Geospatial analysis of Alaskan lakes indicates wetland fraction and surface water area are useful predictors of methane ebullition

Michela Savignano[1], Ethan Kyzivat[1], Laurence Smith[1], and Melanie Engram[2]

[1]Brown University
[2]University of Alaska Fairbanks

January 20, 2023

Abstract

Arctic-Boreal lakes emit methane (CH4), a powerful greenhouse gas. Recent studies suggest ebullition may be a dominant methane emission pathway in lakes but its drivers are poorly understood. Various predictors of lake methane ebullition have been proposed, but are challenging to evaluate owing to different geographical characteristics, field locations, and sample densities. Here we compare large geospatial datasets of lake area, lake perimeter, permafrost, landcover, temperature, soil organic carbon content, depth, and greenness with remotely sensed methane ebullition estimates for 5,143 Alaskan lakes. We find that lake wetland fraction (LWF), a measure of lake wetland and littoral zone area, is a leading predictor of methane ebullition (adj. $R^2 = 0.211$), followed by lake surface area (adj. $R^2 = 0.201$). LWF is inversely correlated with lake area, thus higher wetland fraction in smaller lakes may explain a commonly cited inverse relationship between lake area and methane ebullition. Lake perimeter (adj. $R^2 = 0.176$) and temperature (adj. $R^2 = 0.157$) are moderate predictors of lake ebullition, and soil organic carbon content, permafrost, lake depth, and greenness are weak predictors. The low adjusted R^2 values are typical and informative for methane attribution studies. A multiple regression model combining LWF, area, and temperature performs best (adj. $R^2 = 0.325$). Our results suggest landscape-scale geospatial analyses can complement smaller field studies, for attributing Arctic-Boreal lake methane emissions to readily available environmental variables.

Geospatial analysis of Alaskan lakes indicates wetland fraction and surface water area are useful predictors of methane ebullition

Michela J. Savignano[a,b,*], Ethan D. Kyzivat[a,b], Laurence C. Smith[a,b] and Melanie Engram[c]

[a] *Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, RI, USA;* [b]*Institute at Brown for Environment and Society, Brown University, Providence, RI, USA;* [c]*University of Alaska Fairbanks, Water and Environmental Research Center, Fairbanks, Alaska, United States*

* Corresponding Author: michela_savignano@brown.edu

Geospatial analysis of Alaskan lakes indicates wetland fraction and surface water area are useful predictors of methane ebullition

Arctic-Boreal lakes emit methane (CH_4), a powerful greenhouse gas. Recent studies suggest ebullition may be a dominant methane emission pathway in lakes but its drivers are poorly understood. Various predictors of lake methane ebullition have been proposed, but are challenging to evaluate owing to different geographical characteristics, field locations, and sample densities. Here we compare large geospatial datasets of lake area, lake perimeter, permafrost, landcover, temperature, soil organic carbon content, depth, and greenness with remotely sensed methane ebullition estimates for 5,143 Alaskan lakes. We find that lake wetland fraction (LWF), a measure of lake wetland and littoral zone area, is a leading predictor of methane ebullition (adj. $R^2 = 0.211$), followed by lake surface area (adj. $R^2 = 0.201$). LWF is inversely correlated with lake area, thus higher wetland fraction in smaller lakes may explain a commonly cited inverse relationship between lake area and methane ebullition. Lake perimeter (adj. $R^2 = 0.176$) and temperature (adj. $R^2 = 0.157$) are moderate predictors of lake ebullition, and soil organic carbon content, permafrost, lake depth, and greenness are weak predictors. The low adjusted R^2 values are typical and informative for methane attribution studies. A multiple regression model combining LWF, area, and temperature performs best (adj. $R^2 = 0.325$). Our results suggest landscape-scale geospatial analyses can complement smaller field studies, for attributing Arctic-Boreal lake methane emissions to readily available environmental variables.

Keywords: methane emissions; geospatial analysis, Arctic-Boreal lakes; upscaling; spatial regression

Introduction

Northern lakes are a major source of methane (CH_4), an important greenhouse gas shaping current and future climate change projections (AMAP 2015; Dhakal et al. 2022; Walter, Smith, and Chapin 2007). Currently, the global methane budget is estimated as +551-737 Tg CH_4 yr^{-1} (Lu et al. 2021; Saunois et al. 2020), with northern high latitude lakes, wetlands, and coastal waters contributing ~15 to 112 Tg CH_4 yr^{-1} (AMAP 2015; Bastviken et al. 2011; McGuire et al. 2009). High latitude lakes and ponds account for

~2.4-17.7 Tg CH_4 yr^{-1} (Matthews et al. 2020; Saunois et al. 2020; Wik et al. 2016) despite occupying just 6% of northern landscapes (Olefeldt et al. 2021), thus comprising a particularly potent source of methane to the atmosphere.

Lake methane emissions occur through at least four different pathways: diffusive flux, ebullitive flux, plant-mediated flux, and storage flux (Bastviken et al. 2011; Sanches et al. 2019). Diffusive flux is the best studied of these pathways, due to a relative ease of measurements and more homogeneous production throughout a water body and over time. However, numerous studies indicate that ebullition may be the dominant emission pathway, with plant-mediated and storage fluxes accounting for only a small fraction of emissions (Bastviken et al. 2011; DelSontro et al. 2016; Walter, Smith, and Chapin 2007; Wik et al. 2016). Calculations of global methane emissions based only on diffusive flux therefore underestimate total lake CH_4 flux, perhaps by as much as 277% (Sanches et al. 2019). Improved understanding of the drivers and spatial variability of lake ebullition processes is needed, particularly over landscape scales.

Numerous studies identify lake area, and water, air, or sediment temperature as strong predictors of methane ebullition emissions. Lake area is readily obtained from remote sensing (e.g. Cooley et al. 2019; Kyzivat et al. 2019, 2022; Messager et al. 2016; Muster et al. 2017; Smith et al. 2005) and is often inversely proportional to CH_4 ebullition, with smaller lakes emitting more methane per unit area than large lakes (e.g. Bastviken et al. 2004; Engram et al. 2020; Kuhn et al. 2021; Sanches et al. 2019; Wik et al. 2016). However, this is not a universally accepted conclusion, with other studies finding little correlation between lake area and methane ebullition (e.g. DelSontro et al. 2016; Kohnert et al. 2018). Nonetheless, many regression-based ebullition studies report lake area to be a leading predictor variable (Bastviken et al. 2004; Deemer and Holgerson 2021; Kuhn et al. 2021), along with temperature (air, water, and sediment

temperature are inherently correlated and thus variously used, e.g. Aben et al. 2017; Praetzel, Schmiedeskamp, and Knorr 2021; Sanches et al. 2019; Yvon-Durocher et al. 2014).

Other reported predictors of lake ebullitive flux include lake depth, organic carbon availability, precipitation input, trophic status, littoral zone fraction, and permafrost presence. Depth has a strong impact on ebullition, with shallower lakes having higher ebullitive fluxes than deeper lakes (e.g. Wik et al. 2016). Bastviken et al. (2004) find that the probability of ebullition decreases from ~80% at 0.5-1 m to ~10% at 4-8 m depth, for example. Organic carbon availability, in the form of dissolved organic carbon (DOC) (Bastviken et al. 2004; Deemer and Holgerson 2021; Negandhi et al. 2013; Sanches et al. 2019) and/or sediment carbon (Negandhi et al. 2013; Walter Anthony et al. 2021; Wik et al. 2018) is also an important factor in CH_4 production and emissions. Wik et al. (2016) suggest that temperature is a driving factor in lake methane ebullition only if there is sufficient organic carbon available for CH_4 production. Precipitation and water level sometimes correlate with lake CH_4 emissions, likely due to increased organic carbon availability, turbidity and/or mixing (Liu, Xu, and Li 2018; Sanches et al. 2019). Trophic status is considered a significant factor in methane emissions (because eutrophic waters generally have higher emissions [Zhou et al. 2020]), and can be estimated from remote sensing of lake chlorophyll-a or "greenness" (Bastviken et al. 2004; DelSontro et al. 2016). Other variables such as littoral and vegetated zone area (e.g. Kyzivat et al. 2022; Sanches et al. 2019) and permafrost presence (Walter Anthony et al. 2021) are also proposed as potential drivers of northern lake methane ebullition.

It is challenging to compare the relative importance of environmental predictor variables spanning different studies, time scales, and geographic areas (Julian et al.

2013). Most of the preceding studies identify methane predictors through detailed, field-based studies rather than broad-scale syntheses. Negandhi et al. (2013) and DelSontro et al. (2016), for example, consider few lakes but are detailed and field-intensive, including such variables as depth, DOC, chlorophyll concentration, and sediment temperature. Using individual regression models, DelSontro, Beaulieu, and Downing (2018) find chlorophyll-a to be the best of four predictors of methane ebullition (R^2 = 0.317, n = 65), and Kuhn et al. (2021) find DOC (adj. R^2 = 0.14, n = 72) and surface area (adj. R^2 = 0.08, n = 165) to be the best of 20 predictors. Deemer and Holgerson (2021) find a multiple regression model combining waterbody type, latitude, chlorophyll-a, and lake area to be the best predictor of methane ebullition (R^2 = 0.29, n = 130). In general, all regression studies typically yield low R^2 values, due to the inherent spatial heterogeneity of methane ebullition measurements and processes.

Alternatively, other studies examine thousands of lakes with few variables (e.g. DelSontro, Beaulieu, and Downing 2018), or lakes in aggregate only (e.g Kohnert et al. 2018). Kuhn et al. (2021) strike a balance between these two extremes by considering up to 20 variables for 1,247 lakes and ponds in one of the largest synthesis studies to date. However, ebullitive flux data are available for only 175 of these 1,247 lakes and ponds, limiting the study's ability to evaluate environmental drivers of methane ebullition. This lack of ebullitive flux data is a commonly cited shortcoming of many global CH_4 modeling studies (Sanches et al. 2019).

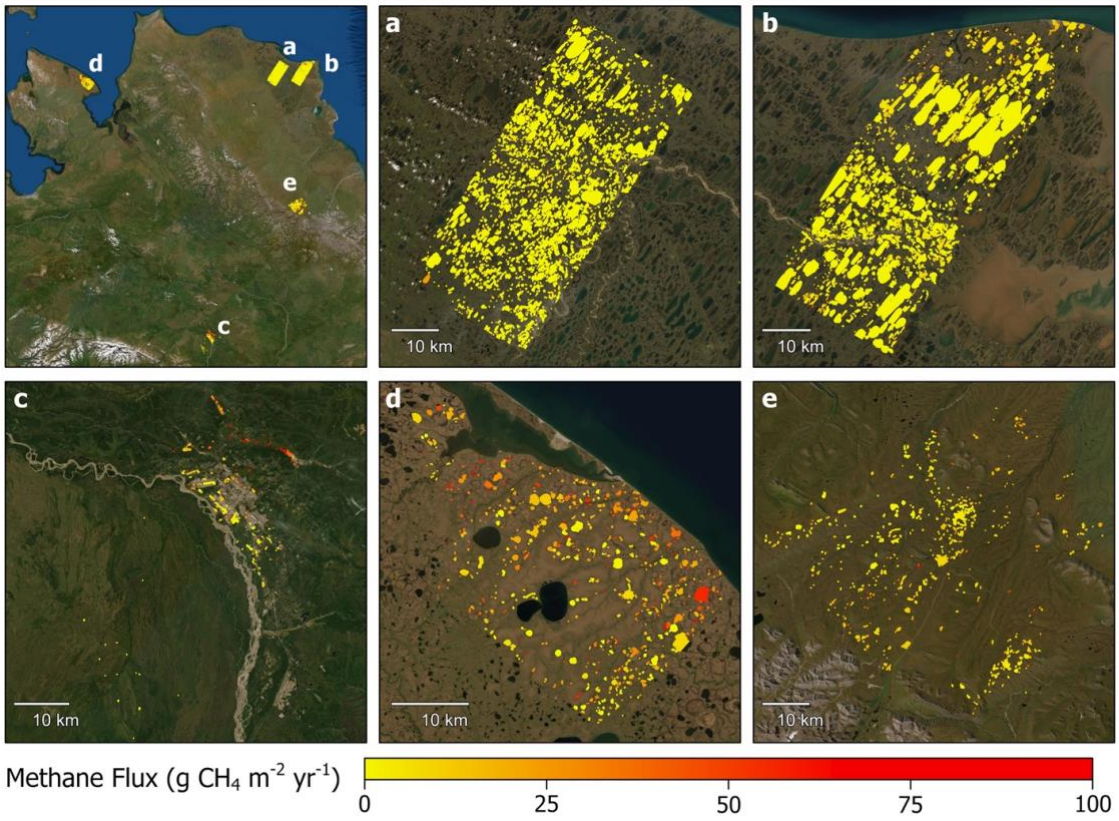


Figure 1. Remotely sensed wintertime methane ebullition flux (g CH_4 m^{-2} yr^{-1}) for 5,143 Alaskan lakes as estimated by Engram, Walter Anthony, and Meyer (2020) for five study regions: (a) Atqasuk, (b) Barrow Peninsula, (c) Fairbanks, (d) Seward Peninsula, and (e) Toolik.

Recently, Engram, Walter Anthony, and Meyer (2020) produced a remotely sensed, field-validated methane ebullitive flux dataset covering 5,143 Arctic-Boreal lakes in Alaska, vastly increasing current knowledge of geospatial variations in methane ebullition at the landscape scale. Here we use individual and multiple linear regression, as in previous studies (e.g. Deemer and Holgerson 2021; DelSontro, Beaulieu, and Downing 2018; Kuhn et al. 2021; **Table SI1**), to examine correlations between this novel lake ebullition product and widely available environmental variables. First, we compile 39 variables from 7 different datasets (**Table SI2**) and run individual regression models for each (**Table SI3**). Due to high multicollinearity between variables (Figs. SI 1-5), we select 8 representative variables for further analysis: lake area, lake perimeter,

permafrost probability, lake wetland fraction (a measure of lake wetland and littoral zone area), temperature, soil organic carbon content (SOC), lake depth, and lake greenness. Our resulting analysis of 5,130 Alaskan lakes thus combines a broad-scale study domain with a fine-scale methane ebullitive flux dataset that is exceptional in both the number of lakes and predictor variables examined. Through individual and multiple linear regression models we identify lake wetland fraction and lake area as leading predictor variables for methane ebullition. We conclude with a discussion of our study's broader implications and limitations, and some promising directions for future research.

Data and Methods

Our study domain is identical to the methane ebullitive flux dataset of Engram, Walter Anthony, and Meyer (2020), which contains ebullitive fluxes for 5,143 lakes in five dispersed regions of Alaska (Barrow, Fairbanks, Seward Peninsula, and Toolik regions, **Figure 1**). The ebullitive methane fluxes from this data product were derived from correlations between satellite-based synthetic aperture radar (SAR) backscatter and field-measured methane gas bubbles and flux trapped by lake ice and measured by on-ice bubble surveys. The SAR backscatter data were obtained from the Phased Array type L-band Synthetic Aperture Radar (PALSAR) instrument on the Advanced Land Observing Satellite (ALOS-1) satellite (Engram, Walter Anthony, and Meyer 2020). Lake areas range from 0.00345 km^2 to 58.1 km^2 with a mean area of 0.176 $\pm$1.13 km^2. Each lake was buffered inward by either 9 or 18 m to exclude high SAR backscatter returns from surrounding vegetation and shore (Engram et al. 2020). The inward buffering step renders the overall ebullition flux estimates conservative, as methane fluxes are typically higher near lake margins than centers (Juutinen et al. 2003; Kyzivat et al. 2022; Walter Anthony et al. 2016). These remotely sensed wintertime flux

estimates were then upscaled to annual values using year-round, semi-automated bubble traps submerged in a smaller subset of lakes for each of the five regions, and a region-specific ice bubble methane fraction (Engram et al. 2020).

Geospatial datasets of lake morphometry, climate, depth, SOC, greenness, landcover, and permafrost are compiled from various sources (**Table 1, Table SI2**). Lake areas and perimeters are obtained from the lake polygons provided by Engram, Walter Anthony, and Meyer (2020). Climate variables come from the NASA Daymet product, which provides daily 1 km gridded estimates of temperature, precipitation, vapor pressure, and other variables, informed by daily meteorological observations (Thornton et al. 2020). Lake depth estimates, calculated using regression analysis based on lake area and surrounding topography, are obtained from the HydroLAKES vector database for lakes 10 ha (0.1 km^2) in area and larger (Messager et al. 2016). SOC estimates, based on interpolated field observations, are obtained at 250 m resolution for six depth intervals between 0 and 200 cm (Poggio et al. 2021). Summer lake greenness data, estimated for 1,063 Alaskan lakes using Landsat imagery, are obtained from Kuhn and Butman (2021). Owing to a 10 ha minimum lake area requirement in both the greenness and depth products, these two variables are only available for 965 of the 5,143 lakes studied here. Annual Landsat-derived land cover maps spanning the NASA Arctic-Boreal Vulnerability Experiment (ABoVE) spatial domain at 30 m spatial resolution are obtained from Wang et al. (2019) and include eight terrestrial classes (evergreen forest, deciduous forest, shrubland, herbaceous, sparsely vegetated, barren), three wetland classes (fen, bog, shallows/littoral), and one water class. This dataset was not developed for purposes of wetland mapping and thus excludes certain wetland subcategories, such as forested wetlands, marshes, and shallow open-water wetlands. Based on inspection, the shallows/littoral class is rarer than expected for these shallow

arctic lakes (area-weighted average = 25% of buffered lakes), suggesting that this wetland class underreports true shallow or littoral areas. Nevertheless, it was developed for an overlapping Arctic-boreal domain and has an appropriate resolution to compare with the ebullition data so is included in our analysis. Near-surface (1 m) Permafrost probability estimates are obtained at 30 m resolution from Pastick et al. (2015).

Table 1. Selected representative variables with sources, variable descriptions, shorthand variable names, data formats, and spatial resolutions. An expanded version of this table presenting all compiled variables is available in the Supplementary Information (Table SI2).

Source	Variable Description	Variable Name	Data Format	Resolution	n
Engram, Walter Anthony, and Meyer 2020	Region (Atqasuk, Barrow, Fairbanks, Seward, or Toolik)	Region	Categorical	0.005 km^2	5,143
	Methane ebullition flux (mg m^{-2} d^{-1})	MassFlxCH4	Vector	0.005 km^2	5,143
	Lake area (km^2)	AreaSqkm			
Derived from Engram, Walter Anthony, and Meyer 2020	Lake perimeter (km)	perimeter	Tabular	0.005 km^2	5,143
Messager et al. 2016	Lake depth (m)	Depth_avg	Vector	0.1 km^2	1,224
Pastick et al. 2015	Mean permafrost probability within 1 m of the surface	Pfrst_mean	Raster	30 m	5,143
Wang et al. 2019	Fraction of combined wetland class (LWF) in lake and surrounding wetlands within 100m of lake	lbf_100n	Raster	30 m	5,139
Thornton et al. 2020	Average winter temperature (˚C)	Tavg_W	Raster	1 km	5,141
Poggio et al. 2021	Soil organic carbon content in the fine earth fraction (dg/kg)	SOC	Raster	250 m	5,130
Kuhn and Butman 2021	Mean Landsat growing season surface reflectance (Rs) in the green wavelengths	Greenness	Tabular	0.1 km^2	1,061

Geospatial datasets were compiled and reprojected to a common equal-area reference system in ArcGIS Pro 2.8.3 for comparison with the lake ebullitive flux data. Each variable from Daymet was extracted to a separate raster using the Make NetCDF Raster Layer, then the Mosaic to New Raster tool was used to merge the tiles for each variable into a single raster for ease of processing. NoData values were removed from

the permafrost probability raster, leaving all values ranging from 0 to 100. The SoilGrids raster was clipped to the extent of Alaska and each band, representing a unique variable for the 0 to 5 cm depth interval, extracted as a new raster. All datasets were reprojected to the Alaska Albers Equal Area Conic projection (EPSG:3338). Finally, these datasets were spatially joined to the original shapefile of Engram, Walter Anthony, and Meyer (2020) to enable comparisons with their methane ebullition data.

Lake perimeters were calculated from lake polygons of Engram, Walter Anthony, and Meyer (2020) using the Calculate Geometry tool in ArcGIS Pro 2.8.3, and the Field Calculator tool used to calculate perimeter-to-area (P/A) ratios. A unique identifier (LAKEID) was created for each lake, and the Zonal Statistics as Table tool used for each of the raster layers (excepting land cover variables, which are described next) to extract the mean raster values within each lake polygon. The LAKEIDs were then used to join these attribute tables containing information on permafrost, climate, and SOC to the original lake polygons. Where available, the vector datasets for lake depth and greenness were also spatially joined to the updated shapefile. Lake areas were already included as attribute information in Engram, Walter Anthony, and Meyer (2020).

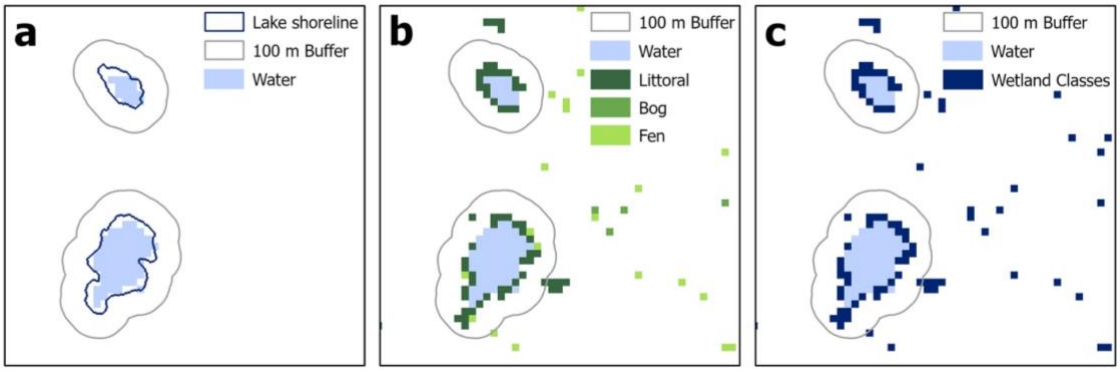


Figure 2. Schematic for LWF calculation. (a) Engram, Walter Anthony, and Meyer (2020) vector lake outline with a 100m buffer and Wang et al. (2019) water class, (b) 100m buffer and Wang et al. (2019) water, littoral, bog, and fen classes;

and (c) 100m buffer and Wang et al. (2019) water and combined wetlands classes. The two classes within the buffer in (c) are used to calculate LWF.

Land cover data of Wang et al. (2019) were processed differently for terrestrial and wetland classes as follows: First, land cover data were extracted and clipped to the extent of Alaska, and binary true/false (1/0) rasters created for each of the ten land cover variables. Lake polygons of Engram et al. (2020) were buffered outward by 100 m (**Figure 2a**) and for each of the 6 terrestrial classes, and pixels corresponding to that class falling within the buffer were summed and normalized by the total number of pixels of any land class within the buffer. The buffer distance of 100 m was chosen to include 3-4 adjacent 30 m land cover pixels, a distance deemed sufficient to robustly sample the surrounding land cover, but not so large as to extend beyond the lake's immediate catchment area. Because lake shorelines commonly differ between raster and vector datasets (which were derived using different methods and time periods) terrestrial classes were normalized to exclude any wetland class pixels present in the buffer. The normalized areal fraction of each terrestrial class within the buffer was calculated as:

$$F_t = \frac{p_t}{\sum_{t=1}^{6} p_t} \quad (1)$$

Where p_t is the number of pixels of an individual terrestrial class within the buffer (t: 1=evergreen forest, 2=deciduous forest, 3=shrubland, 4=herbaceous, 5=sparsely vegetated, 6=barren) and the buffer is defined as the 100 m buffered ring plus the original lake. Wetland classes were normalized to exclude any terrestrial class pixels present within the buffer. The normalized fraction of each wetland class within the buffer was calculated as:

$$F_w = \frac{p_w}{\sum_{w=1}^{4} p_w} \quad (2)$$

Where p_w is the number of pixels of an individual wetland class within the buffer (w: 1=fen, 2=bog, 3=shallows/littoral, 4=water). Given the relatively small areas of wetland classes within a 100 m buffer, an aggregated wetland class attribute was calculated for each lake as the sum of these three classes within the buffer (**Fig 2c**). This combined wetland class, which we call lake wetland fraction (LWF), was calculated as:

$$LWF = \sum_{w=1}^{3} F_w \quad (3)$$

A lake's LWF as defined is highly dependent on the area of its buffer, irrespective of the presence of wetland classes. Therefore, any observed correlation between methane flux and LWF might be attributable to morphological effects (i.e. shallowness, shoreline development). To test the relative contribution of morphology, a buffer ratio (BR), defined as the ratio between the areas of the 100 m buffered ring and the sum of lake area plus ring area, was calculated as follows:

$$BR = \frac{A_b}{A_b + A_l} \quad (4)$$

Where A_b is the area of the buffered ring and A_l is the area of the lake. For all classes, the normalized values were joined to the original lake polygons of Engram, Walter Anthony, and Meyer (2020) using the unique LAKEIDs, and attributes from this shapefile exported as a .csv file for further analysis.

Correlations between environmental variables and the lake methane ebullition fluxes of Engram, Walter Anthony, and Meyer (2020) were tested and compared using both individual and multiple linear regression models. The Statsmodels package in Python was used to perform ordinary least squares regression on log transformed variables, with an individual regression model created for each variable (**Table 2, Table SI3)**. Of the individual models that are not categorical variables, LWF (adj. $R^2 = 0.211$) and lake area (adj. $R^2 = 0.201$) had highest individual correlations so were also used to create multiple regression models (**Table 3**). LWF and lake area were included in all of

the multiple regression models, with a third predictor changed for each model. To avoid overfitting, no more than three variables were considered for each multivariate model (Deemer and Holgerson 2021). To avoid problems with multicollinearity (e.g. Murakami et al. 2018), only one representative variable from each dataset was selected for the multiple regression models, since variables from the same datasets were often highly correlated (**Table SI3, Figure SI 1-5**).

A few multivariate models had some degree of multicollinearity, indicated by condition numbers >30 (Haslwanter 2016). For this reason, we assessed variable importance based on univariate regression metrics: highest adjusted R^2, lowest Akaike Information Criterion (AIC), and low condition number. We also considered plausible physical causes of methane emission and did not consider spatially-autocorrelated variables, such as mean annual water vapor pressure or coverage of the sparse land cover class. In both of these cases, strong metrics likely indicate correlation with more important spatial variables, such as temperature. The categorical variable Region is not physically meaningful, so we excluded it from the multiple regression analysis. Furthermore, Region cannot be replicated, except with a spatially identical dataset, because regions are defined subjectively. All models were compared using adjusted R^2 due to its sensitivity to both the number of observations and the number of predictor variables.

Results

Individual regression models indicate that region, wetland fraction, area, perimeter, and temperature are the strongest individual predictors of lake methane ebullition (**Table 2**). Region alone has the highest predictive power (adj. $R^2 = 0.320$) of all of the variables tested, but is categorical and does not explicitly distinguish physical characteristics so

we do not consider it further. LWF has the second-highest individual predictive power (adj. R^2 = 0.211, **Figure 3a**), followed by lake area (adj. R^2 = 0.201, **Figure 3b**). Of the remaining environmental variables perimeter has moderate predictive power (adj. R^2 = 0.176) but is highly correlated with area (adj. R^2 = 0.899) so was excluded from subsequent multiple regression models. Temperature, both annual (adj. R^2 = 0.147) and winter (adj. R^2 = 0.157), is a moderately important variable in the individual regression models. The buffer ratio (adj. R^2 = 0.166, **Table SI3**) does not correlate as strongly as LWF, demonstrating that LWF is a physically meaningful variable rather than simply a function of lake morphology.

Table 2. Best-performing individual linear regression models for representative variables from each dataset, ordered by adjusted R^2. Both adjusted R^2 and Akaike Information Criterion (AIC) are measures of model quality used for model selection. Adj. R^2 considers the number of observations (n) but AIC does not; a higher adj. R^2 and lower AIC indicates a better model.

Variable	Adj. R^2	AIC	n	*p*-value
Region	0.320	3845	5143	<0.001
LWF	0.211	4562	5132	<0.001
Lake area	0.201	4678	5143	<0.001
Lake perimeter	0.176	4836	5143	<0.001
Average winter temperature	0.157	4949	5141	<0.001
SOC	0.059	5485	5129	0.001
Permafrost	0.038	5630	5143	<0.001
Lake depth*	0.029	838	1224	<0.001
Lake greenness*	0.016	574	1061	<0.001

*Depth and greenness datasets have much smaller sample sizes than other variables.

SOC, permafrost, lake depth, and greenness are weak predictors of lake methane ebullition. These remaining representative predictor variables all have individual adjusted R^2 < 0.15. In the case of SOC, other variables from the same dataset have higher adj. R^2 values (**Table SI3**) but are less physically meaningful and thus were not selected for inclusion in multiple regression models. Depth and greenness were found to

have low adjusted R^2 values (0.029 and 0.016), but they are only available in smaller sample sizes and are biased towards large lakes.

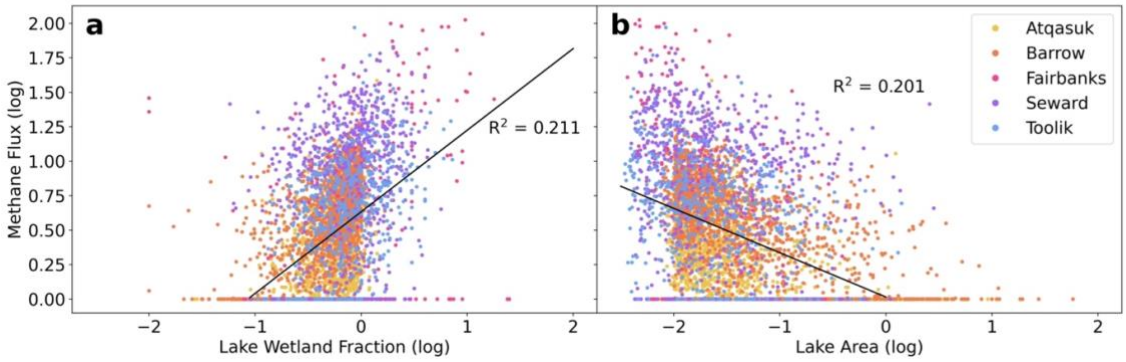


Figure 3. Methane ebullitive flux vs. (a) LWF and (b) lake area with corresponding individual regression models.

A multiple regression model combining LWF, area, and temperature performs best, with an adjusted R^2 of 0.325 (**Table 3**). This model performs better than the model regressed solely on Region (adj. $R^2 = 0.320$), which has the highest individual predictive power and accounts for numerous geographically-dependent variables. This suggests that the model may have higher predictive power than unmeasured variables for which Region is a proxy. Furthermore, for the model using lake area, LWF, and SOC, all variables are significant ($p < 0.005$) except for SOC. Multiple regression models using permafrost and SOC have adjusted R^2 values less than that of Region alone (adj. $R^2 = 0.237$ and 0.227, **Table 2**). The models including lake depth and greenness performed significantly worse than all other models (adj. $R^2 = 0.086$ and 0.037, **Table 3**).

Table 3. Multiple linear regression model results ordered by adjusted R^2. Both adjusted R^2 and Akaike Information Criterion (AIC) are measures of model quality used for model selection. Adj. R^2 considers the number of observations (n) but AIC does not; a higher adj. R^2 and lower AIC indicates a better model. All variables are log transformed.

Equation	Adj. R^2	AIC	n
Lake area + LWF + Average winter temperature	0.325	3765	5130
Lake area + LWF + Permafrost	0.237	4397	5132
Lake area + LWF + SOC	0.227	4434	5118
Lake area + LWF	0.226	4467	5132
Lake area + LWF + Lake depth*	0.086	767.2	1224
Lake area + LWF + Lake greenness*	0.037	554.1	1061

*Depth and greenness datasets have much smaller sample sizes than other variables.

As anticipated (e.g. Murakami et al., 2018), some variables are spatially correlated with each other. Temperature is inherently correlated with Region (adj. R^2 = 0.989, **Figure 4**) and thus spatial correlation of temperature, along with other driving variables, likely accounts for the high predictive power of Region alone. Models that include temperature variables within a restricted domain yield high condition numbers (e.g. AIC=2670 and 2690 for T_{avg}_W or T_{avg}_yr, respectively, see **Figure SI4, Table SI3**), even in univariate models. Such high condition numbers in multivariate models are likely caused by strong correlation between temperature and region. Similarly, LWF, the best individual predictor, is highly correlated with lake area, with an adjusted R^2 of 0.669. Perimeter is similarly correlated with LWF, with an adjusted R^2 of 0.516.

Discussion

Our broad-scale results broadly agree with smaller regression-based studies reporting the importance of lake area and/or temperature to lake methane ebullition emissions. Bastviken et al. (2004) find that a model using both lake area and total phosphorus is best for predicting methane ebullition for 13 lakes in North America and

Eurasia (adj. R^2 = 0.89), and Praetzel, Schmiedeskamp, and Knorr (2021) find that temperature is a strong individual predictor (R^2 = 0.53). DelSontro et al. (2016) find that a model using total phosphorus and sediment temperature is best for predicting ebullition (adj. R^2 = 0.52) for 13 lakes and ponds in Québec, while Deemer and Holgerson (2021) find that combining lake area, latitude, waterbody type, and chlorophyll works best (adj. R^2 = 0.29). DelSontro, Beaulieu, and Downing (2018) find that while chlorophyll a (R^2 = 0.317) is the best predictor of methane ebullition for 65 lakes worldwide, a model combining surface area and total nitrogen is a stronger predictor for a subset of 47 lakes (R^2 = 0.387). Kuhn et al. (2021) find that ebullitive fluxes from 70 lakes are best predicted by a model using lake area alone (adj. R^2 = 0.21). Similar to our own findings, these reported coefficients of determination do not exceed 0.4 for sample sizes of >100 lakes, suggesting that our regression models have predictive power approximately commensurate with other broad-scale studies.

While lake area is a frequently cited predictor for methane ebullition, our individual regression results indicate that LWF is equally important and may even be an underlying driver for the importance of area. The observed collinearity between lake area and LWF indicates that smaller lakes tend to be shallower and contain a greater proportion of land-water interface habitat, which often supports wetlands (**Figure 4**). This high correlation is in part due to the littoral class, which scales with lake perimeter, itself a correlate of lake area. Furthermore, our method for calculating wetland fraction is scale-sensitive and yields larger values for smaller lakes. The classic Shoreline Development Index (SDI) has similar scale dependence (Seekell, Cael, and Byström 2022). Thus, small and/or sinuous lakes have higher wetland fractions, regardless of land cover type. Although regression studies cannot determine causality between predictor variables, wetland presence is better supported as a mechanism for methane

production than lake area *per se* (Juutinen et al. 2003; Kyzivat et al. 2022). Studies reporting high correlations between lake area and area-normalized methane flux (e.g. Engram et al. 2020; Sanches et al. 2019; Stackpoole et al. 2017), may therefore be more appropriately interpreted as describing high correlations with unmeasured shallow and vegetated water. This interpretation is consistent with littoral zone studies reporting higher per unit flux from vegetated littoral zones than from open lake centers (Juutinen et al. 2003; Kyzivat et al. 2022; Walter Anthony et al. 2016).

Like previous broad-scale studies, our regression analyses using widely available geospatial datasets spanning thousands of lakes yield modest but statistically significant empirical correlations, thus offering a complementary way to evaluate the conclusions of smaller field studies. Both lake area and temperature, for example, are commonly cited as correlates of methane ebullition flux in detailed field studies (e.g. Bastviken et al. 2004; Praetzel, Schmiedeskamp, and Knorr 2021), which is also consistent with the results of our broad-domain study. While the reported correlations are stronger than the ones presented here, they also consider far fewer lakes and smaller spatial domains. Studies with large sample sizes are generally expected to yield lower correlations than studies with small sample sizes, consistent with the formula for correlation, which has *n* in the denominator (Ali 1987; Haslwanter 2016). Furthermore, as wetlands and LWF are difficult to classify in remotely sensed imagery, the strength of their correlation can vary based on method. In a similarly broad-scale analysis, Kyzivat et al. (2022) found that statistically significant regional correlations between lake emergent vegetation coverage and lake area never exceeded non-adjusted $R^2 = 0.5$, with a regionally aggregated $R^2 = 0.124$ (considerably lower than our own finding of non-adjusted $R^2=0.669$ for all sites combined, **Figure 4**). This discrepancy is very likely due to different methodologies and quantities being compared (i.e. lake emergent

vegetation coverage versus wetland fraction). Future work should consider more universally-applicable methods for estimating LWF, such as that for littoral area (Seekell et al. 2021). However, for environmental variables that are more readily quantifiable from remote sensing (e.g. lake area, wetland fraction, greenness) we conclude that broad-scale statistical studies such as presented here offer a powerful complement to more detailed field studies sampling smaller areas.

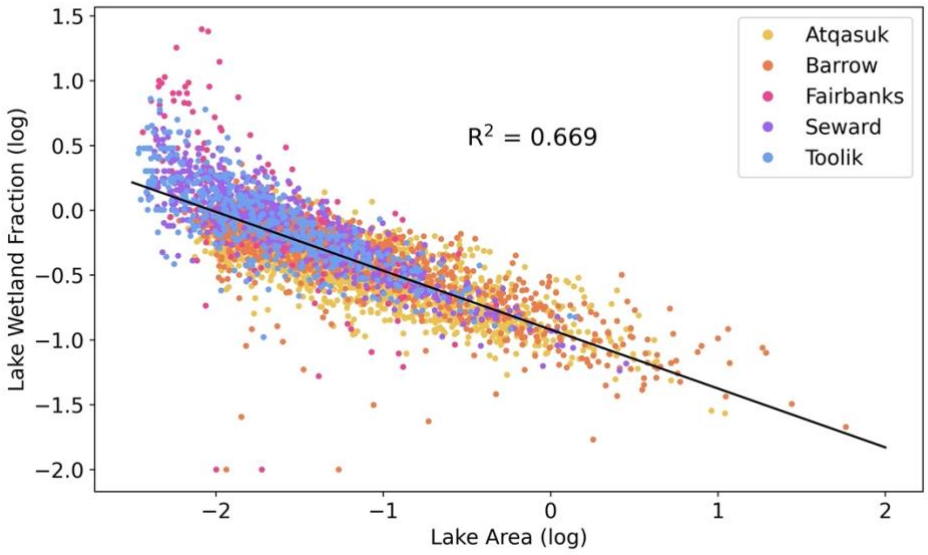


Figure 4. Observed linear relationship between lake area and lake wetland fraction (LWF).

Limitations of this study include a heterogeneous spatial domain and uncertain, multicollinear study variables. Ebullition observations are inherently noisy due to challenges associated with upscaling of field measurements and episodic bubbling events (Engram et al. 2020; Wik et al. 2016). SoilGrids-derived variables like SOC are highly interpolated, so may not correctly characterize the soil margins surrounding individual lakes. HydroLakes depths are modeled estimates, and ABoVE land cover and wetland classes are non-exhaustive. The forthcoming NASA-ISRO Synthetic Aperture

Radar (NISAR) (Kellogg et al. 2020) and Surface Water and Ocean Topography (SWOT) (Fu et al. 2012) satellite radar missions may improve Arctic-Boreal wetland mapping capability over Landsat-based approaches. From a statistical standpoint, the inherent correlations between certain variables, such as temperature and region or lake area and LWF, challenge the reliability of multiple linear regression models built from them. However, from an Earth science perspective, the latter collinearity suggests that LWF (which tends to increase in smaller, shallower lakes) is a likely physical driver of the widely reported correlation between methane flux and lake area (e.g. Sanches et al. 2019; Stackpoole et al. 2017). Furthermore, the Engram, Walter Anthony, and Meyer (2020) dataset purposefully excludes a 9-18 m inner lake buffer, which may omit littoral methane production from the reported ebullitive flux data (Engram et al. 2020). Despite this exclusion, we find a strong, broad-scale regression relationship of LWF with ebullitive flux and speculate that if these buffered areas were included, an even stronger correlation might be expected.

Regardless of these limitations, this study offers a straightforward demonstration of the value of using large environmental datasets to improve understanding of methane emissions over landscape-relevant scales. We conclude that that LWF is an underappreciated yet important predictor of lake methane ebullition, and that broad-scale geospatial studies of known lake attributes can complement conclusions drawn from smaller, field-intensive studies. Future work should continue to accumulate field and/or remotely sensed datasets of other lake attributes such as DOC, CDOM, watershed attributes, topography, and vegetation phenology (Johnston et al. 2020), as the modest predictive power of current regression models suggest that spatial variations in ebullitive methane flux (and likely diffusive, plant-mediated flux, and storage fluxes as well) are not fully captured using existing geospatial datasets. Future ebullition

studies should also consider incorporating LWF in their analyses, as inclusion of this variable should improve landscape-scale assessments of Arctic-Boreal lake methane emissions to the atmosphere.

Acknowledgements

This work was funded by the NASA Terrestrial Ecology Program Arctic-Boreal Vulnerability Experiment (ABoVE, grants 80NSSC19M0104 and 80NSSC22K1237) managed by Dr. Hank Margolis. E.D.K. also acknowledges a Future Investigators in NASA Earth and Space Science and Technology (FINESST) fellowship (80NSSC19K1361), managed by Dr. Allison Leidner.

References

Aben, R.C., N. Barros, E. Van Donk, T. Frenken, S. Hilt, G. Kazanjian, L.P. Lamers, E.T. Peeters, J.G. Roelofs, L.N. de Senerpont Domis, et al. 2017. Cross continental increase in methane ebullition under climate change. Nature communications, 8(1), pp.1-8. doi: 10.1038/s41467-017-01535-y.

Ali, M.A.. 1987. Effect of sample size on the size of the coefficient of determination in simple linear regression. Journal of Information and Optimization Sciences, 8(2), pp.209-219. doi: 10.1080/02522667.1987.10698887.

AMAP. 2015. AMAP Assessment 2015: Methane as an Arctic climate forcer. Arctic Monitoring and Assessment Programme (AMAP). Oslo, Norway. vii + 139 pp.

Bastviken, D., J. Cole, M. Pace, and L. Tranvik. 2004. Methane emissions from lakes: Dependence of lake characteristics, two regional assessments, and a global estimate. Global biogeochemical cycles, 18(4). doi: 10.1029/2004gb002238.

Bastviken, D., L.J. Tranvik, J.A. Downing, P.M. Crill, and A. Enrich-Prast. 2011. Freshwater methane emissions offset the continental carbon sink. Science, 331(6013), pp.50-50. doi: 10.1126/science.1196808.

Cooley, S.W., L.C. Smith, J.C. Ryan, L.H. Pitcher, and T.M. Pavelsky. 2019. Arctic-Boreal lake dynamics revealed using CubeSat imagery. Geophysical Research Letters, 46(4), pp.2111-2120. doi: 10.1029/2018GL081584.

Dhakal, S., J.C. Minx, F.L. Toth, A. Abdel-Aziz, M.J. Figueroa Meza, K. Hubacek, I.G.C. Jonckheere, Y.G. Kim, G.F. Nemet, S. Pachauri, et al. 2022. Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA. doi: 10.1017/9781009157926.004.

Deemer, B.R. and M.A. Holgerson. 2021. Drivers of methane flux differ between lakes and reservoirs, complicating global upscaling efforts. Journal of Geophysical Research: Biogeosciences, 126(4), p.e2019JG005600. doi: 10.1029/2019JG005600.

DelSontro, T., J.J. Beaulieu, and J.A. Downing. 2018. Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change. Limnology and Oceanography Letters, 3(3), pp.64-75. doi: 10.1002/lol2.10073.

DelSontro, T., L. Boutet, A. St-Pierre, P.A. del Giorgio, and Y.T. Prairie. 2016. Methane ebullition and diffusion from northern ponds and lakes regulated by the interaction between temperature and system productivity. Limnology and Oceanography, 61(S1), pp.S62-S77. doi: 10.1002/lno.10335.

Engram, M., K.M. Walter Anthony, T. Sachs, K. Kohnert, A. Serafimovich, G. Grosse, and F.J. Meyer. 2020. Remote sensing northern lake methane ebullition. Nature Climate Change, 10(6), pp.511-517. doi: 10.1038/s41558-020-0762-8.

Engram, M.J., K. Walter Anthony, and F.J. Meyer. 2020. ABoVE: SAR-based Methane Ebullition Flux from Lakes, Five Regions, Alaska, 2007-2010. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1790.

Fu, L.L., D. Alsdorf, R. Morrow, E. Rodriguez, and N. Mognard. 2012. SWOT: the Surface Water and Ocean Topography Mission: wide-swath altimetric elevation on Earth. Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration.

Haslwanter, T. 2016. An Introduction to Statistics with Python With Applications in the Life Sciences. 1st ed. Cham: Springer International Publishing. doi: 10.1007/978-3-319-28316-6.

Johnston, S.E., R.G. Striegl, M.J. Bogard, M.M. Dornblaser, D.E. Butman, A.M. Kellerman, K.P. Wickland, D.C. Podgorski, and R.G. Spencer. 2020. Hydrologic connectivity determines dissolved organic matter biogeochemistry in northern high-latitude lakes. Limnology and Oceanography, 65(8), pp.1764-1780. doi: 10.1002/lno.11417.

Julian, J.P., R.J. Davies-Colley, C.L. Gallegos, and T.V. Tran. 2013. Optical water quality of inland waters: A landscape perspective. *Annals of the Association of American Geographers*, *103*(2), pp.309-318. doi: 10.1080/00045608.2013.754658.

Juutinen, S., J. Alm, T. Larmola, J.T. Huttunen, M. Morero, P.J. Martikainen, and J. Silvola. 2003. Major implication of the littoral zone for methane release from boreal lakes. Global biogeochemical cycles, 17(4). doi: 10.1029/2003GB002105.

Kellogg, K., P. Hoffman, S. Standley, S. Shaffer, P. Rosen, W. Edelstein, C. Dunn, C. Baker, P. Barela, Y. Shen, et al. 2020. NASA-ISRO synthetic aperture radar (NISAR) mission. In 2020 IEEE Aerospace Conference (pp. 1-21). IEEE. doi: 10.1109/AERO47225.2020.9172638.

Kohnert, K., B. Juhls, S. Muster, S. Antonova, A. Serafimovich, S. Metzger, J. Hartmann, and T. Sachs. 2018. Toward understanding the contribution of waterbodies to the methane emissions of a permafrost landscape on a regional scale—A case study from the Mackenzie delta, Canada. Global change biology, 24(9), pp.3976-3989. doi: 10.1111/gcb.14289.

Kuhn, C., and D. Butman. 2021. ABoVE: Lake Growing Season Green Surface Reflectance Trends, AK and Canada, 1984-2019. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 0.3334/ORNLDAAC/1866.

Kuhn, M.A., R.K. Varner, D. Bastviken, P. Crill, S. MacIntyre, M. Turetsky, K. Walter Anthony, A.D. McGuire, and D. Olefeldt. 2021. BAWLD-CH 4: a comprehensive dataset of methane fluxes from boreal and arctic ecosystems. Earth System Science Data, 13(11), pp.5151-5189. doi: 10.5194/essd-13-5151-2021.

Kyzivat, E.D., L.C. Smith, F. Garcia-Tigreros, C. Huang, C. Wang, T. Langhorst, J.V. Fayne, M.E. Harlan, Y. Ishitsuka, D. Feng, and W. Dolan. 2022. The Importance of Lake Emergent Aquatic Vegetation for Estimating Arctic-Boreal Methane Emissions. Journal of Geophysical Research: Biogeosciences, 127(6), p.e2021JG006635. doi: 10.1029/2021JG006635.

Kyzivat, E.D., L.C. Smith, L.H. Pitcher, J.V. Fayne, S.W. Cooley, M.G. Cooper, S.N. Topp, T. Langhorst, M.E. Harlan, C. Horvat, and C.J. Gleason. 2019. A high-resolution airborne color-infrared camera water mask for the NASA ABoVE campaign. Remote Sensing, 11(18), p.2163. doi: 10.3390/rs11182163.

Liu, L., M. Xu, and R. Li. 2018. Modeling temporal patterns of methane effluxes using multiple regression and random forest in Poyang Lake, China. Wetlands ecology and management, 26(1), pp.103-117. doi: 10.1007/s11273-017-9558-7.

Lu, X., D.J. Jacob, Y. Zhang, J.D. Maasakkers, M.P. Sulprizio, L. Shen, Z. Qu, T.R. Scarpelli, H. Nesser, R.M. Yantosca, et al. 2021. Global methane budget and trend, 2010–2017: complementarity of inverse analyses using in situ (GLOBALVIEWplus CH 4 ObsPack) and satellite (GOSAT) observations. Atmospheric Chemistry and Physics, 21(6), pp.4637-4657. doi: 10.5194/acp-21-4637-2021.

Matthews, E., M.S. Johnson, V. Genovese, J. Du, and D. Bastviken. 2020. Methane emission from high latitude lakes: methane-centric lake classification and satellite-driven annual cycle of emissions. Scientific Reports, 10(1), pp.1-9. doi: 10.1038/s41598-020-68246-1.

McGuire, A.D., L.G. Anderson, T.R. Christensen, S. Dallimore, L. Guo, D.J. Hayes, M. Heimann, T.D. Lorenson, R.W. Macdonald, and N. Roulet. 2009. Sensitivity of the carbon cycle in the Arctic to climate change. Ecological Monographs, 79(4), pp.523-555. doi: 10.1890/08-2025.1.

Messager, M.L., B. Lehner, G. Grill, I. Nedeva, and O. Schmitt. 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature communications, 7(1), pp.1-11. doi: 10.1038/ncomms13603. Data is available at www.hydrosheds.org.

Murakami, D., B. Lu, P. Harris, C. Brunsdon, M. Charlton, T. Nakaya, and D.A. Griffith. 2019. The importance of scale in spatially varying coefficient modeling. *Annals of the American Association of Geographers*, *109*(1), pp.50-70. doi: 10.1080/24694452.2018.1462691.

Muster, S., K. Roth, M. Langer, S. Lange, F. Cresto Aleina, A. Bartsch, A. Morgenstern, G. Grosse, B. Jones, A.B.K. Sannel, et al. 2017. PeRL: A circum-Arctic permafrost region pond and lake database. Earth System Science Data, 9(1), pp.317-348. doi: 10.5194/essd-9-317-2017.

Negandhi, K., I. Laurion, M.J. Whiticar, P.E. Galand, X. Xu, and C. Lovejoy. 2013. Small thaw ponds: an unaccounted source of methane in the Canadian High Arctic. PLoS One, 8(11), p.e78204. doi: 10.1371/journal.pone.0078204.

Olefeldt, D., M. Hovemyr, M.A. Kuhn, D. Bastviken, T.J. Bohn, J. Connolly, P. Crill, E.S. Euskirchen, S.A. Finkelstein, H. Genet, et al. 2021. The Boreal–Arctic Wetland and Lake Dataset (BAWLD). Earth system science data, 13(11), pp.5127-5149. doi: 10.5194/essd-13-5127-2021.

Pastick, N.J., M.T. Jorgenson, B.K. Wylie, S.J. Nield, K.D. Johnson, and A.O. Finley. 2015. Distribution of near-surface permafrost in Alaska: Estimates of present and future conditions. Remote Sensing of Environment, 168, pp.301-315. doi: 10.1016/j.rse.2015.07.019.

Poggio, L., L.M. De Sousa, N.H. Batjes, G. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil, 7(1), pp.217-240. doi: 10.5194/soil-7-217-2021.

Praetzel, L.S.E., M. Schmiedeskamp, and K.H. Knorr. 2021. Temperature and sediment properties drive spatiotemporal variability of methane ebullition in a small and shallow temperate lake. Limnology and Oceanography, 66(7), pp.2598-2610. doi: 10.1002/lno.11775.

Sanches, L.F., B. Guenet, C.C. Marinho, N. Barros, and F. de Assis Esteves. 2019. Global regulation of methane emission from natural lakes. Scientific Reports, 9(1), pp.1-10. doi: 10.1038/s41598-018-36519-5.

Saunois, M., A.R. Stavert, B. Poulter, P. Bousquet, J.G. Canadell, R.B. Jackson, P.A. Raymond, E.J. Dlugokencky, S. Houweling, P.K. Patra, and P. Ciais. 2020. The global methane budget 2000–2017. Earth system science data, 12(3), pp.1561-1623. doi: 10.5194/essd-12-1561-2020.

Seekell, D., B. Cael, S. Norman, and P. Byström. 2021. Patterns and variation of littoral habitat size among lakes. Geophysical Research Letters, 48(20), p.e2021GL095046. doi: 10.1029/2021GL095046.

Seekell, D., B.B. Cael, and P. Byström. 2022. Problems With the Shoreline Development Index—A Widely Used Metric of Lake Shape. Geophysical Research Letters, 49(10), p.e98499. doi: 10.1029/2022GL098499.

Smith, L.C., Y. Sheng, G.M. MacDonald, and L.D. Hinzman. 2005. Disappearing arctic lakes. Science, 308(5727), pp.1429-1429. doi: 10.1126/science.1108142.

Stackpoole, S.M., D.E. Butman, D.W. Clow, K.L. Verdin, B.V. Gaglioti, H. Genet, and R.G. Striegl. 2017. Inland waters and their role in the carbon cycle of Alaska. Ecological Applications, 27(5), pp.1403-1420. doi: 10.1002/eap.1552.

Thornton, M.M., R. Shrestha, Y. Wei, P.E. Thornton, S. Kao, and B.E. Wilson. 2020. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1840.

Walter Anthony, K. M., P. Lindgren, P. Hanke, M. Engram, P. Anthony, R.P. Daanen, A. Bondurant, A.K. Liljedahl, J. Lenz, G. Grosse, and B.M. Jones. 2021. Decadal-scale hotspot methane ebullition within lakes following abrupt permafrost thaw. Environmental Research Letters, 16(3), p.035010. doi: 10.1088/1748-9326/abc848.

Walter Anthony, K., R. Daanen, P. Anthony, T. Schneider von Deimling, C.L. Ping, J.P. Chanton, and G. Grosse. 2016. Methane emissions proportional to permafrost carbon thawed in Arctic lakes since the 1950s. Nature Geoscience, 9(9), pp.679-682. doi: 10.1038/ngeo2795.

Walter, K.M., L.C. Smith, and S.F. Chapin. 2007. Methane bubbling from northern lakes: present and future contributions to the global methane budget.

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365(1856), pp.1657-1676. doi: 10.1098/rsta.2007.2036.

Wang, J.A., D. Sulla-Menashe, C.E. Woodcock, O. Sonnentag, R.F. Keeling, and M.A. Friedl. 2019. ABoVE: Landsat-derived Annual Dominant Land Cover Across ABoVE Core Domain, 1984-2014. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1691.

Wik, M., J.E. Johnson, P.M. Crill, J.P. DeStasio, L. Erickson, M.J. Halloran, M.F. Fahnestock, M.K. Crawford, S.C. Phillips, and R.K. Varner. 2018. Sediment characteristics and methane ebullition in three subarctic lakes. Journal of Geophysical Research: Biogeosciences, 123(8), pp.2399-2411. doi: 10.1029/2017JG004298.

Wik, M., R.K. Varner, K.W. Anthony, S. MacIntyre, and D. Bastviken. 2016. Climate-sensitive northern lakes and ponds are critical components of methane release. Nature Geoscience, 9(2), pp.99-105. doi: 10.1038/ngeo2578.

Yvon-Durocher, G., A.P. Allen, D. Bastviken, R. Conrad, C. Gudasz, A. St-Pierre, N. Thanh-Duc, and P.A. Del Giorgio. 2014. Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. Nature, 507(7493), pp.488-491.doi: 10.1038/nature13164.

Zhou, Y., K. Song, R. Han, S. Riya, X. Xu, S. Yeerken, S. Geng, Y. Ma, and A. Terada. 2020. Nonlinear response of methane release to increased trophic state levels coupled with microbial processes in shallow lakes. Environmental Pollution, 265, p.114919. doi: 10.1016/j.envpol.2020.114919.References: see the journal's instructions for authors for details on style

Supplemental Information for Geospatial analysis of Alaskan lakes indicates wetland fraction and surface water area are useful predictors of methane ebullition

Contents

1. Supplementary Tables

Table SI1. Compiled regression results from past ebullition regression studies.

Reference	Model	R^2	Adj. R	rmse	aic	f	n	P<0.005?
Bastviken et al. 2004	log(Area)		0.78				17	yes
	log(Area) + log(TP)		0.89				13	yes, no
	log(TP)		0.46				13	no
DelSontro et al. 2016	Sediment temp (ponds)	0.59					77	yes
	Sediment temp (lakes)	0.015					83	no
	log(TP) + Sediment temp + log(TP)*Sediment temp	0.52					89	yes
Wik et al. 2016	Waterbody type					8.857	51	yes
	Sediment type					8.937		yes
	log(depth)					0.047		no
	waterbody type*sediment type					0.086		no
	waterbody type*log(depth)					8.491		no
	sediment type*log(depth)					5.238		no
	Waterbody type					0.532	51	no
	Sediment type					0.946		no
	log(area)					3.444		no
	waterbody type*sediment type					3.405		no
	waterbody type*log(area)					2.955		no
	sediment type*log(area)					0.444		no
DelSontro, Beaulieu, and Downing 2018	log(TP)	0.292		0.648			101	yes
	log(TN)	0.311		0.647			47	yes
	log(Chl a)	0.317		0.630			65	yes
	log(area)	0.013		0.775			137	no
	log(area)*log(TP)	0.292		0.648			101	yes
	log(area)*log(TN)	0.387		0.610			47	yes
	log(area)*log(chl a)	0.280		0.634			64	yes
Sanches et al. 2019	Estimation method	0.85					78	yes
	Climatic zone							yes
	Min temp							no
	Avg. temp							no
	Estimation method*max temp							no
	Estimation method*avg temp							no

	Climatic zone*max temp Climatic zone*avg temp Climatic zone*year Precipitation Climatic zone*min precip Min temp*avg temp Min temp*min precip Max temp*avg temp							no no no no no no no yes
	Area Landscape Min temp Avg temp Area*landscape	0.98					46	no yes yes yes yes
	Max precip Avg temp Year precip*DOC	0.91					19	no yes no
Kuhn et al. 2021	DOC		0.14			12.25	72	yes
	Area		0.08			13.88	165	yes
	Latitude		0.03			5.38	161	no
	Water temp		0.06			5.55	68	no
	Depth					0.02	152	no
	Area		0.21			19.85	69	yes
Deemer and Holgerson, 2021	Water temp	0.09					134	yes
	Latitude	0.04					218	yes
	ln(Area)	0.00					216	no
	ln(Max depth)	0.07					135	yes
	ln(Mean depth)	0.00					87	no
	ln(DOC)	0.00					85	no
	ln(Chl a)	0.18					143	yes
	Waterbody type + latitude + chl a + area		0.29		481.69		130	
	Waterbody type*chl a + latitude + area		0.29		483.17		130	

Table SI2. All compiled environmental variables with sources, variable descriptions, shorthand variable names, data formats, and spatial resolutions. Selected representative variables which appear in Tables 1-3 of the main text are shown in bold. Representative variable selection was based on the highest adj. R^2 and lowest AIC within each dataset, except in the cases of the climate (Thornton et al. 2020) and soil carbon (Poggio et al. 2021) datasets, where variables were so highly correlated (Figures SI4 and SI5) that physical meaningfulness of the variables was also taken into account.

Source	Variable Description	Variable Name	Data Format	Resolution	n
Engram, Walter Anthony, and Meyer 2020	Region (Atqasuk, Barrow, Fairbanks, Seward, or Toolik)	**Region**	Categorical	0.005 km^2	5,143
	Methane ebullition flux (mg m^{-2} d^{-1})	**MassFlxCH4**	Vector	0.005 km^2	5,143
	Lake area (km^2)	**AreaSqkm**			
Derived from Engram, Walter Anthony, and Meyer 2020	Lake perimeter (km)	**perimeter**	Tabular	0.005 km^2	
	Buffer ratio	bf_pr_wi			
	Perimeter-to-area (P/A) ratio	p_a_ratio			
Messager et al. 2016	Shoreline development	Shore_dev	Vector	0.1 km^2	1,224
	Lake volume (mcm)	Vol_total			
	Lake depth (m)	**Depth_avg**			
	Watershed area (km^2)	Wshd_area			
Pastick et al. 2015	Mean permafrost probability within 1 m of the surface	**Pfrst_mean**	Raster	30 m	5,143
Wang et al. 2019	Fraction of combined wetland class (LWF) in lake and surrounding wetlands within 100m of lake	**lbf_100n**	Raster	30 m	5,139
	Fraction of littoral zone in lake and surrounding wetlands within 100m of lake	littoral_100n			
	Fraction of bog in lake and surrounding wetlands within 100m of lake	bog_100n			
	Fraction of fen in lake and surrounding wetlands within 100m of lake	fen_100n			
	Fraction of sparsely vegetated land among land pixels within 100m of lake	sparseveg_100n			
	Fraction of deciduous forest among land pixels within 100m of lake	decid_100n			

	Fraction of evergreen forest among land pixels within 100m of lake	evrgrn_100n			
	Fraction of barren land among land pixels within 100m of lake	barren_100n			
	Fraction of shrubland among land pixels within 100m of lake	shrub_100n			
	Fraction of herbaceous land among land pixels within 100m of lake	herb_100n			
Thornton et al. 2020	Average winter temperature (˚C)	**Tavg_W**	Raster	1 km	5,141
	Average annual temperature (˚C)	Tavg_yr			
	Average annual vapor pressure (Pa)	Vpavg_yr			
	Maximum annual temperature (˚C)	Tmax_yr			
	Minimum annual temperature (˚C)	Tmin_yr			
	Total annual precipitation (mm)	Precip_yr			
Poggio et al. 2021	Organic carbon stocks (t/ha)	SOCS	Raster	250 m	5,130
	Soil organic carbon content in the fine earth fraction (dg/kg)	**SOC**			
	Proportion of silt particles (≥ 0.002 mm and ≤ 0.05 mm) in the fine earth fraction (g/kg)	Silt			
	Proportion of sand particles (> 0.05 mm) in the fine earth fraction (g/kg)	Sand			
	Soil pH (pHx10)	Soil_pH			
	Organic carbon density (hg/m³)	OCD			
	Total nitrogen (N) (cg/kg)	Nitrogen			
	Cation Exchange Capacity of the soil (mmol(c)/kg)	Cat_Ex			
	Bulk density of the fine earth fraction (cg/cm³)	Bulk_Dens			
	Proportion of clay particles (< 0.002 mm) in the fine earth fraction (g/kg)	Clay			
	Volumetric fraction of coarse fragments (> 2 mm) (cm3/dm3 (vol‰))	Co_Frag			
Kuhn and Butman, 2021	Mean Landsat growing season surface reflectance (Rs) in the green wavelengths	**Greenness**	Tabular	0.1 km^2	1,061

Table SI3. Individual regression results for all assembled variables sorted and shaded by dataset (colors indicate divisions between datasets). Selected representative variables which appear in Tables 1-3 of the main text are shown in bold. Representative variable selection was based on the highest adj. R^2 and lowest AIC within each dataset, except in the cases of the climate (Thornton et al. 2020) and soil carbon (Poggio et al. 2021) datasets, where variables were so highly correlated (Figures SI4 and SI5) that physical meaningfulness of the variables was also taken into account.

Variable	rs	r2_adj	aic	n	cond_no	p < 0.005
Region	**0.321**	**0.320**	**3845**	**5143**	**5.99**	**yes**
AreaSqkm	**0.201**	**0.201**	**4678**	**5143**	**5.47**	**yes**
perimeter	**0.176**	**0.176**	**4836**	**5143**	**30.1**	**yes**
bf_per_wi	0.166	0.166	4895	5143	36.6	yes
p_a_ratio	0.175	0.175	4838	5143	11.8	yes
Shore_dev	0.002	0.001	873.2	1224	8.68	no
Vol_total	0.009	0.008	864.5	1224	2.23	yes
Depth_avg	**0.026**	**0.026**	**842.9**	**1224**	**7.27**	**no**
Wshd_area	0.003	0.003	871.6	1224	1.42	no
Pfrst_mean	**0.038**	**0.038**	**5630**	**5143**	**32.3**	**yes**
lbf_100n	**0.211**	**0.211**	**4564**	**5132**	**8.49**	**yes**
barren_100n	0.007	0.006	5794	5139	33.6	yes
bog_100n	0.000	0.000	5826	5139	2760	no
decid_100n	0.028	0.027	5684	5139	92.9	yes
evrgrn_100n	0.000	0.000	5826	5139	24.0	no
fen_100n	0.030	0.030	5672	5139	38.1	yes
herb_100n	0.073	0.073	5438	5139	14.8	yes
littoral_100n	0.112	0.112	5173	5132	13.9	yes
shrub_100n	0.099	0.099	4964	5017	10.4	yes
sparseveg_100n	0.163	0.163	4915	5139	11.2	yes
Tavg_W	**0.157**	**0.157**	**4949**	**5141**	**2670**	**yes**
Tavg_yr	0.147	0.147	5011	5141	2690	yes

VPavg_yr	0.160	0.160	4934	5141	138	yes
Tmax_yr	0.046	0.046	5586	5141	2340	yes
Tmin_yr	0.043	0.042	5606	5141	6040	yes
Precip_yr	0.132	0.132	5103	5141	39.1	yes
SOCS	0.023	0.023	5679	5129	16.2	yes
SOC	**0.060**	**0.059**	**5485**	**5129**	**42.2**	**yes**
Silt	0.140	0.140	5028	5129	23.1	yes
Sand	0.010	0.010	5747	5129	24.4	yes
Soil_pH	0.075	0.075	5399	5129	14.2	yes
OCD	0.080	0.080	5370	5129	28.7	yes
Nitrogen	0.093	0.093	5297	5129	32.8	yes
Cat_Ex	0.071	0.071	5422	5129	26.4	yes
Bulk_Dens	0.116	0.116	5167	5129	12.3	yes
Clay	0.052	0.052	5527	5129	20.7	yes
Co_Frag	0.097	0.097	5278	5129	7.18	yes
Greenness	**0.017**	**0.016**	**574.4**	**1061**	**28.2**	**yes**

1. Supplementary Figures

Figure SI1. Correlation matrix for lake morphometry variables derived from Engram, Walter Anthony, and Meyer (2020).

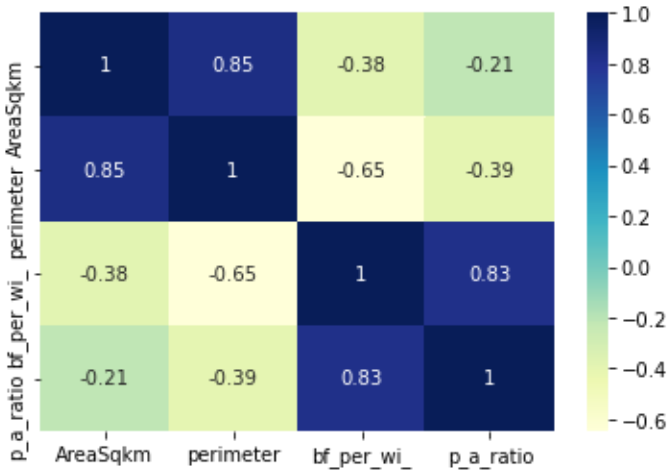


Figure SI2. Correlation matrix for lake morphometry variables from Messager et al. (2016).

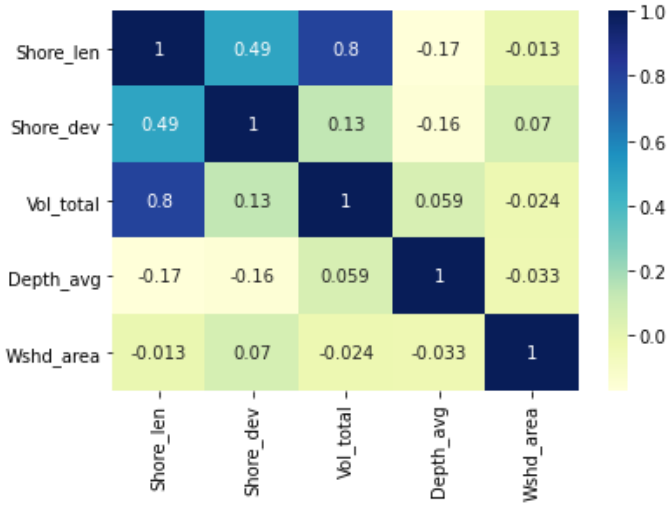


Figure SI3. Correlation matrix for land cover variables from Wang et al. (2020).

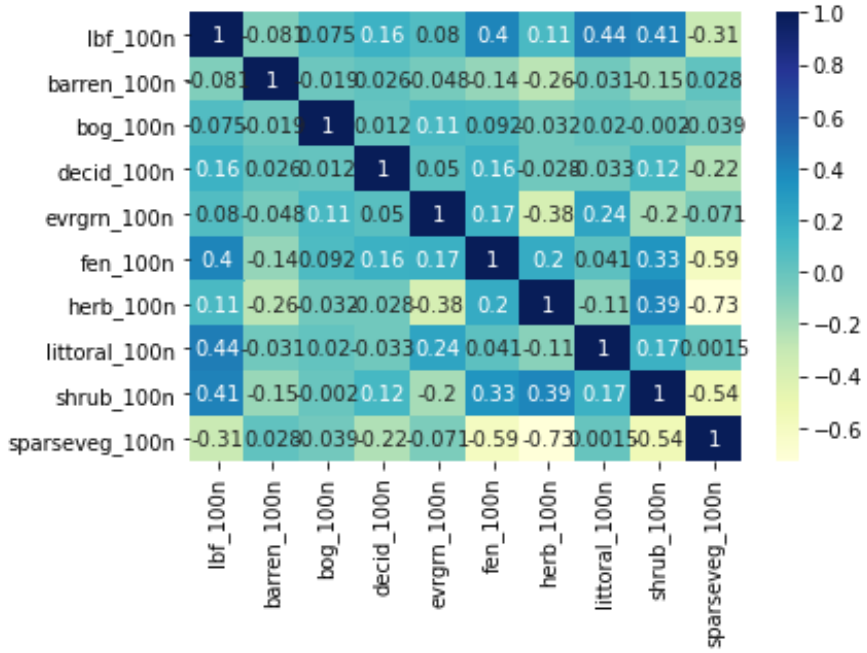


Figure SI4. Correlation matrix for climate variables from Thornton et al. (2020).

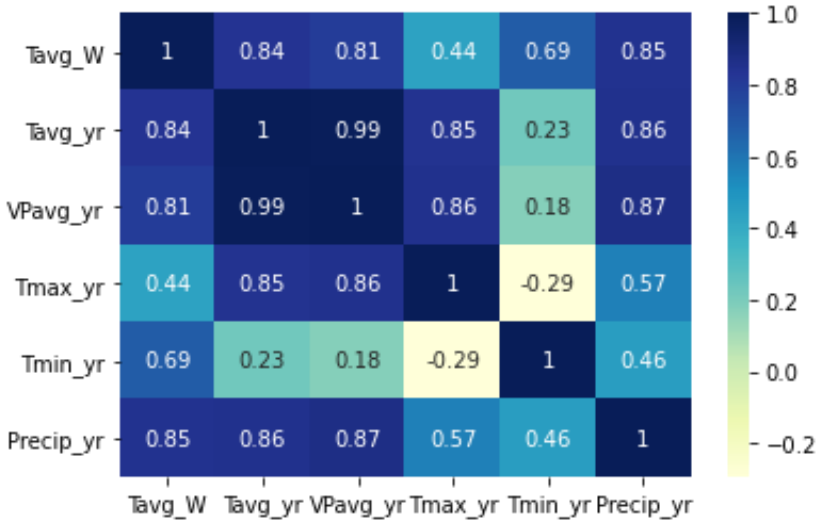


Figure SI5. Correlation matrix for soil variables from Poggio et al. (2021).

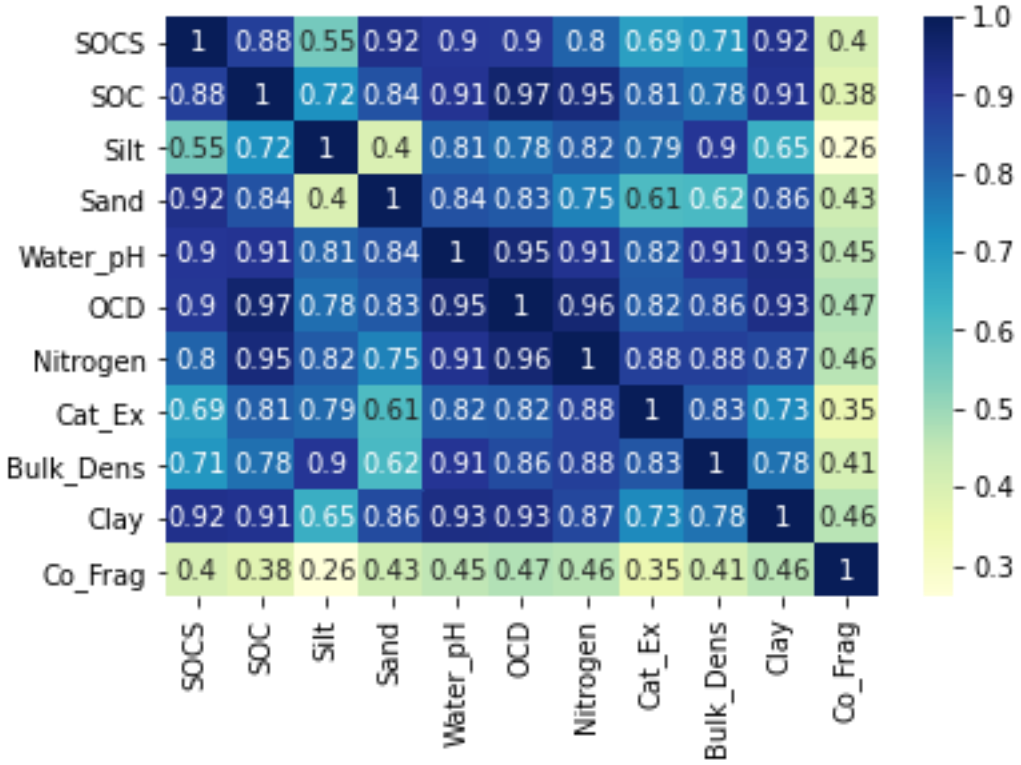
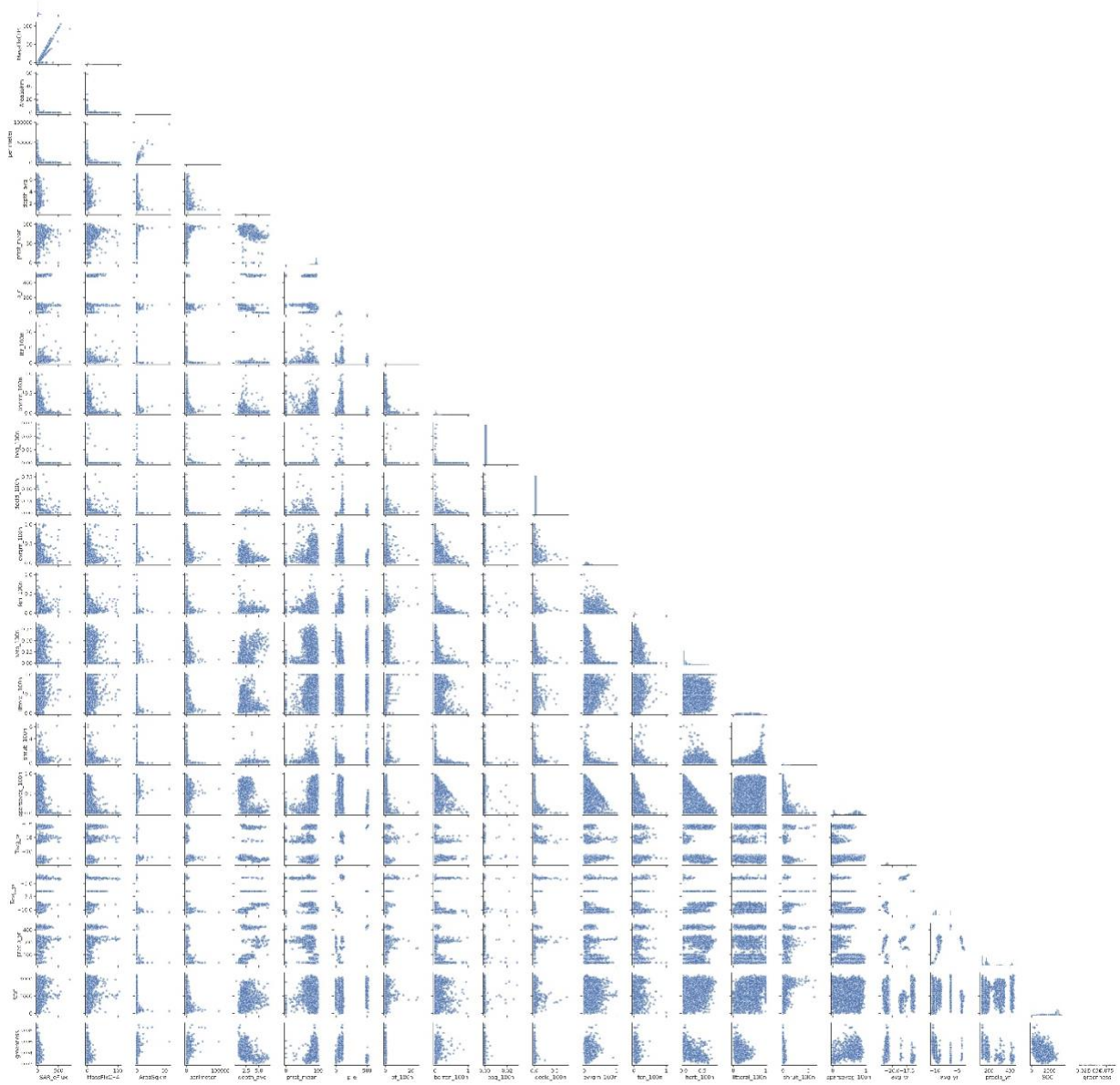


Figure SI6. Pair plots of volumetric and mass-based ebullitive methane fluxes for selected predictor variables. Histograms are shown on the diagonal elements and scatter plots on the lower diagonal elements. Figure is high-resolution and can be zoomed to view individual plots.



3. References

Bastviken, D., J. Cole, M. Pace, and L. Tranvik. 2004. Methane emissions from lakes: Dependence of lake characteristics, two regional assessments, and a global estimate. Global biogeochemical cycles, 18(4). doi: 10.1029/2004gb002238.

Deemer, B.R. and M.A. Holgerson. 2021. Drivers of methane flux differ between lakes and reservoirs, complicating global upscaling efforts. Journal of Geophysical Research: Biogeosciences, 126(4), p.e2019JG005600. doi: 10.1029/2019JG005600.

DelSontro, T., J.J. Beaulieu, and J.A. Downing. 2018. Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change. Limnology and Oceanography Letters, 3(3), pp.64-75. doi: 10.1002/lol2.10073.

DelSontro, T., L. Boutet, A. St-Pierre, P.A. del Giorgio, and Y.T. Prairie. 2016. Methane ebullition and diffusion from northern ponds and lakes regulated by the interaction between temperature and system productivity. Limnology and Oceanography, 61(S1), pp.S62-S77. doi: 10.1002/lno.10335.

Engram, M.J., K. Walter Anthony, and F.J. Meyer. 2020. ABoVE: SAR-based Methane Ebullition Flux from Lakes, Five Regions, Alaska, 2007-2010. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1790.

Kuhn, C., and D. Butman. 2021. ABoVE: Lake Growing Season Green Surface Reflectance Trends, AK and Canada, 1984-2019. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 0.3334/ORNLDAAC/1866.

Kuhn, M.A., R.K. Varner, D. Bastviken, P. Crill, S. MacIntyre, M. Turetsky, K. Walter Anthony, A.D. McGuire, and D. Olefeldt. 2021. BAWLD-CH 4: a comprehensive dataset of methane fluxes from boreal and arctic ecosystems. Earth System Science Data, 13(11), pp.5151-5189. doi: 10.5194/essd-13-5151-2021.

Messager, M.L., B. Lehner, G. Grill, I. Nedeva, and O. Schmitt. 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature communications, 7(1), pp.1-11. doi: 10.1038/ncomms13603. Data is available at www.hydrosheds.org.

Pastick, N.J., M.T. Jorgenson, B.K. Wylie, S.J. Nield, K.D. Johnson, and A.O. Finley. 2015. Distribution of near-surface permafrost in Alaska: Estimates of present and future conditions. Remote Sensing of Environment, 168, pp.301-315. doi: 10.1016/j.rse.2015.07.019.

Poggio, L., L.M. De Sousa, N.H. Batjes, G. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil, 7(1), pp.217-240. doi: 10.5194/soil-7-217-2021.

Sanches, L.F., B. Guenet, C.C. Marinho, N. Barros, and F. de Assis Esteves. 2019. Global regulation of methane emission from natural lakes. Scientific Reports, 9(1), pp.1-10. doi: 10.1038/s41598-018-36519-5.

Thornton, M.M., R. Shrestha, Y. Wei, P.E. Thornton, S. Kao, and B.E. Wilson. 2020. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1840.

Wang, J.A., D. Sulla-Menashe, C.E. Woodcock, O. Sonnentag, R.F. Keeling, and M.A. Friedl. 2019. ABoVE: Landsat-derived Annual Dominant Land Cover Across ABoVE Core Domain, 1984-2014. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1691.

Wik, M., R.K. Varner, K.W. Anthony, S. MacIntyre, and D. Bastviken. 2016. Climate-sensitive northern lakes and ponds are critical components of methane release. Nature Geoscience, 9(2), pp.99-105. doi: 10.1038/ngeo2578.