Lake ice hazards in a changing climate: Lake ice phenology with a smooth, hierarchical, time-to-event approach

Stefano Mezzini Gavin L. Simpson

Abstract

Loss of lake ice has been recognized as an important indicator of climate change. Some work has done on estimating how much ice has been lost in recent decades, but the majority of the studies use data from single sites and linear methods, so they cannot to estimate large-scale spatio-temporal trends, nor can they account for the recent acceleration in loss. This paper estimates the daily hazard of lakes freezing or thawing using a smooth, hierarchical, time-to-event approach. We fit piecewise-exponential additive models to a large dataset of 563 lakes and 623 distinct observation stations from the northern hemisphere to estimate spatio-temporal changes in lake ice onset and offset in the past 70 years. The results presented here demonstrate a widespread and accelerating but heterogeneous shift towards later freezing dates and earlier thaw dates, with an increase in lake ice cover in some areas. The hierarchical approach allowed the models to estimate average spatio-tempotal changes in lake ice, including in areas where little to no data were available. Considerations are made on the effects of the loss of lake ice on Eurasian and North American Peoples, including North American Indigenous People.

1 Introduction

Despite the important consequences the periodic freezing of lakes has for the biota that inhabit them, under-ice ecology in lakes has been until recently relatively under-studied (Hampton *et al.*, 2017). Some recent studies suggest that lake ice loss is widespread and accelerating [**(REF?)**, ***(IPCC-2021?)***], but the change is heterogeneous and some regions have seen a moderate increase in lake, river, and sea ice [***(IPCC-2021?)**, other studies?*].

Summarize what we know from other papers about the loss of lake ice.

Complete IPCC report not finalized yet, it says not to cite it

Lake ice plays an important role in the seasonal cycles of stratification and productivity within lakes. Ice cover decreases wind-induced mixis and decreases the amount of nutrients, oxygen, and light which enter the lake (**REF?**). …

Lake ice is also important from an anthropological viewpoint – many Peoples and communities depend on winter ice cover for economic, cultural, and spiritual activities (Golden, Audet & Smith, 2015; Knoll *et al.*, 2019). Many northern European countries (used to) have annual winter ice skating competitions that are (were) an important part of the local culture, while ice roads often provide essential transportation routes (i.e. ice roads) for many Indigenous Peoples in northern Canada and Alaska. Many communities also rely on ice fishing as a mean of sustenance during winter (Knoll *et al.*, 2019). The annual freezing of some lakes plays an important role in local religion and cultural identity, as is the case of Lake Suwa in Japan, whose freezing dates have been recorded since 1443 by the local Shinto temple and are still recorded today (Arakawa, 1954; Sharma *et al.*, 2016; Knoll *et al.*, 2019). Blue ice in particular has great cultural importance for multiple Indigenous Peoples in northern Canada (Golden, Audet & Smith, 2015).

A number of studies have attampted to estimate the effects of climate change on lake ice phenology, including estimating the change in frequency of lake ice formation (Sharma *et al.*,

2016) and attempting to derive the main drivers of lake ice loss (Sharma *et al.*, 2019; Lopez, Hewitt & Sharma, 2019). A limitation of these past studies, however, is that the analysis has been done on individual lakes, rather than at regional or global scales, and a substantial portion of the studies used statistically inappropriate methods that might have produced biased and inaccurate results. Some authors have analyzed multiple time series, but often times they compared estimated changes by regressing on the estimated coefficients, rather than fitting a single hierarchical model (e.g. Warne *et al.*, 2020; see Pedersen *et al.*, 2019 for hierarchical modelling).

A more efficient model should instead be able to (1) account for accelerating, non-monotonic, and spatially heterogeneous trends, (2) estimate average spatial trends, (3) allow the spatial trends to change over time, and (4) account for variation between lakes. Such a model should also relax some of the (likely problematic) assumptions of linear models, such as linearity, monotonicity, and spatio-temporal uniformity. Hierarchical Generalized Additive Models (HGAMs; see Pedersen *et al.*, 2019) are capable of accounting for (1) nonlinear effects, (2) multiple lakes in a single model, (3) the spatial relationship between lakes, and (4) interactions between smooth terms via tensor product interaction terms.

Fitting a single model to all time series at once reduces the complexity of the analysis and allows us to directly estimate common spatio-temporal trends between lakes while incorporating the variance that exists between lakes. This hierarchical approach is important because the location and morphology of a lake have strong effects on ice phenology (Woolway *et al.*, 2020). In this paper, we estimate the change in lake ice occurrence since 1950 using a hierarchical approach that allows us to fit a single model to many lake time series (Pedersen *et al.*, 2019). The freeze and thaw dates were analyzed using a time-to-event approach which allows us to estimate the probability of a lake freezing or thawing on a given day (Bender, Groll & Scheipl, 2018).

In statistical terms, the *hazard* of an event is the probability of the occurrence of an event

within a period of time Δt. For instance, an ice-free lake has an unknown probability of freezing at a given moment in time, t, and the probability of *being frozen* at time t is equal to the chance of it freezing at time t or any time before t (provided that it has not thawed afterwards). We can then define the cumulative distribution function for the probability of a lake being frozen up to time t as:

$$F_{freeze}(t) = P(T_{freeze} \leq t), \tag{1}$$

where T_{freeze} indicates the unknown true freezing time, and t is a given moment in time. Similarly, let the probability of a lake being ice-free at time t be indicated by

$$F_{thaw}(t) = P(T_{thaw} \leq t). \tag{2}$$

The probability of an event happening at or before t is equal to the complement of the event happening *after* time t, i.e., for a generalized random variable or time T, $P(T \leq t) = 1 - P(T > t)$. From a *survival analysis* perspective (Kleinbaum & Klein, 2012), $P(T > t)$ is the probability that a patient will survive up to time t, so $P(T > t)$ is commonly referred to as the *survival* probability at time t. It is calculated using the estimated survival function $\widehat{S}(t)$. With regards to lake ice phenology, $S(t)$ indicates the probability of a lake freezing or thawing after time t. Therefore, we can state that

$$P(T \leq t) = 1 - P(T > t) = 1 - S(t),$$

T is the random variable for the occurrence time of the event, and $S(t)$ is the survival function (Kleinbaum & Klein, 2012). This is true whether we are estimating the date of freeze or thaw events.

We can also estimate the hazard of an event occurring in a given period of time Δt, such

as a single day or a week. Mathematically, we can write this as $P(t < T \leq t + \Delta t)$. We can compare hazards from time periods of different Δts by dividing $P(t < T \leq t + \Delta t)$ by Δt. If we let Δt be 1 day, we can estimate the daily hazard of the event happening on any day using the *hazard* function, $\lambda(t)$. Generally, however, we allow Δt to approach zero so we can estimate the instantaneous hazard, given that the event did not occur before time t (Kleinbaum & Klein, 2012):

$$\hat{\lambda}(t) \approx \lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3}$$

To estimate the hazard of an event occurring up to and including time t, we can use the estimated cumulative hazard function

$$\widehat{\Lambda}(t) \approx \Lambda(t) = \int_0^t \lambda(T)\ dT = P(0 \leq T \leq t). \tag{4}$$

The probability of an event occurring at or before time t can then be estimated as a function of the cumulative hazard (Kleinbaum & Klein, 2012):

$$F(t) \approx \widehat{F}(t) = 1 - e^{-\widehat{\Lambda}(t)}. \tag{5}$$

Piecewise-exponential Additive Models (PAMs, see Bender, Groll & Scheipl, 2018) are a special case of Generalized Additive Models (GAMs, see Hastie & Tibshirani, 1986, 1999; Wood, 2017) which estimate the log expected hazard of events $\log\left[\mathbb{E}\left(\hat{\lambda}(t)\right)\right] = \log\left[\lambda(t)\right]$. With PAMs, it is possible to estimate equations (1)-(5). PAMs assume that the change in hazard is constant between two consecutive observations. Under these assumptions, it can be shown that the hazard function $\hat{\lambda}(t)$ has a Poisson likelihood, and thus it is possible to model $\hat{\lambda}(t)$ using a Poisson GAM (Bender, Groll & Scheipl, 2018). PAMs standardize $\log\left[\hat{\lambda}(t)\right]$ with an offset term of the log-transformed amount of time between observations (Bender, Groll & Scheipl, 2018).

Before fitting a PAM, it is important to convert the dataset to the Piecewise Exponential Data (PED) format. The PED format rearranges the data such the i^{th} observation is broken into multiple intervals $(\kappa_{j-1}, \kappa_j], j = 1, ..., J$ with their respective event indicators δ_{ij}, and offsets o_{ij}. The event indicators, $\delta_{i,j}$, take the value 1 if the event occurred in the interval $\left(\kappa_{j-1}, \kappa_j\right]$ and 0 otherwise. The offset is equal to the log-transformed interval $(\log(\kappa_j - \kappa_{j-1}))$ if the event occured in the interval $(\delta_{ij} = 1)$ and $\log\left(t_i - \kappa_j\right)$ otherwise, i.e. $o_{ij} = \log\left[\min\left(\kappa_j - \kappa_{j-1}, t_i - \kappa_{j-1}\right)\right]$. Additional information on the PED format is given in section 2.1 and in Figure 3.

In this paper, we fit PAMs to a large ice phenology dataset while accounting for spatio-temporal trends to estimate the change in the daily hazard of freezing and thawing throughout the Northern hemisphere since 1950. We estimate smooth effects of predictors to allow the hazard to vary nonlinearly over time and space. We used a hierarchical Bayesian approach to fit Hierarchical PAMs (HPAMs) which could estimate common spatial trends and the unaccounted variation between lakes (Pedersen *et al.*, 2019).

2 Methods

2.1 Lake ice datasets

2.1.1 Lake Mendota

The Lake Mendota dataset was obtained from the Environmental Data Initiative (EDI) portal (Magnuson, Carpenter & Stanley, 2021) and converted to the PED format. Since the lake froze in December and January, freeze and thaw dates were converted to the number of days post June 30^{th} to avoid the discontinuity that would have occurred if using the numeric day of year (i.e. 1-366, see Figure 3). The waiting time was divided into daily intervals for days between from 150 to 210 (November 27^{th} and January 26^{th} on non-leap years).

The earliest day of freezing was December 3^{rd} in 1976, while the latest day of freezing was January 20^{th} in 2007.

2.1.2 Global lake ice datasets

The lake ice data were obtained from the Global Lake and River Ice Phenology Database (GLRIPD, http://nsidc.org/data/G01377.html, see Benson, 2002). Prior to analysis, the GLRIPD was filtered to only include lakes with known coordinates and observations after 1950, since the majority of the observations occurred after 1950 (SI, Figure 9). Although the analysis could have been performed for the entire dataset, the dataset was reduced to decrease model fitting time and potential sampling bias, since the majority of the data was in the period 1950-1995 (Figure 9). A large portion of the observations are for temperate (45° N) North American lakes and sub-arctic (62° N) Finnish lakes, and the spatial distribution changes substantially after 1950, particularly in North America (Figures 1 and **??**). All observations in the dataset are from lakes that freeze frequently.

Since many lakes froze or thawed in December and January, freeze and thaw dates were converted to the number of days post June 30^{th} and September 30^{th}, respectively, as with the Lake Mendota example (Figure 2, and SI Figure 10). This representation is commonly used in survival analysis and event history modelling and is known as the follow-up period.

The coordinates of the lakes were corrected using Google Maps if the original location was more than 0.01 degrees away from the lake's shore, unless the lake was large and irregular enough that changing the coordinates would not have an appreciable effect. Lake names were changed to a common name if the observation stations were for the same lake (e.g. "LAKE SUWA (ARAKAWA)" and "LAKE SUWA (WEATHER STATION)" were renamed to "LAKE SUWA") or if distinct lakes had the same name (e.g. "TROUT LAKE," Ontario, was changed to "TROUT LAKE, ON" to distinguish it from "TROUT LAKE" in the United States). Finally, the data was converted to the PED format for freeze events and

Figure 1: Map of the lakes retained for analysis. The points in orange indicate lakes for which no data was available after 1995, while blue points indicate lakes with at least one datapoint after 1995. Five lakes (two Swiss, three Canadian) were excluded from the original dataset because they did not have data from after 1950.

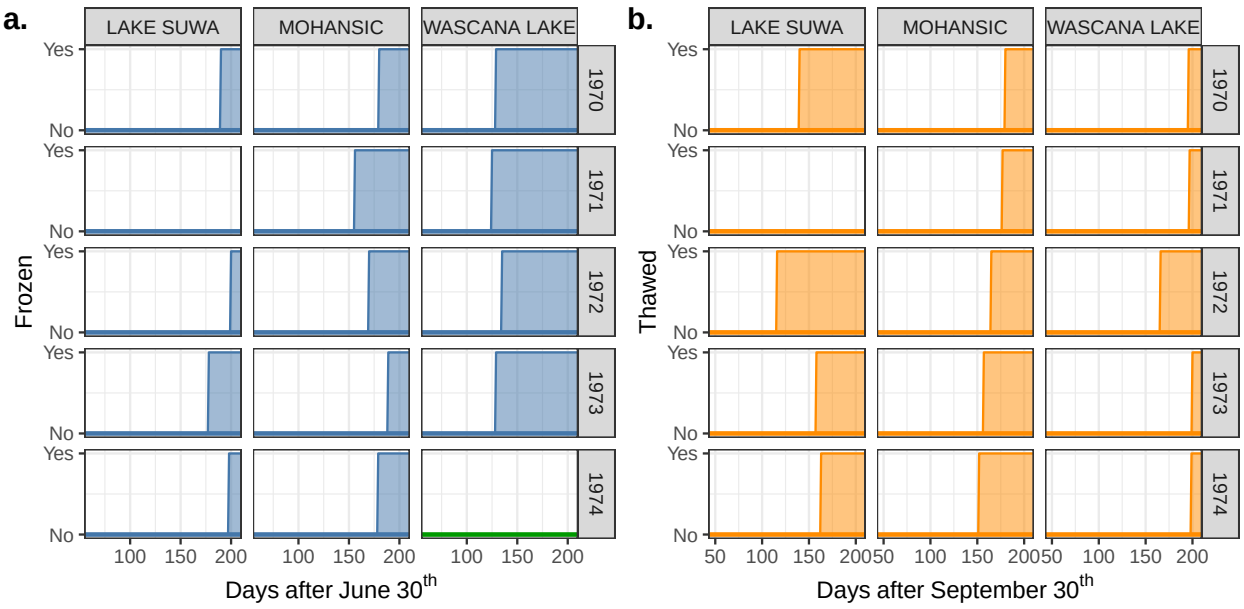


Figure 2: Freeze (a) and thaw (b) events for three lakes in the final dataset for years 1970-1974. Shaded areas indicate when the lake was frozen (a) or ice-free (b); the green baseline for Wascana lake in 1974 indicates that the lake froze, but the date of freezing is unknown. Note that lake Suwa did not freeze in 1971.

thaw events (given that the lake was frozen). The freeze PED had a structure of the form:

Table 1: Example of piecewise exponential data format

for the freezing dates of lakes in Eurasia.

tstart	tend	interval	offset	ped_status	lake	year	long	lat
0	72	(0,72]	4.28	0	SUWA	1950	138.08	36.05
72	81	(72,81]	2.20	0	SUWA	1950	138.08	36.05
81	84	(81,84]	1.10	0	SUWA	1950	138.08	36.05
84	88	(84,88]	1.39	0	SUWA	1950	138.08	36.05
88	90	(88,90]	0.69	0	SUWA	1950	138.08	36.05
90	91	(90,91]	0.00	0	SUWA	1950	138.08	36.05

The columns `tstart` and `tend` are the beginning and end of intervals $\left(\kappa_{j-1}, \kappa_j\right]$ (recorded as the number of days after June 30^{th}) for which the hazard function is estimated. Note that single-day intervals have an offset of $\log(1) = 0$, while longer intervals have non-zero offsets since the hazard of freezing within these periods is higher. The `ped_status` column indicates whether the lake is frozen (1) or not (0), the `year` column indicates the reference year, and `long` and `lat` indicate the lake's location. The thaw PEDs had a similar structure.

All of the data processing was performed in `R` [Versions 3.6.2-4.0.5; R Core Team (2021)], and the script is available in the GitHub repository under `data/freezing-dates.R` (see Appendix). The final dataset contains a total of 568 lakes and 628 distinct observation stations and is available in the GitHub repository as `data/lake-ice-data.rds` (see Appendix). Any lake which did not freeze in a given year was excluded from the thaw dataset for that year.

2.2 Software

All statistical analyses were performed in `R` version 3.6.2 and 4.0.5. HPAMs for freeze and thaw dates were fit using the `pammtools` package [versions 0.2.2-0.5.7; Bender & Scheipl (2018); Bender, Groll & Scheipl (2018)]. Plots were generated using `ggplot2` [versions 3.3.0-3.3.5; Wickham (2016)], `gratia` [versions 0.3.1-0.6.9300; Simpson (2021)], and `cowplot` [version 1.0.0-1.1.1 or higher; Wilke (2020)]. Maps were created using the packages: `sf` [versions *???* <= 1.0.2; Pebesma (2018)], `spData` [versions *???* <= 0.3.10; Bivand, Nowosad & Lovelace (2020)], and `raster` [versions *???* <= 3.4-13; Hijmans (2021)]. When necessary, figures use a palette with colors that are distinguishable by most color-vision-deficient people.

2.3 Model structure

2.3.1 Lake Mendota

The model included a smooth term `year` to count for long-term trends over the years and a smooth term of day of year (`tend`; the end of each observation period). The seasonal term (`s(tend)`) was allowed to vary over the years via a tensor product interaction term of `year` and `tend`.

```
pamm(ped_status ~
       s(tend, k = 5) +        # within-year effect
       s(year, k = 10) +       # between-year effect
       ti(tend, year, k = 5), # change in effect of DOY over years
     data = freeze)
```

2.3.2 Hierarchical models

The North American and Eurasian HPAMs for freeze and thaw dates accounted for the change in hazard between years (`year`) and within years (`tend`), as well as over space. Factor smooths of `tend` and `year` were included in the models to allow both temporal smooths to vary between lakes. Tensor product interaction terms (`ti`) were used to allow the effect of `tend` to vary over the years, and to allow the effects of `tend` and `year` to vary over space. `ti` terms allow the model to account for different rates of ice loss in different locations (Holland & Bitz, 2003):

```
pamm(ped_status ~
       s(tend, bs = 'cr', k = 10) +
       s(year, bs = 'cr', k = 10) +
       s(tend, lake, bs = 'fs', k = 10) +
       s(Year, lake, bs = 'fs', k = 10) +
       s(long, lat, bs = 'ds', k = 20) +
       ti(tend, Year, bs = 'cr', k = c(5, 5)) +
```

```
    ti(tend, long, lat, bs = c('cr', 'ds'), d = c(1, 2), k = c(5, 5)) +
    ti(Year, long, lat, bs = c('cr', 'ds'), d = c(1, 2), k = c(5, 5)),
  data = freeze.na,
  method = 'fREML', # fast restricted marginal likelihood
  engine = 'bam',   # use mgcv::bam() for faster fitting
  discrete = TRUE)  # discretize the covariates for faster fitting
```

The `bs` arguments indicate which basis type each smooth uses. Cubic regression splines (`cr`) are fast-fitting, one-dimensional splines composed of cubic polynomials. Factor smooths bases (`fs`) fit a penalized (thin-plate) smooth for each `lake` factor, such that all smooths have a common smoothness parameter (Pedersen *et al.*, 2019). Duchon splines (`ds`) are two-dimensional splines that avoid excessive spatial extrapolation (i.e. they are well-behaved as they move away from the support of the data, see Duchon, 1977).

The `k` argument sets the maximum complexity of a smooth term, such that the maximum number of degrees of freedom is $k - 1$. Note that `k = c(a, b)` in the `ti` terms indicates that the maximum effective degrees of freedom is $(a - 1)(b - 1)$.

Finally, the `method` argument indicates that the smoothness parameter should be estimated using fast REstricted Marginal Likelihood, while `engine = 'bam'` indicates that `mgcv::bam()` should be used and `discrete = TRUE` discretizes the model's covariates to decrease computational cost and fitting time (Wood, 2011). The structure of the thaw model is essentially identical, with the exception that the response is 0 if the lake is (still) frozen and 1 if the lake is ice-free, given that it was previously frozen.

(Since the `pamm` function from the `pammtools` package is a wrapper function for the `gam` and `bam` functions from the `mgcv` package, one could also use `mgcv::bam` or `mgcv::gam` instead of `pammtools::pamm`, but in that case `family = poisson()` and the `offset` argument need to be specified.)

3 Results

3.1 Lake Mendota

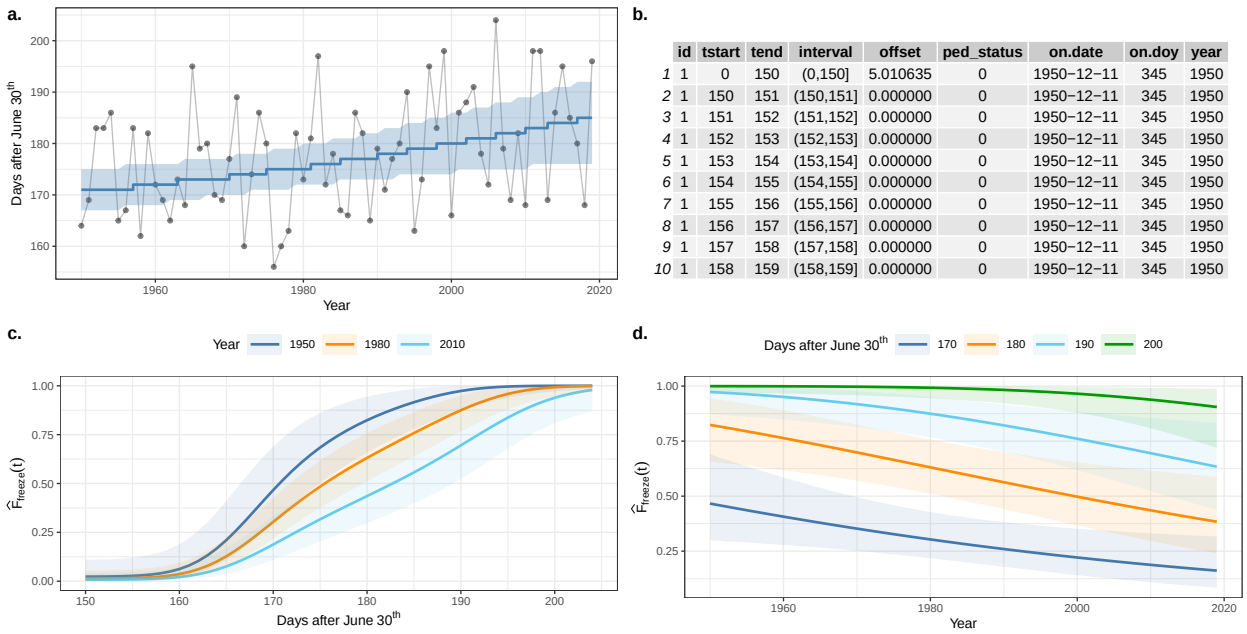
The lake's estimated average freeze date in 2020 was 14 days later than the estimated freeze date in 1950 (Figures 3a, c). Although freeze dates varied by as much as a 32 days between consecutive years, the model explained 46% of the deviance, which indicates that a large portion of the variance in freezing dates is related to effects strongly correlated with time, such as the recent increase in temperature ((**IPCC-2021?**), (**mendota-specific-ref?**)?).

The rightward shift of the estimated cumulative probability function, $\widehat{F}_{freeze}(t)$, over the years (Figure 3c) indicates that, on average, the lake has been freezing later in recent years, while the flattening of the function suggests that the lake has been freezing later and more variably, since $\widehat{F}_{freeze}(t)$ doesn't increase as quickly in 2010 as it did in 1950.

The hazard of freezing on a given day of year decreases nonlinearly over the years (Figure 3d). (Consequently, the average freezing day increases nonlinearly over the years, with an accelerating trend in the later years (Figure 3a), but this may not be as visible with the step function.)

3.2 Global lake ice datasets

The results from the HPAMs fit to the North American and Eurasian datasets are shown in Figures 4 through 8. Since 1950, the average cumulative probability of freezing decreased in both continents, while the cumulative probability of thawing increased. On average, Eurasian lakes in 2010 froze 17 days later and thawed 3 days earlier than in 1950, while North American lakes in 2010 froze 25 days later and thawed 10 days earlier than in 1950 (Figure 4). Most of the change in $\widehat{F}_{freeze}(t)$ and $\widehat{F}_{thaw}(t)$ occurred after 1995, when data was available for only 36% of the lakes in the dataset, and the great majority of them were in the Great Lakes Area and Northern Europe (Figure 1). With respect to the average 1950



	id	tstart	tend	interval	offset	ped_status	on.date	on.doy	year
1	1	0	150	(0,150]	5.010635	0	1950−12−11	345	1950
2	1	150	151	(150,151]	0.000000	0	1950−12−11	345	1950
3	1	151	152	(151,152]	0.000000	0	1950−12−11	345	1950
4	1	152	153	(152,153]	0.000000	0	1950−12−11	345	1950
5	1	153	154	(153,154]	0.000000	0	1950−12−11	345	1950
6	1	154	155	(154,155]	0.000000	0	1950−12−11	345	1950
7	1	155	156	(155,156]	0.000000	0	1950−12−11	345	1950
8	1	156	157	(156,157]	0.000000	0	1950−12−11	345	1950
9	1	157	158	(157,158]	0.000000	0	1950−12−11	345	1950
10	1	158	159	(158,159]	0.000000	0	1950−12−11	345	1950

Figure 3: Results from the PAM of the freeze dates of Lake Mendota. **a.** Freeze dates of Lake Mendota, with estimated mean freeze dates and 95% credible intervals of the mean. The estimated mean day is a step function because observations were assumed to be daily and discrete. **b.** The Lake Mendota dataset in PED format. From a survival analysis perspective, the `id` column can be viewed as a "patient" indicator; it is an alternative representation of the `year` column. The `tstart` and `tend` columns indicate the beginning and end of the observation periods (`interval`), respectively, as the number of days after June 30^{th}. The `offset` column accounts for differences in length between observation `intervals`, and is equivalent to the log-transformed `interval` length (e.g. $\log(150-0) = 5.01$, $\log(151-150) = \log(1) = 0$). The column indicates whether the event occurred (1) or not (0). The `on.date` indicates the date on which the lake froze and is `NA` if the lake did not freeze. Finally, `year` indicates the the year starting on June 30^{th} and `on.doy` indicates the number of days after June 30^{th}. **c.** Estimated cumulative probability of Lake Mendota being frozen on a given day of year for different years. **d.** Estimated probability of Lake Mendota being frozen on a given set of days (December 17^{th} to January 16^{th}) over the years.

dates, in 1995 lakes in North America froze 6 days later and thawed 1 day later, while lakes in Eurasia froze and thawed a day later.

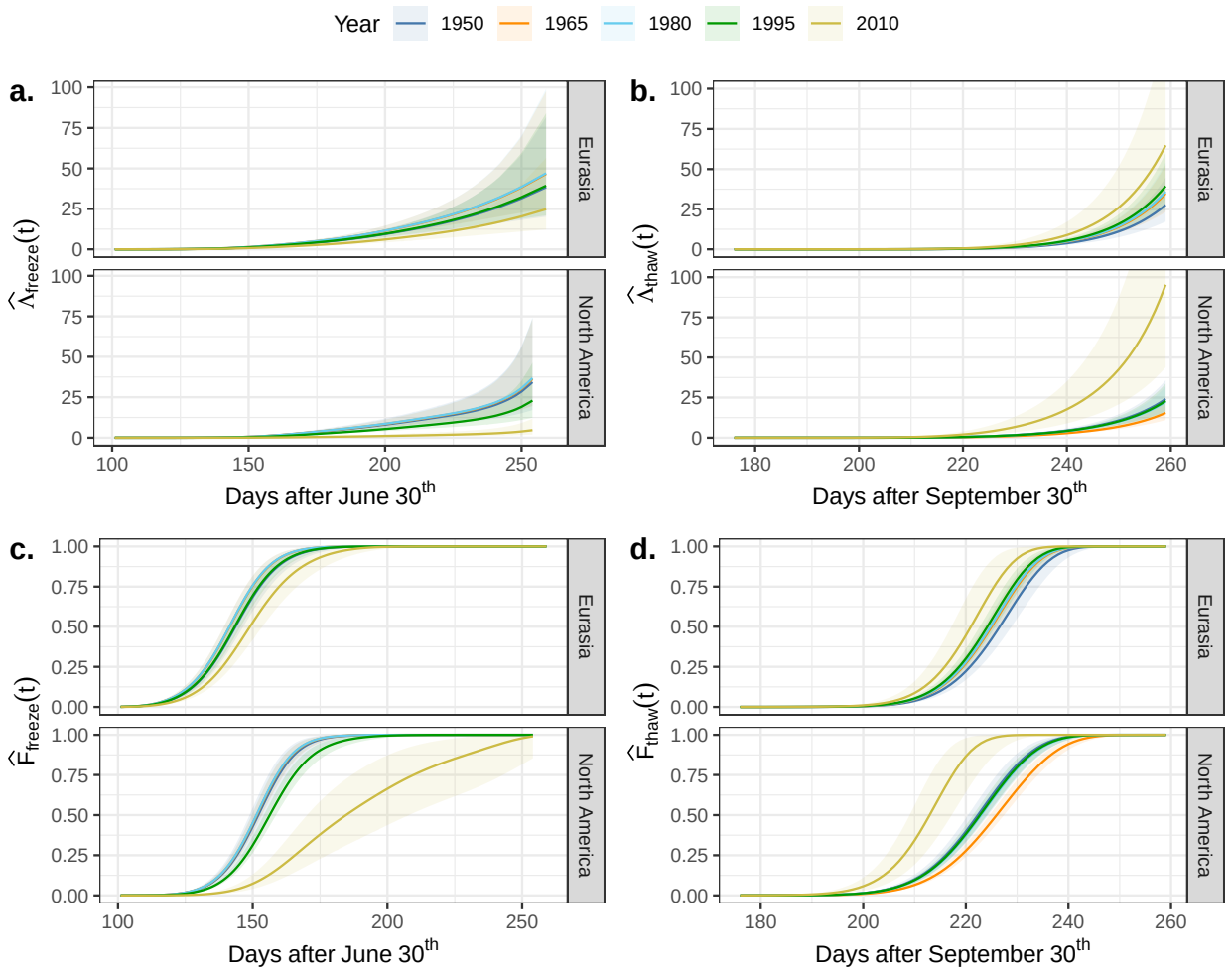


Figure 4: Estimated cumulative hazard (a, b) and probability (c, d) with 89% credible intervals of lakes being frozen (left) or thawed (right) while assuming an average effect of geographic location.

The estimated factor smooth interaction terms for `year` of the four models (Figure 5a) indicates a strong common trend in $\hat{\lambda}(t)$ over the years with little deviation from the global term (`s(year)`) after accounting for the effects of space and `tend`. In contrast, the high spread of the `tend` factor smooths (Figure 5b) indicates that there are strong drivers of seasonal variation that are not accounted for by the models.

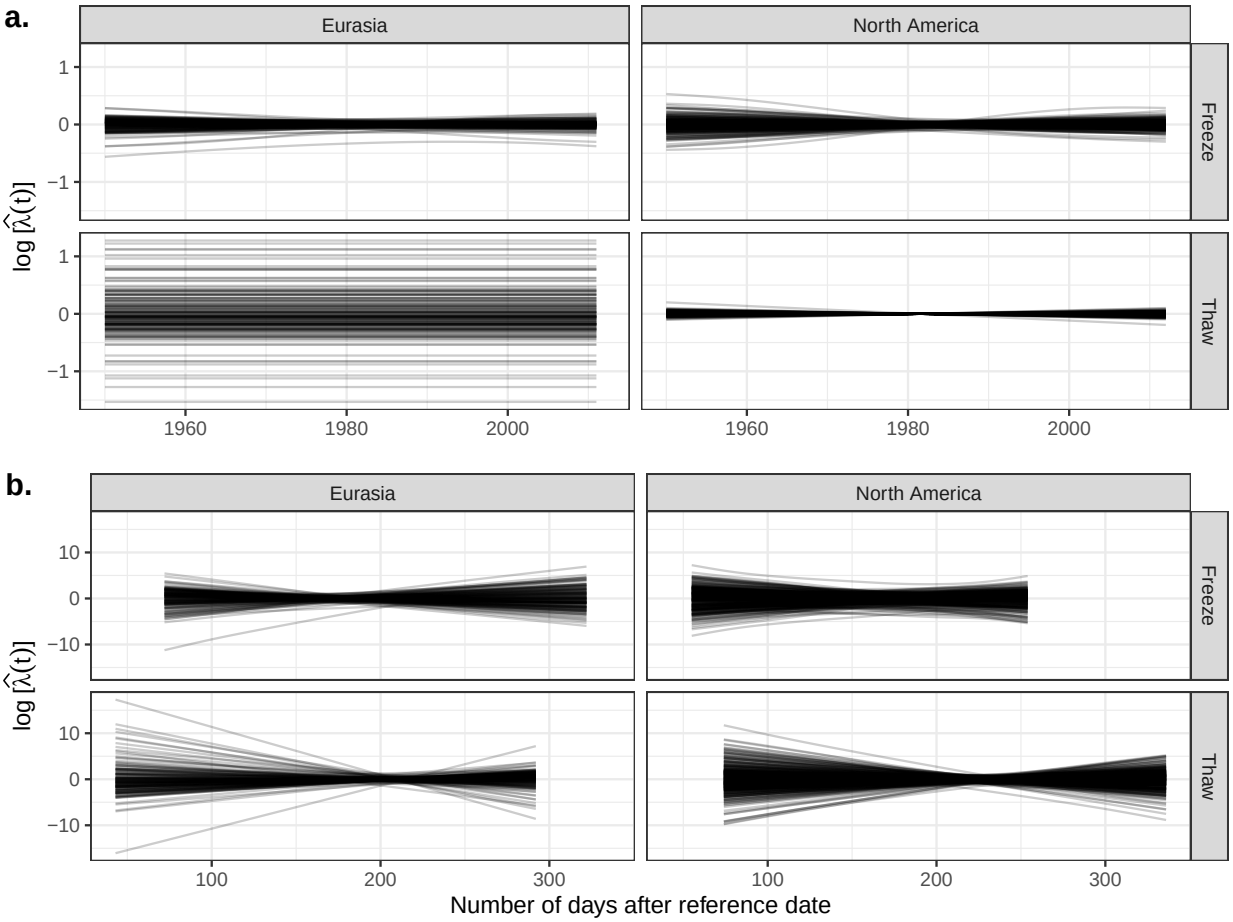


Figure 5: Factor smooths (`fs`) of `year` (a) and `tend` (b) from the HPAMs for the hazard of freezing and thawing in North America and Eurasia. The y axis indicates how much the hazard of each lake deviates from the mean trend, assuming Gaussian random effects of `lake`.

When accounting for the spatial effects in the predictions, the trends are not spatially uniform or monotonic. Overall, European and Western and Northern North American lakes had the greatest shift to later freeze dates, and the majority of the change occurred after 1995. With respect to the average freezing dates in 1950, lakes in 1995 froze 2-7 days later in North America, 1-3 days earlier in Europe, and 3-5 days later in Asia and Russia (Figure 6a). In contrast, changes in 2010 in North America and Europe are much larger and more variable, relative to the average 1950 dates. North American lakes are estimated to have frozen 12-120 days later (IQR = 111 days). European lakes (excluding Russian lakes) are estimated to have frozen 19-31 days later with an average of 25 days later (IQR = 12 days), but Russian

and Asian lakes are not expected to have frozen much earlier or later (mean: - 3 days, IQR = 6 days).

The patterns estimated by the thaw models are similar to the trends in freeze dates but more spatio-temporally uniform (and in the opposite direction). Relative to 1950, lakes in eastern Canada thawed 3-7 days later in 1995 but 4-16 days earlier in 2010. In the rest of Canada, lakes are estimated to have thawed 3-12 earlier in 1995 and 1-2 months earlier in 2010 (median = -38). In contrast, lakes in the United States thawed 1-4 days later in 1995 and 6-19 days later in 2010. Eurasian lakes thawed 0-8 days earlier in 1995 and 4-11 days earlier in 2010, with the exception of eastern Asian lakes, which thawed 2-8 days later in 1995 and 3-12 days later in 2010.

4 Discussion

4.1 Lake Mendota

The model's large proportion of explained deviance (0.462) indicates that although there are unaccounted factors which contribute in the stochasticity of the lake's freezing process, the simple model presented here is able to explain a substantial portion of the variation. Terms such as average daily wind speed or average daily temperature could be included in the model, but the model should ultimately be designed according to the questions the analyst(s) is (are) interested in answering. Any terms which are strongly correlated or have causal connections (e.g. day of year and average daily temperature) should not be included in the same model, since they may result in coefficient inflation or lead one to conflate effects. In our case, the simple model presented here explains sufficient deviance to provide acceptable estimates and predictions with small amounts of data. The inclusion of wind disturbance would likely increase the proportion of explained deviance, but it is beyond the scope of this peper, and such data may not be available for large amounts of lakes, as in the

case of the HPAMs fit here.

The recent nonlinear shift to later freezing dates indicates that a model which assumes linearity (such as a parametric linear model or the nonparametric Kendall rank correlation coefficient, see Kendall, 1938) would fail to capture the acceleration in lake ice loss, and thus over-estimate change in early years and under-estimate change in later years. Using a model which assumes linearity would thus be particularly problematic if one is interested in extrapolating trends forward in time. Such a method would also fail to capture the increase in variance, unless it was a location scale-model, but such a model would require more data and have larger uncertainty.

With this paper, we hope to demonstrate the importance of choosing a model which is adequate for the questions one is interested in answering. Although using a model with appropriate assumptions can require more expertise, it also simplifies the fitting process and allows the analyst(s) to draw complex conclusions from the model. In the case of the Lake Mendota model, the flattening of $\widehat{F}_{freeze}(t)$ indicates that the change is more complex than the simplistic shift to later dates that a simple linear test would support. With the PAM presented here, one is able to create more complex estimates, such as the daily probability of being frozen on a given day, without having to resort to post-hoc methods which may not be statistically sound.

4.2 Global lake ice datasets

The widespread shift to later freezing and earlier thawing of lakes has greatly changed the lives of many many people, including those of Indigenous People in Northern Canada and Alaska. As a consequence, many First Nations report a reduction in food and energy security as well as major changes to traditional activities (Golden, Audet & Smith, 2015). Many of these communities rely on (lake) ice roads as a means of travel and transportation between communities and to obtain food and resources. The loss of lake ice has also resulted in deep

spiritual and emotional damage, since many of Indigenous traditional activities and cultural views are deeply connected with the formation of lake ice, particularly "blue ice" (Golden, Audet & Smith, 2015). The disappearance of blue ice has been suggested as an indicator of the rate of climate change (Orheim & Lucchitta, 1990; Walsh *et al.*, 1998), and the absence of blue ice in lakes has been linked to a decrease in strength and reliability of ice roads (Golden, Audet & Smith, 2015). In addition, youth, particularly Indigenous youth, have expressed stress due to changes in climate and loss of ice (Petrasek MacDonald *et al.*, 2013).

Lake ice loss has also had a strong impact on the lives of many Eurasian people. Religious practices such as the transferring of a statue of John the Apostle across Lake Constance (central Europe) or the Shinto purification rituals on Lake Suwa (Japan) are unlikely to occur in the coming years (Knoll *et al.*, 2019). Changes in lake ice phenology have also caused severe damage to the economies of many nations. For instance, ice fishing on Lake Peipsi, Estonia, has attracted as many as 3000 anglers on a single weekend, and it is an important source of income and food for many people (Orru *et al.*, 2014). However, the amount of fish that is caught each year can vary by as much as a factor of 10 due to variable ice conditions (Orru *et al.*, 2014). In Northern Europe, the unpredictability of lake ice has lead to the cancellation of many skating events and competitions, including the Swedish "Viking ride," a long-distance skating race that was held annually from 1999 until 2018 (Knoll *et al.*, 2019).

Despite the low data availability after 1995, the estimates produced by the models agree with recently published results (e.g. Brown & Duguay, 2011). However, the hierarchical and nonlinear time-to-event approach allowed for finer spatio-temporal detail. The large number of lakes in the dataset and their wide spatial range allowed informed estimates of the daily hazard of freezing or thawing for the majority of the two continents while

providing a statistically sound measure of uncertainty and variance. Although the data scarcity after 1995 resulted in a substantial increase in the recent estimates' uncertainty, the low degrees of freedom (`k`) in the spatial terms ensured that predictions would be sufficiently generalizable during the years with little data. While the estimated loss of ice in Western and Northern Canada may seem excessive, the 89% credible intervals still contain the estimates of recent peer-reviewed studies (e.g. Brown & Duguay, 2011). Additional data post 1995 would produce more informed (and less uncertain) estimates, but the current estimates are reasonable, and they show the accelerating trends estimated by others [Warne *et al.* (2020); ***(additional-refs?)***].

The spatial term `s(long, lat)` decreased the potential bias from the non-random and opportunistic nature of the sample of lakes, while the Duchon splines remained sufficiently constrained outside the spatial range of the data. Allowing the long-term and seasonal trends to vary over space with the inclusion of spatio-temporal tensor product interaction terms (`ti(tend, long, lat)`, and `ti(year, long, lat)`), the models were able to estimate the changes in hazard over time without assuming trends that with the same magnitude, direction, and slope in all regions. Instead, the spatio-temporal `ti()` terms allowed the estimated loss to increase at faster rates in more northern regions (e.g. North-Western Canada, Scandinavia) and decrease in other areas (e.g. Eastern United States and Eastern Asia). Allowing `s(year)` and `s(tend)` to vary over space ensured that the models would be able to produce realistic estimates (with measures of uncertainty) even in areas where data was scarce or unavailable after 1995.

Without the spatial terms, the models would likely have produced substantially different results. Without any information about the (marginal and partial) effects of space, the models would only estimate the average change in hazard over time and would lack any information on the spatial distribution of the lakes and the effects of latitude and continentality on climate. While the spatial effect would be partially accounted for by the random effects in the `fs` terms of `year` and `tend`, such terms are not appropriate for estimating the full effect of

location, as they are best for (relatively small) amounts of stochastic variation, rather than large, deterministic, and spatially autocorrelated effects. Still, a simple HPAM with smooths of `tend` and `year`, i.e.

$$\log\left[\hat{\lambda}(t)\right] = f_1(year) + f_2(tend) + \epsilon, \tag{6}$$

would likely still produce a better estimate of the common temporal trends than an ensamble of individual models. This is because the hierarchical model would produce a single posterior distribution for `s(tend)` and `s(year)` using a single common likelihood function for all lakes. This way, the model would contain information about the variation between lakes due to the spatial location of the lakes and other characteristics such as lake morphology. The estimated coefficients would also be conditional on the entirety of the dataset rather than a single lake, so they would be more statistically meaningful than the average of many individual models which lack any information on differences between lakes. Averaging the coefficients of siglelake models would thus fail to provide any information on the spatial heterogeneity of lake ice loss, and it would not be possible to reasonably extrapolate for lakes which are not in the dataset.

Averaging linear coefficients from multiple linear models (e.g. Warne *et al.*, 2020) is particularly inefficient, since although it would be mathematically simple, the estimates are unable to account for any nonlinearity in the trends, such as the accelerating warming of lakes (Velicogna, 2009; Zhong *et al.*, 2016).

Hierarchical Generalized Additive Models (HGAMs) and, more specifically, Hierarchical Piece-wise exponential Additive Models (HPAMs) are particularly effective and appropriate models for analyzing lake ice phenology data at a regional and global scale, since they can account for (1) nonlinear effects, (2) multiple lakes in a single model, (3) the spatial relationship between lakes, and (4) interactions between smooth terms via tensor product interaction terms. Ultimately, however, the accuracy and precision of any model depend on

data availability.

An increase in data, particularly in periods and regions with low data availability or with complex and accelerating trends, would reduce the uncertainty in the estimates and the potential bias. However, the hierarchical approach used here allows the models to produce reasonable and well-informed predictions even with fragmented and short records. Since the models account for the spatial relationship and variance between lakes, multiple short time series (e.g. 2-5 years) from low-data areas would likely have a more appreciable effect on the models than few but long time series. By incorporating the spatial relationship between lakes into the models, spatial and temporal gaps in the data can be filled using information from nearby lakes. Thus, even small data contributions can greatly improve hierarchical models such as those presented here.

With the recent increase in open and "big data," it is imperative that large datasets such as the GLRIPD be analyzed using hierarchical methods. Failing to use hierarchical methods ultimately forfeits a large portion of the value that comes with such large, and spatially and temporally diverse datasets.

Model accuracy and precision could also be improved by adding predictors to the models. Predictors with a strong effect, such as lake elevation would account for the earler freezing and later thawing of lakes at higher altitudes, and it can also offset the underreppresentation of lakes at low altitudes, which do not freeze often. Although lake altitude was present in the GLRIPD, it was unavailable for a large amount of the lakes, so it was not included in the models. We hope to include lake elevation in future versions of the models, whether the missing values are entered manually or are estimated using functions such as the `GNgtopo30()` function from the `geonames` package [version 0.999; Rowlingson (2019)], which estimates altitude using the `GTOPO30` global digital elevation model (Center, 2017).

Other variables such as lake depth and surface area would also help explain a portion of the unexplained seasonal variation between lakes Magee & Wu (2017). Currently, these effects

and other stochastic effects (e.g. wind disturbance and snow cover) are accounted for as random between-lake variation by the factor smooths of `tend` and `year`, since they likely affect the seasonal trends in freezing and thawing hazard, and shallower and larger lakes appear to be more resilient to loss of ice (Warne *et al.*, 2020).

The explanatory power of the models could also be increased by accounting for the exposure to above-zero or sub-zero temperatures. The air temperature term could also be lagged or smoothed to account for delays in the effect. However, additional care should be taken to ensure the collinearity of air temperature with `year`, `tend`, and space is not excessive. Ultimately, the model should be designed to address the question of interest. While the addition of a smooth of air temperature would likely increase the predictive power (and R^2) of the models, the models may no longer be appropriate for estimating whether lake ice cover is changing over time, since they estimate the changes in lake ice over time once the effect of air temperature is accounted for. If the main interest is to estimate the loss of lake ice over time, then it may be counterproductive to include time-varying and cyclical variables in the model, such as air temperature and wind disturbance. Instead, the models should only have smooths for `year`, `tend`, and variables that can be assumed to be uncorrelated with time during the period of analysis, such as lake depth, surface area, and geographic location.

5 Conclusion

There is currently little understanding about the large-scale effects of the loss of lake ice on biological systems, but the damage to many people's lifestyle, traditions and economy are evident. Although some work has been done on the recent changes in lake ice phenology, the subject remains understudied, and there is a need for more statistically appropriate and efficient analysis. To appropriately estimate large-scale changes in lake ice phenology, it is necessary to use smooth, hierarchical, and flexible models which allow trends to vary spatio-temporally, such as the models presented in this paper. The hierarchical structure

of the models allowed us to estimate complex and heterogenous spatio-temporal changes in lake ice, even when little to no data were available. The results presented here demonstrate a widespread and accelerating but heterogeneous change in lakes' freeze and thaw dates. Many large areas exhibit a shift towards later freezing and earlier thawing which has caused substantial cultural, spiritual, and economic loss to different People worldwide.

Code and data availability

All code and data used in this project can be found in its GitHub repository at https://github.com/simpson-lab/lake-ice-event-history-honours. The repository contains separate folders for the data, code, custom functions, and plots used in the project.

References

Arakawa H. (1954). Fujiwhara on five centuries of freezing dates of Lake Suwa in the Central Japan. *Archiv für Meteorologie, Geophysik und Bioklimatologie Serie B* **6**, 152–166. https://doi.org/10.1007/BF02246747

Bender A., Groll A. & Scheipl F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling* **18**, 299–321. https://doi.org/10.1177/1471082X17748083

Bender A. & Scheipl F. (2018). Pammtools\: Piece-wise exponential Additive Mixed Modeling tools. *arXiv:1806.01042 [stat]*

Benson B. (2002). Global Lake and River Ice Phenology Database. https://doi.org/10.7265/n5w66hp8

Bivand R., Nowosad J. & Lovelace R. (2020). *spData: Datasets for Spatial Analysis*.

Brown L.C. & Duguay C.R. (2011). The fate of lake ice in the North American Arctic. *The Cryosphere* **5**, 869–892. https://doi.org/10.5194/tc-5-869-2011

Center E.R.O.A.S. (EROS) (2017). Global 30 Arc-Second Elevation (GTOPO30)

Duchon J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: *Constructive Theory of Functions of Several Variables*. (Eds A. Dold, B. Eckmann, W. Schempp & K. Zeller), pp. 85–100. Springer Berlin Heidelberg, Berlin, Heidelberg.

Golden D.M., Audet C. & Smith M.A.(Peggy). (2015). "Blue-ice": Framing climate change and reframing climate change adaptation from the indigenous peoples' perspective in the northern boreal forest of Ontario, Canada. *Climate and Development* **7**, 401–413. https://doi.org/10.1080/17565529.2014.966048

Hampton S.E., Galloway A.W.E., Powers S.M., Ozersky T., Woo K.H., Batt R.D., *et al.* (2017). Ecology under lake ice. *Ecology Letters* **20**, 98–111. https://doi.org/10.1111/ele.12699

Hastie T. & Tibshirani R. (1999). *Generalized additive models*. Chapman & Hall/CRC, Boca Raton, Fla.

Hastie T. & Tibshirani R. (1986). Generalized Additive Models. *Statistical Science* **1**, 297–310. https://doi.org/10.1214/ss/1177013604

Hijmans R.J. (2021). *Raster: Geographic Data Analysis and Modeling*.

Holland M.M. & Bitz C.M. (2003). Polar amplification of climate change in coupled models. *Climate Dynamics* **21**, 221–232. https://doi.org/10.1007/s00382-003-0332-6

Kendall M.G. (1938). A NEW MEASURE OF RANK CORRELATION. *Biometrika* **30**, 81–93. https://doi.org/10.1093/biomet/30.1-2.81

Kleinbaum D.G. & Klein M. (2012). *Survival analysis: A self-learning text*, 3rd ed. Springer, New York.

Knoll L.B., Sharma S., Denfeld B.A., Flaim G., Hori Y., Magnuson J.J., *et al.* (2019). Consequences of lake and river ice loss on cultural ecosystem services. *Limnology and Oceanography Letters* **4**, 119–131. https://doi.org/10.1002/lol2.10116

Lopez L.S., Hewitt B.A. & Sharma S. (2019). Reaching a breaking point: How is climate change influencing the timing of ice breakup in lakes across the northern hemisphere? *Limnology and Oceanography* **0**. https://doi.org/10.1002/lno.11239

Magee M.R. & Wu C.H. (2017). Effects of changing climate on ice cover in three morphometrically different lakes: Climate change on ice cover in three morphometrically different lakes. *Hydrological Processes* **31**, 308–323. https://doi.org/10.1002/hyp.10996

Magnuson J.J., Carpenter S.R. & Stanley E.H. (2021). North Temperate Lakes LTER: Ice Duration - Madison Lakes Area 1853 - current

Orheim O. & Lucchitta B. (1990). Investigating Climate Change by Digital Analysis of Blue Ice Extent on Satellite Images of Antarctica. *Annals of Glaciology* **14**, 211–215. https://doi.org/10.3189/S0260305500008600

Orru K., Kangur K., Kangur P., Ginter K. & Kangur A. (2014). Recreational ice fishing on the large Lake Peipsi: Socioeconomic importance, variability of ice-cover period, and possible implications for fish stocks. *Estonian Journal of Ecology* **63**, 282. https://doi.org/10.3176/eco.2014.4.06

Pebesma E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**, 439–446. https://doi.org/10.32614/RJ-2018-009

Pedersen E.J., Miller D.L., Simpson G.L. & Ross N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* **7**, e6876. https://doi.org/10.7717/peerj.6876

Petrasek MacDonald J., Harper S.L., Cunsolo Willox A., Edge V.L. & Rigolet Inuit Community Government (2013). A necessary voice: Climate change and lived experiences of youth in Rigolet, Nunatsiavut, Canada. *Global Environmental Change* **23**, 360–371. https://doi.org/10.1016/j.gloenvcha.2012.07.010

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rowlingson B. (2019). Geonames: Interface to the "Geonames" Spatial Query Web Service

Sharma S., Blagrave K., Magnuson J.J., O'Reilly C.M., Oliver S., Batt R.D., *et al.* (2019). Widespread loss of lake ice around the Northern Hemisphere in a warming world. *Nature Climate Change* **9**, 227–231. https://doi.org/10.1038/s41558-018-0393-5

Sharma S., Magnuson J.J., Batt R.D., Winslow L.A., Korhonen J. & Aono Y. (2016). Direct observations of ice seasonality reveal changes in climate over the past 320–570 years. *Scientific Reports* **6**, 25061

Shuter B.J., Minns C.K. & Fung S.R. (2013). Empirical models for forecasting changes in the phenology of ice cover for Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences* **70**, 982–991. https://doi.org/10.1139/cjfas-2012-0437

Simpson G.L. (2021). Gratia: Graceful 'ggplot'-Based Graphics and Other Functions for GAMs Fitted Using 'mgcv'

Velicogna I. (2009). Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE. *Geophysical Research Letters* **36**, L19503. https://doi.org/10.1029/2009GL040222

Walsh S.E., Vavrus S.J., Foley J.A., Fisher V.A., Wynne R.H. & Lenters J.D. (1998). Global patterns of lake ice phenology and climate: Model simulations and observations. *Journal of Geophysical Research: Atmospheres* **103**, 28825–28837. https://doi.org/10.1029/98JD02275

Warne C.P.K., McCann K.S., Rooney N., Cazelles K. & Guzzo M.M. (2020). Geography and Morphology Affect the Ice Duration Dynamics of Northern Hemisphere Lakes Worldwide. *Geophysical Research Letters* **47**. https://doi.org/10.1029/2020GL087953

Wickham H. (2016). *ggplot2: Elegant graphics for data analysis*, Second edition. Springer, Cham.

Wilke C.O. (2020). Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'

Wood S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood S.N. (2017). *Generalized additive models: An introduction with R*, Second edition. CRC Press/Taylor & Francis Group, Boca Raton.

Woolway R.I., Kraemer B.M., Lenters J.D., Merchant C.J., O'Reilly C.M. & Sharma S. (2020). Global lake responses to climate change. *Nature Reviews Earth & Environment*. https://doi.org/10.1038/s43017-020-0067-5

Zhong Y., Notaro M., Vavrus S.J. & Foster M.J. (2016). Recent accelerated warming of the Laurentian Great Lakes: Physical drivers: Physical drivers of Great Lakes' warming. *Limnology and Oceanography* **61**, 1762–1786. https://doi.org/10.1002/lno.10331

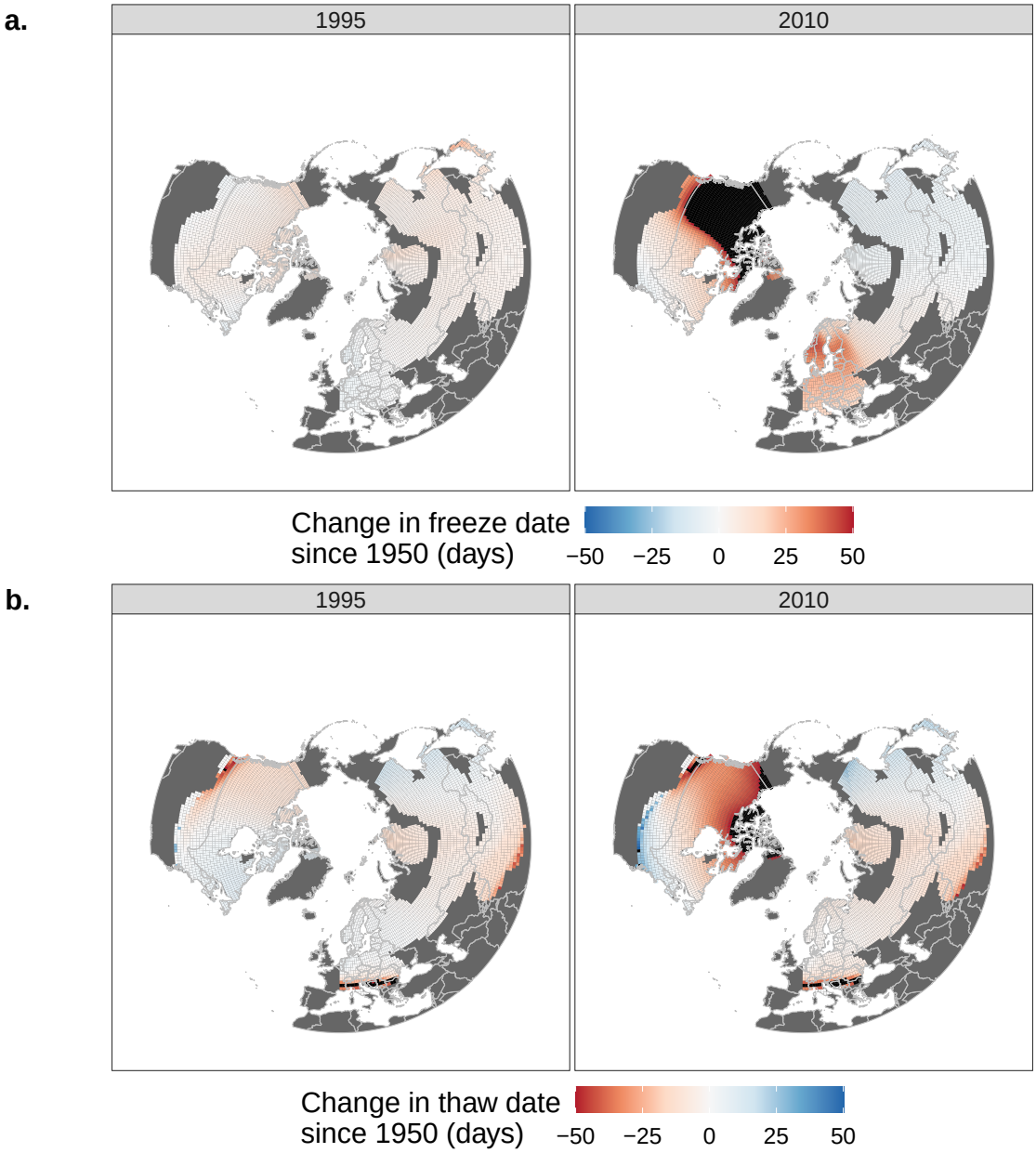


Figure 6: Estimated change in the average freeze (a) and thaw (b) dates relative to 1950. Areas in black were estimated to have an absolute change greater than 50 days. 1995 was chosen as a midpoint between the beginning and the end of the record because a large portion of the records ended in 1994 and 1995.

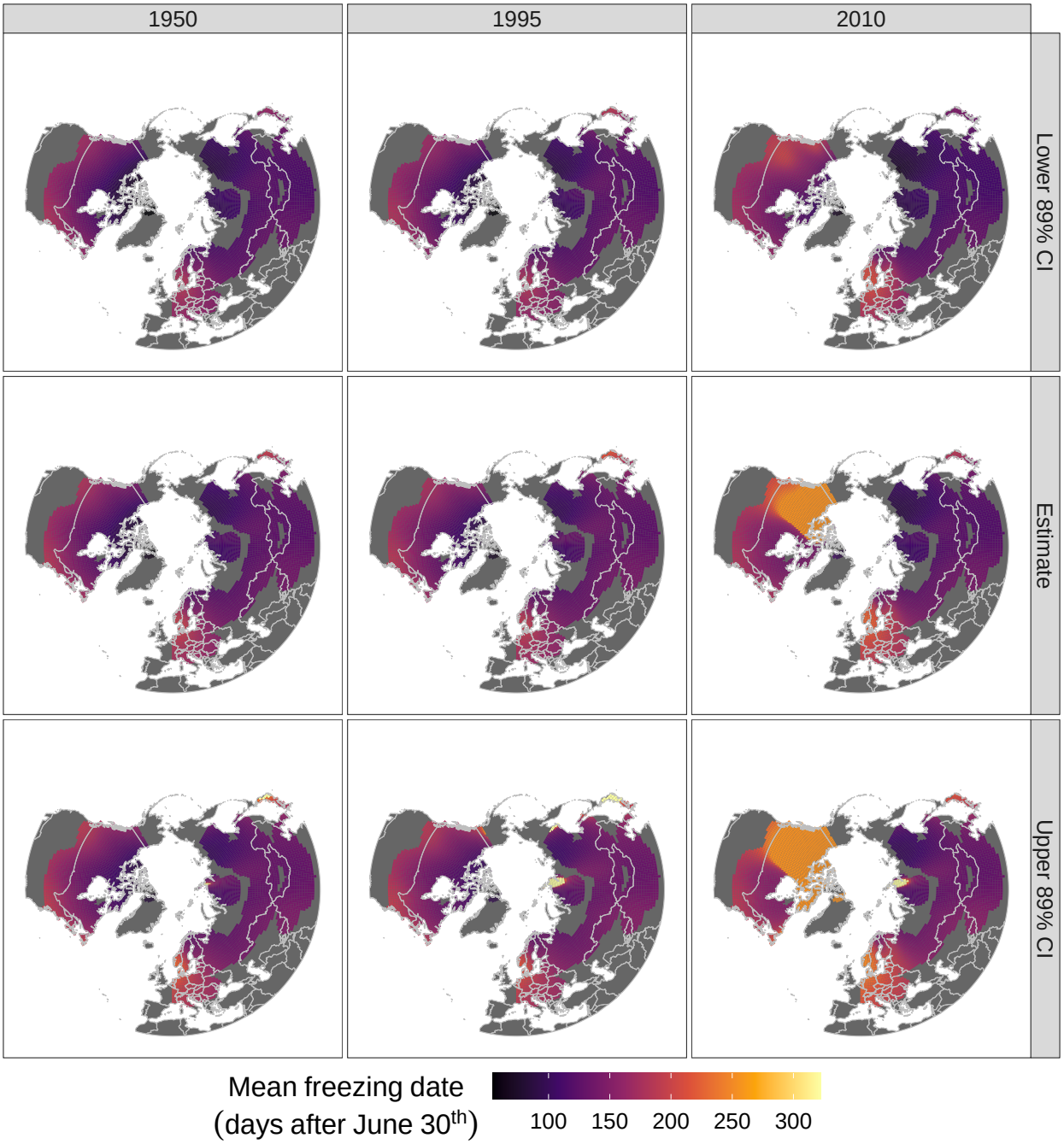


Figure 7: Estimated freeze dates and relative 89% credible intervals. 1995 was chosen as a midpoint between the beginning and the end of the record because a large portion of the records ended in 1994.

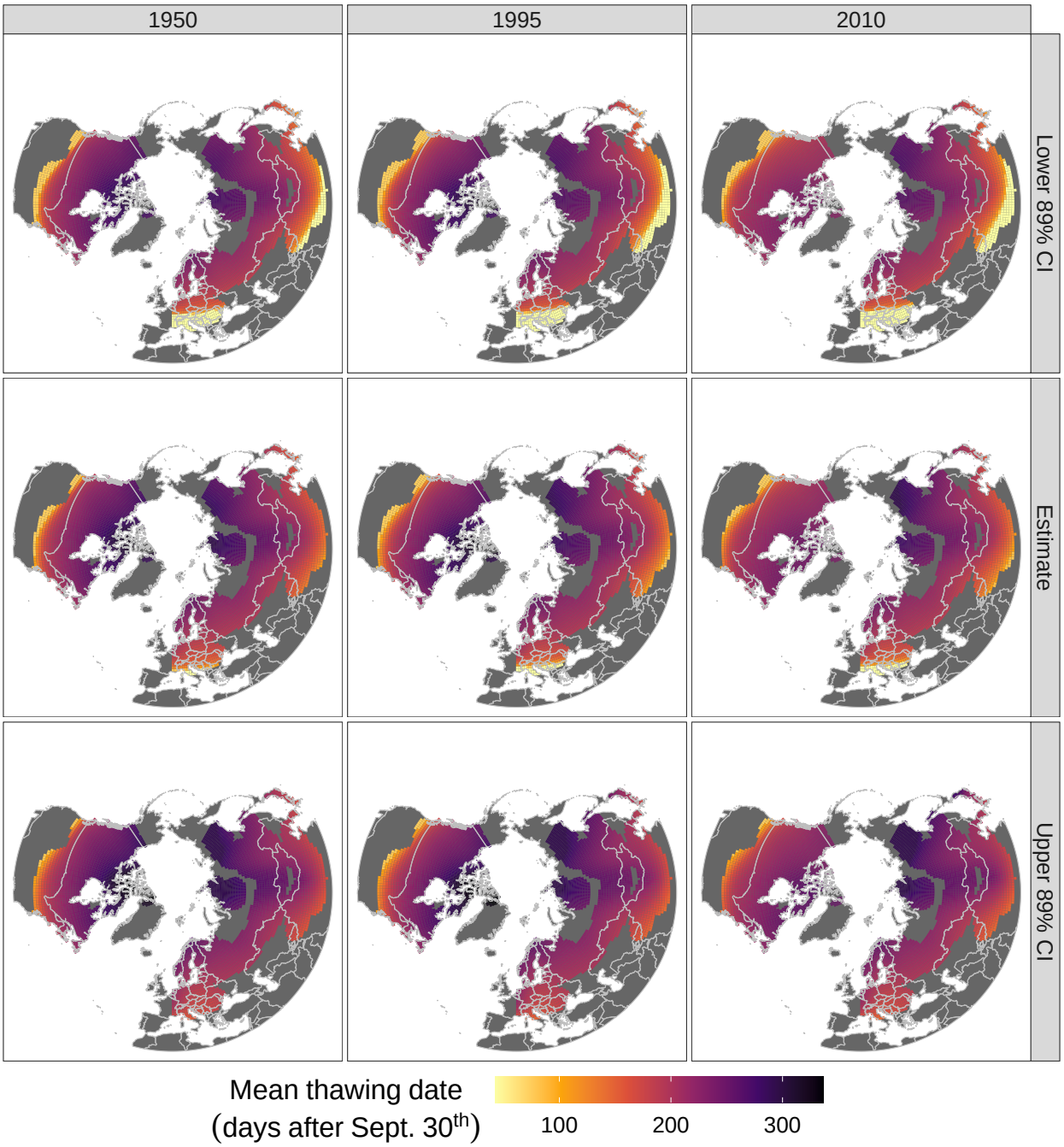


Figure 8: Estimated thaw dates and relative 89% credible intervals. 1995 was chosen as a midpoint between the beginning and the end of the record because a large portion of the records ended in 1994.

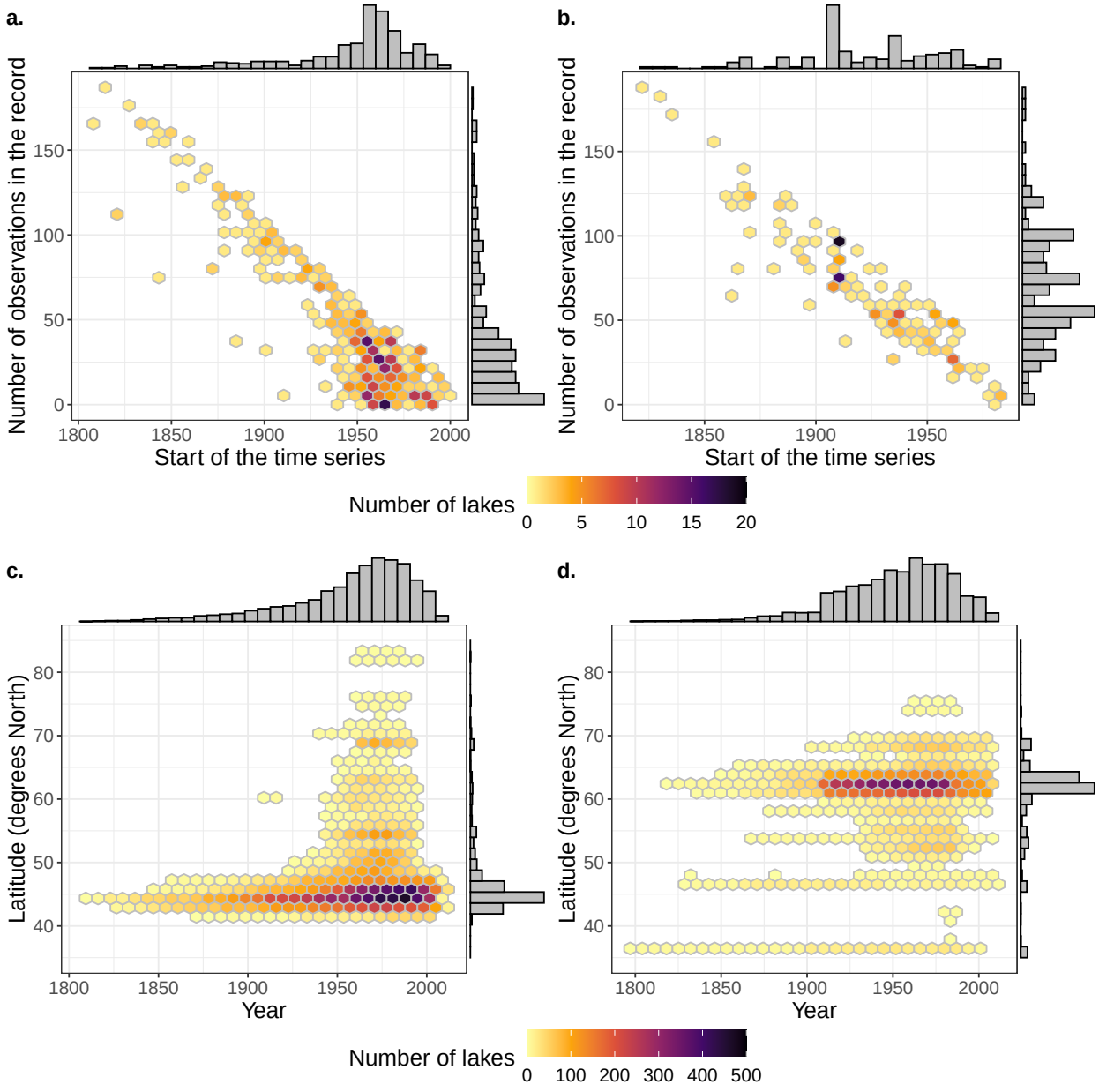


Figure 9: Number of lakes in the final North American (a) and Eurasian (b) datasets for a given record length and starting year, and number of observations in the final North American (c) and Eurasian (d) datasets for a given latitude and year. The marginal histograms are relative to the axis they are opposite to.

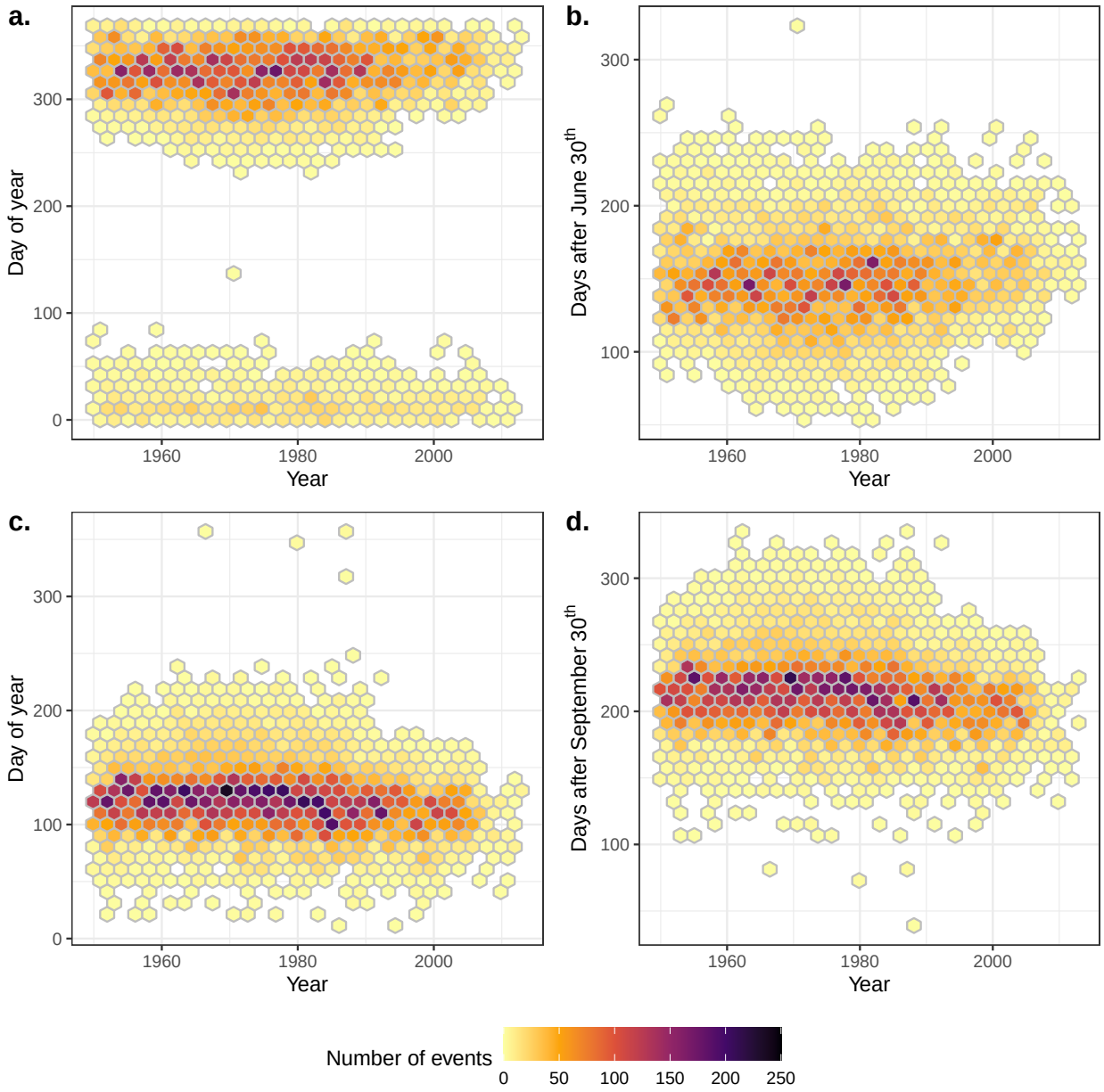


Figure 10: Number of freezing (a, b) and thawing (c, d) events on a given day. Panels (a) and (c) indicate the day of year of the event, such that January 1^{st} is 1. Plot (b) indicates the days of freezing as the number of days after June 30^{th}, such that July 1^{st} is 1. Plot (d) indicates the days of thawing as the number of days after September 31^{st}, such that October 1^{st} is 1.