



OPEN

DATA DESCRIPTOR

A harmonized river-ocean coupled database for the northern Gulf of Mexico

Bailey Armos¹, Shuang Zhang¹✉, Tao Wen², Ellie Walker³ & Prabir Daripa⁴

The northern Gulf of Mexico (nGoM) receives water from over 50 rivers which are highly influenced by humans and include the largest river in the United States, the Mississippi River. To support large-scale data-driven research centered on the dynamic river-ocean system in the region, this study consolidated hydrogeochemical river and ocean data from across the nGoM. In particular, we harmonized 35 chemical solute parameters from 54 rivers and incorporated river discharge data to derive daily solute concentration and flux estimates throughout the nGoM. By integrating this river data with 17 ocean parameters, we generated a pre-processed and time-averaged River-Ocean coupled Database for the nGoM, known as ROcD-nGoM with the goals to streamline and enhance diverse research efforts in the nGoM, and also to showcase the value of making hydrological and oceanographic data FAIR (Findable, Accessible, Interoperable, Reusable). Moreover, the script developed in this study can be easily adapted for analyzing other chemical solutes and exploring other regions of interest.

Background & Summary

The Northern Gulf of Mexico (nGoM) holds significant ecological and economic importance. It spans coastlines of five United States (US) states (i.e., Florida, Alabama, Mississippi, Louisiana, Texas) and has witnessed a population increase of 164% from 1950 to 2000¹. This region is not only home to a thriving fishing industry but also accounts for approximately 97% of US oil and gas production that comes from the outer continental shelf (OCS)². Widespread anthropogenic influences in the nGoM have far-reaching effects on its ecosystem and environment. Among the ecological and environmental challenges faced by the nGoM, many of them stem from urbanization and human-induced inputs transported via rivers into the coastal waters. These anthropogenic nutrient inputs not only contribute to coastal eutrophication in the Gulf of Mexico but also pose a global threat to coastal ecosystems³. Moreover, the influx of alkalinity and other carbonate system species can influence the coastal ocean's susceptibility to acidification^{4,5}. Additionally, the release of heavy metals through anthropogenic activities has polluted estuaries surrounding the nGoM, such as the Grand Bay National Estuarine Reserve⁵. In the wake of challenges, the nGoM has still demonstrated remarkable resilience, evident in its recovery from the ecosystem-wide damage caused by the Deepwater Horizon oil spill in 2010^{6,7} as well as hurricanes such as Hurricane Katrina⁸.

One limitation in further understanding the nGoM's past, present, and future states is the accessibility and usability of both the ocean and river data surrounding this region. The Gulf of Mexico Coastal Ocean Observing System (GCOOS) offers on-demand information about the nGoM's coastal and open ocean waters, with different focuses including ecosystems and long-term changes⁹. However, researchers interested in utilizing GCOOS data face the substantial task of finding, downloading, exploring, harmonizing (the process of synthesizing different data fields, formats, dimensions, and columns into a single integrated and coherent database), and organizing the data to suit their specific projects. The United States Geological Survey (USGS) provides river chemical and discharge data for stations across the country, with the Water Quality Portal (WQP) serving as the database for publicly available water quality data^{10,11}. While the WQP hosts data from various sources, mainly the USGS, EPA, and USDA, it lacks proper assimilation of water quality data into a harmonized database, creating a gap between data availability and reusability^{11,12}. As a result, additional datasets such as the Standardized Nitrogen

¹Department of Oceanography, Texas A&M University, College Station, TX, 77843, USA. ²Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, 13244, USA. ³Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, VA, 23062, USA. ⁴Department of Mathematics, Texas A&M University, College Station, TX, 77843, USA. ✉e-mail: shuang-zhang@tamu.edu

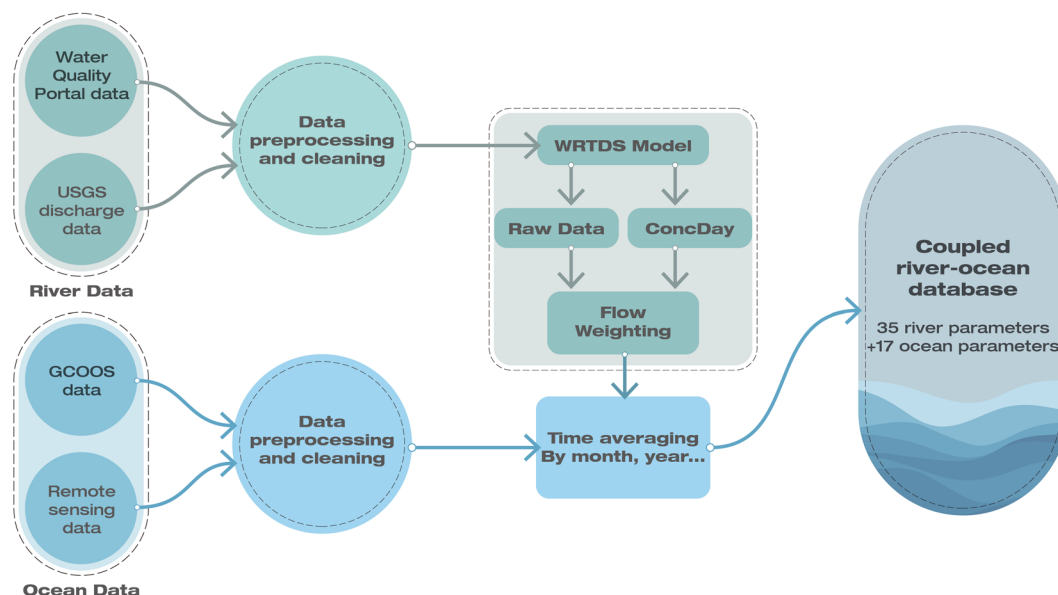


Fig. 1 The figure demonstrates the workflow used to create the ROcD-nGoM database by mining and compiling data from river and ocean datasets in the nGoM region. This process begins with initial data preprocessing and cleaning, followed by applying the WRTDS model to river data and flow-weighting the outputs. The resulting database combines raw and modelled ConcDay data, which are subsequently time-averaged to produce monthly and yearly averages. A detailed description of each step is provided in the Methods section.

and Phosphorus Dataset (SNAPD)¹³ and the River Chemistry for the U.S. Coast (RC4USCoast)¹⁴ have been developed to remove the barrier of utilizing WQP data. However, these two datasets are confined to river data with a limited range of parameters and lack integration with ocean data.

This study introduces a novel database that builds on enhanced processing steps for water quality data. These steps include pairing river chemistry data with discharge and running it through a regression model to estimate daily concentrations and fluxes of various parameters (Fig. 1). Moreover, the database harmonizes river and ocean data, with the potential of serving as a central hub for the nGoM research community. The primary objective of this database is to expedite and support diverse research related to the nGoM environmental and ecological conditions while highlighting the significance of improving reusability of publicly available hydrological and oceanographic data.

The database enables local, regional, and Gulf-wide analyses of crucial environmental and ecological parameters, facilitating the tracking of river fluxes to the nGoM and the assessment of its ecosystem health. It will benefit earth and environmental sciences researchers working on the nGoM region by significantly reducing the time required for data compilation and processing. Example research projects that will be facilitated by this database include (1) investigating nonpoint source pollution and its impact on coastal primary production, (2) analyzing time-series riverine carbonate chemistry and its influence on ocean acidification, (3) exploring the supply of essential and toxic metals to the coastal ocean (e.g., iron, lead, aluminum, arsenic, and selenium), (4) understanding the plausible impact of estuarine processes affecting the fluvial flux of materials to the coastal ocean, and (5) examining the fluxes of rock weathering in this region. Built on a robust framework for data integration and analysis, this database empowers researchers to conduct more comprehensive studies surrounding the nGoM region, ultimately advancing our understanding and management of the nGoM ecosystem.

Methods

River site selection. The Northern Gulf of Mexico watershed is widely known for high nutrient loading from both agricultural and industrial runoff as well as having a large human impact in general^{15,16}, yet there have been few regional studies. This study encompasses a total of 54 river sites, which are representative of various rivers, streams, and bayous that flow into the nGoM (Fig. 2, Table S1). The rivers in this analysis were selected based on two previous studies^{14,17} which encompasses about 87% of the drainage area to the nGoM from the continental US. Specifically, we selected the nearest USGS river monitoring sites along the nGoM coast that offer the longest available time-series data. The river sites were split into three regional groups with the West group representing sites in Texas, the East group representing sites in Florida, and the North group representing sites in Louisiana, Mississippi, and Alabama.

River data mining and analysis. To facilitate the data mining and formatting of all the USGS river data, a script was developed in the R programming language, utilizing the EGRET and dataRetrieval packages¹⁸. A total of 35 water quality parameters were collected from each river site, if available (Table 1: Step 1:2). In cases where the flag of “non-detects” were encountered for the parameter, they were quantified as half of the parameter’s

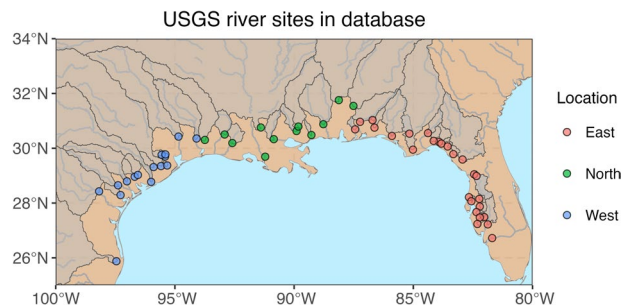


Fig. 2 USGS river sites included in the ROCd-nGoM database plotted with their corresponding watershed shapefiles. Note that some watersheds (e.g., the big Mississippi watershed) are only partially displayed due to the focus on the Gulf Coast region.

detection limit concentration¹⁸. However, if the units of a parameter could not be determined for the non-detect measurement due to multiple units in the sample data, the non-detect was retained with its initial NA value from the WQP portal (Table 1: **Step 3:4**).

To ensure data integrity, several cleaning steps were implemented. Samples taken from mediums other than water, estimated samples instead of measured ones, quality control replicates, and samples without an associated unit were all removed from the database (Table 1: **Steps 5:10**).

Harmonization of units was performed, with all parameters (excluding temperature) standardized to mg/l (Table 1: **Step 11**). Nitrogen data was converted to mg/l as total N to simplify analysis and create the longest possible time series by combining data with different units. Similarly, phosphorus data was converted to mg/l as total P. For alkalinity, the USGS conducted a study that found no statistical difference between 123 pairs of filtered and unfiltered alkalinity samples in the Red River¹⁹. This finding has also been observed in open ocean waters²⁰. Therefore, filtered and unfiltered alkalinity measurements were merged to generate a longer time series. Each parameter was further categorized by fraction (e.g., dissolved, total, suspended) (Table 1: **Step 12**).

To estimate the daily concentration and flux for each parameter, each river sample with that specific parameter was paired with discharge data (Table 1: **Steps 13:15**). Discharge data were mined using the EGRET package. Three sites that did not have discharge data from the USGS station were matched with the closest (within 22 km) discharge station on the same river with sufficient data (Table S2). In addition, the discharge for the Mississippi and Atchafalaya Rivers was downloaded from the Army Corps of Engineers (<https://www.usace.army.mil/>), and the discharge for the Rio Grande River was downloaded from the International Boundary and Water Commission (<https://www.ibwc.gov/>), as there were no USGS long-term daily discharge data reported for these river outlets. Subsequently, the parameter, daily discharge, and site information were fed into the Weighted Regression on Time, Discharge, and Season (WRTDS) model (Table 1: **Step 16**). This weighted regression approach allows the coefficients of Eq. 1 (see below) to vary across the calibration period and over the range of streamflow values. An equation is fitted for each day in the calibration period, with observations weighted based on their similarity in terms of time, season, and discharge to the day being calibrated²¹. The minimum number of observations (minNumObs) was set to 50 which is the threshold for the reliability of the regression, while the minimum number of uncensored observations (minNumUncen) was set to half of the minimum number of observations¹⁸. However, for parameters with a limited range of data, such as observations from 1997–1999, the number of observations was insufficient to run the regression. In such cases, adjustments were made to the model settings for the main nutrients (silica, mixed nitrogen, or phosphorus). Specifically, the settings were adjusted to minNumObs = 20 and minNumUncen = 10. These adjustments are detailed in Table S3.

$$\ln(c) = \beta_0 + \beta_1 t + \beta_2 \ln(Q) + \beta_3 \sin(2\pi t) + \beta_4 \cos(2\pi t) + \epsilon \quad (1)$$

Equation 1. *Weighted regression model for estimating daily concentration where c is concentration, β_n are fitted coefficients, t is time, Q is mean daily discharge, and ϵ is the unexplained residual.*

The WRTDS model was used to generate estimated continuous daily concentrations (ConcDay) and fluxes (FluxDay). One of the advantages of the WRTDS model is its ability to generate flow-normalized data, effectively eliminating random year-to-year variations in discharge from the sample concentrations. In addition to generating continuous records with the WRTDS model, we also maintained the raw daily average concentration, denoted as “ConcAve”. This metric represents the measured average concentration for each day, typically assessed at weekly to monthly intervals. We then calculated the raw daily fluxes, “FluxAve”, by multiplying “ConcAve” by the daily discharge (Q) as described in **Step 17**. We further manually weighted the concentration data by discharge to calculate a flow-weighted mean concentration per month, season, and year (Eq. 2) (Table 1: **Step 18:21**). This is a useful method especially for time-series analysis because it directly accounts for the influence of stream inflow which can significantly vary over different time scales.

$$[Conc_{FN}] = \frac{\sum_{i=1}^n ([Conc_i] * Q_i)}{\sum_{i=1}^n Q_i} \quad (2)$$

Harmonization step		Details
Step 1	Read in WQP data	readWQPqw() function
Step 2	Read in Sample format	readWQPsample(), this is the format for running WRTDS model but lacks important metadata
Step 3	Identify units in data	Reporting units can change due to changing protocol, methods...
Step 4	Non-detects	-Identifying if non-detects are in data -readWQPqw() has a flag for non-detects, but no unit or detection limit info for non-detects -readWQPsample() has detection limit info for samples, but no unit info -If there is only one unit in readWQPqw, non-detects are kept with flag -If there are multiple units and units for non-detect are unknown, these values were dropped
Step 5	Result Type	Samples that were estimated instead of measured were dropped
Step 6	Medium	Samples that were from any other media than water were dropped
Step 7	Activity type	Samples that were for quality control replication were dropped
Step 8	Units	If unit information was missing or could not be converted to mg/l, sample was dropped
Step 9	Fraction	Identifying if sample has distinct fractions (filtered/unfiltered) in data, samples from bed sediment were dropped
Step 10	Assigning fraction info	Temperature data assigned fraction = None Total and dissolved fraction combined for alkalinity
Step 11	Conversions	Units converted to mg/l
Step 12	Separate fractions	Each fraction of data run through WRTDS model separately
Step 13	Calculate daily concentration	WRTDS takes daily concentration, some sites have multiple samples per day
Step 14	Read in daily discharge	readNWISDaily(), within sample
Step 15	Merge report	eList with INFO, Daily, and Sample data
Step 16	Run modelEstimation	Run WRTDS model to derive continuous daily concentration
Step 17	Calculate FluxAve	FluxAve (kg/day) = ConcAve*Q*86.4
Step 18	Add metadata	Add data source, site number, parameter, unit, fraction, drainage area, latitude and longitude to dataframe
Step 19	Add time columns	Add columns for year, month, yearMonth, Season, yearSeason
Step 20	Time-averaging	Averaging parameters per different time range
Step 21	Flow Weighting	Manual flow weighting of ConcAve and ConcDay averaged by different time ranges

Table 1. Summary of river database harmonization.

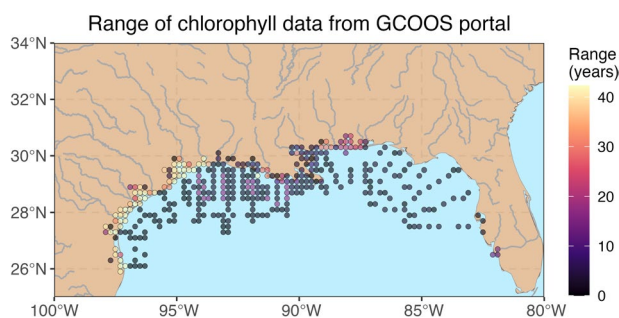


Fig. 3 Spatial distribution of GCOOS Chl-a data. The color bar shows the range of years for the available data.

Equation 2. *Manual Flow weighted (normalized) equation. Q is the discharge in cubic meters per second and Conc is concentration in mg/l.*

GCOOS data. GCOOS is a component of the larger Integrated Ocean Observing System (IOOS) in the U.S. It encompasses a vast network of over 2,000 sensors deployed by numerous data providers, e.g. universities, state and federal government, and energy/oil companies. The GCOOS nutrient portal (<https://wq.gcoos.org/nutrients/>) serves as a repository for data from 80 organizations, including Chl-a (Fig. 3), oxygen, nutrients, pH, salinity, and water temperature. Additionally, GCOOS provides data on currents, wind, sea surface height, turbidity, and wave characteristics. Similar to the river data, the GCOOS data is organized based on the five-time columns as seen in Table 1: **Step 19**. To ensure consistency, the data was cleaned and standardized to match the format of the river data. All units of chemical species were converted to mg/l, particularly for chlorophyll and other chemical measurements. GCOOS also provides acceptable ranges of values for most water quality parameters (https://data.gcoos.org/certification/GCOOS_DMAC_DMS.pdf), which were utilized to exclude concentrations or values that fell outside the acceptable range. Our analysis focused on surface water, specifically the top 100 meters of the water column, while samples from deep water were excluded from consideration.

Satellite data. Satellite remote sensing is a widely used technique in oceanography for estimating chlorophyll levels. The SeaWiFS instrument on the OrbView-2 satellite offers data records spanning approximately 13 years, from 1997-09-04 to 2010-12-11. With eight spectral bands ranging from 412 to 865 nm, this instrument collects data at a spatial resolution of 4 kilometers with a revisit time of 1 day. For this project, monthly stage 3 mapped data (in the NetCDF format) with a spatial resolution of 9 kilometers was obtained from the NASA Ocean Color portal (<https://oceancolor.gsfc.nasa.gov/>) for the Gulf of Mexico region. Another source of remotely sensed chlorophyll data is the MODIS instrument on the Aqua satellite, which was launched in 2002²². This instrument provides data records from 2002-07-04 to the present, covering approximately 20 years. The MODIS instrument captures data across 36 spectral bands with spatial resolutions of 250 m, 500 m, and 1000 m, depending on the spectral band. Similarly, for this project, monthly stage 3 mapped data (in the NetCDF format) with a spatial resolution of 4 kilometers was acquired from the NASA Ocean Color portal for the Gulf of Mexico region. To extract chlorophyll data from NetCDF files, an R script was adapted²³. After cleaning, the extracted data from satellite sources was standardized to match the format of the river data and GCOOS data. The units were converted to mg/l to align with the units used for the GCOOS chlorophyll data.

Data Records

The final database, ROCD-nGoM, is publicly available on Zenodo²⁴ with a Creative Commons Attribution 4.0 International (CC BY 4.0) license. A README file provides guidance and information for the database. The data is provided in two formats: (1) raw with outlier flags and (2) cleaned and time averaged. The first raw format includes the “river_flagged” and “ocean_flagged” folders, and the second clean format includes the “river_time_averaged” and “ocean_time_averaged” folders. In the raw format, daily averaged concentrations are provided for the USGS and GCOOS data, while the satellite product offers monthly data. Two specific flags are employed: “outlier_99.5” for concentrations exceeding the 99.5th percentile, and “outlier_0.5” for concentrations falling below the 0.5th percentile. These percentiles are calculated across all sites for each respective parameter. For the USGS river data, outlier flags are applied to “ConcAve,” “ConcDay,” and “Q.” The range designation for “ConcDay” is determined based on the distribution of the actual raw measurements (“ConcAve”). In the case of ocean data, outlier flags are assigned to each parameter except for wind/current direction, which ranges from 0 to 360 degrees.

Table 2 presents the time range and number of observations for the raw river data, while Table 3 provides the same details for the ocean parameters.

The following variables are included in the river_flagged folder:

- **data_source:** source of the water quality data.
- **siteNumber:** the site number of the USGS site the data was collected from.
- **latitude:** latitude in decimal degrees format.
- **longitude:** longitude in decimal degrees format.
- **drainage_area:** the area of the drainage basin in km².
- **parameter:** the water quality parameter name.
- **conc_unit:** the concentration unit which was converted to mg/l except for temperature.
- **fraction:** filtration status of sample.
- **Date:** date the sample was collected (format: YYYY-MM-DD).
- **Month:** month of the year that the sample was collected (format: MM).
- **Season:** season of the year that the sample was collected. Each season is indicated by each number, with number 3 representing months 3,4, and 5 (spring), 6 representing months 6,7, and 8 (summer), 9 representing months 9,10, and 11 (fall), and 12 representing months 12,1, and 2 (winter). (format: MM).
- **Year:** calendar year the sample was collected (format: YYYY).
- **yearMonth:** unique year + month combination (format: YYYY-MM-01). 01 is a placeholder for the day.
- **yearSeason:** unique year + season combination (format: YYYY-MM-01). 01 is a placeholder for the day.
- **Q:** daily river discharge in units m³/s
- **ConcAve:** the raw daily average concentration measurement of parameter.
- **ConcDay:** the WRTDS model daily concentration estimate of parameter.
- **FluxAve:** the ConcAve * Q in units of kg/day.
- **FluxDay:** the WRTDS daily flux estimate in units of kg/day.
- **Uncen:** represents the detection limit. Uncen = 1 means the concentration is above the detection limit, while Uncen = 0 means the concentration is below the detection limit (the concentration then is estimated as half of the detection limit).
- **outlier_99.5_concAve:** outlier flag for ConcAve concentration. If the value is above the 99.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.
- **outlier_99.5_concDay:** outlier flag for ConcDay concentration based on distribution of raw ConcDay values. If the value is above the 99.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.
- **outlier_99.5_Q:** outlier flag for daily Q. If the value is above the 99.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.
- **outlier_0.5_concAve:** outlier flag for ConcAve concentration. If the value is below the 0.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.
- **outlier_0.5_concDay:** outlier flag for ConcDay concentration based on distribution of raw ConcDay values. If the value is below the 0.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.

Parameter	Fraction	Time Range	# of ConcAve observations	# of ConcDay estimates
Temperature, water	None	1939–2021	12,034	780,448
Temperature, air	None	1969–2021	1,265	118,709
Oxygen	Dissolved	1965–2021	9,867	664,441
H ion	Total	1939–2021	14,985	784,763
Alkalinity	Combined	1931–2021	13,046	678,341
Carbon	Total	1971–1978	122	6,335
Carbon	Suspended	1966–2021	1,323	94,892
Organic carbon	Total	1968–2021	7,917	711,198
Organic carbon	Dissolved	1966–2021	2,880	266,522
Organic carbon	Suspended	1966–2021	1,555	159,132
Inorganic carbon	Total	1971–1978	67	3,550
Inorganic carbon	Dissolved	2018–2021	57	1,243
Inorganic carbon	Suspended	1969–2021	1,952	135,448
Carbon dioxide	Dissolved	1944–2021	13,563	671,243
Bicarbonate	Total	1931–2021	14,924	729,840
Bicarbonate	Dissolved	1966–2021	2,233	155,438
Carbonate	Dissolved	1966–2021	351	26,561
Mixed nitrogen	Total	1966–2021	14,308	943,673
Mixed nitrogen	Dissolved	1966–2021	2,488	234,288
Organic nitrogen	Total	1966–2021	16,083	1,051,439
Organic nitrogen	Dissolved	1966–2021	2,430	229,907
Nitrate + Nitrite	Total	1966–2021	7,882	585,068
Nitrate + Nitrite	Dissolved	1966–2021	4,451	326,492
Kjeldahl nitrogen	Total	1966–2021	21,391	1,271,620
Kjeldahl nitrogen	Dissolved	1966–2021	6,107	403,969
Kjeldahl nitrogen	Suspended	1979–1981	22	940
Ammonia + Ammonium	Total	1966–2021	8,651	663,683
Ammonia + Ammonium	Dissolved	1966–2021	4,705	489,685
Nitrate	Total	1950–2021	5,850	390,608
Nitrate	Dissolved	1931–2021	9,952	841,209
Nitrite	Total	1966–2021	4,934	445,358
Nitrite	Dissolved	1939–2021	3,393	339,303
Phosphorus	Total	1966–2021	16,184	994,499
Phosphorus	Dissolved	1966–2021	6,445	415,793
Orthophosphate	Total	1963–2021	3,117	273,300
Orthophosphate	Dissolved	1939–2021	7,070	538,502
Silica	Dissolved	1944–2021	13,392	681,384
Calcium	Dissolved	1931–2021	13,458	665,822
Magnesium	Dissolved	1944–2021	13,789	678,403
Sodium	Dissolved	1950–2021	12,041	654,578
Potassium	Dissolved	1950–2021	10,198	636,726
Chloride	Dissolved	1931–2021	15,142	713,021
Sulfate	Dissolved	1931–2021	14,788	727,231
Iron	Total	1952–1973	644	30,808
Iron	Dissolved	1950–2021	14,912	1,236,904
Iron	Recoverable	1969–2021	1076	188,961
Lead	Suspended	1969–2021	471	115,614
Lead	Recoverable	1969–2016	173	38,538
Aluminum	Dissolved	1963–2021	897	131,097
Fluoride	Dissolved	1939–2021	11,415	624,546
Arsenic	Total	1969–2021	1,220	119,746
Arsenic	Dissolved	1966–2021	8,006	697,097
Arsenic	Suspended	1969–2021	322	57,834
Boron	Dissolved	1966–2021	1,355	96,100
Selenium	Total	1969–2021	156	44,090
Selenium	Dissolved	1966–2021	1,283	146,183
Total			378,342	24,523,351

Table 2. River parameter summary for the database.

Parameter	Source	Time range	# of observations
Chlorophyll	GCOOS	1972–2021	48,333
Chlorophyll	MODIS	2002–2021	9,826,629
Chlorophyll	seaWiFS	1997–2010	1,724,074
Oxygen	GCOOS	1922–2021	396,270
Phosphorus	GCOOS	1968–2014	28,917
Nitrogen	GCOOS	1969–2013	931
Nitrate	GCOOS	1972–2014	14,277
Nitrite	GCOOS	1968–2014	9,929
Nitrate + Nitrite	GCOOS	1969–2014	4,289
Wave	GCOOS	2008–2019	819,601
Turbidity	GCOOS	2008–2021	48,006
Sea surface height	GCOOS	2008–2019	206,607
Wind	GCOOS	2001–2021	365,600
Current	GCOOS	1995–2021	69,004
Temperature	GCOOS	1956–2021	369,913
pH	GCOOS	1962–2014	398,085
Salinity	GCOOS	1922–2021	338,261

Table 3. Ocean parameter summary for the database.

- **outlier_0.5_Q:** outlier flag for daily Q. If the value is below the 0.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.

The following unique variable groups (not included in previous description) are included in the river_time_averaged folder:

- **Q_yearMonth, Q_yearSeason, Q_year, Q_month, Q_Season:** The daily discharge averaged by yearMonth, yearSeason, year, month, and season for each site.
- **ConcDay_yearMonth, ConcDay_yearSeason, ConcDay_year, ConcDay_month, ConcDay_season:** the WRTDS model daily concentration averaged by yearMonth, yearSeason, year, month, and season for each site.
- **FluxDay_yearMonth, FluxDay_yearSeason, FluxDay_year, FluxDay_month, FluxDay_season:** the WRTDS model daily flux averaged by yearMonth, yearSeason, year, month, and season for each site.
- **FN_ConcDay_yearMonth, FN_ConcDay_yearSeason, FN_ConcDay_year, FN_ConcDay_month, FN_ConcDay_season:** the WRTDS model daily concentration averaged and flow weighted by yearMonth, yearSeason, year, month, and season for each site.
- **ConcAve_yearMonth, ConcAve_yearSeason, ConcAve_year, ConcAve_month, ConcAve_season:** the raw daily concentration averaged by yearMonth, yearSeason, year, month, and season for each site.
- **ConcAve_counts_yearMonth, ConcAve_counts_yearSeason, ConcAve_counts_year, ConcAve_counts_month, ConcAve_counts_season:** counts of the number of observations of the raw daily concentration in each time average for each site. This is included because some time-averages are only based on one or two observations to represent a whole month, season, or year.
- **FluxAve_yearMonth, FluxAve_yearSeason, FluxAve_year, FluxAve_month, FluxAve_season:** the ConcAve * Q in units of kg/day, averaged by yearMonth, yearSeason, year, month, and season for each site.
- **FN_ConcAve_yearMonth, FN_ConcAve_yearSeason, FN_ConcAve_year, FN_ConcAve_month, FN_ConcAve_season:** the raw daily concentration averaged and flow weighted by yearMonth, yearSeason, year, month, and season for each site.

The following variables are included in the ocean_flagged folder (excluding variables previously introduced):

- **conc_yearMonth:** for the satellite data, this is the raw observation provided by NASA in yearMonth.
- **conc_daily:** for the GCOOS data, this is the daily mean of the raw observations.
- **direction_daily:** both ocean currents and winds have multiple parameters associated with them. They are given separate titles other than conc_daily in order to distinguish them. Direction is provided in units of degrees (0–359° N).
- **speed_daily:** For both daily current and wind speed. Current speed in units (cm/s) while wind speed in units (m/s).
- **upwell_daily:** This parameter is only in the currents file and represents vertical currents in units (cm/s).
- **outlier_99.5:** outlier flag for concentration. If the value is above the 99.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.
- **outlier_0.5:** outlier flag for concentration. If the value is below the 0.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.

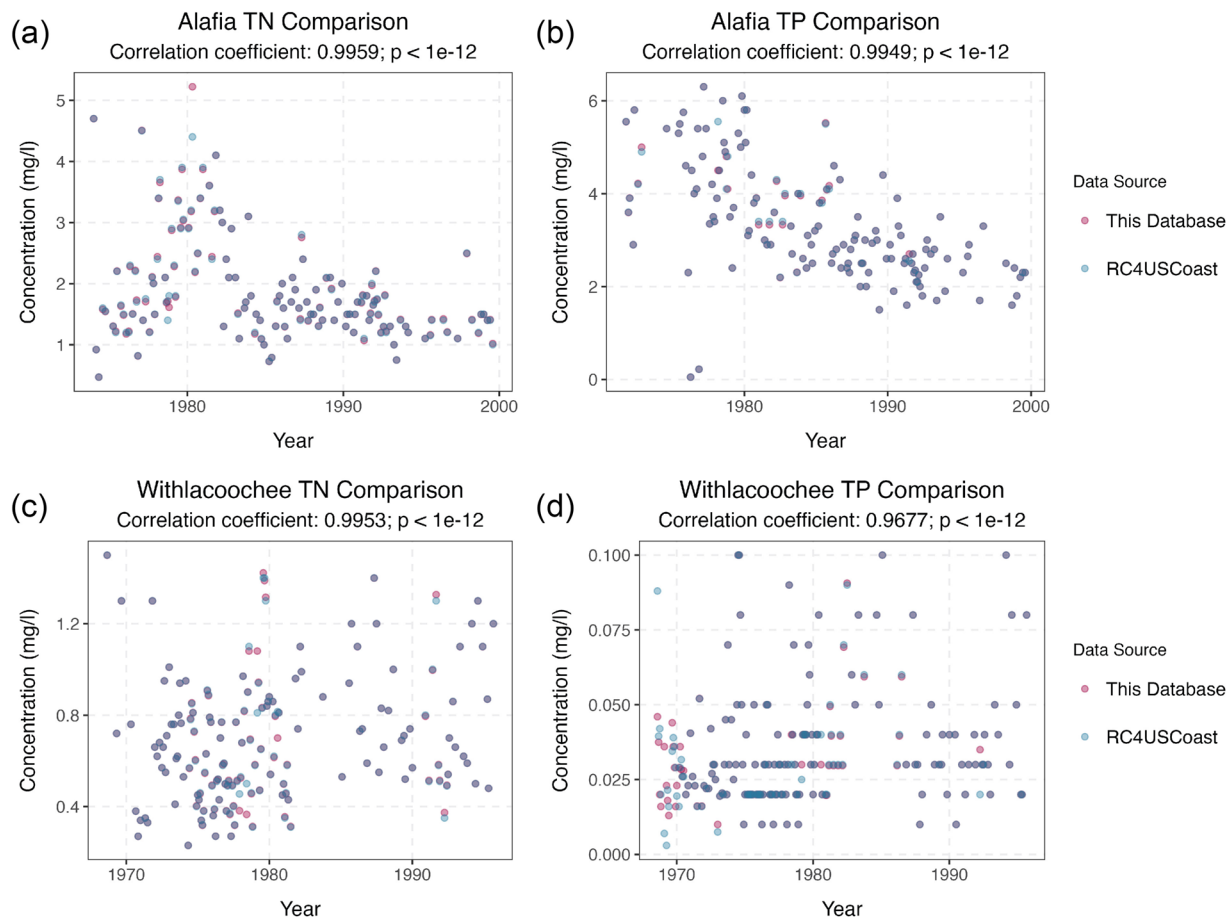


Fig. 4 ROcD-nGoM (this Database) and RC4USCoast data comparison. Year-month unfiltered raw TN and TP comparison for two sites: (a) Alafia River TN, (b) Alafia River TP, (c) Withlacoochee River TN, (d) Withlacoochee River TP. The data from our ROcD-nGoM are shown in red, and the data from RC4USCoast are shown in blue. The purple colour indicates the data points that overlap between the two databases.

- **outlier_99.5_speed, outlier_0.5_speed, outlier_99.5_upwell, outlier_0.5_upwell:** outlier flag columns for speed and upwelling. If the value is above the 99.5th percentile, the flag column will say “Outlier”, if it is not an outlier, the column is left blank.
- **network:** this column represents the original data source of the data provided to GCOOS.
- **platform:** unique identifiable platform from which the measurement was collected.

The following unique variable groups (not included in previous description) are included in the ocean_time_averaged folder:

- **conc_yearMonth, conc_yearSeason, conc_year, conc_month, conc_season:** the raw daily concentration of each parameter averaged by yearMonth, yearSeason, year, month, and season for each site.
- **speed_yearMonth, speed_yearSeason, speed_year, speed_month, speed_season:** the raw daily current (or wind) speed averaged by yearMonth, yearSeason, year, month, and season for each site.
- **direction_yearMonth, direction_yearSeason, direction_year, direction_month, direction_season:** the raw daily current (or wind) direction averaged by yearMonth, yearSeason, year, month, and season for each site.
- **upwell_yearMonth, upwell_yearSeason, upwell_year, upwell_month, upwell_season:** the raw daily upwell speed averaged by yearMonth, yearSeason, year, month, and season for each site.

Technical Validation

Raw data (ConcAve) validation. Random river sites (Alafia River and Withlacoochee River) from both this database and the existing RC4USCoast database¹⁴ were chosen for comparison. Total unfiltered nitrogen and phosphorus from the RC4USCoast database were converted to mg/l before comparison. For pairwise analysis, we merged the two datasets by aligning entries from the same year and month. Figure 4 shows the database comparison with ROcD-nGoM (this study) in red and RC4USCoast in blue. Overlapping points are shown in purple. Even with added uncertainties for converting units, both datasets visually display a notable agreement, indicating that our data mining and harmonization is accurate. The correlation analysis further revealed a high correlation

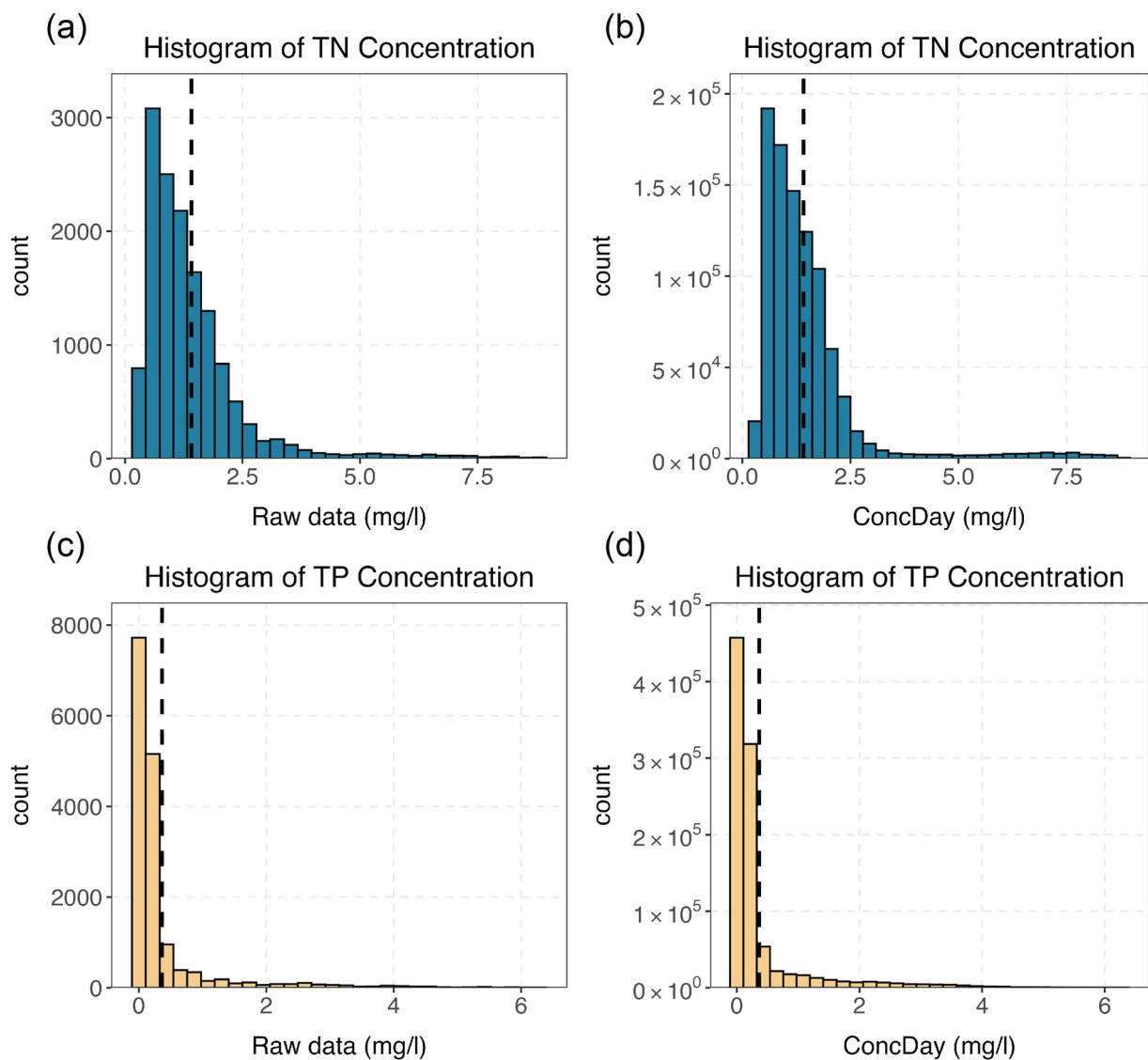


Fig. 5 Histograms of raw data (ConcAve) and ConcDay of nutrient parameters. **(a)** raw TN data, **(b)** TN ConcDay, **(c)** raw TP data, **(d)** TP ConcDay. The vertical dashed lines represent the mean concentration in mg/l.

between the two datasets for both total nitrogen and phosphorus at the random river sites. The correlation coefficient (r) exceeds 0.96 with a p -value smaller than $1e-12$ across all scenarios (Fig. 4), confirming the alignment of our compiled data with the RC4USCoast database.

ConcDay validation. In order to ensure the reliability of the ConcDay estimates, the distributions of the raw observations (ConcAve) and daily estimates (ConcDay) for all of the data across all of the river sites are provided along with time-series plots. Figure 5 shows the distributions of the unfiltered TN and TP concentration across all of the sites. The ConcDay estimated concentration follows the same distribution as the ConcAve raw data (Fig. 5). The ConcDay of TN demonstrates a distribution of 0.23 mg/L at the 0% quantile, 1.18 mg/L at the 50% quantile, and 8.75 mg/L at the 100% quantile. This distribution closely mirrors that of the ConcAve of TN, which shows 0.23 mg/L at the 0% quantile, 1.11 mg/L at the 50% quantile, and 8.75 mg/L at the 100% quantile. Similarly, ConcDay of TP demonstrates a distribution of 0.01 mg/L at the 0% quantile, 0.12 mg/L at the 50% quantile, and 6.30 mg/L at the 100% quantile. This distribution also closely mirrors that of the ConcAve of TP, which shows 0.01 mg/L at the 0% quantile, 0.11 mg/L at the 50% quantile, and 6.30 mg/L at the 100% quantile. The capability of the modelled ConcDay to capture seasonal cycles was assessed by comparing its monthly averaged value with monthly averaged river raw data (ConcAve). Figure 6 shows the seasonal difference between the raw data (ConcAve) and ConcDay in the TN and TP data. Seasonal and interannual variations are captured by both the raw data the modelled ConcDay data. Notice the larger fluctuations with the monthly raw data (ConcAve) compared to the monthly modelled data (ConcDay). This observation aligns with the smooth nature of the WRTDS model, which effectively eliminates random year-to-year variations in discharge from the sample concentrations. Accordingly, there should be a correlation between ConcAve and ConcDay, though the correlation coefficient

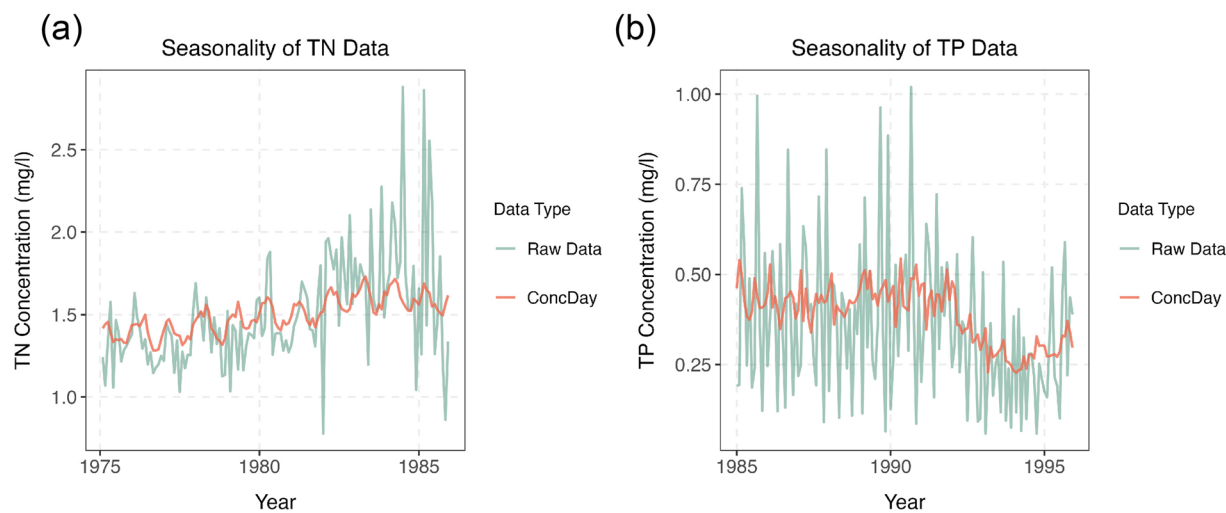


Fig. 6 Average monthly comparison of ConcAve and ConcDay. **(a)** example timeframe with TN data, **(b)** example timeframe with TP data.

is not expected to reach 1. For the TN and TP examples, the correlation analysis between the two variables, ConcAve and ConcDay, revealed a strong correlation for nitrogen ($r = 0.57$; p -value $< 1e-10$) and a weaker, yet statistically significant, correlation for phosphorus ($r = 0.27$; p -value = 0.002).

Limitations and directions. It is important to note that the WRTDS model's reliability decreases for sample sizes lower than 50 observations¹⁸. In this analysis, certain sites did not meet the 50-observation threshold, leading to adjustments in the model settings. These specific sites where the model settings were adjusted were identified in the supplementary info (see Table S1) and users can use at their own discretion. Another important consideration is the handling of non-detects in the database. Dealing with non-detects is a challenging task, and while there is no definitive approach, the standard practice in environmental science is to utilize half of the detection limit²⁵. The `readWQPsample()` function provided information on the detection limits, allowing for the inclusion of non-detects in the WRTDS model whenever possible. Our analysis reveals that the majority of parameters (88%) in our compiled database have fewer than 10% of observations below the detection limit, underscoring the completeness and reliability of our database (Table S4). Specifically, essential river parameters including alkalinity, pH, bicarbonate, calcium, magnesium, sodium, potassium, sulfate, chloride, kjeldahl nitrogen, mixed nitrogen, nitrate, nitrite, orthophosphate, and phosphorus all show fewer than 10% of observations below this threshold. When analyzing data of certain elements such as suspended inorganic carbon, dissolved aluminum, and dissolved selenium, it is crucial to consider that more than 15% (but still less than 40%) of observations fall below the detection limit. This relatively large proportion of non-detects could influence the interpretation and application of the data, and should be carefully accounted for in any analyses or conclusions drawn from these parameters.

Moreover, it is essential to acknowledge potential sources of variability stemming from evolving methodologies, technologies, *in situ* sensor calibrations, and satellite algorithms used for measuring chemical parameters and environmental metrics over time. Our database leverages historical data dating back to the 1930s, offering a unique perspective on relatively pristine environmental conditions. However, using older data necessitates careful consideration of the methodologies employed at the time of collection. It has been argued that the measurement of pH has improved over time²⁶. The accuracy of trace metal measurements before year 2000, particularly in their dissolved forms, can be affected by errors in the preparation and filtering processes, as outlined in previous studies^{27–29}. Advances in analytical techniques and instrumentation can introduce nuances and biases into the database, affecting the comparability of data collected across different periods. Moreover, despite advances in measurements, there has been a persistent bias in the measurement of certain key river parameters, such as pH and alkalinity. The reliability of pH data from low ionic strength freshwaters has been questioned due to shifts toward lower values, which impact associated CO_2 calculations along with uncorrected high organic alkalinity³⁰. The same study also provides a roadmap for correcting errors, which can be particularly useful for users looking to improve the accuracy of carbon speciation reconstruction from USGS data. Users are encouraged to consult this reference to mitigate errors effectively.

Given these complexities, users are encouraged to approach the database with a critical mindset. Detailed documentation for each station's data source and parameters within our database allows users to thoroughly explore data quality and methodological specifics. It is essential for users to consider the variations in methodologies and technologies used in chemical parameter measurements, along with changes in sensor calibrations and satellite algorithms. These factors, along with previously outlined uncertainties, emphasize the need for meticulous interpretation and application of the data in environmental research and policymaking. Our efforts have been focused on maintaining transparency in data compilation and thoroughly documenting data sources

to facilitate this process. Looking ahead, we recommend that data scientists and environmental scientists collaborate closely to develop innovative methods for enhancing the assessment and improvement of water quality measurements.

Usage Notes

While this database focuses on the nGoM rivers, the script and harmonization process can be used/adapted for other regional rivers with WQP/USGS data.

Code Availability

The open-source programming language R was used for data mining, harmonization, and database production. The R script for compiling the USGS data and the final river-ocean database are hosted on Zenodo²⁴ (<https://zenodo.org/records/10152141>), as well as on GitHub (<https://github.com/OceanArmos/ROcD-nGoM/tree/main>).

Received: 23 November 2023; Accepted: 20 December 2024;

Published online: 30 December 2024

References

- Adams, C. M., Hernandez, E. & Cato, J. C. The economic significance of the Gulf of Mexico related to population, income, employment, minerals, fisheries and shipping. *Ocean & Coastal Management* **47**, 565–580 (2004).
- BOEM. Bureau of Ocean Energy Management. <https://www.boem.gov/regions/gulf-mexico-ocs-region/oil-and-gas-gulf-mexico>.
- Malone, T. C. & Newton, A. The Globalization of Cultural Eutrophication in the Coastal Ocean: Causes and Consequences. *Frontiers in Marine Science* **7** (2020).
- Savoie, A. M. *et al.* Impact of local rivers on coastal acidification. *Limnology and Oceanography* **67**, 2779–2795 (2022).
- McComb, J. Q. *et al.* Trace elements and heavy metals in the Grand Bay National Estuarine Reserve in the northern Gulf of Mexico. *Marine Pollution Bulletin* **99**, 61–69 (2015).
- McKinney, L. *et al.* The Gulf of Mexico: An Overview. *Oceanography* **34**, 30–43 (2021).
- Beyer, J., Trannum, H. C., Bakke, T., Hodson, P. V. & Collier, T. K. Environmental effects of the Deepwater Horizon oil spill: A review. *Marine Pollution Bulletin* **110**, 28–51 (2016).
- Piazza, B. P. & La Peyre, M. K. The effect of Hurricane Katrina on nekton communities in the tidal freshwater marshes of Breton Sound, Louisiana, USA. *Estuarine, Coastal and Shelf Science* **83**, 97–104 (2009).
- GCOOS. GCOOS - Gulf of Mexico Coastal Ocean Observing System. <https://gcoos.org/> (2023).
- Read, E. K. *et al.* Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research* **53**, 1735–1745 (2017).
- Sprague, L. A., Oelsner, G. P. & Argue, D. M. Challenges with secondary use of multi-source water-quality data in the United States. *Water Research* **110**, 252–261 (2017).
- Shaughnessy, A. R., Wen, T., Niu, X. & Brantley, S. L. Three Principles to Use in Streamlining Water Quality Research through Data Uniformity. *Environmental Science & Technology* **53**, 13549–13550 (2019).
- Krasovich, E. *et al.* Harmonized nitrogen and phosphorus concentrations in the Mississippi/Atchafalaya River Basin from 1980 to 2018. *Scientific Data* **9**, 524 (2022).
- Gomez, F. A. *et al.* RC4USCoast: A river chemistry dataset for regional ocean model applications in the U.S. East, Gulf of Mexico, and West Coasts. *Earth System Science Data Discussions* 1–19 (2022).
- Lehrter, J. C., Murrell, M. C. & Kurtz, J. C. Interactions between freshwater input, light, and phytoplankton dynamics on the Louisiana continental shelf. *Continental Shelf Research* **29**, 1861–1872 (2009).
- Bianchi, T. S. *et al.* The science of hypoxia in the Northern Gulf of Mexico: A review. *Science of The Total Environment* **408**, 1471–1484 (2010).
- Dunn, D. E. *Trends in Nutrient Inflows to the Gulf of Mexico from Streams Draining the Conterminous United States, 1972-93*. *Water-Resources Investigations Report* <https://pubs.usgs.gov/publication/wri964113> (1996).
- Hirsch, R. M. & Cicco, L. A. D. User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data. in *Techniques and Methods* (U.S. Geological Survey, Reston, VA, 2015).
- Menheer, M. A. & Brigham, M. E. *Ground-Water Sampling Methods and Quality-Control Data for the Red River of the North Basin, Minnesota, North Dakota, and South Dakota, 1993-95*. *Water-Resources Investigations Report* <https://pubs.usgs.gov/publication/wri964317> (1997).
- Chanson, M. & Millero, F. J. Effect of filtration on the total alkalinity of open-ocean seawater. *Limnology and Oceanography: Methods* **5**, 293–295 (2007).
- Oelsner, G. P. *et al.* *Water-Quality Trends in the Nation's Rivers and Streams, 1972–2012—Data Preparation, Statistical Methods, and Trend Results*. *Scientific Investigations Report* <https://pubs.usgs.gov/publication/sir20175006> (2017).
- NASA Ocean Biology Processing Group. Moderate-resolution Imaging Spectroradiometer (MODIS) Aqua Chlorophyll Data. NASA Ocean Biology Distributed Active Archive Center <https://doi.org/10.5067/AQUA/MODIS/L3B/CHL/2022> (2022).
- da Silva, A. R. script to extract Aqua MODIS Chlorophyll a data from a nc file. *Gist* <https://gist.github.com/aolinto/7d3f282c9fde96133daa4cdd4f1bbcab> (2016).
- Armos, B. & Zhang, S. ROcD-nGoM: A River-Ocean Coupled Database for the Northern Gulf of Mexico. *Zenodo* <https://doi.org/10.5281/zenodo.10152141> (2023).
- Hornung, R. W. & Reed, L. D. Estimation of Average Concentration in the Presence of Nondetectable Values. *Applied Occupational and Environmental Hygiene* **5**, 46–51 (1990).
- Raymond, P. A. & Cole, J. J. Increase in the Export of Alkalinity from North America's Largest River. *Science* **301**, 88–91 (2003).
- Windom, H. L., Byrd, J. T., Jr, R. G. S. & Huan, F. Inadequacy of NASQAN data for assessing metal trends in the nation's rivers. *Environmental Science & Technology* (1991).
- Shiller, A. M. & Taylor, H. E. Comment on “Problems associated with using filtration to define dissolved trace element concentrations in natural water samples. *Environmental Science & Technology* **30**, 3397–3398 (1996).
- Horowitz, A. J. *et al.* Response to Comments on “Problems associated with using filtration to define dissolved trace element concentrations in natural water samples. *Environmental Science & Technology* **30**, 3398–3400 (1996).
- Liu, S., Butman, D. E. & Raymond, P. A. Evaluating CO₂ calculation error from organic alkalinity and pH measurement error in low ionic strength freshwaters. *Limnology and Oceanography: Methods* **18**, 606–622 (2020).

Acknowledgements

S.Z. acknowledges support from the DOE EarthShots award (DE-SC0024709) and the Type I collaboration grant from the College of Arts and Sciences at Texas A&M University. T.W. and S.Z. acknowledge the support from the National Science Foundation under Grant No. OAC-2209864. E.W. was supported by an NSF-funded REU program (OCE-1849932). We would also like to thank Shaoda Liu and Jens Hartmann for their insightful discussions on documented carbonate species in the river.

Author contributions

B.A. and S.Z. conceived the idea. B.A. led data mining, harmonization, and database production as well as manuscript drafting. S.Z. and T.W. provided background information, advice and resources for data collection and harmonization process. All co-authors (B.A., S.Z., T.W., E.W., P.D.) contributed to manuscript editing and revising.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-024-04338-1>.

Correspondence and requests for materials should be addressed to S.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024