scientific data

OPEN

DATA DESCRIPTOR

A 30-year dataset of CO_2 in flowing freshwaters in the United States

Timothy R. Toavs[1], Caleb T. Hasler[2], Cory D. Suski[3] & Stephen R. Midway[1]✉

Increasing atmospheric carbon dioxide (CO_2) concentrations have been linked to effects in a wide range of ecosystems and organisms, with negative effects of elevated CO_2 documented for marine organisms. Less is known about the dynamics of CO_2 in freshwaters, but the potential exists for freshwater organisms to be challenged by elevated CO_2. In flowing freshwaters CO_2 exhibits more variability than in lakes or the ocean, yet spatiotemporally extensive direct measures of CO_2 in freshwater are rare. However, CO_2 can be estimated from pH, temperature, and alkalinity—commonly collected water quality metrics. We used data from the National Water Quality Monitoring Council along with the program PHREEQC to estimate CO_2 in flowing freshwaters across 35,000 sites spanning the lower 48 US states from 1990 through 2020. Site data for water chemistry measurements were spatially joined with the National Hydrology Dataset. Our resulting dataset, CDFLOW, presents an opportunity for researchers to add CO_2 to their datasets for further investigation.

Background & Summary

Climate change caused by anthropogenically produced carbon dioxide (CO_2) is an issue that poses challenges around the world, including within marine and freshwater ecosystems. CO_2 concentrations in the atmosphere have been steadily increasing since the mid-nineteenth century with a total increase of around 40%[1] in that time. While CO_2 concentrations in the atmosphere have fluctuated throughout time, the rate of increase recorded since the 1850s is greater than any rate of increase that has occurred in the last million years[2]. As CO_2 in the atmosphere rises, dissolution of CO_2 into the ocean increases, thus interacting with the ocean carbonate system and ultimately leading to a decrease in ocean pH and a decrease in surface calcium carbonate ($CaCO_3$) concentrations, a process known as ocean acidification[3]. Dissolved CO_2 in marine and freshwater environments is measured as the partial pressure of CO_2 (pCO_2)[4]. This rise in pCO_2 has been shown to affect a wide range of ecosystems and organisms, with negative effects of elevated pCO_2 documented for marine and freshwater organisms. More specifically, ocean acidification caused by increasing atmospheric CO_2 has been shown to alter fish behaviour and physiology[5] and affect planktonic primary producers[6,7]. Outcomes of the effects are difficult to predict due to the variability across taxa. However, possible outcomes include reduced fish populations[5,8] and declines in ocean primary productivity[9]. While the effects of elevated pCO_2 in marine environments are well documented, less is known about the dynamics of pCO_2 in freshwaters, but the potential exists for freshwater organisms to be challenged[10].

While less is known about pCO_2 dynamics in freshwater, some general characteristics and processes have been documented. Flowing freshwaters have many different potential sources of pCO_2 and show high variability from one water body to another. Cole *et al.*[11] showed that pCO_2 in North American lakes was rarely at equilibrium with CO_2 in the atmosphere and found a range of concentration differences from 175 times lower pCO_2 than atmospheric CO_2 to 57 times greater. Flowing freshwaters show more variability and are typically supersaturated compared to the atmosphere, and have even been identified as sources of atmospheric CO_2[12]. Butman and Raymond[13] verified supersaturation in US flowing freshwaters and found that there is a relationship between pCO_2 and stream order suggesting a proportional relationship between stream size and pCO_2. Typically, pCO_2 in flowing freshwater is influenced by the water source of the flowing freshwater systems coupled with characteristics of that system including surrounding geologic conditions, pCO_2 residence time, and the gas transfer velocity[14]. Other contributing factors to pCO_2 include (but are not limited to) the balance between photosynthetic and respiration rates[15] and terrestrial respiration[12]. No matter the source, flowing freshwaters display high variability

[1]Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, 70803, USA. [2]Department of Biology, The University of Winnipeg, Winnipeg, Manitoba, Canada. [3]Department of Natural Resources, University of Illinois, Urbana, IL, 61801, USA. ✉e-mail: smidway@lsu.edu

in pCO_2 and while not much is known about the potential impacts on freshwater organisms and ecosystems it is important to understand pCO_2 spatiotemporal trends to identify potential impacts.

Considering the high variability displayed in flowing freshwaters a large spatiotemporal dataset is needed for understanding patterns and trends. Direct measures of pCO_2 in flowing freshwaters are extremely limited making it challenging to define spatial or temporal pCO_2 trends. However, pCO_2 can be estimated from a combination of water quality metrics including pH, temperature, and alkalinity—commonly collected water quality metrics, and has been done numerous times throughout the literature[13,16–18]. Our dataset (referred to as CDFLOW)[19] fills the need for a large spatiotemporal dataset using pH, temperature, and alkalinity measurements from across the lower 48 United States (CONUS) from 1990 through 2020. To our knowledge, CDFLOW[19] is the largest publicly available pCO_2 database with over 750,000 pCO_2 estimates coming from over 35,000 sites. CDFLOW[19] is also integrated with the National Hydrologic Dataset (NHD)[20] allowing for the addition of other environmental and geospatial variables[21] and ease when incorporating with other databases related to the NHD. CDFLOW[19] provides an opportunity for spatiotemporal analysis of pCO_2 across the CONUS and the possibility of adding pCO_2 data to other researchers' data.

Methods

Data query. Water quality measurements and their respective site-data (see below for site definition) were queried separately by each of the 48 CONUS states from the Water Quality Data Portal[22] using the following filters:

- Country = "United States of America"
- Site Type = "Stream"
- Date Range from = "01-01-1990"
- Date Range to = "12-31-2020"
- Sample media = "Water"
- Characteristics = "Alkalinity, total", "Alkalinity", "pH", and "Temperature, water"

The "Total alkalinity" and "Alkalinity" characteristic parameters are equivalent measurements but represent the different labels that respective reporting agencies use. The separate data queries for each state were merged using a shared variable called "MonitoringLocationIdentifier". The data queries and subsequent data merges resulted in 48 water quality measurement datasets with matching site data, representing each state within the CONUS.

***p*CO2 estimation.** The 48 datasets were processed and formatted separately then combined into one dataset for estimating pCO_2. The first step was to subset the datasets for quality and consistency among measurements. The following filters were applied:

- Removing non-numeric measurement values; e.g., "alkalinity <1 mg/l"
- Removing measurement values represented as statistical summaries and not observations; e.g., "average temp = 21 °C"
- Removing measurements not taken at the surface of the respective waterbody.
- Removing extreme water temperature measurements e.g., temperature ≤ 0 °C and temperature ≥ 40 °C
- Removing impossible pH values e.g., pH >14
- Removing pH values below 5.4

Hunt *et al*[23]. found that when pH is under 5.4 there is an increased risk of overestimating pCO_2 due to the possibility of non-carbonate anions contributing to the total pH, thus filtering out pH values less than 5.4. pH over 14 was excluded because the standard pH scale goes from 0–14. No filters were applied to alkalinity measurements.

Next, we grouped temperature, pH, and alkalinity measurements by location, date, and time. Grouping was done by creating a key identification by concatenating the following columns: "MonitoringLocationIdentifier", "ActivityStartDate", and "ActivityStartTime". If time data were not available for water quality measurements, they were still included but were grouped with water quality measurements also without time data. In grouping water quality measurements this way, they are grouped by the highest time/date resolution available, with day being the coarsest acceptable resolution. CDFLOW[19] requires all three of the queried water quality metrics to be present in each group to estimate pCO_2.

Finally, if a group had records of temperature, pH, and alkalinity, a single pCO_2 value was estimated using the United States Geological Survey's program PHREEQC v3[24]. PHREEQC quantitatively accounts for the chemical composition of a solution by relying on mole-balancing equations and in solving the mole-balance equations it derives the most likely pCO_2 estimation[25]. It should be noted that PHREEQC calculates pCO_2 under the assumption that alkalinity and pH in a system are determined by the current state of the carbonate system. PHREEQC can detect when this carbonate system assumption cannot be safely made in which case that group of observations was discarded. In cases where multiple measurements of a single water quality measurement were grouped with one or more of the two other required measurements, a measurement was chosen at random to be grouped for a pCO_2 estimate. All measurements not grouped were then discarded. Also, we excluded extreme outliers in the pCO_2 estimates which exceeded 2 standard deviations from the mean. The combination of the 48 processed, formatted, and estimated datasets resulted in a single dataset representing all our pCO_2 estimates across the CONUS.

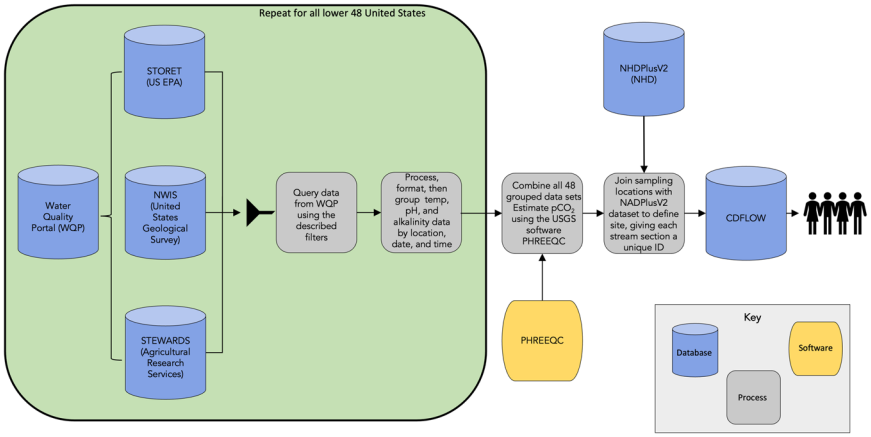

Fig. 1 Workflow for developing CDFLOW[19].

Defining sites. The site data that was merged with water quality measurements included latitude and longitude coordinates. These coordinates corresponded with the location identifier for each water quality measurement, now a pCO_2 estimate, and labeled as "MonitoringLocationIdentifier" (referred to as MID). We created a separate dataset using our dataset of pCO_2 estimates across the CONUS created above, and this new dataset included each of the unique MIDs along with latitude and longitude coordinates. Using the dataset of unique MIDs, we spatially joined each unique MID with the Environmental Protection Agencies National Hydrological Dataset Plus V2[20,26] (NHD) based on the closest stream catchment feature within NHD. Stream catchment features were labeled with a unique code called a COMID[27]. The spatial join resulted in a dataset with each unique MID now being associated with a COMID and was merged with our dataset of pCO_2 estimates across the CONUS. We also calculated the distance between MID's and the associated COMID, when the distance was greater than 100 meters the associated pCO_2 estimate(s) was excluded from our dataset of pCO_2 estimates across the CONUS. Finally, we spatially joined MID coordinates with Hydrologic Unit (HUC12)[28] polygons included in the NHD. The result of the two spatial joins is the ability to group pCO_2 estimates at any Hydrologic Units Code level and now sites within CDFLOW[19] are defined as what COMID the estimate resides.

All data queries, manipulations, and calculations were done using the statistical program R version 4.1.2[29]. A visual representation of the workflow to create CDFLOW can be found in Fig. 1.

Data Records

CDFLOW[19] exists as a single CSV file that has 779,186 pCO_2 estimates (rows) and 10 variables (columns) across the CONUS from 1990 through 2020 (Table 1). All 48 states within the CONUS are represented across 35,855 sites. CDFLOW[19] and all supporting code needed to generate and validate the dataset can be downloaded from a public repository on Figshare (https://doi.org/10.6084/m9.figshare.19787326).

While CDFLOW[19] has representation across all 48 states and 18 major watersheds within the CONUS, some areas are more represented than others. To display the spatial variability of CDFLOW we grouped estimates by hydrological unit codes (HUC2) and mapped them (Fig. 2). The South Atlantic Gulf and Mid Atlantic Watersheds had the most representation in CDFLOW[19] followed by the Missouri and Arkansas-White-Red watersheds. Also, we normalized the quantity of estimates within HUC2s by calculating the number of estimates per 5,000 km of stream distance within the HUC2. Total stream distance was calculated by taking the sum of COMID distances within the NHD for each HUC2. The normalized quantity of stream estimates followed similar patterns to the total number of estimates (Fig. 2). Leading us to conclude that estimates are not proportional to quantity of water but other non-environmental factors. We also looked at the temporal scale of CDFLOW[19] (Fig. 3). Generally, estimates increased going from the 1990's to the early 2000 were they remained constant then started to decrease from 2015 to 2020. Finally, we inspected spatiotemporal trends of estimates across the CONUS by splitting CDFLOW[19] into three decades (1990–2000, 2001–2010, 2011–2020). We found that the same spatial trends as the total number of estimates in Fig. 2 held constant across the three decades.

Technical Validation

Data validation. pCO_2 values in flowing freshwaters from the literature range widely with typical values falling between 1,300 to 4,300 micro atmospheres, but values in excess of 10,000 micro atmospheres have been reported[30–33] (micro atmospheres being the unit of the partial pressure of CO_2). CDFLOW estimates fall within the listed range with mean HUC2 values ranging from 1,200 to 4,500 micro atmospheres and a total interquartile range (25% to 75%) of 1,000 to 3,450 micro atmospheres. Also, CDFLOW does have values that reach in excess of 10,000 micro atmospheres as reported above. Although we find that CDFLOW estimates compare adequately to what is found in the literature, the majority of pCO_2 reported (including those cited here) come from estimated values using similar methods as CDFLOW. In a recent study, Liu *et al.*[34] assembled a data set of direct measurements of pCO_2 from other published studies. Liu *et al.*[34] calculated average pCO_2 values in different global ecoregions at 1810, 1540, and 2560 micro atmospheres in the arctic, temperate, and tropics respectively, and again CDFLOW had similar averages.

Column Names	Date type	Description	Source
Comid	Character	Stream Catchment code where the *p*CO2 estimates exists, code is derived from the EPA StreamCat dataset within the NHD	NHD
Date	Character	Date of temperate, pH and Alkalinity observations	WQP
Time	Character	Time of temperate, pH and Alkalinity observations	
HUC_12	Character	12-digit hydrological unit code	NHD
State	Character	US state where data point was located, given in standard state abbreviation code	WQP
Temp.C	Numeric	Temperature in units Celsius	WQP
pH.std_units	Numeric	pH given in standard pH units	WQP
Alkalinity.ueq/kgw	Numeric	total alkalinity in units - micro equivalence per kilogram water	WQP
*p*CO2.uatm	Numeric	*p*CO2 given in units - micro atmospheres	PHREEQC
CO2.mg/l	Numeric	Concentration of CO_2 given in milligrams per liter	PHREEQC

Table 1. Description of the data included in CDFLOW[19]. Variables are indicated under column names with the data type, a description, and the original source of that column.

We downloaded the dataset assembled by Liu *et al.*[34] and compared it with CDFLOW. However, first, we did the same site join as done in CDFLOW to assign the direct measurements COMIDs and Hydrologic Unit Codes. We then filtered CDFLOW to the months that data from the direct measurements were from and the HUC8s data was located. Both datasets were then filtered so that each HUC8 had a minimum of 10 data points (in order to avoid comparing very low sample sizes). We then did a separate ANOVA comparing the data from CDFLOW and Liu *et al.*[34] for each HUC8. This resulted in 26 within-HUC8 comparisons. Of those comparisons, less than half (46%) were significantly different ($p < 0.05$), suggesting that most of the time our estimates were distributed the same as those in Liu *et al.* (2022). We also inspected the direction of the bias between the estimates and direct measurements by finding the difference between the median pCO_2 values in each HUC8. This result is akin to examining residuals from a linear model, in which we expect the differences to be centered on 0 and normally distributed. We found that the bias difference (i.e., residuals) between the medians was homoscedastic, which is strong evidence that neither our data or the Liu *et al.*[34] data was over- or under-estimating *p*CO2.

Site ground truth. To test the accuracy of the site join procedure used to define sites in CDFLOW we created a procedure to ground truth the site join. The procedure worked by randomly choosing 50 CDFLOW sites and mapping the original latitude and longitude as well as the given COMID *and* all COMID stream features within 0.025 degrees latitude and 0.025 degrees longitude of the original coordinates in 50 separate plots. The resulting 50 plots were then checked manually by 2 observers to demonstrate how often the unsupervised procedure led to a reliable result. Both observers independently found that 50/50 (100%) of the random sites were correctly assigned. The R-script for the analysis is available at the Figshare link (https://doi.org/10.6084/m9.figshare.19787326).

Water quality data portal. The Water Quality Data Portal is a water quality data repository hosted by the United States Geological Survey[22]. Users can interface and download data *via* the Water Quality Data Portal website (https://www.waterqualitydata.us). The Water Quality Data Portal is a dynamic data repository with over 290 million standardized records. A record being a single collected water quality metric. Contributing agencies include all water quality records reported to the United States Geological Survey, the United States Department of Agriculture, and the Environmental Protection Agency.

National hydrological dataset. The National Hydrological Dataset (NHD) is a national geospatial surface water framework hosted by the Environmental Protection Agency building in conjunction with the United States Geological Survey[20,26]. NHD includes shapefiles mapping all flowing water systems throughout the United States.

StreamCat. The StreamCat dataset is incorporated into the NHD, which maps stream segments and their associated catchment within the CONUS[27].

PHREEQC. PHREEQC Version 3 is a computer program written in the C++ programming language that is designed to perform a wide variety of aqueous geochemical calculations[24]. PHREEQC quantitatively accounts for the chemical composition of a solution by relying on mole-balancing equations. It is free and available (e.g. https://www.usgs.gov/software/phreeqc-version-3).

Usage Notes

Estimation uncertainty. PHREEQC relies on the equilibrium of the carbonate system in water in order to estimate pCO_2[25] and uncertainty has been documented for pCO_2 estimates that rely on carbonate equilibrium. When error is present in pCO_2 estimation using carbonate equilibria, overestimation is usually the error[23,35,36]. We applied filters to data that went into pCO_2 estimation to mitigate overestimation (see methods). Further filters can be applied to data to further mitigate overestimation risks at the discretion of the user; e.g., removing pCO_2 estimates greater than 100,000 parts per million volume, and removing alkalinity values below 1,000 micro equivalents per kilogram water[36]. While absolute values of CDFLOW[19] pCO_2 estimates may be subject to overestimation relative values and trends are still valid.

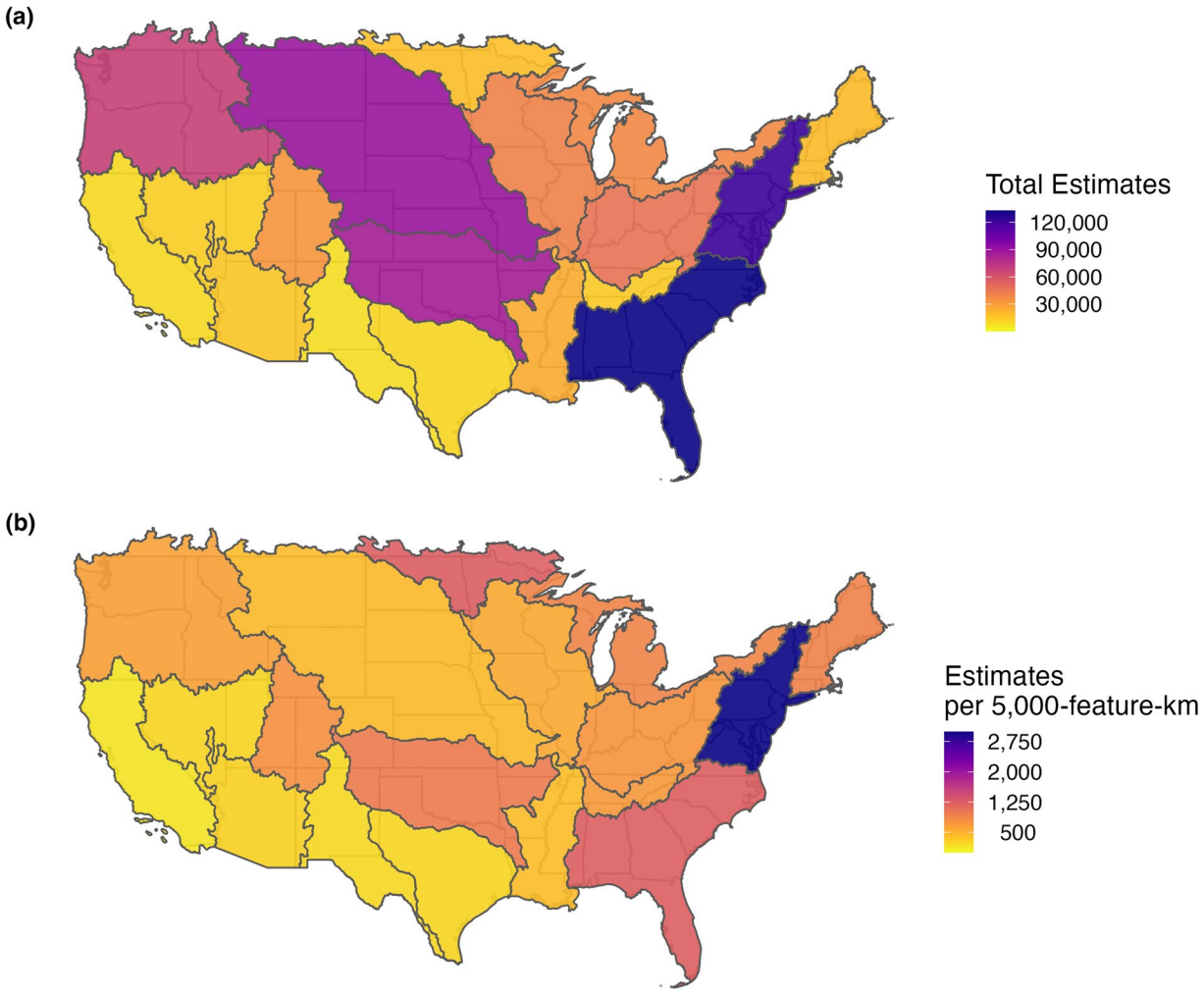

Fig. 2 Spatial distribution of pCO_2 estimates within CDFLOW[19]. Panel (**a**) shows the total number of estimates in each Hydrological unit code-2 (HUC2)[28] within the CONUS. Panel (**b**) shows the total number of estimates divided by the number of 5000-feature-km in each HUC2 within the CONUS.

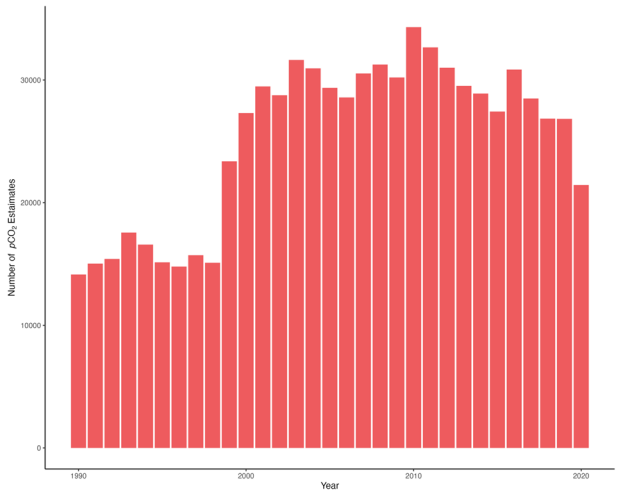

Fig. 3 Counts of pCO_2 estimates by year within CDFLOW[19].

Uncertainty estimates from PHREEQC are available as mole balance percent errors. However, when only including three metrics to compute pCO_2 this error term is always quite high but does not necessarily reflect a poor estimate. As discussed in Potter *et al.*[37] which compares modeled pCO_2 estimates using PHREEQC to

direct measurements, they conclude that although mole change balance percent errors are high PHREEQC still provides a good estimate of pCO_2 using pH, temperature, and alkalinity. So, we have decided to exclude mole change balance percent error from the dataset as they are not relevant for modeling purposes and do not negate the validity of CDFLOW pCO_2 estimates.

Extra parameters. PHREEQC does allow for the inclusion of extra parameters when estimating pCO_2, and more specifically the inclusion of other dissolved inorganic species. However, data on other dissolved inorganic species that matches the same date, time, and location of the pH, temperature, and alkalinity is only available to a limited number of observations. Due to the limited number of other dissolved inorganic species for observation they were excluded from the PHREEQC estimation. However, the use of other dissolved inorganic species in estimating pCO_2 using PHREEQC would potentially allow for more robust estimates. If CDFLOW users are interested in the inclusion of other dissolved inorganic species a supporting script can be found at the Figshare link (https://doi.org/10.6084/m9.figshare.19787326) that describes and gives examples of the changes required to do so.

Expanding data. By defining sites in CDFLOW[19] by which COMID they fall into gives each site all the data that corresponds to that COMID. COMID data can be accessed *via* the NHD (see technical validation). COMID data can also be accessed *via* R package NHD Tools[38].

Code availability

Code for the creation of CDFLOW is available as a series of R scripts *via* public repository on Figshare19 (https://doi.org/10.6084/m9.figshare.19787326).

Received: 2 June 2022; Accepted: 15 December 2022;
Published online: 11 January 2023

References

1. Hartmann, D. L. *et al.* in *Climate change 2013 the physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change* 159–254 (Cambridge University Press, 2013).
2. Doney, S. C. & Schimel, D. S. Carbon and Climate System Coupling on Timescales from the Precambrian to the Anthropocene. *Annual Review of Environment and Resources* **32**, 31–66, https://doi.org/10.1146/annurev.energy.32.041706.124700 (2007).
3. Doney, S. C., Fabry, V. J., Feely, R. A. & Kleypas, J. A. Ocean acidification: the other CO_2 problem. *Annual Review of Marine Science* **1**, 169–192 (2009).
4. Solomon, S., Manning, M., Marquis, M. & Qin, D. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*. **Vol. 4** (Cambridge University Press, 2007).
5. Munday, P. L., Jarrold, M. D. & Nagelkerken, I. Ecological effects of elevated CO_2 on marine and freshwater fishes: from individual to community effects. *Fish Physiology* **37**, 323–368, https://doi.org/10.1016/bs.fp.2019.07.005 (2019).
6. Ross, P. M., Parker, L., O'Connor, W. A. & Bailey, E. A. The Impact of Ocean Acidification on Reproduction, Early Development and Settlement of Marine Organisms. *Water* **3**, 1005–1030, https://doi.org/10.3390/w3041005 (2011).
7. Orr, J. C. *et al.* Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature* **437**, 681–686, https://doi.org/10.1038/nature04095 (2005).
8. Munday, P. L. *et al.* Replenishment of fish populations is threatened by ocean acidification. *Proceedings of the National Academy of Sciences* **107**, 12930–12934, https://doi.org/10.1073/pnas.1004519107 (2010).
9. Flynn, K. J. *et al.* Changes in pH at the exterior surface of plankton with ocean acidification. *Nature Climate Change* **2**, 510–513 (2012).
10. Hasler, C. T., Butman, D., Jeffrey, J. D. & Suski, C. D. Freshwater biota and rising pCO_2. *Ecology Letters* **19**, 98–108, https://doi.org/10.1111/ele.12549 (2016).
11. Cole, J. J., Caraco, N. F., Kling, G. W. & Kratz, T. K. Carbon dioxide supersaturation in the surface waters of lakes. *Science* **265**, 1568–1570 (1994).
12. Cole, J. J. *et al.* Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10**, 172–185 (2007).
13. Butman, D. & Raymond, P. A. Significant efflux of carbon dioxide from streams and rivers in the United States. *Nature Geoscience* **4**, 839–842, https://doi.org/10.1038/ngeo1294 (2011).
14. Wetzel, R. G. *Limnology: lake and river ecosystems*. (Gulf Professional Publishing, 2001).
15. Sobek, S., Algesten, G., Bergström, A. K., Jansson, M. & Tranvik, L. J. The catchment and climate regulation of pCO_2 in boreal lakes. *Global Change Biology* **9**, 630–641 (2003).
16. Lauerwald, R., Laruelle, G. G., Hartmann, J., Ciais, P. & Regnier, P. A. Spatial patterns in CO_2 evasion from the global river network. *Global Biogeochemical Cycles* **29**, 534–554, https://doi.org/10.1002/2014gb004941 (2015).
17. Jones, J. B. Jr, Stanley, E. H. & Mulholland, P. J. Long-term decline in carbon dioxide supersaturation in rivers across the contiguous United States. *Geophysical Research Letters* **30**, https://doi.org/10.1029/2003gl017056 (2003).
18. Liu, S. & Raymond, P. A. Hydrologic controls on pCO_2 and CO_2 efflux in US streams and rivers. *Limnology and Oceanography Letters* **3**, 428–435 (2018).
19. Toavs, T. M. Steve; Hasler, Caleb; Suski, Cory. *Figshare*. https://doi.org/10.6084/m9.figshare.19787326 (2022).
20. McKay, L. *et al.* (ed US Environmental Protection Agency) (2012).
21. Wieczorek, M., Jackson, S. & Schwarz, G. Select attributes for NHDPlus version 2.1 reach catchments and modified network routed upstream watersheds for the conterminous United States. *US Geological Survey*. https://doi.org/10.5066/F7765D7V (2018).
22. Read, E. K. *et al.* Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research* **53**, 1735–1745, https://doi.org/10.1002/2016wr019993 (2017).
23. Hunt, C. W., Salisbury, J. E. & Vandemark, D. Contribution of non-carbonate anions to total alkalinity and overestimation of pCO_2 in New England and New Brunswick rivers. *Biogeosciences* **8**, 3069–3076, https://doi.org/10.5194/bg-8-3069-2011 (2011).
24. Parkhurst, D. L. & Appelo, C. in *US geological survey techniques and methods* **Vol. 6** (ed USGS) 497 (2013).
25. Parkhurst, D. L. & Appelo, C. User's guide to PHREEQC (Version 2): A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. *Water-Resources Investigations Report* **99**, 312 (1999).
26. (ed U.S. Geological Survey) (USGS, 2018).
27. Hill, R. A., Weber, M. H., Leibowitz, S. G., Olsen, A. R. & Thornbrugh, D. J. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. *JAWRA Journal of the American Water Resources Association* **52**, 120–128, https://doi.org/10.1111/1752-1688.12372 (2016).

28. Seaber, P. R., Kapinos, F. P. & Knapp, G. L. (ed USGS) (1987).
29. Team, R. C. (2013).
30. Richey, J. E., Melack, J. M., Aufdenkampe, A. K., Ballester, V. M. & Hess, L. L. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO_2. *Nature* **416**, 617–620 (2002).
31. Johnson, M. S. *et al*. CO_2 efflux from Amazonian headwater streams represents a significant fate for deep soil respiration. *Geophysical Research Letters* **35** (2008).
32. Humborg, C. *et al*. CO_2 supersaturation along the aquatic conduit in Swedish watersheds as constrained by terrestrial respiration, aquatic respiration and weathering. *Global Change Biology* **16**, 1966–1978 (2010).
33. Cole, J. J. & Caraco, N. F. Carbon in catchments: connecting terrestrial carbon losses with aquatic metabolism. *Marine and Freshwater Research* **52**, 101–110 (2001).
34. Liu, S. *et al*. The importance of hydrology in routing terrestrial carbon to the atmosphere via global streams and rivers. *Proceedings of the National Academy of Sciences* **119**, e2106322119 (2022).
35. Golub, M., Desai, A. R., McKinley, G. A., Remucal, C. K. & Stanley, E. H. Large uncertainty in estimating pCO_2 from carbonate equilibria in lakes. *Journal of Geophysical Research: Biogeosciences* **122**, 2909–2924 (2017).
36. Abril, G. *et al*. Large overestimation of pCO_2 calculated from pH and alkalinity in acidic, organic-rich freshwaters. *Biogeosciences* **12**, 67–78, https://doi.org/10.5194/bg-12-67-2015 (2015).
37. Potter, L., Tollrian, R., Wisotzky, F. & Weiss, L. C. Determining freshwater pCO_2 based on geochemical calculation and modelling using PHREEQC. *MethodsX* **8**, 101430, https://doi.org/10.1016/j.mex.2021.101430 (2021).
38. nhdplusTools: Tools for Accessing and Working with the NHDPlus (2022).

Author contributions

All authors were active in the development of this manuscript. S.R.M., C.T.H. and C.D.S. conceived of the idea and S.R.M. developed the estimates. T.R.T. led the effort for nationwide estimates, in addition to leading the writing of the manuscript. All authors were active in reviewing and editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.R.M.

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.