Regional Drivers of Stream Chemical Behavior: Leveraging Lithology, Land Use, and Climate Gradients across the Colorado River, Texas USA

Grace M Goldrich-Middaugh[1], Lin Ma[2], Mark A Engle[3], Jason Ricketts[4], Paola Soto-Montero[4], and Pamela L Sullivan[1]

[1]Oregon State University
[2]University of Texas at El Paso
[3]University of Texas
[4]University of Texas, El Paso

November 22, 2022

Abstract

Understanding relationships between stream chemistry and watershed factors: land use/land cover, climate, and lithology are crucial to improving our knowledge of critical zone processes that influence water quality. We compiled major ion data from more than 100 monitoring stations collected over 60 years (1958-2018) across the Colorado River Watershed in Texas (103,000 km^2). We paired this river chemistry data with complementary lithology, land use, climate and stream discharge information. A combination of graphical geochemistry and machine learning techniques were used to produce new insights on controls of stream water chemical behavior. Studies on stream flow and chemistry in the American west and globally have shown strong relationships between major ion chemical composition and lithology, which hold true for the Colorado River basin in this study. Reactive minerals, including carbonates and evaporites, dominate major ion chemistry across the upper watershed. Upstream and central reaches of the Colorado River showed shifts from Na-Cl-SO_4 dominated water from multiple sources including dissolution of gypsum and halite in shallow groundwater, agricultural activities, and oil and gas development, to Ca-HCO_3 water types controlled by carbonate dissolution. In the lower portion of the watershed multiple analyses demonstrate that stream chemistry is more influenced by greater precipitation and the presence of relatively fewer reactive silicate minerals than middle and upstream reaches. This study demonstrates the power of applying machine learning approaches to publicly available long term water chemistry datasets to improve the understanding of water and nutrient cycling, salinity sources, and water use.

Regional Drivers of Stream Chemical Behavior: Leveraging Lithology, Land Use, and Climate Gradients across the Colorado River, Texas USA

Authors:

G. M. Goldrich-Middaugh[1], goldricg@oregonstate.edu, ORCid: 0000-0002-5899-8016

Ma, L.[2], email: lma@utep.edu, ORCid: 0000-0001-6240-2564

Engle, M.A.[2], email: maengle@utep.edu, ORCid: 0000-0001-5258-7374

Ricketts, J.W.[2], email: jricketts@utep.edu

Soto-Montero, P.[2], email: pisotomontero@miners.utep.edu, ORCid: 0000-0003-1047-5192

Sullivan, P. L.[1] , email: pamela.sullivan@oregonstate.edu, ORCid: 0000-0001-8780-8501

[1]Oregon State University, College of Earth, Ocean, and Atmospheric sciences.

[2]University of Texas at El Paso, Department of Earth, Environmental and Resource Sciences.

Corresponding author: Pamela Sullivan (pamela.sullivan@oregonstate.edu)

Key Points (140 characters)

- The distribution of reactive minerals (e.g., evaporites & carbonates) dictates the chemical behavior of the Colorado River
- Climate factors act as a secondary control, with increasing precipitation leading to decreased overall concentrations and dilution behavior
- Agriculture, urban development, reservoirs, and oil and gas wells lead to noise in chemical measurements and obscures natural signals

Abstract (250 words)

Understanding relationships between stream chemistry and watershed factors: land use/land cover, climate, and lithology are crucial to improving our knowledge of critical zone processes that influence water quality. We compiled major ion data from more than 100 monitoring stations collected over 60 years (1958-2018) across the Colorado River Watershed in Texas (103,000 km^2). We paired this river chemistry data with complementary lithology, land use, climate and stream discharge information. A combination of graphical geochemistry and machine learning techniques were used to produce new insights on controls of stream water chemical behavior. Studies on stream flow and chemistry in the American west and globally have shown strong relationships between major ion chemical composition and lithology, which hold true for the Colorado River basin in this study. Reactive minerals, including carbonates and evaporites, dominate major ion chemistry across the upper watershed. Upstream and central

reaches of the Colorado River showed shifts from Na-Cl-SO_4 dominated water from multiple sources including dissolution of gypsum and halite in shallow groundwater, agricultural activities, and oil and gas development, to Ca-HCO_3 water types controlled by carbonate dissolution. In the lower portion of the watershed multiple analyses demonstrate that stream chemistry is more influenced by greater precipitation and the presence of relatively fewer reactive silicate minerals than middle and upstream reaches. This study demonstrates the power of applying machine learning approaches to publicly available long term water chemistry datasets to improve the understanding of water and nutrient cycling, salinity sources, and water use.

Plain Language Summary (200 words)

Across the United States public and private users rely on large rivers for access to potable water, making water quality of crucial concern. Water quality measurements are widely available but require intensive pre-treatment due to their irregular collection across space and time. Here, public water quality measurements from the Colorado River basin were used to understand the influence of different land use, geologic, and climate factors on water quality. The Colorado River runs across a range of rock types, land uses, and precipitation regimes and therefore displays complex interactions with the land surface that produce changes in water quality. We found that the upper Colorado River is dominated by Na^+, Cl^-, and SO_4^{2-} which are derived from multiple sources including agriculture, oil and gas activity, and rock salt dissolution in shallow groundwater. The middle reaches of the Colorado are dominated by Ca^{2+} and HCO_3^- which are mainly contributed by the large areas of limestone bedrock. Downstream reaches of the river show more inputs from precipitation as well as potential seawater mixing in coastal areas. Overall, these techniques were effective in demonstrating large scale trends across this watershed and could be improved with more detailed datasets.

Keywords: critical zone, machine learning, historical data, water quality, regional hydrology

1. Introduction

River chemical composition and behavior reflects geochemical, hydrological, and anthropogenic factors of its upstream contributing areas (Ameli et al., 2017; Gaillardet et al., 1999; Godsey et al., 2009). Understanding how spatially varying watershed factors influence stream chemistry at large scales is crucial for illuminating how hydrologic processes influence global biogeochemical cycles. Particularly, synthetic analyses of coexisting river chemistry datasets with other datasets such as discharge, bedrock, soil types, and land uses are needed to improve understanding of how spatial heterogeneity in lithology, climate, and land use/land cover (LULC) impacts overall chemical behavior and transport of materials within large watersheds (i.e., critical zone function).

Water flow paths, residence times, and water-rock interactions within the critical zone govern weathering reactions, the rate of solute transport, and stream water chemical composition and behavior (Ameli et al., 2017; Baronas et al., 2017; Bouchez et al., 2017; Chen et al., 2014;

Chorover et al., 2017; Dupré et al., 2003; Gaillardet et al., 1999; Stewart et al., 2022; Torres et al., 2017). Flow paths and residence times are controlled by a combination of topography, lithology, climate, and anthropogenic factors. The degree of influence of each of these factors varies both spatially and temporally. Thus, predicting stream chemistry at any one point within a watershed is challenging. Many studies have attempted to capture the degree of influence of different watershed factors through concentration-discharge (C-Q) analysis at both event, and watershed scales (Abbott et al., 2018; Ameli et al., 2017; Evans & Davies, 1998; Godsey et al., 2009; Herndon et al., 2015; F. Liu et al., 2017; Minaudo et al., 2019; Torres et al., 2017). Significant research has also been conducted focusing on multivariate analyses that capture landscape influences of stream chemistry at a range of scales (Kaushal et al., 2013; X. Liu et al., 2021; Park & Lee, 2020; Sliva & Williams, 2001; Tiwari et al., 2017; Xu et al., 2021). More recently, machine learning techniques, such as random forest algorithms, have been employed to understand the degree to which topography, lithology, and climate govern stream discharge behavior (Addor et al., 2018; Hammond et al., 2021; Konapala & Mishra, 2020; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Oppel & Schumann, 2020; Rice et al., 2016), but to the best of our knowledge their use in explaining stream water chemistry has been limited (Nearing et al., 2021). Indeed, machine learning techniques offer a promising and flexible approach to quantify and represent relationships in complex systems and to identify important drivers across a range of environments.

Arising from many studies is the consensus that lithology is one of the most influential watershed factors controlling the composition and behavior of geogenic solutes in streams (Dupré et al., 2003; Gaillardet et al., 1999). Widespread homogeneity and availability of reactive minerals, in combination with long transit times relative to dissolution rates, have been proposed to account for chemostatic behavior across watersheds, meaning concentrations are relatively constant across a wide range of flows (Chorover et al., 2017; Godsey et al., 2009; Kirchner, 2003; Knapp et al., 2020; Maher, 2011). For example, studies in the Andes-Amazon basin have shown that in large river basins, longer residence times are associated with more chemostatic behavior, and that relationships between solute concentrations and flow change in response to spatial variations in watershed factors as well as contributions from major tributaries (Baronas et al., 2017; Bouchez et al., 2017; Torres et al., 2017). Chemical composition of river waters also generally reflects subsurface characteristics including lithology and geological structures (Stewart et al., 2022). However, large river systems have complex spatiotemporal interactions within the critical zone which makes quantitative models of direct lithologic (or structural) influence on stream chemical composition and behavior difficult to develop.

A second highly influential factor for stream solute behavior is climate (Shen et al., 2021; White & Blum, 1997). Precipitation and evapotranspiration (or the difference of the two, i.e., *effective precipitation*) dictate water availability and residence time in river channels and the shallow subsurface and hence control the depth distribution of reactive mineral fronts over long time periods (Ameli et al., 2017; Bouchez et al., 2017; Brantley & Lebedeva, 2020; Maher,

2011; Torres et al., 2017). Effective precipitation (defined as [precipitation – evapotranspiration] / precipitation) represents the fraction of precipitation that remains as available water. Negative values indicate actual evapotranspiration exceeds precipitation while positive values indicate that precipitation exceeds actual evapotranspiration and will infiltrate or run-off. Increased available precipitation leads to decreased water residence times, increased solute dilution behavior, and decreased coupling of river chemical behavior with lithologic influences (Maher, 2011; Torres et al., 2017). Hence, under low effective precipitation lithology and land use are more prominent factors influencing stream water chemistry compared to areas of high effective precipitation.

Finally, anthropogenic factors influence riverine chemistry and behavior through processes such as agricultural runoff, waste water discharge, regulation of flows by reservoirs, extraction of stream water, and alterations to hydrologic function via land use changes (Kaushal et al., 2013; Liu et al., 2021). Anthropogenic factors encompass myriad potential effects as they are broad and varied. For example, stream chemistry and discharge can be significantly different from natural conditions where urban and agricultural land represent a large fraction of watershed's area due to inputs of fertilizers, industrial additions, agricultural withdrawals, industrial and municipal withdrawals, regulation via reservoirs (Aitkenhead-Peterson et al., 2011; Chen et al., 2014; Musolff et al., 2015). Outside of direct impacts to flow regimes, studies have shown that special C-Q behaviors can occur in anthropogenically modified systems. Specifically, addition behavior (increase in concentration over large increase in discharge) has been observed for some nutrient-based solutes, including NO_3-N and PO_4 (Musolff et al., 2015). At a sub-watershed scale, agricultural areas have been shown to influence surface and shallow groundwater quality (Liu et al., 2021; Park & Lee, 2020). Understanding the magnitude of human impacts on chemical behavior in large watersheds that cross many land use classes remains elusive.

To improve our understanding of how spatially heterogenous catchment characteristics, including lithology, climate, and anthropogenic (e.g., LULC) factors, influence general stream chemical behavior at a regional scale, the objective of this research is to address our main research question: What are the most influential watershed factors governing variations in stream water chemistry in large rivers that cross a range of lithology, climate, and LULC types? More specifically, we apply a top-down approach to test the following three hypotheses through examination of coexisting water chemistry, discharge, lithology, climate, and LULC datasets in the Colorado River watershed, Texas:

Hypothesis 1: Lithology is the strongest control on stream chemistry where there is an abundance of reactive minerals in the near-surface. This setting leads to high concentrations of geogenic solutes, and chemostatic behavior.

Hypothesis 2: Climatic condition, measured as effective precipitation, governs flow paths, water residence times, and solute generation. In areas of high effective precipitation (P>>ET), low topographic relief, and/or high permeability, reactive mineral weathering fronts occur

at greater depths, water residence times are shorter, and systems are source limited and exhibit more dilution behavior.

Hypothesis 3: Anthropogenic influences obscure natural controls on stream water chemistry but the intensity and spatial spread of influence depends on the type and degree of development or contamination.

To test these hypotheses, we analyzed a 60 year-long (1958-2018), publicly available major ion chemical and discharge dataset using graphical techniques (Piper and Gibb's diagrams), C-Q relationships, geochemical modeling, and machine learning methods. We analyzed these data with spatially explicit land cover, climate, and lithology data from various public sources. We then used random forest algorithms to assess the importance of these factors in explaining the stream water chemistry. Overall, the data suggested that lithology was the first order control of stream water chemistry at regional scales, while climate exerted a secondary influence. Anthropogenic factors (particularly agricultural activities and oil and gas development) were found to be important for some geogenic species but their influence was difficult to separate from lithologic sources using only major ion chemical data.

2 Materials and Methods

2.1 Study area: Contrasting lithologic and land use patterns across the Colorado River Basin, Texas

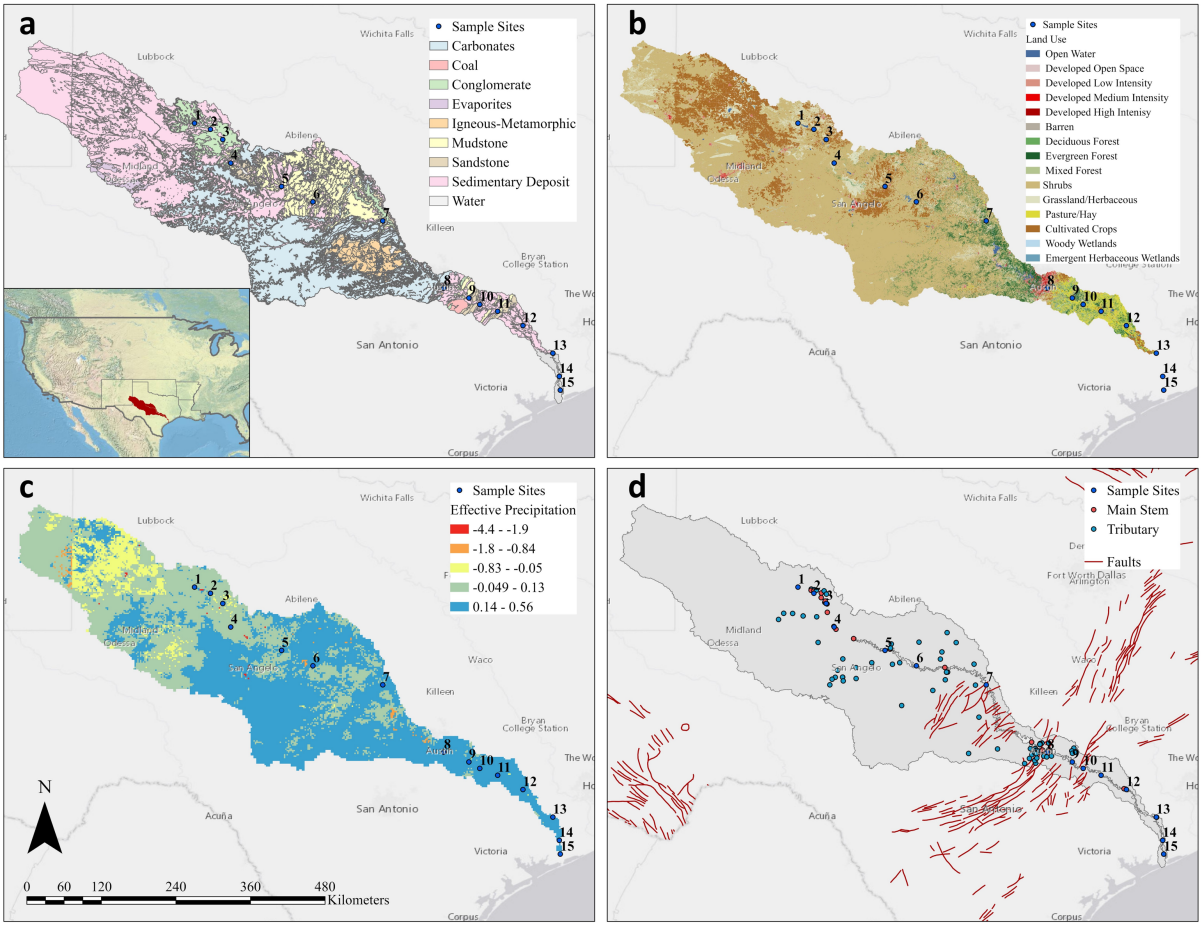

Figure 1 Study area: Colorado River a. simplified lithology, b. distribution of LULC classes with the Austin metropolitan area shown in red underlying sampling point 8 c. distribution of effective precipitation, and d. location of water chemistry sampling points on tributaries (blue) and main stem (red) with 15 flow sampling sites labelled (1-15) and faults shown in light red

We focus on the Colorado River of New Mexico and Texas because climate, LULC, and lithology changes across the watershed set the stage for spatially contrasting chemical behaviors. The river is approximately 1,288 km long, with 5 major tributaries and a watershed area that covers 15% of Texas (Colorado River Alliance, 2021, Texas Natural Resource Conservation Commission, 1999). It is also a crucial municipal water source for large and small communities in the state. Lithologic units across the watershed are roughly oriented NE to SW, perpendicular to flow in the main stem (Figure 1a). This systematic change in lithology and thus mineralogy across the watershed may control solute contributions to the Colorado River. Quaternary sedimentary deposits (conglomerates, sandstones, and mudstones), Cretaceous sedimentary carbonates, and gypsum and halite units are common rock types across the Colorado River basin (Clark et al., 2020). Karst features are common and host several large and productive aquifers

that intersect the Colorado River watershed near the Gulf Coast. The major structural feature affecting the Colorado River is the inactive late Oligocene to early Miocene Balcones fault zone, which is a network of NW-striking high-angle normal faults (Figure 1d) that predominantly dip southeast (Clark et al., 2020). The upper Colorado River is dominated by carbonate outcrops with gypsum inclusions (Leifeste & Lansford, 1968; Richter et al., 1991). The lower portions of the watershed are dominated by sedimentary deposits of varying lithologies. Overall, the lithologic shifts lead to general trends from more (evaporite and carbonates) to less (silicate minerals) reactive mineralogy moving downstream in the watershed. Thus, we expect that the upstream reaches of the Colorado River will be strongly influenced by the abundance of reactive mineral assemblages including evaporites and carbonates, the middle reaches will be influenced by a combination of reactive mineralogy (carbonates) and increased agricultural and urban influences, and the outlet of the river will be influenced dominantly by other factors as the lithology (silicate sediments) is more resistant to chemical weathering.

Spatial changes in climate across the Colorado River watershed strongly influence the hydrodynamics of the system, and thus can exert control on the chemical composition of the river. Similar to lithologic shifts perpendicular to the river course across the watershed, precipitation increases from NW to SE with the semi-arid western portion receiving an average of 250 mm of rain per year, and the sub-humid east receiving more than 1,500 mm (Harwell et al., 2020; Texas Natural Resource Conservation Commission, 1999; Texas Water Development Board, 2012). Across this area, the annual average temperature changes but not as significantly as the precipitation. Average annual evapotranspiration also increases in accordance with the increase in available water (NW to SE), except in human-modified areas where management of reservoirs and irrigated agriculture occur. A survey of trends in mean annual precipitation, air temperature, and streamflow shows moderate increases in precipitation in the eastern portion of the state and increases in air temperature of 0.6 ^{o}C per 50 years across the state leading to increased potential evapotranspiration (Harwell et al., 2020). Overall, the combination of precipitation and evapotranspiration leads to a spatial trend of increasing effective precipitation (-4.4 to 0.56), likely decreasing water residence times from northwest to southeast across the watershed (Figure 1c).

Anthropogenic influences are variable in distribution and impact across the watershed and encompass oil and gas development activities in the northwest and southeast, agriculture in the center and southeast, reservoir construction in the upper and middle reaches, and intensive urban development in the southeast, each of which can have distinctive controls on streamsolute load and behavior. The northwestern portions of the watershed are generally dominated by shrubs, grassland, and cultivated crops with the presence of diffuse, moderately developed areas. Evergreen and deciduous forest, high intensity development, pasture and hay roughly follow lithologic trends and appear near the Balcones fault zone (Figure 1b and d). Population gradients are present with many major cities centered in the southeast portion of the watershed including the Austin metropolitan area (population of 2.3 million) (US Census Bureau, 2019) which is shown in red around flow sampling site 8 (Figure 1c). Population growth over the next 50 years

is also projected at a much higher rate in the eastern half of the watershed than the western half (Scanlon et al., 2005; Texas Water Development Board, 2017). Agricultural production is central to the state and irrigation made up about 25.8% of total water use in the state in 2010 totaling 20.78 million m^3 per day (Dieter et al., 2018). Many dams are present in the watershed with six major reservoirs on the main stem of the Colorado River including E.V. Spence, O.H. Ivie, and J.B. Thomas. These reservoirs are dominantly for flood control, municipal, and agricultural water use. Trace metal concentrations along the river are also higher in areas of increased development (Chen et al., 2014). It is projected that over the next 50 years the population of the state will increase by 70% while the demand for water will increase by 17% and available water resources will decrease by 11% (Texas Water Development Board, 2017). These changes are likely to impact not only available water and flow regimes, but also the chemical composition of waters across the state.

2.2 Data collection and processing

2.2.1 Temporal Data

Daily flow measurements were extracted from the USGS NWIS database using R package dataRetrieval for the earliest recorded date to the current date (or the full period of record) whenever available (De Cicco et al., 2018). The Colorado River has 15 USGS-maintained flow measurement sites (Sites 1-15 on Figure 1), 14 of which (Sites 1-14) have daily flow measurements on the river main stem. In total, 360,670 flow measurements are available across the 14 sites between 1916 and 2020.

All water chemistry data were obtained using the water quality portal (WQP) R tool dataRetrieval (De Cicco et al., 2018). In the Colorado River basin, which includes the main stem and several major tributaries, we used a total of 108 sites with water chemistry data ranging from 1958 to 2018 for our analysis. The data were spatially and temporally irregular with, on average, at least one and at maximum 39 samples collected at each site per year. Here water chemistry measurements of interest included: alkalinity as HCO_3^-, K^+, Na^+, Mg^{2+}, P, SO_4^{2-}, Ca^{2+}, Cl^-, Si, NO_3^-, pH, and water temperature. The 108 chemical parameter sampling sites (Figure 1d) were grouped by proximity to the 15 USGS-flow sampling sites and labelled zones 1-15 (upstream to downstream) for further referencing continuity (Figure 1a-d). Measurements were filtered to include sites where all solutes were measured on the same date. 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)

2.2.2 Spatial data

Sub-watersheds were delineated using the NASA Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) digital elevation model (30 m resolution) accessed using the EarthData Portal. Watershed delineation was conducted using ArcGIS Pro 2.9 functions. Flow directions were generated using the D8 method in which flow is routed to one of 8 neighboring cells selected by calculating 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. Contributing areas for each sampling point were then delineated using the watershed tool. Next, each watershed factor dataset (lithology, LULC, and effective precipitation) was clipped to reflect the upstream contributing area for each of the sampling sites. A variety of spatial datasets were employed to examine factors influencing stream chemistry at each point within the watershed. Lithologic data is from the USGS via the Texas Water Development Board (TWDB) spatial data services portal. Rock types were merged to convey major classes (i.e., carbonates, sandstone, mudstone, etc.). Comprehensive land use data are from the National Land Cover Database (NLCD) for the most recent year released (2016). Annual evapotranspiration (2000-2013) at 800 m resolution is estimated using an empirical regression equation for long-term water balance ET data at 679 gaged watersheds as a function of land cover, precipitation, and daily temperature (Reitz et al., 2017). Both evapotranspiration and precipitation data (1971-2000) as annual average values (Crawford et al., 2006) were included to represent climate variations across the watershed. Lithology and land use classes were expressed as percent cover for each sub-basin. An areal weighted average of effective precipitation was calculated for the upstream contributing area of each sampling point.

2.3 Data Analyses

2.3.1 Piper diagrams and geochemical modeling of the chemical composition of the Colorado River

Piper diagrams were used to compare spatial trends in sample chemical composition to underlying lithology, land use, and climate metrics. Piper diagrams show the distribution of the relative proportion of major cations and anions in two trilinear diagrams, and one quadrilinear diagram which shows overall chemical composition (Piper, 1944).

To further understand the potential lithologic controls on stream solute concentrations, R package phreeqc (which interfaces the PHREEQC modeling code of Parkhurst & Appelo, 2013) was used to determine saturation indices (SI) of possible expected minerals comprised of major ions (gypsum, halite, aragonite, and chalcedony) for each charge balanced sample. pH measurements were also included and temperature was assumed to be constant at 25°C as actual stream temperature measurements were limited to 73% of observations (see S1). A sample was considered to be undersaturated when SI is < 0 (dissolution is likely), saturated when SI $= 0$ (showing that the system is at equilibrium), and supersaturated when SI > 0 (mineral precipitation is likely) (Drever, 1982). This analysis gives an idea of the degree of mineral dissolution contributing to stream chemical composition across lithologic and climate gradients. It also illustrates mineral presence not captured at the scale of USGS lithologic datasets, such as gypsum and halite. Further analysis using a larger number of measured solutes could increase our understanding of silicate dissolution in particular, as aluminosilicate minerals are likely the main contributors of silica rather than chalcedony. However, this dataset includes very limited measurements of Fe and Al to calculate the SI values of aluminosilicate minerals.

2.3.2 Concentration-discharge analysis

Concentration-discharge (C-Q) relationships were employed to examine spatial patterns in the behavior of individual solutes. Concentration discharge relationships can be related to watershed characteristics and commonly follow power-law relationships, as shown in equation 1 (Abbott et al., 2018; Baronas et al., 2017; Bouchez et al., 2017; Chorover et al., 2017; Godsey et al., 2009; Herndon et al., 2015; Moatar et al., 2017; Musolff et al., 2015; Sullivan et al., 2019):

$$C = aQ^b \quad (1)$$

where C is the concentration of a given constituent, Q is the discharge, and a and b are model parameters. When the measurements are log-transformed they can therefore be described by a linear model where a slope (b) of 0 indicates chemostatic behavior, a positive slope indicates addition behavior, and a slope of -1 indicates simple dilution behavior. Dilution occurs when the concentration of the stream decreases with increased flow, which occurs at sites where the baseflow of a river carries higher concentrations of each solute. Chemostatic behavior occurs where concentrations are invariant across a large range of flow values. The addition behavior occurs when concentrations increase systematically with increased flow. Concentration measurements from charge balanced samples were matched to flow measurements taken at the same site and date and converted to mmol/L for all solutes. Flow measurements were converted to L/s and the logarithm (base 10) of both concentration and discharge was calculated. The measurements were grouped by site and solute. Of 108 sites with charge balanced chemical measurements, 88 included flow measurements and 51 had at least 5 observations. Only sites with at least 5 measurements were included in this analysis. The mean number of observations per site was 73 and the maximum was 442. A linear model was fit for each site and solute using the "stats" package in R with the slope describing the concentration-discharge (C-Q) behavior (R Core Team, 2021).

2.3.3 Principal Components Analysis (PCA) of compositional stream chemical data

Multivariate techniques were implemented to expand on observations of individual solutes and to further refine sample groupings and relative influence of watershed factors. Specifically, PCA was used to group samples that were biogeochemically similar and to examine sample distribution relative to trends in lithology, LULC, and climate. To accomplish this goal, data from charge balanced samples were analyzed as compositional measurements. Compositional measurements represent the contribution of each solute to the overall concentration of the sample in mmol/L. Compositional measurements show change in solute concentrations relative to each other rather than outside factors, including dilution (Filzmoser et al., 2009; Filzmoser & Hron, 2008; Shelton et al., 2018). This means that further transformations must be applied to convert the variables into an unconstrained space. PCA weights eigenvectors by the square root of the corresponding eigenvalue to transform a large number of variables into orthogonal components that represent as much variation in the dataset as possible (Christophersen & Hooper, 1992; Jolliffe, 2002; Burke, 1997; Walter et al., 2019). In our PCA,

matrix Z was comprised of all compositional data Loadings of each variable on the retained PCs was tested for significance using the student's t-test, and the original data was reprojected using the PCs. The PCs were converted from isometric to centered log ratio format to improve interpretability. The latter log-ratio provides a one-to-one correspondence between input variables and log-ratios but produces a singular covariance matrix, which limits its use in PCA. PCs explaining up to 80% of cumulative variance in the dataset were considered significant for interpretation.

2.4 Machine learning for determination of watershed factor influence on stream chemistry

Multivariate models including spatial data were also employed to quantify the behavior of individual solutes in relation to changing watershed factors at a sub-watershed scale. This process allows for improved understanding of differences in behavior across solute groups in response to variation in lithologic, LULC, and climate distribution. A random forest approach was used to predict the proportion of each watershed factor using sample chemical composition as an input, and to quantify the strength of the relationship between these variables. Stream chemical data was chosen as an input variable rather than a predictive output variable in this investigation because of the greater variety in measurement values as compared to the summarized watershed factor data (multiple water samples were collected from each site).

Random forests are a supervised machine learning algorithm consisting of ensembles 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 (Addor et al., 2018). R package randomForest (R Core Team, 2021) was used for algorithm generation on 80% of the data and tested using the remaining 20% (Liaw & Wiener, 2002). The strength of the relationship between each solute and watershed factor was quantified using the percent increase in mean squared error (%IncMSE) if the predictor variable (solute) was removed from consideration. A high %IncMSE value indicates a stronger relationship between the two variables. This gives a relative metric for comparison of the influence of each factor (lithology, LULC, and climate) on the behavior of solutes across the watershed (Addor et al., 2018; Hammond et al., 2021; Konapala & Mishra, 2020; Oppel & Schumann, 2020). Relationships between all variables were also represented using the Spearman's rank correlation coefficient.

3 Results

3.1 Piper diagrams show shift from upstream Na-Cl dominated to downstream Ca-HCO_3 dominated samples.

Surface water chemistry shifts from Na-Cl dominated to Ca-HCO_3 dominated from northwest to southeast across the watershed (Figure 2). From zone 2 to zone 6, relative

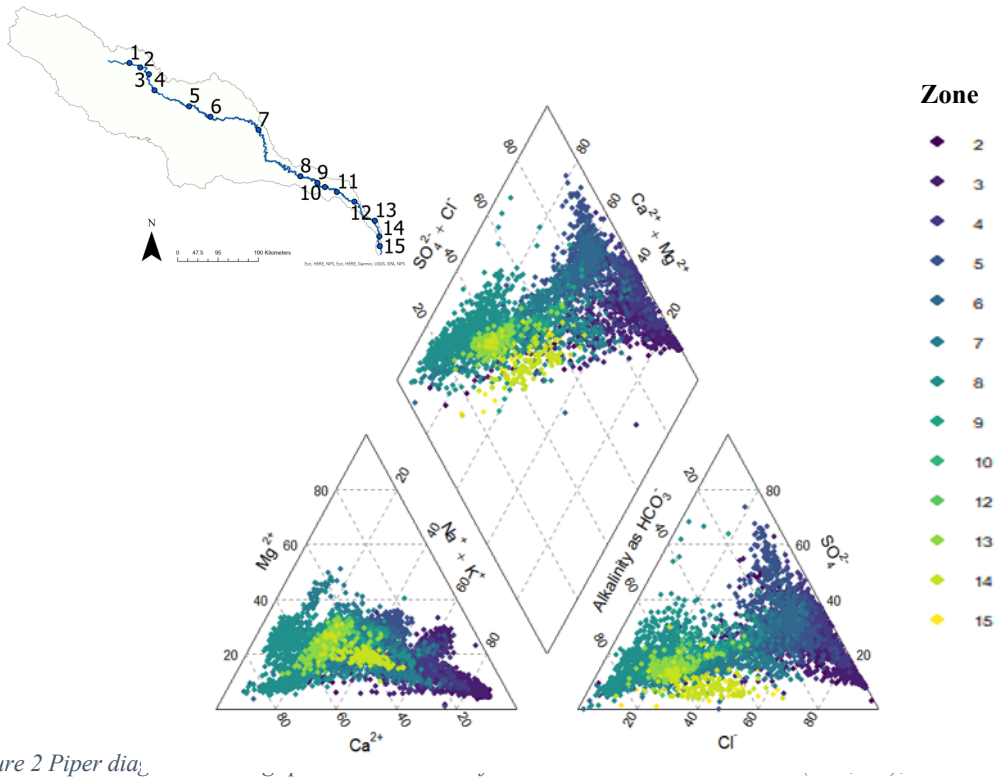

Figure 2 Piper dia[illegible] d represent the 14 zones along the Colorado river (inset map upper left). Darker blue are sites located upstream while lighter yellow indicate sites located near the coast.

concentration of Na^++K^+ and Cl^- are very high. Additionally, the relative concentration of SO_4^{2-} is highly variable. Relative proportions of Ca^{2+} and Mg^{2+}, increase from zones 2 to 6 as Cl^- decreases corresponding to an increase in HCO_3^-. Measurements from zones 7-14 generally show the highest relative proportion of Ca^{2+} and HCO_3^-. However, after zone 8 there is a slight decrease in Ca^{2+} and a corresponding increase in Na^++K^+ and Cl^- (Figure 2).

3.2 Geochemical modeling to refine conceptual model of landscape factor influences

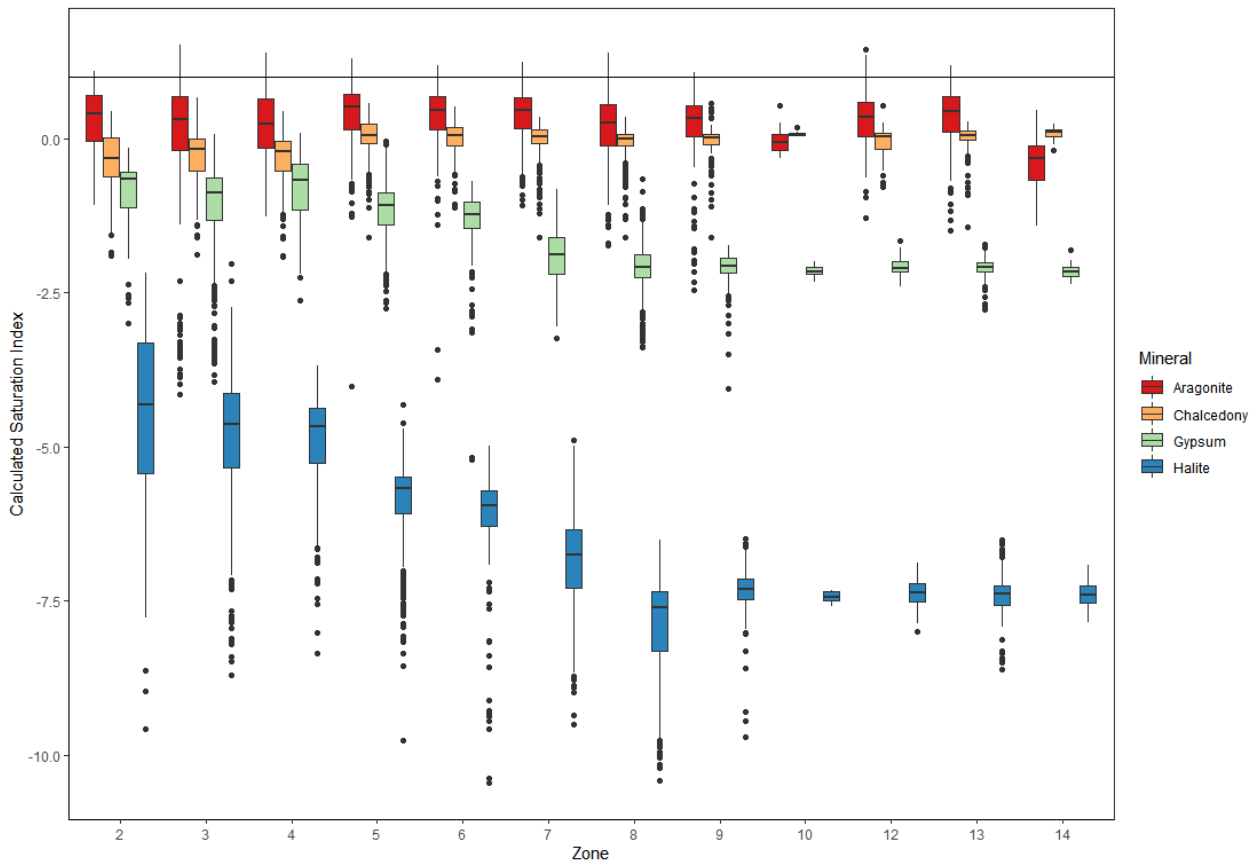

Figure 3 Box plots of the calculated saturation index across the 14 river zones for Aragonite (red), Chalcedony (orange), Gypsum (green), and Halite (blue). Box plots illustrating the median and inter quartile range as well as outlying points for calculated SI in each zone. A saturation index >0 indicates the potential for mineral precipitation, while a value <0 indicates the potential for mineral dissolution.

Saturation indices of aragonite ($CaCO_3$), chalcedony (SiO_2 used here to evaluate silicate minerals as SI of aluminosilicates was not available), gypsum ($CaSO_4$), and halite (NaCl) were calculated using PHREEQC software with major ion chemistry and pH measurements at a standard temperature of 25ºC (Figure 3; see S1 for a subset of the data where field temperatures are available, overall trends remain consistent). Carbonate and silicate species are near saturated or slightly oversaturated while sulfate and evaporite species are dominantly undersaturated. Saturation indices for aragonite and chalcedony are relatively consistent from upstream to downstream zones. In contrast, gypsum and halite saturation indices are highest in the upstream portions of the river and decrease until zone 9 then remain relatively constant.

3.3 Concentration-Discharge relationships reveal overarching spatial trends

Slopes (parameter b) of the linear regression of log(concentration) vs log(discharge) were calculated for each site with at least 4 observations and are generally within error but still demonstrate trends in space (Figure 4). In summary, the Colorado River demonstrates 4 trends in slope of the C-Q relationships. The first trend includes HCO_3^- and Ca^{2+} which have slopes close

to 0 (chemostatic behavior) for zones 1-6 followed by increasing dilution behavior (slopes close to -0.5) from zones 7-14. Trend 2 represents the majority of the geogenic solutes Cl^-, Mg^{2+}, SO_4^{2-} and Na^+ have negative slopes (~-0.5; dilution behavior) in the northwest zones, with slopes trending towards 0 in zones 6-7 and returning to dilution behavior for zones 8-14 (slopes <-0.5). Of these solutes Ca^{2+} is more chemostatic than the others in zone 2 with slopes between 0 and -0.25. Trend 3 includes K^+ and Si which are very close to chemostatic across all zones. Trend 4 shows NO_3^- and P which are more highly variable than other solutes and have fewer

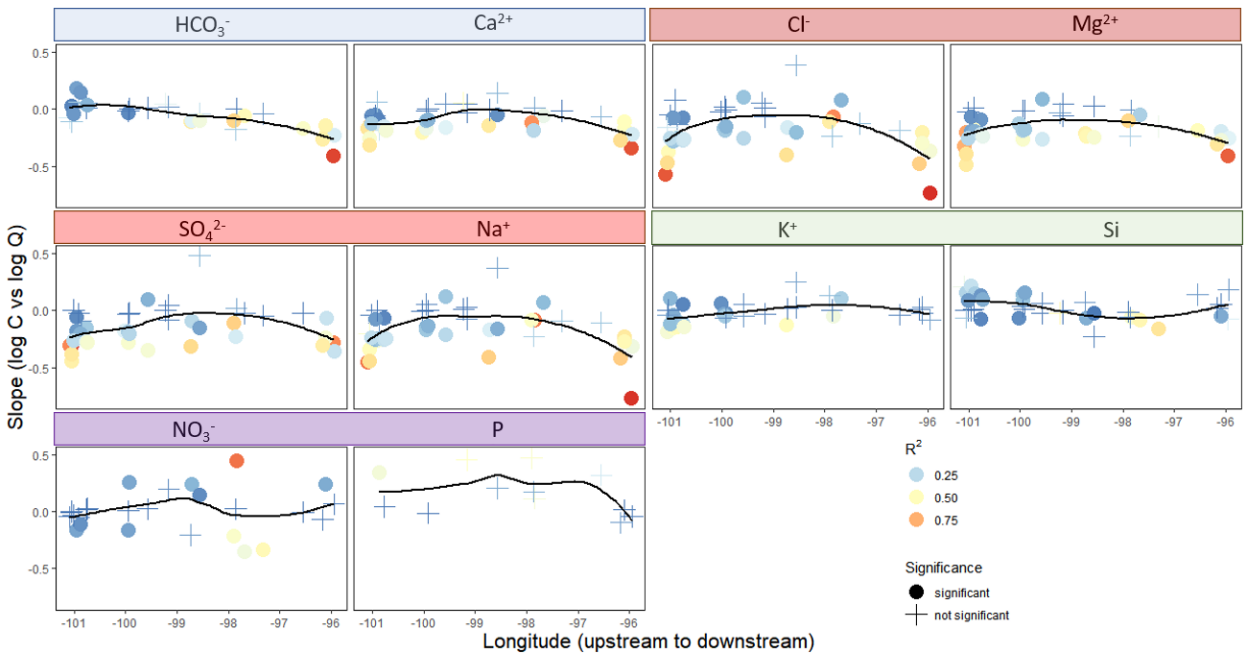

Figure 4 Slope of linear regression of log (C) versus log(Q) across the Colorado River Basin. Each point represents the slope calculated for measurements at a single site which are organized from upstream to downstream (left to right). Point shape indicates wether the model coefficient is significantly different from 0 at alpha=0.05. Point color indicates the R^2 of the regression. Black lines indicate the smoothed conditional mean slope to illustrate overall trends for each solute. Title color indicates trends 1-4 with trend 1 in blue, trend 2 in red, trend 3 in green, and trend 4 in purple.

measurements included for analysis. NO_3^- shows some addition behavior in zones 4-6 with one outlying site in zone 8 showing significant addition behavior with a slope of almost +0.5.

3.4 Compositional PCA mirrors piper diagram trends from Na-Cl to Ca-HCO_3

PCA was conducted on all chemical measurements collected within the Colorado River Watershed for a total of 4,863 points. The first two principal components explain 88.7% of the variation in the dataset (PC1 = 76.4% and PC2 = 12.3 %) and the third explains an additional 9.76%. The primary loadings of interest for the first principal component are Na^+, Cl^-, and SO_4^{2-}

which are negative, and HCO_3^-,and Si which are positive (Table 1). The primary loadings of interest for the second principal component are K^+ (positive) and Mg^{2+} (negative).

Table 1 Loadings from compositional PCA for all measurements (n=4,863) in the Colorado River Watershed.

	PC1	PC2	PC3
Bicarbonate	0.448	-0.176	0.485
Calcium	0.146	-0.262	0.208
Chloride	-0.461	0.142	-0.168
Potassium	0.149	0.668	0.338
Magnesium	-0.015	-0.537	-0.116
Sodium	-0.442	0.222	0.039
Sulfate	-0.319	-0.247	-0.038
Silica	0.496	0.190	-0.750

Examination of the first 2 principal components reveals a pattern which orders the samples from upstream (left) to downstream (right and down) across PC1. Corresponding changes in the ions are also shown, indicating evolution from a Na-Cl dominated composition to a Ca-HCO_3 dominated one (Figure 5). Higher relative abundance of Si on the centered log ratio biplots (Figure 5) generally indicates lower salinity composition due to the dilution of the other elements relative to Si, whose concentration is relatively constant (Engle et al., 2011). Sites 6-8 are widely distributed and include many tributary measurements. Sites 9-15 include fewer measurements from each zone, are dominantly collected on the main stem, and are more closely clustered. The outlying cluster of 30 points on PC2 is from 2 USGS sites 08120700 and 08121000 charge balanced samples collected in 1966 and 1967 and represents anomalously high values for K^+. A total of 274 observations from site 08120700 collected from 1965-2002, and 298 measurements collected from site 08121000 from 1963-2003 are included and measurements from other years are not outliers. Analysis of the spread of points in each zone based on season

and tributary position revealed no significant trends supporting the idea that this long-term dataset is more reflective of spatial variations than temporal trends (S1).

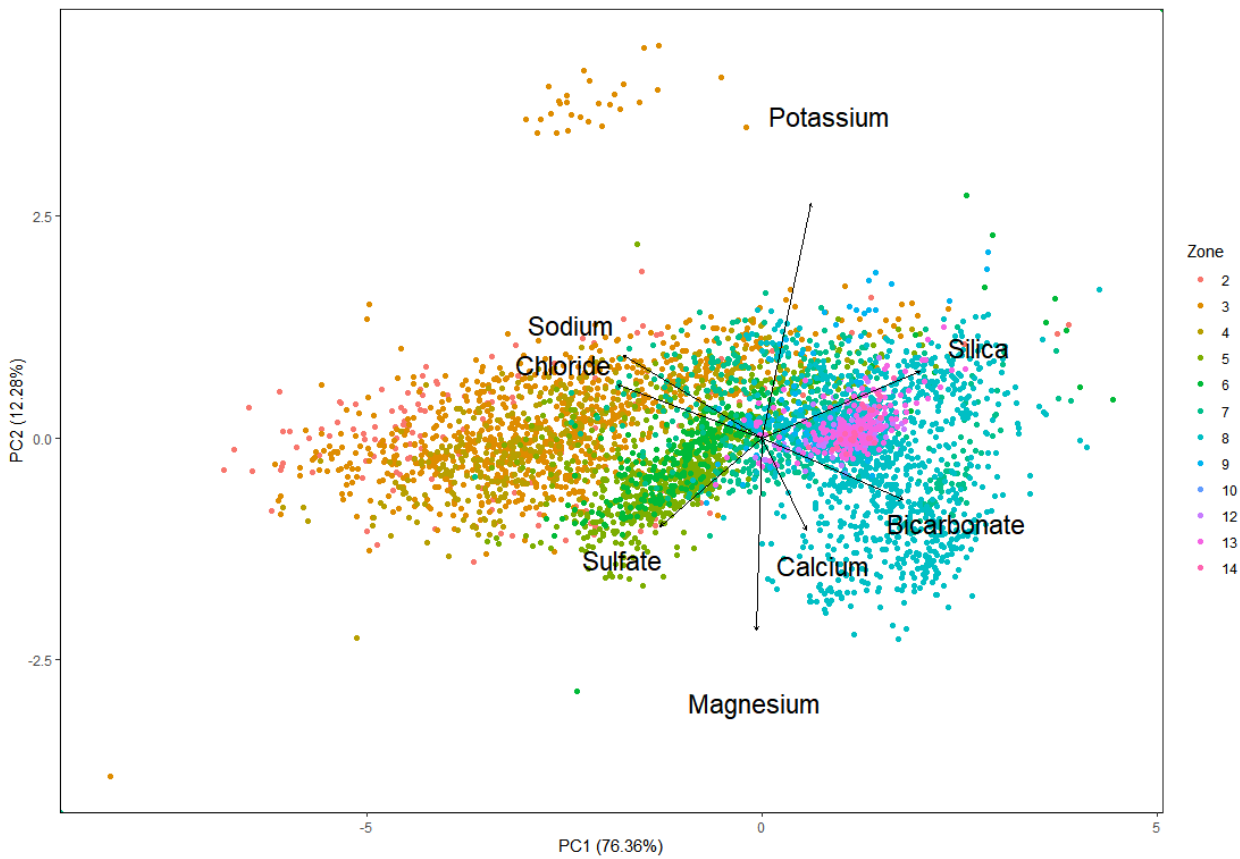

Figure 5 Colorado river water chemistry projected onto first two principal components shows spatial variations, points are colored by proximity to sampling sites 2-14.

3.5 Random forest analysis gives insight for LULC influences on stream water composition

The random forest regression algorithms were used to assess relationships between solutes and watershed factors through comparison of percent increase in mean squared error (%IncMSE) for each solute/watershed factor combination (Figure 6). Here we focused on three groups of watershed factors: lithology, LULC, and climate. Of all LULC classes, cultivated crops were the strongest predictor across all solutes, while sedimentary deposits and evaporites had the strongest relationships to solute behavior in terms of lithology classes. Specifically, the results show a strong relationship between cultivated crops and K^+, Mg^{2+}, and SO_4^{2-} and to a lesser degree Cl^-, Ca^{2+}, and HCO_3^-. Cultivated crops are also positively correlated with Na^+ and Cl^- but negatively correlated with Mg^{2+}, Ca^{2+}, and HCO_3^-. There are large areas covered by cultivated crops in the upper portions of the Colorado River Watershed and smaller stretches covered by pasture near zones 9-14. Additionally, there appears to be a strong relationship between open areas (including barren, grass, and shrubland) with Na^+, Mg^{2+}, Ca^{2+}, and to a lesser

extent Cl^- and HCO_3^-. Open areas are also most prevalent in zones 2-7. Correlation coefficients are smaller than for agriculture but are positive for Mg^{2+}, Ca^{2+}, and HCO_3^- and negative for Na^+ and Cl^- (opposite to agriculture). Forest and developed areas were both shown to have generally smaller %incMSE with most of the solutes.

Lithologic influences are strongest for sedimentary deposits, including the conglomerate, sandstone, carbonate, and evaporite classes. Sedimentary deposits and carbonates show the highest correlation coefficients. Sedimentary deposits have the highest %IncMSE for Ca^{2+} and Mg^{2+} with additional high %IncMSE for HCO_3^-, Na^+, and Cl^-. Conglomerate has the highest %IncMSE for Na^+ and Mg^{2+} with additional large values for HCO_3^- and Cl^-.

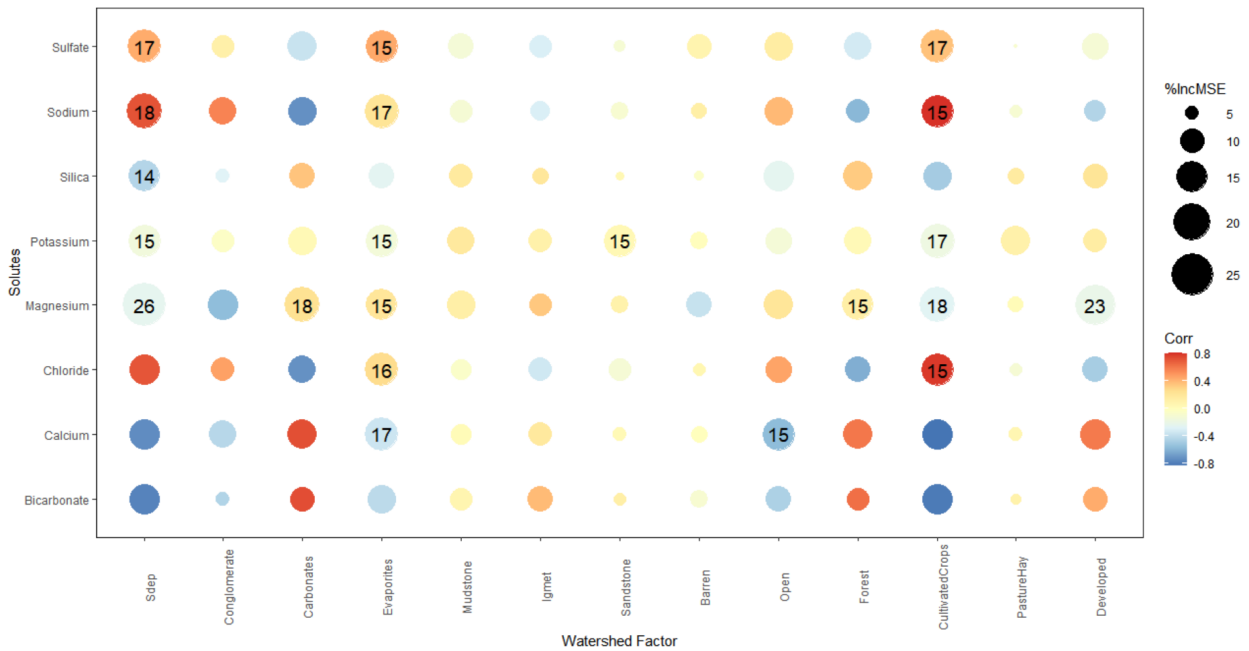

Figure 6 Plot of solutes versus lithology and LULC factors with point size showing %incMSE and color indicating the spearman's correlation coefficient, numeric labels show %incMSE value for top 80th percentile %incMSE scores

4 Discussion

Long term river chemistry data sets, especially in combination with spatial data, have been used at a range of scales to reveal underlying processes that control chemical behavior (Godsey et al., 2009; Park & Lee, 2020; Tiwari et al., 2017; Torres et al., 2017). Yet, the use of publicly available historical stream chemical measurements has been limited because of temporal and spatial irregularity. Here, we used PCA, C-Q relationships, geochemical modeling, and supervised machine learning (random forest) analyses to investigate the influence of watershed factors in the Texas Colorado River Basin. Our analyses reveal three main regions of behavior: the upper (zones 2-4), middle (5-8), and lower (9-15) Colorado River. From this analysis we find that abundant reactive mineralogy, saline shallow groundwater, agricultural and oilfield activities dominate chemical signatures in low-development, low-precipitation, and low-flow reaches of the upstream Colorado River watershed. Conversely, the downstream reaches are increasingly

controlled by climate factors rather than lithology or anthropogenic signatures. Below these inferences are described in detail.

4.1 Evaporites and brines exert a strong influence on river chemistry in the upper reaches (zones 2-4) of the Colorado River while carbonates dominate middle reaches (zones 5-8)

Several lines of evidence show that most of the spatiotemporal variation of solute chemistry in the upper Colorado basin is controlled by the presence of easily weatherable evaporite and carbonate minerals with shifts downstream from zone 2 to 8 in water chemistry related to changes in this lithologic distribution. First, the spread and separation of measurements across piper diagrams (from Na-Cl and Ca–SO_4 to Ca-HCO_3 water types; Figure 2) and PCA (loadings of -6 to 0 across PC1 where Na^+, Cl^-, and HCO_3^- loaded significantly; Figure 5, Table 1) illustrates the control of evaporites (gypsum and halite) and brines across this upstream portion of the Colorado River. Spatially, the proportion of evaporites is high in zones 2-5 as compared to downstream zones while carbonates are ubiquitous across zones 2-8 (Figure 1a), which contributes to the shift from Na-Cl and Ca-SO_4 to Ca-HCO_3 water types shown by piper diagrams and PCA. Geochemical modeling of the saturation state of the stream water also suggests water is near equilibrium with carbonate minerals throughout the upper extent of the Colorado but is consistently undersaturated with respect to evaporite minerals. SI values for evaporite minerals are highest in zones 2-3 and consistently decrease until zone 7 (Figure 3). One interpretation of the decline in SI values for halite and gypsum is that evaporite mineral dissolution acts as a solute source while these minerals consistently decline in proportion moving downstream.

This strong lithologic control on stream water chemistry is echoed by other analyses of historic water chemistry in the upper portions of the Colorado River (Scanlon et al., 2005; Slade & Buszka, 1994; Slade et al., 2002), which show that a combination of anthropogenic and natural factors leads to increased salinity in the upstream reaches of the Colorado River. Particularly, abundance of near-surface evaporites and upwelling of brines in this region have been shown to influence shallow groundwater composition and discharge which occurs laterally and through salt-springs to the Colorado River. Analysis of water types using piper diagrams, major ion ratios, and stable isotope analysis of surface and shallow groundwater samples all support these conclusions (Dutton et al., 1989; Leifeste & Lansford, 1968; Reed, 1961). Gibbs diagrams show extremely high TDS, Na^+, and Cl^-, particularly in the upstream reaches of the Colorado River (zones 2-5), which suggest evaporation dominated systems (including

agricultural influences), brines, and halite/gypsum dissolution, but do not allow for distinction between these mechanisms (Figure 7).

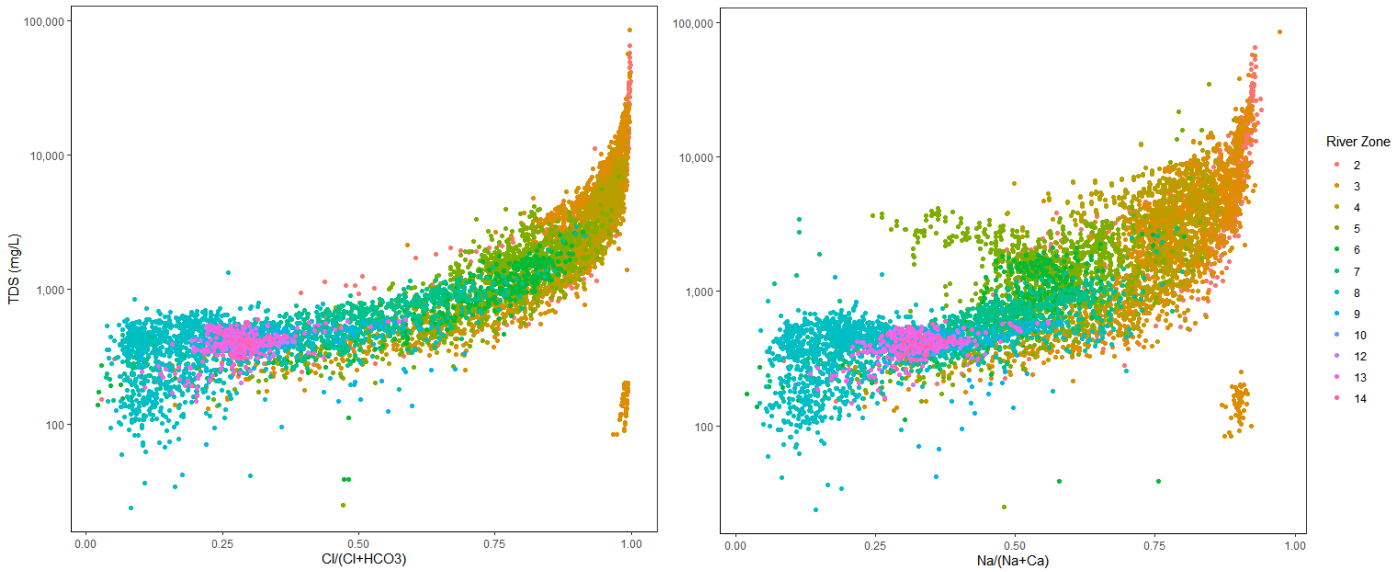

Figure 7 Gibb's diagrams showing the distribution of calculated total dissolved solids compared to Cl, Na, Ca, and HCO_3 concentrations across the Colorado River. Evaporation dominated areas are characterized by high TDS, Cl, and Na. Rock weathering dominated areas are characterized by moderate TDS (100-1,000 mg/L) and low Cl and Na.

Along with lithologic controls in this area, a combination of potential leakage from oil and gas development sites, naturally upwelling brines, and enrichment of solutes in shallow groundwater contribute to high overall concentrations of Na^+, Cl^-, and SO_4^{2-}, dilution behavior of evaporite related solutes, and undersaturated conditions with respect to halite and gypsum (Dutton et al., 1989; Leifeste & Lansford, 1968; Paine et al., 1999; Richter et al., 1991; Slade & Buszka, 1994). Particularly, analysis of shallow groundwater and river water major ion and isotopic chemistry as compared to brines in local formations has shown significant similarity in composition between surface waters and brines, with other areas showing more direct influence from evaporite dissolution (Richter & Kreitler, 1986).

Concentration discharge behavior analyses also provide evidence of a strong lithologic control on solute behavior but suggest important differences in the spatial heterogeneity of the lithology. Specifically, dilution behavior occurs in zones 2-4 for in Na^+, Cl^-, and SO_4^{2-} (slopes of -0.5 to -0.1), while HCO_3^- and Ca^{2+} exhibit chemostatic behavior across these zones (Figure 4). This difference in C-Q behaviors across the upper Colorado River suggests that evaporites in zones 2-4 are distributed more heterogeneously in space, limiting solute sources and leading to the observed dilution behavior while the abundance of carbonates supports chemostatic behavior. Previous studies have shown that heterogeneity in lithologic distribution in combination with hydrologic factors (activation of different flow pathways and water sources during precipitation) can lead to non-chemostatic behavior due to mixing of different source waters (Baronas et al., 2017; Torres et al., 2017).

Hydrogeologic conditions in the semi-arid upper Colorado basin are conducive to dominant lithologic influences on stream chemistry. Low precipitation and high potential evapotranspiration can reduce pore flushing and lead to a greater degree of water-rock interaction and elevated concentrations of geogenic solutes in subsurface water (Brantley & Lebedeva, 2020; Stewart et al., 2022; White & Blum, 1997). Additionally, this region's shallow groundwater bodies are subject to elevated solute concentrations as a result of evapotranspiration (Dutton et al., 1989; Paine et al., 1999; Richter et al., 1991). Finally, the depth to reactive minerals may be shallower in areas where effective precipitation is lower (Ameli et al., 2017; Kim et al., 2017).

4.2 Lower reaches (zones 9-14) of the Colorado River are more dominated by climate and sedimentary deposits

As effective precipitation increases downstream along the Colorado River watershed, studies suggest the increase in available water will yield faster transit times (shorter residence time) that support more dilution behavior (Brantley et al., 2017; Chorover et al., 2017; Maher, 2011). All geogenic solutes (such as HCO_3^-, Ca^{2+}, Cl^-, Na^+, SO_4^{2-}, and Mg^{2+}) included here reflect this climatic shift with increasing dilution behavior (decreasing slope b values towards -1) beginning in zone 8 and continuing until the river outlet (zone 15). In the upper Colorado River watershed HCO_3^- was highly chemostatic until zones 7 and 8 where the slopes declined showing increased dilution. Evaporite and brine related solutes (such as Cl^-, Na^+, and SO_4^{2-}) show the same trends toward dilution in these zones. In zone 8 the river flow increases markedly, the local watershed receives more precipitation, and the Colorado river passes the last of the major reservoirs. Here, long-term weathering resulting from greater precipitation may have led to deepening of reactive mineral fronts (Brantley et al., 2017; Brantley & Lebedeva, 2020). This, along with increased contributions of meteoric water, directly via precipitation, and indirectly through shallow groundwater contributions to streamflow and decreased water residence time in the shallow subsurface lead to slightly increased dilution behavior. This is consistent with previously developed C-Q and reactive transport conceptual models which show strong dependence on thermodynamic and hydrologic controls by stream chemical behavior (Brantley & Lebedeva, 2020; Maher, 2011; Wlostowski et al., 2021).

As the watershed moves through areas of increased effective precipitation, chemical weathering occurs at a faster rate leading to clustering of points on PC1 and in the Piper diagram, changes in calculated SI for evaporites, and increased chemostatic behavior of Na^+, Cl^-, and SO_4^{2-} (Figures 2-5). Variations in C-Q relationships and PCA results are consistent with previous findings which suggest that thermodynamic (reaction rates of underlying lithology) and hydrologic (flow and climatic) controls exert strong influence on stream chemical composition and behavior except where this influence is obscured by contributions of different source waters, climate factors, and anthropogenic influences which all increase moving downstream in this watershed (Baronas et al., 2017; Bouchez et al., 2017; Knapp et al., 2020; Maher, 2011; Torres et al., 2017). These findings show that, for evaporite and carbonate related solutes in this

watershed, lithology is the dominant influencing factor until increased contributions from precipitation and urban development (zone 8) shift river chemistry away from lithologic signals.

4.3 Possible anthropogenic inputs in the upper Colorado watershed

While lithology and climate dominantly control river chemical composition across the Colorado River, anthropogenic factors, particularly agricultural activities and oil and gas development, are also influential in both the upstream and central reaches of the watershed. Oil and gas development has been posited to increase surface water salinity via a variety of mechanisms. Orphaned and improperly abandoned oil and gas wells and poorly managed disposal of produced waters allows for inputs of brines in the shallow subsurface (Dutton et al., 1989). Additionally, salts accumulate below historic brine disposal pits and are leached into shallow groundwater or laterally moved into surface waters (Dutton et al., 1989; Richter et al., 1991). While it is possible that this occurs on the Colorado River it is difficult to distinguish between the impacts of oil and gas activities (either historical or recent) versus natural contributions like evaporite dissolution, concentrated shallow groundwater, natural upwelling of deep brines, or other human activities including agriculture. Gibb's diagrams and analysis of major ion ratios give some support to brine influence (Figure 7) but further analyses (such as U, Sr, and B isotope tracers) need to be conducted using a greater number of chemical metrics to determine the degree and extent of oil and gas versus other salinity sources across this region (Engle et al., 2011; Osborn et al., 2011).

Agricultural land (cultivated crops) is most abundant in zones 2-6 of the Colorado River basin and may contribute to increased salinity due to agricultural amendments and increased groundwater use. Groundwater use for irrigation is particularly common in this region. The strong relationship between Na^+ and agricultural land is shown by the random forest analysis (Figure 6) where there is a large %IncMSE for Na^+, K^+, Cl^-, and Mg^{2+} all components of subsurface brines and agricultural amendments. A potential mechanism for this is the ET-based accumulation and following dissolution and percolation of naturally occurring salts accumulated under agricultural irrigation areas (Kondash et al., 2020; Yurtseven et al., 2018). Further, agricultural irrigation may lead to increased soil salinity over time due to the use of saline groundwater.

Urban areas, including the developed region around the city of Austin, impact both flow and water quality (Aitkenhead-Peterson et al., 2011). Urban regulation of discharge and influence of water chemistry in combination with increased precipitation obscures the underlying lithologic influences and cause clustering of measurements on principal component 1 (Chen et al., 2014; Kaushal et al., 2013; Raymond et al., 2008). Particularly in large cities (i.e., Austin) significant municipal water needs require importing water from multiple sources. Discharges of this water for various uses (irrigation, treated wastewater, road runoff) may be of entirely different composition than the naturally occurring water sources (Aitkenhead-Peterson et al., 2011). However, these impacts are not always well-captured through analysis of major ion chemistry. Analyses including stable isotope ratios (H, O, and B) and/or nutrients that are

common indicators of anthropogenic activity (e.g., NO_3^-, NH_4^+, PO_4^{3-}) could better capture the influence of urban development on stream chemical composition (Lyon et al., 2008; Raymond et al., 2008). While reservoirs are likely impacting the system, they do not appear to be significantly impacting the major ion composition. Reservoir storage has also been shown to be a weak predictor of flow metrics across varying environments (Hammond et al., 2021). Differing LULC types including oil and gas acreage, urban areas, and reservoirs alter natural signals along the Colorado River. The influence of brines associated with oil and gas production activities was discussed earlier (see 4.1 and 4.3) as their chemical signature with the available data is hard to distinguish from evaporite dissolution or naturally upwelling brines (Leifeste & Lansford, 1968). Land use and intensive urban development are the next most influential factors, as illustrated by the shift in data distribution occurring after zone 8 in the PCA biplot, contributing to the spread of measurements occurring in the downstream zones of the Colorado River. From zone 8 (Austin area) to the river outlet the position of measurements on the PCA biplot displays a scattered rather than linear trend. This again suggests that intensive development, while it does not display a dominant signature in major ion chemical composition of surface waters, tends to add variability in other hydrologic signals.

Along the central reaches of the Colorado River, influences from tributaries, reservoirs, and differing LULC become increasingly dominant, shifting C-Q behavior away from dilution and towards chemostatic and leading to data clustering on the PCA biplot. Some elements may even display addition behavior such as P and NO_3^- that can be linked to LULC and human impacts. Major tributaries dilute elements transported from upstream and contribute deep groundwater derived carbonates drawn from underlying lithology on both the main stem and tributaries of the Colorado River in connection with the nearby Llano uplift (Bruun et al., 2016; Leifeste & Lansford, 1968; Slade et al., 2002). This is shown through increases in relative concentration of HCO_3^- and chemostatic behavior of C-Q slopes for all solutes in zones 5-8. In the Amazon basin, contributions from relatively homogeneous water sources from major tributaries were also shown to produce chemostatic behavior of major ions (Baronas et al., 2017; Torres et al., 2017). Similar to the behavior of HCO_3^- and Ca^{2+} in the upstream zones of the Colorado River, the influence of geologic features (including the Llano uplift) near major tributaries serves to increase the availability of reactive carbonate minerals through connections to deep groundwater and increases chemostatic behavior and homogeneity of chemical composition.

4.4 Broader applications and limitations due to reliance on major ion concentrations

The analyses presented here rely on publicly available major ion chemical data collected from over 100 sites in the Colorado River watershed, collected at irregular spatial and temporal intervals. This valuable resource allows for analysis of many influences of stream water quality across lithology, LULC, and climate gradients but does not provide sufficient information to distinguish among some of these influences with great detail. With additional major, trace element, and isotope data (e.g., δ18O, δD, δ11B, $^{87}Sr/^{86}Sr$ and $^{234}U/^{238}U$), other relationships

could be examined and more details could be distinguished concerning existing processes and influences. Salinity in the Colorado River watershed encompasses much of the overall chemical composition and is derived from multiple sources including deep brines, shallow groundwater, evaporite dissolution, and agricultural influences (Leifeste & Lansford, 1968; Paine et al., 1999; Richter et al., 1991). Further measurements of stable isotopes, age-dating methods, and bromide could be used in zones 2-4 of the Colorado river to better define the degree of each of these influences and their variation in space.

The simplified lithology classes analyzed here provided valuable information about the degree of influence of various reactive minerals on stream chemistry. Strong relationships were evident in the upper Colorado River watershed between evaporites (halite, gypsum), carbonates (aragonite, dolomite), and stream water chemistry. Further studies could leverage the information and analytical techniques presented here and include additional measurements of trace metals (e.g., Fe, Al, etc.) with more detailed lithology maps to better understand the distribution of different lithologies, mineralogic influences on stream chemistry, and sources of solutes across the watershed. These analyses can also focus on the potential structural controls of stream chemistry from possible groundwater inputs from the Balcones fault zone, which crosses the Colorado River at zone 8. Finally, the application of random forest algorithms provided useful insight to watershed factor influences on stream chemical composition and behavior in this catchment. Future work could include larger-scale public datasets in combination with a variety of non-linear data analysis techniques to refine and generalize understanding of influences of stream water chemistry across diverse environments.

5 Conclusions

Over 5,280 publicly available water quality observations were paired with daily discharge measurements and high resolution geospatial (e.g., lithology and land cover) data from the Colorado River (103,000 km^2) over a 60 year period (1958-2018) to determine the degree of influence of key watershed factors on stream chemical behavior. Comparison of trends in chemical composition using Piper and Gibb's diagrams and compositional PCA set a baseline for interpretation of changing contributions of evaporite and carbonates vs. silicate minerals across the watershed. Further analyses including C-Q relationships and their changes in space showed that not all solutes behaved similarly at the watershed scale and C-Q behavior changes could be attributed to watershed processes such as shallow groundwater contributions, long term precipitation influences, and the distribution of reactive mineralogy. Finally, random forest analysis provided a metric for interpretation of the relative influence of watershed factors on each solute and showed strong relationships between sedimentary units, contribution of carbonates, and Na^+, Cl^-, Ca^{2+}, Mg^{2+}, and SO_4^{2-}. Further, human factors such as LULC types including cultivated crops, grass, and shrublands were shown to have strong relationships with

various solutes across the watershed while developed areas and forestland did not appear to have a similarly strong relationship in this watershed.

Overall, this research highlighted the importance of reactive mineral abundance and dissolution, the effect of climate (precipitation), and the impacts of multiple sources of salinity in the Colorado River watershed. Broader implications are that regional, historical major ion chemical datasets at the watershed scale can be leveraged to improve understanding of lithologic and climate impacts on stream water chemistry at relatively large spatial and temporal scales. Further analyses should be employed to better-determine the processes occurring at key locations with shorter time scales ("hotspots and hot moments") using a broader and combined set of chemical, climate, lithologic, and land use parameters. Additionally, these techniques could be applied to examine processes controlling water chemical composition and behavior at larger scales leveraging national databases.

Acknowledgments, Samples, and Data

We acknowledge funding support from the National Science Foundation (EAR-1933261; EAR-1933259). 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 contributions Sean Fleming, Nuria Andreu Garcia, and Jennifer Herrera to project design and analysis techniques.

Stream discharge data can be accessed from the USGS National Water Information System (https://waterdata.usgs.gov/nwis/rt), stream water quality data can be accessed from the USGS Water Quality Portal (https://www.waterqualitydata.us/), spatial data can be accessed via the TCEQ Spatial Database (https://gis-tceq.opendata.arcgis.com/) and Multi-resolution land characteristics consortium (https://www.mrlc.gov/).

References

Abbott, B. W., Gruau, G., Zarnetske, J. P., Moatar, F., Barbe, L., Thomas, Z., Fovet, O., Kolbe, T., Gu, S., Pierson-Wickmann, A. C., Davy, P., & Pinay, G. (2018). Unexpected spatial stability of water chemistry in headwater stream networks. *Ecology Letters*, *21*(2), 296–308. https://doi.org/10.1111/ele.12897

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 Resources Research*, *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

Ameli, A., Beven, K., Erlandsson, M., Creed, I. F., McDonnell, J. J., & Bishop, K. (2017). Primary weathering rates, water transit times, and concentration-discharge relations: A theoretical analysis for the critical zone. *Water Resources Research*. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

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

Bouchez, J., Moquet, J. S., Espinoza, J. C., Martinez, J. M., Guyot, J. L., Lagane, C., Filizola, N., Noriega, L., Hidalgo Sanchez, L., & Pombosa, R. (2017). River Mixing in the Amazon as a Driver of Concentration-Discharge Relationships. *Water Resources Research*, *53*(11), 8660–8685. https://doi.org/10.1002/2017WR020591

Brantley, S. L., & Lebedeva, M. I. (2020). Relating land surface, water table, and weathering fronts with a conveptual valve model for headwater catchments. *Hydrological Processes*.

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

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

Chen, J.-B., Gaillardet, J., Bouchez, J., Louvat, P., & Wang, Y.-N. (2014). Anthropophile elements in river sediments: Overview from the Seine River, France. *Geochemistry, Geophysics, Geosystems*, 4526–4546. https://doi.org/10.1002/2014GC005516.Received

Chorover, J., Derry, L. A., & McDowell, W. H. (2017). Concentration-Discharge Relations in the Critical Zone: Implications for Resolving Critical Zone Structure, Function, and Evolution. *Water Resources Research*, *53*(11), 8654–8659. https://doi.org/10.1002/2017WR021111

Christophersen, N., & Hooper, R. P. (1992). Multivariate analysis of stream water chemical data: The use of principal components analysis for the end-member mixing problem. *Water Resources Research*, *28*(1), 99–107. https://doi.org/10.1029/91WR02518

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

Crawford, C., Hamilton, P., & Hoos, A. (2006). *National Water-Quality Assessment Program - Modifications to the status and trends network and assessments of streams and rivers*. US Geological Survey. https://water.usgs.gov/GIS/metadata/usgswrd/XML/nhd_ppt30yr.xml#stdorder

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.

Dieter, C. A., Maupin, M. A., Caldwell, R. R., Harris, M. A., Ivahnenko, T. I., Lovelace, J. K., Barber, N. L., & Linsey, K. S. (2018). Estimated Use of Water in the United States in 2015. In *US Geological Survey Circular* (Issue 1441). https://doi.org/10.3133/cir1441

Drever, J. (1982). *The Geochemistry of Natural Waters*. Prentice-Hall.

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

Dutton, A. R., Richter, B. C., & Kreitler, C. (1989). Brine Discharge and Salinization , Concho River Watershed , West Texas. *Ground Water*. https://doi.org/10.1111/j.1745-6584.1989.tb00461.x

Engle, M. A., Bern, C. R., Healy, R. W., Sams, J. I., Zupancic, J. W., & Schroeder, K. T. (2011). Tracking solutes and water from subsurface drip irrigation application of coalbed methane-produced waters, Powder River Basin, Wyoming. *Environmental Geosciences*, *18*(3), 169–187. https://doi.org/10.1306/eg.03031111004

Evans, C., & Davies, T. D. (1998). Causes of concentration/discharge hysteresis and its potential as a tool for analysis of episode hydrochemistry. *Water Resources Research*, *34*(1), 129–137. https://doi.org/10.1029/97WR01881

Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, *40*(3), 233–248. https://doi.org/10.1007/s11004-007-9141-5

Filzmoser, P., Hron, K., & Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, *20*(6), 621–632. https://doi.org/10.1002/env.966

Gaillardet, J., Dupre, B., Louvat, P., & Allegre, C. J. (1999). Global silicate weathering and CO2 consumption rates deduced from the chemistry of large rivers. *Chemical Geology*, *159*, 3–30.

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

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. *Hydrogeology Journal*, *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. *Geophysical Research Letters*, *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*. http://pubs.er.usgs.gov/publication/sir20195137

Herndon, E. M., Dere, A. L., Sullivan, P. L., Norris, D., Reynolds, B., & Brantley, S. L. (2015). Landscape heterogeneity drives contrasting concentration-discharge relationships in shale headwater catchments. *Hydrology and Earth System Sciences*, *19*(8), 3333–3347. https://doi.org/10.5194/hess-19-3333-2015

Jolliffe, I. T. (2002). Principal components analysis. In *Springer Series in Statistics*.

https://doi.org/10.1016/B978-0-08-044894-7.01358-0

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. *Environmental Science and Technology*, *47*(18), 10302–10311. https://doi.org/10.1021/es401046s

Kim, H., Dietrich, W. E., Thurnhoffer, B. M., Bishop, J. K. B., & Fung, I. Y. (2017). Controls on solute concentration-discharge relationships revealed by simultaneous hydrochemistry observations of hillslope runoff and stream flow: The importance of critical zone structure. *Water Resources Research*, 1424–1443. https://doi.org/10.1002/2016WR019722.Received

Kirchner, J. W. (2003). A double paradox in catchment hydrology and geochemistry. *Hydrological Processes*, *17*(4), 871–874. https://doi.org/10.1002/hyp.5108

Knapp, J. L. A., Von Freyberg, J., Studer, B., Kiewiet, L., & Kirchner, J. W. (2020). Concentration-discharge relationships vary among hydrological events, reflecting differences in event characteristics. *Hydrology and Earth System Sciences*, *24*(5), 2561–2576. https://doi.org/10.5194/hess-24-2561-2020

Konapala, G., & Mishra, A. (2020). Quantifying Climate and Catchment Control on Hydrological Drought in the Continental United States. *Water Resources Research*, *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. *Science of the Total Environment*, *733*, 139392. https://doi.org/10.1016/j.scitotenv.2020.139392

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Leifeste, D. K., & Lansford, M. W. (1968). *Reconnaissance of the chemical quality of surface waters of the Colorado River Basin, Texas* (Issue March).

Liaw, A., & Wiener, M. (2002). *Classification and regression by random forest* (pp. 2(3), 18–22).

Liu, F., Conklin, W. H., & Shaw, G. . (2017). Insights into hydrologic and hydrochemical processes based on concentration-discharge and end-member mixing analyses in the mid-Merced River Basin, Sierra Nevada, California. *Journal of the American Water Resources Association*. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Liu, X., Zhang, Y., Li, Z., Li, P., Xu, G., Cheng, Y., & Zhang, T. (2021). Response of water quality to land use in hydrologic response unit and riparian buffer along the Dan River, China. *Environmental Science and Pollution Research*, *28*(22), 28251–28262. https://doi.org/10.1007/s11356-021-12636-z

Lyon, S. W., Seibert, J., Lembo, A. J., Steenhuis, T. S., & Walter, M. T. (2008). Incorporating landscape characteristics in a distance metric for interpolating between observations of stream water chemistry. *Hydrology and Earth System Sciences*, *12*(5), 1229–1239. https://doi.org/10.5194/hess-12-1229-2008

Maher, K. (2011). The role of fluid residence time and topographic scales in determining chemical fluxes from landscapes. *Earth and Planetary Science Letters*, *312*(1–2), 48–58. https://doi.org/10.1016/j.epsl.2011.09.040

Minaudo, C., Dupas, R., Gascuel-odoux, C., Roubeix, V., Danis, P., & Moatar, F. (2019). Seasonal and event-based concentration-discharge relationships to identify catchment controls on nutrient export regimes. *Advances in Water Resources*, *131*(July 2018), 103379. https://doi.org/10.1016/j.advwatres.2019.103379

Moatar, F., Abbott, B. W., Minaudo, C., Curie, F., & Pinay, G. (2017). Elemental properties, hydrology, and biology interact to shape concentration-discharge curves for carbon, nutrients, sediment, and major ions. *Journal of the American Water Resources Association*. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Musolff, A., Schmidt, C., Selle, B., & Fleckenstein, J. H. (2015). Catchment controls on solute export. *Advances in Water Resources*, *86*, 133–146. https://doi.org/10.1016/j.advwatres.2015.09.026

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 Resources Research*, *57*(3). https://doi.org/10.1029/2020WR028091

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

Osborn, S. G., Vengosh, A., Warner, N. R., & Jackson, R. B. (2011). Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(20), 8172–8176. https://doi.org/10.1073/pnas.1100682108

Paine, J. G., Dutton, A. R., & Blum, M. U. (1999). *Using Airborne Geophysics to Identify Salinization in West Texas*. https://doi.org/10.23867/RI0257D

Park, S. R., & Lee, S. W. (2020). Spatially varying and scale-dependent relationships of land use types with stream water quality. *International Journal of Environmental Research and Public Health*, *17*(5). https://doi.org/10.3390/ijerph17051673

Parkhurst, D. L., & Appelo, C. A. J. (2013). *Description of input and examples for PHREEQC version 3 - A computer program for speciation and batch-reaction, one-dimensional transport, and inverse geochemical calculations: U.S. Geological Survey Techniques and Methods, book 6, chap. A43*. U.S. Geological Survey.

Piper, A. M. (1944). A graphic procedure in the geochemical interpretation of water-analyses. *Eos Transactions, American Geophysical Union*, 914–928.

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. *International*

Journal of Geographical Information Science.

Rao, A., Burke, T. (1997). Principal Component Analysis of Hydrologic Data. In *Integrated Approach to Environmental Systems* (pp. 275–290). Springer, Dordrecht.

Raymond, P. A., Oh, N. H., Turner, R. E., & Broussard, W. (2008). Anthropogenically enhanced fluxes of water and carbon from the Mississippi River. *Nature*, *451*(7177), 449–452. https://doi.org/10.1038/nature06505

Reed, E. L. (1961). *A study of the salt water pollution of the Colorado River, Scurry and Mitchell Counties, Texas: Consultant's report to Colorado River Municipal Water District, Big Spring, Texas*.

Reitz, M., Sanford, W. ., Senay, G. B., & Cazenas, J. (2017). *Annual Estimates of recharge, quick-flow runoff and ET for the contiguous US using empirical regression equations*. US Geological Survey Data Release. https://www.sciencebase.gov/catalog/item/55d3730fe4b0518e35468e1e

Rice, J. S., Emanuel, R. E., & Vose, J. M. (2016). The influence of watershed characteristics on spatial patterns of trends in annual scale streamflow variability in the continental U.S. *Journal of Hydrology*, *540*, 850–860. https://doi.org/10.1016/j.jhydrol.2016.07.006

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

Scanlon, B. R., Tachovsky, J. A., Keese, K., Slade, R. M., Howard, M. T., Wells, G. L., & Mullins, G. J. (2005). *Groundwater - Surface water interactions: Texas* (Issue August).

Shelton, J. L., Engle, M. A., Buccianti, A., & Blondes, M. S. (2018). The isometric log-ratio (ilr)-ion plot: A proposed alternative to the Piper diagram. *Journal of Geochemical Exploration*, *190*(September 2017), 130–141. https://doi.org/10.1016/j.gexplo.2018.03.003

Shen, B., Wu, J., Zhan, S., Jin, M., Saparov, A. S., & Abuduwaili, J. (2021). Spatial variations and controls on the hydrochemistry of surface waters across the Ili-Balkhash Basin, arid Central Asia. *Journal of Hydrology*, *600*(May), 126565. https://doi.org/10.1016/j.jhydrol.2021.126565

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*.

Slade, R. M. J., Bentley, J. T., & Michaud, D. (2002). Results of stream-flow gain-loss studies in Texas, with emphasis on gains from and losses to major and minor aquifers, Texas, 2000. *U.S. Geol. Surv. Open File Rep., 02-068*, 136.

Sliva, L., & Williams, D. D. (2001). Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Water Research*, *35*(14), 3462–3472. https://doi.org/10.1016/S0043-1354(01)00062-8

Stewart, B., Shanley, J. B., Kirchner, J. W., Norris, D., Adler, 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 Subsurface Water Chemistry From Stream Chemistry Water

Resources Research. *Water Resources Research*, 1–20. https://doi.org/10.1029/2021WR029931

Sullivan, P. L., Stops, M. W., Macpherson, G. L., Li, L., Hirmas, D. R., & Dodds, W. K. (2019). How landscape heterogeneity governs stream water concentration-discharge behavior in carbonate terrains (Konza Prairie, USA). *Chemical Geology*, *527*(April). https://doi.org/10.1016/j.chemgeo.2018.12.002

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

Texas Natural Resource Conservation Commission. (1999). *Surface Water/Groundwater Interaction Evaluation for 22 Texas River Basins*. http://www.twdb.texas.gov/publications/reports/contracted_reports/doc/Surface-Groundwater_Interaction.pdf

Texas Water Development Board. (2012). *2012 State Water Plan*. 1–314. http://www.twdb.texas.gov/publications/state_water_plan/2012/2012_SWP.pdf

Texas Water Development Board. (2017). *2017 State Water Plan Water for Texas*. www.twdb.texas.gov

Tiwari, T., Lidman, F., Laudon, H., Lidberg, W., & Agren, A. (2017). GIS-based prediction of stream chemistry using landscape composition, wet aeras, and hydrological flow pathways. *Journal of Geophysical Research: Biogeosciences*.

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

Walter, J., Chesnaux, R., Gaboury, D., & Cloutier, V. (2019). Subsampling of regional-scale database for improving multivariate analysis interpretation of groundwater chemical evolution and ion sources. *Geosciences (Switzerland)*, *9*(3). https://doi.org/10.3390/geosciences9030139

White, A. F., & Blum, A. E. (1997). Effects of climate on chemical weathering in watersheds. *Geochimica et Cosmochimica Acta*, *59*(9), 1729–1747. https://doi.org/https://doi.org/10.1016/0016-7037(95)00078-E.

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 Resources Research*, *57*(3). https://doi.org/10.1029/2019WR026635

Xu, H., Cai, C., Du, H., & Guo, Y. (2021). Responses of water quality to land use in riparian buffers: a case study of Huangpu River, China. *GeoJournal*, *86*(4), 1657–1669. https://doi.org/10.1007/s10708-020-10150-2

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

Supplemental Information for

Regional Drivers of Stream Chemical Behavior: Leveraging Lithology, Land Use, and Climate Gradients across the Colorado River, Texas USA

G. M. Goldrich-Middaugh[1]

Ma, L.[2]

Engle, M.A.[2]

Soto-Montero, P.[2]

Ricketts, J.W.[2]

Sullivan, P. L.[1]

[1]Oregon State University, College of Earth, Ocean, and Atmospheric sciences.

[2]University of Texas at El Paso, Department of Earth, Environmental, and Resource Sciences.

Contents of this file

Text S1

Site selection and Data

Data for all WQP sites within the Colorado River basin (defined as HUC codes "120800", "120901", "120902", "120903", "120904") was downloaded and included 490 sites and 27,7433 observations. Charge balance was calculated and only sites with a charge balance error of less than 10% were retained. This included 155 sites, and 8,066 observations with a mean absolute charge balance error of 1.32%. These observations were used for concentration-discharge analysis as measurements were assumed to be relatively representative and it was not necessary to include observations of every solute at each site.

Samples were further subset to include dates and locations where Ca^{2+}, Cl^{-}, K^{+}, Na^{+}, HCO_3^{-}, SO_4^{2-}, Si, and Mg^{2+} were all measured (termed complete cases). This subset included 117 sites and 4,686 observations. These observations were used for piper diagrams, PCA, and random forest analysis because it was necessary to include observations for each variable at every site and date.

Analysis of seasonal, tributary, and temperature variations

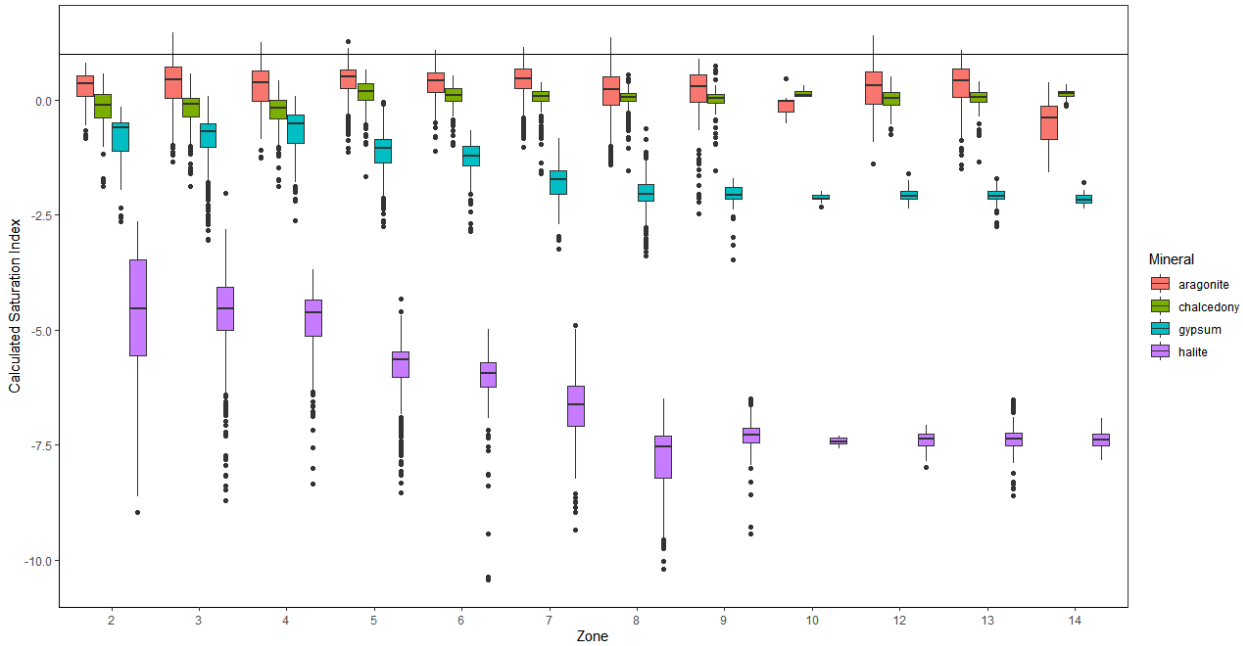

Figure S1 Box plots of the calculated saturation index across the 14 river zones for Aragonite (red), Chalcedony (green), Gypsum (blue), and Halite (purple). Box plots illustrating the median and inter quartile range as well as outlying points for calculated SI in each zone for observations with temperature measurements. A saturation index >0 indicates the potential for mineral precipitation, while a value <0 indicates the potential for mineral dissolution.

Temperature measurements are available for 73% of all complete, charge balanced samples. The mean measured temperature is 19.27^{o}C. Trends in calculated SI by zone for the four minerals of interest are extremely similar when using a subset of observations with temperature measurements as compared to a constant temperature of 25°C.

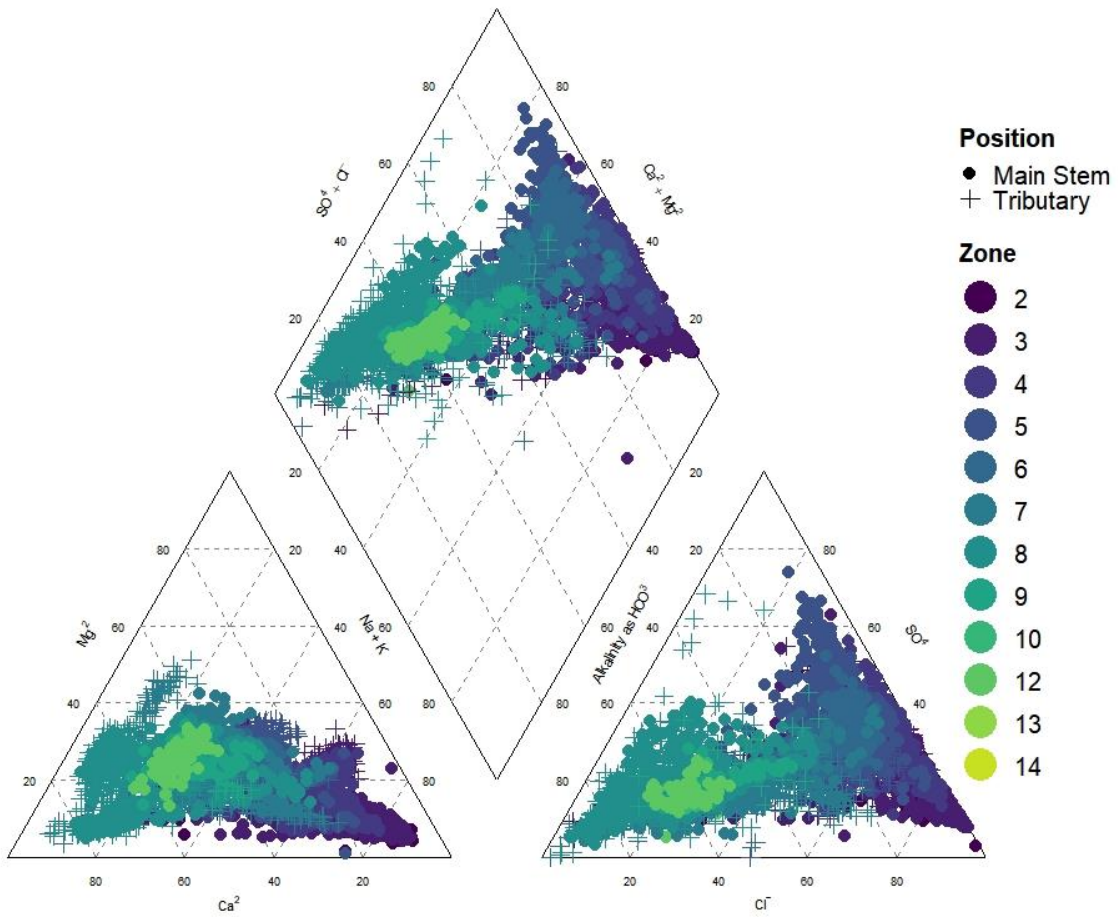

Figure S2 Piper diagram from the Colorado River showing observations based on river zone (2-14) and watershed position (tributary or main stem)

Distinguishing between tributary and non-tributary observations on the piper diagram (Figure S2) shows that overall variations are controlled by watershed position (zone) while some localized variations are influenced by tributary position. For example, samples collected on tributaries in zones 3 and 4 are somewhat higher in Mg^{2+} than samples collected from the main stem in these locations. Additionally, samples collected in tributaries within zone 8 are generally higher in Ca^{2+} and Cl^{-} than those collected from the main stem.

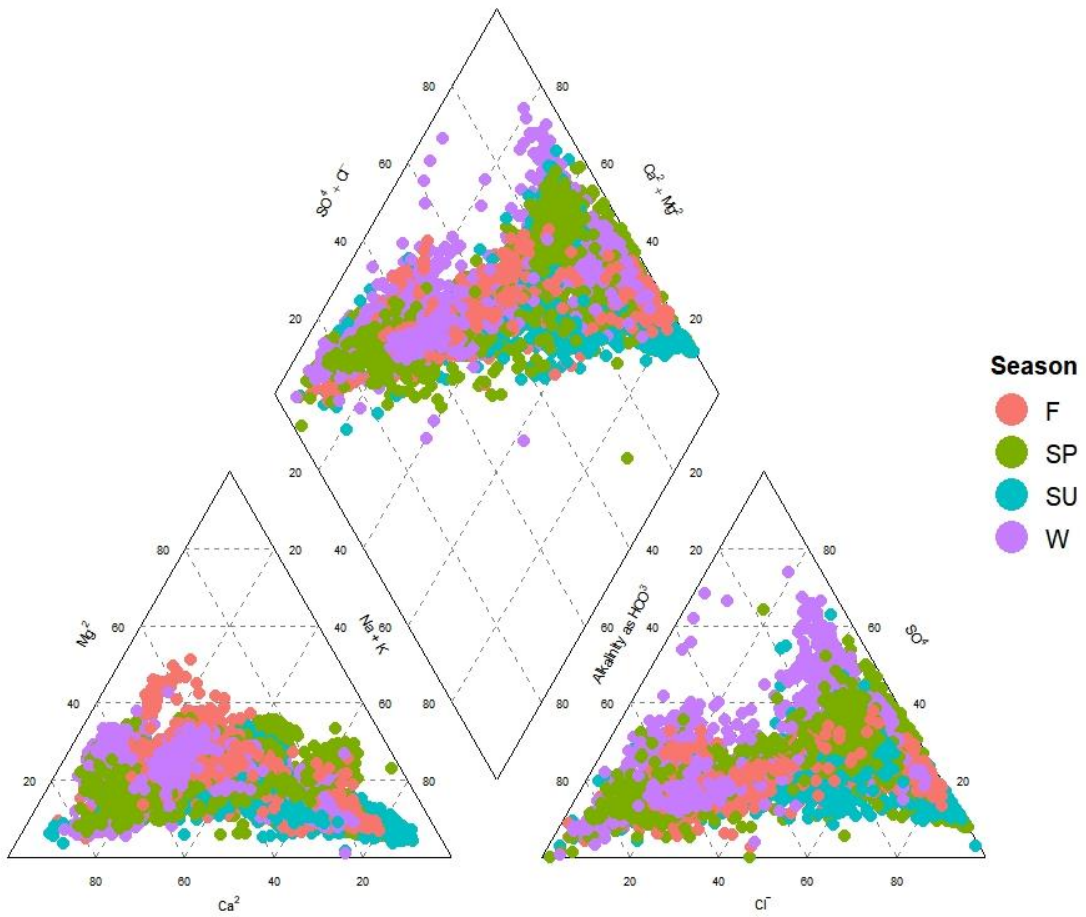

Figure S3 Piper diagram from the Colorado River showing observations colored by season (F=Fall (September, October, November), SP=Spring (March, April, May), SU=Summer (June, July, August), W=Winter (December, January, February))

The composition of stream water across seasons (Figure S3) did not show overarching variations in composition by season which suggests that variations are dominantly dictated by watershed position.

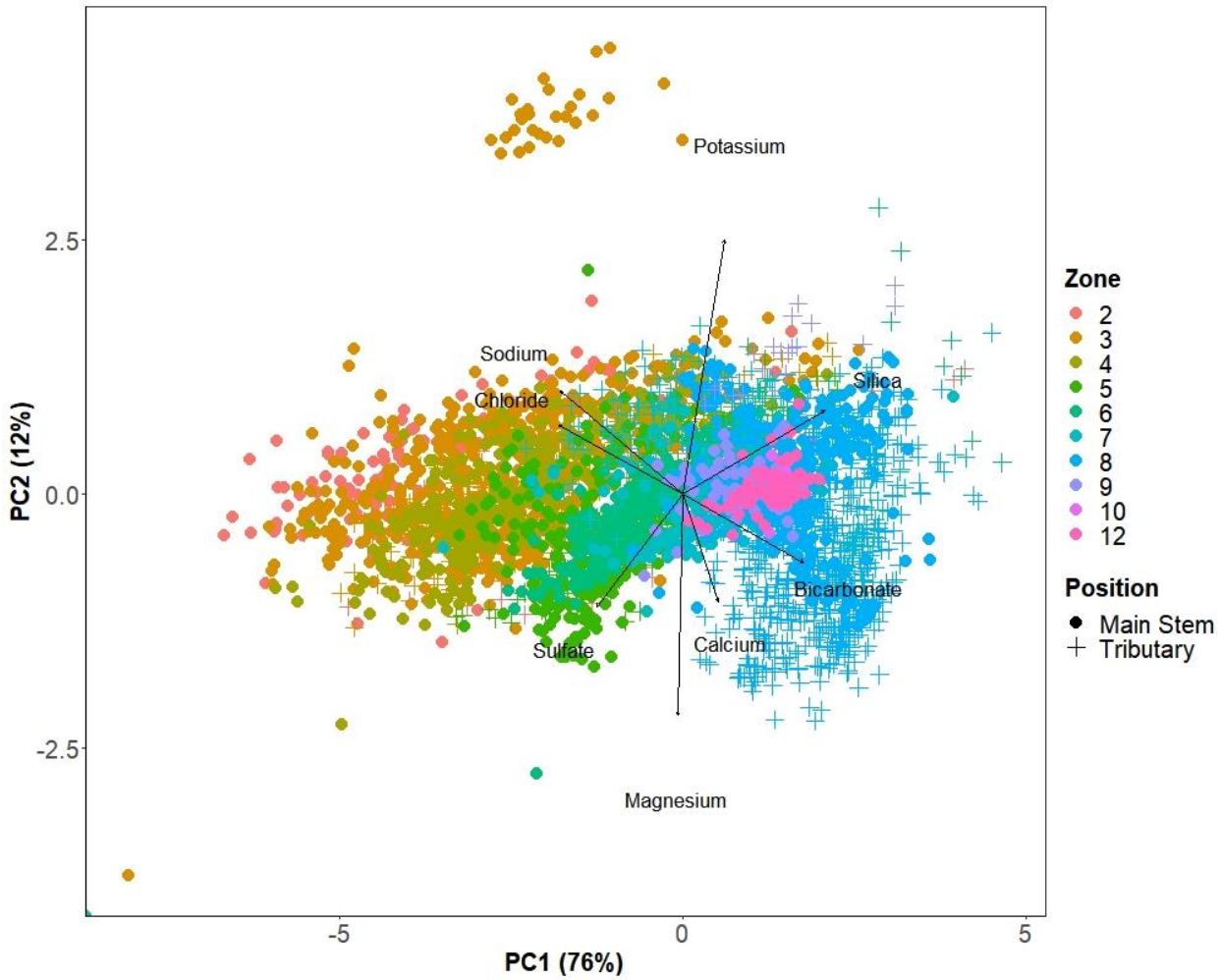

Figure S4 PCA Biplot for the Colorado River showing distribution of observations across PC1 and PC2 colored by river zone (2-12) and watershed position (tributary or main stem)

The compositional PCA analysis also shows that stream type (tributary vs main stem) does not play a dominant role on the distribution of variance observed across PC1 and PC2 (Figure S4). The area where most variation between main stem and tributary points occurs is in zone 8 which is dominated by measurements collected on tributaries near the main stem of the Colorado River.

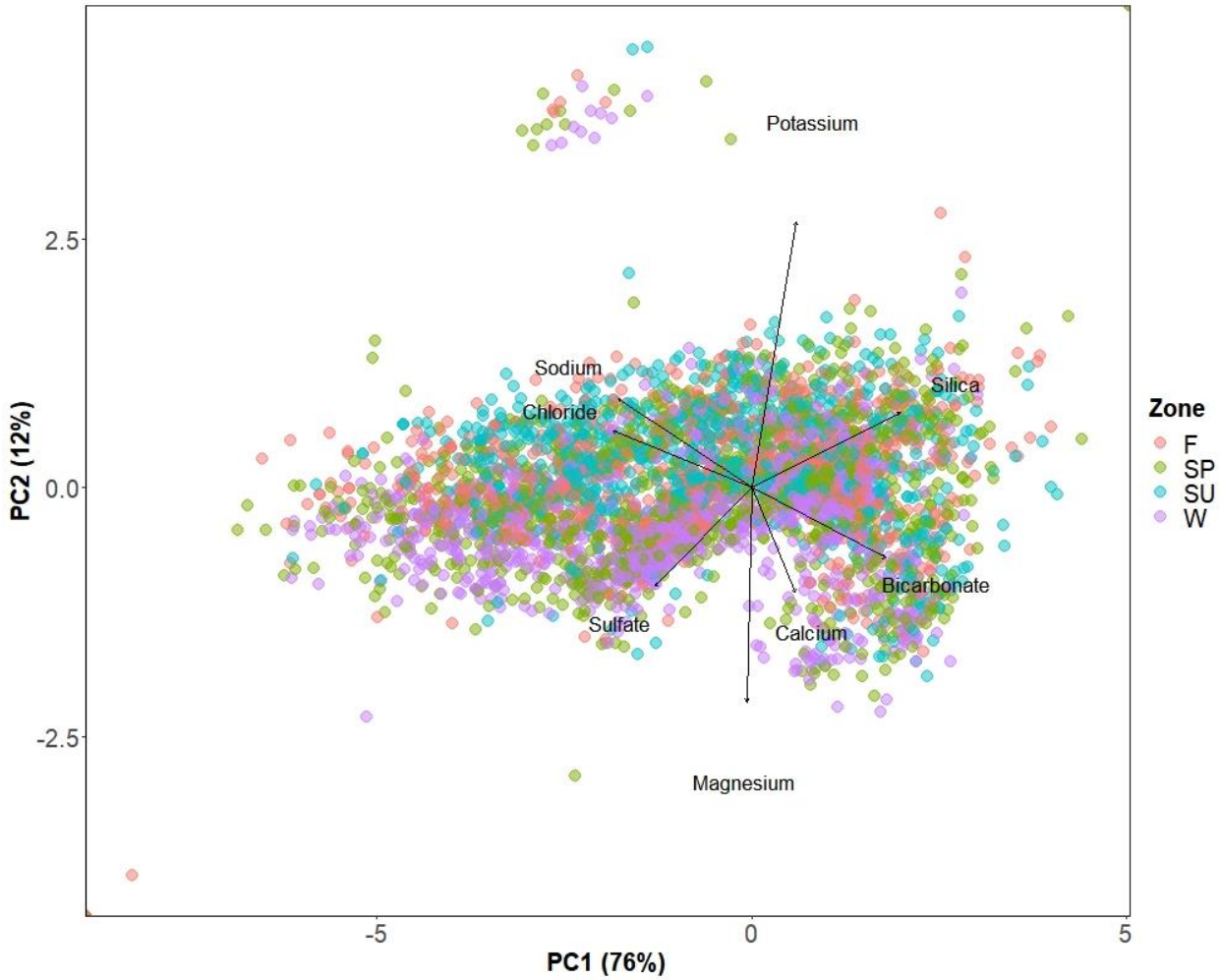

Figure S5 PCA biplot for the Colorado River showing distribution of observations across PC1 and PC2 colored by season (F=Fall (September, October, November), SP=Spring (March, April, May), SU=Summer (June, July, August), W=Winter (December, January, February))

There is not a strong seasonal trend across PC1 which explains 76% of the variation. The distribution of points across PC2 shows some division between measurements collected in summer vs winter indicating that seasonal influence may contribute a very small proportion of the overall variance in the dataset.

Random Forest and Verification

Lithology, LULC (Land use/land cover, and stream chemical data was split into training (80%) and testing (20%) subsets. Individual random forest models were constructed using the training dataset for each watershed factor compared to all solutes using the randomForest package in R (Liaw & Wiener, 2002). The strength of the relationship between each solute and land use factor was assessed using the percent increase in mean square error (%incMSE) if that solute was removed from consideration. A larger %IncMSE indicates a stronger relationship between the two variables Each model was constructed with 500 trees and a minimum terminal node size of 5.

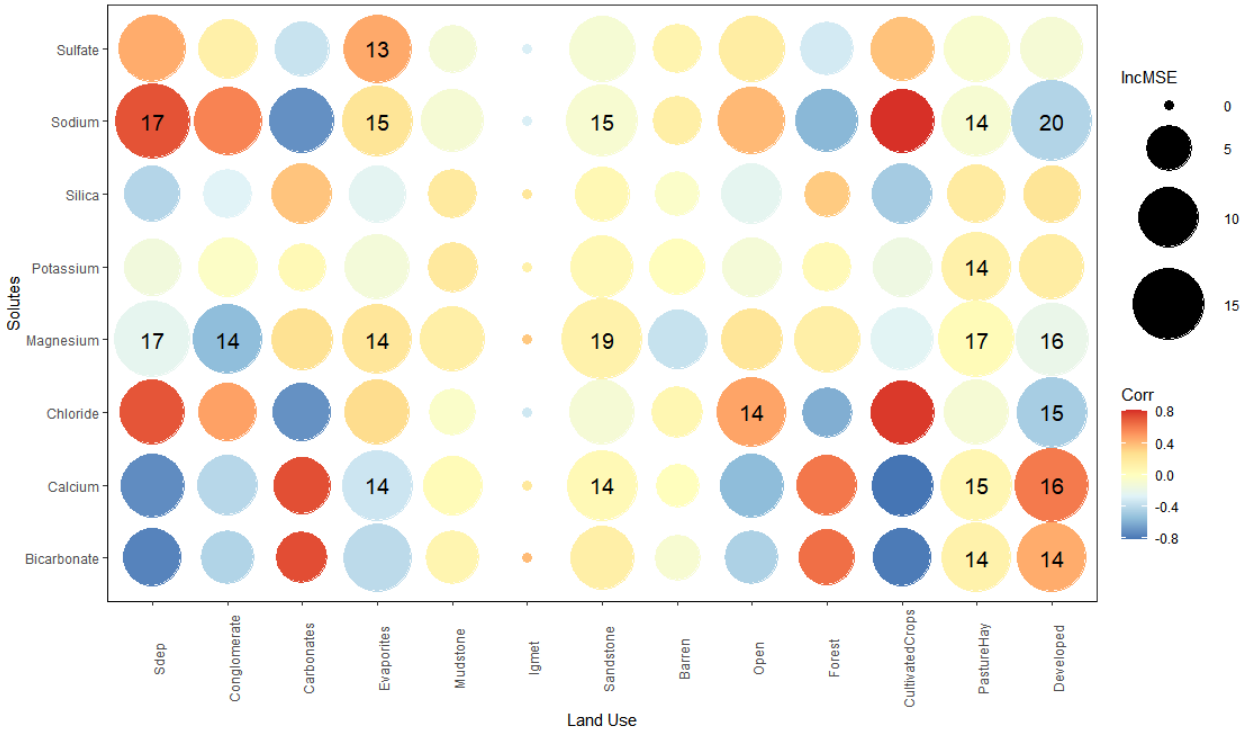

Figure S6 Random forest output for upstream zones (2-5) of the Colorado River including 25 total sites and 2,951 observations. X axis shows lithology and LULC classes and y axis shows solutes. Point size represents magnitude of %IncMSE and point color represents correlation between the two variables. Points with the top 85th percentile and above of %IncMSE are labelled

For the upstream zones (2-5 only) no igneous-metamorphic rocks are present in this region and evaporites are most abundant as compared to the rest of the watershed. Additionally, cultivated crops are common and several small towns represent the largest developed areas. The random forest revealed (Figure S6) that lithologic relationships to stream water chemistry are mostly as-expected with evaporites having strong relationships with Na^+, Ca^{2+}, Mg^{2+}, and SO_4^{2-}. Correlations are low between evaporites and all solutes with the highest correlation between evaporites and sulfate. Sedimentary deposits have the strongest relationships of all lithology classes with %incMSE of 17 for both Na^+ and Mg^{2+}. Sedimentary deposits also have high positive correlations to Na^+ and Cl^- and negative correlations to Ca^{2+} and HCO_3^- .

When we examine all sites (Figure S7) the developed areas and pasture/hay are the most strongly related to all solutes for LULC classes but have weak correlations. Cultivated crops are more

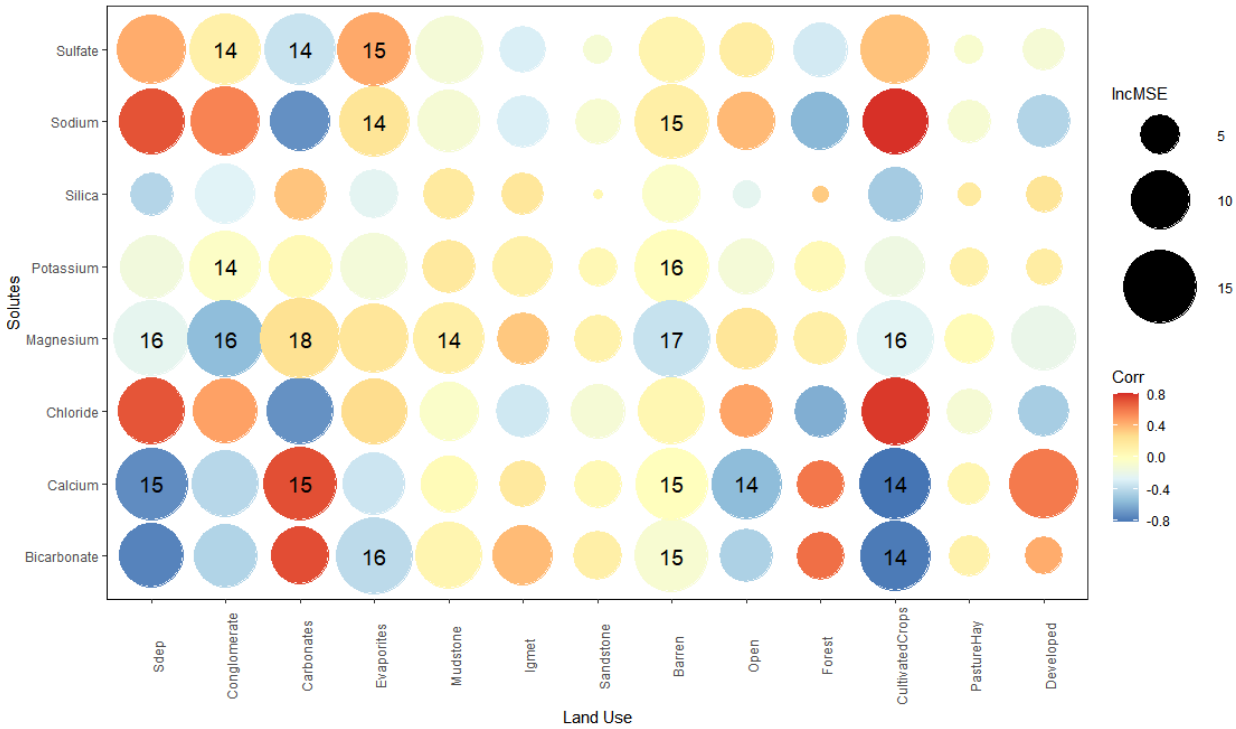

Figure S7 Random forest output for main stem sites on the Colorado River including 36 total sites and 2,686 observations. X axis shows lithology and LULC classes and y axis shows solutes. Point size represents magnitude of %IncMSE and point color represents correlation between the two variables. Points with the top 85th percentile and above of %IncMSE are labelled

highly correlated to Na^+, Cl^-, Ca^{2+} and HCO_3^- but were not shown to have strong relationships via this analysis.

Analysis of sites on the main stem of the Colorado river shows disproportionate influence of Barren areas on many solutes including Na^+, K^+, Mg^{2+}, Ca^{2+}, and HCO_3^- even though correlations are low. Cultivated crops represent a second stand out LULC class with high %incMSE for Mg^{2+}, Ca^{2+}, and HCO_3^- and high correlations for Na^+, Cl^-, Ca^{2+}, and HCO_3^-. A variety of lithologic classes have strong relationships for a few solutes. Sedimentary deposits show high %incMSE for Ca^{2+} and Mg^{2+} and high correlations with Na^+, Cl^-, Ca^{2+} and HCO_3^-. Conglomerate shows high %incMSE for Mg, K, and SO_4. With low to moderate correlations and higher correlations to Na^+. Carbonates have high %incMSE for SO_4^{2-}, Mg^{2+}, and Ca^{2+} and have high correlations for Na^+, Cl^-, Ca^{2+}, and HCO_3^-. Evaporites have weaker correlations but high %incMSE for SO_4^{2-}, Na^+, and HCO_3^-.

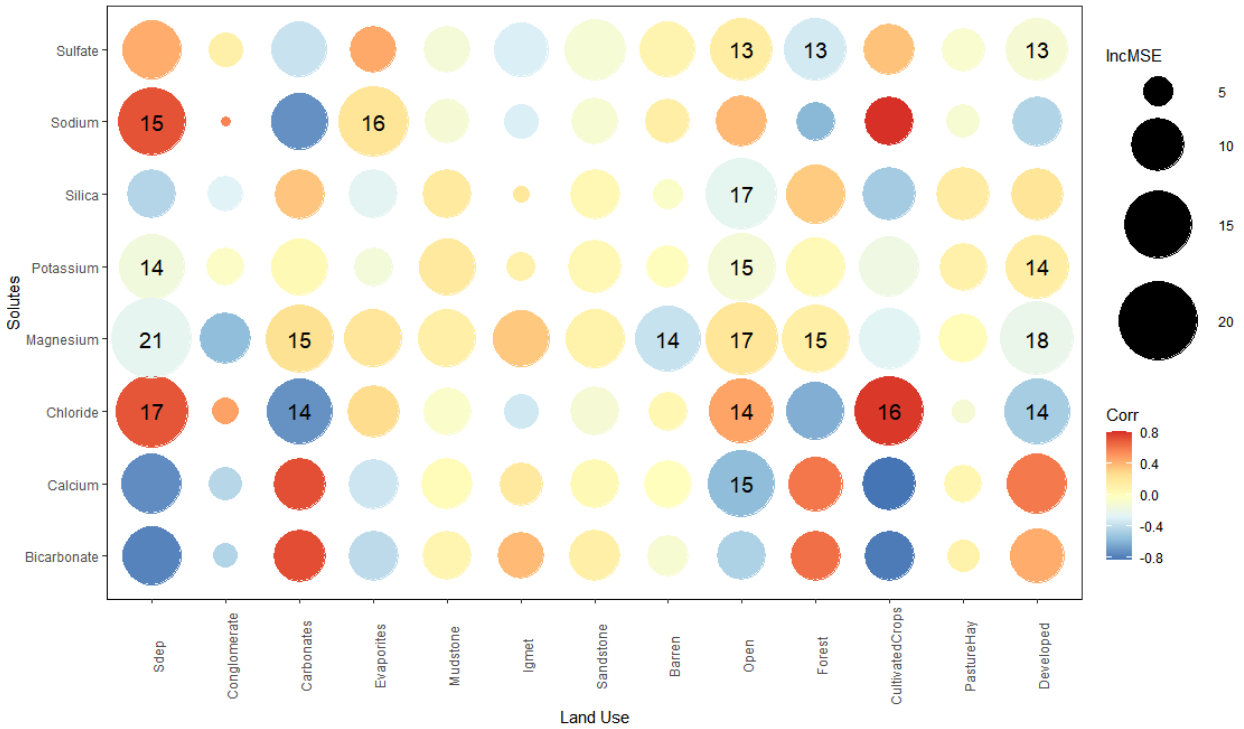

Figure S8 Random forest output for tributary sites on the Colorado River including 44 total sites and 1,766 observations. X axis shows lithology and LULC classes and y axis shows solutes. Point size represents magnitude of %IncMSE and point color represents

It was expected that tributary measurements might have stronger relationships to the underlying watershed factors due to limited mixing and upstream influences as compared to measurements collected from the main stem. The output of the random forest models (Figure S8) shows that some relationships are stronger for tributary sites to a small degree, but the difference is not large. This suggests that water chemical composition between tributaries and the main stem is relatively homogeneous. This is supported by observations using piper diagrams and compositional PCA.