ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Scaling waterbody carbon dioxide and methane fluxes in the arctic using an integrated terrestrial-aquatic approach

To cite this article: Sarah M Ludwig *et al* 2023 *Environ. Res. Lett.* **18** 064019

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Scaling waterbody carbon dioxide and methane fluxes in the arctic using an integrated terrestrial-aquatic approach

Sarah M Ludwig[1,2,*], **Susan M Natali**[3], **John D Schade**[3], **Margaret Powell**[4], **Greg Fiske**[3], **Luke D Schiferl**[2,5] and **Roisin Commane**[1,2]

[1] Department of Earth and Environmental Science, Columbia University, New York, NY, United States of America
[2] Lamont-Doherty Earth Observatory, Palisades, NY, United States of America
[3] Woodwell Climate Research Center, Woods Hole, MA, United States of America
[4] Department of Earth and Planetary Sciences, Harvard College, Cambridge, MA, United States of America
[5] Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: Ludda.ludwig@columbia.edu

Keywords: carbon, scaling, methane, lake, arctic

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
11 December 2022

REVISED
2 May 2023

ACCEPTED FOR PUBLICATION
11 May 2023

PUBLISHED
19 May 2023

Abstract

In the Arctic waterbodies are abundant and rapid thaw of permafrost is destabilizing the carbon cycle and changing hydrology. It is particularly important to quantify and accurately scale aquatic carbon emissions in arctic ecosystems. Recently available high-resolution remote sensing datasets capture the physical characteristics of arctic landscapes at unprecedented spatial resolution. We demonstrate how machine learning models can capitalize on these spatial datasets to greatly improve accuracy when scaling waterbody CO_2 and CH_4 fluxes across the YK Delta of south-west AK. We found that waterbody size and contour were strong predictors for aquatic CO_2 emissions, attributing greater than two-thirds of the influence to the scaling model. Small ponds (<0.001 km^2) were hotspots of emissions, contributing fluxes several times their relative area, but were less than 5% of the total carbon budget. Small to medium lakes (0.001–0.1 km^2) contributed the majority of carbon emissions from waterbodies. Waterbody CH_4 emissions were predicted by a combination of wetland landcover and related drivers, as well as watershed hydrology, and waterbody surface reflectance related to chromophoric dissolved organic matter. When compared to our machine learning approach, traditional scaling methods that did not account for relevant landscape characteristics overestimated waterbody CO_2 and CH_4 emissions by 26%–79% and 8%–53% respectively. This study demonstrates the importance of an integrated terrestrial-aquatic approach to improving estimates and uncertainty when scaling C emissions in the arctic.

1. Introduction

Regions of permafrost soils, perennially frozen ground, store approximately twice as much carbon as is currently in the entire atmosphere [1, 2]. With accelerated warming in high latitudes, permafrost is thawing across the Arctic, leading to increasing emissions of CO_2 and CH_4 [3–6] as it is decomposed directly to CO_2 and CH_4 or transported through landscapes to inland waterbodies [4, 7–9]. Globally, inland waterbodies receive 2–3 Pg-C yr^{-1} from terrestrial landscapes, most of which is emitted as CO_2 fluxes to the atmosphere (0.5–2.1 Pg-C yr^{-1}) [10–14], an amount comparable to the global net terrestrial carbon sink [15–17]. Inland waterbodies are also a globally significant source of CH_4 (70–150 Tg-C yr^{-1}), accounting for up to half of all CH_4 emissions from natural sources and close to a quarter of global CH_4 emissions [13, 14, 18–23]. Lateral carbon transport accounts for ~20% of terrestrial net ecosystem productivity in high latitude ecosystems (compared to 1% for temperate and tropical ecosystems) due to the abundance of lakes and ponds, highlighting the importance of CO_2 and CH_4 fluxes from inland waters [11, 24–26]. Despite their importance in global carbon cycling, inland waterbody CO_2

and CH_4 budgets remain uncertain due to the wide range in fluxes reported and the uncertainty in waterbody areal estimates, particularly for small lakes and ponds [12, 16, 27].

Observations of inland aquatic CO_2 and CH_4 fluxes are highly variable, caused by differing contributions from processes across a hierarchy of scales from watershed transport, carbon cycling within waterbodies, to microcosms of microbial productivity. The drivers of inland aquatic carbon dynamics are often complex and non-linear. Observations of these drivers lack the spatial representation necessary for scaling, e.g. water temperature, dissolved organic carbon concentration and lability [28–32]. Traditional bottom–up scaling estimates of inland waterbody CO_2 and CH_4 fluxes applies an average or median flux to an estimated total water surface area on local, regional, or global scales [10, 11, 14, 18, 21, 26]. Several scaling studies have shown improved estimates of inland waterbody CH_4 emissions by including lake size and landscape history as categorical drivers [14, 19, 21, 23, 33]. Lake productivity (calibrated from a remotely sensed analog) has also been used as a linear predictor of lake dissolved CO_2 concentrations [14, 34]. However, a large amount of variation in CO_2 and CH_4 waterbody fluxes remains unexplained in scaling studies, resulting in high uncertainty in regional and global carbon budget estimates [10, 11, 14, 19, 21]. Applying an average flux to waterbodies within a region or lake size-class could also create a biased carbon estimate, for example if smaller high-flux lakes are more abundant in the observation dataset than they are in the landscape [22, 35, 36].

Top–down carbon estimates from inversion studies rarely consider inland waterbodies. Most often, inversion studies mask inland water, functionally attributing a zero flux, or categorize waterbodies as wetlands (but see Tan *et al* [37]). In lake-rich regions this mis-attribution will cause either over or under estimation of fluxes in wetland or terrestrial environments to compensate, a source of uncertainty in the attributed carbon fluxes that is rarely quantified or discussed. A recent top-down and bottom-up comparison of CO_2 fluxes from the North Slope of Alaska found that inland waters were likely a significant source of CO_2 during the early cold season, and attributing a flux to waterbodies in bottom-up estimates was necessary to match airborne observations [38]. Correctly attributing waterbodies in top-down inversion analyses requires a gridded waterbody carbon flux map for use as a flux prior, which are not often available or produced in bottom-up studies (but see Tan and Zhuang [39]).

Advances in remote sensing and computational abilities have led to steady improvements in inland waterbody areal estimates in recent years [19, 40, 41], but accurately mapping small lakes and ponds still remains a challenge [13, 42]. Delineating open waterbodies from vegetated wetland is particularly important because both are critical ecosystems for carbon emissions but with differing governing processes [22]. Vegetated wetlands are often narrow features bordering waterbodies along shorelines or channel networks that cannot be detected without high resolution imagery (<30 m) [43–46]. Furthermore, wetlands and lakes are often mapped separately (but see Olefeldt *et al* [47]), which can lead to double counting. This uncertainty around the waterbody and wetland area could explain why bottom-up CH_4 budgets in the Arctic are twice as high as top-down atmospheric inversion estimates [20, 48].

Here we create accurate scaling models of waterbody CO_2 and CH_4 emissions that reflect the underlying processes driving carbon cycling in these ecosystems and reduce uncertainty and bias in carbon budget estimates. We use an integrated terrestrial-aquatic approach by combining high-resolution remote-sensing imagery of watershed-level and waterbody drivers from an object-based imagery analysis. We include waterbody size and contour characteristics and the surrounding landcover, hydrology, terrain, and landscape heterogeneity as possible variables affecting carbon fluxes. We train boosted regression tree models, a type of machine learning, to predict waterbody CO_2 and CH_4 diffusive fluxes. While ebullition is a large component of aquatic methane fluxes, we do not include it here as ebullitive fluxes are unlikely to be driven by watershed level processes and are stochastic in nature, and therefore remain a unique challenge to scaling. We demonstrate this scaling technique using the Yukon–Kuskokwim (YK) Delta of Alaska as our study region. The YK Delta is subarctic tundra abundant in lakes and wetlands and underlain by discontinuous permafrost [49]. Atmospheric inversion models of high latitudes have shown the YK Delta to be a regional hotspot of CO_2 and CH_4 emissions [50–53]. Despite this, the YK Delta has been historically understudied with few ground-based observations of carbon fluxes [54, 55]. We leverage recent datasets of high-density observations of CO_2 and CH_4 measurements from waterbodies in the YK Delta to train scaling models and map waterbody CO_2 and CH_4 emissions [56, 57].

2. Methods

2.1. CH_4 and CO_2 observations

This study uses a dataset of surface water samples ($n = 364$) analyzed for dissolved CO_2 ($n = 235$) and CH_4 ($n = 294$), collected from various waterbodies in the central-interior of the YK Delta from the first half of July 2016–2019 [56, 57]. The majority (>85%) of the samples in these datasets were from

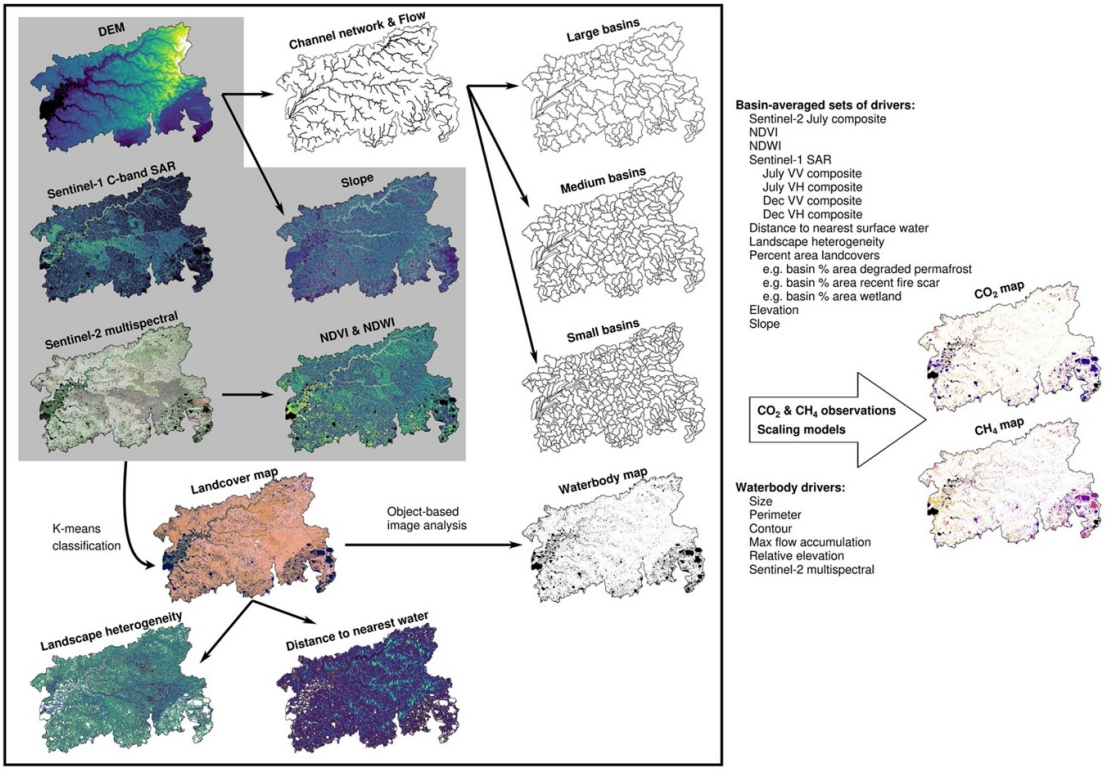

Figure 1. Schematic diagram of the geospatial analysis and remote sensing imagery used in this study. The thumbnail imagery examples were created from remote sensing layers used in the study but are not presented at accurate scales. The layers in the grey box were used to create the landcover map via k-means. The raw bands and derived layers (including landcovers) were averaged over each set of basins to create landscape-level drivers. An object-based image analysis of the waterbody product from the landcover map was used to create waterbody size, perimeter, and contour drivers, as well as waterbody reflectance, elevation, and maximum flow accumulation.

lakes and waterbodies within wetlands. All waterbodies were sampled at the water surface, either from shore for small waterbodies or from a boat. While all waterbodies were sampled in triplicate for dissolved gases with the average reported, the largest few waterbodies were sampled multiple years and at multiple locations. The average of all observations was used for those waterbodies with multiple samples. Variation in concentrations between waterbodies was far greater than interannual variability or spatial variation within waterbodies. The waterbodies in this region are uniformly shallow (less than 2 m deep) and flat, and consequently well mixed [32]. As a result, we do not consider lake depth as driving variable of CO_2 and CH_4 concentrations. A further description of the site, dataset, sample processing, and size distribution of sampled waterbodies can be found in the supplement (figure S1) and Ludwig *et al* [32].

2.2. Geospatial waterbody and sub-basin analyses

We use remote sensing imagery to: (1) identify waterbodies and quantify waterbody contour characteristics, (2) quantify watershed characteristics that might be related to landscape-level drivers, hydrology, or indirectly affect waterbody CO_2 or CH_4 biogeochemistry, and (3) scale results to map diffusive fluxes of CO_2 and CH_4 (figure 1).

2.2.1. Remote sensing imagery

We use Google Earth Engine to select imagery in the YK Delta from 2016 through 2019 to coincide with the timing of the samples in the dissolved CO_2 and CH_4 observation dataset. We used a composite of cloud-free level-2A Sentinel-2 multispectral images (red, green, blue, near-infrared (NIR), and short-wave infrared (SWIR) bands) from July 2019 (within 1 week of water sample collection timing). Level-2A are surface reflectance Sentinel-2 products, with cloud removal, orthorectification, and sen2cor atmospheric corrections applied using the Sentinel-2 Toolbox. We use Sentinel-1 C-band short-aperture radar (SAR) images, pre-processed using the Sentinel toolbox [58, 59]. We created four SAR images; mean composites of July and December, VV and VH backscatter (10 m resolution). We use elevation from the mosaiced ArcticDEM (2 m resolution), after filling and detrending the digital elevation model (DEM) using System for Automated Geoscientific Analyses (SAGA) [60]. From these remote sensing layers, we derive further layers such as slope and flow accumulation (from the DEM before detrending), the

normalized difference vegetation index (NDVI), and normalized difference water index [61, 62].

2.2.2. Landcover map

We use a 5 by 10 m resolution landcover map created for the region (https://doi.org/10.3334/ORNLDAAC/2178) [63]. For detailed methods, see Ludwig *et al* [32]. We used Google Earth Engine's 'entropy' function to create a spatial texture layer from the landcover map. Higher entropy values occur at borders and transitions between landcover types, and we interpret the average entropy in an area as a metric of landscape heterogeneity. We used Google Earth Engine's 'fastDistanceTransform' function to create a gridded layer of distance to nearest water for the study region, using the landcover category identified as 'surface water' and excluding waterbodies >10 km across. While the surface water landcover was highly accurate (balanced accuracy >0.95; Ludwig *et al* [32]), the validation procedure was not stratified by waterbody size. We compared the size distribution of waterbodies used here to a higher resolution, independent waterbody product validated by size against ground truth points in an overlapping region of the YK Delta (Mullen *et al* [64]). The mapped waterbodies were remarkably similar despite differences in seasonality and interannual variation in the underlying remote sensing imagery used in the two maps (figure S2). Most notably, the waterbody product used here did not under-sample small ponds despite approaching the limit of detection. Therefore, we chose to include waterbodies <0.001 km^2 though they become pixelated, as omitting them would lead to an underestimate of carbon emissions.

2.2.3. Waterbody object-based image classification and shape analysis

We used an object-based imagery analysis of the surface water classification from the landcover map to identify and then quantify aspects of waterbodies in the region. We used the 'reduceConnectedComponents' algorithm in Google Earth Engine, thresholding to exclude waterbodies larger than 10 km in width (due to lack of *in situ* data), to calculate the area, perimeter, and ratio of area:perimeter of each uniquely identified waterbody. We specified a maximum neighborhood of 1024 pixels in any dimension using eight-way connectedness with a scale of 10 m. We created gridded layers of the average red, green, blue, NIR, and SWIR reflectance from each waterbody by reducing the Sentinel-2 composite imagery in section 2.2.1 over the waterbody objects. We similarly created an average waterbody elevation layer, using the detrended DEM, which reflects each waterbody's relative landscape position on (higher elevation) or between (lower elevation) peat plateaus. In QGIS (Quantum Geographic Information System, we used the 'catchment area' algorithm to create a flow accumulation layer from the DEM, and then reduced this over the waterbody objects to create a maximum flow accumulation per waterbody gridded product. A higher flow accumulation within a waterbody indicates more pixels 'pouring' into it, which correlates with a larger watershed. All layers were reprojected to 10 m resolution while reduced to averages over waterbody objects.

2.2.4. Sub-basin analyses

We split the study region into non-nested contiguous sub-basins that are distinct hydrologic units. We used the 'channel network' SAGA algorithm in QGIS inputting the filled-DEM and flow accumulation map as the channel initialization to construct these sub-basins. We created three sets of sub-basins using different channel thresholds: (1) $>1 \times 10^5$ m^2 created 16 036 small sub-basins with an average area of 0.15 km^2, (2) $> 1 \times 10^6$ m^2 created 1988 medium sub-basins with an average area of 1.2 km^2, and (3) $> 1 \times 10^7$ m^2 created 278 large sub-basins with an average area of 8.5 km^2. These sizes were chosen to span the range of watershed sizes from the waterbodies in the observation dataset [32]. For each set of sub-basins, we masked out waterbodies and then calculated the sub-basin average values for the remote sensing imagery and derived indices described in section 2.2.1, the sub-basin average entropy (which we term 'landscape heterogeneity'), and average distance to nearest waterbody. We also calculated the percent cover of each landcover type within each of the three sets of sub-basins. Finally, each set of these sub-basin metrics were assigned to the waterbodies located within those sub-basins, to create images of pseudo-watershed level drivers. We used this sub-basin approach to enable all the potential drivers to be organized as a stack of images, which is efficient to operate over for predicting and scaling. The alternative approach, using actual watersheds delineated for each waterbody such as in Ludwig *et al* [32], would be computationally prohibitive. We identified 17 071 distinct waterbodies used in this study, and even if watersheds for all 17 071 could be derived efficiently, the nested nature of most of the watersheds would prevent operating over watershed-level variables as images. Not all waterbodies will exhibit similar connectivity to the sub-basins used to aggregate remote-sensing imagery. While flow accumulation is related to connectivity, better metrics, particularly of lateral or sub-surface flow, could improve predictability in future scaling models. We further expect connectivity between waterbodies and sub-basins to vary seasonally: for example, different landcover variables might be retained when using late season flux observations and imagery where deeper thaw depths change the hydrologic regime. Future scaling studies using

seasonally representative datasets will likely need to take this seasonality into account.

2.3. Statistical modeling and scaling

We use boosted regression tree models to predict and scale waterbody CO_2 and CH_4 fluxes. Recent studies have used machine learning to accurately model and predict aquatic carbon cycling [31, 32, 65]. Machine learning methods are particularly useful when using surface reflectance over inland waters, where atmospheric corrections can sometimes cause artifacts that lead to nonsensical results in approaches that predict using band ratios, band linear combinations, or other empirical or analytical models [66–68]. For example, gradient boosting models, such as used here, were effective in predicting Chl-a using Sentinel 2 and 3 imagery of inland waters [69, 70]. We use a similar approach to Ludwig *et al* [32], using the gradient boosting machine ('gbm') package in R v.3.6.1 [71]. Both CO_2 and CH_4 values were log-transformed to achieve normality. Our potential drivers include all waterbody-specific variables (section 2.2.3), sub-basin averages and percent landcover areas (section 2.2.4). We sampled this stack of image layers at every waterbody observation point to create a tabular dataset of drivers for model training. We used 'gbm.step' as described in Elith *et al* [71] to tune the number of trees and drop variables to avoid overfitting, using ten-fold cross validation, with a bag-fraction of 0.65. We tested the learning rate ($lr = 0.005$) and tree complexity ($tc = 2$) manually to optimize. We used percent deviance improvement over null model from cross-fold validation to determine model predictability, and regression between observations and fitted values of dissolved gases to determine model fit ('lm' function in R). We repeated this analysis with ten random seeds for bag-fraction to quantify data uncertainty in model training. We calculated the relative influence of each predictor variable, which were scaled to sum to 100 [71, 72]. We used partial dependence plots (the mean and standard deviation of the ten model runs with different random seeds) to investigate the average predicted response across all observations for a given predictor variable [72]. All partial dependence plots were centered on zero μM predicted CO_2 or CH_4.

We used the 'predict' function in the 'raster' package in R to apply our boosted regression tree scaling models to the study region using the stack of imagery described in sections 2.2.3 and 2.2.4. These predictions were then back-transformed to convert into μM, with a prediction-based back-transformation bias correction applied [73]. Dissolved gas concentrations were converted to diffusive fluxes (mg C m^{-2} d^{-1}) using the relationship in equation (1) [74, 75], where C_{aq} is the surface dissolved gas concentration, C_{atm} is the atmospheric concentration, F is the diffusive flux, and k is the gas transfer velocity,

$$F = k\left(C_{aq} - C_{atm}\right). \quad (1)$$

Gas transfer velocity values were obtained from paired observations of chamber-based diffusive fluxes and dissolved surface water concentrations from waterbodies in the YK Delta ($n = 55$ for CO_2, $n = 65$ for CH_4 [56, 57]). Gas transfer velocities were calculated separately for CO_2 and CH_4 and normalized to k_{600} using Schmidt numbers [75]. High outliers of k_{600} were deemed to be contaminated by ebullitive fluxes and removed [33]. Since there was no detectable relationship to lake size or wind speed we used the mean, 25th, and 75th percentiles of gas transfer velocity from the observations in the region.

3. Results and discussion

3.1. Model performance

Our models were able to accurately fit and predict dissolved CO_2 and CH_4 observations using only remotely sensed drivers. Model fit for dissolved CO_2 and CH_4 was not significantly different from a slope of one based on 95% confidence intervals, with high coefficients of determination (CO_2; $r^2 = 0.76$, RMSE = 175 μM, CH_4; $r^2 = 0.75$, RMSE = 4.8 μM, figure 2). The average of residuals from both the model predictions (figures 2(a) and (b)) and the back-transformed results (figures 2(c) and (d)) were zero (p-value $\gg 0.05$ in one-sample t-tests). Both scaling models performed equally well across waterbody sizes, with no relationship between model residuals and waterbody area (figure S3). Models that also consider biogeochemical and mechanistic drivers (e.g. dissolved oxygen, dissolved organic carbon), which cannot be detected by remote sensing, outperform our scaling models (e.g. Ludwig *et al* [32]: CO_2; $r^2 = 0.94$, CH_4; $r^2 = 0.88$).

We measured the predictive strength of our models using ten-fold cross-validation using percent reduction in deviance as the metric of success. Our CO_2 model had the best predictability with 63% reduction in predictive deviance, while our CH_4 model had 39% reduction in predictive deviance. This is only slightly reduced predictive ability compared to the models used to describe dissolved CO_2 and CH_4 in unburned watersheds in the YK Delta (79% and 52% respectively) that incorporated numerous biogeochemical but non-scalable drivers, and is better predictability than the equivalent models for burned watersheds (61% and 36% respectively) [32].

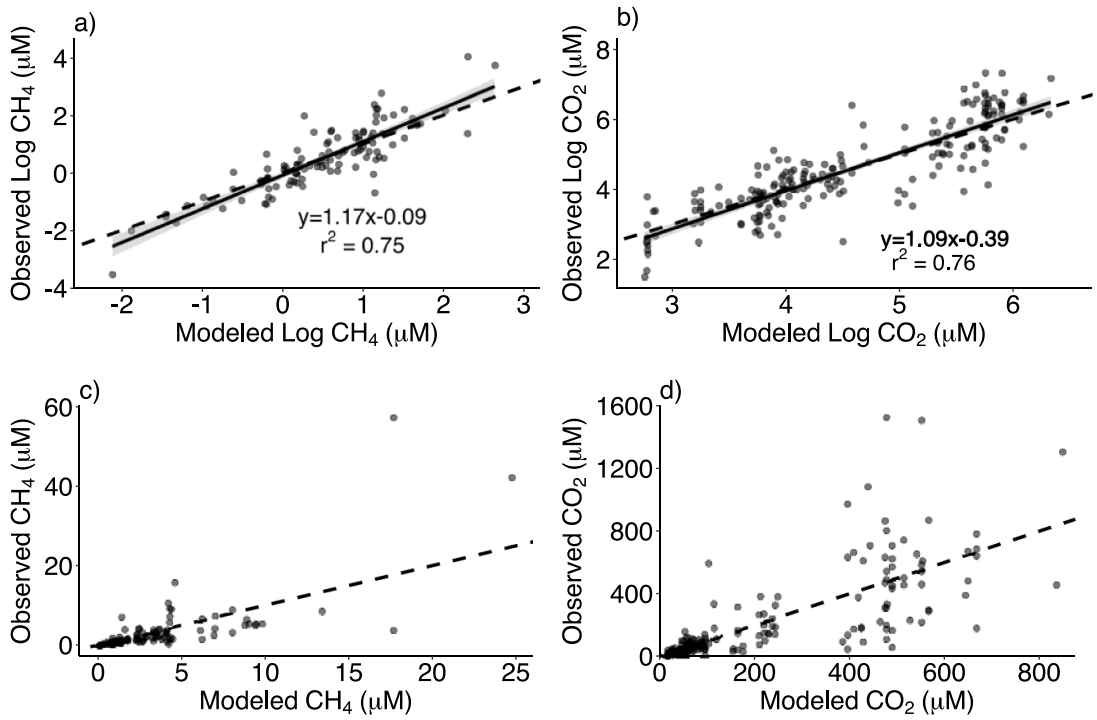

Figure 2. Modeled and observed dissolved gas concentrations for CH_4 (a) and (c) and CO_2 (b) and (d). Model goodness-of-fit is shown with log-transformed data (a) and (b) and un-transformed and back-transformed data with bias corrections (c) and (d). Solid lines and shading depict the fitted slope and 95% CI from linear regressions, with the one-to-one lines as dashed lines.

3.2. Spatial drivers of dissolved CO_2 and CH_4

Dissolved CO_2 was primarily driven by variables related to waterbody contour, with waterbody perimeter, area, and the ratio of area:perimeter accounting for more than two-thirds of the total explanatory power of the model (figure 3(a)). In contrast, the influence of waterbody contour drivers on dissolved CH_4 was similar to watershed landcover and watershed hydrology drivers (figure 3(b)). For both CO_2 and CH_4, smaller waterbodies and those with more complex contours (smaller area:perimeter) exhibited higher dissolved gas concentrations (figures 5(a), (b), S4(a) and S6(d)), consistent with patterns observed globally and in the Arctic [13, 19, 21]. Higher carbon emissions from waterbodies with complex contours could be caused by more relative abundance of lake-edge landcovers: the transitions between wetland and aquatic ecosystems have long been recognized as biogeochemical hotspots [76].

Wetland landcover was the most important driver of dissolved CH_4 (figure 3(b)), with a threshold effect where more CH_4 was predicted as wetland percent area rose above 15% (figures S4(f) and (j)). Basin averaged NDVI, SWIR reflectance, and SAR backscatter were also significant drivers of dissolved CH_4 in waterbodies and likely also depict the importance of wetlands on downstream CH_4. Wetland areas were distinctly visible in NDVI, SWIR, and SAR imagery, identifiable as the greenest areas (NDVI effect), while SAR backscatter and SWIR reflectance are commonly used in wetland mapping due to relationships to soil water content and canopy structure [59, 77–80]. Basin averaged NDVI, SAR, and SWIR correlated with basin wetland percent area in the study region (Pearson's $r = 0.81, 0.56, -0.58$ respectively). Wetlands are often a significant, if not the dominant, source of natural CH_4 emissions in ecosystems. Our results show wetlands also impact CH_4 concentrations in nearby and downstream waterbodies (figure 3(b)). This downstream effect could be caused by groundwater flow transporting dissolved CH_4 from where it was produced in a wetland to an open waterbody [81]. Wetland-related drivers were also significant but less important for predicting CO_2 (figure 3(a)). Wetlands may encourage conditions towards CH_4 and CO_2 production downstream in the watershed through longer water residence times, depletion of oxygen, and increased dissolved organic carbon inputs [82]. Regardless of the mechanism, cohesive high-resolution mapping of waterbodies and wetlands is important for scaling CH_4 emissions.

Basin hydrology plays an important role for both dissolved CO_2 and CH_4. Predicted CH_4 concentrations peak when surface waters are present in ~10% of the surrounding basin, but decline if there is either less or more surface water present (figure S4(c)).

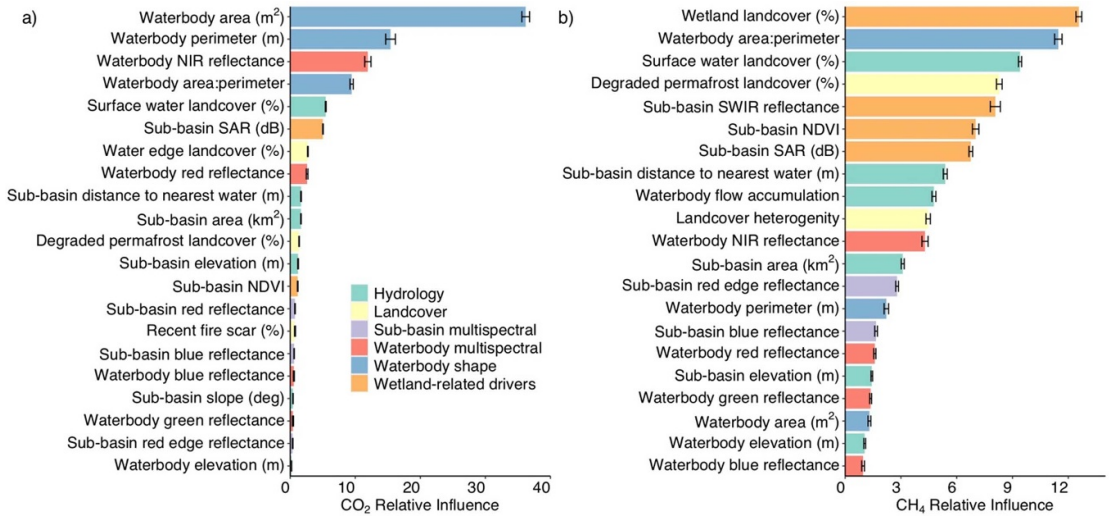

Figure 3. Relative influence of variables in boosted regression tree models predicting CO_2 (a) and CH_4 (b) concentrations in waterbodies in the YK Delta of Alaska. Error bars indicate standard deviation of relative influence from ten boosted regression tree model runs with different random seeds. Relative influences are scaled to 100%, with similar variables combined (e.g. watershed SAR backscatter from VV, VH, July, and December composites). The full listing of variable relative influence can be found in tables S1 and S2 in the supporting information.

Predicted CO_2 strictly declined with increasing surface water area in the surrounding basin (figure S6(e)). Predicted CH_4 and CO_2 concentrations both increased as a function of increasing basin-average distance to nearest water (figures S4(g) and S6(l)). Flow accumulation had an overall negative effect on CH_4 concentrations (figure S4(h)). Surface water area, distribution, and flow accumulation relate to water residence times in the surrounding landscape, and could indicate basins with longer water residence times can promote higher waterbody CH_4, possibly through increased interaction with soil pore water, more oxygen depletion, or more leaching of carbon substrates [83, 84].

Waterbody surface reflectance in blue, red, and NIR bands contributed significantly to predicting dissolved CO_2 and CH_4 (figure 3). Combinations of these bands from Landsat surface reflectance have been used to remotely sense chromophoric dissolved organic matter (CDOM) in rivers and inland waters in other environments [85–87]. Previous results from the YK Delta have indicated that CDOM is an important driver of both dissolved CO_2 and CH_4 [32]. Small waterbodies (<0.01 km^2) may have land-adjacency effects in surface reflectance that are complicating signals from waterbody color. It is possible for these smaller waterbodies that the role of surface reflectance instead indicates increasing edge-effects through land-adjacency, which is a demonstrated driver of both CO_2 and CH_4. To test how small-waterbody reflectance values, land-adjacency, and size metrics affect our scaling models, we re-ran them using the same hyper-parameters, drivers, and random-seeds while replacing waterbody reflectance, size, contour, and perimeter variables with NA's for those waterbodies under 0.01 km^2 in area. There was a slight decrease in predictive performance (percent deviance explained decreased by 3% and 2% for CH_4 and CO_2). We regressed the predicted results from our original models against those with small lake data withheld, with little discernable difference (slopes of 1, intercepts of 0, and r^2 of 0.97 and 0.99). The remote sensing-based models in this study could be capturing both CDOM and edge-effects (for the smallest waterbodies) indirectly through waterbody surface reflectance.

3.3. Lake size effects

Using the boosted regression tree models developed from observations of dissolved CO_2 and CH_4 and remotely sensed drivers (figure 1), we created maps of waterbody CO_2 and CH_4 fluxes for our study region in the YK Delta (figure 4). The effect of waterbody size on CH_4 and especially CO_2 fluxes can be seen clearly in the mapped predictions (figures 5(a) and (b)). The smallest waterbodies account for the highest fluxes of both CO_2 and CH_4, and fluxes decline precipitously with increasing waterbody size for CO_2. The smallest waterbodies (<0.001 km^2) have disproportionately high fluxes and comprise 38% of the 17 071 uniquely identified waterbodies in the region, but account for only 1% of the overall surface area of water (table 1). Because of this, the smallest waterbodies do not contribute significantly to the total CO_2 and CH_4 fluxes (figures 5(c) and (d)), accounting for only 3% and 2.8% of the total fluxes of CO_2 and CH_4 in the region respectively (table 1). In contrast, Holgerson and Raymond [13] estimated very

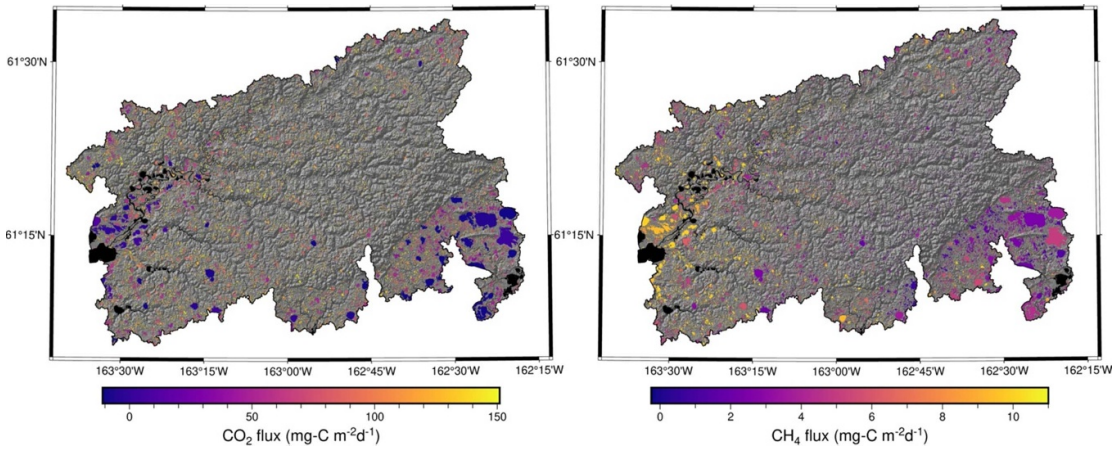

Figure 4. Maps of predicted CO_2 and CH_4 in the study region in the Yukon-Kuskokwim Delta, AK. Large waterbodies (>10 km width) and rivers, which were not sampled in the observation dataset and are depicted in black in the maps. Non-water is represented with a grey hillshade image created from the DEM.

small ponds (<0.001 km^2) account for ~15.1% and 40.6% of global CO_2 and CH_4 emissions from lentic inland waters. Holgerson and Raymond [13] estimated the distribution of very small ponds by extrapolating the Pareto distribution, a log-abundance log-size regression relationship [88], yielding a much greater area of very small ponds relative to total water area. While our waterbody mapping could be missing sub-pixel very small ponds (<0.00005 km^2), our lake distribution are similar to higher resolution products from the same region [42, 64]. Muster *et al* [42] found that waterbody distributions in permafrost lowland regions were not representable by a power-law relationship. Consequentially, the Pareto distribution would overestimate very small pond abundance. Small lakes (0.001–0.1 km^2) are the most abundant in the study area, have high CO_2 and CH_4 emissions relative to their area, and contribute the majority of total CO_2 and CH_4 emissions (table 1). Medium and large lakes are the least abundant, and while they are still significant contributors to CH_4 emissions, they have a small or negative contribution to CO_2 emissions (table 1). Our results suggest that very small ponds in the Arctic need to be explicitly mapped using high-resolution techniques to avoid underestimating aquatic C emissions from their exclusion, or overestimating aquatic C emissions from inaccurate areal estimates.

3.4. Scaling CO_2 and CH_4

Carbon emissions from inland waterbodies remains one of the least certain portions of the global C cycle. We compared the total fluxes when scaled using our boosted regression models to two simpler, more traditional approaches. The simplest scaling method multiplies the average areal flux rate by the total waterbody surface area [10, 11, 18]. The other approach applies the average observed flux in a waterbody size-class to the total area of water in that size-class [13, 14, 21, 26, 33]. With the simplest approach, CO_2 total fluxes are overestimated by 79% and CH_4 total fluxes are overestimated by 53% (figure 6) compared to our approach. When scaling using the average flux by size-class, CO_2 and CH_4 fluxes were overestimated in the three smallest size-classes of waterbodies (figure 6). The largest lakes (>1 km^2) were underestimated for CO_2 and CH_4 in the size-class average scaling, as well as the mid-sized lakes (0.1–1 km^2) for CO_2 (figure 6). Overall the scaling approach using the average flux by lake size-class overestimated the diffusive waterbody fluxes of CO_2 and CH_4 from the region by 26% and 8% respectively. This bias in simpler scaling results is likely due to a lack of spatial representativeness in the observation dataset, a common problem in arctic flux datasets. Synthesis studies in particular will lack spatially representative flux observations. For example, Kuhn *et al* [21] attribute the relatively poor fit of their aquatic CH_4 model compared to their terrestrial models' performance to spatial under-sampling.

Recent efforts to improve arctic CH_4 estimates include the Boreal-Arctic Wetland and Lake methane Dataset (BAWLD), a circumpolar database of wetland landcover jointly mapped with lake size and abundance [22, 47]. The BAWLD database solves issues of double-counting and the representativeness of empirical flux observations and is well disposed for large-scale modeling of circumpolar CH_4 emissions. However, the random forest models used in BAWLD has the lowest predictive power when mapping small lakes (<0.1 km^2), and did not differentiate waterbody sizes below 0.1 km^2. Our results are complimentary to the BAWLD approach and demonstrate how high-resolution remote sensing, machine learning models,

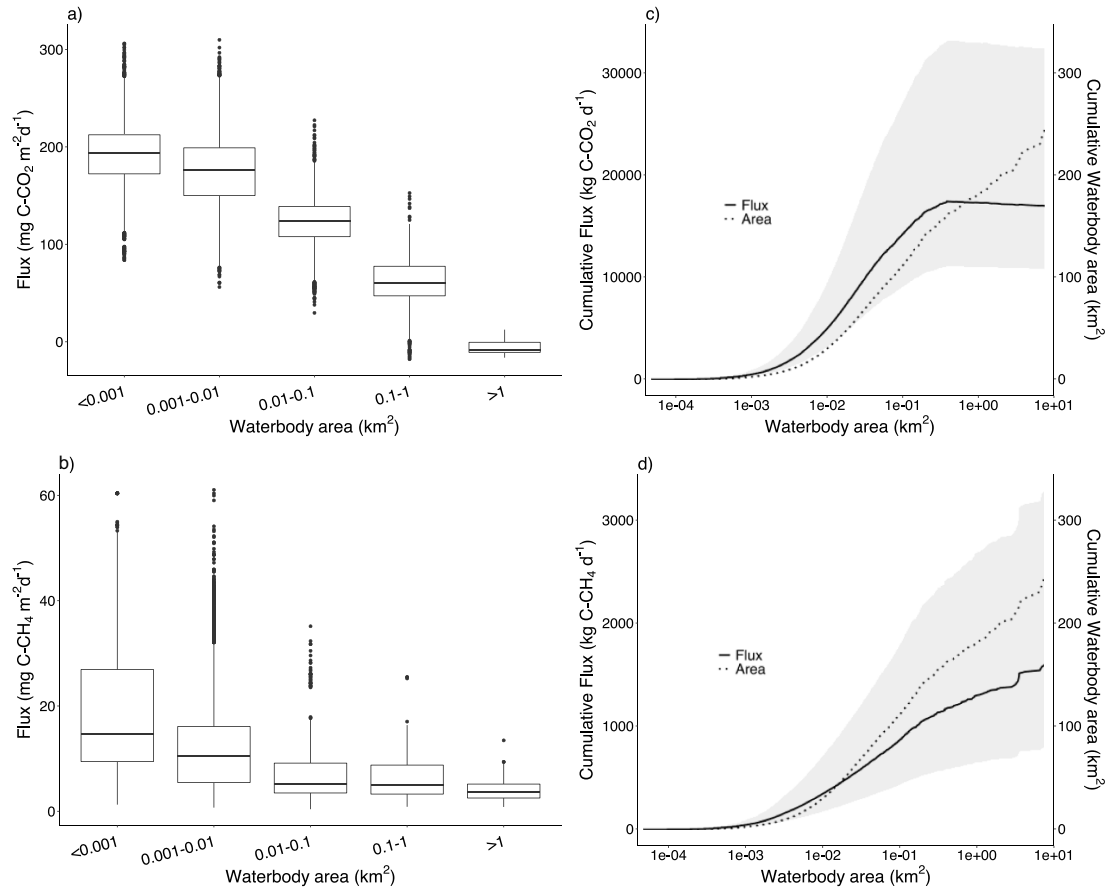

Figure 5. Box plots of predicted fluxes of CO_2 (a) and CH_4 (b) as a function of waterbody size. The lower and upper hinges correspond to the first and third quartiles, the whiskers extend to 1.5 times the interquartile range, with outliers indicated as points. Cumulative CO_2 (c) and CH_4 (d) fluxes as a function of waterbody size, with cumulative waterbody area as a secondary *y*-axis. Note the log10 scale *x*-axis. Shading indicates the predicted flux using the 75th and 25th percentile of k_{600} from observations in the region.

Table 1. Waterbody distribution and area by size-class. CO_2 and CH_4 emissions are the predicted dissolved concentrations from our scaling models, converted to diffusive fluxes using the mean gas transfer velocity (*k* in equation (1)) from observations, multiplied by each waterbody's surface area, and summed over all the waterbodies within each size-class. The range in parentheses for CO_2 and CH_4 emissions is calculated using the 25th and 75th percentile of gas transfer velocities to calculate fluxes.

Waterbody size-class	Abundance within study region	Area within study region (km^2)	CO_2 emissions (kg C d^{-1})	CH_4 emissions (kg C d^{-1})
< 0.001 km^2	6425	2.57	447.4 (284 to 853)	29.15 (14.52 to 60.14)
0.001–0.01 km^2	7268	27.66	4042 (2569 to 7707)	215.1 (107.2 to 443.8)
0.01–0.1 km^2	3037	81.65	8321 (5289 to 15 864)	424.5 (211.5 to 875.8)
0.1–1.0 km^2	317	68.62	2722 (1730 to 5189)	347.5 (173.1 to 716.9)
1.0–10 km^2	24	63.51	−481.1 (−305 to −917)	248.2 (123.6 to 512)
Total	17 071	244	15 051 (9567 to 28 696)	1264 (630 to 2609)

and extensive field observations can be leveraged to explicitly and accurately map open-water diffusive C fluxes at the regional scale.

Our models do not account for temporal variation in fluxes, but rather provide a snapshot map of peak growing season fluxes. Typically, waterbody scaling studies extrapolate an average flux to a seasonal or annual budget estimate. Recent syntheses have improved carbon budget estimates by accounting for ice-free days and large fluxes from ice-off events [19]. Studies of specific lakes have documented diurnal patterns in dissolved CO_2 and CH_4 in surface waters, though sampling regimes rarely capture this temporal variation [89]. Similarly, variable gas evasion from wind and convection affecting turbulence is an important source of temporal variability in lake fluxes [89–91]. While our models improve the spatial accuracy of waterbody fluxes, we recommend considering seasonal and diurnal variation in dissolved gas concentrations and gas transfer velocities

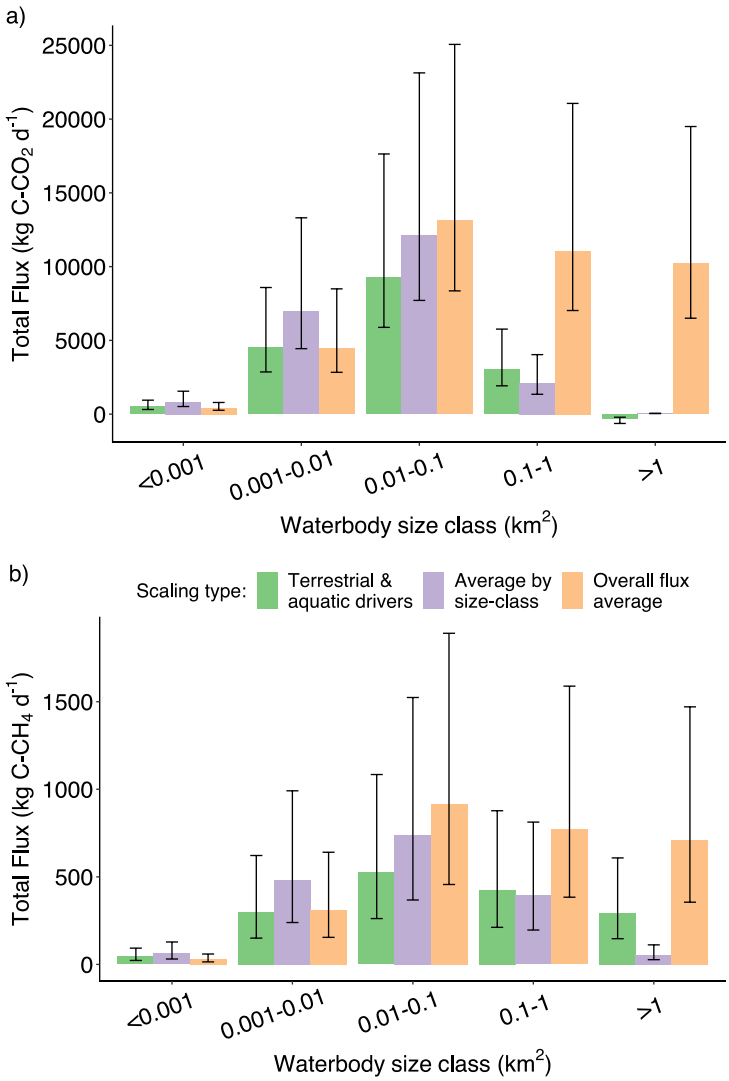

Figure 6. Total carbon emissions from each waterbody size-class using different scaling methods: this study, using an integrated terrestrial-aquatic approach (green), using the average flux by size-class (purple), and the overall average flux (orange). Scaled CO_2 total fluxes are in the top panel (a) and scaled CH_4 total fluxes are in the bottom panel (b). Carbon fluxes from the smaller water bodies are overestimated and largest lakes are underestimated by the simpler scaling methods. Error bars depict the range flux totals derived using the 25th–75th percentile of k_{600} from observations in the region.

and seasonal variability in water body size when using our flux maps.

4. Conclusion and implications

Inland aquatic carbon emissions remain one of the most uncertain components of the global carbon budget. We can reduce this uncertainty using scaling models driven by watershed and waterbody processes. Aquatic CO_2 emissions can be predicted well using lake size and contour as continuous variables, which could be applied to larger regions with suitably high-resolution waterbody maps. Very small ponds in the Arctic need to be explicitly mapped using high-resolution techniques to avoid biasing aquatic carbon emissions from their exclusion or inaccurately extrapolated areal estimates. Waterbody size metrics are insufficient for scaling CH_4 emissions, which were primarily driven by wetland landcover and wetland-related variables within watersheds. Our terrestrial-aquatic integrated approach using watershed landcovers and hydrology-related remote sensing drivers improved our ability to scale both CO_2 and CH_4 waterbody emissions. Concurrent wetland landcover and waterbody mapping is necessary to avoid double-counting open water areas and for integrating terrestrial effects on aquatic emissions. As the Arctic warms with climate change, new waterbodies will form from thawing permafrost while others will drain and be replaced by wetlands. Our results imply the increased abundance of small ponds and replacement of large lakes with wetlands would lead to higher emissions of CO_2 and CH_4 for the YK Delta.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.18739/A23775V7T and https://doi.org/10.18739/A22804Z8M. Maps of CO_2 and CH_4 produced in this study can be found at https://doi.org/10.3334/ORNLDAAC/2178.

Acknowledgments

We would like to acknowledge that our research work took place on the traditional land of the Yup'ik, who have stewarded this land through many generations. Our work would not have been possible without the support of the Yukon Delta National Wildlife Refuge, U S Fish and Wildlife Service. This study was supported with funding from a National Aeronautics and Space Administration FINESST grant (80NSSC19K1301) to S M L, and National Science Foundation grants (NSF-1044610 and NSF-1561437) to S M N. This work would not have been possible without the efforts of the Polaris Project students from 2017 to 2019 in sample collection and processing. We would like to thank Max Holmes and Paul Mann for their roles in bringing the Polaris Project to the YK Delta and as mentors for aquatic sampling. We thank Max Jones for guidance on pygmt in the creation of figures 1 and 4. This study was part of the NASA Arctic-Boreal Vulnerability Experiment.

ORCID iDs

Sarah M Ludwig https://orcid.org/0000-0002-2873-479X
Luke D Schiferl https://orcid.org/0000-0002-5047-2490
Roisin Commane https://orcid.org/0000-0003-1373-1550

References

[1] Hugelius G, Strauss J, Zubrzycki S, Harden J W, Schuur E A G Ping C L and Kuhry P 2014 Improved estimates show large circumpolar stocks of permafrost carbon while quantifying substantial uncertainty ranges and identifying remaining data gaps *Biogeosci. Discuss.* **11** 4771–822
[2] Hugelius G *et al* 2020 Large stocks of peatland carbon and nitrogen are vulnerable to permafrost thaw *Proc. Natl Acad. Sci.* **117** 20438–46
[3] Belshe E F, Schuur E A G and Bolker B M 2013 Tundra ecosystems observed to be CO_2 sources due to differential amplification of the carbon cycle *Ecol. Lett.* **16** 1307–15
[4] Schuur E A G *et al* 2015 Climate change and the permafrost carbon feedback *Nature* **520** 171–9
[5] Virkkala A-M *et al* 2021 Statistical upscaling of ecosystem CO_2 fluxes across the terrestrial tundra and boreal domain: regional patterns and uncertainties *Glob. Change Biol.* **27** 4040–59
[6] Watts J *et al* 2021 Soil respiration strongly offsets carbon uptake in Alaska and Northwest Canada *Environ. Res. Lett.* **16** 084051
[7] Schuur E A G *et al* 2008 Vulnerability of permafrost carbon to climate change: implications for the global carbon cycle *BioScience* **58** 701–14
[8] Vonk J E *et al* 2015 Reviews and syntheses: effects of permafrost thaw on Arctic aquatic ecosystems *Biogeosciences* **12** 7129–67
[9] Tank S E, Vonk J E, Walvoord M A, McClelland J W, Laurion I and Abbott B W 2020 Landscape matters: predicting the biogeochemical effects of permafrost thaw on aquatic networks with a state factor approach *Permafr. Periglac. Process.* **31** 358–70
[10] Cole J J *et al* 2007 Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget *Ecosystems* **10** 172–85
[11] Tranvik L J *et al* 2009 Lakes and reservoirs as regulators of carbon cycling and climate *Limnol. Oceanogr.* **54** 2298–314
[12] Raymond P A *et al* 2013 Global carbon dioxide emissions from inland waters *Nature* **503** 355–9
[13] Holgerson M A and Raymond P A 2016 Large contribution to inland water CO_2 and CH_4 emissions from very small ponds *Nat. Geosci.* **9** 222–6
[14] DelSontro T, Beaulieu J J and Downing J A 2018 Greenhouse gas emissions from lakes and impoundments: upscaling in the face of global change *Limnol. Oceanogr. Lett.* **3** 64–75
[15] Tian H *et al* 2016 The terrestrial biosphere as a net source of greenhouse gases to the atmosphere *Nature* **531** 225–8
[16] Regnier P *et al* 2013 Anthropogenic perturbation of the carbon fluxes from land to ocean *Nat. Geosci.* **6** 597–607
[17] Friedlingstein P, Jones M W, O'Sullivan M, Andrew R M, Bakker D C E Hauck J and Zeng J 2022 Global carbon budget 2021 *Earth Syst. Sci. Data* **14** 1917–2005
[18] Bastviken D, Tranvik L J, Downing J A, Crill P M and Enrich-prast A 2011 Freshwater methane emissions offset the continental carbon sink *Science* **331** 50
[19] Wik M, Varner R K, Anthony K W, MacIntyre S and Bastviken D 2016 Climate-sensitive northern lakes and ponds are critical components of methane release *Nat. Geosci.* **9** 99–105
[20] Saunois M, Stavert A R, Poulter B, Bousquet P, Canadell J G Jackson R B and Zhuang Q 2020 The global methane budget 2000–2017 *Earth Syst. Sci. Data* **12** 1561–623
[21] Rosentreter J A *et al* 2021 Half of global methane emissions come from highly variable aquatic ecosystem sources *Nat. Geosci.* **14** 225–30
[22] Kuhn M A, Varner R K, Bastviken D, Crill P, MacIntyre S, Turetsky M, Walter Anthony K, McGuire A D and Olefeldt D 2021 BAWLD-CH_4: a comprehensive dataset of methane fluxes from boreal and arctic ecosystems *Earth Syst. Sci. Data* **13** 5151–89
[23] Johnson M S, Matthews E, Du J, Genovese V and Bastviken D 2022 Methane emission from global lakes: new spatiotemporal data and observation-driven modeling of methane dynamics indicates lower emissions *J. Geophys. Res.: Biogeosci.* **127** e2022JG006793
[24] Kling G W, Kipphut G W and Miller M C 1991 Lakes and streams for tundra carbon budgets atmosphere: implications *Science* **251** 298–301
[25] Chapin F I *et al* 2006 Reconciling carbon-cycle concepts, terminology, and methods *Ecosystems* **9** 1041–50
[26] Stackpoole S M, Butman D E, Clow D W, Verdin K L, Gaglioti B V, Genet H and Striegl R G 2017 Inland waters and their role in the carbon cycle of Alaska *Ecol. Appl.* **27** 1403–20
[27] Matthews E, Johnson M S, Genovese V, Du J and Bastviken D 2020 Methane emission from high latitude lakes: methane-centric lake classification and satellite-driven annual cycle of emissions *Sci. Rep.* **10** 12465
[28] Lapierre J-F and Giorgio P A D 2012 Geographical and environmental drivers of regional differences in the lake pCO_2 versus DOC relationship across northern landscapes *J. Geophys. Res.: Biogeosci.* **117**
[29] Tan Z, Zhuang Q and Walter Anthony K 2015 Modeling methane emissions from arctic lakes: model development and site-level study *J. Adv. Model. Earth Syst.* **7** 459–83

[30] Tan Z, Zhuang Q, Shurpali N J, Marushchak M E, Biasi C, Eugster W and Walter Anthony K 2017 Modeling CO_2 emissions from Arctic lakes: model development and site-level study *J. Adv. Model. Earth Syst.* **9** 2190–213
[31] Toming K, Kotta J, Uuemaa E, Sobek S, Kutser T and Tranvik L J 2020 Predicting lake dissolved organic carbon at a global scale *Sci. Rep.* **10** 8471
[32] Ludwig S M, Natali S M, Mann P J, Schade J D, Holmes R M, Powell M, Fiske G and Commane R 2022 Using machine learning to predict Inland Aquatic CO_2 and CH_4 concentrations and the effects of wildfires in the Yukon-Kuskokwim Delta, Alaska *Glob. Biogeochem. Cycles* **36** e2021GB007146
[33] Bastviken D, Cole J, Pace M and Tranvik L 2004 Methane emissions from lakes: dependence of lake characteristics, two regional assessments, and a global estimate *Glob. Biogeochem. Cycles* **18** 1–12
[34] Kuhn C, Bogard M, Johnston S E, John A, Vermote E, Spencer R, Dornblaser M, Wickland K, Striegl R and Butman D 2020 Satellite and airborne remote sensing of gross primary productivity in boreal Alaskan lakes *Environ. Res. Lett.* **15** 105001
[35] Bruhwiler L, Dlugokencky E, Masarie K, Ishizawa M, Andrews A, Miller J, Sweeney C, Tans P and Worthy D 2014 CarbonTracker-CH_4: an assimilation system for estimating emissions of atmospheric methane *Atmos. Chem. Phys.* **14** 8269–93
[36] Wik M, Thornton B F, Bastviken D, Uhlbäck J and Crill P M 2016 Biased sampling of methane release from northern lakes: a problem for extrapolation *Geophys. Res. Lett.* **43** 1256–62
[37] Tan Z, Zhuang Q, Henze D K, Frankenberg C, Dlugokencky E, Sweeney C, Turner A J, Sasakawa M and Machida T 2016 Inverse modeling of pan-Arctic methane emissions at high spatial resolution: what can we learn from assimilating satellite retrievals and using different process-based wetland and lake biogeochemical models? *Atmos. Chem. Phys.* **16** 12649–66
[38] Schiferl L D, Watts J D, Larson E J L, Arndt K A, Biraud S C Euskirchen E S and Commane R 2022 Using atmospheric observations to quantify annual biogenic carbon dioxide fluxes on the Alaska North Slope *Biogeosci. Discuss.* **19** 1–26
[39] Tan Z and Zhuang Q 2015 Arctic lakes are continuous methane sources to the atmosphere under warming conditions *Environ. Res. Lett.* **10** 054016
[40] Downing J A *et al* 2006 The global abundance and size distribution of lakes, ponds, and impoundments *Limnol. Oceanogr.* **51** 2388–97
[41] Pekel J-F, Cottam A, Gorelick N and Belward A S 2016 High-resolution mapping of global surface water and its long-term changes *Nature* **540** 418–22
[42] Muster S *et al* 2019 Size distributions of Arctic waterbodies reveal consistent relations in their statistical moments in space and time *Front. Earth Sci.* **7** 5
[43] Virtanen T and Ek M 2014 The fragmented nature of tundra landscape *Int. J. Appl. Earth Obs. Geoinf.* **27** 4–12
[44] Liljedahl A K *et al* 2016 Pan-Arctic ice-wedge degradation in warming permafrost and its influence on tundra hydrology *Nat. Geosci.* **9** 312–8
[45] Cooley S W, Smith L C, Stepan L and Mascaro J 2017 Tracking dynamic northern surface water changes with high-frequency planet CubeSat imagery *Remote Sens.* **9** 1306
[46] Wickland K P, Jorgenson M T, Koch J C, Kanevskiy M and Striegl R G 2020 Carbon dioxide and methane flux in a dynamic Arctic Tundra landscape: decadal-scale impacts of ice wedge degradation and stabilization *Geophys. Res. Lett.* **47** e2020GL089894
[47] Olefeldt D *et al* 2021 The Boreal–Arctic Wetland and Lake Dataset (BAWLD) *Earth Syst. Sci. Data* **13** 5127–49
[48] Thornton B F, Wik M and Crill P M 2016 Double-counting challenges the accuracy of high-latitude methane inventories *Geophys. Res. Lett.* **43** 12569–77
[49] Zolkos S, MacDonald E, Hung J K Y, Schade J D, Ludwig S Mann P J and Natali S 2022 Physiographic controls and wildfire effects on aquatic biogeochemistry in tundra of the Yukon-Kuskokwim Delta, Alaska *J. Geophys. Res.: Biogeosci.* **127** e2022JG006891
[50] Chang R Y-W *et al* 2014 Methane emissions from Alaska in 2012 from CARVE airborne observations *Proc. Natl Acad. Sci.* **111** 16694–9
[51] Chen X, Bohn T J and Lettenmaier D P 2015 Model estimates of climate controls on pan-Arctic wetland methane emissions *Biogeosciences* **12** 6259–77
[52] Miller S M *et al* 2016 A multiyear estimate of methane fluxes in Alaska from CARVE atmospheric observations *Glob. Biogeochem. Cycles* **30** 1441–53
[53] Commane R *et al* 2017 Carbon dioxide sources from Alaska driven by increasing early winter respiration from Arctic tundra *Proc. Natl Acad. Sci.* **114** 5361–6
[54] Bartlett K B, Crill P M, Sass R L, Harriss R C and Dise N B 1992 Methane emissions from tundra environments in the Yukon-Kuskokwim Delta, Alaska *J. Geophys. Res.* **97** 16645
[55] Fan S M, Wofsy S C, Bakwin P S, Jacob D J, Anderson S M Kebabian P L and Fitzjarrald D R 1992 Micrometeorological measurements of CH_4 and CO_2 exchange between the atmosphere and subarctic tundra *J. Geophys. Res.* **97** 16627–43
[56] Ludwig S, Holmes R, Natali S, Schade J and Mann P 2018 Yukon-Kuskokwim Delta fire: aquatic data, Yukon-Kuskokwim Delta Alaska, 2015–2016 *Arctic Data Center* (https://doi.org/10.18739/A22804Z8M)
[57] Ludwig S, Holmes R, Natali S, Schade J and Mann P 2018 Polaris Project 2017: aquatic isotopes, carbon, and nitrogen Yukon-Kuskokwim Delta: Alaska *Arctic Data Center* (https://doi.org/10.18739/A23775V7T)
[58] Hird J, DeLancey E, McDermid G and Kariyeva J 2017 Google earth engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping *Remote Sens.* **9** 1315
[59] LaRocque A, Phiri C, Leblon B, Pirotti F, Connor K and Hanson A 2020 Wetland mapping with Landsat 8 OLI, Sentinel-1, ALOS-1 PALSAR, and LiDAR Data in Southern New Brunswick, Canada *Remote Sens.* **12** 2095
[60] Wang L and Liu H 2006 An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling *Int. J. Geogr. Inf. Sci.* **20** 193–213
[61] Cihlar J, St.-Laurent L and Dyer J A 1991 Relation between the normalized difference vegetation index and ecological variables *Remote Sens. Environ.* **35** 279–98
[62] Gao B 1996 NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space *Remote Sens. Environ.* **58** 257–66
[63] Ludwig S M *et al* 2023 CO_2 and CH_4 fluxes from waterbodies, and landcover map, YK Delta, Alaska, 2016–2019 *ORNL DAAC* (https://doi.org/10.3334/ORNLDAAC/2178)
[64] Mullen A, Watts J D, Rogers B M, Carroll M L, Caraballo-Vega J A, Noomah J and Natali S 2022 ABoVE: lake and pond extents in Alaskan Boreal and Tundra Subregions, 2019–2021 *ORNL DAAC* (https://doi.org/10.3334/ORNLDAAC/2134)
[65] Chen S, Hu C, Barnes B B, Wanninkhof R, Cai W-J, Barbero L and Pierrot D 2019 A machine learning approach to estimate surface ocean pCO_2 from satellite measurements *Remote Sens. Environ.* **228** 203–26
[66] Topp S N, Pavelsky T M, Jensen D, Simard M and Ross M R V 2020 Research trends in the use of remote sensing for inland water quality science: moving towards multidisciplinary applications *Water* **12** 169
[67] Kim Y W, Kim T, Shin J, Lee D-S, Park Y-S, Kim Y and Cha Y 2022 Validity evaluation of a machine-learning model for chlorophyll a retrieval using Sentinel-2 from inland and coastal waters *Ecol. Indic.* **137** 108737

[68] Saberioon M, Brom J, Nedbal V, Souček P and Císař P 2020 Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters *Ecol. Indic.* **113** 106236
[69] Shen M, Luo J, Cao Z, Xue K, Qi T, Ma J, Liu D, Song K, Feng L and Duan H 2022 Random forest: an optimal chlorophyll-a algorithm for optically complex inland water suffering atmospheric correction uncertainties *J. Hydrol.* **615** 128685
[70] Li S *et al* 2021 Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm *Sci. Total Environ.* **778** 146271
[71] Elith J, Leathwick J R and Hastie T 2008 A working guide to boosted regression trees *J. Anim. Ecol.* **77** 802–13
[72] Goldstein A, Kapelner A, Bleich J and Pitkin E 2015 Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation *J. Comput. Graph. Stat.* **24** 44–65
[73] More S 2022 Identifying and overcoming transformation bias in forecasting models *Proc. of the 8th ACM SIGKDD Int. Workshop on Mining and Learning from Time Series—Deep Forecasting: Models, Interpretability, and Applications*
[74] Cole J J and Caraco N F 1998 Atmospheric exchange of carbon dioxide in a low-wind oligotrophic lake measured by the addition of SF6 *Limnol. Oceanogr.* **43** 647–56
[75] Cole J J, Bade D L, Bastviken D, Pace M L and Van de Bogert M 2010 Multiple approaches to estimating air-water gas exchange in small lakes: gas exchange in lakes *Limnol. Oceanogr.: Methods* **8** 285–93
[76] McClain M E *et al* 2003 Biogeochemical hot spots and hot moments at the interface of terrestrial and Aquatic ecosystems *Ecosystems* **6** 301–12
[77] Schmidt K S and Skidmore A K 2003 Spectral discrimination of vegetation types in a coastal wetland *Remote Sens. Environ.* **85** 92–108
[78] Davranche A, Lefebvre G and Poulin B 2010 Wetland monitoring using classification trees and SPOT-5 seasonal time series *Remote Sens. Environ.* **114** 552–62
[79] Bartsch A, Widhalm B, Leibman M, Ermokhina K, Kumpula T, Skarin A and Pointner G 2020 Feasibility of Tundra Vegetation Height Retrieval from Sentinel-1 and Sentinel-2 Data *Remote Sens. Environ.* **237** 111515
[80] Mahdianpari M, Jafarzadeh H, Granger J E, Mohammadimanesh F, Brisco B, Salehi B, Homayouni S and Weng Q 2020 A large-scale change monitoring of wetlands using time series Landsat imagery on Google earth engine: a case study in Newfoundland *GIScience Remote Sens.* **57** 1102–24
[81] Dabrowski J S, Charette M A, Mann P J, Ludwig S M, Natali S M, Holmes R M, Schade J D, Powell M and Henderson P B 2020 Using radon to quantify groundwater discharge and methane fluxes to a shallow, tundra lake on the Yukon-Kuskokwim Delta, Alaska *Biogeochemistry* **148** 69–89
[82] Laudon H, Berggren M, Ågren A, Buffam I, Bishop K, Grabs T, Jansson M and Köhler S 2011 Patterns and dynamics of dissolved organic carbon (doc) in boreal streams: the role of processes, connectivity, and scaling *Ecosystems* **14** 880–93
[83] Pacific V J, McGlynn B L, Riveros-Iregui D A, Welsch D L and Epstein H E 2011 Landscape structure, groundwater dynamics, and soil water content influence soil respiration across riparian-hillslope transitions in the Tenderfoot Creek Experimental Forest, Montana *Hydrol. Process.* **25** 811–27
[84] Covino T 2017 Hydrologic connectivity as a framework for understanding biogeochemical flux through watersheds and along fluvial networks *Geomorphology* **277** 133–44
[85] Kutser T, Pierson D C, Kallio K Y, Reinart A and Sobek S 2005 Mapping lake CDOM by satellite remote sensing *Remote Sens. Environ.* **94** 535–40
[86] Brezonik P L, Olmanson L G, Finlay J C and Bauer M E 2015 Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters *Remote Sens. Environ.* **157** 199–215
[87] Griffin C G, McClelland J W, Frey K E, Fiske G and Holmes R M 2018 Quantifying CDOM and DOC in major Arctic rivers during ice-free conditions using Landsat TM and ETM+ data *Remote Sens. Environ.* **209** 395–409
[88] Seekell D A and Pace M L 2011 Does the Pareto distribution adequately describe the size-distribution of lakes? *Limnol. Oceanogr.* **56** 350–6
[89] Natchimuthu S, Panneer Selvam B and Bastviken D 2014 Influence of weather variables on methane and carbon dioxide flux from a shallow pond *Biogeochemistry* **119** 403–13
[90] Natchimuthu S, Sundgren I, Gålfalk M, Klemedtsson L, Crill P, Danielsson Å and Bastviken D 2016 Spatio-temporal variability of lake CH_4 fluxes and its influence on annual whole lake emission estimates *Limnol. Oceanogr.* **61** S13–S26
[91] MacIntyre S, Bastviken D, Arneborg L, Crowe A T, Karlsson J, Andersson A, Gålfalk M, Rutgersson A, Podgrajsek E and Melack J M 2021 Turbulence in a small boreal lake: consequences for air–water gas exchange *Limnol. Oceanogr.* **66** 827–54