Climate signatures on lake and wetland size distributions in arctic deltas

Supplementary Material

Lawrence Vulis[1], Alejandro Tejedor[2,1], Ilya Zaliapin[3], Joel Rowland[4], and
Efi Foufoula-Georgiou[1,5]

[1]Department of Civil and Environmental Engineering, University of California Irvine, lvulis@uci.edu
[2]Department of Science and Engineering, Sorbonne University Abu Dhabi
[3]Department of Mathematics and Statistics, University of Nevada Reno
[4]Earth and Environmental Sciences Division, Los Alamos National Laboratory
[5]Department of Earth System Science, University of California Irvine

Table of Contents

I. Quality control of the Global Surface Water dataset

Thorough quality control of the water masks is necessary to reduce uncertainty in the estimated pixel water occurrence w_i and therefore the waterbody classification scheme. In particular, misclassified or poorly classified masks, e.g. where land pixels are classified as water or vice-versa, particularly in the presence of abundant unresolved pixels (i.e. pixels unable to be classified as land or water due to cloud cover, Landsat 7 striping, or other issues), introduce errors into the estimate of w_i, which lead to waterbody misclassification. To address this, we performed the following quality control procedure composed of a combination of quantitative rules and visual inspection on the GSW monthly water masks for all 12 deltas. First, for every delta we discarded from the analysis any mask over the period of record that had less than 10% of the study region resolved, as we observed misclassification errors for such poor-quality data. Second, we performed a visual inspection for significant misclassification errors, e.g. stripes of pixels classified as land or water or large swaths of the region appearing to be land only for a single year, and found only July 2016 on the Lena delta had to be discarded. Third, we identified and estimated mis-collocation errors in the GSW dataset of at least 1 pixel (30 meters) over the Yana delta from 2016 to 2018 and Lena delta from 2017 to 2018 relative to the masks from 1999 to 2015. These years were discarded from the computation of the July water pixel occurrence, w_i, but were used to estimate the average water cover since mis-collocation does not imply features were misclassified, only that their locations were shifted. No miscollocation on the order of one pixel (30-m) was observed on the other 10 deltas from 1999 to 2018. Note that the Pechora delta has not been considered in this work because of a large collocation error even in GSW v1.0 (i.e. years prior to 2016).

An example of the collocation errors is shown for the Yana delta, where waterbodies extracted from July 2018 are shifted to the north-west compared to waterbodies extracted from July 2011 (Figure S1). Due to interannual variability in surface water extent and a lack of ground control points, we were not able to compute the exact collocation error over the region and to correct the masks. Therefore, to estimate the magnitude of the miscollocation, we looked at the distribution of differences in waterbody centroids between different years, $(\Delta C_x, \Delta C_y)$. We found that the median of $(\Delta C_x, \Delta C_y) = (C_{x,2011} - C_{x,2018}, C_{y,2011} - C_{y,2018})$ was (29.24, -11.06) m, i.e. the median centroid difference between the two masks was approximately one pixel in the horizontal direction and a third of a pixel in the vertical. By examining the whole distribution of differences in waterbody centroids, we quantified that over 88% percent of waterbodies in 2018 were shifted to the southwest relative to the position of the same waterbodies in 2011 (i.e. over 88% of the centroids lay within the lower right quadrant of Figure S1b).

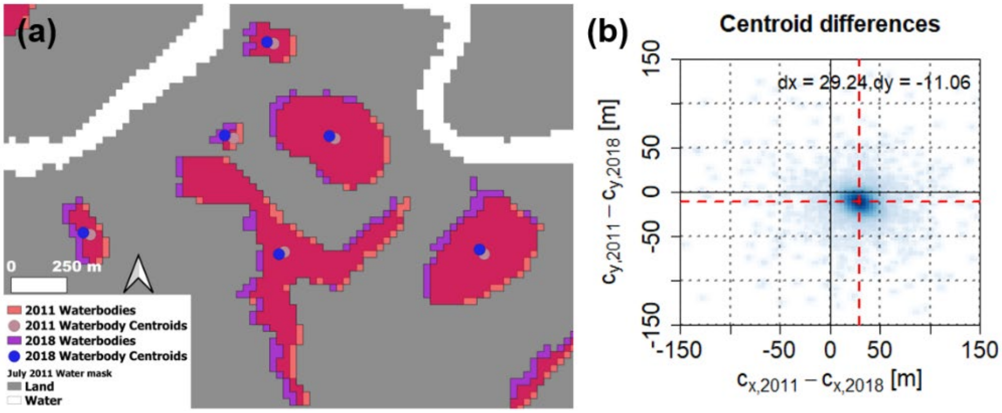

Figure S1. Collocation errors in the GSW dataset on the Yana delta. (a) Waterbodies from 2011 (red) and 2018 (purple) overlaid over the July 2011 water mask, with a clear offset between the two. The corresponding waterbody centroids are shown in brown and blue, respectively. (b) The distribution of centroid differences is shown with the median difference in each direction given by the red dashed line.

II. Hydrology of the deltas and choice of the year for waterbody mask extraction

To choose the reference year y^* in which to extract waterbody extents as objects and classify perennial lakes and ephemeral wetlands based on their year-to-year variability, we first computed for each delta and year the water cover, i.e. the fraction of valid (i.e. resolved as water or land) pixels that are classified as water over the subaerial delta, defining time series of July water cover from 1999 to 2018 (Figures S2 and S3). Then, we computed for each delta the average water cover over the period of record using the total number of valid pixels in each year as weights. Finally, y^* was chosen as the year with water cover closest to the average and at least 99% valid pixels. To test the robustness of the results, an alternative reference year, y_{alt}^* was also selected for each delta with a similar water cover to y^* and high data quality and the analysis repeated (S4 and Figures S4 and S5). To account for the heterogeneity in data quality across the range of analyzed systems, exceptions to these criteria had to be made for the Yukon, Lena, and Indigirka deltas. On the Yukon delta, the only two years satisfying the 99% valid pixel criterion were the 2008 and 2014, but these two are the wettest years on record, not years with typical hydrology. Therefore, 2017 and 2016 which had 98.7% and 98.9% valid pixels (slightly less than the 99% criterion), but close to average water cover were chosen as y^* and y_{alt}^*, respectively (Figure S2). On the Lena and Indigirka deltas only 2013 and 2016, respectively, had at least 99% valid pixels for the period of record. To perform the replication analysis, we relaxed the 99% valid pixels criterion to identify an alternative reference year y_{alt}^*. We found that 2007 had 98.5% valid pixels over the Lena delta and 98.7% valid pixels over the Indigirka delta, and therefore chose 2007 as y_{alt}^* for both deltas.

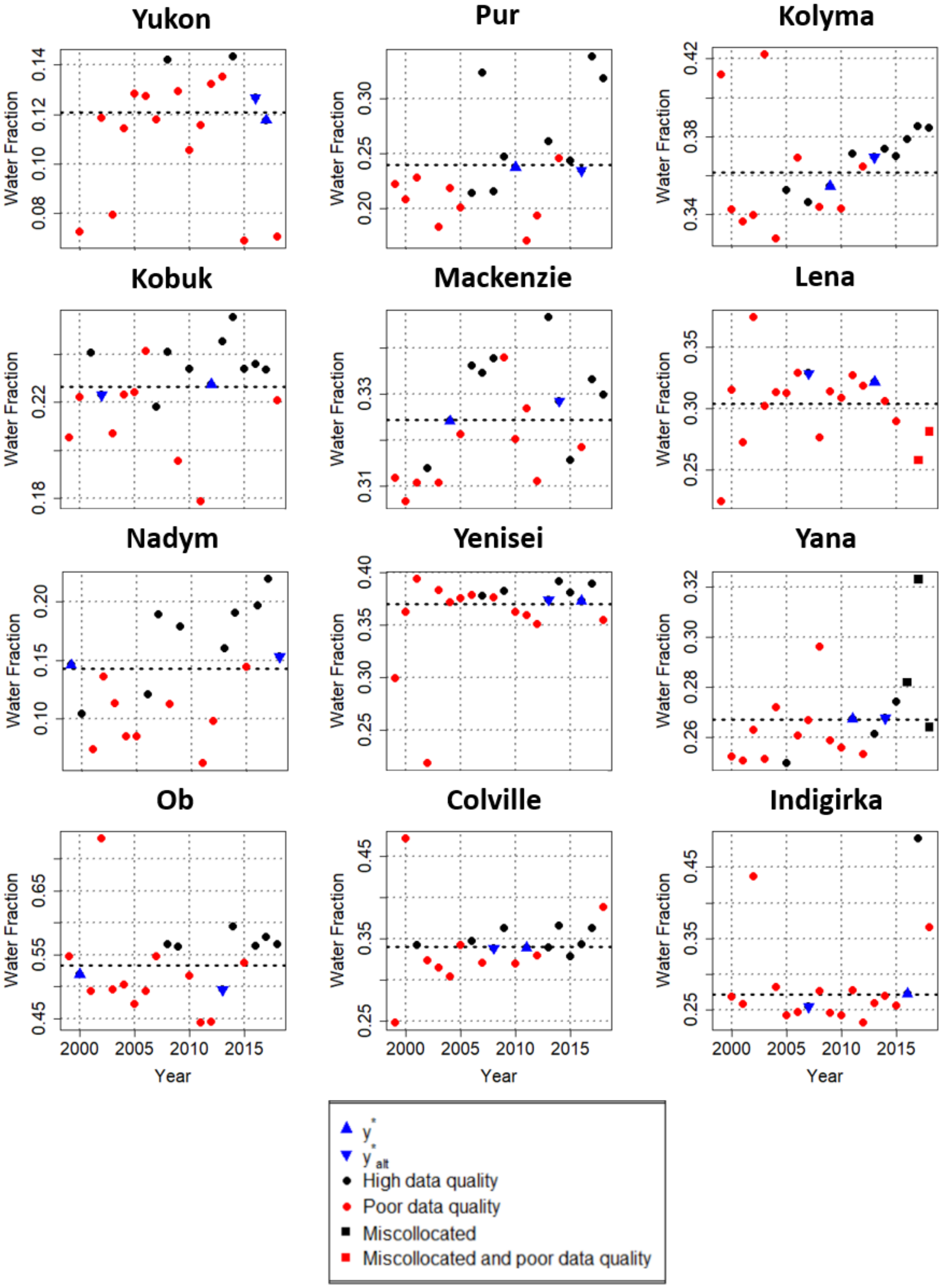

Figure S2. Surface water hydrology of arctic deltas. Time series of July water cover for every delta from 1999 to 2018. Years with at least 99% valid pixels are marked in black and years with less than 99% valid pixels in red, while years chosen for waterbody extraction are in blue triangles. Miscollocated years are shown with squares. The time series of percent valid pixels for each delta is shown in Figure S3.

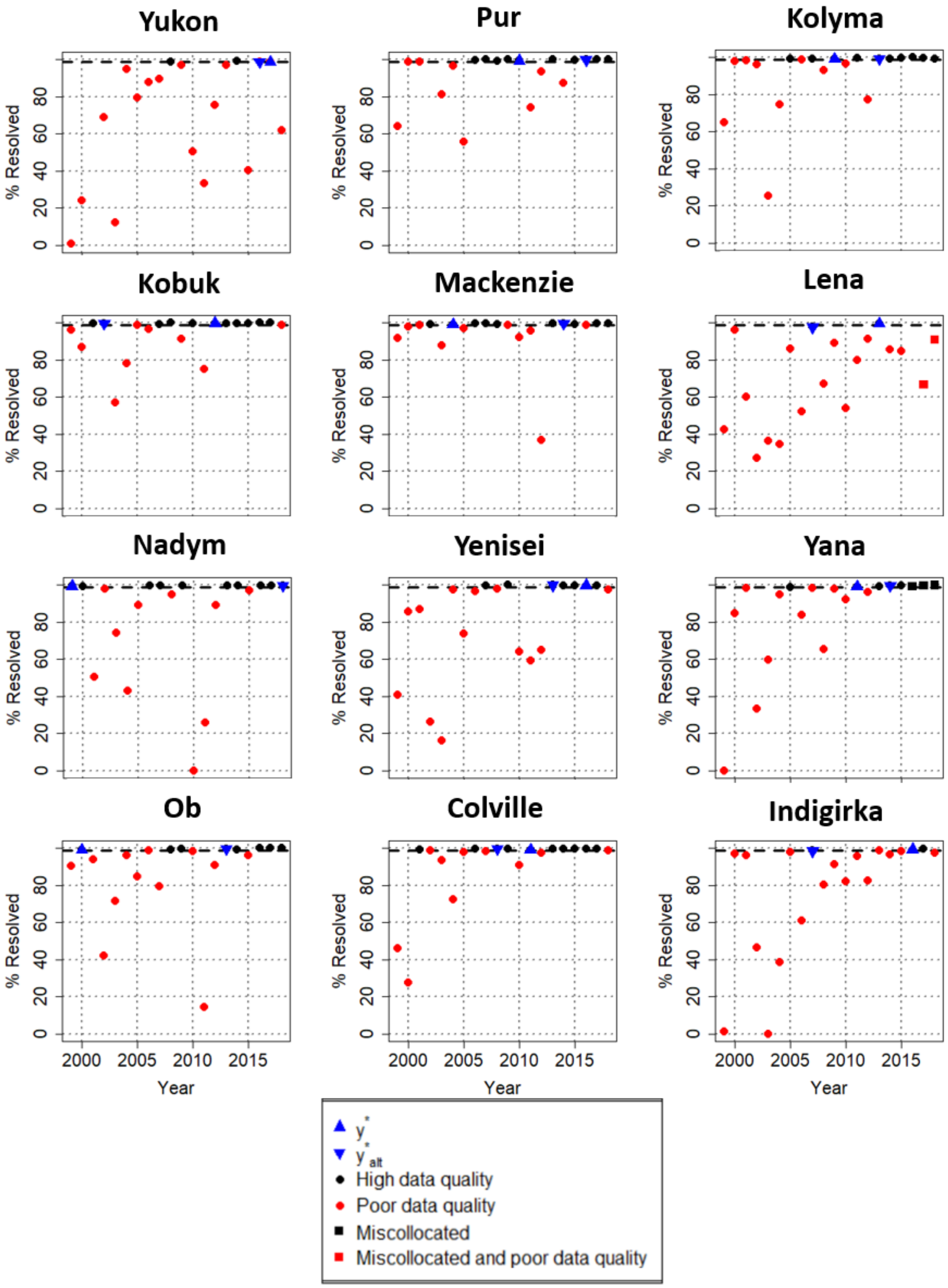

Figure S3. Observational data quality. The percent of pixels resolved in every year on the period of record for the deltas, with symbology the same as in Figure S2.

III. Proportionate growth model

Proportionate growth models, which describe processes where objects grow proportional to their size but the growth is stochastic, have seen widespread applications e.g. in modelling micro-organism sizes, income distribution, and city sizes (Crow & Shimizu, 1989; Mitzenmacher 2004). An interesting property of the proportionate growth models is that they result in a lognormal distribution of the size of the objects, with the parameters related to the parameters of the stochastic growth rate. On the basis that the greater thermal inertia of larger lakes results in lake waters remaining unfrozen for longer and maintaining greater lake to soil temperature gradients, we assume that lake growth is proportional to the size of the lake, which has been observed in Alaska (Jones et al. 2011). Then for a lake with radius r_j at the beginning of a time period j of length Δt, its growth rate $\frac{\Delta r_j}{\Delta t}$ is given by Equation (S1).

$$\frac{\Delta r_j}{\Delta t} = r_j k_j. \tag{S1}$$

We can assume that the proportional growth rate k_j at each timestep is an independent and identically distributed random variable characterized by mean γ and variance φ^2, reflecting the variability in water and soil temperature, precipitation, and soil ice content and matrix properties all of which impact lateral heat fluxes. It is easy to show from Equation (S1) that the distribution of the lake radii after some time period t (arising as the sum of the initial lake radius and its subsequent incremental growths Δr_j over the cumulative period of time) will approach a lognormal distribution (31), i.e., $\ln(r) \sim N(\gamma t, \varphi^2 t)$ (see Equation 1 with no lower bound). Assuming a circular shape of the lake it follows that $\ln(A) = \ln(\pi r^2) \sim N(2\gamma t + \ln(\pi), 4\varphi^2 t) = N(\nu, \beta^2)$, i.e. lake areas are also lognormally distributed with parameters, ν and β^2, and similarly for the volume. A similar model was proposed by Victorov et al. (2019) for thermokarst lakes although empirical testing did not reveal ubiquity of the lognormal size distribution likely due to the mixing of lakes and wetlands in the studied domains.

IV. Fitted distribution parameters and climate trends for lakes, wetlands, and waterbodies

This section contains tables and plots of the fitted distributions and climate trends for lakes, wetlands, and all waterbodies in the reference and alternative reference years. The fitted distribution parameters of lakes and wetlands for a range of waterbody occurrence index thresholds θ used to classify waterbodies extracted in y^* are in Tables (S1 to S3), lake and wetland distribution properties for waterbodies extracted in an alternative reference year y^*_{alt} in Table (S4), the fitted lognormal distribution parameters for waterbody sizes extracted in y^* in Table (S5), the plots of fitted distributions and climate trends of lakes and wetlands extracted in y^*_{alt} (Figures S4 and S5), fitted distributions and climate trends of waterbody sizes extracted in y^* (Figure S6) and boxplots of the waterbody, wetland, and lake size distributions extracted in y^* in Figure (S7).

Table S1. Properties of lake and wetland size distributions at occurrence index threshold $\boldsymbol{\theta}$ = 0.85. For each delta, the fitted lognormal parameters ν and β, and number of lakes, N_{Lake}, and p-value (p_{lake}) from a Lilliefors-corrected Kolmogorov Smirnov test (KS test), as well as the fitted power law exponent α, fitted minimum lake size x_0, observed maximum wetland size A_{max}, the number of wetlands $N_{wetland}$ in the range [x_0, A_{max}], and p-value ($p_{Wetland}$) from a KS test. We report the parameters ν and β in log$_{10}$ scale rather than in Napierian logarithmic scale (ln) as they are easier to interpret. The fitted distributions which cannot be rejected at the 5% significance level **(p > 0.05) are bolded**.

Delta	N_{Lake}	ν [-]	β [-]	p_{Lake}	$N_{Wetland}$ (above x_0)	x_0 [10^5 m^2]	A_{max} [10^5 m^2]	α [-]	$p_{Wetland}$
Yukon	1,511	3.87	0.80	**0.278**	401	0.135	2.835	2.55	**0.052**
Kobuk	1,272	4.40	0.82	**0.688**	196	0.09	3.924	2.30	**0.105**
Nadym	866	4.46	0.70	**0.404**	1,005	0.144	52.092	1.91	**0.143**
Ob	1,567	4.32	0.82	**0.843**	940	0.054	31.428	1.77	**0.306**
Pur	2,407	4.24	0.75	0.008	556	0.117	21.411	1.81	**0.289**
Mackenzie	20,318	4.37	0.75	0.025	1,404	0.189	30.168	2.39	**0.636**
Yenisei	4,058	4.62	0.60	0.038	1,028	0.153	10.620	2.47	0.049
Colville	338	4.57	0.79	**0.326**	105	0.162	7.731	2.30	**0.532**
Kolyma	3,084	4.19	0.82	**0.283**	555	0.135	14.202	2.29	**0.576**
Lena	11,265	4.49	0.74	0.008	1,353	0.477	27.783	2.63	**0.253**
Yana	10,297	4.21	0.88	**0.403**	1,563	0.144	37.872	2.07	**0.511**
Indigirka	4,875	3.91	1.08	**0.162**	1,830	0.099	42.930	1.91	**0.540**

Table S2. Properties of lake and wetland size distributions at occurrence index threshold θ = 0.80. Same as Table S1 but with waterbody classification threshold $\theta = 0.8$. Bolded p-values refer to distributions which cannot be rejected at the 5% significance level.

Delta	N_{Lake}	ν [-]	β [-]	p_{lake}	$N_{Wetland}$ (above x_0)	x_0 [10^5 m^2]	A_{max} [10^5 m^2]	α [-]	$p_{wetland}$
Yukon	1,829	3.80	0.79	**0.137**	252	0.126	1.863	2.74	**0.117**
Kobuk	1,417	4.22	0.87	**0.663**	185	0.054	3.924	2.22	**0.709**
Nadym	1,311	4.31	0.73	**0.645**	1,452	0.063	42.876	1.89	0.019
Ob	1,773	4.17	0.88	**0.825**	734	0.054	21.483	1.82	**0.298**
Pur	2,796	4.07	0.81	**0.168**	784	0.054	21.411	1.85	0.001
Mackenzie	22,495	4.24	0.79	0.016	1,019	0.153	19.620	2.37	**0.824**
Yenisei	4,889	4.50	0.62	0.023	765	0.126	9.090	2.65	**0.773**
Colville	407	4.38	0.84	**0.215**	109	0.108	7.731	2.22	**0.720**
Kolyma	3,613	3.98	0.87	**0.435**	692	0.072	14.202	2.31	**0.995**
Lena	14,156	4.35	0.76	0.047	637	0.540	19.008	2.63	**0.481**
Yana	11,567	4.08	0.91	**0.756**	2,015	0.072	12.789	2.10	**0.251**
Indigirka	5,440	3.74	1.12	**0.062**	1,433	0.099	25.299	1.91	**0.879**

Table S3. Properties of lake and wetland size distributions at occurrence index threshold θ = 0.9. Same as Table S1 but with waterbody classification threshold $\theta = 0.9$. Bolded p-values refer to distributions which cannot be rejected at the 5% significance level.

Delta	N_{Lake}	ν [-]	β [-]	p_{lake}	$N_{Wetland}$ (above x_0)	x_0 [10^5 m^2]	A_{max} [10^5 m^2]	α [-]	$p_{wetland}$
Yukon	1,118	3.95	0.81	**0.279**	185	0.369	15.993	2.69	**0.985**
Kobuk	1,022	4.62	0.76	**0.827**	100	0.333	4.311	2.76	**0.597**
Nadym	433	4.61	0.71	**0.395**	262	1.008	52.092	2.23	**0.610**
Ob	1,275	4.50	0.75	**0.677**	1,232	0.054	43.704	1.75	**0.641**
Pur	1,753	4.47	0.69	0.025	1,356	0.081	23.697	1.85	**0.816**
Mackenzie	16,395	4.55	0.70	**0.091**	2,941	0.198	30.168	2.30	0.000
Yenisei	2,883	4.76	0.58	**0.625**	497	0.486	10.620	2.73	**0.281**
Colville	248	4.77	0.78	**0.382**	167	0.162	7.731	2.22	**0.255**
Kolyma	2,218	4.42	0.79	**0.730**	352	0.378	14.202	2.37	**0.946**
Lena	7,438	4.67	0.73	0.000	2,369	0.495	27.783	2.43	**0.339**
Yana	8,286	4.34	0.86	0.016	2,806	0.144	37.872	1.96	0.000
Indigirka	3,973	4.06	1.08	**0.264**	1,113	0.270	73.431	1.93	**0.276**

Table S4. Properties of lake and wetland size distributions for waterbody extents identified in a duplicate year. Same as Table S1 but for waterbody extent identified in an alternative reference year, y^*_{alt}, with close to average water cover, and using an occurrence index threshold $\theta = 0.85$. Bolded p-values refer to distributions which cannot be rejected at the 5% significance level.

Delta	N_{Lake}	ν [-]	β [-]	p_{lake}	$N_{Wetland}$ (above x_0)	x_0 [10^5 m^2]	A_{max} [10^5 m^2]	α [-]	$p_{wetland}$
Yukon	1,340	3.90	0.81	**0.767**	961	0.081	3.357	2.10	0.003
Kobuk	1,421	4.28	0.83	**0.302**	196	0.054	2.025	2.25	**0.517**
Nadym	867	4.40	0.72	**0.396**	1,358	0.108	50.175	1.81	0.001
Ob	1,440	4.49	0.78	0.007	361	0.288	8.766	2.45	**0.238**
Pur	2,132	4.58	0.63	0.002	404	0.234	15.867	2.59	**0.106**
Mackenzie	18,256	4.46	0.73	**0.080**	2,084	0.189	28.251	2.41	0.001
Yenisei	4,040	4.62	0.60	**0.072**	344	0.324	8.127	2.84	**0.385**
Colville	441	4.25	0.88	**0.312**	140	0.072	2.934	2.20	**0.687**
Kolyma	2,321	4.38	0.80	**0.511**	988	0.153	15.183	2.10	0.029
Lena	12,467	4.37	0.77	**0.059**	1,633	0.324	48.402	2.34	**0.360**
Yana	10,145	4.31	0.84	**0.331**	2,011	0.126	16.470	2.21	**0.090**
Indigirka	5,892	3.90	1.05	**0.197**	866	0.117	23.193	2.31	**0.052**

Table S5. Lognormal waterbody size distribution parameters. Fitted lognormal parameters ν and β, for the waterbody size distribution in the reference year y^*, the number of waterbodies, $N_{waterbody}$, and KS test p-values ($p_{Waterbody}$) used to evaluate the goodness of fit. Bolded p-values refer to distributions which cannot be rejected at the 5% significance level.

Delta	$N_{Waterbody}$	ν [-]	β [-]	$p_{Waterbody}$
Yukon	2,610	2.97	0.97	**0.350**
Kobuk	1,602	3.92	0.97	**0.130**
Nadym	2,945	3.26	1.01	**0.417**
Ob	2,507	3.51	1.08	0.012
Pur	3,580	3.63	0.95	**0.251**
Mackenzie	25,995	3.96	0.88	0.000
Yenisei	6,981	3.97	0.81	0.005
Colville	606	3.50	1.09	**0.417**
Kolyma	4,557	3.35	1.04	**0.674**
Lena	25,604	3.20	1.06	0.000
Yana	14,283	3.53	1.06	0.000
Indigirka	7,807	2.70	1.36	0.043

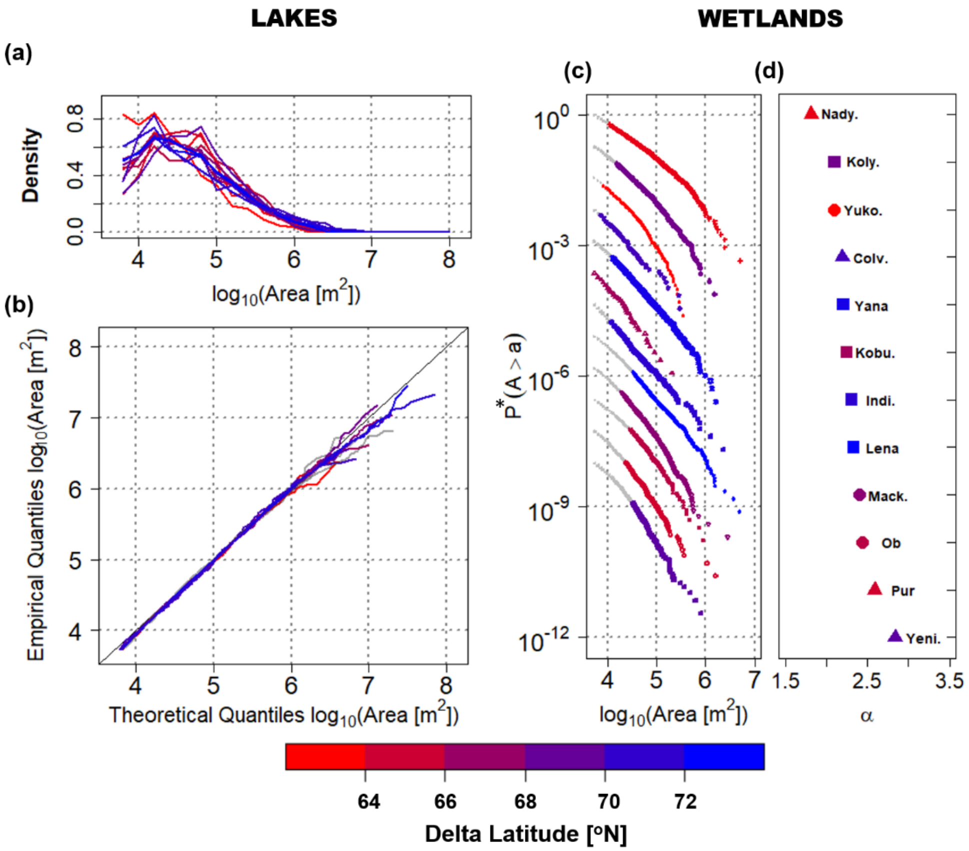

Figure S4. Lake and wetland size distributions extracted in an alternative year. Same as Figure 3 but for waterbody extents identified in an alternative reference year, y^*_{alt}, for all 12 deltas. A truncated lognormal distribution is significant for the lake area distribution at the 5% significance level (KS test) for 10 deltas. The KS test does not reject a power law for the upper tails of the wetland size distributions on 8 out of 12 deltas at a 5% significance level.

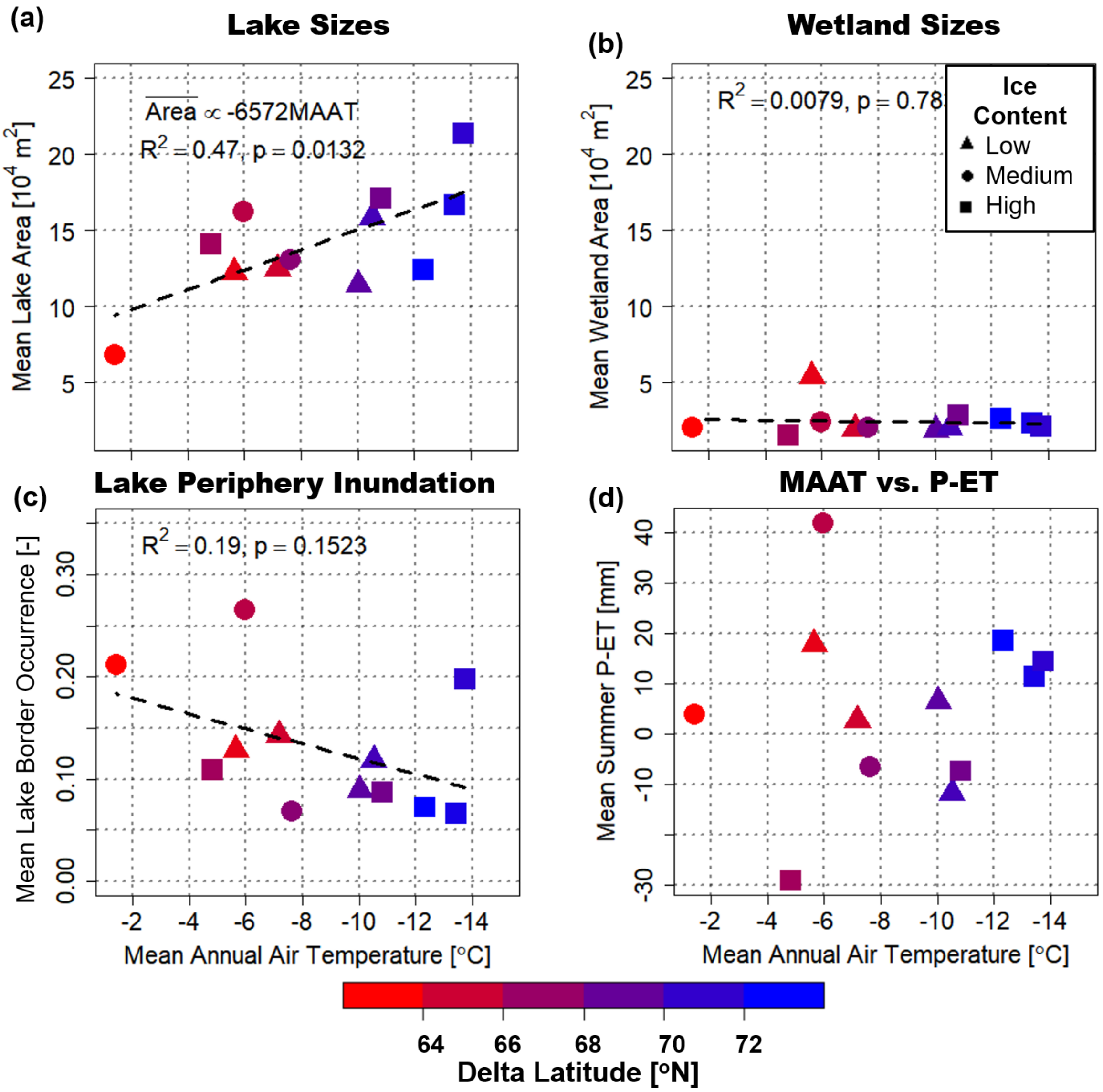

Figure S5. Climate trends for lakes and wetlands extracted in an alternative reference year. (a-c) are the same as Figures 4a, 4b, and 4e, but for waterbody extents identified in an alternative reference year, y^*_{alt}, for all 12 deltas. In (c), the presence of two large outliers (Ob and Indigirka) in (c) renders the trend non-significant. Excluding them to evaluate the relationship among the rest of deltas yields a significant trend (R^2 = 0.66, p = 0.005), supporting a possible relationship. (d) Scatterplot of 2000-2016 mean June to July precipitation minus evapotranspiration (P-ET) over the deltas versus MAAT (27), indicating vertical hydrologic budget is unrelated to differences in MAAT (R^2 = 0.013) and therefore doesn't explain the relationship in (c) or in Figure 4e.

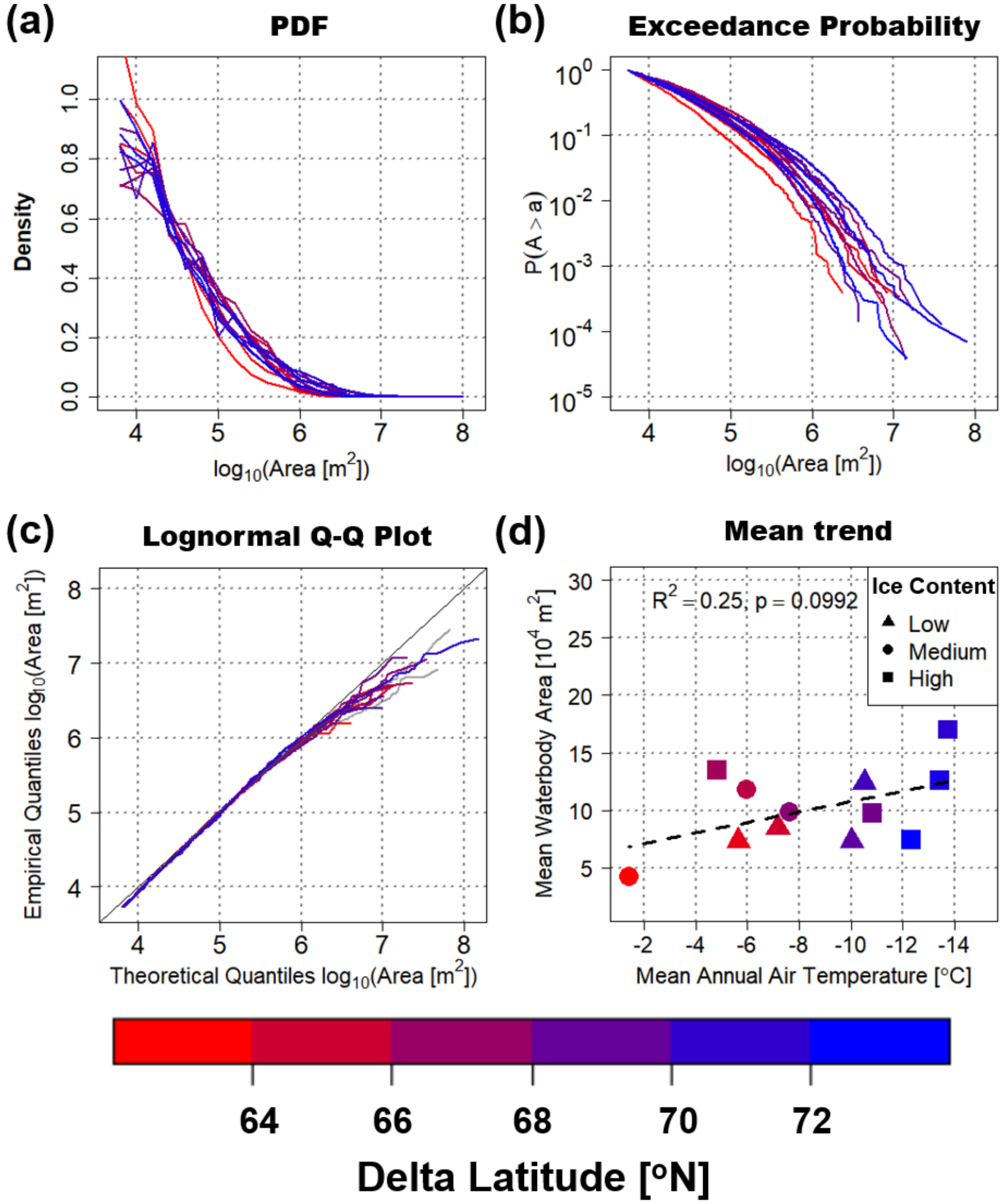

Figure S6. Waterbody size distributions and goodness of fit: (a) The PDF and (b) exceedance probability curves of the waterbody size distributions extracted in the reference year y^*, for all 12 deltas. (c) Q-Q plots of the lognormal distribution fit to the waterbody sizes, for all 12 deltas, with the fitted distributions which are not statistically significant at the 5% significance level (KS test) in grey. (d) Scatterplot of mean waterbody area and MAAT, with delta ice content indicated by point symbol shows no significant linear relationship between the two.

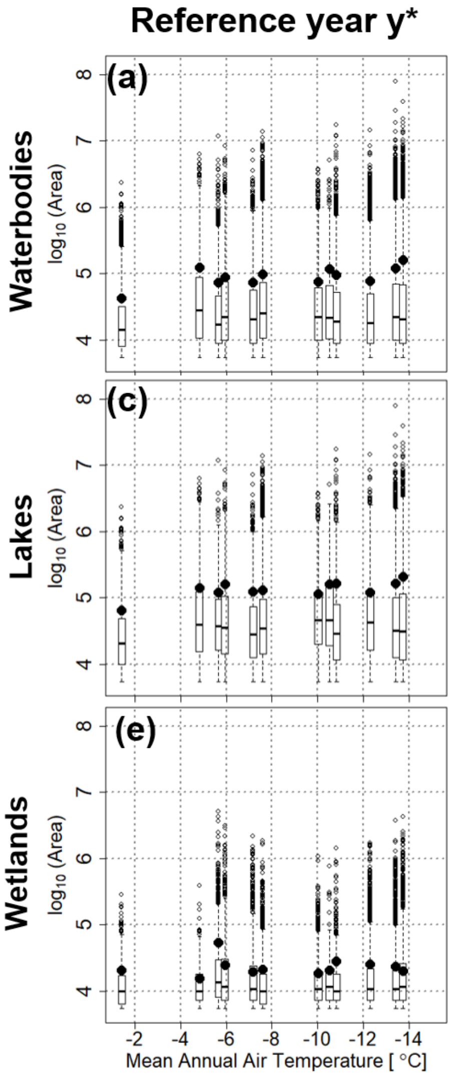

Figure S7. Waterbody, lake, and wetland size distribution boxplots. (a-c) Boxplots of size distribution for all waterbodies (a), lakes (b), and wetlands (c), with boxes representing the interquartile range, whiskers 1.5x the interquartile range, horizontal lines the sample median, and black dots the sample mean. No trend in the median lake size is observed, and a significant trend in the 90th percentile of lake sizes was also found ($p = 0.041$, $R^2 = 0.36$).

V. Relationships between the first three conditional moments

Muster et al. (2019) analyzed 30 regional size distributions of ponds and lakes from the circum-Arctic Permafrost Region Pond and Lake (PeRL) database, and found a linear relationship between the sample mean and the variance, and a hyperbolic relationship between the sample mean and the skewness coefficient of the empirical distributions when estimating these moments over a bounded range, e.g. a lower bound a and an upper bound b, also called the conditional sample moments. They also found that the statistical moments of waterbody sizes identified by inundating a digital elevation model exhibited similar relationships, and therefore determined that pond and lake sizes likely reflect landscape inundation level, rather than reflecting temperature driven growth due to climate. We compared the conditional moments of the 30 PeRL regional size distributions and the lake size distributions on the 12 arctic deltas to investigate if they displayed similar scaling relationships. In Muster et al. (2019) the bounds to compute the conditional sample moments used were $a = 100$ m^2, the minimum reliable lake size from PeRL, and $b = 10^6$ m^2 an upper bound to account for poor sample size for large lakes. We used for both the PeRL regions and the 12 deltas $a = 5.4 \cdot 10^3$ m^2, the minimum reliable lake size estimate in our study and $b = 10^6$ m^2, the same upper bound used in their study. We observed nearly identical relationships between the conditional moments from both data sets (Figure S8). As the relationships between the conditional sample moments of a fitted LN size distribution arising from proportionate growth are indistinguishable from those in the PeRL database, such relationships cannot be used to differentiate between probability distributions and the different mechanisms underlying wetland (inundation) and lake (proportionate growth) formation.

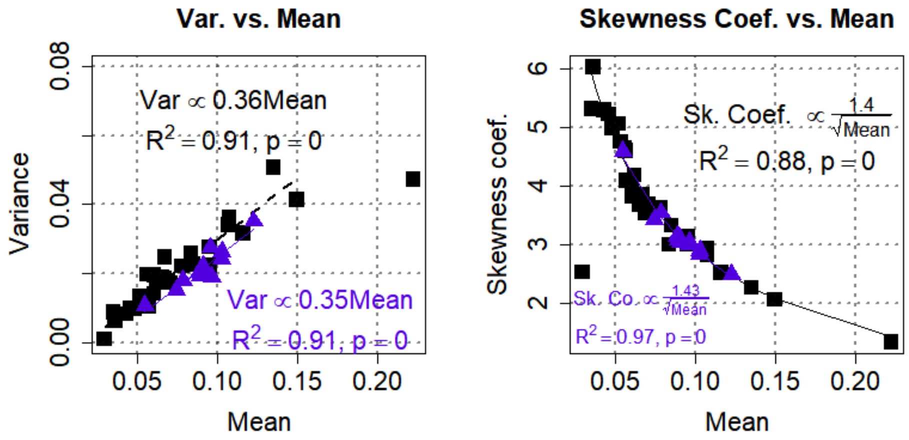

Figure S8. Lake size conditional moment scaling compared with PeRL lake and pond size conditional moment scaling. The conditional mean and conditional variance (a) and the conditional mean and the conditional skewness coefficient (b) of the lakes on arctic deltas (purple triangles) and lakes and ponds examined in Muster et al. (2019) (black squares). The outlier at (0.23 km^2, 0.05 km^4) was discarded to fit the mean and variance relationship (a) and the outlier at (.01 km^2, 2.2) were discarded to fit the mean and skewness relationship (b) for the PeRL data.