GRiMeDB: The global river database of methane concentrations and fluxes

Emily H. Stanley\textsuperscript{1}, Luke C. Loken\textsuperscript{1,2}, Nora J. Casson\textsuperscript{3}, Samantha K. Oliver\textsuperscript{2}, Ryan A. Sponseller\textsuperscript{4} Marcus B. Wallin\textsuperscript{5} Liwei Zhang\textsuperscript{6} Gerard Rocher-Ros\textsuperscript{7,8}

1Center for Limnology, University of Wisconsin-Madison, Madison WI, USA 53706
2U.S. Geological Survey, Upper Midwest Science Center, Madison, WI USA 53726
3Department of Geography, University of Winnipeg, Winnipeg, MB Canada R3B 2E9
4Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden 90736
5Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, 75007, Sweden
6Sino-French Institute for Earth System Science, College of Urban and Environmental Sciences, Peking University, 100871, Beijing, China
7Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, 90183 Umeå, Sweden
8Integrative Freshwater Ecology Group, Centre for Advanced Studies of Blanes (CEAB-CSIC), 17300 Girona, Spain

Correspondence to: Emily H. Stanley (ehstanley@wisc.edu)
Abstract. Despite their small spatial extent, fluvial ecosystems play a significant role in processing and transporting carbon in aquatic networks, which results in substantial emission of methane (CH$_4$) to the atmosphere. For this reason, considerable effort has been put into identifying patterns and drivers of CH$_4$ concentrations in streams and rivers and estimating fluxes to the atmosphere across broad spatial scales. Yet progress toward these ends has been slow because of pronounced spatial and temporal variability of lotic CH$_4$ concentrations and fluxes and by limited data availability across diverse habitats and physicochemical conditions. To address these challenges, we present a comprehensive database of CH$_4$ concentrations and fluxes for fluvial ecosystems along with broadly relevant and concurrent physical and chemical data. The Global River Methane database (GRiMeDB; [https://doi.org/10.6073/pasta/f48cdb77282598052349e969920356ef](https://doi.org/10.6073/pasta/f48cdb77282598052349e969920356ef), Stanley et al., 2023) includes 24,024 records of CH$_4$ concentration and 8,205 flux measurements from 5,029 unique sites derived from publications, reports, data repositories, unpublished data sets, and other outlets that became available between 1973 and 2021. Flux observations are reported as diffusive, ebullitive, and total CH$_4$ fluxes, and GRiMeDB also includes 17,655 and 8,409 concurrent measurements of concentrations and 4,444 and 1,521 of fluxes for CO$_2$ and nitrous oxide (N$_2$O) respectively. Most observations are date-specific (i.e., not site averages) and many are supported by data for 1 or more of 12 physicochemical variables and 6 site variables. Site variables include codes to characterize marginal channel types (e.g., springs, ditches) and/or presence of human disturbance (e.g., point source inputs, upstream dams). Overall, observations in GRiMeDB encompass the broad range of the climatic, biological, and physical conditions that occur among world river basins, although some geographic gaps remain (arid regions, tropical regions, high latitudes and altitude systems). The global median CH$_4$ concentration (0.20 μmol L$^{-1}$) and diffusive flux (0.44 mmol m$^{-2}$ d$^{-1}$) in GRiMeDB are lower than estimates from prior site-averaged compilations, although ranges (0-456 μmol L$^{-1}$ and -136-4057 mmol m$^{-2}$ d$^{-1}$) and standard deviations (10.69 and 86.4) are greater for this larger and more temporally resolved database. Available flux data are dominated by diffusive measurements despite the recognized importance of ebullitive and plant-mediated CH$_4$ fluxes. Nonetheless, GRiMeDB provides a comprehensive and cohesive resource for examining relationships between CH$_4$ and environmental drivers, estimating the contribution of fluvial ecosystems to CH$_4$ emissions, and contextualizing site-based investigations.

1 Introduction

Despite their small areal extent, running-water (fluvial) ecosystems play a significant role in processing and transporting carbon (C) in and through aquatic networks, including the production, consumption, transport, and evasion of carbon dioxide (CO$_2$) and methane (CH$_4$). The profound planetary warming effects of CH$_4$ in the atmosphere, its erratic but accelerating rate of increase over recent years (NOAA, 2022), the significant contributions of natural sources to the growing atmospheric pool (Turner et al., 2019), and improvements in gas measurement technologies have all contributed to a rapid increase in studies of CH$_4$ dynamics in aquatic environments in general, and fluvial ecosystems in particular. These studies reveal widespread
supersaturation of CH₄ in running waters that underlies their larger than expected contribution to the atmospheric pool (Stanley et al., 2016).

Efforts to quantify fluvial CH₄ dynamics at regional, continental, and global scales have been fraught with uncertainty, reflecting the inherent variability of this gas in surface waters combined with a notable limitation in data availability. Sources and sinks of CH₄ are often unevenly distributed over space and time within drainage systems and, as a result, concentrations can vary over 1-3 orders of magnitude over short time periods (sub-daily to sub-weekly; e.g., Natchimuthu et al., 2017; Smith and Böhlke, 2019) or relatively small spatial extents (<10 - <100 m for small streams and large rivers; e.g., Anthony et al., 2012; Crawford et al., 2017; Bretz et al., 2021; Robison et al., 2021). Similarly, several drivers or predictors of CH₄ have been identified in the literature, and these properties also have variable spatial and temporal distributions. Thus, efforts to estimate the total emissions from world rivers have relied on relatively small data sets composed of site-specific values that have been averaged over time, and then used upscaling strategies based on Monte Carlo techniques or extrapolations using predictor variables that have little or no significant statistical relationships with large-scale patterns of gas concentrations or fluxes (Hutchins et al. 2020). Consequently, current global scale estimates of riverine emissions are poorly constrained and highly uncertain (Saunois et al., 2020; Rosentreter et al., 2021).

The combination of rapidly increasing atmospheric concentrations of CH₄, the significant role of fluvial systems in emitting this gas, and, critically, current difficulties in explaining or predicting concentrations and fluxes with reasonable certainty inspired the central goal of this paper: to assemble a comprehensive database of CH₄ concentrations and fluxes for fluvial ecosystems that includes broadly relevant concurrent physical and chemical data. This effort expands upon a prior compilation of CH₄ and CO₂ data (MethDB; Stanley et al., 2015) that was constructed to emphasize among-site differences and included 1,496 concentration records and 532 flux records from 1,080 sites. In this more comprehensive Global River Methane database (GRiMeDB), most data are date-specific (i.e., not averaged over time), the breadth of site types is expanded to include marginal fluvial habitats as well as disturbed and artificial waterways, and CH₄ data are supported by a broad suite of site-specific physical and chemical attributes along with concurrent measurements of CO₂ and N₂O where available. Given the more finely resolved scale of the data and the growth of the field in the past decade, GRiMeDB represents a significant expansion beyond MethDB. Building GRiMeDB with greater detail and breadth of data was done with the intent of increasing opportunities to identify and predict spatial and temporal variation in CH₄, to test hypotheses related to greenhouse gas dynamics, and to reduce uncertainty in future upscaled estimates of gas emissions. In this paper, we (1) provide a detailed description of the components of the database and its construction; (2) summarize some basic patterns of gas concentrations and fluxes from GRiMeDB; and (3) highlight critical data gaps and possible future research opportunities for improving current understanding of CH₄ dynamics in streams and rivers.
2 Database components and assembly

GRiMeDB is composed of four tables that contain information related to (1) data sources, (2) sites, (3) gas concentrations and supporting physicochemical data, and (4) gas fluxes. All tables are linked by unique data source identifiers, and all concentration and flux observations are also linked to unique site numbers (Fig. 1). Data included in GRiMeDB were gathered from scientific journals, government reports, public data repositories, theses, dissertations, and unpublished data sets provided by individual investigators. Sources were discovered via searches of bibliographic databases and data repositories (Web of Science, Google Scholar, ProQuest Dissertations & Theses Global, China National Knowledge Infrastructure, Environmental Data Initiative, USGS ScienceBase, Natural Environment Research Council (NERC) Environmental Information Data Centre, Arctic Data Center, PANGAEA, Zenodo) using the keywords: methane and stream* or river* or ditch* or canal*, and searches were repeated numerous times prior to December 2021 for completeness. We also used informal ‘word of mouth’ approaches to discover additional, often unpublished data sets.

![Diagram of GRiMeDB connections between tables]

Figure 1: General structure of the GRiMe database and connections between its four tables. Information flow began with entering information about each data source into the Sources Table and assigning a unique Source_ID. Site information for each site within a data source was then entered into the Sites table. The site was given a unique Site_ID and linked to its data source by the Source_ID. Source_IDs and Site_IDs were carried over to all concentration and flux observations in their respective tables. Methane (CH₄) observations include site-date combinations with only concentration data (orange), only flux data (green), or both concentration and flux data (brown). Concentrations and available supporting data (described in Sect. 2.3) were entered into the Concentration Table, and each observation was given a unique observation (obs) name. For site-date combinations that had both concentration and flux observations, the Source_ID, Site_ID, observation name, and date information were copied to the Flux Table for data entry. Site-date combinations with flux data only were entered into the Flux Table and given a unique observation name. If a flux observation had associated supporting data, the Source_ID, Site_ID, observation name, and date information were copied to the Concentration Table for supporting data entry. However, if there were no supporting data, matching rows were not added to the Concentration Table.

All potential data sources were first screened to determine their appropriateness for inclusion in GRiMeDB. Several criteria were established a priori to ensure the usability of the data and that it was derived from inland running water systems. Coastal sites with
>1 ppt salinity were considered estuarine and thus were excluded. Similarly, sites that were situated in reservoirs or immediately upstream of small dams, dam spillways, beaver ponds, or lake outlets or were subject to experimental manipulation were omitted. We did not enter fluxes derived from chambers attached to collars or inserted into sediments because we could not be certain that such measurements were capturing air-water fluxes. Sources that reported minimum and maximum gas concentrations or fluxes only were not included. Finally, rates expressed on an annual basis were also excluded to avoid introducing uncertainty associated with different upscaling assumptions and methods.

2.1 Sources Table

The Sources Table contains the list of all sources used to build GRiMeDB, a unique identification number (Source_ID) for each CH₄ data source, and basic bibliographic information for the data source (Title, Author, Source, publication year [Pub_year], and digital object identifiers [Paper.DOI or Data.DOI_primary] or other persistent identifier; all column titles for this table are defined in Table A1). In several cases, data sources were supplemented with additional supporting information (e.g., associated physicochemical data) from separate sources (described further in Sect. 2.3) or additional or corrected information from authors (Fig. 2). In the latter case, we contacted authors if questions arose regarding their data (e.g., clarification regarding units) and/or to request supporting information or site- or date-specific concentrations or fluxes if published values were aggregated. Inclusion of additional unpublished data from authors is noted in the Sources Table along with a description of the addition or correction. If supporting data from separate published sources were used, the DOI or other persistent identifier for the secondary source was listed in a separate column [Data.DOI_supporting].

Figure 2: Workflow for entering data into the Sources Table of GRiMeDB.
2.2 Sites Table

The Sites Table reports basic information on attributes for all sites where CH$_4$ was sampled. Each site has a unique identification code (Site_ID) and name (usually taken directly from the data source) and is linked to the Sources Table via the Source_ID (see Table A2 for detailed descriptions of all columns in the Sites Table). What compose a ‘site’ (i.e., the spatial extent of data collection) varied among data sources and includes (1) discrete sampling points, (2) geomorphically distinct study reaches, and/or (3) larger channel sections, drainage networks, or other geographic units. The second case typically corresponded to reaches such as riffles or pools in small streams. In the third case, multiple points were often sampled within the ‘site’ and data were then presented as averages. The distance between sampling points that had been averaged varied widely, but were typically >1 km, and in some cases exceeded 100 km. Because land use, channel order, slope, etc., can vary substantially across such distances, we included fields to indicate if a site was an aggregation of widespread points (‘aggregated’), and if so, the number of locations in the aggregation (if available). We also limited the resolution of latitude and longitude for these sites to < 3 decimal places. At the opposite extreme, gas sampling at points very close to one another (a ‘high density site’ sensu Fig. 3) has the potential to create ambiguities for site delineation and data analysis. To avoid these pitfalls, we combined points with slightly different latitude-longitude values to represent a single site for three specific cases. First, multiple samples collected at different points and/or depths within a channel cross-section were averaged to form a single site. Second, some drainages or regions were surveyed repeatedly (particularly the Congo River basin and streams in Pennsylvania, USA) and it was not always clear if closely-situated (ca. 10-50 m) points from different surveys were intended to be a repeated sampling of the same location or sampling of discrete sites. Some judgment was involved in choosing between these two possibilities, and in a subset of cases, points in close proximity to one another that were sampled on separate dates were treated as a single site. What constituted ‘close proximity’ varied between small streams and large rivers but was always <100 m, and typically <50 m. Finally, three data sources had extremely high sampling densities within discrete reaches (50 - >20,000 samples per reach; Crawford et al., 2016; Call et al., 2018; Loken et al., 2018). Because closely adjacent gas samples can be spatially autocorrelated (Crawford et al., 2017) and including all individual values from these studies would have resulted in their over-representation in the database, individual point measurements were treated as within-reach replicates.
Figure 3: Workflow for entering and checking data for the GRiMeDB Sites Table. ‘Lat-Lon’ is an abbreviation for latitude and longitude.

For a site used in multiple studies, the Site_ID was assigned to the earliest paper and a comment was added to the site entry noting its use in other data sources (Fig. 3). Latitude and longitude coordinates were available for most sites; however, in several cases, location information was acquired from authors or estimated from study site figures using Google Earth (© Google Earth 2020). All sites were plotted on Google Earth and inspected (Fig. 3) to identify and correct data errors. If a site’s coordinates were immediately adjacent to, but not on a channel, the coordinates were adjusted to fall on the channel and this
modification was noted in the Comments field. If available, additional variables drawn from the data sources were entered to characterize the site, including stream name, basin or region name, elevation, channel slope, Strahler order, basin area, and codes denoting distinct channel or site types (described below). To supplement the available elevation data, we also estimated elevation for all sites except aggregated sites or sites with poorly-resolved coordinates (<3 decimal places for both latitude and longitude) after snapping coordinates to the nearest stream. To determine the adjusted within-channel coordinates, we first downloaded a digital elevation model (DEM) for each site using the function get_elev_raster() from the package “elevatr” (version 0.4.2; Hollister et al., 2021) for R statistical software (version 4.2, R Core Team 2021) at a resolution of 6-9 m depending on the location in the globe. Second, the DEM was processed for hydrological correctness using the package “whitebox” (version 1.2.0, Wu, 2020) by filling single cell pits (fill_single_cell_pits() function) and breaching depressions (breach_depressions() function) to obtain a flow direction model (d8_pointer() function). Finally, we calculated a flow accumulation model (d8_flow_accumulation() function). If the coordinates reported in the data source had a flow accumulation <10 cells (indicating that they were not located in a preferential flow path), new coordinates were assigned to the cell with the highest flow accumulation within a 50 m radius. If the initial site had a high flow accumulation value (>10 cells), we assumed the site was in a stream channel. Typically, the snapping procedure resulted in very minor changes to a site’s location (median <3 m).

Many studies of CH₄ dynamics have been undertaken to determine if and how specific phenomena such as presence of upstream reservoirs, point source discharges, thermokarst features, or oil and gas extraction potentially affect fluvial CH₄ (and other constituents), usually with an expectation of a net enhancement of concentrations and fluxes. Similarly, other studies have examined sites that may be expected to be enriched in CH₄, but whose fluvial identity might be considered marginal or ambiguous (e.g., springs, floodplain backwaters, ditches, canals). Inclusion of such ‘methane hunting’ studies has the potential to bias the dataset toward higher values (Stanley et al., 2016). Nonetheless we included these studies in GRiMeDB because they provide an opportunity to investigate the consequences of human activity and gain a more comprehensive understanding of fluvial CH₄ dynamics. However, to accommodate future analyses in which use of such data might be unsuitable, or alternatively, when these sites might be the sole focus of a study, we generated a set of channel codes to identify targeted site types (Table 1). Information about four of the codes was not consistently available among data sources and thus their assignment often involved judgment calls. The first case involved determining if the presence of an upstream dam (code DD) was relevant for sites of varying distances downstream. We used a distance of 7 km as a cut-off for this category, although the zone of influence may be far shorter or extend far beyond this distance depending on dam size and operation (Kemenes et al., 2007), respectively. To provide some context for this code, a site’s distance from a dam was acquired from the data source or estimated in Google Earth using the Path tool and reported in the Comments field whenever possible. The second case involved straight, symmetrical channels that are common in many agricultural and urban areas. Frequently, it was not known if this unnatural geometry was due to channelization (straightening) of a stream (code CH) or creation of a new channel (ditches and canals; codes DIT and CAN). In the absence of specific information, straight channels were classified as CH. Third, channels draining or passing through wetlands (WS) were
often difficult to identify, particularly given seasonal variation in wetland appearance. Finally, floodplain channels presented a distinct challenge because of the complex nature of these environments and their potential to be classified as either riverine or wetland systems. We used the FP code to indicate habitats that were described as, or appeared to be lentic (i.e., backwaters or connected floodplain lakes) but were persistently connected to the main river channel and thus were part of the fluvial system. Given these ambiguities, we recommend that these four codes be viewed and used with care.

2.3 Concentration Table and Flux Table

The Concentration Table and the Flux Table contain the primary gas data central to GRiMeDB, and the Concentration Table also hosts physical and chemical variables associated with concentration and/or flux observations (see Tables A3 and A4 for the full list of columns and their descriptions). The vast majority of concentration and flux data were extracted from tables within data sources, data repositories, or provided by authors. However, in some cases, values were acquired from figures using graphical digitizing software (WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/), GetData (http://getdata-graph-digitizer.com/), or DigitizeIt (https://www-digitizeit.xyz/)). Plots with log scales or that were difficult to interpret were not digitized. The accuracy and consistency of this method were evaluated by comparing data generated by different individuals digitizing a set of common figures and by comparing digitized results to known results. Agreement between both comparisons was strong (average slope = 0.994, average $R^2 = 0.9996$ for 5 comparisons between individuals digitizing the same dataset, and average slope = 0.998, average $R^2 = 0.997$ for digitized versus actual data for 7 datasets; see Table S1 for further details), demonstrating the reliability of this method of data gathering.

Whenever possible, concentrations and fluxes were entered as values for individual sites on individual days (i.e., not averaged across sites or days) (Fig. 4). Because 1 day represented the lowest level of temporal resolution in GRiMeDB, repeated measurements made on a sub-daily scale were averaged and expressed as a daily value and were not considered to be aggregated over time. If multiple replicates were collected at different times on the same day (e.g., a study of diurnal gas dynamics), this was noted in the Comments fields and measurements prior to and after 12:00 a.m. (local time) were entered as separate, consecutive days. Observations resolved to the daily scale can be identified using either a “No” in the Aggregated_Time field or by having the same reported starting (Date_start) and ending (Date_end) dates. If the specific start and end dates were not specified in the data source, we entered the day as the 15th of the month and noted this approximation in the Comments field. If available, we also reported minimum and maximum values and standard deviations (SD) for entries that were aggregated over space and/or time. SDs, but not minima and maxima were reported for replicates from non-aggregated sampling when available, except for reach-averaged entries with multiple within-reach measurements and diel studies with multiple within-day values. In these cases, minima and maxima were also included.
Table 1. Codes denoting distinct site or channel attributes or presence of conditions that potentially affect methane (CH$_4$) concentrations or fluxes. Assignment of codes to a site is based on information provided in the data source and/or physical appearance of a site and a site may have more than one code. Codes are reported in the Channel_type field of the Sites Table.

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAN</td>
<td>Canal or other artificial channel with hardened channel boundaries</td>
</tr>
<tr>
<td>CH</td>
<td>Channelized; a channel that has long straight-line sections of uniform width and changes in channel direction are typically distinct angular features rather than curves</td>
</tr>
<tr>
<td>DC</td>
<td>Channel in a river delta</td>
</tr>
<tr>
<td>DD</td>
<td>Downstream (within 7 km) of a dam. Samples from spillways were excluded.</td>
</tr>
<tr>
<td>DIT</td>
<td>Ditch, typically for agricultural drainage, without channel hardening</td>
</tr>
<tr>
<td>FP</td>
<td>Site in a floodplain water body connected to the main channel that appears lentic or is described as a floodplain lake or backwater. This category does not include side-channels within floodplains or tributary channels transecting a floodplain</td>
</tr>
<tr>
<td>GT</td>
<td>Site below the toe or terminus of a glacier</td>
</tr>
<tr>
<td>IMP</td>
<td>Presence of multiple and typically small impoundments in a site’s vicinity (e.g., various European rivers, Mississippi River)</td>
</tr>
<tr>
<td>PI</td>
<td>Permafrost influenced; this refers specifically to sites at or immediately below thermokarst outflows and not to sites in areas underlain with permafrost</td>
</tr>
<tr>
<td>PS</td>
<td>Immediately (&lt;1 km) downstream of a point source discharge</td>
</tr>
<tr>
<td>SP</td>
<td>Spring channel; this does not include sites characterized as seeps (features with low flow volume adjacent to channels)</td>
</tr>
<tr>
<td>TH</td>
<td>Site receiving inputs of thermogenic CH$_4$, either naturally or as a result of mining, fracking, oil extraction, and other related activities.</td>
</tr>
<tr>
<td>WS</td>
<td>Wetland stream; site is in a wetland or immediately downstream from the outlet of a wetland</td>
</tr>
<tr>
<td>NORM</td>
<td>Non-targeted site</td>
</tr>
</tbody>
</table>

Dealing with concentration data reported as a negative value, zero, or below a detection limit (BDL) is problematic because of inconsistencies in detection limits and reporting practices, and any decision about handling these records introduces some bias (Stow et al., 2018). For example, using a non-numerical format such as BDL or <0.01 is likely to lead to the elimination of these entries during data analysis and thus would introduce a bias against low-value observations. Alternatively, converting any such value to zero would introduce a bias in the opposite direction. As a compromise solution, concentrations recorded as zero in the
original data source were entered as zero in GRiMeDB and other below-detection values were entered as -999999. In this latter case, the original data entry format was noted in the Comments column. For fluxes, negative and zero values were entered without modification or comment.

The Flux Table reports diffusive, ebullitive, and total CH₄ fluxes along with CO₂ and N₂O diffusive fluxes. Given the diverse strategies for measuring each of the three CH₄ flux pathways and potential biases associated with different approaches (Lorke et al., 2015; Chen et al., 2021), values are accompanied by brief categorical descriptions of methods used for each flux type as well as for CO₂ fluxes and the gas exchange coefficient k. For a small number of entries, CH₄ fluxes were not directly reported in the data source but information was available (dissolved gas concentration, temperature, and a corresponding k value) that allowed us to calculate these fluxes. We also entered BDL values for flux for one data source in which fluxes had been calculated from concentration, but fluxes associated with BDL concentrations had been omitted from the results. Finally, a small number of observations listed diffusive and ebullitive but not total fluxes, so diffusion and ebullition were summed and entered as total flux. In all cases, the added calculations are noted in the Comments field.

The GRiMeDB Concentration Table includes physicochemical measurements in support of concentration and flux observations (Figs. 1, 4, Table A3). Availability of this supplemental information varied widely among data sources, and was limited to data collected concurrently with gas samples. For data sources with gas fluxes and physicochemical data but not gas concentrations, we created rows in the Concentration Table to capture the supporting data. These records are identified by a “Yes” in the FluxYesNo column, SampleCount = 0, and NA in the CH₄mean column. Finally, water temperature was estimated for entries if it was needed to convert gas units and entered in the WaterTemp_degC_estimated column. Estimates were typically based on values from adjacent sites or the same site at a a similar time (e.g., averages of temperature from the prior and subsequent dates, or from the same month in an adjacent year). Error introduced from these estimates should be small; e.g., ca. <10% of the actual value if the estimated temperature is off by 3°C.

Following completion of all data entry, gas and physicochemical variables were converted to ‘new’ standard units (Tables A3, A4). The identities of the new and original units are included in both the Concentration Table and Flux Table for clarity. Elevation was used to estimate atmospheric pressure if needed for unit conversions. We used Henry’s Law, water temperature, and atmospheric pressure to convert dissolved gas values reported in ppm, ppb, µatm, and % saturation (~13% of observations). For observations that reported gas values as percent saturation (<1% of all observations), we also used the global average CH₄, CO₂, and the N₂O atmospheric concentrations from the NOAA Global Monitoring Laboratory (https://gml.noaa.gov/ccgg/) for the year 2013, which corresponds to median observation year in the database.
Figure 4: Workflow for entering and checking data for the GRiMeDB Concentration Table and Flux Table.

2.4 Assessment of representativeness

We assessed the representativeness of sites in GRiMeDB relative to the global distribution of biological, physical, and climatic properties following van den Hoogen et al. (2021). Briefly, we first assigned each site to a corresponding river reach in
HydroSHEDS (Linke et al., 2019), which is a global hydrological network database that contains spatial data for a wide array of hydrological, physiographical, climatic, land cover, geological, edaphic and anthropogenic variables for each river reach. HydroSHEDS thus provides a multidimensional characterization of global rivers that is well suited for assessing how representative GRiMeDB sites are in terms of key biophysical and anthropogenic features. After excluding non-numerical variables (e.g., biome) and variables with monthly values (e.g., monthly precipitation), we performed a principal component analysis (PCA) on all HydroSHEDS subcatchments using all possible combinations of the 54 remaining HydroSHEDS variables. From this, we selected all principal components (PCs) needed to explain 90% of the variance in the PCA, which corresponded to 28 PCs and 378 possible bivariate combinations of these PCs. For each unique PC pair, we computed the convex hull of all sampled sites to determine the distribution of these sites relative to all global river subcatchments for the specified PCs (Fig. 5). Each HydroSHEDS subcatchment was then assigned a value of 1 or 0 if it fell within or outside the convex hull, respectively. This process was repeated for each of the 378 possible PC combinations. To collapse this information, we calculated the fraction of cases that a given subcatchment fell within the convex hull for all PC combinations to obtain a dimensionless summary value ranging from 0 to 1. A subcatchment with a value of 1 for this index of “representativeness” means that it fell within the convex hull for 100% of the PC combinations, indicating that its overall characteristics are well captured in the database. It is important to note that this analysis only captures average catchment properties of relatively large river reaches (average subcatchment area: 130 km$^2$). Given the strong local controls on CH$_4$ concentrations and fluxes, interpretations from this analysis should be made with some caution.

Figure 5: Example of a representative principal components analysis (PCA) hexagon plot based on variability in HydroSHEDS river subcatchment attributes. Hexagon colour indicates the number of subcatchments per hexagon. Subcatchments hosting GRiMeDB sites are plotted in red and contained within the convex hull delineated by red lines. Subcatchments that fall within this polygon are assigned a value of 1 and those outside the perimeter are given a value of 0 to indicate the representativeness of sampled reaches for this pair of PC axes. See Sect. 2.4 for further explanation.
2.5 Data checking and data analysis

Several approaches were taken to check the accuracy of data in GRiMeDB. This included evaluation of the reliability of digitized data (Sect. 2.3) along with several additional inspection steps. Entries were error checked by a co-author other than the individual who entered the data, including confirmation of site location information, validating units for all variables, and spot- or complete checking of entered gas data (independent units and data check in Fig. 4), depending on dataset length and if data were manually entered or imported directly from a file. Once values had been converted to standard units, all variables were plotted to identify outliers (outlier check; Fig. 4), and extreme values were checked against the original data source. In cases in which errors were present in the original data, if possible, authors were contacted for clarification. In the few rare cases in which issues could not be resolved, the data were excluded. These and all other calculations and analyses were performed in R (version 4.2, R Core Team 2021), using the “dplyr” package (version 1.0.7, Wickham et al., 2021) and “data.table (Dowle and Srinivasin, 2021) for data analysis, “sf” package (version, 1.0, Pebesma, 2018) for spatial data processing, and “ggplot2” (version 3.3.5, Wickham, 2016) and patchwork (Pedersen, 2020) packages for visualization.

3. RESULTS

3.1 Overview of GRiMeDB data

GRiMeDB includes 24,024 records of CH₄ concentration and 8,205 CH₄ flux values from 5,037 unique sites, along with 17,655 and 8,409 concurrent measurements of concentration and 4,444 and 1,521 of flux for CO₂ and N₂O, respectively (Table S2). Although the first concentration and flux values in GRiMeDB were published in 1973 (Lamontagne et al., 1973) and 1987 (de Angelis and Lilley, 1987), respectively, over 70% of all CH₄ concentrations and 80% of flux observations became available after 2015 (the year of publication of MethDB; Fig. 6, Fig. S1). This growth in data availability has occurred predominantly along the spatial axis, as almost two thirds of all sites were added in or after 2015 and over half of all sites in the database have a single concentration and/or flux observation. Conversely, long timeseries are rare, with only 8% of the 5,037 sites having >10 concentration observations and 4% having >10 diffusive flux records (Fig. 6, Fig. S1). The longest concentration record includes 590 observations distributed over 28 years (Toolik Inlet, Site_ID 9025; Kling, 2019a; 2022) while the longest flux record has 82 observations of diffusive flux over 4 years (Site_ID 3644; Aho et al. 2021). Further, among the 15 sites with time series >5 years, 12 are situated in either the Toolik Lake region of Alaska, USA (Kling, 2019a; 2019b; 2022) or within the Krycklan watershed in Sweden (Wallin et al., 2018; Wallin, 2021).
Figure 6: Distribution of the number of methane (CH$_4$) observations per site. Brown bars indicate sites with both concentration (conc) and flux observations. Orange and green bars show sites with only concentration and only flux observations, respectively. Inset: Cumulative observations of CH$_4$ concentration and flux data based on the year of publication of the data source. The vertical line (2015) indicates the year of MethDB (Stanley et al., 2015) publication. See Fig. S1 for data accumulation and length resolved by CH$_4$ flux type.

3.2 Spatial and temporal distribution of data

Spatially, 40% of all sites and 52% of all CH$_4$ concentration observations are in North America, followed by Europe (25% of all sites and 26% of all CH$_4$ concentration values; Table S2). Conversely, there are vast geographic areas with moderate to high channel densities with few or no observations, such as central Canada, Central America, South America beyond the Amazon mainstem area, most of Russia, central and western Asia, New Zealand, and the Malay Archipelago (Fig. 7a). Geographic limitations in availability of flux data, particularly of ebullition, are pronounced given smaller number of observations and domination of diffusion measurements. Observations of ebullition are absent or limited to 1-2 studies for
Africa, Oceania, Central America, South America, and Russia (Fig. S2). Despite these gaps, there is surprisingly good representation in terms of the range of hydrological, physiographical, climatic, land cover, geological, edaphic, and anthropogenic conditions that exist globally (Fig. 7b). Areas that are poorly represented are characterized by very low channel density associated with arid or polar climates as well as high altitude regions (Greenland, northern Canada, northern Africa, central Australia, Middle Eastern nations, western China, Mongolia, Chile, southern Argentina). Evaluating the distribution or representativeness of sites in terms of system size is difficult given the limited availability of relevant information such as Strahler stream order or basin area, which were reported for only 26% and 28%, respectively, of all sites (Table S2). For sites with these data, counts of observations decline with increasing stream order (Fig. 8) in a log-linear fashion ($R^2 = 0.92$ for concentration and 0.90 for flux; $P <0.0005$ for both regressions after excluding zero-order counts), consistent with Horton’s Law of Stream Numbers (Horton, 1945). Thus, other than the extreme under-representation of zero-order channels, this predictable decline suggests reasonable representation by order. Nonetheless, this result should be interpreted with caution given the scarcity of relevant data. The distribution of counts by basin size follows a similar pattern of under-representation of sites draining very small basins and also indicates a potential over-representation of some large basin sizes (basins of ca. 10,000 km$^2$; Fig. 8).

The distribution of observations among months illustrates seasonal sampling regimes dominated by summer sampling in northern ($>40^\circ$) and southern ($< -20^\circ$) latitudes contrasted by even or erratic sampling at mid-latitudes (Fig. 9). Consistent with the lower representation of southern hemisphere rivers and streams, several months lack concentration and/or flux measurements south of $-10^\circ$ latitude, particularly during winter months. Beyond these gaps, the only months missing data in the northern hemisphere are fluxes in January and February at sites north of $60^\circ$ latitude and several missing months north of $70^\circ$, presumably due to pervasive ice and snow cover.
Figure 7: (a) Global distribution of methane (CH$_4$) observations in the database, colour coded for sites with concentration data only, flux data only, or both concentration and flux data. Top and right panels show, respectively, longitudinal and latitudinal patterns of the density of CH$_4$ observations (grey bars) and the density of river area (blue bars). These bars have been aggregated at a 1 latitudinal or longitudinal degree and rescaled from 0 to 1 for this visualization. River area was obtained from BasinAtlas (Linke et al. 2019). (b) Representativeness (dimensionless) of the database based on a wide array of biological, physical, hydrological and land cover variables (see Sect. 2.4 for details). Values close to 1 indicate a high representativeness, with only 4% of the global river surface below a threshold of 0.9. See Fig. S2 for data distribution resolved by CH$_4$ flux type.
Figure 8: Number of sites with concentration (a, b) or diffusive flux (c, d) observations as a function of stream order (a, c) and basin size (b, d) for the subset of sites with channel order and/or basin size information.
Figure 9: Number of observations of concentration (left) and flux (right) by month for 10° latitude bands.
3.3 CH₄ flux methodology

Records of CH₄ flux are dominated by diffusive flux measurements, which represent 85% of all flux values in the database, with ebullition (8%) and total flux (7%) accounting for the remaining entries (Fig. 10). Not surprisingly, a variety of methods have been used to quantify each flux type, although diffusive flux methods are dominated by calculations based on dissolved gas concentration and a gas exchange coefficient (k) (74% of all observations), while chamber-based methods are most common for quantifying total flux (93% of all observations). Similarly, k is most commonly estimated via physical models (n = 3188). Several models have been employed for this calculation, as indicated by >25 different references for k model sources listed in GRiMeDB.

All CH₄ flux observations (8205)

Diffusion (6997) → Concentration + k (5183)
  → Chamber (1810)
  → Other (4)
  → Chamber – [Concentration + k] (326)
  → Bubble rate + CH₄ content (266)
  → Departure from linear gas increase (28)
  → Other (1)
  → Chamber (545)
  → Chamber and ebullition (19)
  → [Concentration + k] and ebullition (18)
  → Mass balance (5)

Ebullition (621) → Physical model(s) (3188)
  → CO₂ chamber flux + dissolved concentration (884)
  → Physical model and tracer addition (19)
  → Tracer addition (221)
  → Assigned values (230)
  → Other (536)
  → Not available (105)

Total flux (587)

Figure 10: Counts of methane (CH₄) flux observations by type (left), by major methodological categories for each pathway (middle), and for method type used to estimate the gas exchange coefficient k (right). For clarity, the chamber category includes all chamber types and patterns of gas increase in the chamber unless specified; more resolved methodological data are presented in the GRiMeDB Flux Table. See Table A4 for further details about category definitions.

3.4 Overview of concentration and flux data

Concentrations and fluxes of all three gases are characterized by log-normal distributions that vary over several orders of magnitude (Fig. 11) and large coefficients of variation (CVs) for CH₄ and especially N₂O (Table 2). The vast majority (~95%) of CH₄ and CO₂ concentrations appear to be supersaturated, in contrast to N₂O concentrations with 67% of observations above this threshold. Reports of concentrations below detection are scarce for all gases, including N₂O (Table 2).
Figure 11: Histograms of gas concentrations and fluxes in GRiMeDB, excluding values reported as below detection or zero; counts of these values are reported in Table 3. Dashed vertical lines in the concentration histograms indicate the 100% saturation concentration based on the median estimated elevation (250 m) and water temperature (12.5) for all sites and atmospheric concentrations of 1.83, 400, and 0.325 ppm for methane (CH₄), carbon dioxide (CO₂), and nitrous oxide (N₂O), respectively.
The fraction of observations with zero, below detection, or negative fluxes (5%, 5%, and 19% for diffusive CH₄, CO₂, and N₂O fluxes, respectively), were similar to the frequency of subsaturated concentrations. At the other extreme, the highest CH₄ concentrations (>200 µmol L⁻¹) paradoxically occur in either anthropogenically influenced large rivers of the warm tropics (e.g., Amazon basin: Kemenes et al., 2007; Ganges, Mekong: Begum et al., 2021) or in small boreal headwater streams (e.g., Campeau et al., 2018; Wallin et al, 2018).

There were no meaningful univariate relationships between variables that may be used for upscaling (latitude, basin area, and stream order) and mean site concentration or flux (Fig. 12, Table S3). Linear regressions indicated that latitude and flux accounted for a very small percent of the variation in both concentration ($R^2 = 0.006$ and $0.002$, respectively) and flux ($R^2 = 0.036$ and $0.055$) among sites. Similarly, concentration and flux among stream orders suggested possible differences for concentration ($\chi^2 = 47.165$, $df = 8$, $P < 0.001$) and flux ($\chi^2 = 14.777$, $df = 8$, $P = 0.070$). However, results of corrected pairwise comparisons (using the method of Benjamini and Hochberg, 1995) among orders were ambiguous, suggesting no differences among orders for flux. For concentration, these comparisons indicated possible differences in distributions only between 7th order channels and all other orders, and between 6th vs 1st order sites for concentration. Collectively, these results indicate a lack of a consistent change in CH₄ magnitude across channel orders for flux.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Gas (and type)</th>
<th>Mean</th>
<th>Median</th>
<th>Max</th>
<th>Min</th>
<th>SD</th>
<th>CV</th>
<th>%BDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentration (µmol L⁻¹)</td>
<td>CH₄</td>
<td>1.49</td>
<td>0.20</td>
<td>456</td>
<td>0</td>
<td>10.69</td>
<td>718</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>CO₂</td>
<td>135</td>
<td>81.7</td>
<td>5,479</td>
<td>0</td>
<td>174.8</td>
<td>130</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>N₂O</td>
<td>0.058</td>
<td>0.017</td>
<td>32.9</td>
<td>0</td>
<td>0.602</td>
<td>1,042</td>
<td>0.59</td>
</tr>
<tr>
<td>Flux (mmol m⁻² d⁻¹)</td>
<td>CH₄-diffusive</td>
<td>7.31</td>
<td>0.44</td>
<td>4,057</td>
<td>-136</td>
<td>86.4</td>
<td>1,182</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CH₄-ebullitive</td>
<td>5.42</td>
<td>0.28</td>
<td>366</td>
<td>0</td>
<td>24.02</td>
<td>443</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CH₄- total</td>
<td>8.71</td>
<td>0.63</td>
<td>366</td>
<td>-0.05</td>
<td>31.90</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CO₂</td>
<td>319</td>
<td>128</td>
<td>23,749</td>
<td>-1625</td>
<td>770</td>
<td>242</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N₂O</td>
<td>0.082</td>
<td>0.008</td>
<td>31.3</td>
<td>-11.3</td>
<td>0.981</td>
<td>1,199</td>
<td></td>
</tr>
</tbody>
</table>
As with relationships between CH₄ and physical site attributes, relationships between CH₄ concentration or flux and water chemistry parameters are also characterized by substantial variability. Representative examples indicate increasing, decreasing, and ambiguous relationships between CH₄ concentrations and fluxes and selected chemical constituents (Fig. 13). One source of the variation in the relationship shown in Fig. 13 may be attributed to differences among sites, as is illustrated for the case of CH₄ concentration versus discharge. The cluster of points in this plot (Fig. 14a) does not suggest an obvious linear relationship between concentration and discharge; however, resolving the data to the site level for sites with multiple observations reveals several significant trends (Fig. 14b). Among 57 sites with >30 observations, 42 had significant relationships ($P < 0.05$) between concentration and discharge and 30 of these 42 trends were negative.

Median site concentrations for most categories of targeted channels (Fig. 15) differed from “normal” (NORM) sites (Kruskal-Wallis test $\chi^2 = 460.1$, $df = 12$, $P < 0.0001$ after dropping channel types with <10 observations to improve test reliability).
Pairwise Wilcoxon comparisons adjusted to account for multiple comparisons (Benjamini and Hochberg, 1995) indicated that springs (SP) and delta channels (DC) were similar to NORM sites ($P > 0.4$) and impoundment-influence (IMP) sites were marginally different ($P = 0.053$). Concentrations in channels at glacial termini (GT) and floodplain backwaters (FP) were lower ($P < 0.0001$), whereas all other site types had higher site average CH$_4$ concentrations than NORM sites. Fluxes also varied among channel type (Kruskal-Wallis test $\chi^2 = 126.4$, $df = 8$, $P < 0.0001$ after dropping channel types with <10 observations), and similar to concentration, fluxes in delta channels (DC) and channelized sites (CH) were similar to NORM channels while all other channel types considered had higher median fluxes.

Figure 13: Methane (CH$_4$) concentration (a-d) and diffusive flux (e-h) versus concurrent measures of dissolved oxygen (O$_2$; $n = 8,529$ and 2,316 for concentration and flux, respectively), dissolved organic carbon (DOC; $n = 14,441$ and 1,901), total nitrogen (Total N; $n = 8,378$ and 467) and total phosphorus (Total P; $n = 6,904$ and 240). Three outliers were excluded from the DOC plots, and because of the log scale for CH$_4$, negative and zero values have been omitted.
Figure 14: Methane (CH$_4$) concentration versus concurrent measures of discharge for (a) all sites with discharge data and (b) sites with >30 observations (57 sites) with trend lines denoting within-site relationships between concentration and discharge. Each site is represented by a separate colour. Because of the log scale for CH$_4$, negative and zero values are omitted.
4. Discussion

The rapid increase in availability of aquatic CH$_4$ (as well as CO$_2$ and N$_2$O) data over the past 5-10 years has been remarkable and creates new opportunities for examining patterns and drivers of these gases in lotic ecosystems across broad spatial scales. Similarly, constructing GRiMeDB provided us with an unprecedented opportunity to identify tendencies in when, where, and
how CH₄ has been sampled in streams and rivers. Examination of such data collection tendencies can reveal important biases and gaps within a field (Stanley et al. 2019, Gomez-Gener et al. 2021b) and thus points to future research needs and opportunities. Below, we discuss the distribution of sampling efforts and methodological issues, preliminary data analyses, and consider questions that GRiMeDB can help to answer.

4.1 When and where: sampling effort considerations

The growth of GHG studies in flowing water systems in the past decade includes a geographic expansion beyond the large body of historic and current work in temperate regions of North America and Europe. In particular, recent research in Africa, Australia, and especially southeast Asia has greatly improved the global coverage of available data. However, studies in arid drainages remain scarce—even beyond what would be expected given their small river surface area. A possible explanation for the limited study of CH₄ in these systems may be the pervasive focus on the contribution of streams and rivers to the global atmospheric CH₄ pool, and the corresponding assumption that aridland systems play a minor role in this context. Yet we suggest that limited study in arid and semi-arid drainages represents a missed opportunity to understand metabolism and carbon cycling in a set of streams and rivers that drain nearly half of the global land surface, are increasingly stressed by growing human water demands (e.g., Sabo et al., 2010; Lian et al., 2021; Stringer et al., 2022), and support ecosystem process rates that are amplified by warm temperatures and highly variable flow regimes (Fisher et al., 1982; Ran et al., 2021). Beyond arid and semi-arid basins, further research emphasis in tropical and high-latitude regions would also be beneficial even given recent improvements in data availability and geographic representation of both areas. Existing data for tropical forests and grasslands are dominated by studies of African rivers (especially the Congo drainage) and the Amazon River system. In fact, observations from tropical areas of the Indomalayan and northern Australasian region represent <3% of all sites, and Central America is represented by a single study. Tropical drainages are frequently characterized by high CH₄ concentrations and fluxes, along with rapid changes in land use and river regulation that are affecting C cycling and GHG dynamics (Park et al., 2018; Flecker et al. 2022). However, understanding or detecting the magnitude and consequences of these anthropogenic changes on fluvial CH₄ is constrained by these current sampling limitations. Finally, while high latitude regions (north of the Arctic Circle) are well represented in GriMeDB with >3,600 concentration observations, more than 80% of these values are derived from studies in the vicinity of the Toolik Field Station in Alaska, USA, and thus do not capture the full biophysical diversity of Arctic biomes (Metcalf et al. 2018). Given that climate change at high latitudes is progressing faster than elsewhere on the planet (IPCC, 2021), and that the global north stores massive quantities of C in soils (Hugelius et al., 2014), more extensive coverage of CH₄ across Arctic drainage systems is warranted.

Although the spatial coverage of CH₄ data has improved markedly over the past decade, expansion across temporal dimensions has lagged. The predominant mode of sample collection has been and continues to be through surveys that yield one or a few observations from individual sites (e.g., Bouillon et al., 2012; Kuhn et al., 2017; Jin et al., 2018; Ho et al., 2022), and studies characterizing seasonal dynamics or responses to a site-specific environmental change are limited. Indeed, long-term (>5 years)
CH\textsubscript{4} datasets in general are extremely rare (Leng et al., 2021); no such data are currently available for fluxes and most long-term concentration records are derived from just a few clustered locations. Determining the consequences of changes in land use or habitat attributes on fluvial CH\textsubscript{4} dynamics have instead relied on space-for-time substitutions (e.g., Smith et al., 2017; Gatti et al., 2018; Woda et al. 2020) rather than on direct observations of change over time. Although this strategy has been successful in revealing variation in GHG dynamics among different site types, current knowledge about how gases vary over time and respond to perturbations is poorly developed because of these data limitations. This deficit may be particularly consequential in the case of climate change, as the broad scope of this phenomenon will inevitably limit the effectiveness of spatial sampling approaches.

The discussion above regarding the ‘when’ and ‘where’ of sampling emphasizes large spatial and relatively long temporal scales, consistent with the extent of GRiMeDB. However, another current deficit in our understanding relates to the degree of heterogeneity of this gas at fine spatial and temporal scales, and thus if current sampling strategies are missing meaningful variation. Recent studies of CO\textsubscript{2} provide a cautionary tale in this context, as failure to account for diurnal variation in this gas results in a consistent under-estimation of fluvial emissions that is quantifiable at regional (Attermeyer et al., 2021) and global (Gómez-Gener et al., 2021b) scales. Similar questions may arise for spatial variation; that is, what is the minimum grain size or appropriate spatial scale for sampling CH\textsubscript{4} in running waters (Crawford et al., 2017; Lupon et al., 2019)? The potential to examine very short-term variation is not possible using GRiMeDB data because of our decision to average within-day measurements given the current small number (ca. 20) of these temporally detailed studies. Assessment of fine-scale spatial variation is also limited because of limited fine-scale sampling in general, as well as by decisions made both by investigators and during database construction. For example, geomorphologically distinct units (e.g., an individual riffle or pool) are often used as a basic sampling unit and results are presented as averages of replicates collected at different points within the study reach (e.g., Hlaváčová et al., 2006; Smith et al., 2017). In general, information about replication was frequently omitted, or if reported, information about variability among replicates was frequently absent. In addition to this limitation, our decision to combine replicates taken at different points in a channel cross-section and within individual channel units that had hundreds to thousands of datapoints to avoid ambiguities for site delineation and data analysis also constrains the opportunity to examine variation at fine spatial scales. However, we anticipate that this situation will change over the next few years, as in situ sensors or other devices capable of collecting high-frequency/high density gas measurements become more widely available. Recent papers signal this new frontier and have highlighted the presence (e.g., Lamarche-Gagnon et al., 2019; Smith and Bohlke, 2019; Chen et al., 2021; Taillardat et al., 2022) and absence (e.g., Castro-Morales et al., Chen et al., 2021; 2022; Rovelli et al., 2022; Zhang et al., 2021) of predictable diel variation in CH\textsubscript{4} concentrations and fluxes, and varying degrees of within-reach spatial variability (Crawford et al., 2016; 2017; Call et al., 2018; Bussman et al., 2022).
4.2 How: methodological considerations

Measuring dissolved GHG concentrations or fluxes involves multiple steps and calculations. Field and laboratory protocols vary widely in the literature, and methodological variety is particularly conspicuous for flux determination. Ironically, even though many studies of lotic CH$_4$ dynamics are framed in terms of understanding the contribution of these ecosystems to the rapidly increasing atmospheric CH$_4$ pool, flux measurements lag far behind those of concentration and the vast majority (ca. 85%) of observations are of the diffusive pathway alone. Further, the most common method for estimating this pathway involves combining dissolved CH$_4$ concentration with $k$, the gas exchange coefficient. Quantifying $k$ is notoriously challenging (Hall and Ulseth, 2019) and the large number of approaches for calculating $k$ used among data providers is concerning and undoubtedly introduces substantial uncertainty. A more in-depth consideration of the consequences of different models or strategies for arriving at a $k$ value was beyond the scope of this paper, but inclusion of methodological information should be useful for such an analysis in the future.

Ebullition measurements are notably scarce despite the potential of this pathway to account for a large fraction of total emissions in some streams (e.g., from 30-98% of total CH$_4$ emissions, as shown in Baulch et al., 2011; Crawford et al., 2014; Chen et al., 2021). The conventional approach to quantifying ebullition involves a combination of capturing bubbles just below the water surface to determine the area and time-specific rate of bubble volume reaching the surface and measuring CH$_4$ content of recently-erupted bubbles. The episodic nature and extreme spatial heterogeneity of ebullition (Crawford et al., 2014; Spawn et al., 2015; Chen et al., 2021; Robison et al., 2021) require multiple bubble trap replicates that need to be deployed over several days to generate reliable measurements. Given the logistic challenges and labour-intensive work involved, indirect approaches are becoming more common. These approaches typically use the difference between a chamber-based measurement of flux, which is assumed to represent total flux (diffusion + ebullition) and diffusion calculated from dissolved CH$_4$ and $k$ (i.e., the ‘chamber – [concentration + $k$]’ method in Fig. 10) to estimate ebullition (e.g., Campeau et al., 2014; Zhang et al., 2020; Ran et al., 2021). We suggest that this approach should be used cautiously, however. This strategy is arguably inappropriate for situations in which the gas content within a chamber increases in a linear fashion during the measurement period, consistent with the occurrence of diffusive flux alone. Second, it is not clear if it is reasonable to assume that chamber-based measurements capture both diffusion and ebullition, even if a chamber-based flux value is greater than that calculated from dissolved CH$_4$ concentration. Relatively short chamber deployments are likely to miss or incompletely capture bubble releases, while long-term deployments are vulnerable to sampling artefacts associated with altered concentration gradients within, and/or turbulence around the chamber (Sawakuchi et al., 2014; Lorke et al., 2015). Given these challenges, it is not altogether surprising that comparisons between direct and indirect measurements of ebullition can yield substantially different results (e.g., Yang et al., 2012; Bednářík et al., 2017; Chen et al., 2021).

The final and most profound knowledge gap in the collection of flux data is the absence of measurements of plant-mediated emissions. Plant-mediated fluxes can account for a substantial fraction of total emissions from wetlands and shallow lake...
habitats (Bodmer et al., 2021), but the contribution of this pathway is unknown in fluvial systems. Indeed, we did not include plant mediated fluxes in GRiMeDB because we encountered only two papers that had explicitly quantified this pathway in streams (Sanders et al., 2007; Wilcock and Sorrell, 2008). Although aquatic macrophytes are sparse or absent from many streams and rivers, they can be abundant in low-gradient, low-disturbance environments (Riis and Biggs, 2003; Gurnell et al., 2010) where diffusive fluxes would be constrained by low gas exchange rates. Sediment trapping and venting by macrophytes enhances both methanogenesis and methane emission in these systems (Sanders et al., 2007), but the significance of such processes and the contribution of plant-mediated fluxes at larger spatial scales remain to be determined for fluvial systems (Bodmer et al., 2021).

### 4.3 Concentration and flux patterns

Not surprisingly, the massive increase in data availability has led to differences in averages and measures of variability for CH$_4$ concentrations and fluxes compared to our previous efforts. Median values for all three CH$_4$ flux pathways in GRiMeDB are 1.2-2.2 times lower than those reported by Stanley et al. (2016), as well as those from Rosentreter et al. (2021). Conversely, measures of variability (SD, CV) in GRiMeDB are almost 3-fold greater than previous estimates, undoubtedly due to the far larger number of observations, the associated expansion of geographic scope and channel types, and the inclusion of temporally resolved data. It is not yet clear if the sample sizes are sufficient to capture the true global-scale variability of fluvial concentrations and fluxes, and future database updates could be used to examine this possibility.

Despite the slight lowering of median values compared to previous estimates, supersaturated concentrations and positive fluxes are the norm for CH$_4$ as well as for CO$_2$ and N$_2$O. However, it is likely that CH$_4$ concentrations and fluxes below detection limits (BDLs) are under-reported, as is common with environmental data in general (Stow et al., 2018), so these latest estimates may still be slight overestimations of true population medians. Even given the modest number of zero or undetectable CH$_4$ concentrations in GriMeDB (<2.5%), decisions about handling BDLs can have a small but detectable effect on the estimation of global averages. For example, if these observations are excluded, median CH$_4$ concentrations for all other observations increases from 1.49 to 1.51 µmol L$^{-1}$. If we keep all of these observations and assign them a value of zero (an unlikely scenario, but used here to provide a lower limit for this example), then the overall median declines to 1.46 µmol L$^{-1}$. Although these differences are relatively small, it would likely be consequential for upscaling estimates. At a minimum, we urge GRiMeDB users to be aware of how these values are handled and encourage future researchers to determine and report detection limits and include samples that fall below these limits in their results.

A goal of assembling GRiMeDB was to centralize CH$_4$ data to foster future research efforts. To this end, we included information about habitat conditions that allows the exploration of relationships between CH$_4$ and potential explanatory variables and covariates. To demonstrate this opportunity, we provided a limited number of examples of CH$_4$ versus variables that have been identified as potential predictors or drivers of CH$_4$ production, concentration, or flux (Figs. 12-14), and these
plots suggest both the presence and absence of relationships. For example, increasing CH$_4$ concentrations have been associated with low or decreasing dissolved oxygen and/or increasing organic carbon (e.g., Borges et al., 2018; Jin et al. 2018; Begum et al., 2021) and these relationships are recognizable for concentration but ambiguous for flux across the entirety of the GRiMeDB dataset. Similarly, increased CH$_4$ production and emissions tend to be elevated in nutrient-rich (eutrophicication) lakes (DelSontro et al., 2018) and polluted rivers (Rajkumar et al. 2008; Ho et al., 2022), consistent with positive relationships between CH$_4$ flux and TN and TP. However, nutrient enrichment in rivers often occurs concurrently with fine sediment and organic matter input; thus it remains to be determined if positive relationships in Figs. 13g and 13h are correlative or reflect a causal mechanism. Finally, increases in discharge have been linked to declines in gas concentration, likely due to source limitation (i.e., dilution) of terrestrial supply (Aho et al., 2021; Gómez- Gómez-Gener et al., 2021a) and/or greater water turbulence, which increases gas exchange and thus reduces supersaturated CH$_4$ stocks (Billett and Harvey, 2013; Kokic et al. 2018). This relationship is not obvious when all data were considered en masse, but became more apparent when examining within-site dynamics. In contrast to these three confirmatory examples, although latitude and channel size have also been identified as determinants of CH$_4$ concentrations or used to extrapolate site-specific gas measurements to larger (even global) scales (e.g., Bastviken et al., 2011; Li et al., 2020; Rosentreter et al., 2021), evidence for such relationships is absent from our analysis. Further, even for the former examples that indicated relationships between CH$_4$ concentration and DO, DOC, or discharge, there is substantial variability present in these relationships, the strength of these predictors is likely to vary across scales, and they explain little of the variability for diffusive fluxes. In short, substantial opportunities exist to identify multivariate relationships between different predictors and CH$_4$ concentrations and fluxes across different scales, and pursuit of these opportunities should be improved by the substantial increase in data for both gases and potential predictor variables.

The disproportionate contribution of streams and rivers to atmospheric inputs together with the utility of CH$_4$ as an indicator of anthropogenic influences on drainage systems have inspired several studies that focus on fluvial habitats that are expected to have high concentrations and fluxes. Many of these ‘methane hunting’ studies have demonstrated significant increases in CH$_4$ concentrations and/or fluxes associated with phenomena such as point source inputs (Alshboul et al., 2016) ditch and canal construction (Peacock et al., 2021), oil and gas extraction (Woda et al. 2020), or passage through wetlands (Taillardat et al., 2022). Such signals persist at the global scale (Fig. 15), highlighting widespread human enhancement of CH$_4$ emissions from lotic ecosystems. Not all targeted sites are CH$_4$-rich however. Low concentrations in glacial outflows (GT) likely reflect the effects of cold temperatures and/or low organic carbon availability (Crawford et al., 2015; Burns et al., 2018) while low values at floodplain (FP) sites may be attributable to their more characteristically lentic conditions, which favor higher rates of CH$_4$ oxidation in the water column. Indeed, oxidation has been shown to represent a significant CH$_4$ sink in floodplain lakes associated with the Amazon River (Barbosa et al., 2018) and most of the FP sites in GRiMeDB are part of the Amazon system.

As noted in Sect. 3.4, the availability of supporting information is inconsistent, as, for example, only ~25% of data sources provided data on channel order or basin size. However, open-access regional and global geospatial datasets that provide
information about site characteristics (e.g., Linke et al., 2019; Yang et al., 2020) have increased rapidly in the past decade, to the benefit of analyses seeking to link landscape attributes to CH₄ distribution among sites. Recent upscaling efforts analyses (Rosentreter et al., 2021; Liu et al., 2022; Rocher-Ros et al., accepted) have, for example, capitalized on improved estimates of the surface area of world streams and rivers (Allen and Pavelsky, 2018; Yang et al., 2020), while the diverse datasets in HydroSHEDS (Linke et al., 2019) allowed us to evaluate the global representativeness of GRiMeDB sites. As new global-scale datasets become available and become more spatially resolved, we anticipate that their pairing with GRiMeDB data will result in significant improvements in the strength and certainty of data assimilation models, regional to global-scale analyses of CH₄ distribution and drivers, and quantification of fluvial emissions to the atmosphere.

5. Data and code availability

GRiMeDB and its associated metadata are available from the the Environmental Data Initiative (Stanley et al., 2023): [https://doi.org/10.6073/pasta/f48cdb77282598052349e969920356ef](https://doi.org/10.6073/pasta/f48cdb77282598052349e969920356ef).

6. Conclusion

The data gathered in GRiMeDB highlight many new opportunities, both through analysis of CH₄ and supporting data in the database, and by revealing gaps that currently exist across fluvial CH₄ studies. The most conspicuous data limitations include deficits in measurements of non-diffusive flux pathways and underrepresentation of sites from arid, tropical, and arctic biomes. Challenges associated with quantifying ebullition discussed earlier also emphasize the need for more intercomparisons among flux methodologies. Regardless of pathway, flux is a difficult process to quantify and can be highly sensitive to methods or gas exchange model choices, yet there are few comparisons (such as Raymond et al., 2012; Lorke et al., 2015) available to inform these decisions. Finally, we highlight that the expansion of GHG data over the past decade has proceeded largely across spatial rather than temporal dimensions. While this expansion has vastly improved the geographic representativeness of the data, long-term datasets are rare despite their power for generating ecological understanding and informing policy/management in the face of environmental change (Hughes et al., 2017). GHG’s are rarely included as routine components of water quality monitoring programs. Thus, we emphasize the compelling need to establish such sampling efforts and perpetuate those few that do exist.

Given the rapid growth in both research interest and data in fluvial GHG dynamics, we imagine future updates and expansion of GRiMeDB and we welcome datasets and associated research products (e.g., theses, journal publications, reports, etc.). To facilitate the data acquisition and updating process, a downloadable spreadsheet template and detailed information about its use and submission are available at [https://stanley.limnology.wisc.edu/GRiMe](https://stanley.limnology.wisc.edu/GRiMe). Regardless of database updates, we recommend that the minimum basic information to collect along with GHG data that would be most valuable for later analyses include:
well-resolved site location data (latitude and longitude); information about site size (Strahler order and/or basin size at the sampling site), disturbance or modification relevant to GHGs (e.g., categories listed in Table 1); specific sample dates and times; discharge, dissolved oxygen, and temperature at the time of sample collection; and clear information about units and method(s) used to measure gas flux. Finally, we strongly encourage data package (sensu Gries et al., 2022) publication in a trustworthy public data repository such as the Environmental Data Initiative that requires metadata to meet FAIR data principles and increase data findability, accessibility, and re-use (Wilkinson et al., 2016).

Despite highlighting areas of data limitation, it is important to underscore the opportunities that the growth in GHG data availability—especially of CH$_4$ data—now provide. Assembly of GRiMeDB was motivated by the goal of having a centralized, standardized resource to facilitate further studies of CH$_4$ pattern and process in flowing water systems. Our strategy in developing this database was to maximize opportunities for identifying patterns and relationships involving this gas in future analyses. Past difficulties with such efforts may well be a product of the common practice of averaging values over time or among sites and/or of including non-fluvial sites in analyses. Thus, we carefully documented the data and resolved observations to individual sites and dates whenever possible to match the pronounced spatial and temporal variance of this gas. Similarly, while we included a range of habitat types in GRiMeDB, unconventional or targeted sites are easily identifiable. Further, we carefully examined sites to ensure that they were not situated within reservoirs/impoundments or estuaries where distinct processes such as methane oxidation, tidal cycles, or elevated sulphate reductions may obscure or overtake relationships present in inland flowing water systems. Thus, we are optimistic that analysis of GriMeDB data by itself, or in concert with other complementary datasets, will provide new and unprecedented opportunities to examine relationships between CH$_4$ and environmental drivers or correlates, as well as providing broad contextual information for site-based studies of fluvial carbon and GHG dynamics.
## Appendix A. GRiMeDB tables and variables

### Table A1. Column titles and description of their content for the GRiMeDB Sources Table.

<table>
<thead>
<tr>
<th>Column Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title of data source</td>
</tr>
<tr>
<td>Author</td>
<td>Lead author last name</td>
</tr>
<tr>
<td>Source</td>
<td>Identity of the outlet for the data (e.g., journal, data repository, agency that presented the data). For titles with published papers paired with published datasets, the journal is listed in this column</td>
</tr>
<tr>
<td>Pub_year</td>
<td>Year of publication, data release, or acquisition of an unpublished dataset</td>
</tr>
<tr>
<td>Source_ID</td>
<td>Unique data source identifier</td>
</tr>
<tr>
<td>Additional_data</td>
<td>“Yes” in this column indicates that additional data were acquired directly from the author for any field. Additions are described in the Comments field</td>
</tr>
<tr>
<td>Comments</td>
<td>Additional information or clarification about the data source</td>
</tr>
<tr>
<td>Paper.DOI</td>
<td>DOI or hyperlink for journal article or other publication based on the CH$_4$ data</td>
</tr>
<tr>
<td>Data.DOI.primary</td>
<td>DOI or hyperlink for CH$_4$ data posted in a data repository</td>
</tr>
<tr>
<td>Data.DOI.supporting</td>
<td>DOI or hyperlink for separate datasets providing supporting data</td>
</tr>
</tbody>
</table>
Table A2. Column titles and content description for the GRiMeDB Sites Table.

<table>
<thead>
<tr>
<th>Column Title</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source_ID</td>
<td>Unique data source identifier from the Sources Table</td>
</tr>
<tr>
<td>Site_ID</td>
<td>Unique site identifier</td>
</tr>
<tr>
<td>Site_Name</td>
<td>Unique site name</td>
</tr>
<tr>
<td>Stream_Name</td>
<td>Stream or river name; taken or modified from the data source or generated <em>de novo</em> when a name was not specified in the data source</td>
</tr>
<tr>
<td>Aggregated</td>
<td>Yes or No; “Yes” if CH₄ data entered are averages from &gt;1 site</td>
</tr>
<tr>
<td>N_sites_aggregated</td>
<td>Number of sites that were averaged for aggregated sites</td>
</tr>
<tr>
<td>Basin_Region</td>
<td>Name of the larger drainage basin or region that contains the site. This information is included to facilitate site grouping during data analysis</td>
</tr>
<tr>
<td>Latitude</td>
<td>Decimal degrees, WGS84 ensemble: EPSG:4326 coordinate system</td>
</tr>
<tr>
<td>Longitude</td>
<td>Decimal degrees, WGS84 ensemble: EPSG:4326 coordinate system</td>
</tr>
<tr>
<td>Elevation_m</td>
<td>Reported meters above sea level</td>
</tr>
<tr>
<td>Slope_m_per_m</td>
<td>Reported channel slope expressed as m m⁻¹</td>
</tr>
<tr>
<td>Strahler_order</td>
<td>Reported Strahler stream order</td>
</tr>
<tr>
<td>Basin_size_km²</td>
<td>Reported basin size in square kilometers</td>
</tr>
<tr>
<td>Channel_type</td>
<td>Codes denoting distinct site or channel attributes or presence of specified conditions. See Table 1 for categories and their definitions</td>
</tr>
<tr>
<td>Latitude_snapped</td>
<td>Latitude in decimal degrees for site location after snapping to the closest channel for elevation determination</td>
</tr>
<tr>
<td>Longitude_snapped</td>
<td>Longitude in decimal degrees for site location after snapping to the closest channel for elevation determination</td>
</tr>
<tr>
<td>Elevation_estimated_m</td>
<td>Elevation (meters above sea level) calculated from the DEM. See Sect. 2.2 for details</td>
</tr>
<tr>
<td>Comments</td>
<td>Additional information or clarification about the site source</td>
</tr>
</tbody>
</table>
## Table A3. Column titles and definitions for the GRiMeDB Concentration Table

<table>
<thead>
<tr>
<th>Column Title</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source_ID</td>
<td>Unique paper identifier from the Sources Table</td>
</tr>
<tr>
<td>Site_ID</td>
<td>Unique site identifier from the Sites Table</td>
</tr>
<tr>
<td>Site_Name</td>
<td>Unique site name from the Sites Table</td>
</tr>
<tr>
<td>Conc_Name</td>
<td>Unique name for the sampling event at the site; same as Flux_Name in the Flux Table if both concentration and flux data for the same site-date combination are available</td>
</tr>
<tr>
<td>Date_start</td>
<td>First sampling date</td>
</tr>
<tr>
<td>Date_end</td>
<td>Last sampling date; this is the same date as the Date_start if data are not aggregated over time</td>
</tr>
<tr>
<td>Aggregated_Space</td>
<td>Yes or No; “Yes” if CH₄ data entered are averages from &gt;1 site</td>
</tr>
<tr>
<td>Aggregated_Time</td>
<td>Yes or No; “Yes: if CH₄ data entered are averages from &gt;1 date</td>
</tr>
<tr>
<td>FluxYesNo</td>
<td>Yes or No; “Yes” if there is a corresponding flux measurement associated with this site-date combination</td>
</tr>
<tr>
<td>SampleCount</td>
<td>Number of samples or observations corresponding to the mean or median concentration</td>
</tr>
<tr>
<td>CH4min</td>
<td>Minimum measured CH₄ concentration in µmol L⁻¹ if data are aggregated spatially or temporally, has multiple within-day measurements (e.g., a diel study), or are from a data-dense spatial study</td>
</tr>
<tr>
<td>CH4max</td>
<td>Maximum measured CH₄ concentration in µmol L⁻¹ if data are aggregated spatially or temporally, has multiple within-day measurements (e.g., a diel study), or are from a data-dense spatial study</td>
</tr>
<tr>
<td>CH4mean</td>
<td>Mean or sole reported CH₄ concentration in µmol L⁻¹ for the sampling event</td>
</tr>
<tr>
<td>CH4_SD</td>
<td>Standard deviation of the mean CH₄ concentration</td>
</tr>
<tr>
<td>CH4_median</td>
<td>Median CH₄ concentration in µmol L⁻¹</td>
</tr>
<tr>
<td>CO2min</td>
<td>Minimum measured CO₂ concentration in µmol L⁻¹ if data are aggregated spatially or temporally, has multiple within-day measurements (e.g., a diel study), or are from a data-dense spatial study</td>
</tr>
<tr>
<td>CO2max</td>
<td>Maximum measured CO₂ concentration in µmol L⁻¹ if data are aggregated spatially or temporally, has multiple within-day measurements (e.g., a diel study), or are from a data-dense spatial study</td>
</tr>
<tr>
<td>CO2mean</td>
<td>Mean or sole reported CO₂ concentration in µmol L⁻¹ for the sampling event</td>
</tr>
<tr>
<td>CO2_SD</td>
<td>Standard deviation of the mean concentration</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>CO2median</td>
<td>Median CO₂ concentration in µmol L⁻¹</td>
</tr>
<tr>
<td>N2Omin</td>
<td>Minimum measured N₂O concentration in µmol L⁻¹ if data are aggregated spatially or temporally, has multiple within-day measurements (e.g., a diel study), or are from a data-dense spatial study</td>
</tr>
<tr>
<td>N2Omax</td>
<td>Maximum measured N₂O concentration in µmol L⁻¹ if data are aggregated spatially or temporally, has multiple within-day measurements (e.g., a diel study), or are from a data-dense spatial study</td>
</tr>
<tr>
<td>N2Omean</td>
<td>Mean or sole reported N₂O concentration in µmol L⁻¹ for the concentration for the sampling event</td>
</tr>
<tr>
<td>N2O_SD</td>
<td>Standard deviation of the mean N₂O concentration</td>
</tr>
<tr>
<td>N2Omedian</td>
<td>Median N₂O concentration in µmol L⁻¹</td>
</tr>
<tr>
<td>WaterTemp_degC</td>
<td>Water temperature in degrees C measured concurrently with CH₄</td>
</tr>
<tr>
<td>WaterTemp_degC_estimated</td>
<td>Estimated water temperature in degrees C. This field was populated only for cases in which temperature was needed for gas unit conversion. Most estimates were based on temperatures from adjacent sites, averaging temperatures from prior and proceeding sample dates, or from an adjacent day of the year but from another year.</td>
</tr>
<tr>
<td>Cond_uScm</td>
<td>Specific conductance in µS cm⁻¹</td>
</tr>
<tr>
<td>pH</td>
<td>pH</td>
</tr>
<tr>
<td>DO_mgL</td>
<td>Dissolved oxygen in mg L⁻¹</td>
</tr>
<tr>
<td>DO_percentsat</td>
<td>Percent saturation of dissolved oxygen</td>
</tr>
<tr>
<td>Q</td>
<td>Discharge in m³ s⁻¹ measured at the time of sample collection</td>
</tr>
<tr>
<td>NO₃</td>
<td>NO₃ or NO₂+NO₃ concentration in µmol L⁻¹ measured concurrently with CH₄</td>
</tr>
<tr>
<td>NH₄</td>
<td>NH₄ concentration in µmol L⁻¹ measured concurrently with CH₄</td>
</tr>
<tr>
<td>TN</td>
<td>TN or TDN concentration in µmol L⁻¹ measured concurrently with CH₄</td>
</tr>
<tr>
<td>SRP</td>
<td>SRP or PO₄ concentration in µmol L⁻¹ measured concurrently with CH₄</td>
</tr>
<tr>
<td>TP</td>
<td>TP or TDP concentration in µmol L⁻¹ measured concurrently with CH₄</td>
</tr>
<tr>
<td>DOC</td>
<td>DOC or TOC concentration in µmol L⁻¹ measured concurrently with CH₄</td>
</tr>
<tr>
<td>Comments</td>
<td>Any additional relevant information regarding data</td>
</tr>
<tr>
<td>new_CH4_unit</td>
<td>Current common units for all CH₄ concentrations</td>
</tr>
<tr>
<td>new_CO2_unit</td>
<td>Current common units for all CO\textsubscript{2} concentrations</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>new_N2O_unit</td>
<td>Current common units for all N\textsubscript{2}O concentrations</td>
</tr>
<tr>
<td>new_NO3_unit</td>
<td>Current common units for all NO\textsubscript{3} or NO\textsubscript{2}+NO\textsubscript{3} concentrations</td>
</tr>
<tr>
<td>new_NH4_unit</td>
<td>Current common units for all NH\textsubscript{4} concentrations</td>
</tr>
<tr>
<td>new_TN_unit</td>
<td>Current common units for all TN or TDN concentrations</td>
</tr>
<tr>
<td>new_SRP_unit</td>
<td>Current common units for all SRP or PO\textsubscript{4} concentrations</td>
</tr>
<tr>
<td>new_TP_unit</td>
<td>Current common units for all TP or TDP concentrations</td>
</tr>
<tr>
<td>new_DOC_unit</td>
<td>Current common units for all DOC or TOC concentrations</td>
</tr>
<tr>
<td>new_Q_unit</td>
<td>Current common units for all discharge measurements</td>
</tr>
<tr>
<td>orig_CH4_unit</td>
<td>Original units for CH\textsubscript{4} concentration</td>
</tr>
<tr>
<td>orig_CO2_unit</td>
<td>Original units for CO\textsubscript{2} concentration</td>
</tr>
<tr>
<td>orig_N2O_unit</td>
<td>Original units for N\textsubscript{2}O concentration</td>
</tr>
<tr>
<td>orig_NO3_unit</td>
<td>Original units for NO\textsubscript{3} or NO\textsubscript{2}+NO\textsubscript{3} concentration</td>
</tr>
<tr>
<td>orig_NH4_unit</td>
<td>Original units for NH\textsubscript{4} concentration</td>
</tr>
<tr>
<td>orig_TN_unit</td>
<td>Original units for TN concentration</td>
</tr>
<tr>
<td>orig_SRP_unit</td>
<td>Original units for SRP or PO\textsubscript{4} concentration</td>
</tr>
<tr>
<td>orig_TP_unit</td>
<td>Original units for TP concentration</td>
</tr>
<tr>
<td>orig_DOC_unit</td>
<td>Original units for DOC concentration</td>
</tr>
<tr>
<td>orig_Q_unit</td>
<td>Original units of discharge</td>
</tr>
</tbody>
</table>
### Table A4. Column titles and definitions for the GRiMeDB Flux Table

<table>
<thead>
<tr>
<th>Column Title</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source_ID</strong></td>
<td>Unique paper identifier from the Sources Table</td>
</tr>
<tr>
<td><strong>Site_ID</strong></td>
<td>Unique site identifier from the Sites Table</td>
</tr>
<tr>
<td><strong>Site_Name</strong></td>
<td>Unique site name from the Sites Table</td>
</tr>
<tr>
<td><strong>Flux_Name</strong></td>
<td>Unique name for the sampling event at the site; same as Conc_Name in the Concentration Table if both concentration and flux data for the same site-date combination are available</td>
</tr>
<tr>
<td><strong>Date_start</strong></td>
<td>First sampling date</td>
</tr>
<tr>
<td><strong>Date_end</strong></td>
<td>Last sampling date; this is the same date as the Date_start if data are not aggregated over time</td>
</tr>
<tr>
<td><strong>Aggregated_Space</strong></td>
<td>Yes or No; “Yes” if CH4 data entered are averages from &gt;1 site</td>
</tr>
<tr>
<td><strong>Aggregated_Time</strong></td>
<td>Yes or No; “Yes: if CH4 data entered are averages from &gt;1 date</td>
</tr>
<tr>
<td><strong>Diffusive_CH4_Flux_Min</strong></td>
<td>Minimum measured CH4 diffusive flux in mm m(^{-2}) d(^{-1}) if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td><strong>Diffusive_CH4_Flux_Max</strong></td>
<td>Maximum measured CH4 diffusive flux in mm m(^{-2}) d(^{-1}) if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td><strong>Diffusive_CH4_Flux_Mean</strong></td>
<td>Mean or sole reported CH4 diffusive flux in mm m(^{-2}) d(^{-1}) for the sampling event</td>
</tr>
<tr>
<td><strong>Diffusive_CH4_Flux_SD</strong></td>
<td>Standard deviation of the mean CH4 diffusive flux</td>
</tr>
<tr>
<td><strong>Diffusive_CH4_Flux_Median</strong></td>
<td>Median CH4 diffusive flux in mm m(^{-2}) d(^{-1})</td>
</tr>
<tr>
<td><strong>SampleCount_Diffusive</strong></td>
<td>Number of samples or observations corresponding to the mean or median diffusive CH4 flux</td>
</tr>
<tr>
<td><strong>Diff_Method</strong></td>
<td>Methodological category used to measure diffusive gas flux. Categories (with brief explanations in italics) are:</td>
</tr>
<tr>
<td></td>
<td><strong>chamber (unspecified)- unspecified response</strong></td>
</tr>
<tr>
<td></td>
<td>use of an unspecified type of chamber (suspended, tethered, or free-floating) and pattern of change gas concentration over time during flux measurements is also not specified</td>
</tr>
<tr>
<td></td>
<td><strong>chamber (unspecified)- linear response</strong></td>
</tr>
<tr>
<td></td>
<td>unspecified type of chamber with a linear increase in chamber gas concentration over time or use of a linear model to calculate flux</td>
</tr>
<tr>
<td></td>
<td><strong>suspended/tethered chamber-unspecified response</strong></td>
</tr>
<tr>
<td></td>
<td>chamber is restrained to maintain its position and not float downstream during flux measurement</td>
</tr>
<tr>
<td></td>
<td><strong>suspended/tethered chamber- linear response</strong></td>
</tr>
<tr>
<td></td>
<td><strong>floating chamber- unspecified response</strong></td>
</tr>
<tr>
<td></td>
<td>chamber is unrestrained and is able to float downstream during flux measurement</td>
</tr>
<tr>
<td></td>
<td><strong>floating chamber- linear response</strong></td>
</tr>
<tr>
<td></td>
<td><strong>conc+k</strong></td>
</tr>
<tr>
<td></td>
<td>diffusive flux calculated using the equation:</td>
</tr>
<tr>
<td></td>
<td>flux = (k(C_w-C_{eq})), where</td>
</tr>
<tr>
<td></td>
<td>(k) = gas exchange coefficient</td>
</tr>
<tr>
<td></td>
<td>(C_w) = CH4 concentration measured in water</td>
</tr>
</tbody>
</table>
Table A4. Continued.

<table>
<thead>
<tr>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceq = CH₄ concentration in water in equilibrium with the atmosphere</td>
<td></td>
</tr>
<tr>
<td>Other methods other than those described above</td>
<td></td>
</tr>
<tr>
<td>Eb_CH4_Flux_Min</td>
<td>Minimum measured CH₄ ebullitive flux in mm m⁻² d⁻¹ if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td>Eb_CH4_Flux_Max</td>
<td>Maximum measured CH₄ ebullitive flux in mm m⁻² d⁻¹ if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td>Eb_CH4_Flux_Mean</td>
<td>Mean or sole reported CH₄ ebullitive flux in mm m⁻² d⁻¹ for the sampling event</td>
</tr>
<tr>
<td>Eb_CH4_Flux_SD</td>
<td>Standard deviation of the mean CH₄ ebullitive flux</td>
</tr>
<tr>
<td>Eb_CH4_Flux_Median</td>
<td>Median CH₄ ebullition flux in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>SampleCount_Eb</td>
<td>Number of samples or observations corresponding to the mean or median ebullitive CH₄ flux</td>
</tr>
<tr>
<td>Echoosounder + bubble analysis</td>
<td></td>
</tr>
<tr>
<td>gas bubble volume determined using echosounder and combined with CH₄ content of recently collected bubbles</td>
<td></td>
</tr>
<tr>
<td>Departure from linear increase during measurement</td>
<td></td>
</tr>
<tr>
<td>non-linear change in gas concentrations during chamber-based flux measurements taken as evidence of ebullition; various approaches used to quantify ebullition from these departures</td>
<td></td>
</tr>
<tr>
<td>Other methods other than those described above</td>
<td></td>
</tr>
<tr>
<td>Total_CH4_Flux_Min</td>
<td>Minimum measured total CH₄ flux in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>Total_CH4_Flux_Max</td>
<td>Maximum measured total CH₄ flux in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>Total_CH4_Flux_Mean</td>
<td>Mean or sole reported total CH₄ flux for the sampling event in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>Total_CH4_Flux_SD</td>
<td>Standard deviation of the mean total CH₄ flux</td>
</tr>
<tr>
<td>Total_CH4_Flux_Median</td>
<td>Median measured total CH₄ flux in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>Total_Method</td>
<td>Methodological category used to measure total CH₄ flux. Categories (with brief explanations in italics) are:</td>
</tr>
<tr>
<td>Conc+k and ebullition</td>
<td></td>
</tr>
<tr>
<td>total flux calculated as the sum of separate measurements of diffusion determined by the Conc+k method plus ebullition determined from the bubble trap or echosounder approach combined with bubble CH₄ analysis</td>
<td></td>
</tr>
<tr>
<td>Floating chamber</td>
<td></td>
</tr>
<tr>
<td>free-floating chamber is assumed to capture diffusive flux and ebullitive flux (if present)</td>
<td></td>
</tr>
<tr>
<td>Suspended/tethered chamber</td>
<td></td>
</tr>
<tr>
<td>suspended or tethered chamber is assumed to capture diffusive flux and ebullitive flux (if present)</td>
<td></td>
</tr>
<tr>
<td>Chamber and ebullition</td>
<td></td>
</tr>
<tr>
<td>total flux calculated as the sum of separate measurements of diffusion determined using a floating or suspended/tethered chamber plus ebullition determined from the bubble trap or echosounder approach combined with bubble CH₄ analysis</td>
<td></td>
</tr>
</tbody>
</table>
mass balance

total flux represents the difference between all measured inputs to a reach (e.g., dissolved CH₄ from upstream flow, groundwater discharge, and methanogenesis) minus all outputs other than efflux to the atmosphere (e.g., downstream export, methane oxidation)

other

methods other than those described above

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO2_Flux_Min</td>
<td>Minimum measured CO₂ flux in mm m⁻² d⁻¹ if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td>CO2_Flux_Max</td>
<td>Maximum measured CO₂ flux in mm m⁻² d⁻¹ if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td>CO2_Flux_Mean</td>
<td>Mean or sole reported CO₂ diffusive flux in mm m⁻² d⁻¹ for the sampling event</td>
</tr>
<tr>
<td>CO2_Flux_SD</td>
<td>Standard deviation of the mean CO₂ flux</td>
</tr>
<tr>
<td>CO2_Flux_Median</td>
<td>Median CO₂ flux in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>N2O_Flux_Min</td>
<td>Minimum measured N₂O flux in mm m⁻² d⁻¹ if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td>N2O_Flux_Max</td>
<td>Maximum measured N₂O flux in mm m⁻² d⁻¹ if data are aggregated or are from diel or data-dense spatial studies</td>
</tr>
<tr>
<td>N2O_Flux_Mean</td>
<td>Mean or sole reported N₂O diffusive flux in mm m⁻² d⁻¹ for the sampling event</td>
</tr>
<tr>
<td>N2O_Flux_Stddev</td>
<td>Standard deviation of the mean N₂O flux</td>
</tr>
<tr>
<td>N2O_Flux_Median</td>
<td>Median N₂O flux in mm m⁻² d⁻¹</td>
</tr>
<tr>
<td>k_Method</td>
<td>Methodological category used for estimating the gas exchange coefficient, k, Categories (with brief explanations in italics) are:</td>
</tr>
<tr>
<td></td>
<td>physical model</td>
</tr>
<tr>
<td></td>
<td>k calculated using equations based on physical variables such as channel slope, water velocity, etc.</td>
</tr>
<tr>
<td></td>
<td>chamber + conc</td>
</tr>
<tr>
<td></td>
<td>k determined by chamber-based measurements of flux, dissolved gas concentration, and re-arrangement of the flux equation</td>
</tr>
<tr>
<td></td>
<td>flux = k(C_w - C_eq)</td>
</tr>
<tr>
<td></td>
<td>to solve for k. Typically, these measurements are made for CO₂, and then k_CO₂ is converted to k_CH₄</td>
</tr>
<tr>
<td></td>
<td>tracer addition</td>
</tr>
<tr>
<td></td>
<td>paired conservative and gas tracer additions used to calculate k from concentration declines along a stream reach</td>
</tr>
<tr>
<td></td>
<td>assigned k value</td>
</tr>
<tr>
<td></td>
<td>use of k values from other dates or sites in the same study or k values considered to be characteristic of the site</td>
</tr>
<tr>
<td></td>
<td>other</td>
</tr>
<tr>
<td></td>
<td>methods other than those described above</td>
</tr>
<tr>
<td></td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>method to determine k is not described</td>
</tr>
</tbody>
</table>

k_ref                  | k method citation reported in the data source         |

Comments               | Any additional relevant information regarding data entered in this row |

new_Diffusive_Flux_unit| Current common units for all diffusive CH₄ flux data |

new_Eb_CH4_Flux_unit   | Current common units for all ebullitive CH₄ flux data |
<table>
<thead>
<tr>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>new_Total_Flux_unit</td>
<td>Current common units for all total CH$_4$ flux data</td>
</tr>
<tr>
<td>new_CO2_Flux_unit</td>
<td>Current common units for all CO$_2$ flux data</td>
</tr>
<tr>
<td>new_N2O_Flux_unit</td>
<td>Current common units for all N$_2$O flux data</td>
</tr>
<tr>
<td>orig_Diffusive_Flux_unit</td>
<td>Original units for diffusive CH$_4$ flux</td>
</tr>
<tr>
<td>orig_Eb_CH4_Flux_unit</td>
<td>Original units for ebullitive CH$_4$ flux used</td>
</tr>
<tr>
<td>orig_Total_Flux_unit</td>
<td>Original units for total CH$_4$ flux</td>
</tr>
<tr>
<td>orig_CO2_Flux_unit</td>
<td>Original units for CO$_2$ flux</td>
</tr>
<tr>
<td>orig_N2O_Flux_unit</td>
<td>Original units for N$_2$O flux</td>
</tr>
</tbody>
</table>
Appendix B. Citations for data sources in GRiMeDB, including citations for unpublished datasets. Dates for unpublished dataset correspond to the year the data were provided by data authors.


Aho, K., Cawley, K., DelVecchia, A., Stanley, E., and Raymond, P.: Dissolved greenhouse gas concentrations derived from the NEON dissolved gases in surface water data product (DP1.20097.001), Environmental Data Initiative, https://doi.org/10.6073/pasta/47d7cb6d374b6662cce98e42122169f8, 2021a


Bowden, W. B.: Arctic LTER streams chemistry Toolik Field Station, Alaska 1978 to 2019., Environmental Data Initiative, \url{https://doi.org/10.6073/pasta/3faacd18b63b3bacc5a0dbd6f09660e1}, 2021.


Dinsmore, K. J., Billett, M. F., and Dyson, K. E.: Five year record of aquatic carbon and greenhouse gas concentrations from Auchencorth Moss, NERC Environmental Information Data Centre, https://doi.org/10.5285/3f0820a7-a8c8-4dd7-a058-8db79ba9c7fe, 2013a.


Dodd, A.: Flow regime influences on stream and riparian soil carbon dynamics in the Ozark Highlands and Boston Mountains of Arkansas, Ph.D., University of Arkansas, Fayetteville, AR, USA, 143 pp., https://scholarworks.uark.edu/etd/2911/, 2018.


Kling, G.: Toolik Inlet discharge data collected in summer 1996, Arctic LTER, Toolik Research Station, Alaska., Environmental Data Initiative, https://doi.org/10.6073/pasta/6e9d9bd807d8ec133e91d0e665a1550d, 2016d.

Kling, G.: Toolik Inlet discharge data collected in summer 1997, Arctic LTER, Toolik Research Station, Alaska., Environmental Data Initiative, https://doi.org/10.6073/pasta/33f027ad109d650964a0a084e5df7b11, 2016e.


Kling, G.: Tussock Watershed stream discharge, electrical conductivity, and temperature measurements from 1992, Environmental Data Initiative, https://doi.org/10.6073/pasta/1e224958e278841f9a7a035007c65f21, 2016r.


Kling, G.: Tussock Watershed stream discharge, electrical conductivity, and temperature measurements from 1999, Environmental Data Initiative, https://doi.org/10.6073/pasta/4b943b5a2de08aca8b7dd4854276f12, 2016x.

Kling, G.: Tussock Watershed stream discharge, electrical conductivity, and temperature measurements from 2000, Environmental Data Initiative, https://doi.org/10.6073/pasta/53a45c5a110f0af13c5ae0ed3154b8ca, 2016y.


Kling, G.: Tussock Watershed stream discharge, electrical conductivity, and temperature measurements from 2003, Environmental Data Initiative, https://doi.org/10.6073/pasta/b24b8bb901a4b1b825e09c7ab494b39d, 2016aa.


Li, L., Yan, R., and Xue, B.: Methane levels of a river network in Wuxi City, China and response to water governance, Water, 12, 2617, https://doi.org/10.3390/w12092617, 2020a.


Taniwaki, R. H.: Methane concentrations and fluxes in agricultural and preserved tropical headwater streams, Environmental Data Initiative, https://doi.org/10.6073/pasta/82b1aa19ad5b4ef88939303f79c1e74c, 2022.


Wynn, P.: Direct isotopic evidence of biogenic methane production and efflux from beneath a temperate glacier, Lancaster University, https://doi.org/10.17635/lancaster/researchdata/246, 2018b.


Author contributions.
EHS conceived of the project idea, and led data entry, manuscript preparation, and data curation. LCL developed the code used for unit conversions, was responsible for data conversion and QA/QC, and contributed to data visualization, data analysis, and code curation. GRR was responsible for spatial analyses, and contributed to data visualization, code curation, and manuscript preparation. The structure and composition of the manuscript were resulted from collaborative discussions among EHS, GRR, LCL, NJC, SKO, and RAS. All authors contributed to data acquisition, data entry, data checking, and substantial manuscript revising and editing. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Competing Interests.
The authors declare that they have no conflict of interest.

Acknowledgements.
We are grateful to all the authors who generously shared data or made their data public; our science is better or this transparency. We are also grateful to several of these investigators for patiently and thoughtfully responding to emails from E. Stanley. Corinna Gries provided advice and assistance with data publication. Stanley thanks K. Forest and K. McMahon for guidance during manuscript preparation. Great thanks to M. W. Diebel, S. N.Topp, Y. Yao, K. Castro-Morales, and B. Deemer for their thoughtful and detailed comments that improved this paper.

Financial Support.
Emily H. Stanley acknowledges support from the National Science Foundation #DEB-2025982, NTL LTER. Nora J. Casson was supported in part by the Canada Research Chair program. Gerard Rocher-Ros was supported by the Swedish Research Council (2021-06667) and the Department of Ecology and Environmental Science, Umeå University. Marcus B. Wallin was supported by the Swedish Research Council’s VR (2021-04058) and Formas (2019-01105). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.
References


R Core Team: R: A language and environment for statistical computing, 2021.


