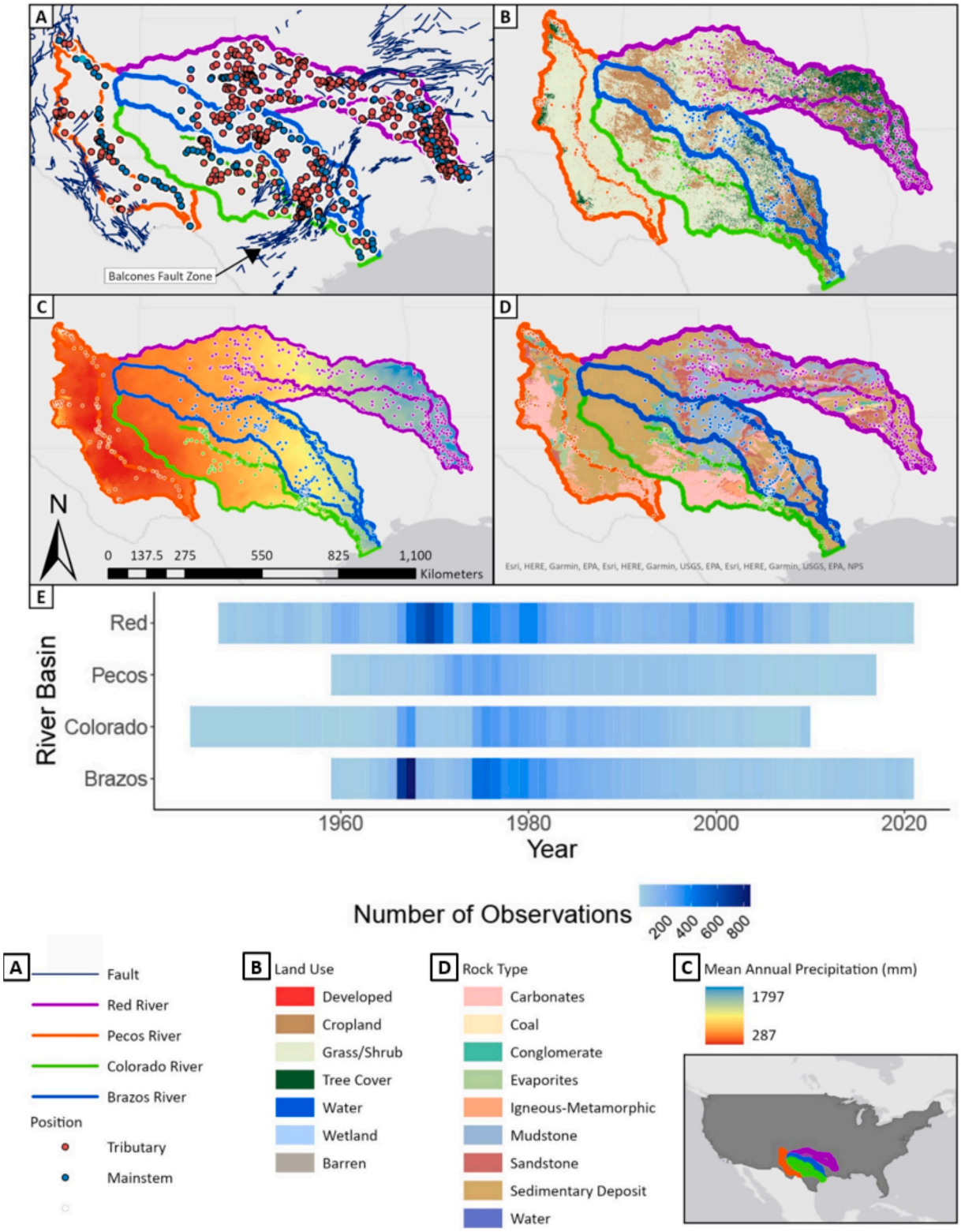


Fig. 1. Maps showing A) Sampling point locations colored by tributary position and mapped faults including the Balcones Fault Zone, B) Land use and land cover data at 30 m resolution from 2019, C) Mean annual precipitation (PRISM) in mm from 1985 to 2019, and D) Simplified lithology class. For reference of each site to the contributing tributaries see Fig. S1. E) Depiction of the number of observations for any given year in each of the four watersheds.

watershed area of over 100,000 km^2 and a main-stem river length of over 1,300 km (Table 1). The rivers flow across diverse lithologies, with geologic units oriented roughly perpendicular to flow. Climate, LULC, and lithology across the watersheds set the stage for spatially contrasting chemical compositions. A total of 670 surface water sampling sites across these watersheds were accessed using the National Water Quality Monitoring Council water quality portal (WQP).

2.1. Land use and land cover

Land use varies roughly perpendicular to flow directions, particularly across the Colorado, Brazos, and Red river watersheds. The westernmost portion of these watersheds and most of the Pecos River watershed are dominated by barren and grass/shrub cover (Fig. S2). The headwaters of the Colorado, Brazos, and Red rivers also support large swaths of agriculture. Forested land covers the central reaches of these watersheds and the Pecos headwaters and western most section. Larger portions of developed area occur closer to the outlets of the Colorado and Brazos rivers surrounding Houston and Austin (Fig. 1B). The density distribution of sites across these landcover types for each watershed can be found in the supplemental material (Fig. S2). Access to LULC data between 1985 and 2022 (see methods), show that across the sub-basins in the four watersheds the LULC change in any given land cover type ranged from < 1 % to nearly 40 % (Fig. S3), with the largest increases observed for impervious surfaces in the Brazos river watershed (38 %), while the largest decreases were observed for shrub and grass in the Pecos river watershed (31 % decline) and forest across all four watersheds (19–25 % decline). Over the same period the sub-basins in the Brazos, Colorado, and Pecos watersheds showed a general increase in cropland, while the Red showed a decline. The inverse pattern was generally observed for grass and shrub land. Finally, both the Brazos and the Colorado river watershed showed some degree of increase in the amount of impervious surface (Fig. S3).

2.2. Climate

The western portion of the study area, where a greater density of sites in the Pecos are located, is dominantly semi-arid receiving approximately 300 mm of precipitation per year (Fig. 1, Fig. S3). Mean annual precipitation (MAP) increases to the east until the Gulf Coast, which receives approximately 1,800 mm of precipitation per year (Fig. 1C), particularly in the Red River. These regional climate patterns strongly influence the availability of surface water resources and the distribution of land cover. The Red river watershed encompasses the largest climatic gradient and receives the highest MAP in the study area near its outlet (Fig. 1C) where the Pecos river watershed is the most homogeneous in terms of climate and receives little precipitation along its entire length. The Colorado and Brazos watersheds receive MAP with intermediate gradients of the four watersheds.

Table 1
Summary of watershed characteristics and available data for the Brazos, Colorado, Pecos, and Red Rivers. Complete cases indicate the observations where all solutes were present for the analysis. Mean observations indicates the number of samples collected on average at a site in each watershed.

	Brazos	Colorado	Pecos	Red
Complete Cases: Observations	8,980	5,441	4,249	12,654
Sites	153	112	92	313
Mean Observations per Site	59	49	46	40
Watershed Area	115,565 km^2	103,340 km^2	115,000 km^2	169,900 km^2
River Length	1,352 km	1,387 km	1,490 km	2,189 km

2.3. Lithology

Texas exposes a wide variety of lithologic units that span more than a billion years of geologic time. Mesozoic and Cenozoic sedimentary rocks form linear belts that trend SW-NE and become progressively younger towards the Gulf of Mexico (Fig. 1D; Collins, 1993). This pattern is interrupted by exposures of Proterozoic igneous and metamorphic rock of the Llano uplift in central Texas. Most of the recent tectonic activity is concentrated along the western edge of Texas, although the Balcones fault zone sweeps NE across the state and crosses the watersheds of interest in this study (Fig. 1A). The Pecos river (westernmost watershed) has large surface expressions of carbonates in the northern portion of the watershed and evaporites near the center. Small areas of igneous-metamorphic rock are present in the extreme north and south. The Colorado river watershed has smaller areas of evaporites in the headwaters with large portions of the central reaches dominated by carbonates. The Brazos river is dominantly covered by sedimentary units of varying textures underlain by carbonate rocks which show surface expression in the central reaches of the watershed. The Red river is also dominantly covered by sedimentary units with some large areas of evaporites and gypsum within Permian sandstone and surface expression of the underlying carbonates. Additionally, coal dominated areas are present in the lower-middle reaches of the Red river (Fig. 1D). The density distribution of sites across these lithologies for each watershed can be found in the supplemental material (Fig. S2).

3. Methods

3.1. Data Collection and Processing

3.1.1. Hydrologic data

All water chemistry data were obtained using the water quality portal (WQP) R package *dataRetrieval* to access all chemical measurements within the Pecos, Colorado, Brazos, and Red river watersheds (De Cicco et al., 2018). This dataset included 31,324 measurements of alkalinity as HCO_3^-, K^+, Na^+, Mg^{2+}, SO_4^{2-}, Ca^{2+}, Cl^-, and Si collected at 670 sites between 1944–2021 (Fig. 1A). Stream discharge was not available at each of the sites, and thus stream discharge was not included in the analyses. Water chemistry data were used for the development of the ESOM. The water quality data were spatially and temporally irregular with a minimum of one and a maximum of 39 samples collected at each site per year. Of the sites included in our dataset, 91 have at least 100 total samples collected over the period of record and 175 sites have at least 50 observations (see Fig. S4 for the distribution of observation per site). Measurements were filtered to include sites where all solutes were measured on the same date (termed complete cases). Samples included in analysis were assumed to be representative of major constituents and free of significant analytical error if they had a low charge balance error (≤ 10 %; Godsey et al., 2009; Güler et al., 2002). Samples were transformed into compositional data by converting measurements to meq/L and calculating the contribution to the overall charge of the sample. Observations were then centered to a mean of zero and unit variance so that parameters with larger magnitudes did not dominate the training of the ESOM. Piper diagrams of these data show that while latitude (proxy for aridity) plays a role in water chemistry patterns (Fig. S5), nonlinearity in the data warrants the use of machine learning and clustering algorithms.

3.1.2. Spatial data

The contributing area for each sampling point (termed sub-basin) was delineated using the NASA Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) digital elevation model (30 m resolution) (NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team, 2019). Sub-basin delineation was conducted using ArcGIS Pro 2.9 (ESRI, 2021) functions. Flow directions were generated using the D8 method in which flow is routed to one of eight neighboring

cells selected by calculating maximum downhill steepness (Qin et al., 2007). Next, flow accumulation was calculated based on the flow direction raster with zones of high accumulation representing stream channels. Water chemistry sampling points were used as pour points and were snapped to nearest zones of flow accumulation with a maximum snapping threshold. Contributing areas for each sampling point were then delineated using the watershed tool. Snapping of pour points and delineation of watersheds was conducted iteratively using a range in maximum snapping distances, 0.005–0.25 degrees, to obtain the appropriate delineation. We verified the delineation of each sampling point visually. Results were assessed for goodness of fit with reference measurements such as contributing areas referenced for USGS stations, where available. This approach was employed to ensure that summarized characteristics were site-specific for each sampling point, rather than summarized at a sub-watershed scale such as the National Hydrography Dataset (NHD) HUC 12 subbasins or similar. A variety of spatial datasets were employed to examine factors influencing stream chemistry at each point within the watershed. Lithologic data was obtained from the state-level preliminary integrated geologic map databases for the United States for Texas, Oklahoma, Arkansas, and Louisiana (Stoesser et al., 2007). Rock types were merged to convey major classes (i.e., carbonates, sandstone, mudstone, etc.; Schweitzer, 2011; Texas Water Development Board – https://www.twdb.texas.gov). Comprehensive land use data are from the Global Land Cover Change Dataset (GLC-FCSD 30; Liu et al., 2023; Zhang et al., 2024) that documents land use change across seven major land use‘ classes from 1985 to 2022 at a 30-meter resolution. This data product provides land cover data every five years from 1985 to 2000, annually thereafter. As such we set all samples with dates at or prior to 1985 equal to the 1985 land cover, we then applied a linear interpolation using the five-year data (e.g., 1985–1990) for every year between 1985–2000, we then used the annual data thereafter in the model. Land cover types were then merged to convey major classes. Annual precipitation data was accessed from PRISM (PRISM Climate Group, 2014) using the USGS GEODATA portal and provides gridded estimates of annual precipitation at the 4 km scale. Mean annual precipitation across the study area was calculated from the annual datasets for 1985–2019 (Fig. 1C), thus overlapping with the same period of record as LULC. Average effective precipitation (i.e., precipitation – evapotranspiration) was calculated for the upstream contributing area of each sampling point. Lithology and land use classes were expressed as percent cover for the contributing area upstream of each sampling point.

3.2. Machine learning analyses

We used stream chemical measurements of eight major solutes across nearly 32,000 observations from 670 distinct geographic sampling locations to develop an ESOM, whose structure was then clustered using K-means. Finally, we used sub-basin characteristics to predict cluster membership using random forest classification. Below we provide more details on this machine learning framework.

3.2.1. Emergent Self-Organizing maps ESOMs

We used an ESOM to examine the spatial and temporal structure of stream chemical observations. ESOMs were constructed with toroidal structure and 50 x 82 nodes for a total of 4,100 nodes using the R package Umatrix (Lerch et al., 2020; Fig. 2). The ESOM nodes were initialized with random samples selected from a normal distribution around the mean of the data. A gaussian neighborhood function was used, the search radius was initiated at 24 nodes and decreased linearly to 1, and the learning rate was initiated at 0.5 and decreased linearly to 0.1. The ESOM was trained on 100 epochs where each epoch represents one presentation of all observations to the map (Sinha et al., 2010; Vesanto et al, 2000). The distance between the observation vectors and each of the node vectors was compared, where a vector represents the normalized, compositional chemistry for each observation. The observation was then matched to the node vector with the minimum distance (Vesanto et al, 2000). Thus, each node can represent more than one of the nearly 32,000 samples and multiple sites based on similarity in chemistry. Toroidal structures were utilized to eliminate any edge effects, where the neurons at the edges of the map have much fewer observations than those in the center, by creating a cyclical structure in the output SOM (Thrun et al., 2016). The resulting unified distance matrix (Umatrix) was then plotted and assessed for stability. Umatrices for ESOMs depict the sum of the distance for each weight between each node and its neighbors, highlighting the topology of the dataset wherein valleys show zones of high similarity and peaks show zones of higher dissimilarity (Thrun et al., 2016; Ultsch, 1999). Stability of the Umatrix was assessed by examining root mean square error (RMSE) between each observation and its “BestMatch”, which was the node with the smallest distance to a given observation, as described above. The acceptability of a RMSE value depends on the range of input data; thus, a normalized RMSE (RMSE divided by the range of population data) that is close to 0 would indicate a good model fit, while a value closer to 1 indicates a poor fit.

3.2.2. Clustering

Clustering of ESOM outputs assists in interpretation of the map output and further analysis of differences in chemical composition across map regions (Gamble & Babbar-Sebens, 2012; Haselbeck et al., 2019; Melo et al., 2019). Cluster analysis is also a powerful tool for identifying spatial and temporal gradients and their drivers at the regional scale (Brown et al., 2014). Final node weights (each a vector of dimension 8x1, where each weight in the vector represents one solute)

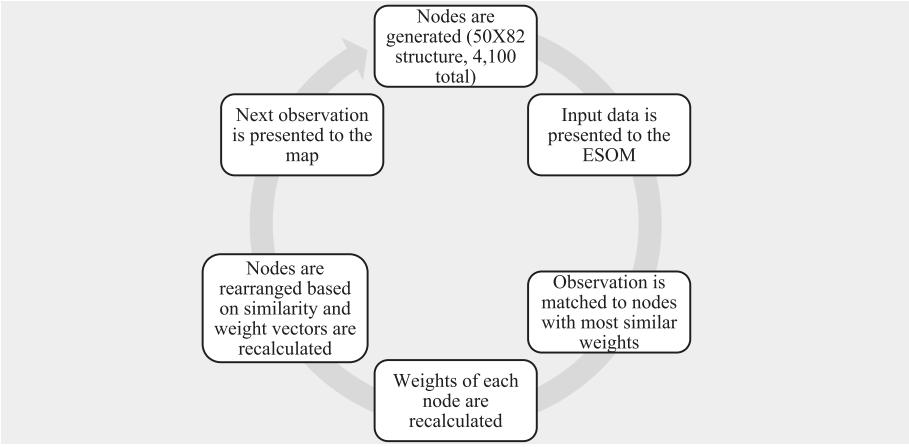


Fig. 2. Steps for training of emergent self organizing maps (ESOM).

from the ESOM were clustered using the k-means clustering algorithm using R package stats (R Core Team, 2021). Using the node weights rather than the raw data for k-means clustering reduced noisiness in the dataset while reflecting key properties of the underlying data (Vesanto et al., 2000). K-means clustering partitions observations into k number of clusters, where k is set by the user. The optimal k value was selected using a scree plot of within sum of squares, which is the sum of the variance from each observation to its centroid (Fig. S9). The structure of the resulting clusters was analyzed using silhouette plots showing the distinctness of each cluster. Silhouette plots were constructed using the squared Euclidean distance between each observation and others within the same cluster. Values closer to 1 show cluster similarity, while values closer to −1 indicate disagreement in cluster membership. A site's cluster membership was also displayed across the four river watersheds to explore spatial patterns. Here a site could be displayed more than once if it fell into more than one cluster over its period of record.

3.2.3. Random forests and SHapley Additive exPlanations (SHAP)

A random forest classification model was used to predict the cluster membership of each observation based on mean annual precipitation, lithology, and LULC using the R package randomForest (Liaw and Wiener, 2002). The model was trained on a random split of 70 % of the data and tested on the other 30 % of the data. Random forests models can be tuned to improve model performance by altering the number of trees generated in each forest (ntree) and the number of parameters evaluated at each split (mtry). The model was tuned to minimize the out-of-bag error (OOB) on the training data. Tuning began by searching for the optimal mtry value, the model was initialized with the default mtry (square root of number of input variables) and then integer values on either side of the default mtry were tried. A new value of mtry was only retained if improved the OOB error by 0.01 %. During mtry tuning, ntree was set to 500 trees. Once the optimal mtry value was found, ntree values between 1 and 1000 were tested using the optimal mtry value (Fig. S1) and the smallest number of trees where the OOB was minimized and stable was selected. Due to the unequal distribution of observations in each cluster, a stratified sampling method was used to train the random forests model on an equal number of observations from each cluster. The number of observations in the smallest cluster was used to set the sample size of the number of training observations. Observations were sampled with replacement. Model performance was evaluated using the accuracy of cluster membership prediction on the test dataset.

To interpret the importance of each feature to the cluster membership predictions of our random forest, we used SHapley Additive exPlanations (SHAP) values, a permutation method that relies on game-theory to provide consistent and locally accurate feature importance measures (Shapley 1953; Lundberg and Lee, 2017; Merrick et al., 2020). SHAP values are distinct in their ability to quantify feature importance within subsets of the data. This contrasts with other permutation methods, which typically provide feature importance information based solely on the average prediction across the entire dataset (Aas et al., 2021; Lundberg et al., 2019). SHAP values can indicate the degree to which input features (e.g., mean annual precipitation) contribute positively or negatively to the random forest model's prediction, in this case the tendency of a given sample to be assigned to one of the clusters. To understand the overall measure of a given feature's importance, we averaged the absolute value of the SHAP values for a given feature across all clusters. To understand how a feature's values (e.g., MAP at a site) contributed to its corresponding SHAP values we constructed strip plots to examine potential dependencies.

4. Results

4.1. ESOM

The trained Umatrix achieved an RMSE of 0.4 (normalized value of 0.03) between each node and its BestMatch (Fig. S6). Continuous low-lying valleys (blue and green) in the ESOM-derived topology indicates zones of similar water chemistry, while ridgelines of brown and white indicate strong differences in water chemistry and help to differentiate distinct chemical signatures across the four watersheds (Fig. 3).

The distribution of solutes across the ESOM (Fig. 4) reveals trends and structure present in the ESOM that become obscured in the overall Umatrix (Fig. 3). In the solute maps, white and yellow colors represent generally lower compositional concentrations, while orange and red colors represent high compositional concentrations (Fig. 4). The most homogeneous zones in the overall Umatrix were located on the edges of the rendered 2-D map of the toroidal shape (Fig. 3) and were concurrent with high Na^+, Cl^-, and SO_4^{2-} and low HCO_3^-, and Ca^{2+} compositional concentrations (displayed on individual maps; Fig. 4). Other solutes K^+, Mg^{2+}, and Si have smaller spatial extents with high values occurring dominantly in the center (Mg^{2+}) and at the upper (Si, K^+) and lowermost extents of the center (K^+) which overlap with areas of high HCO_3^- and Ca^{2+}.

4.2. Clustering

Six clusters emerged as the optimal number of clusters to minimize the within sum of squares distance from each observation to its cluster centroid and fit the heterogenous structure of the ESOM (Fig. S9). All clusters show a positive average silhouette width, with an average silhouette width of 0.514, which is deemed satisfactory given a silhouette width of 1 indicates each data point is unlikely to be assigned to another cluster, while values closer to −1 indicates each data point is misclassified (Haselbeck et al., 2019; Oppel and Schumann, 2020). Fig. 5 shows the cluster membership of ESOM nodes projected onto the trained ESOM. Cluster 1 (yellow, n = 13,817) has the largest number of observations and occurs on the sides of the map. Cluster 2 (light blue, n = 5,834) occurs mainly on the right side of the map, with another grouping on the bottom left side of the map. Cluster 3 (green, n = 4,627) occurs on the upper left side of the map. Cluster 4 (dark blue, n = 3,074) and cluster 6 (pink, n = 1,255) occur in the center of the map near the top and bottom. Lastly, cluster 5 (orange, n = 2,717) occurs in the middle of the map. Evaluating solute distributions across the clusters aids in understanding dominant chemical makeups of observations within a given cluster (Fig. 6). In general, Cluster 1 shows high Cl^-, Na^+, and SO_4^{2-}; Cluster 2 shows high Na^+ and Cl^- but lower SO_4^{2-} than Cluster 1; Cluster 3 shows high SO_4^{2-} but low in Cl^- and Na^+; Cluster 4 is high in HCO_3^- and Ca^{2+}; Cluster 5 is dominated by Mg^{2+}, and Cluster 6 shows observations high in Si. When the temporal characteristics of the nodes within each cluster were evaluated (Fig. S7), neither mean month nor mean decade coincided with any given cluster indicating clusters do not represent seasonal nor long term temporal variability in observations.

Observations from the same site could be assigned to more than one cluster over their period of record. Thus, we evaluated cluster membership over time at the sites to understand a sites stability. Here we selected sites with at least 10 years of data and at least 6 observations per year, which resulted in 90 unique sites of the original 670 (Fig. S8). We found that 9 sites had membership in one cluster, 28 sites had membership in two clusters, 23 sites had membership in 3 clusters, 23 sites had membership in 4 clusters, and 7 had membership in 5 clusters. However, when we evaluated the proportion of time sites spent in each cluster, we found that 44 % of sites spent at least 90 % of their time in their modal cluster, and 83 % of sites spend at least 50 % of their time in their modal cluster. This suggest that while sites generally have one dominant chemical composition, they can change over time though the patterns in cluster membership for the 90 sites with enough data do not show systematic changes over time (Fig. S8).

4.3. Random forest classification

Random forest classification was applied to the dataset using MAP, lithologic classes, and LULC classes from each sub-basin to predict

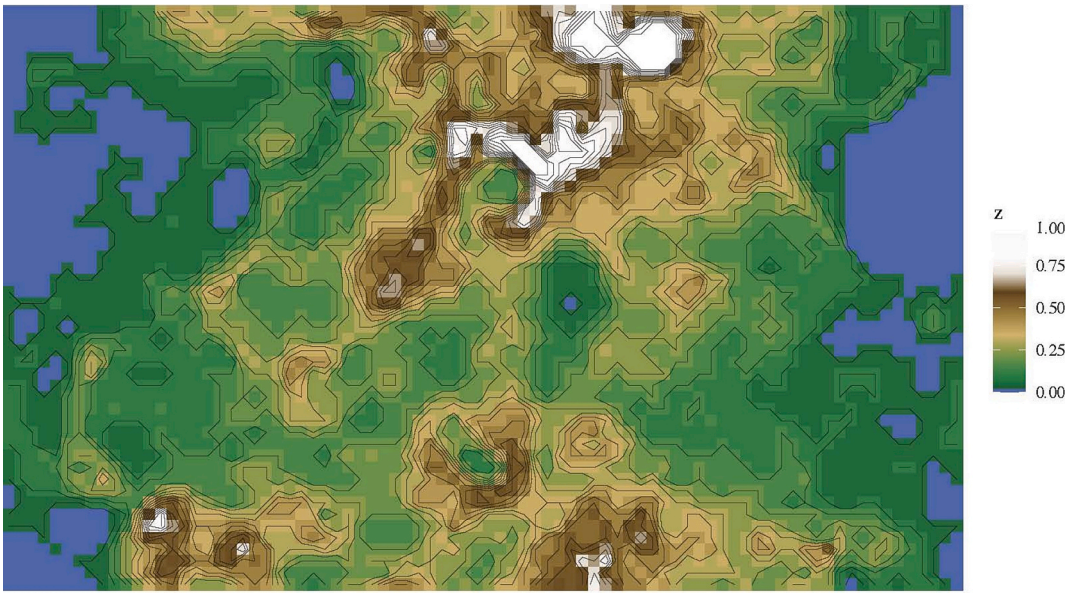


Fig. 3. Umatrix showing the distance from each node to its neighbors in the ESOM after training (RMSE of 0.40). z represents the maximum distance from a node to each of its four neighbors (up, down, right, left) as a proportion (Thrun et al., 2016). The topographic color scales represent the degree of similarity among nodes where blues and greens indicate small distances and homogeneous zones and brown and white indicate larger distances and areas that are anomalous (Lerch et al., 2020). The Umatrix is a toroidal structure that has been unwrapped in this depiction to eliminate the influence of edge effects, with the standard arrangement of a 50 x 82 node mesh.

cluster membership. The optimal performance of the model occurred with a ntree value of 500 and a mtry value of 1 (Fig. S10) and accurately classified sites into their clusters 78 % of the time. Within-cluster error rates varied across classes with cluster 5 having the highest accuracy rate (89 %), and cluster 2 having the lowest accuracy rate (61 %) (Fig. 7, green boxes). Assessment of variable importance conducted using the SHAP values showed that MAP was the most important predictor variable. This was followed by the percent of marsh and swamp land, then evaporite, mudstone and carbonate lithologies (Fig. 7).

5. Discussion

The application of multiple machine learning methods is improving our understanding of controls on stream and groundwater chemistry at a range of scales (Haselbeck et al., 2019; Melo et al., 2019; Nguyen et al., 2015; Vesanto et al., 2000). When paired with clustering algorithms and random forest classification, ESOMs can help to identify major critical zone drivers of hydrogeochemical processes (Addor et al., 2018; Hammond et al., 2021). Here we applied these techniques to regional scale watersheds, focusing on four parallel rivers across Texas with long-term, temporally and spatially variable hydrochemical datasets and their associated spatial attributes (e.g., MAP, lithology, and LULC). Even given spatial and temporal heterogeneity in the 60-year dataset with over 31,000 observations collected from an area spanning 500,000 km^2, the multivariate structures identified in the ESOM and k-means clustering led to six groups with distinct chemical compositions that were well predicted by random forest classification (testing accuracy of 78.9 %). Below we discuss these findings, the potential drivers of each cluster, and implications of this framework for management.

5.1. ESOMs revealed six stream signatures across four large watersheds in Texas

Our results show the potential application of pairing ESOMs and k-means clustering to large datasets to help inform watershed managers on stream quality patterns. When examined spatially across the four watersheds, four of the clusters (1, 3, 5, and 6) demonstrated spatial relationships to known watershed characteristics (Fig. 8), while two clusters (2 and 4) were distributed more heterogeneously in space and are discussed in section 5.3 below.

First, Cluster 1 exhibited chemistry that was dominated by Na^+, Cl^-, and SO_4^{2-} and comprised a homogeneous zone (blue and green). It formed the cluster with the largest number of observations that were predominately located in the lower main stem of the Pecos river, the upper main stem of the Colorado and Brazos rivers, and the upper and middle tributaries and main stem of the Red river (Fig. 8; Cluster 1 – yellow). These regions include substantial areas of cropland, and large bodies of evaporite bedrock outcrops and semi-arid soils containing both halite and gypsum (Richter et al., 1991; Richter & Kreitler, 1986). The presence of evaporite bedrock outcrops can contribute Na^+, Cl^-, and SO_4^{2-} to streams via flushing from rainfall events in these watersheds. In addition, irrigated agriculture and related amendments have been hypothesized to contribute Na^+ and Cl^- to soils and nearby surface waters in this region (Kondash et al., 2020; Yurtseven et al., 2018). Furthermore, irrigated agriculture in the region relies heavily on extracted groundwater which contains higher concentrations of Na^+, Cl^-, and SO_4^{2-} (Bruun et al., 2016).

Second, Cluster 3 had a stream chemistry that was dominated by high SO_4^{2-} and Ca^{2+}. When mapped back onto the watersheds, Cluster 3 occurred on tributaries in the upper portions of all four rivers but was most extensive in the upper portion of the Red river (Fig. 8– green). While other watersheds contain sandstone, the Red has multiple subbasins with over 50 % sandstone (Fig. S2b). The chemical signature here is likely derived from the unique gypsum deposits known to occur in the Permian sandstone underlying this region (Clark et al., 2020; Slade & Buszka, 1994).

Third, Cluster 5 was marked by samples with higher Ca^{2+}, HCO_3^- and Mg^{2+} concentrations. When mapped across Texas, it appeared these samples were located along major mapped faults, particularly in the Colorado and Red river watersheds (Fig. 8; Cluster 5– orange). These samples are located within and surrounding the Llano uplift and the Balcones fault zone (Fig. 8). It has been suggested that faults in these areas may act as conduits to bring deeper groundwater sources to the surface, facilitating local changes in water chemistry (e.g., Ferrill et al., 2008; Schindel, 2019). For example, existing spring samples from along the Balcones fault zone have distinctly higher $^{234}U/^{238}U$ values

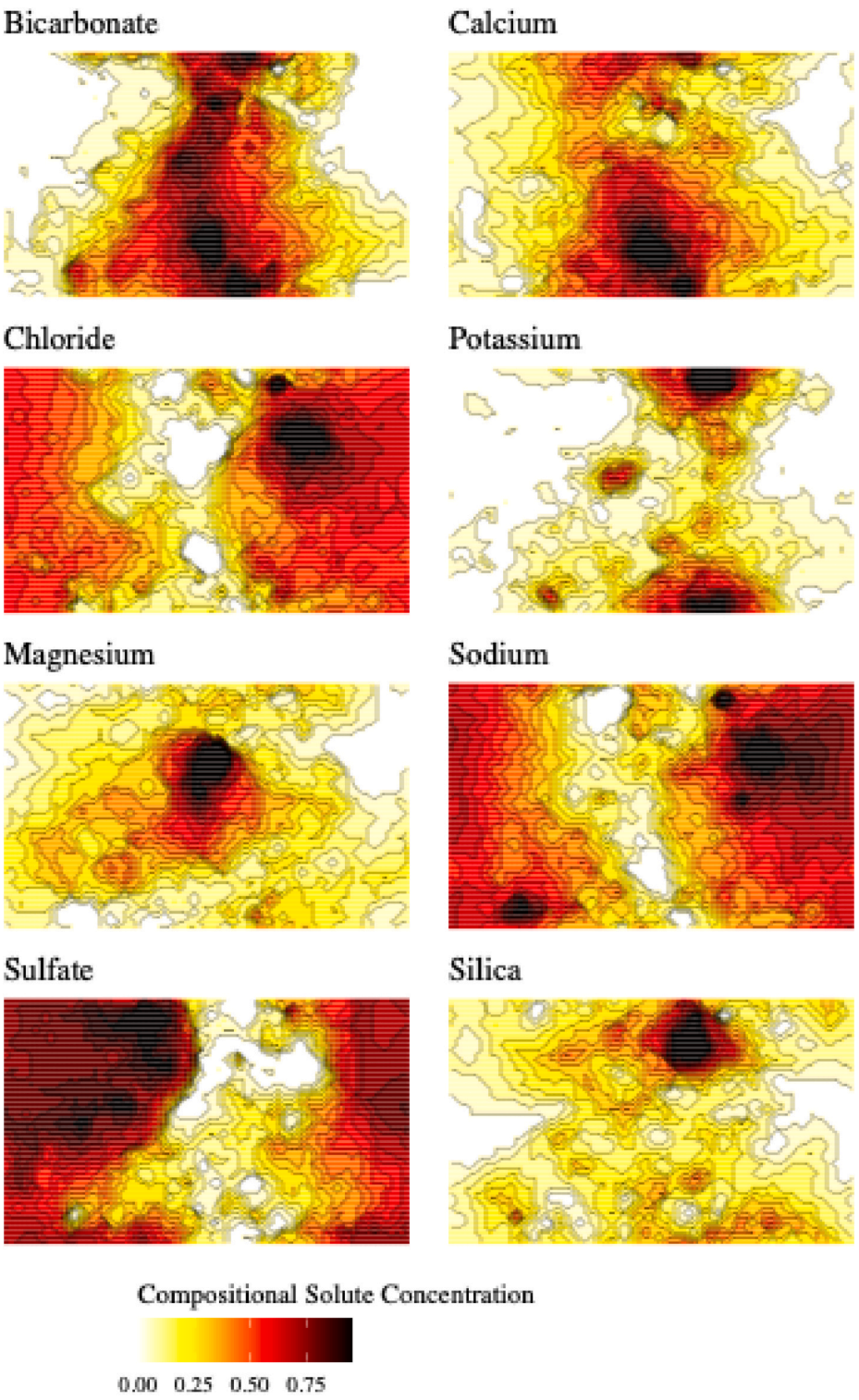


Fig. 4. Maps showing the distribution of each solute (as the proportion of total composition) across the trained ESOM, whites and yellows indicate lower proportional composition, while reds and blacks indicate higher value.

compared to surrounding springs, indicating deeper fluid sources (Kronfeld, 1974). The Balcones fault zone has a prolonged tectonic history spanning 100′s of millions of years, although it is not tectonically active today (Collins, 1993; Ewing, 2005).

Finally, Cluster 6 was distinguished by high Si concentrations and occurred predominantly in the lower Red river watershed (Fig. 8– pink). This region has distinctly higher MAP than the other regions of the four rivers (Fig. 1C). Cluster 6 also within the Colorado and Brazos river watersheds, and again these samples are concentrated within the lower reaches of each river where lithology is dominated by siliciclastic units, and carbonate units are notably absent.

Our overall results support the distinctness of approximately four major water chemical classes (as well as two classes that are less well-defined by chemical variations) and when examined spatially across the four watersheds the chemistry can be discussed in terms of known spatial differences in MAP, lithology, and spatial variability in land use.

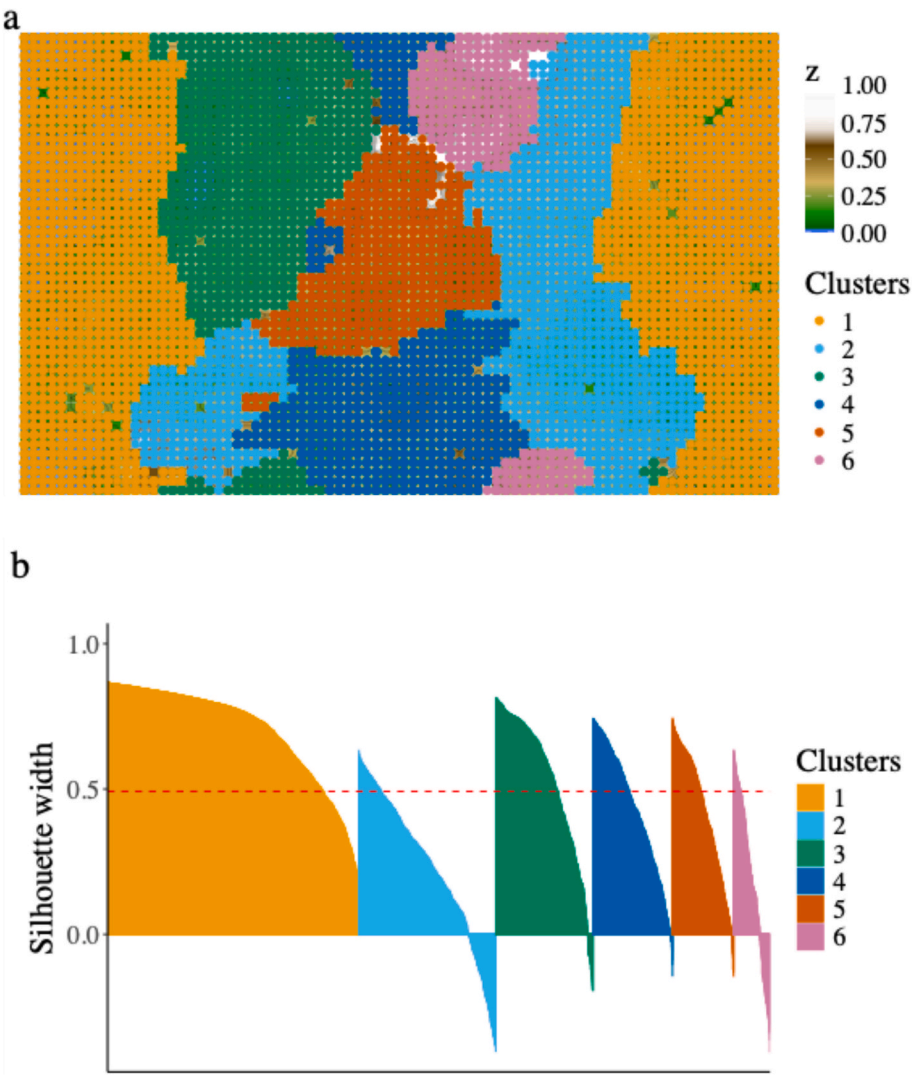


Fig. 5. (a) Umatrix with best matching units colored by cluster membership (k = 6), and (b) silhouette plot showing cluster widths as squared Euclidean distance. Red line shows mean cluster silhouette width (0.514).

5.2. Random forests accurately predicted membership of the stream signatures using sub-basin characteristics

Random forest classification was shown to be a useful tool in combination with ESOM node clustering (Ultsch and Lötsch, 2017; Ultsch et al., 2016; Park et al., 2013), and predicted cluster membership with an acceptable overall accuracy rate using long-term MAP and lithologic and land use classes. Our model performed comparably (78.9 % accuracy) to other random forest classification applications in hydrology, which have been shown to range from 59 % to 90 % (e.g., Johnson et al., 2024; Baudron et al., 2013; Mobley et al., 2021), and allowed for identification of important watershed variables controlling variations in river water chemistry (Addor et al., 2018; Hammond et al., 2021; Olson and Hawkins, 2012).

By examining the range in SHAP and feature values across each of the clusters (Fig. 9), we were able to attribute the critical zone features that best predicted each cluster. Overall, clusters where the ESOM had the smallest distance between nodes were also those best-predicted by the random forests. Samples high in Ca^{2+}, HCO_3^- and Mg^{2+} concentrations were the best predicted by the random forest algorithm (i.e., cluster 5 – orange). In this region, SHAP values indicate carbonate lithology has the most positive contribution to the prediction of cluster five, while the feature value revealed the proportion of carbonate rocks exceeds that of any of the other clusters (Fig. 9). Cluster 1, 3 and 6 were predicted with similar accuracies. SHAP values indicate the presence of evaporites, and lack of sedimentary deposits contributed to the highest SHAP values for Cluster 3, while greater precipitation, forest cover and a lack of cropland contributed the most to the prediction of Cluster 6. Interestingly, Cluster 1 showed several inverse patterns to that of Cluster 5 and 6, where lower precipitation, greater crop cover and a lower concentration of carbonates contributed positively to its prediction. Overall, differences in MAP, distinct differences in underlying lithology, and vegetation patterns which mirror the available amount of water, emerged across the six clusters.

5.3. Esom-random forest framework offers a tool for identifying unknown controls on stream chemistry

One advantage of using ESOMs compared to more traditional multivariate tools such as piper diagrams and principal components analysis is the ability to capture non-linear relationships or structures in the data (Melo et al., 2019; Thrun et al., 2016). In addition, ESOMs help to define biogeochemical groups of large, complex, and highly variable data. As discussed above, our analysis revealed six distinct biogeochemical groups in these four Texas rivers, which could be predicted by simple watershed parameters. Areas where watershed parameters less

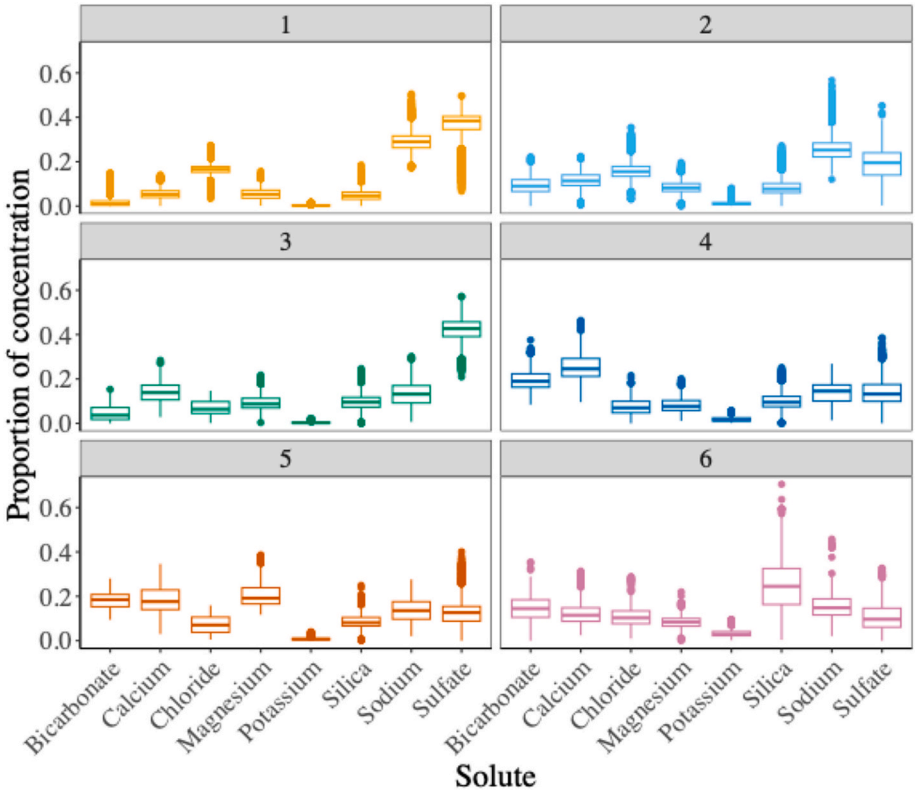


Fig. 6. Boxplots showing the distribution of composition for all observations assigned to Clusters 1–6. Boxplots indicate the mean, 25th, and 75th percentiles as horizontal lines and points show outliers. Boxplot colors indicate clusters as shown on Umatrix.

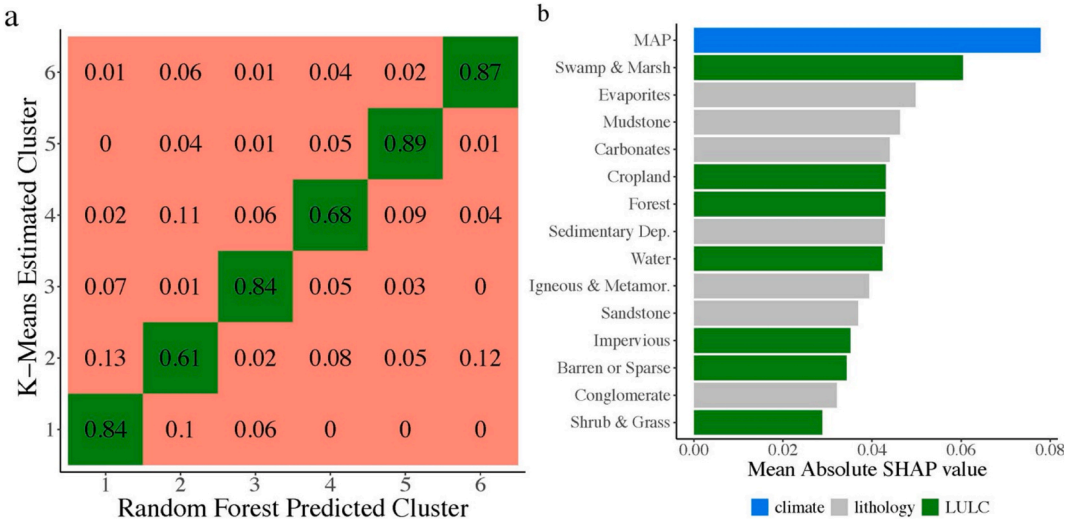


Fig. 7. (a) Confusion matrix showing performance of random forest algorithm on the testing dataset, where the overall accuracy rate was 78.9%. Green diagonal boxes give the proportion of sites, within a given cluster, that were accurately classified. Salmon colored boxes in the same row show the proportion of time sites within a given cluster were misclassified into another cluster. For example, Cluster 1 was accurately classified 84% of the time, while 10% of the time, sites in Cluster 1 were classified as Cluster 2, and 6% of the time sites in Cluster 1 were classified as Cluster 3. For the confusion matrix all rows sum to 1. (b) Mean absolute SHAP values of each considered variable explaining the prediction of a given cluster. Larger values indicate the variable contributes more to the prediction and is thus more important to the model's performance. Bars are colored by their general feature class: climate (blue), lithology (gray), and land use land cover (LULC; green). Gini impurity is provided in Fig. S11 and the correlation matrix of input features is provided in Fig. S12.

accurately predicted the cluster groups (Cluster 2 and 4) may warrant further investigation of underlying controls on water chemistry.

Potential explanations of chemistry observed in Cluster 2 and Cluster 4 are deep brines where reduction has removed SO_4^{2-} increasing the ratio of Na^+ and Cl^- to SO_4^{2-} (McMahon et al., 2016) or by water discharging from carbonate aquifers that is not associated with the fault zone, respectively (elevated Ca^{2+} and HCO^{3-}). Other processes that can control water chemistry are large reservoirs and wetlands, which were rudimentarily incorporated into our model as open water and swamp/marsh, may explain the challenges in predicting these clusters. Violin plots (Fig. 9) reveal that the greatest SHAP values for Cluster 2 were linked to a greater degree of open water and land classified as swamp/

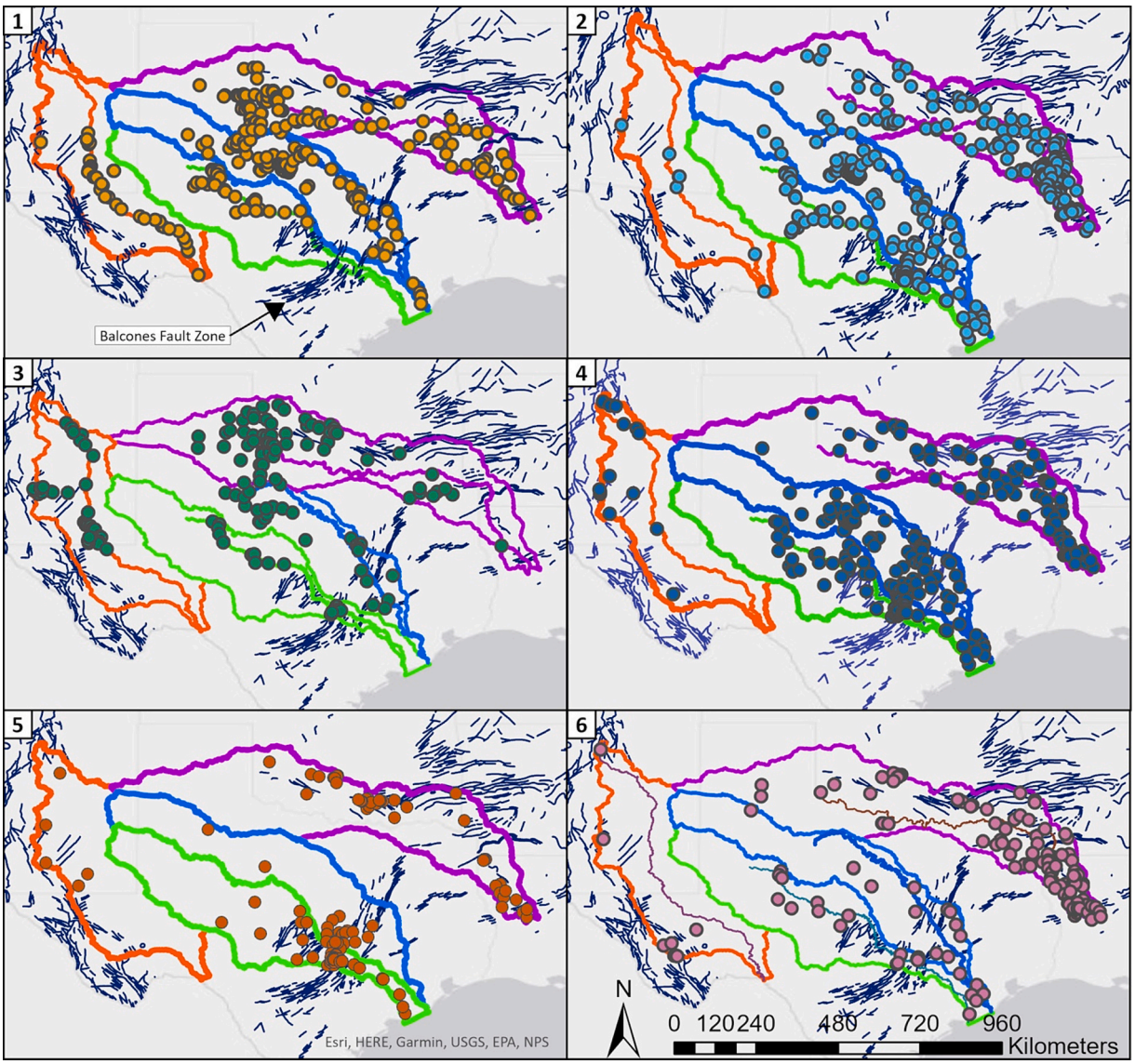


Fig. 8. Map of observations across the four watersheds showing the ESOM informed k-means cluster membership.

marsh, yet this was not observed for Cluster 4. Land cover changes such as the increase in cropland and impervious surface may also contribute to changes in water chemistry. While we were able to account for land cover changes between 1985-present, data availability did not support analysis of earlier land cover changes. In our study, sub-basins across all watersheds have experienced some change in landcover with the greatest degree of impervious surface created in the Brazos and Colorado river sub-basins and some increase in cropland cover in all but the Red river, yet these two land cover types did not emerge as highly important in Cluster 2 or 4. Additional monitoring and analysis of different watershed characteristics is required to better understand the dominant controls on stream chemistry represented by Clusters 2 and 4.

While this work accounts for spatial variability in MAP and lithology, and both spatial and temporal variability in LULC, there are additional metrics that would be valuable to include in future work. Stream discharge and precipitation metrics beyond MAP were not incorporated into the model, which could provide insight into the temporal dynamics in flowpath variability that contribute to river chemical composition (Bush et al., 2023; Warix et al., 2023). Additionally, the included LULC and MAP data only captured dynamics from 1985 to present, while our stream chemistry data stretch back to 1944, this temporal incongruity increases the uncertainty in our models and likely contributes to underperformance. However, given that the dominant control on ESOM structure was spatial rather than temporal (Fig. S7), the combination of machine learning analyses presented above provide a robust framework for assessing regional controls on stream chemical composition. Lastly, our work hints that some sites exhibit temporal variability in their chemical composition and cluster membership; future work could leverage the frequency of observations of stream chemistry, LULC, and climate measurements to understand temporal variations in stream chemistry composition and behavior including interannual and seasonal shifts.

6. Conclusions

We present a robust framework for predicting stream chemistry from spatially and temporally inconsistent hydrochemical data by leveraging emergent self-organizing maps (ESOMs), k-means clustering, and random forest classification algorithms. These methods support analysis of one-off and disparate time period samples; as such we are able to learn from traditionally excluded data that is typically removed due to low sampling frequency or temporal irregularity. Our results show that water chemistry across Texas can be described using six major classes that are generally well-predicted using only a few climate, land use, and

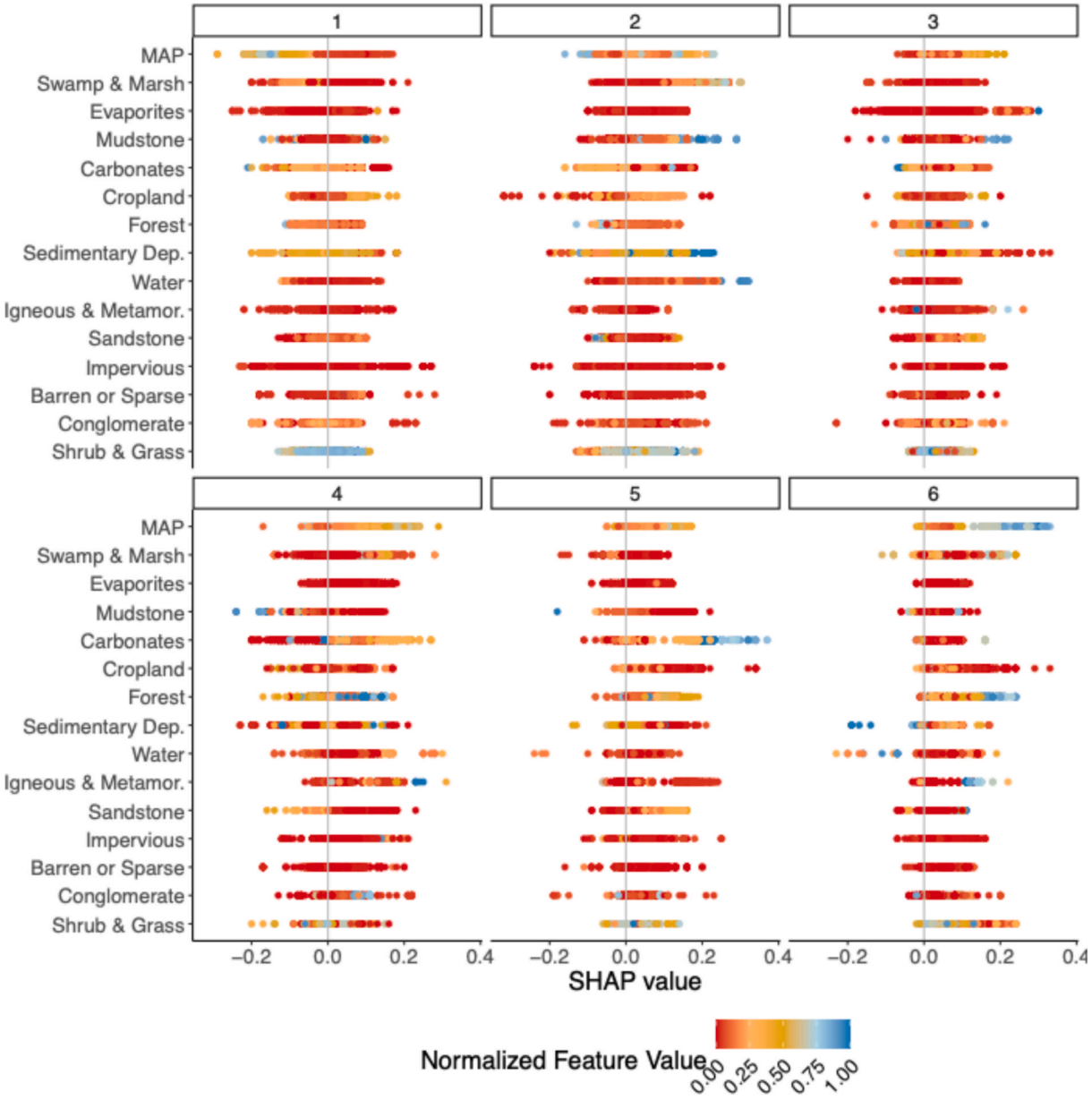


Fig. 9. SHAP strip plots for each cluster. Each point represents one observation, and the attributes are ordered by overall importance. Positive SHAP values indicates the feature contributes towards the prediction of that cluster. Larger SHAP values indicate a greater contribution of the feature toward the prediction. Color ramp indicates the normalized feature value with 0 indicating low values (red) and 1 indicating high values (blue). For example, a MAP value of 400 mm would be low, while a value of 1600 mm would be high.

lithology factors. Important factors in predicting classes were dominated by MAP, proportion of land classified as swamp and marsh, and reactive lithologies.

Improved understanding of dominant controls on stream chemistry, especially using widely available spatial data products, could support improved decision making based on fewer observations. For example, this type of analysis could show that controls on stream water chemistry for given reaches are dominated by underlying lithology and climate patterns, while other areas are more directly influenced by land use. This information could highlight areas where restoration efforts focused on ameliorating land use impacts on water quality have the largest potential benefits. These techniques also allow us to identify areas that are more chemically variable or more uncertain and facilitate improved sampling protocols to target these areas. For example, our work highlights areas where management and additional monitoring is needed to assess controls on water quality not captured by sub-basin characteristics such as climate, land use, and lithology (Clusters 2 and 4).

Our framework demonstrates significant potential for applying machine learning to explore the relationships between watershed factors and stream chemistry, thereby enhancing regional water management strategies. Additionally, previous studies utilizing random forest and clustering analyses (Bolotin et al., 2023; Johnson et al., 2024b) have successfully elucidated large-scale controls on river chemistry, indicating that this framework may be adaptable to larger scales. In addition, this study demonstrates machine learning has value not only in its stereotypical use as a black-box predictor, but also as a toolkit for understanding physical processes, such as the potential controls of open water and swamp and marshes (Cluster 2) or cropland (Cluster 1) on stream chemistry across varied lithology. This is consistent with an emerging body of literature on interpretable AI, and in particular the use of machine learning for knowledge-discovery, in the Earth and

environmental sciences (for reviews and syntheses, refer to e.g., Fleming et al., 2021b; McGovern et al., 2019; Nearing et al., 2021; Reichstein et al., 2019).

CRediT authorship contribution statement

G.M. Goldrich-Middaugh: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **K.R. Johnson:** Writing – review & editing, Visualization, Methodology. **L. Ma:** Writing – review & editing, Funding acquisition. **M.A. Engle:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **S.W. Fleming:** Writing – review & editing, Methodology, Formal analysis. **J.W. Ricketts:** Writing – review & editing, Funding acquisition. **P.L. Sullivan:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

We acknowledge funding support from the National Science Foundation (EAR-1933261; EAR-1933259; EAR - 2012796). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Lourdes Moreu for her support in initial site selection and project setup. We are grateful for the thoughtful feedback from Florian Lerch, Nuria Andreu Garcia, and Jennifer Herrera. Finally, we would like to acknowledge the thoughtful and constructive comments from the three reviewers that made this manuscript stronger.

Data Availability Statement

Stream water quality data can be accessed from the USGS Water Quality Portal (https://www.water-qualitydata.us/), rock types can be found at the Texas Water Development Board (www.twdb.texas.gov), comprehensive land use data are from the Global Land Cover Change Dataset (https://gee-community-catalog.org/projects/glc_fcs/), annual precipitation data can be accessed from PRISM (https://prism.oregonstate.edu/). Codes and input files used for analysis can be found on Zenoto (Johnson, 2025).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.133319.

Data availability

All code is available on Gi

References

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N., Clark, M.P., 2018. A Ranking of Hydrological Signatures Based on Their Predictability in Space. Water Resour. Res., i 8792–8812. https://doi.org/10.1029/2018WR022606.

Aitkenhead-Peterson, J.A., Nahar, N., Harclerode, C.L., Stanley, N.C., 2011. Effect of urbanization on surface water chemistry in south-central Texas. Urban Ecosystems 14 (2), 195–210. https://doi.org/10.1007/s11252-010-0147-2.

Allan, J.D., 2004. Landscapes and Riverscapes: The Influence of Land Use on Stream Ecosystems. Annu. Rev. Ecol. Evol. Syst. 35 (1), 257–284. https://doi.org/10.1146/annurev.ecolsys.35.120202.110122.

Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif. Intell. 298, 103502.

Baronas, J. J., Torres, M. A., Clark, K. E., & West, A. J. (2017). Mixing as a driver of temporal variations in river hydrochemistry: 2. Major and trace element concentration dynamics in the Andes-Amazon transition. Water Resources Research. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x.

Baudron, P., Alonso-Sarría, F., García-Aróstegui, J.L., Cánovas-García, F., Martínez-Vicente, D., Moreno-Brotóns, J., 2013. Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification. J. Hydrol. 499, 303–315.

Bolotin, L.A., Summers, B.M., Savoy, P., Blaszczak, J.R., 2023. Classifying freshwater salinity regimes in central and western U.S. streams and rivers. Limnol. Oceanogr. Lett. 8 (1), 103–111. https://doi.org/10.1002/lol2.10251.

Brantley, S.L., Lebedeva, M.I., Balashov, V.N., Singha, K., Sullivan, P.L., Stinchcomb, G., 2017. Toward a conceptual model relating chemical reaction fronts to water flow paths in hills. Geomorphology 277, 100–117. https://doi.org/10.1016/j.geomorph.2016.09.027.

Brooks, K.N., Ffolliott, P.F., Magner, J.A., 2012. Hydrology and the Management of Watersheds. John Wiley & Sons.

Brown, S.C., Lester, R.E., Versace, V.L., Fawcett, J., Laurenson, L., 2014. Hydrologic landscape regionalisation using deductive classification and random forests. PLoS ONE 9 (11). https://doi.org/10.1371/journal.pone.0112856.

Bruun, B., Jackson, K., Lake, P., & Walker, J. (2016). Texas Aquifers Study: Groundwater Quantity, Quality, Flow, and Contributions to Surface Water.

Bush, S.A., Birch, A.L., Warix, S.R., Sullivan, P.L., Gooseff, M.N., McKnight, D.M., Barnard, H.R., 2023. Dominant source areas shift seasonally from longitudinal to lateral contributions in a montane headwater stream. J. Hydrol. 617, 129134.

Chavoshi, S., Azmin Sulaiman, W.N., Saghafian, B., Sulaiman, M.N.B., Latifah, A.M., 2012. Soft and hard clustering methods for delineation of hydrological homogeneous regions in the southern strip of the Caspian Sea Watershed. J. Flood Risk Manage. 5 (4), 282–294. https://doi.org/10.1111/j.1753-318X.2012.01149.x.

Chen, I.T., Chang, L.C., Chang, F.J., 2018. Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps. J. Hydrol. 556, 131–142. https://doi.org/10.1016/j.jhydrol.2017.10.015.

Clark, A. K., Morris, R. E., & Pedraza, D. E. (2020). Geologic framework and hydrostratigraphy of the edwards and trinity aquifers within northern medina county, Texas. In U.S. Geological Survey Scientific Investigations (Vol. 2020, Issue 3461). https://doi.org/10.3133/sim3461.

Collins, E. W. (1993). Fracture zones between overlapping en echelon fault strands: Outcrop analogs within the Balcones Fault Zone, Central Texas.

Council, N.R., 2012. Challenges and opportunities in the hydrologic sciences. National Academies Press.

De Cicco, L. A., Hirsch, R. M., Lorenz, D., & Watkins, W. D. (2018). dataRetrieval: R packages for discovering and retrieving water data available from Federal hydrologic web services (2.7.6). U.S. Geological Survey.

Dupré, B., Dessert, C., Oliva, P., Goddéris, Y., Viers, J., François, L., Millot, R., Gaillardet, J., 2003. Rivers, chemical weathering and Earth's climate. Comptes Rendus - Geoscience 335 (16), 1141–1160. https://doi.org/10.1016/j.crte.2003.09.015.

Esri, 2021. ArcGIS Pro: Ver 2.9. Environmental Systems Research Institute, Redlands, CA.

Ewing, T.E. (2005). Phanerozoic Development of the Llano Uplift: Vol. 49(9).

Ferrill, D.A., Morris, A.P., Sims, D.W., Green, R., Franklin, N., Waiting, D.J., 2008. Geologic controls on interaction between the Edwards and Trinity Aquifers. Balcones fault system, Texas.

Fleming, S.W., Garen, D.C., Goodbody, A.G., McCarthy, C.S., Landers, L.C., 2021a. Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence. J. Hydrol. 602, 126782.

Fleming, S. W., Titus, M., Watson, J. R., & Doring, D. (2019). Technology Demonstration for One-Week-Ahead Forecasting of Toxic Algal Blooms in the US Army Corps of Engineers Reservoir at Detroit Lake using Machine Learning. 2019, GH44A-12.

Fleming, S.W., Watson, J.R., Ellenson, A., Cannon, A.J., Vesselinov, V.C., 2021b. Machine learning in Earth and environmental science requires education and research policy reforms. Nat. Geosci. 14 (12), 878–880. https://doi.org/10.1038/s41561-021-00865-3.

Gaillardet, J., Dupre, B., Louvat, P., Allegre, C.J., 1999. Global silicate weathering and CO2 consumption rates deduced from the chemistry of large rivers. Chem. Geol. 159, 3–30.

Gamble, A., Babbar-Sebens, M., 2012. On the use of multivariate statistical methods for combining in-stream monitoring data and spatial analysis to characterize water quality conditions in the White River Basin, Indiana, USA. Environ. Monit. Assess. 184 (2), 845–875. https://doi.org/10.1007/s10661-011-2005-y.

Giri, S., 2021. Water quality prospective in Twenty First Century: Status of water quality in major river basins, contemporary strategies and impediments: A review. Environ. Pollut. 271, 116332.

Godsey, S.E., Kirchner, J., Clow, D., 2009. Concentration-discharge relationships reflect chemostatic characteristics of US Catchments. Hydrol. Process. 32 (18), 2829–2844. https://doi.org/10.1002/hyp.13235.

Goldrich-Middaugh, G.M., Ma, L., Engle, M.A., Ricketts, J.W., Soto-Montero, P., Sullivan, P.L., 2022. Regional Drivers of Stream Chemical Behavior: Leveraging Lithology, Land Use, and Climate Gradients Across the Colorado River, Texas USA. Water Resour. Res. 58 (11), e2022WR032155. https://doi.org/10.1029/2022WR032155.

Güler, C., Thyne, G.D., McCray, J.E., Turner, A.K., 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrgeol. J. 10 (4), 455–474. https://doi.org/10.1007/s10040-002-0196-6.

Hammond, J.C., Zimmer, M., Shanafield, M., Kaiser, K., Godsey, S.E., Mims, M.C., Zipper, S.C., Burrows, R.M., Kampf, S.K., Dodds, W., Jones, C.N., Krabbenhoft, C.A., Boersma, K.S., Datry, T., Olden, J.D., Allen, G.H., Price, A.N., Costigan, K., Hale, R., Allen, D.C., 2021. Spatial Patterns and Drivers of Nonperennial Flow Regimes in the Contiguous United States. Geophys. Res. Lett. 48 (2), 1–11. https://doi.org/10.1029/2020GL090794.

Harwell, G. R., McDowell, J., Gunn-Rosas, C., & Garrett, B. (2020). Precipitation, temperature, groundwater-level elevation, streamflow, and potential flood storage trends within the Brazos, Colorado, Big Cypress, Guadalupe, Neches, Sulphur, and Trinity River Basins in Texas Through 2017. Scientific Investigations Report, April.

Haselbeck, V., Kordilla, J., Krause, F., Sauter, M., 2019. Self-organizing maps for the identification of groundwater salinity sources based on hydrochemical data. J. Hydrol. 576 (June), 610–619. https://doi.org/10.1016/j.jhydrol.2019.06.053.

Jager, H.I., Baskaran, L.M., Schweizer, P.E., Turhollow, A.F., Brandt, C.C., Srinivasan, R., 2015. Forecasting changes in water quality in rivers associated with growing biofuels in the Arkansas-White-Red river drainage, USA. GCB Bioenergy 7 (4), 774–784. https://doi.org/10.1111/gcbb.12169.

Jankowski, K.J., Johnson, K., Sethna, L., Julian, P., Wymore, A.S., Shogren, A.J., Thomas, P.K., Sullivan, P.L., McKnight, D.M., McDowell, W.H., Heindel, R., Jones, J. B., Wollheim, W., Abbott, B., Deegan, L., Carey, J.C., 2023. Long-Term Changes in Concentration and Yield of Riverine Dissolved Silicon From the Poles to the Tropics. Global Biogeochem. Cycles 37 (9), e2022GB007678. https://doi.org/10.1029/2022GB007678.

Jiang, W., Xu, X., Hall, R., Zhang, Y., Carroll, K.C., Ramos, F., Engle, M.A., Lin, L., Wang, H., Sayer, M., Xu, P., 2022. Characterization of Produced Water and Surrounding Surface Water in the Permian Basin, the United States. J. Hazard. Mater. 430 (February), 128409. https://doi.org/10.1016/j.jhazmat.2022.128409.

Johnson, K (2025) hydrokeira/Texas_Rivers_ESOM: Texas River Geochemistry. (v1.0). [Software]. Zenodo. https://doi.org/10.5281/zenodo.15058457.

Johnson, K., Jankowski, K. J., Carey, J., Lyon, N. J., McDowell, W. H., Shogren, A., Wymore, A., Sethna, L., Wollheim, W. M., Poste, A. E., Kortelainen, P., Heindel, R., Laudon, H., Räike, A., Jones, J. B., McKnight, D., Julian, P., Bush, S., & Sullivan, P. L. (2024a). Establishing fluvial silicon regimes and their stability across the Northern Hemisphere. Limnology and Oceanography Letters, n/a(n/a). https://doi.org/10.1002/lol2.10372.

Johnson, K., Jankowski, K.J., Carey, J.C., Sethna, L.R., Bush, S.A., McKnight, D., Sullivan, P.L., 2024. Climate, hydrology, and nutrients control the seasonality of Si concentrations in rivers. J. Geophys. Res. Biogeosciences 129 (9), e2024JG008141.

Keen, R.M., Nippert, J.B., Sullivan, P.L., Ratajczak, Z., Ritchey, B., O'Keefe, K., Dodds, W. K., 2023. Impacts of Riparian and Non-riparian Woody Encroachment on Tallgrass Prairie Ecohydrology. Ecosystems 26 (2), 290–301. https://doi.org/10.1007/s10021-022-00756-7.

Kirchner, J.W., 2009. Catchments as Simple Dynamical Systems: Catchment Characterization , Rainfall-Runoff Modeling , and Doing Hydrology Backward 45, 1–34.

Kohonen, T., 2001. Self-Organizing Maps (3 edition.). 30 in Springer Series in Information Sciences.

Konapala, G., Mishra, A., 2020. Quantifying Climate and Catchment Control on Hydrological Drought in the Continental United States. Water Resour. Res. 56 (1), 1–25. https://doi.org/10.1029/2018WR024620.

Kondash, A.J., Redmon, J.H., Lambertini, E., Feinstein, L., Weinthal, E., Cabrales, L., Vengosh, A., 2020. The impact of using low-saline oilfield produced water for irrigation on water and soil quality in California. Sci. Total Environ. 733, 139392. https://doi.org/10.1016/j.scitotenv.2020.139392.

Kronfeld, J., 1974. Uranium deposition and Th-234 alpha-recoil: An explanation for extreme U-234/U-238 fractionation within the Trinity aquifer. Earth Planet. Sci. Lett. 21 (3), 327–330.

Lerch, F., Thrun, M., Pape, F., Paebst, R., & Ultsch, A. (2020). Umatrix: Visualization of Structures in High-Dimensional Data (R package version 3.3).

Li, L., Knapp, J.L.A., Lintern, A., Ng, G.-H.-C., Perdrial, J., Sullivan, P.L., Zhi, W., 2024. River water quality shaped by land–river connectivity in a changing climate. Nat. Clim. Chang. 14 (3), 225–237. https://doi.org/10.1038/s41558-023-01923-x.

Li, L., Sullivan, P.L., Benettin, P., Cirpka, O.A., Bishop, K., Brantley, S.L., Knapp, J.L., van Meerveld, I., Rinaldo, A., Seibert, J., 2021. Toward catchment hydro-biogeochemical theories. Wiley Interdiscip. Rev. Water 8 (1), e1495.

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

Liu, L., Zhang, X., Zhao, T., 2023. GLC_FCS30D: the first global 30-m land-cover dynamic monitoring product with fine classification system from 1985 to 2022. Zenodo.

Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree Ensembles. Epub ahead of print 7 March 2019. DOI: 10.48550/arXiv.1802.03888.

Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (2017, accessed 9 March 2025).

McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., Smith, T., 2019. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. Bull. Am. Meteorol. Soc. 100 (11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1.

McMahon, P.B., Böhlke, J.K., Dahm, K.G., Parkhurst, D.L., Anning, D.W., Stanton, J.S., 2016. Chemical Considerations for an Updated National Assessment of Brackish Groundwater Resources. Groundwater 54 (4), 464–475. https://doi.org/10.1111/gwat.12367.

Melo, D.S., Gontijo, E.S.J., Frascareli, D., Simonetti, V.C., Rosa, A.H., Friese, K., 2019. Self - Organizing Maps for Evaluation of Biogeochemical Processes and Temporal Variations in Water Quality of Subtropical Reservoirs. Water Resour. Res. 268–281. https://doi.org/10.1029/2019WR025991.

Merrick, L., Taly, A., 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In: Holzinger, A., Kieseberg, P., Tjoa, A.M. (Eds.), Machine Learning and Knowledge Extraction. Springer International Publishing, Cham, pp. 17–38.

Mobley, W., Sebastian, A., Blessing, R., Highfield, W.E., Stearns, L., Brody, S.D., 2021. Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast Texas. Nat. Hazards Earth Syst. Sci. 21 (2), 807–822.

NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team (2019). ASTER Global Digital Elevation Model V003 . NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2024-12-06 from https://doi.org/10.5067/ASTER/ASTGTM.003.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. J. Water Process Eng. 48, 102920. https://doi.org/10.1016/j.jwpe.2022.102920.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What Role Does Hydrological Science Play in the Age of Machine Learning? Water Resour. Res. 57 (3). https://doi.org/10.1029/2020WR028091.

Nguyen, T.T., Kawamura, A., Tong, T.N., Nakagawa, N., Amaguchi, H., Gilbuena, R., 2015. Clustering spatio-seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta. VietnamJournal of Hydrology 522, 661–673. https://doi.org/10.1016/j.jhydrol.2015.01.023.

Olson, J.R., Hawkins, C.P., 2012. Predicting natural base-flow stream water chemistry in the western United States. Water Resour. Res. 48 (2). https://doi.org/10.1029/2011WR011088.

Oppel, H., Schumann, A.H., 2020. Machine learning based identification of dominant controls on runoff dynamics. Hydrol. Process. 34 (11), 2450–2465. https://doi.org/10.1002/hyp.13740.

Park, Y.-S., Chung, Y.-J., Moon, Y.-S., 2013. Hazard ratings of pine forests to a pine wilt disease at two spatial scales (individual trees and stands) using self-organizing map and random forest. Eco. Inform. 13, 40–46. https://doi.org/10.1016/j.ecoinf.2012.10.008.

PRISM Climate Group, . Oregon State University. accessed Dec 2021. https://prism.oregonstate.edu. .

Qin, C., Zhu, A.X., Pei, T., Li, B., Zhou, C., Yang, L., 2007. An adaptive approach to selecting a flow partition exponent for a multiple flow direction algorithm. Int. J. Geogr. Inf. Sci.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Richter, B. C., Dutton, A. R., & Kreitler, C. W. (1991). Identification of sources and mechanisms of salt-water pollution affecting ground-water quality: A case study, west Texas.

Richter, B. C., & Kreitler, C. W. (1986). Identification of sources of ground-water salinization using geochemical techniques.

Sadayappan, K., Keen, R., Jarecke, K.M., Moreno, V., Nippert, J.B., Kirk, M.F., Sullivan, P.L., Li, L., 2023. Drier streams despite a wetter climate in woody-encroached grasslands. J. Hydrol. 627, 130388. https://doi.org/10.1016/j.jhydrol.2023.130388.

Sadayappan, K., Kerins, D., Shen, C., Li, L., 2022. Nitrate concentrations predominantly driven by human, climate, and soil properties in US rivers. Water Res. 226, 119295. https://doi.org/10.1016/j.watres.2022.119295.

Sahour, H., Gholami, V., Vazifedan, M., 2020. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. J. Hydrol. 591 (July), 125321. https://doi.org/10.1016/j.jhydrol.2020.125321.

Schindel, G. M. (2019). Genesis of the Edwards (Balcones fault zone) Aquifer.

Schweitzer, P., 2011. Combined geologic map data for the conterminous US derived from the USGS state geologic map compilation. US Geological Survey, Available at Https://mrdata. Usgs. Gov/geology/state/geol_poly. Zip.

Shapley, L.S., 1953. Stochastic Games*. Proc. Natl. Acad. Sci. 39, 1095–1100.

Shi, P., Zhang, Y., Li, Z., Li, P., Xu, G., 2017. Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales. Catena 151, 182–190. https://doi.org/10.1016/j.catena.2016.12.017.

Singh, S. K., Pahlow, M., Booker, D. J., Shankar, U., & Chamorro, A. (2019). Towards baseflow index characterisation at national scale in New Zealand. Journal of Hydrology, 568(July 2018), 646–657. https://doi.org/10.1016/j.jhydrol.2018.11.025.

Singha, K., Sullivan, P.L., Billings, S.A., Walls, L., Li, L., Jarecke, K.M., Barnard, H.R., Gasparini, N.M., Madoff, R.D., Dhital, S., Jones, C., Kastelic, E.C., Ma, L., Perilla-Castillo, P., Song, B., Zhu, T., 2024. Expanding the Spatial Reach and Human Impacts of Critical Zone Science. Earth's. Future 12 (3), e2023EF003971. https://doi.org/10.1029/2023EF003971.

Sinha, S., Singh, T.N., Singh, V.K., Verma, A.K., 2010. Epoch determination for neural network by self-organized map (SOM). Comput. Geosci. 14 (1), 199–206. https://doi.org/10.1007/s10596-009-9143-0.

Slade, R. M., & Buszka, P. M. (1994). Characteristics of Streams and Aquifers and Processes Affecting the Salinity of Water in the Upper Colorado River Basin, Texas.

Stoesser, D.B., Green, G. M., Morath, L. C., Heran, W. D., Wilson, A. B., Moore, D. W., and Van Gosen, B.S. (2007) Preliminary integrated geologic map databases for the United States: Central States: Montana, Wyoming, Colorado, New Mexico, North Dakota, South Dakota, Nebraska, Kansas, Oklahoma, Texas, Iowa, Missouri, Arkansas, and Louisiana.U.S. Geological Survey. https://pubs.usgs.gov/of/2005/1351/.

Stewart, B., Shanley, J.B., Kirchner, J.W., Norris, D., Alder, T., Bristol, C., Harpold, A.A., Perdrial, J.N., Rizzo, D.M., Sterle, G., Underwood, K.L., Wen, H., Li, L., 2022. Streams as mirrors: Reading the subsurface water chemsitry from stream chemsitry. Water Resources Research 58 (1). https://doi.org/10.1029/2021WR029931.

Team, R. c, 2021. R: A language environment for statistical computing (4.0.4). R Foundation for Statistical Computing.

Thrun, M.C., Lerch, F., Lötsch, J., Ultsch, A., 2016. Visualization and 3D Printing of Multivariate Data of. Biomarkers. 4617, 7–16.

Torres, M. A., West, A. J., Clark, K. E., & Feakins, S. J. (2017). Mixing as a driver of temporal variations in river hydrochemistry: 1. Insights from conservative tracers in the Andes-Amazon transition. Water Resources Research. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x.

Ultsch, A., 1999. Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. In Knowledge Creation Diffusion Utilization.

Ultsch, A., 2007. Emergence in self organizing feature maps. International Workshop on Self-Organizing Maps. Proceedings.

Ultsch, A., Behnisch, M., & Lötsch, J. (2016). ESOM visualizations for quality assessment in clustering. 39–48.

Ultsch, A., Lötsch, J., 2017. Machine-learned cluster identification in high-dimensional data. J. Biomed. Inform. 66, 95–104. https://doi.org/10.1016/j.jbi.2016.12.011.

Vesanto, J., Alhoniemi, E., Member, S., 2000. Clustering of the Self-Organizing Map. IEEE Trans. Neural Netw. 11 (3), 586–600.

Wai, K.P., Chia, M.Y., Koo, C.H., Huang, Y.F., Chong, W.C., 2022. Applications of deep learning in water quality management: A state-of-the-art review. J. Hydrol. 613, 128332. https://doi.org/10.1016/j.jhydrol.2022.128332.

Warix, S.R., Navarre-Sitchler, A., Manning, A.H., Singha, K., 2023. Local topography and streambed hydraulic conductivity influence riparian groundwater age and groundwater-surface water connection. Water Resour. Res. 59 (9), e2023WR035044.

Wehrens, R., Buydens, L.M.C., 2007. Self- and Super-organizing maps in R: The kohonen Package. JSS Journal of Statistical Software 21 (5).

Wlostowski, A.N., Molotch, N., Anderson, S.P., Brantley, S.L., Chorover, J., Dralle, D., Kumar, P., Li, L., Lohse, K.A., Mallard, J.M., McIntosh, J.C., Murphy, S.F., Parrish, E., Safeeq, M., Seyfried, M., Shi, Y., Harman, C., 2021. Signatures of Hydrologic Function Across the Critical Zone Observatory Network. Water Resour. Res. 57 (3). https://doi.org/10.1029/2019WR026635.

Xu Fei, E., & Harman, C. J. (2020). Technical Note: A data-driven method for estimating the composition of end-members from streamwater chemistry observations. Hydrology and Earth System Sciences Discussions, June, 1–15. https://doi.org/10.5194/hess-2020-250.

Yang, M., Olivera, F., 2023. Classification of watersheds in the conterminous United States using shape-based time-series clustering and Random Forests. J. Hydrol. 620, 129409. https://doi.org/10.1016/j.jhydrol.2023.129409.

Yurtseven, E., Çolak, M.S., Öztürk, A., Öztürk, H.S., 2018. Drainage water salt load variations related to the salinity and leaching ratios of irrigation water. Tarim Bilimleri Dergisi 24 (3), 394–402. https://doi.org/10.15832/ankutbd.456667.

Zhang, X., Zhao, T., Xu, H., Liu, W., Wang, J., Chen, X., Liu, L., 2024. GLC_FCS30D: the first global 30 m land-cover dynamics monitoring product with a fine classification system for the period from 1985 to 2022 generated using dense-time-series Landsat imagery and the continuous change-detection method. Earth Syst. Sci. Data 16, 1353–1381. https://doi.org/10.5194/essd-16-1353-2024.

Zhi, W., Appling, A.P., Golden, H.E., Podgorski, J., Li, L., 2024. Deep learning for water quality. Nat. Water. https://doi.org/10.1038/s44221-024-00202-z.

Zhi, W., Ouyang, W., Shen, C., Li, L., 2023. Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. Nat. Water 1 (3), 249–260. https://doi.org/10.1038/s44221-023-00038-z.