CAMELS-Chem: Augmenting CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) with Atmospheric and Stream Water Chemistry Data

Gary Sterle[1], Julia Perdrial[2], Li Li[3], Thomas Adler[2], Kristen Underwood[4], Donna Rizzo[4], Hang Wen[5] and Adrian Harpold[1]

[1] Department of Natural Resources and Environmental Science, University of Nevada, Reno, USA
[2] Department of Geology, University of Vermont, USA
[3] Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, USA
[4] Department of Civil and Environmental Engineering, University of Vermont, USA
[5] School of Earth System Science, Tianjin University, Tianjin, China

Correspondence to: Adrian A. Harpold (aharpold.unr.edu)

Abstract. Large sample datasets are transforming hypothesis testing and model fidelity in the catchment sciences, but few large stream water chemistry datasets exist with complementary streamflow, meteorology, and catchment physiographic attributes. Here, we pair atmospheric deposition and water chemistry related information with the existing CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) dataset. The newly developed dataset, CAMELS-Chem, comprises U.S. Geological Survey water chemistry data and instantaneous discharge over the period from 1980 through 2014 in 506 minimally impacted headwater catchments. The CAMELS-Chem dataset includes 18 common stream water chemistry constituents: Al, Ca, Cl, Dissolved Organic Carbon, Total Organic Carbon, HCO_3, K, Mg, Na, Total Dissolved Nitrogen [nitrate + nitrite + ammonia + organic-N], Total Organic Nitrogen, NO_3, Dissolved Oxygen, pH (field and lab), Si, SO_4, and water temperature. We also provide an annual wet deposition loads from the National Atmospheric Deposition Program over the same catchments that includes: Ca, Cl, H, K, Mg, and Total Nitrogen from deposition [precipitation NO_3 + NH_4, dry deposition of particulate NH_4, + NO_3, and gaseous NH_3], Na, NH_4, NO_3, SO_4. We release a paired instantaneous discharge (and mean daily discharge) measurement for all chemistry samples. To motivate wider use by the larger scientific community, we develop three example analyses: 1. Atmospheric-aquatic linkages using atmospheric and stream SO4 trends, 2. Hydrologic-biogeochemical linkages using concentration-discharge relations, and 3. Geological-biogeochemical linkages using weathering relations. The retrieval scripts and final dataset of $> 412,801$ individual stream water chemistry measurements are available to the wider scientific community for continued investigation.

1 Introduction

While hydrochemical datasets are increasingly available at continental scales, these datasets have often provided scientific insights focused on detailed knowledge of a single catchment rather than transformative knowledge across many catchments.

To address the need for 'balancing breadth with depth' (Gupta et al., 2014, Hubbard et al., 2020), the hydrological sciences have developed large sample size datasets that are of high quality and then made these available to the research community. One of the key advantages of aggregating and harmonizing data into larger sample size datasets is to test how model hypotheses reproduce observed behavior across variable conditions and sites to reduce the uniqueness of place-based issues that plague catchment science (Gupta et al., 2014) (Hubbard et al., 2020). As a result, several recent efforts have focused on generating and using datasets across the continental U.S. (CONUS) where observation networks are relatively dense. For example, the Model Parameter Estimation Experiment (MOPEX, Duan et al., 2006) dataset has been used to detect the effects of shifts from snow to rain on streamflow (Berghuijs et al. 2014) and to better diagnose the cause of catchment-scale water budgets (Brooks et al., 2015). Recent efforts have extended the record and detail of auxiliary data of older efforts (e.g. MOPEX) to develop longer-term streamflow and hydrometeorological forcing data for a larger number of minimally disturbed catchments, including the continental U.S (Newman et al.), Chile (Alvarez-Garreton et al., 2018), and Brazil (Chagas et al., 2020). In particular, Catchment Attributes for Large-Sample Studies (CAMELS, Newman et al., 2014) compile high quality streamflow measurement in 671 unimpaired catchments, as well as climate forcing datasets (e.g. daily precipitation and temperature) and physiographic properties (e.g. land cover, topography, etc., Addor et al., 2017). CAMELS has seen widespread adoption by the hydrological community as a benchmarking tool for hydrological models (Melsen et al., 2018; Mizukami et al., 2019; Pool et al., 2019), in the development of hydrological signatures and new information theory-based approaches, and the application of novel machine learning tools (Kratzert et al., 2018). However, the generation of matching datasets on stream water chemistry has not seen as much development.

Indeed, some of the most important insights across Earth science disciplines over the last several decades derived from investigations that combined several datasets such as long-term stream chemistry data, stream discharge, hydroclimatology, and catchment properties (e.g. vegetation, geology, topography). For example, global analyses of CO2 evasion from headwater streams have helped to quantify global fluxes (Gaillardet et al., 2018; Horgby et al., 2019; Lauerwald et al., 2015; Raymond et al., 2013). Changes in dissolved organic carbon (DOC) in stream water were partially related to changes in atmospheric deposition and acidity only viewable with longer records (Monteith et al., 2007). Stream flow and chemistry data, in particular combined(?) concentration and discharge data, have also been shown to illuminate subsurface flow paths and chemical vertical stratification (Stewart et al., 2022; Zhi and Li, 2020). Applying an integrative dataset in the Northeastern U.S., a recent study showed differential sensitivity of headwater catchments to reductions in SO4 and NO3 and resulting variations in stream DOC efflux (CITE Adler et al. 2021). Importantly, this study confirmed that much of the long-term recovery from acid rain is mediated by catchment-scale processes in variable soils and bedrock as well as variable hydrological and climatic forcings. Only by aggregating data across many catchments could these interacting factors affecting stream chemistry be determined because transferability from a small number of regional catchments is very limited (Clow et al, 1999; Garmo et al., 2014; Harpold et al., 2010). Long-term water chemistry datasets have also given insight into rock weathering and solute flux estimates. For example, Godsey et al. (2009) showed the ubiquity of 'chemostasis', where solute fluxes are primarily driven

by stream discharge and not variations in concentrations in 59 catchments but did not investigate the role of catchment properties or climate. More recent work using 2,186 catchments from the Global River Chemistry Dataset (GLORICH) began to illustrate the role of aridity and catchment properties in controlling the concentration-discharge (C-Q) relationship in key solutes (Godsey et al., 2019).

70

An opportunity exists to inform our understanding of the inter-relationships between water chemistry, hydrology, and biogeochemistry by developing new datasets that link stream chemistry to discharge and other catchment properties. A notable example is the GLORICH dataset that focuses on large, human impacted riverine systems and contains corresponding catchment properties and streamflow data from over 17,000 sampling locations (Hartman et al., 2019). However, GLORICH does not include complete observation of streamflow (and instead supplements with modeled data); its climate forcing datasets and catchment properties are scant, significantly limiting its application to process-based modeling and data analysis. As a result, we lack understanding on how stream chemistry responds to climatic forcings and catchment structure characteristics at the continental to global scales (Godsey et al., 2019). The continued development of large sample datasets that include stream chemistry has the potential to transform catchment science and address grand scientific challenges.

80

To enable and advance these (bio)geo-chemistry and catchment sciences, we compiled a new dataset that takes advantage of one of the most comprehensive collections of catchment attributes across the US i.e., CAMELS (Addor et al., 2017; Newman et al., 2014), and amended it with paired quality-controlled water chemistry data, instantaneous discharge, and atmospheric deposition data, hereafter referred to CAMELS-Chem. We harvested >412,801 individual water chemistry measurements from the U.S. Geological Survey (USGS) National Water Information System (NWIS), covering 506 catchments and 16 constituents, over the period from 1980 to 2014 as well as annual catchment-scale atmospheric deposition from a product developed by the National Atmospheric Deposition Program (NADP) for the same time period. The publicly released dataset also includes scripts that increase reproducibility and the potential to update the record with more current data. We first introduce the methods used to develop the CAMELS-Chem dataset (Section 2) and then summarize the availability and spatial patterns of key water chemistry constituents (Section 3). Finally, we develop example analyses using the large hydrochemistry dataset to demonstrate its utility for applications in (bio)geo-chemistry, ecology, and hydrology (Section 4).

2 Materials and Methods

2.1 Data Sources

Two different data sources were identified for acquisition 1) NWIS instantaneous and daily mean discharge data and 2) NWIS water quality data for the (bio)geo-chemical records. NWIS time series data significantly reduce the need to resolve semantic and structural conflicts due to the data having already been normalized and quality controlled. NWIS data also have unique gauge identifiers for each observation reducing the complexity of merging the two datasets. The geographical coverage for the

dataset encompasses the gauge IDs of the original 671 CAMELS watersheds and uses the same temporal period of the CAMELS dataset from water years 1980 to 2014. The resulting data capture rendered a water quality record comprising 493 watersheds wherein 178 gauges did not have any associated water quality data. All 671 CAMELS gauges returned values for both instantaneous and daily mean discharge.

2.2 Applied Methods

The temporal and geographical span of this dataset posed two distinct data management needs: 1) the integration of a large amount of data from disparate data repository systems such that different gauges could be compared, and 2) an integrated data repository to group and browse large amounts of data.

Data integration required the development of processes for the acquisition and integration of the NWIS data; its transformation into appropriate subsets and subsequent storage of the transformed data was solved using a data integration process called Extract, Transform, and Load (ETL). ETL generally involves extracting data from outside sources (Figure 1), transforming the data to fit operational needs (which can include quality definitions) and loading the data into an end target (e.g., a database or data warehouse). ETL is an especially useful method in the creation and management of efficient and consistent databases and data warehouses. Moreover, ETL tools allow for the fast, straightforward, and reproducible loading of new (and reloading of old) datasets, which facilitates continuous data updates, and auditing of the data acquisition processes. For this project, Pentaho Data Integration (PDI) ETL (Pentaho Data Integration 2020) was used.

The second need, an integrated data repository, was solved by using an online transactional processing (OLTP) relational database management system (RDBMS). For the data repository, we selected Oracle's free and open-source database, PostgreSQL (PostgreSQL 2020). OLTP databases are optimized to prevent data anomalies by only storing data in one place and using keys to relate different tables and data to each other. A relational database provides fast query processing, enforces data integrity, provides detailed information about current data and schema, and a flexible platform to export data for external analysis. Statistics were calculated for key parameters such as probability of exceedance, standard deviation, low/medium/high flow conditions, mean daily and annual flow. These were used in the following figures and are also available as part of this data release.

2.3 Related Discharge Dataset

In addition to a daily average discharge, we solve several challenges to release an instantaneous discharge dataset. Approximately 93% of the samples in CAMELS-Chem have a paired daily discharge value; however, the coverage for paired instantaneous discharge values is lower at 15% because not all gauges provide hourly observations or they provide discharge only from a more recent part of the record. The biogeochemical sample time was adjusted forward/backward to pair with the closest quarter -hour (or sometimes hourly) instantaneous discharge observation. Missing time stamps for (bio)geo-chemical

130   sampling is more challenging to correct. We address this by using the fact that field technicians generally visit and sample once per day so the recorded time from one sample is as representative for all samples taken on that day. By assuming one sampling time every day per site we obtained another 397 instantaneous discharge (out of a total of 8,975 records without timestamps).

2.4 Wet Deposition Dataset

135   Wet deposition data were obtained from NADP for the period from 1985 until 2019. The data product is generally an inverse distance weighted interpolation of the wet deposition observations; more details are provided on their website. Data rasters were aligned with CAMELS catchment shape files to determine total watershed deposition for 9 species for a given year. Quantum GIS 3.12 was used to calculate total (mean) deposition for a given year by catchment. Minimum, maximum, and mean values were calculated for each catchment (Table 2).

140   **Table 1: The stream water chemistry datasets in the CAMELS-Chem dataset including the name, units, and description that includes the abbreviation used in this paper.**

Attribute	Name	Description	Units
Al	Aluminum	Water, filtered	mg/l
Ca	Calcium	Water, filtered	mg/l
Cl	Chloride	Water, filtered	mg/l
DOC	Dissolved Organic Carbon	Water, filtered	mg/l
TOC	Total Organic Carbon	Water, unfiltered	mg/l
HCO$_3$	Bicarbonate	Water, filtered, field, inflection-point (incremental titration method)	mg/l
K	Potassium	Water, filtered	mg/l
Mg	Magnesium	Suspended sediment, total	mg/l
TDN	Total Dissolved Nitrogen	Water, filtered [nitrate + nitrite + ammonia + organic-N]	mg/l
TON	Organic Nitrogen	Water, unfiltered	mg/l
Na	Sodium	Water, filtered	mg/l
NO$_3$	Nitrate	Water, filtered	mg/l
DO	Dissolved Oxygen	Water, unfiltered	mg/l
pH	pH	Water, unfiltered, field	std units
pH_l	pH	Water, unfiltered, laboratory	std units
Si	Silica	Water, filtered	mg/l

SO₄	Sulfate	Water, filtered	mg/l

Table 2: NADP depositional dataset, denoting concentration for each constituent, including the name, units, and description that includes the abbreviation used in this paper. *TN_d denotes Total Nitrogen depositional observation [precipitation NO3 + NH4, dry deposition of particulate NH4, + NO3, and gaseous NH3].

Attribute	Name	Units
Ca	Calcium	mg/l
Cl	Chloride	mg/l
H	Hydrogen	mg/l
K	Potassium	mg/l
Mg	Magnesium	mg/l
*TN_d	Total Nitrogen	mg/l
Na	Sodium	mg/l
NH₄	Ammonium	mg/l
NO₃	Nitrate	mg/l
SO₄	Sulfate	mg/l

3 Dataset Description

The CAMELS-Chem dataset comprises 17 water chemistry and property values (Table 1), instantaneous discharge, and annual wet deposition loads (Table 2). A comparison of the stream water and wet deposition constituents are provided in Table 3. The focus of CAMELS-Chem is stream water chemistry, which we organize into cations (Al, Ca, K, Mg, Na, Si), anions and nutrients (Cl, DOC, HCO3, NO3, SO₄, TDN, TOC) and general water quality parameters (pH, DO, and water temperature). The additional wet deposition dataset can also be organized into cations (Ca, H, K, Mg, Na, NH4,), anions (Cl, NO3, SO4), and TN. In the next two sections we describe and exemplify the scope and overall spatial patterns of the stream chemistry (Section 3.1) and wet deposition data (Section 3.2) for a subset of 8 constituents (Table 2: Ca, DO, DOC, Na, NO3, pH, Si, SO4) and discharge. Information for other constituents is provided in the supplementary materials.

Table 3: NADP depositional dataset, denoting concentration for each constituent, including the name, units, and description that includes the abbreviation used in this paper. *TN_d denotes Total Nitrogen depositional observation [precipitation NO3 + NH4, dry deposition of particulate NH4, + NO3, and gaseous NH3].

Source	Group	Water Chemistry or Property

USGS NWIS Surface water chemistry	Cations	Al, Ca, Mg, Na, K, Si
	Anion and nutrients	Cl, DOC, HCO₃, NO₃, SO₄, TDN, T(
	General water quality	pH, DO, temperature
NADP wet deposition	Cations	Ca, K, Mg, Na, H (pH)
	Anions and nutrients	Cl, TN_d, NO₃, SO₄

3.1 Water Chemistry Constituents and Data Availability

While the lengths of discharge and climate records extend up to or greater than 30 years for many CAMELS-Chem catchments,

165 the water chemistry data are not as continuous and spatially consistent (Figure 2 and S1). The total number of stream water samples varies substantially depending on the variable of interest, which should be taken into consideration when using this dataset. For example, the dataset has 34,704 total Si measurements from 477 catchments, compared to only 11,101 DOC measurements from 189 catchments (Figure 2). In general, weathering-related solutes (Si, Ca, K, Mg, Na) are more comprehensively sampled (i.e. longer records and more catchments) than biologically driven constituents (NO3, TDN, DOC).

170 Temperature and DO are sampled more often because they are part of the standard USGS field measurements at the time of water chemistry sampling. Because USGS sampling foci varied over each decade, temporal biases in the sampling record exist. For example, many of the stream water constituents impacted by acid rain (e.g. SO₄, Ca) were sampled less frequently starting in the late 1990's. In contrast, sampling frequency for many solutes related to local water quality issues (i.e. NO3, and K) have increased in recent decades (Figure 2). The NO3 data are more focused in the Midwest and east coast where

175 agriculture is generally more intense. In contrast, DOC observations highlight the location of long-term sites focused on minimally human-impacted catchments (e.g., UGSG Hydrologic Benchmark Network) that have the most comprehensive sampling of all solutes (Gupta et al., 2014; Mast, 2013). We explore how the consistency of spatial and temporal sampling can affect analysis techniques and inferences in Section 4.

180

3.2 Atmospheric Deposition Dataset

Annual average atmospheric deposition maps developed by the National Atmospheric Deposition Program (NADP) are reported for each catchment (Figure 4). The wet deposition data are important to explain some water chemistry patterns. For example, Cl deposition near the coasts (Figure 3) is evident in higher streamflow Cl concentrations along the coast relative to more inland catchment (Figure 3). Conversely, atmospheric deposition of NO3 is much higher in the Midwest and other places where anthropogenic inputs of fertilizer are high (Figure 4) and are consistent with patterns in stream chemistry (Figure 2). The annual resolution of the wet deposition data is sufficient to investigate trends in air and water quality (see example 4.3) and could be supplemented as additional datasets become available (i.e., total wet and dry deposition; Schwede et al., 2014).

4 Example Analyses Using CAMELS-Chem

We illustrate the research potential of CAMELS-Chem using three themes and examples: 1. by investigating trends in SO_4 deposition and stream chemistry over the continental U.S we show highlight atmospheric-aquatic linkages, 2. by investigating continental-scale C-Q relations we highlight hydrological and biogeochemical linkages, and 3. by investigating weathering patterns we highlight geological-biogeochemical linkages. Our goal for this section is to demonstrate the unique value of the CAMEL-Chem dataset, its potential limitations, and to motivate future work.

4.1 Trends in Atmospheric SO_4 Deposition and Stream Chemistry

4.1.1 Background and Motivation

The Industrial Revolution caused a rapid increase in fossil fuel emissions, which introduced acid anions (SO_4, $NO3$) in excess of background conditions leading to acidic precipitation throughout many industrialized regions (Newell & Skjelkvåle, 1997). The Clean Air Act in 1970 and subsequent amendment in 1990 led to major reduction in air pollution as apparent in progressively decreasing deposition amounts (Accessed September 20, 2020: https://www.epa.gov/air-trends/sulfur-dioxide-trends). These significant changes in atmospheric composition served as an experiment in stream chemistry response (Stoddard et al, 1990) and motivated several studies across the US and Europe. For example, Stoddard et al. (1990) showed that stream water acid anion concentration declined more in the 1990s than the 1980s and later research indicated that effects were larger in Europe than North America (Garmo et al., 2014). In the Northeastern U.S., SO_4 stream chemistry has generally responded to declines in SO_4 deposition (McHale et al., 2017; Siemion et al., 2018). However, catchment properties and flow regime were shown to impact the sensitivity of stream response to shifts in deposition (Clow et al., 2018; Murdoch et al., 2006). The lack of co-located long-term deposition data, stream water chemistry and discharge data, and catchment attributes hinders investigation of the legacy of effects of acid rain on stream chemistry at continental scales.

4.1.2 Analysis and Findings

CAMELS-Chem provides new potential to analyze the effects of acid deposition on long-term stream chemistry trends across
215   a range of hydrological conditions. Many (but not all) sites have >100 SO4 stream samples spanning over 20 years of record
(Figure 2). Our analyses of the flow duration curves (FDC, divided into terciles of probability of exceedance < 0.33) show
that bias towards part of the flow regime when SO4 samples is common and depends on the location of interest (Figure 5).
For example, some streams in central Arizona and North Dakota have no low flow samples (FDC>66%), whereas streams in
Oregon and Washington tend to have more moderate flows (33%>FDC<66%) than high flows (FDC<33%; Figure 5).

220

We performed a Mann-Kendall test to investigate trends in atmospheric SO4 deposition and resulting trends in stream water
chemistry (Hirsch et al., 1984). Our results are in agreement with previous findings of declining SO4 deposition following the
1990 Clear Air Amendments (Figure 5, Stoddard et al., 2003). As expected, plotting trends in SO4 stream chemistry and wet
deposition just after the Clean Air Act (Figure 5a) show a predominance of decreasing trends in SO4 deposition and
225   corresponding decreasing trends in SO4 stream chemistry. Wet deposition trends remain decreasing in the following two
decades (1995-2005 and 2005-2015), but with much response in SO4 stream chemistry. This initial analysis provides a starting
point for hypothesis testing - for example, on the role of catchment attributes that are shown for the dominant geology in Figure
5a-c. Furthermore, we hope to motivate more sophisticated statistical and machine learning techniques to investigate coupled
trends in solutes and the role of climate physiographic information as predictor variables.

230   ## 4.2 Concentration-Discharge Relationships

4.2.1 Background and Motivation

Concentration-discharge (C-Q) relationships are routinely used to compute solute loads for mass budgets (Cohn et al., 1989)
and have been used to infer catchment effects on biogeochemical cycling (Basu et al., 2010, Musolff et al., 2015). When
event-scale data is not available, routine sampling can be used to compute a log-log C-Q relationship to understand whether
235   concentration has a clear dependency on discharge. The log-log C-Q relationship is considered chemostatic if the regression
slope is near zero, implying that export of that solute is negligibly influenced by discharge (Godsey et al., 2009, Underwood
et al., 2017, Zhi and Li, 2020).

This chemostatic relation is more likely for geogenic solutes, like Si and Ca, and less so for biogenic solutes, like nutrients,
240   with large anthropogenic or natural legacy effects (Musolff et al., 2015). Recent work by Godsey et al., (2019) suggests that
groundwater and fast chemical reactions buffer C-Q relationships towards chemostasis, but that the baseline concentrations
reflect catchment differences in geology, land use, etc. Godsey et al. (2019) also found that nutrients can be chemostatic, but
their long-term mean concentrations correlated more to land use than climate. Moreover, sampling bias occurring at high and
low streamflow (Figure 5, as well as rising and falling limb of hydrograph, season, etc.) may bias effects towards chemostatic

behavior (Zhi et al., 2019). These efforts with CAMELS-Chem show the value of large-sample datasets for inferring C-Q relations and their uncertainty.

4.2.2 Analysis and Findings

The CAMELS-Chem dataset offers potential to resolve C-Q relationships by improving discharge estimates and pairing with substantial existing physiographic information. Here we repeat the methods and analysis of Godsey et al. (2019) to show similar event-scale (i.e. instanaeous water chemistry and discharge) C-Q slopes across the overlapping solutes (Figure 6). The existing physiographic data compiled by (Addor et al., 2017) allows us to easily filter by the aridity index to match the previous analysis of Godsey et al. 2019. Arid catchments were defined as those in which potential evapotranspiration was greater than actual evapotranspiration, the reverse was used to determine humid catchments. In general, the findings for CONUS sites are consistent with the global-scale analysis of Godsey et al. (2019), showing near-chemostatic behavior of DOC and Si. In contrast to Godsey et al. (2019), Na for the CAMELS sites shows more pronounced negative slopes in arid catchments than in humid catchments. This analysis could be expanded to include additional solutes, or physiographic data, in more complex statistical or machine learning based approaches that attempt to infer association with biogeochemical process and uncertainty due to irregular and limited sampling.

The CAMELS-Chem dataset has several advantages in this C-Q analysis: 1) instantaneous (hourly) discharge data were used to supplement the NWIS database and 2) 30-year daily discharge records are a reference for C-Q ranges. Because Q values are more likely to change over daily periods at high flows than low flows, instantaneous discharge measurements are more critical on high flow days than low flow days. Instantaneous discharge during the chemistry sampling can be > 50% higher or lower than the daily mean discharge for a range of important solutes, despite little overall bias (Figure S4). Thus, the CAMELS-Chem discharge data could improve accuracy in previous C-Q analysis that rely on daily measured or modeled discharge data. The 30-year daily discharge records can be used to assess the degree of discharge variation covered by chemistry sampling dates. We show this as the percent of the long-term FDC covered by the minimum and maximum discharge during sampling in Figure 7. While many sites cover > 90% of the FDC, like those in the upper Midwest, other locations cover less than 50% of the FDC, like many of the Gulf Coast sites. The response of the concentration discharge relationship can be different at high and low flows (Li et al., 2020). The diverse number of samples with uneven collection dates, as well as the lack of discharge variation during solute sampling, suggest that future work should explore C-Q uncertainty estimates using more sophisticated statistical techniques with CAMELS-Chem.

4.3 Weathering Response

4.3.1 Background and Motivation

275 Weathering (i.e. the breakdown of bedrock and regolith) involves key physical and chemical processes that shape the critical zone and provide important nutrients for all ecosystems. Weathering of silicate minerals regulates global atmospheric CO_2 levels over geologic time scales (Spence and Telmer, 2002), and weathering of carbonate rocks modulate CO_2 levels at the human to century time scales (Martin, 2018). Riverine composition is often used as an indicator for weathering rates and to draw conclusions on implications at larger scales. A classic example is the study by Gaillardet et al. (1999) where 60 of the

280 world largest rivers were used to show a strong role of bedrock lithology on weathering rates from Si effluxes. To display the impact of major rock classes (i.e. silicates, carbonates, evaporites) on riverine composition, the use of molar ratios for geogenic species (Ca/Na, Mg/Na, HCO3/Na) instead of absolute concentrations is useful, because large differences in concentrations between solid and liquid phases make comparisons difficult. CAMELS-Chem provides complete and co-sampled water chemistry datasets, with known lithology, making these types of weathering analyses possible across CONUS.

285 #### 4.3.2 Analysis and Findings

CAMELS-Chem has sufficient data on the dissolved load of geogenic species to be used in a regional scale weathering study across variable lithology and climate. We display these molar ratios of stream water composition for all CAMELS-Chem sites (with a sample number larger than 5) as a function of bedrock lithology including igneous, metamorphic and sedimentary rocks (Fig 5). We included only samples in the lower tercile of daily discharge values because the low flow discharge is

290 more likely from longer residence groundwater that carries the weathering signal. The lower tercile encompasses more Ca and Mg samples than Na and HCO3 samples based on overall sampling frequency (Figure 9). Similar to previous studies (Gaillardet et al. 1999), we observe organization as a function of lithology, with catchments underlain by carbonate plotting in the upper right (i.e., high Ca/Na, Mg/Na, and HCO3/Na ratios) and unconsolidated sediments plotting in the lower left (e.g., low Ca/Na, Mg/Na, and HCO3/Na ratios). These results are consistent with the high weathering rates of carbonates, where

295 even small amounts of carbonate lithology lead to significant shifts to higher Ca/Na (calcite endmember) and Mg/Na (presence of dolomite) ratios. Although beyond the scope of this work, CAMELS-Chem gives sufficient samples to provide uncertainty estimates in Figure 8, particularly given the uneven number of samples and distribution across solutes (Figure 9). CAMELS-Chem also allows consideration of catchment properties to better understand connections to climate, hydrology, and other physiographic effects. For example, we visualize the baseflow index to show generally higher baseflow in carbonates-underlain

300 catchments (Figure 8), which is consistent with fractures and highly conductive conduits that are common in carbonate aquifers (Hartmann et al., 2014). In contrast, unconsolidated sediments tend to have low weathering rates and low baseflow index (Figure 8). The control of lithology on geogenic species concentrations needs to be taken into account when investigating other processes, like baseline concentrations found in the C-Q relationships (i.e. Godsey et al. 2019), particularly when using a single or small number of catchments to infer large-scale response. CAMELS-Chem includes these important data on

305   lithology across many catchments, which we hope motivates the low-temperature geochemistry research community to take
another look at weathering processes across CONUS.

5 Summary and Conclusions

We developed and released the CAMELS-Chem dataset comprising > 44,000 water chemistry measurements, corresponding
instantaneous discharge, and wet deposition data.   The dataset has several advantages because it is paired with

310   hydrometeorological and hydrological modeling data (Newman et al., 2014), as well as catchment physiographic properties
(Addor et al. 2017).   In addition, the consistency of USGS water chemistry analysis increases the comparability of the dataset
across regions and decades, which is a well-recognized problem (Godsey et al., 2019; Niu et al., 2018).

CAMELS-Chem has several limitations that were prominently illustrated through examples.   In general, the sampling

315   frequency and chemical variables of interest are not consistent over time, reflecting changing priorities and budgets within the
USGS. This may require subsetting the data into the catchments with longer records or using new machine learning and
statistical techniques that can take advantage of sparse data within a larger data analysis. The CAMELS-Chem dataset shows
that bi-weekly or coarse water sampling reflects somewhat random discharge conditions on the flow duration curve (Figure
S4). An advantage of CAMELS-Chem would be to use long-term discharge to constrain and propagate uncertainty in water

320   chemistry-based flux and modeling efforts.

We illustrate several basic applications of CAMELS-Chem to motivate future research efforts across a variety of fields. One
advantage of a comprehensive dataset such as CAMELS-Chem is to target new measurement campaigns that either extend
long-term records, add new chemical species, or add new sites all together. These long-term records should guide future large-

325   scale sampling campaigns and/or new gauges. For example, consistent co-sampling of solutes across the streamflow regime
provides more information than limited discharge conditions, which should be paired with regional or larger-scale consistency
in the variables of interest. Another exciting application of CAMELS-Chem is use in 'big data' applications of advanced
statistical or machine learning tools that are focused on specific research questions. For example, large data approaches can
develop new predictive tools for ungauged locations to help with water quality management (Zhi et al., 2020) or make process

330   inferences and improve predictive models (Nearing et al., 2020). All of the research directions take advantage of the many
unique aspects of CAMELS-Chem, which bodes well for its adoption and use by a variety of scientific disciplines.

Code and Data Availability

Data are made available on this google drive with plans to publish a more permenant place in the next round of reviews.
https://drive.google.com/drive/folders/1AF37U3jXW8nxIe195bb2nN2HDpDsdKVr?usp=sharing

Author Contributions

G.S., A.H., and J.P. conceived and planned the study. G.S., T.A. and K.U. contributed to the analysis. All authors contributed to the writing of the paper

Competing Interests

The authors declare no competing interests.

References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and earth system sciences 21, 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Adler, T., Underwood, K., Rizzo, D., Harpold, A., Sterle, G., Li, L., Wen, H., Stinson, L., Bristol, C., Lini, A. and Perdrial, N., 2021. Drivers of Dissolved Organic Carbon Mobilization from Forested Headwater Catchments: A Multi Scaled Approach. Frontiers in Water, 3(63). https://doi.org/10.3389/frwa.2021.578608

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., ... & Ayala, A. (2018). The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies–Chile dataset. Hydrology and Earth System Sciences, 22(11), 5817-5846. https://hess.copernicus.org/articles/22/5817/2018/

Basu, N.B., Destouni, G., Jawitz, J.W., Thompson, S.E., Loukinova, N.V., Darracq, A., Zanardo, S., Yaeger, M., Sivapalan, M., Rinaldo, A., Rao, P.S.C., 2010. Nutrient loads exported from managed catchments reveal emergent biogeochemical stationarity. Geophysical research letters 37. https://doi.org/10.1029/2010gl045168

Berghuijs, W.R., Woods, R.A., Hrachowitz, M., 2014. A precipitation shift from snow towards rain leads to a decrease in streamflow. Nature climate change 4, 583–586. https://doi.org/10.1038/nclimate2246

Brooks, P.D., Chorover, J., Fan, Y., Godsey, S.E., Maxwell, R.M., McNamara, J.P., Tague, C., Univ. of Colorado, B., CO (United States), 2015. Hydrological partitioning in the critical zone: Recent advances and opportunities for developing transferable understanding of water cycle dynamics. Water resources research 51, 6973–6987. https://doi.org/10.1002/2015wr017039

Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., & Siqueira, V. A. (2020). CAMELS-
365   BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. Earth System Science Data, 12(3),
2075-2096. https://essd.copernicus.org/articles/12/2075/2020/

Clow, D.W., Mast, M.A., 1999. Long-term trends in stream water and precipitation chemistry at five headwater basins in the
northeastern United States. Water resources research 35, 541–554. https://doi.org/10.1029/1998WR900050

370

Clow, D.W., Mast, M.A., Sickman, J.O., 2018. Linking transit times to catchment sensitivity to atmospheric deposition of
acidity and nitrogen in mountains of the western United States. Hydrological processes 32, 2456–2470.
https://doi.org/10.1002/hyp.13183

375   Cohn, T.A., Delong, L.L., Gilroy, E.J., Hirsch, R.M., Wells, D.K., 1989. Estimating constituent loads. Water resources research
25, 937–942. https://doi.org/10.1029/wr025i005p00937

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue,
T., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E.F.,
380   2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second
and third workshops. Journal of hydrology (Amsterdam) 320, 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Gaillardet, J., Calmels, D., Romero-Mujalli, G., Zakharova, E., Hartmann, J., 2018. Global climate control on carbonate
weathering intensity. Chemical geology 527, 118762. https://doi.org/10.1016/j.chemgeo.2018.05.009

385

Garmo, Ø.A., Skjelkvåle, B.L., de Wit, H.A., Colombo, L., Curtis, C., Fölster, J., Hoffmann, A., Hruška, J., Høgåsen, T.,
Jeffries, D.S., Keller, W.B., Krám, P., Majer, V., Monteith, D.T., Paterson, A.M., Rogora, M., Rzychon, D., Steingruber, S.,
Stoddard, J.L., Vuorenmaa, J., Worsztynowicz, A., 2014. Trends in Surface Water Chemistry in Acidified Areas in Europe
and North America from 1990 to 2008. Water, air, and soil pollution 225, 1–14. https://doi.org/10.1007/s11270-014-1880-6
390   Godsey, S.E., Hartmann, J., Kirchner, J.W., 2019. Catchment chemostasis revisited: Water quality responds differently to
variations in weather and climate. Hydrological processes 33, 3056–3069. https://doi.org/10.1002/hyp.13554

Godsey, S.E., Kirchner, J.W., Clow, D.W., 2009. Concentration-discharge relationships reflect chemostatic characteristics of
US catchments. Hydrological processes 23, 1844–1864. https://doi.org/10.1002/hyp.7315

395

Gupta, H.V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V., 2014. Large-sample hydrology: a need to balance depth with breadth. Hydrology and earth system sciences 18, 463–477. https://doi.org/10.5194/hess-18-463-2014

400   Harpold, A.A., Burns, D.A., Walter, T., Shaw, S.B., Steenhuis, T.S., 2010. Relating hydrogeomorphic properties to stream buffering chemistry in the Neversink River watershed, New York State, USA. Hydrological processes 24, 3759–3771. https://doi.org/10.1002/hyp.7802

Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., Weiler, M., 2014. Karst water resources in a changing world: Review
405   of hydrological modeling approaches. Reviews of geophysics (1985) 52, 218–242. https://doi.org/10.1002/2013rg000443
Hartmann, J., Lauerwald, R., Moosdorf, N., 2014. A Brief Overview of the GLObal RIver Chemistry Database, GLORICH. Procedia earth and planetary science 10, 23–27. https://doi.org/10.1016/j.proeps.2014.08.005

Hirsch, R.M., Slack, J.R., 1984. A Nonparametric Trend Test for Seasonal Data With Serial Dependence. Water resources
410   research 20, 727–732. https://doi.org/10.1029/wr020i006p00727

Horgby, Å., Segatto, P.L., Bertuzzo, E., Lauerwald, R., Lehner, B., Ulseth, A.J., Vennemann, T.W., Battin, T.J., 2019. Unexpected large evasion fluxes of carbon dioxide from turbulent streams draining the world's mountains. Nature communications 10, 4888–9. https://doi.org/10.1038/s41467-019-12905-z

415
Hubbard, S.S., Varadharajan, C., Wu, Y., Wainwright, H., Dwivedi, D., 2020. Emerging technologies and radical collaboration to advance predictive understanding of watershed hydrobiogeochemistry. Hydrological processes 34, 3175–3182. https://doi.org/10.1002/hyp.13807

420   Kahl, J.S., Stoddard, J.L., Haeuber, R., Paulsen, S.G., Birnbaum, R., Deviney, F.A., Webb, J.R., DeWalle, D.R., Sharpe, W., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, P.S., Roy, K., Webster, K.E., Urquhart, N.S., 2004. Have U.S. surface waters responded to the 1990 Clean Air Act amendments? Environmental science & technology 38, 484A.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term
425   Memory (LSTM) networks. Hydrology and earth system sciences 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018
Lauerwald, R., Laruelle, G.G., Hartmann, J., Ciais, P., Regnier, P.A.G., 2015. Spatial patterns in CO2 evasion from the global river network. Global biogeochemical cycles 29, 534–554. https://doi.org/10.1002/2014GB004941

Li, L., Sullivan, P.L., Benettin, P., Cirpka, O.A., Bishop, K., Brantley, S.L., Knapp, J.L., van Meerveld, I., Rinaldo, A., Seibert,
430   J. and Wen, H., 2021. Toward catchment hydro-biogeochemical theories. Wiley Interdisciplinary Reviews: Water, 8(1).
https://doi.org/10.1002/wat2.1495

Martin, J.B., 2017. Carbonate minerals in the global carbon cycle. Chemical geology 449, 58–72.
https://doi.org/10.1016/j.chemgeo.2016.11.029

435

Mast, M.A., 2013. Evaluation of stream chemistry trends in US Geological Survey reference watersheds, 1970–2010.
Environmental monitoring and assessment 185, 9343–9359. https://doi.org/10.1007/s10661-013-3256-6

McHale, M.R., Burns, D.A., Siemion, J., Antidormi, M.R., 2017. The response of soil and stream chemistry to decreases in
440   acid deposition in the Catskill Mountains, New York, USA. Environmental pollution (1987) 229, 607–620.
https://doi.org/10.1016/j.envpol.2017.06.001

Melsen, L.A., Addor, N., Mizukami, N., Newman, A.J., Torfs, P.J.J.F., Clark, M.P., Uijlenhoet, R., Teuling, A.J., 2018.
Mapping (dis)agreement in hydrologic projections. Hydrology and earth system sciences 22, 1775–1791.
445   https://doi.org/10.5194/hess-22-1775-2018

Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H.V., Kumar, R., 2019. On the choice of
calibration metrics for "high-flow" estimation using hydrologic models. Hydrology and earth system sciences 23, 2601–2614.
https://doi.org/10.5194/hess-23-2601-2019

450

Monteith, D.T., Stoddard, J.L., Evans, C.D., de Wit, H.A., Forsius, M., Høgåsen, T., Wilander, A., Skjelkvåle, B.L., Jeffries,
D.S., Vuorenmaa, J., Keller, B., Kopácek, J., Vesely, J., 2007. Dissolved organic carbon trends resulting from changes in
atmospheric deposition chemistry. Nature (London) 450, 537–540. https://doi.org/10.1038/nature06316

455   Moosdorf, N., Hartmann, J., Dürr, H.H., 2010. Lithological composition of the North American continent and implications of
lithological map resolution for dissolved silica flux modeling. Geochemistry, geophysics, geosystems : G3 11, n/a-n/a.
https://doi.org/10.1029/2010GC003259

Murdoch, P.S., Shanley, J.B., 2006. Detection of water quality trends at high, median, and low flow in a Catskill Mountain
460   stream, New York, through a new statistical method. Water Resources Research 42, W08407-n/a.
https://doi.org/10.1029/2004WR003892

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

465

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2014. Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance. Hydrology and earth system sciences discussions 11, 5599–5631. https://doi.org/10.5194/hessd-11-5599-2014

470

Niu, X., Wen, T., Li, Z., Brantley, S.L., 2018. One Step toward Developing Knowledge from Numbers in Regional Analysis of Water Quality. Environmental science & technology 52, 3342–3343. https://doi.org/10.1021/acs.est.8b01035

Pentaho Data Integration - Kettle [WWW Document], n.d. URL https://community.hitachivantara.com/s/article/data-integration-kettle (accessed 11.22.20b).

475

Pool, S., Viviroli, D., Seibert, J., 2019. Value of a Limited Number of Discharge Observations for Improving Regionalization: A Large-Sample Study Across the United States. Water resources research 55, 363–377. https://doi.org/10.1029/2018wr023855

480

Postgres, P.G.D., 2020. PostgreSQL [WWW Document]. PostgreSQL. URL https://www.postgresql.org/ (accessed 11.22.20).

Siemion, J., McHale, M.R., Lawrence, G.B., Burns, D.A., Antidormi, M., 2018. Long-term Changes in Soil and Stream Chemistry across an Acid Deposition Gradient in the Northeastern United States. Journal of environmental quality 47, 410–418. https://doi.org/10.2134/jeq2017.08.0335

485

Underwood, K. L., Rizzo, D. M., Schroth, A. W., & Dewoolkar, M. M. (2017). Evaluating spatial variability in sediment and phosphorus concentration-discharge relationships using Bayesian inference and self-organizing maps. Water Resources Research, 53(12), 10293-10316. https://doi.org/10.1002/2017WR021353

490

U.S.G.S. 1994. USGS Water Data for the Nation. https://doi.org/10.5066/F7P55KJN

Zhi, W., & Li, L. 2020. The shallow and deep hypothesis: subsurface vertical chemical contrasts shape nitrate export patterns from different land uses. Environmental Science & Technology, 54(19), 11915-11928. https://doi.org/10.1021/acs.est.0c01340

495

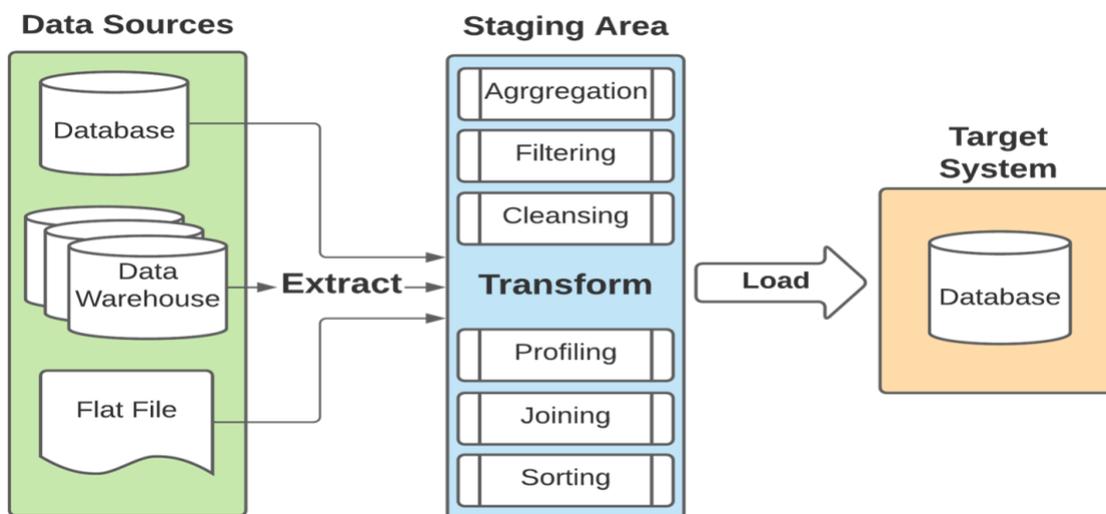



Figure 1: ETL process flow from source system to the data transformers and loading into target system.

500

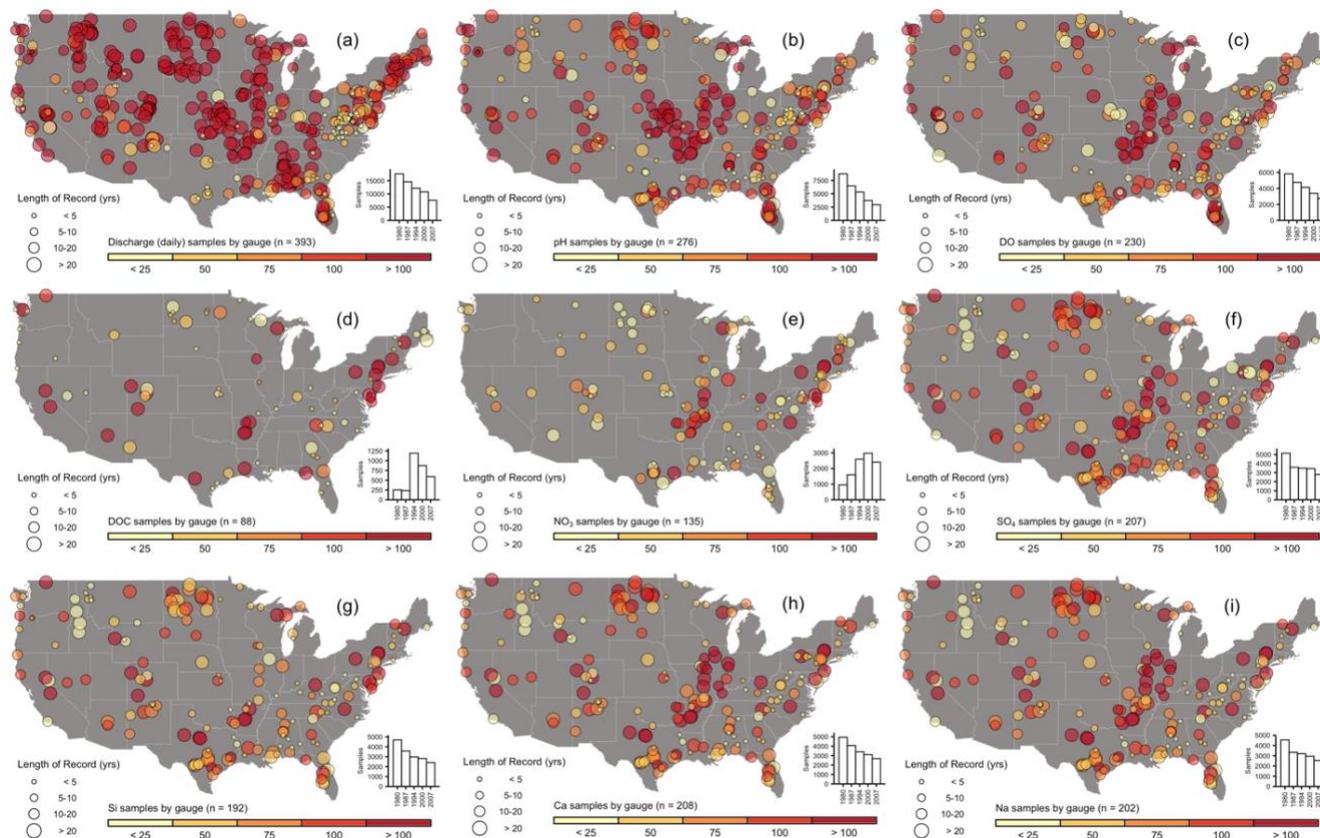



Figure 2: The number of samples (symbol color) and length or record (symbol size) for: (a) Daily Discharge, (b) pH, (c) DO, (d) DOC, (e) NO3, (f) SO₄, (g) Si, (h) Ca, (i) Na. The inset histogram shows the number of samples by roughly 7-year periods.

505

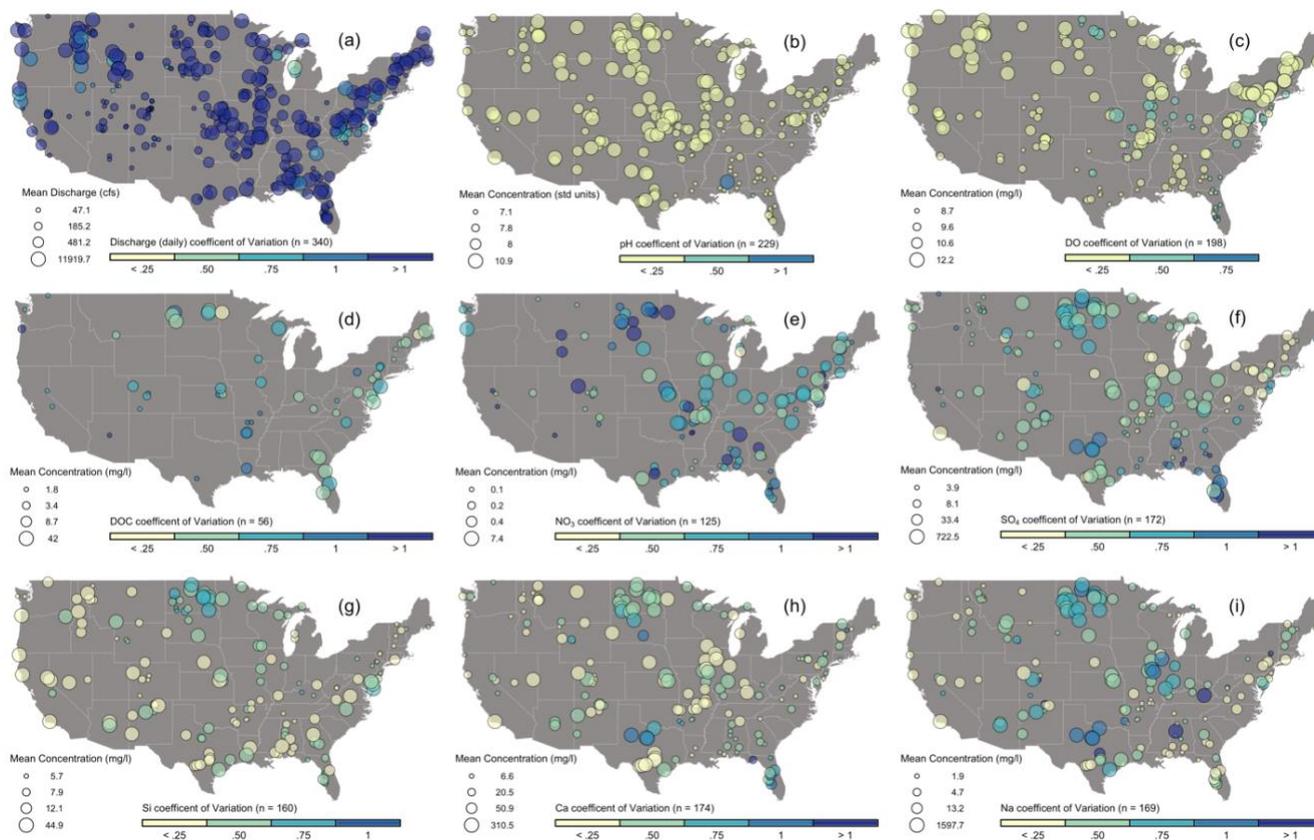



Figure 3: The mean concentration (symbol size) and coefficient of variation (symbol color) for (a) Daily Discharge, (b) pH, (c) DO, (d) DOC, (e) NO3, (f) SO₄, (g) Si, (h) Ca, and (i) Na.

510

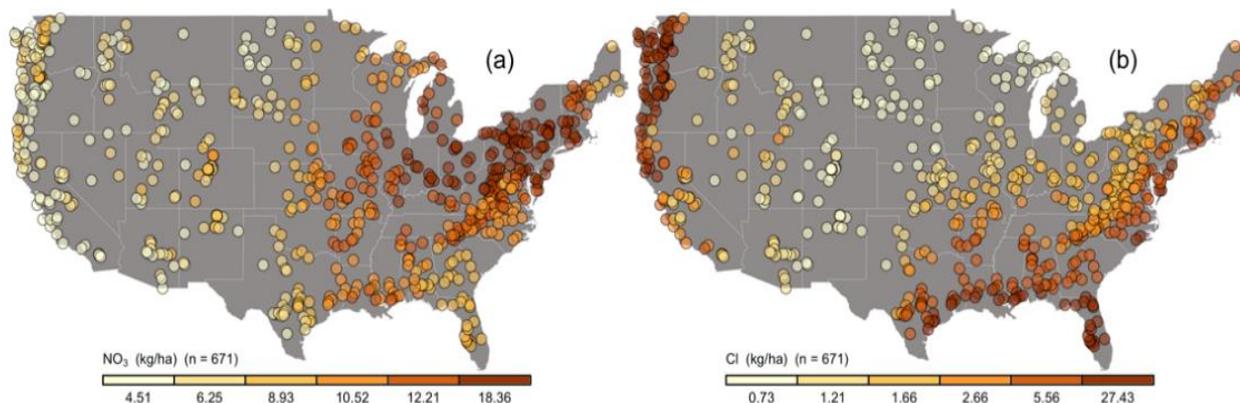



Figure 4: Average atmospheric wet deposition for (a) NO3, (b) Cl using data from the NADP network from 1981-2014.

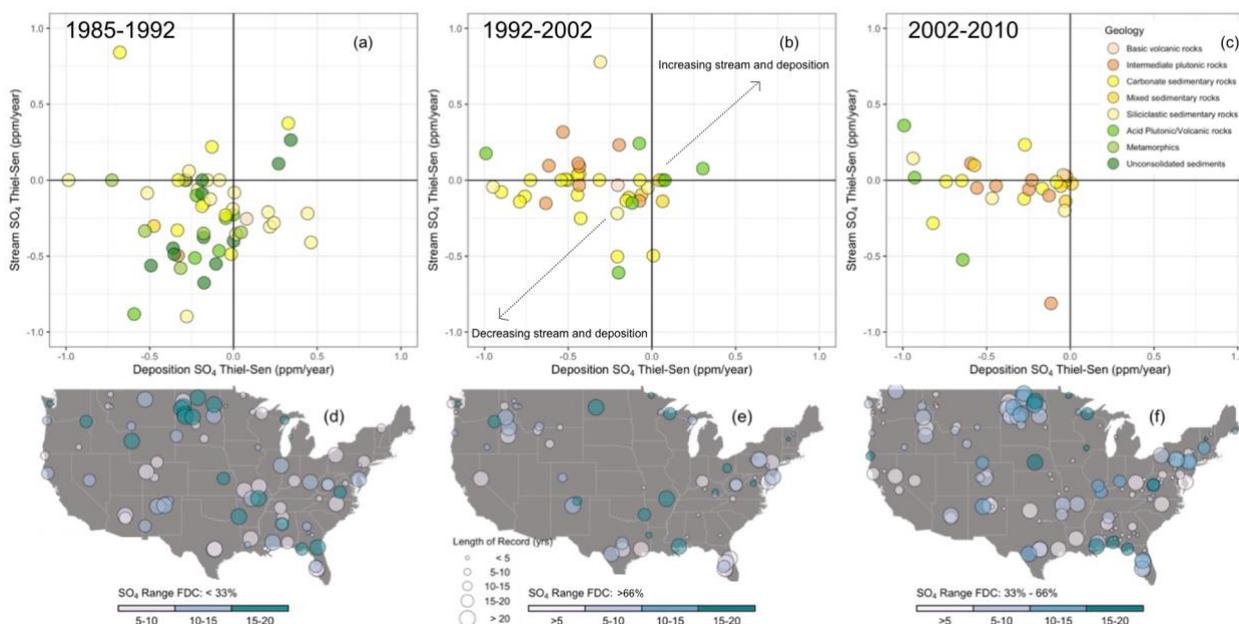



Figure 5: Long-term Sen slope (e.g. average trend) from Mann Kendall analysis of SO4 deposition (a) from 1985-1992, (b) 1992-2002, and (c) 2002-2010. Symbol color shows dominant geology. The maps show the corresponding number of streams SO4 samples for (d) higher flows (FDC <33%), (e), lower flows (FDC >66%), and (f) middle streamflow values (FDC 33%-66%).

Solutes derived from mineral weathering

Slopes of best fit lines from concentration-discharge

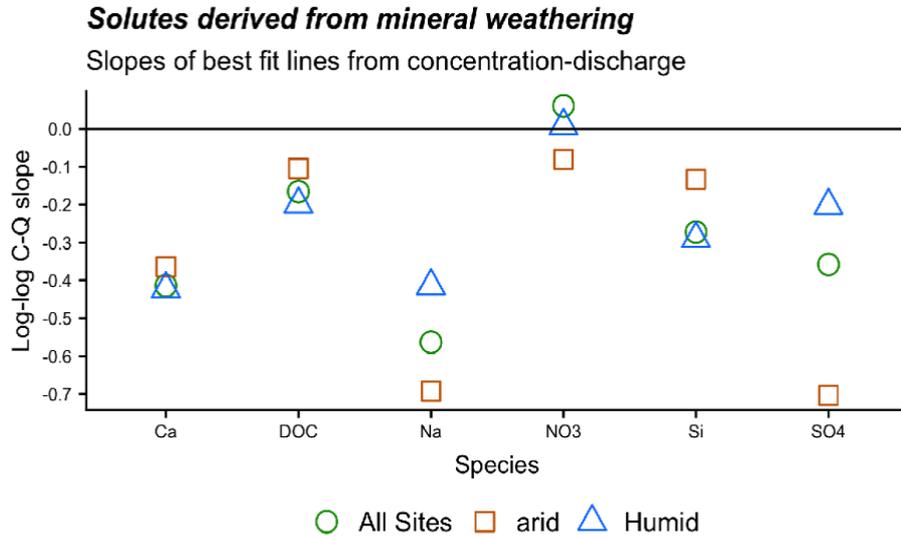



Figure 6: Power law slope of the event-scale concentration discharge relationships averaged across catchments based on aridity for common solutes.

525

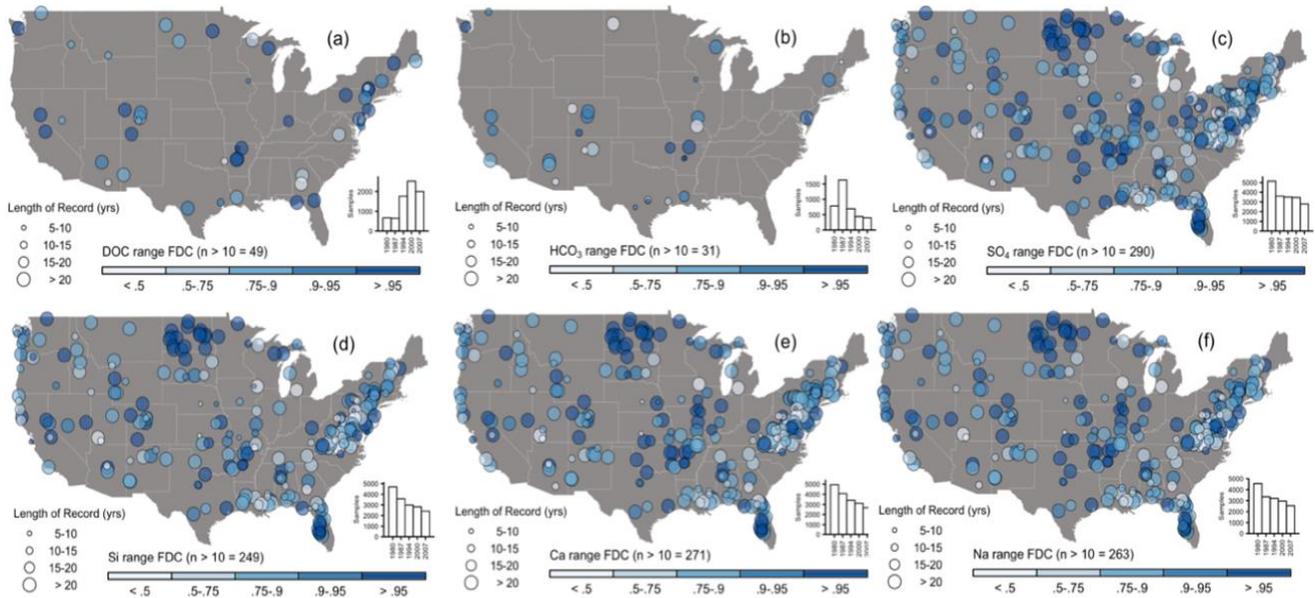



Figure 7: The fraction of the flow duration curve the water chemistry data spanned by the observed data (color, 1=observed high and low daily flow) and length of record (symbol size) for (a) DOC, (b) HCO3, (c) SO4, (d) Si, (e) Ca, and (f) Na.

530

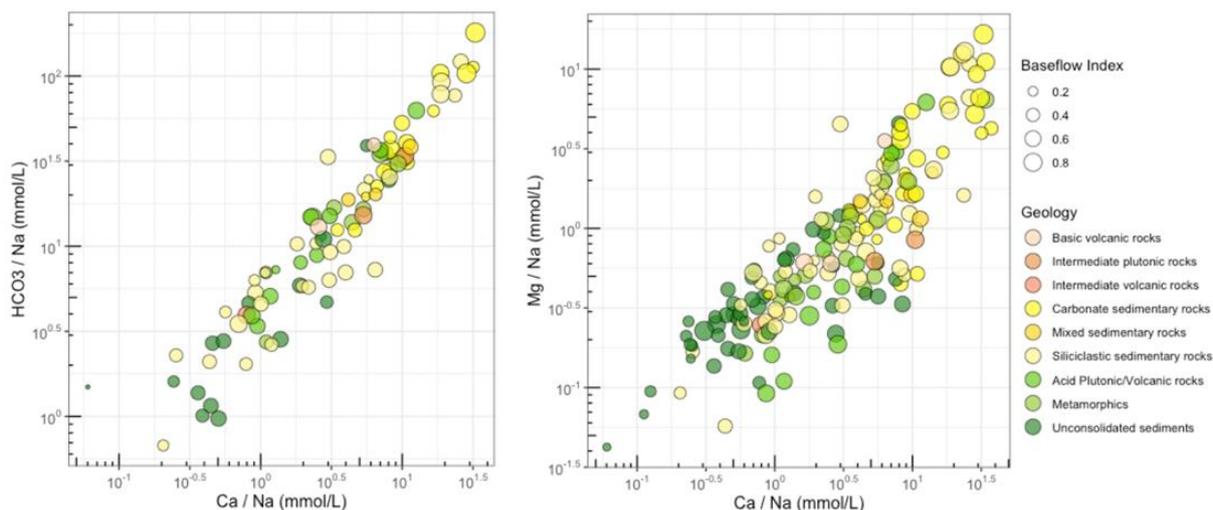
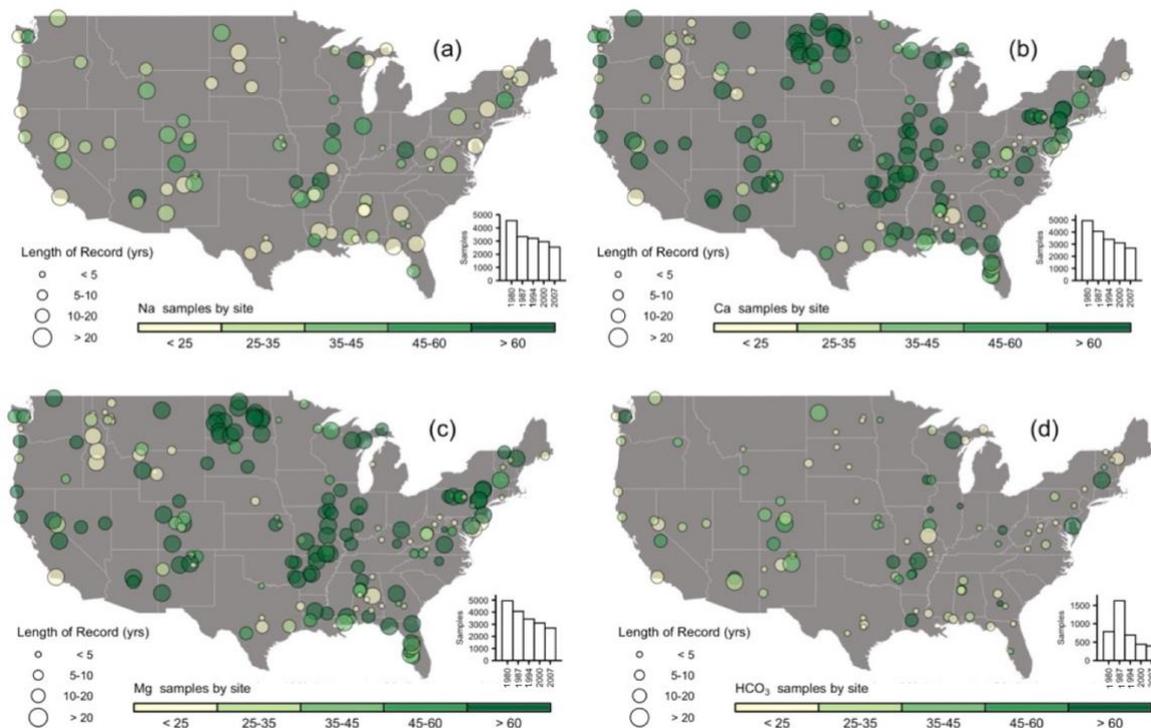



Figure 8: Molar ratios of HCO₃ and Mg to Na as a function of the molar ratio of Mg/Na using the FDC>66% (low flows). Symbol color is the dominant geology and symbol size is the baseflow index.

535 **Figure 9: The number of samples in the lower tercile of discharge values (symbol size) and their length of record (symbol color) for (a) Na, (b) Ca, (c) Mg and, (d) HCO₃.**