# 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 the loss of lake ice, but the majority of the studies use data from single sites and linear methods, which are unable to estimate large-scale spatio-temporal trends, nor can they account for the recent acceleration in loss. This paper estimates the daily hazard of a lake freezing or thawing using a smooth, hierarchical, time-to-event approach. We use piecewise-exponential additive models to estimate spatio-temporal changes in lake ice onset and offset in the past 70 years using a large dataset of 568 lakes and 628 distinct observation stations from the northern hemisphere. 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 changes in lake ice 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 First Nations.

# 1 Introduction

Approximately half of the Earth's lakes freeze periodically (Verpoorter *et al.*, 2014), with important consequences for the biota that inhabit them (Hampton *et al.*, 2017). Despite this, under-ice ecology in lakes has been relatively under-studied (Hampton *et al.*, 2017). It is known that net primary productivity (NPP) in lakes is generally lower in winter than in summer, but the main drivers of this seasonality are hard to identify (Hampton *et al.*, 2017). Lower winter NPP may be due to multiple factors, including lower inputs of heat, nutrients, photosynthetic radiation, and oxygen which often result form seasonal ice cover (Vincent & Laybourn-Parry, 2008; Hampton *et al.*, 2017). Still, there are some lakes that do not exhibit significant seasonality in NPP despite winter ice cover (Hampton *et al.*, 2017).

Lake ice is also important from an anthropological viewpoint – many peoples and communities depend on winter ice cover for economic, cultural, and spiritual activities . 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. Many communities also rely on ice fishing as a mean of sustenance during winter (Knoll *et al.*, 2019). The annual freezing of some lakes has an important role in local religions and cultural identities, 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 many First Nations People in in northern Canada (Golden, Audet & Smith, 2015).

Some research has been done on estimating the effects of a warming climate on lake ice phenology, including estimating the change in frequency of lake ice formation (Sharma *et al.*, 2016) and the main drivers of lake ice loss (Sharma *et al.*, 2019; Lopez, Hewitt & Sharma, 2019). Unfortunately, however, most of this research has been done on individual lakes, rather than at regional or global scales, and a substantial portion of the studies used inappropriate methods that might have produced biased and inaccurate results. Some authors

heve analyzed multiple time series, but often times they compared estimated changes by regressing on the estimated coefficents, rather than fitting a single common model (e.g. Warne *et al.*, 2020).

Fitting a single model to all time series at once reduces the complexity 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 $\Delta t$. 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 it (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:

$$\widehat{F}_{freeze}(t) = P(T_{freeze} \leq t), \tag{1}$$

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

$$\widehat{F}_{thaw}(t) = P(T_{thaw} \leq t). \tag{2}$$

The probability of an event happening at or before time $t$ is equal to the complement of the event happening *after* time $t$, i.e. $P(T \leq t) = 1 - P(T > t)$. From a survival analysis

perspective, $P(T > t)$ 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 estimated using the survival function $\widehat{S}(t)$. With regards to lake ice phenology, $\widehat{S}(t)$ indicates the probability of a lake freezing or thawing after time $t$. Therefore, we can state that

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

where $t$ is a specific point in time, $T$ is the random variable for time, and $\widehat{S}(t)$ 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 window of time $\Delta t$, such as a single day or a week, given that the event has not yet happened. Mathematically, we can write this as $P(t < T \leq t + \Delta t)$. We can estimate the hazard of an event happening by dividing $P(t < T \leq t + \Delta t)$ by $\Delta t$, so that we can compare hazards from time periods of different $\Delta t$s. If we let $\Delta t$ be 1 day, we can estimate the daily hazard of the event happening on any day using the *hazard* function (Kleinbaum & Klein, 2012):

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

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

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

Intuitively, the probability of an event happening after time $t$ and the cumulative hazard of the event are closely related. Using the fact that $P(A|B) = \frac{P(A \wedge B)}{P(B)}$, we can re-write the hazard function in the form:

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

and since $\widehat{S}(t) = P(T > t)$, we can further re-write $\hat{\lambda}(t)$ as

$$\hat{\lambda}(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t)}{P(T \geq t)\Delta t} = \lim_{\Delta t \to 0} \frac{1 - \widehat{S}(t + \Delta t)}{\widehat{S}(t)\Delta t}.$$

Next, we can use the limit definition of derivatives to show that

$$\hat{\lambda}(t) = \lim_{\Delta t \to 0} \frac{1 - \widehat{S}(t + \Delta t)}{\widehat{S}(t)\Delta t} = \frac{\lim_{\Delta t \to 0} \frac{1 - \widehat{S}(t+\Delta t)}{\Delta t}}{\widehat{S}(t)} = -\frac{\frac{\partial \widehat{S}(t)}{\partial t}}{\widehat{S}(t)}.$$

Therefore, we can define the hazard function to be the negative change in survival over time, divided by the survival itself:

$$\hat{\lambda}(t) = -\frac{\frac{\partial S(t)}{\partial t}}{S(t)}.$$

Finally, by solving for $\widehat{S}(t)$, we can estimate the survival function from the hazard function (Kleinbaum & Klein, 2012):

$$\widehat{S}(t) = \exp\left[-\int_0^t \hat{\lambda}(T)\ dT\right] = \exp\left[-\widehat{\Lambda}(T)\right]. \tag{5}$$

This allows us to estimate the probability of an event occurring at or before time $t$ as a function of the cumulative hazard:

$$\widehat{F}(t) = 1 - e^{-\widehat{\Lambda}(t)}. \tag{6}$$

Under these assumptions, events are expected to occur (i.e. $P(T \leq t) = P(T \geq t) = 0.5$) when the hazard is equal to $\lambda(t) = 1.177$ (and thus $\log[\lambda(t)] = 0.1633$.

Piecewise-exponential Additive Models (PAMs, see Bender *et al.*, 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 $\log\left[\mathbb{E}\left(\hat{\lambda}(t)\right)\right]$. With PAMs, it is possible to estimate functions (1)-(6). 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 *et al.*, 2018). PAMs standardize $\log\left[\hat{\lambda}(t)\right]$ using an offset term equal to the log-transformed number of days between observations, $\log(\frac{t_i - t_{i-1}}{\Delta t})$.

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 that the $i^{th}$ observation corresponds to the $i^{th}$ row with the interval $[t_i, \Delta t_i)$, the corresponding offset, and a binary indicator variable $j_i$ which is equal to 0 if the event did not happened and 1 if the event occurred in the interval (Bender *et al.*, 2018). An example of a PED structure is given in section 2.1.

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 used smooth 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 the the spatial component of the model and the unaccounted variation between lakes (Pedersen *et al.*, 2019).

# 2 Methods

## 2.1 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, 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 8). 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. 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 (SI, Figures 9 and 1). 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 after June $30^{th}$ and September $30^{th}$, respectively, to avoid the discontinuity that would have occurred if using the day of year (Figure 2, and SI Figure 10).

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

Table 1: Example of piecewise exponential data format for the freezing dates of lakes in North America.

| 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 |

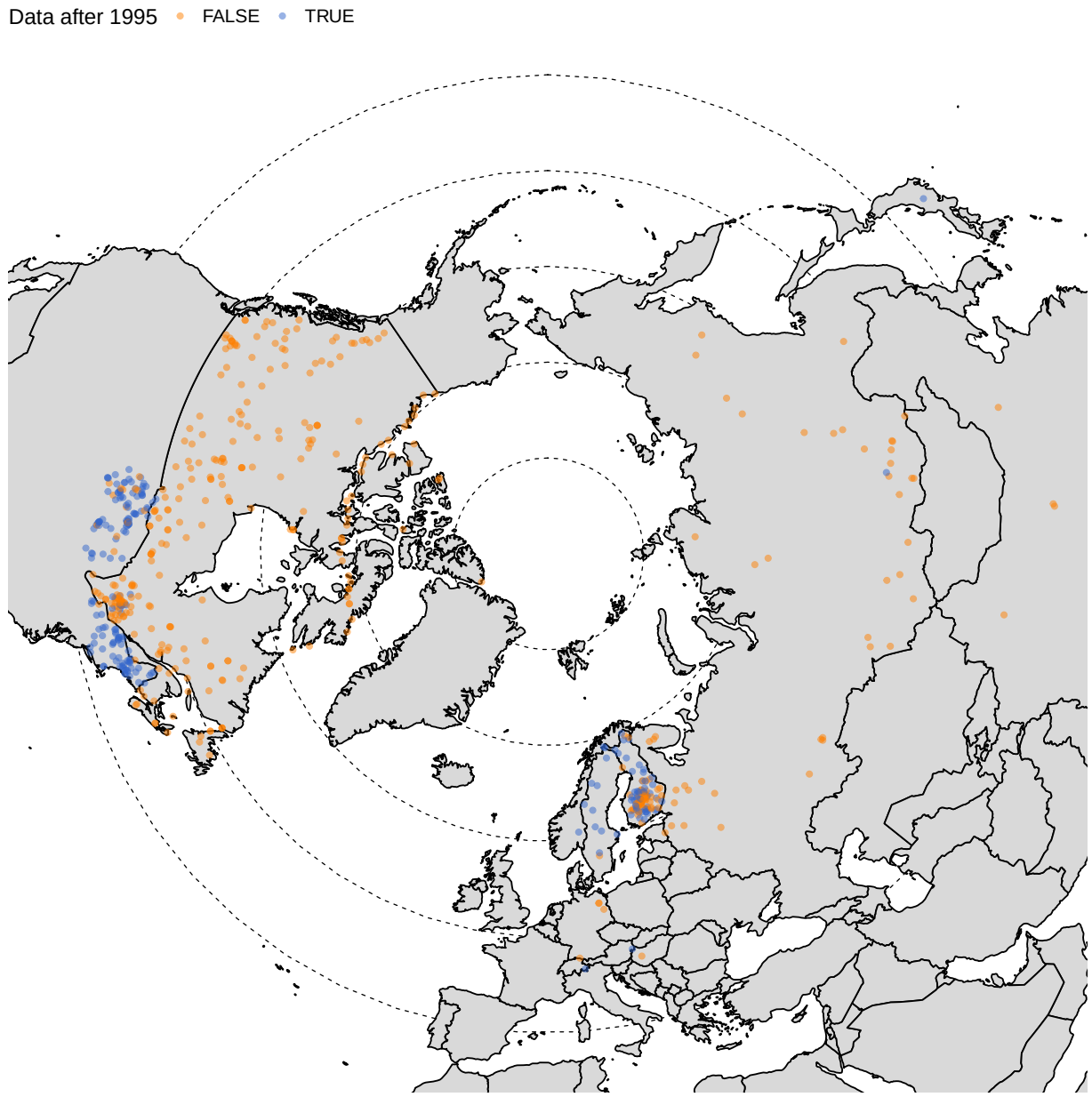


Figure 1: Polar projection of the lakes that in the final dataset. The points in orange indicate lakes for which no data was available after 1995, while blue points indicate lakes with at least one datapoint after 1995.

The columns `tstart` and `tend` are the beginning and end of intervals $[t_{i-1}, t_i)$ (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), and `Year` indicates the reference year. The thaw PEDs had a similar structure.

All of the data processing was performed on in `R` (R Core Team, 2020), 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). Years in which a lake did not freeze were excluded from the thaw dataset (for that lake alone, all other observations for that year were kept).

## 2.2 Software

All statistical analyses were performed in `R` version 3.6.2 or higher. HPAMs for freeze and thaw dates were fit using the `pammtools` package (version 0.2.2 or higher; Bender & Scheipl, 2018; Bender *et al.*, 2018). All plots were generated using `ggplot2` (version 3.3.0; Wickham, 2016), `gratia` (version 0.3.1; Simpson & Singmann, 2020), and `cowplot` (version 1.0.0; Wilke, 2019). When necessary, plots use a palette with colors that are distinguishable by most color-vision-deficient people.

## 2.3 Model structure

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 slightly between lakes. Tensor 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).

The `bs` arguments indicate which basis type each smooth uses. Cubic regression splines (`cr`) are fast-fitting and one-dimensional splines composed of cubic polynomials. Factor smooths bases (`fs`) fit a penalized 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 the 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 $(a-1)(b-1)$.

Finally, the `method` argument indicates that the smoothness parameter should be estimated using restricted marginal likelihood, while `engine = 'bam'` indicates that `mgcv::bam()` should be used and `discrete = TRUE` discretizes the posterior to decrease computational cost and fitting time (Wood, 2011). The structure of the thaw model is essentially identical, with the exception that the response $Y$ is 0 if the lake is 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

The results from the HPAMs fit to the North American and Eurasian datasets are shown in Figures 3 through 7. 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 5.5 days later and thawed 5.5 days earlier than in 1950, while North American lakes in 2010 froze 33 days later and thawed 10 days earlier than in 1950 (Figure 3). 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).

The high smoothness of the `year` factor smooths in each of the four models (Figure 4a) indicates that the trend in $\hat{\lambda}(t)$ over the years is close to the average (`s(year)`), once the effects of space and `tend` are accounted for. In contrast, the high spread of the `tend` factor smooths (Figure 4b) indicates that there are strong drivers of seasonal variation that are not accounted for by the models.

When accounting for the spatial effects in the predictions, the trends are not spatially uniform or monotonic. Areas in Western and Northern North America were estimated to have much later freezing dates (or not freeze at all) in 2010, whith most of the change occurring after 1995, while freeze dates in the Eastern US did not change substantially. European lakes were estimated to have frozen somewhat earlier in 1995 than in 1950 but have frozen three weeks later in 2010. Asian lakes froze somewhat later in 1995 but sligthly earlier in 2010, compared to the 1950 dates (Figure 5a). The North American thaw model estimated an earlier thaw of about 1-2 months in north-western Canada by 2010 and a 10-day shift towards later thaw in the Eastern US. Overall, the Eurasian model estimated little to no changes in thaw dates in 1995, followed by a widespread seven-day shift to earlier thawing in 2010, with the exception of Eastern Asia, where lakes thawed a week later on average.

# 4 Discussion

The widespread shift to later freezing and earlier thawing dates has greatly changed the lives of many many people, including those of Indigenous People in Northern Canada and Alaska. Many First Nations report a reduction in food and energy security as well as

major changes to traditional activities (Golden *et al.*, 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, and they have a deep spiritual and emotional connection with the seasonal freezing of lakes, since many of their traditional activities and cultural views are deeply connected with the formation of lake ice, particularly "blue ice" (Golden *et al.*, 2015). Blue ice occurs in coastal and sloped locations when winds are sufficiently strong and persistent to cause the surface of the ice to melt and re-freeze slowly. The slow freezing process and the ablation cause the gas bubbles commonly present in ice to escape, which results in more "pure" ice with the characteristic blue hue (Winther, Jespersen & Liston, 2001). 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 *et al.*, 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. Ice fishing on Lake Peipsi, Estonia, attracts 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), but 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).

***linebreak***

Despite the low data availability after 1995, the estimates produced by the models agree

with recently published results (e.g. Brown & Duguay, 2011), but the time-to-event approach and the model's nonlinearity allowed for finer spatiotemporal 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. Although the data scarcity after 1995 resulted in a substantial increase in the uncertainty after 1995, 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 would produce more informed (and less uncertain) estimates, but the current estimates are not unreasonable.

The spatial term `s(long, lat)` and its tensor interaction terms (`ti(tend, long, lat)`, and `ti(Year, long, lat)`) decreased the potential bias that arised from the non-random and opportunistic nature of the sample of lakes. By accounting for the effect of space and how the effect changed over time, the model was able to estimate the changes in $\lambda(t)$ over time without assuming an equal change in all regions. Even in areas where data was scarce or unavailable, the model was still able to produce realistic estimates with a measure of uncertainty. The Duchon spatial spline appears to remain sufficiently constrained outside the spatial range of the data, even near the edges of the spatial range of the data. The tensor interaction terms `ti(tend, long, lat)`, and `ti(Year, long, lat)` allowed the estimated loss to increase at faster rates in areas where the change is accelerating (e.g. North-Western Canada, Scandinavia) and decrease in other areas (e.g. Eastern US and Eastern Asia).

While the estimated effect of `year` after 1995 is driven mostly by lakes in the Great Lakes Area and Scandinavia, the spatial bias in data availability was reduced substantially by allowing `s(year)` and `s(tend)` to vary over space. 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 $\hat{\lambda}(t)$ over time and would lack any information on the spatial distribution of the lakes.

While a part of the spatial effect would be accounted for by the random effects in the `fs` terms of `year` and `tend`, `fs` 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.

$$\log\left[\hat{\lambda}(t)\right] = f_1(year) + f_2(tend) + \epsilon, \tag{7}$$

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 that cannot be explained without information on the spatial location of the lakes, or other characteristics such as lake depth and area.

In addition, estimating the average trends from multiple models would prove rather difficult and likely inapproriate. While estimating the average coefficients of linear models would be mathematically simple, these estimates are unable to account for any nonlinearity in the trends, such as the accelerating warming of lakes (Velicogna, 2009; Zhong *et al.*, 2016).

Estimating the average change in $\hat{\lambda}(t)$ or $\widehat{F}(t)$ (or the average freezing date) using linear models and then estimating the average coefficients (e.g. Warne *et al.*, 2020), prevents the model from accounting for the between-lake variation that arises due to lake depth, area and spatial location. In addition, assuming linear trends prevents the model from estimating any acceleration in the loss of ice.

A hierarchical, nonlinear model which accounts for the effects of space and how the yearly and seasonal trends change over space is best suited for estimating changes in lake ice phenology, as it relaxes many assumptions that are likely problematic, and it avoids post-hoc

estimates of average trends. However, the accuracy of any model depends on the data to which it is fit, so even the models presented in this paper could be improved.

An increase in data, particularly post 1995 and in areas with low data availability, would reduce the uncertainty in the estimates and the potential bias in the temporal marginal smooths (i.e. `s(tend)` and `s(year)`). In addition, since each model is fit to all data from the relative continent simultaneously, the observations do not need to be many years in length. 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 space into the models, spatial and temporal gaps in the data can be filled using data 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 datasets of this size.

Besides adding more rows of data (i.e. more observations), models can also be improved by including more columns of data (i.e. predictors) in the models. 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 that do not freeze often (such as lakes at low altitudes). Although lake altitude was present in the GLRIPD, it was not available 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 (Shuter, Minns & Fung, 2013, @magee_effects_2017). Currently, these effects and other stochastic effects (e.g. wind dis-

turbance, 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 $\hat{\lambda}_{freeze}(t)$ and $\hat{\lambda}_{thaw}(t)$, and shallower and larger lakes appear to be more resilient to loss of ice.

The explanatory power of the models could also be increased by accounting for the exposure to above-zero or sub-zero temperatures. The smooth of air temperature 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 of the models, the models would no longer be appropriate for estimating whether lake ice cover is changing over time, since they estiamte 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 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 time-invariant (i.e. uncorrelated) during the period of analysis, such as lake depth and surface area.

# 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. To appropriately estimate changes in lake ice phenology, it is necessary to use smooth, hierarchical, and flexible models such as those presented in this paper to allow the trends to vary spatio-temporally. While many different types of modelling approaches are available, it is important to design the models according to the question that is being adressed.

The results presented here demonstrate a widespread and accelerating but heterogeneous

shift towards later freezing and earlier thaw, but also an increase in lake ice cover in some areas. The hierarchical structure of the models allowed us to estimate changes in lake ice in areas where little to no data were available.

# 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

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

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

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

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

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

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 (2020). *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., Minns C. & Fung S. (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. & Singmann H. (2020). 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

Verpoorter C., Kutser T., Seekell D.A. & Tranvik L.J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters* **41**, 6396–

6402. https://doi.org/10.1002/2014GL060641

Vincent W.F. & Laybourn-Parry J. eds (2008). *Polar Lakes and Rivers: Limnology of Arctic and Antarctic Aquatic Ecosystems*. Oxford University Press, Oxford.

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. (2019). Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'

Winther J.-G., Jespersen M.N. & Liston G.E. (2001). Blue-ice areas in Antarctica derived from NOAA AVHRR satellite data. *Journal of Glaciology* **47**, 325–334. https://doi.org/10.3189/172756501781832386

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

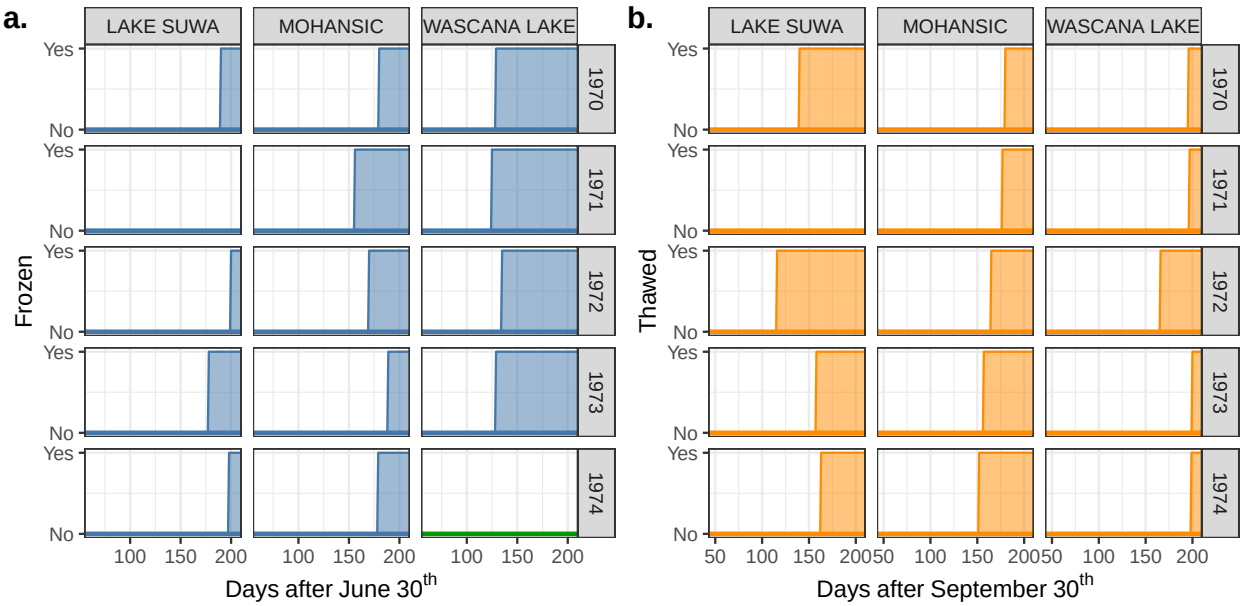


Figure 2: Freeze (a) and thaw (b) events for three lakes in the final dataset for years 1970-1974. Shaded areas indicate when the lake was frozen (a) or ice-free (b); the green baseline for Wascana lake in 1974 indicates that the lake froze, but the date of freezing is unknown. Note that lake Suwa did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

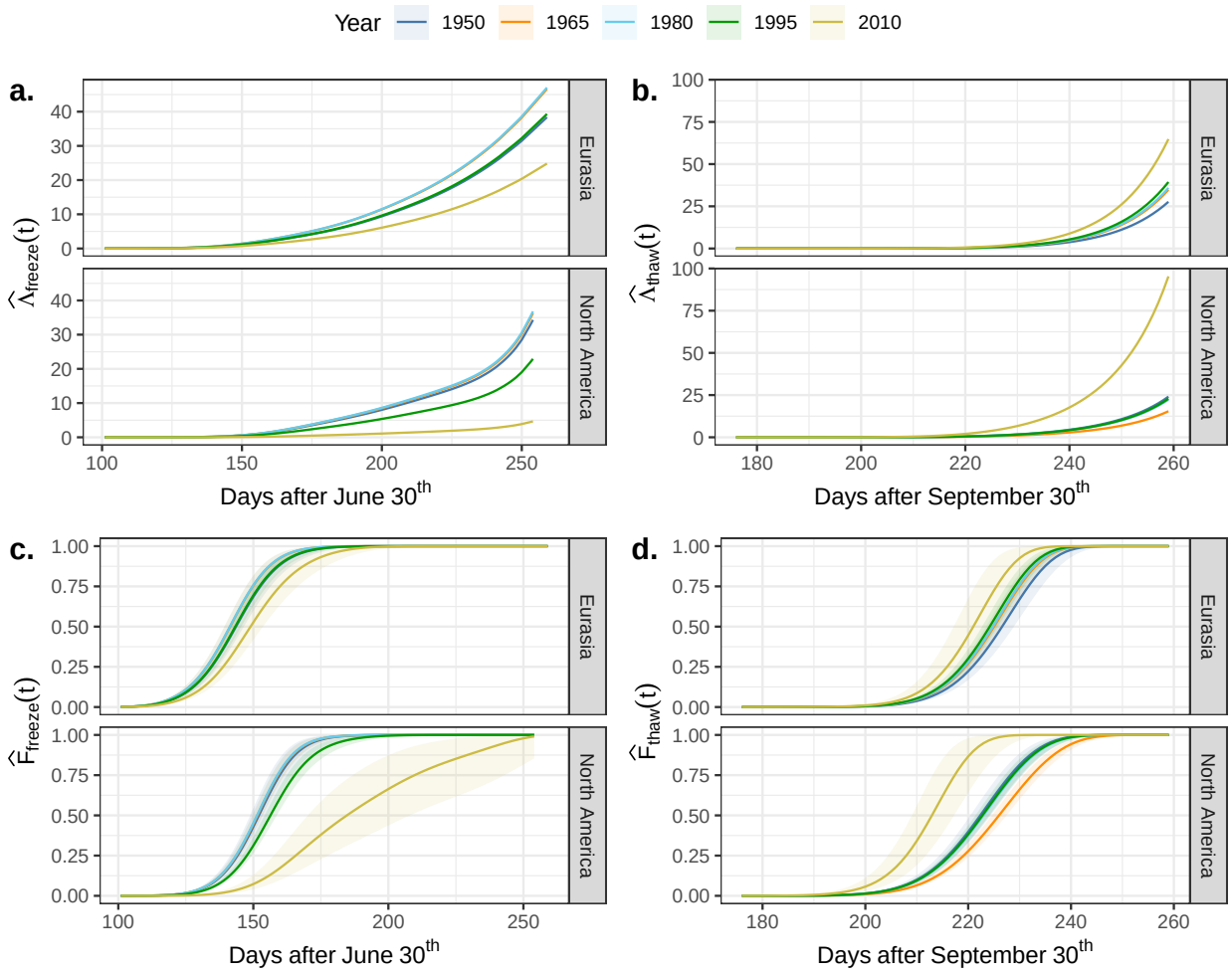


Figure 3: Estimated cumulative hazard (a, b) and cumulative probability (c, d) of lakes being frozen (left) or thawed (right) while assuming an average effect of geographic location.

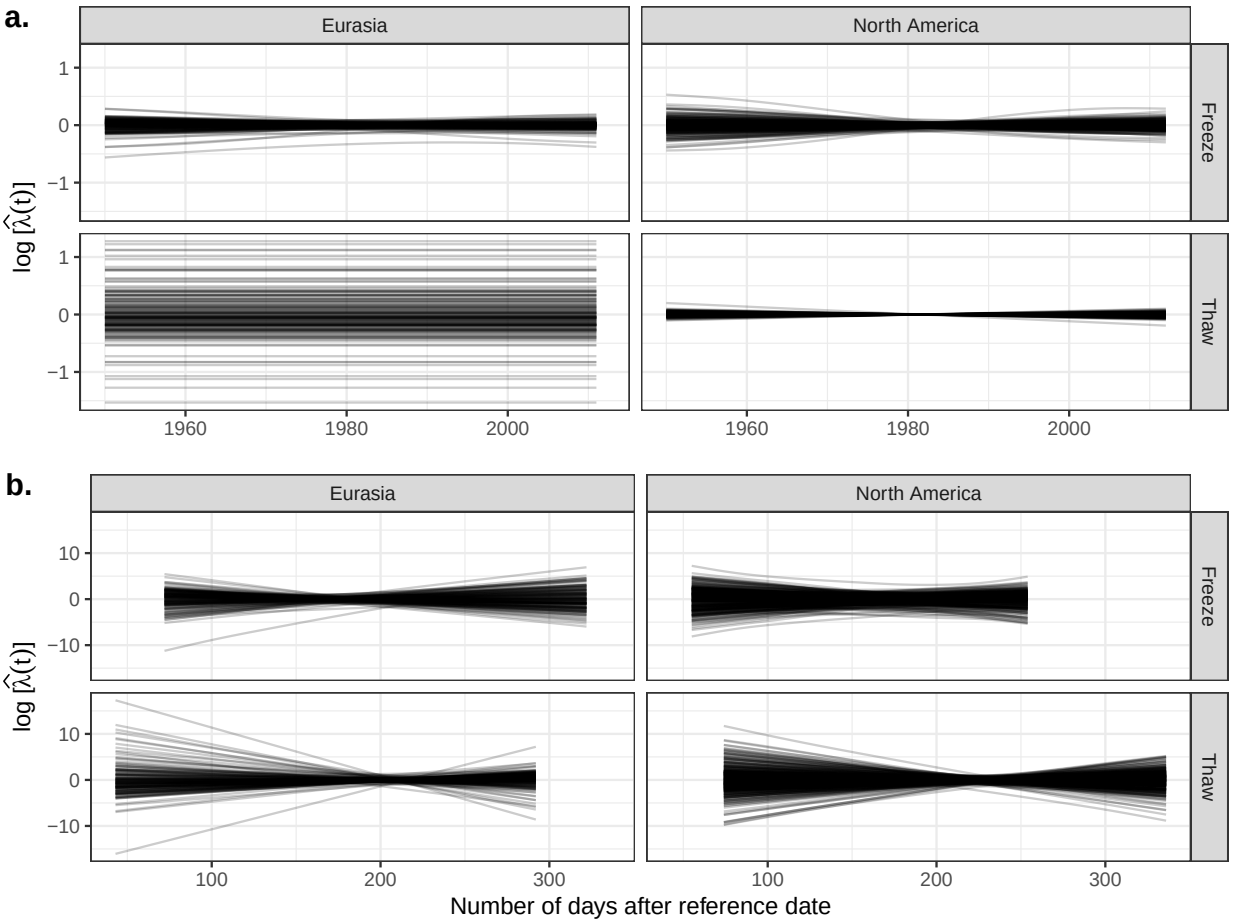


Figure 4: Factor smooths (‘fs‘) of ‘year‘ (b) and ‘tend‘ (a) from the HPAMs for the hazard of freezing and thawing in North America and Eurasia. The y axis indicates how much the hazard of each lake deviated from the mean trend, assuming Gaussian random effects of lake.

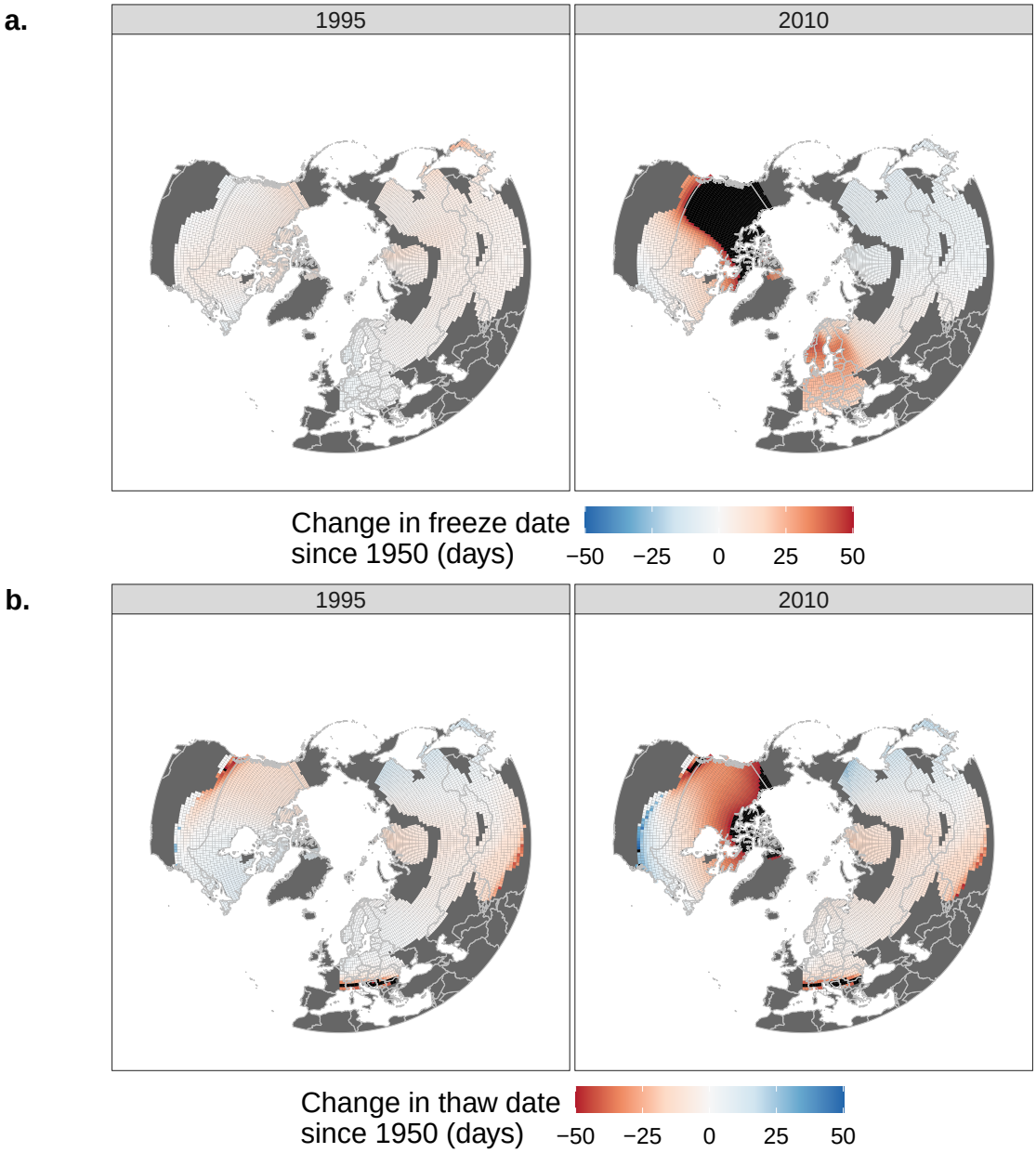


Figure 5: Estimated change in the average freeze (a) and thaw (b) dates relative to 1950. Areas in black were estimated to have an absolute change greater than 50 days. 1995 was chosen as a midpoint between the beginning and the end of the record because a large portion of the records ended in 1994 and 1995.

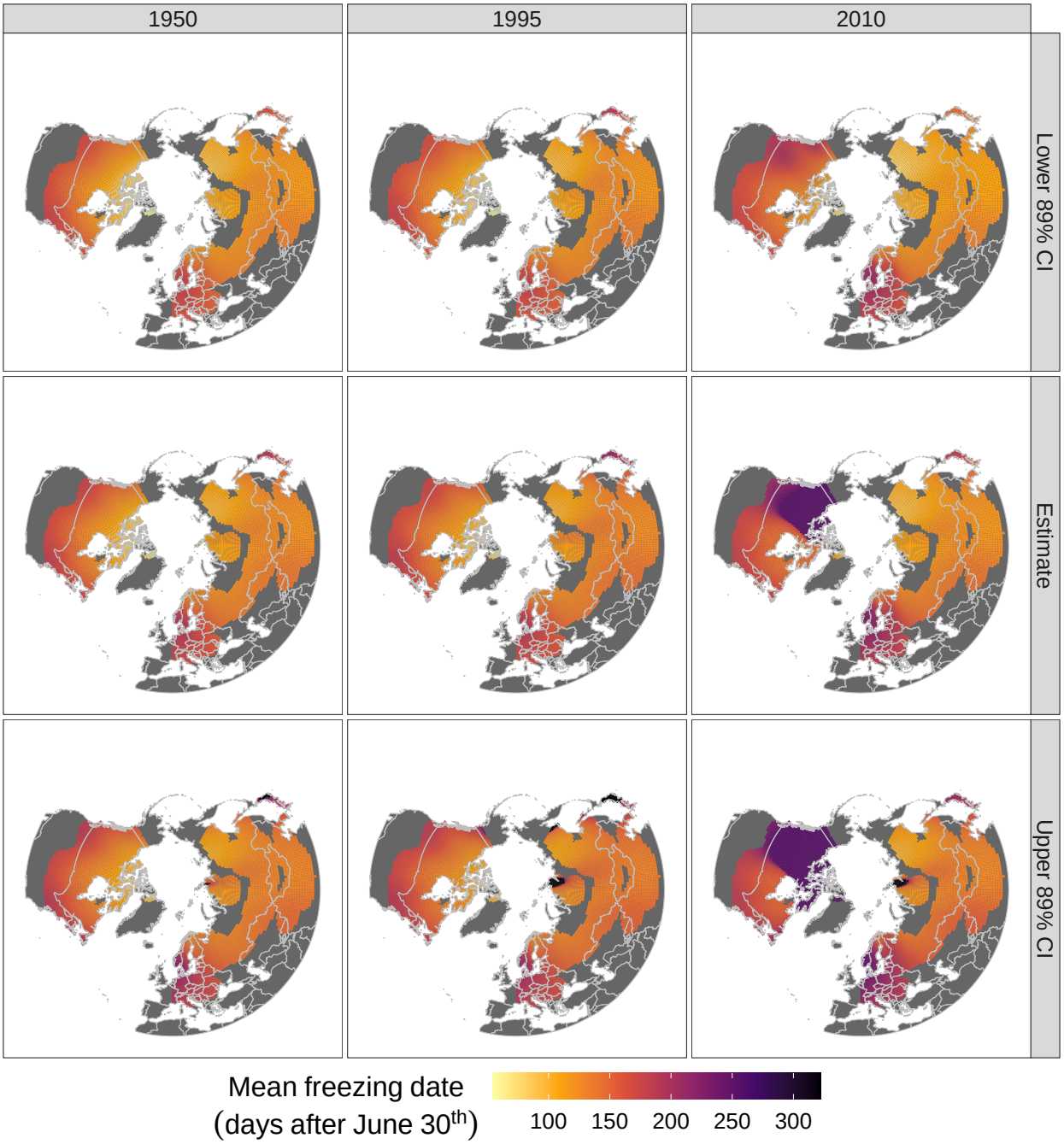


Figure 6: Estimated freeze dates and relative 89% credible intervals. 1995 was chosen as a midpoint between the beginning and the end of the record because a large portion of the records ended in 1994.

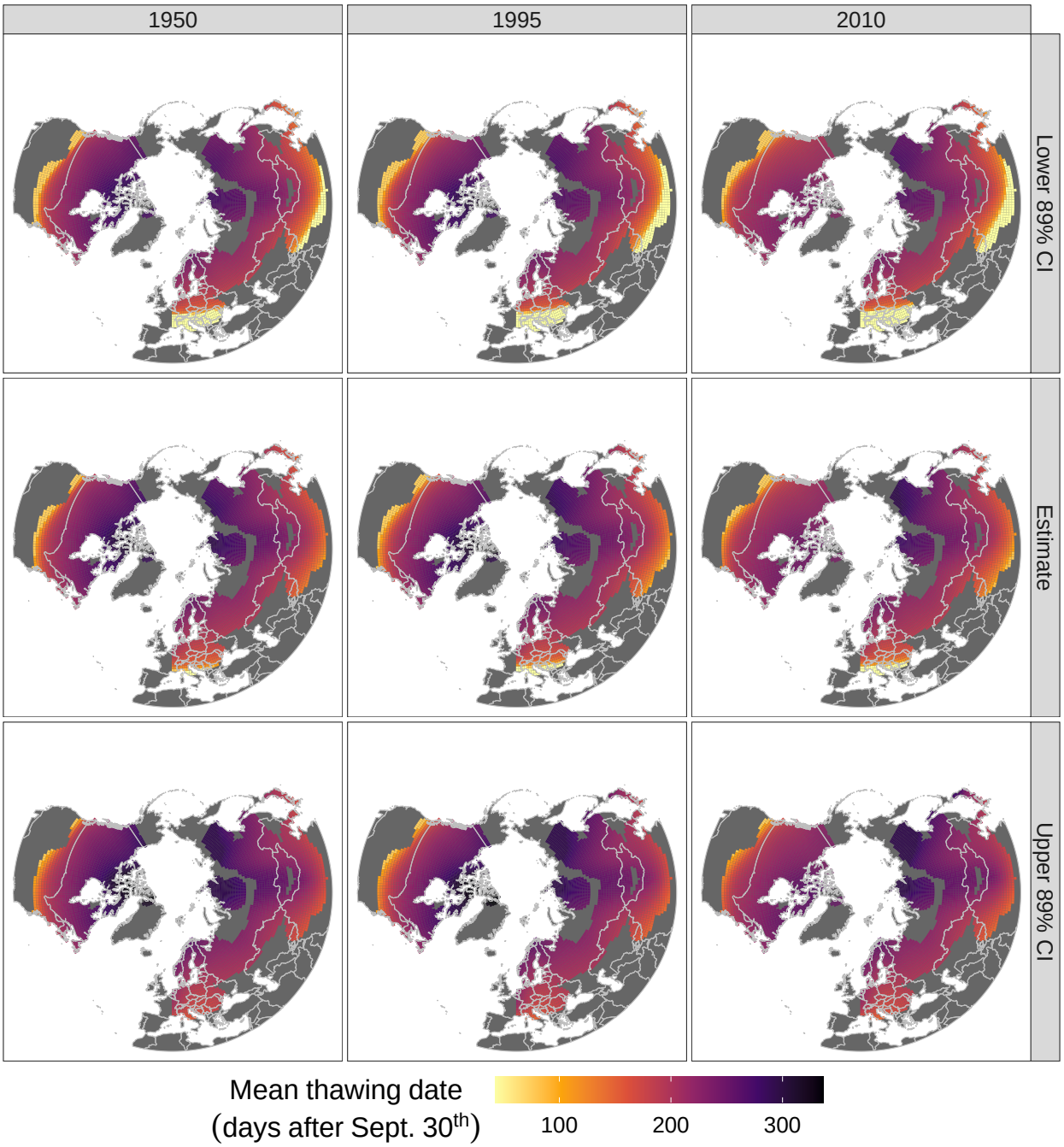


Figure 7: Estimated thaw dates and relative 89% credible intervals. 1995 was chosen as a midpoint between the beginning and the end of the record because a large portion of the records ended in 1994.

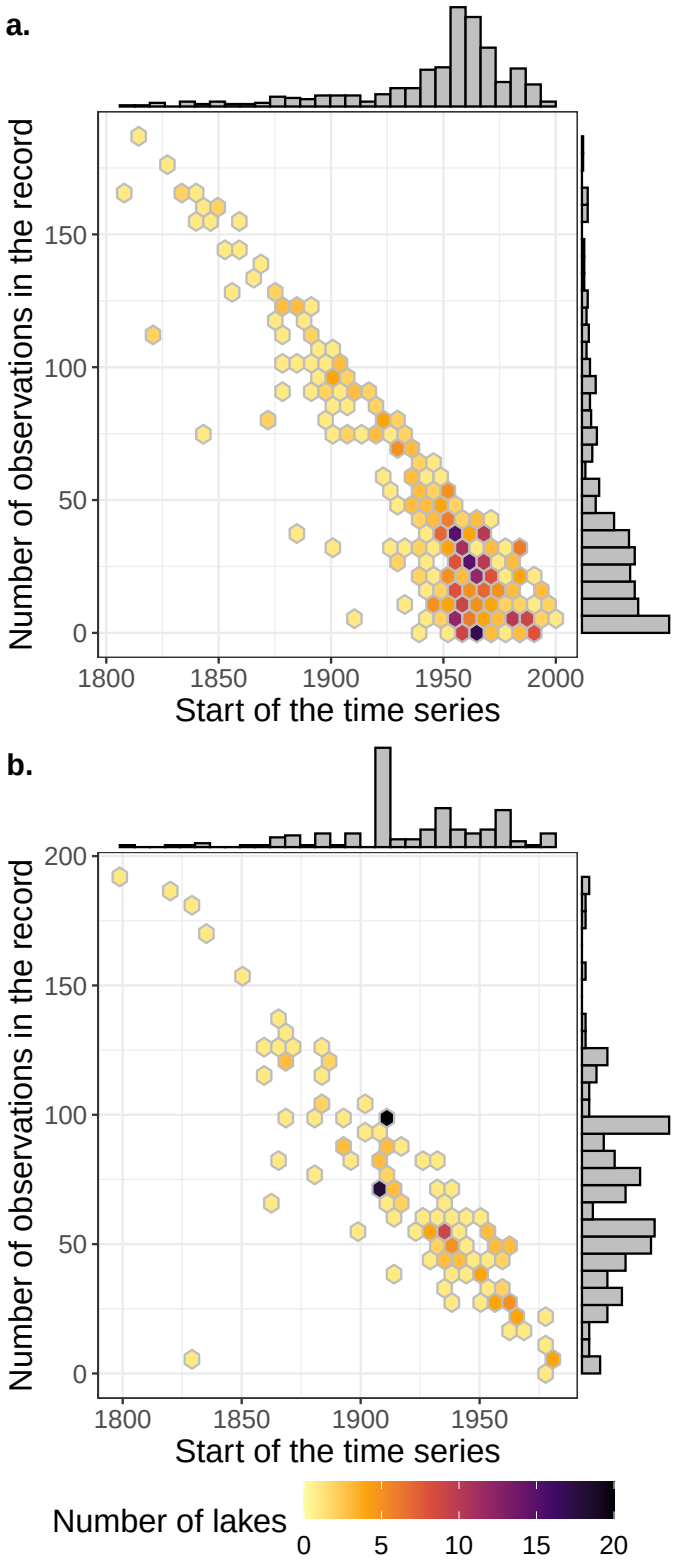


Figure 8: Number of lakes in the final North American (a) and Eurasian (b) datasets for a given record length and starting year; the marginal histograms are relative to the axis they are opposite to.

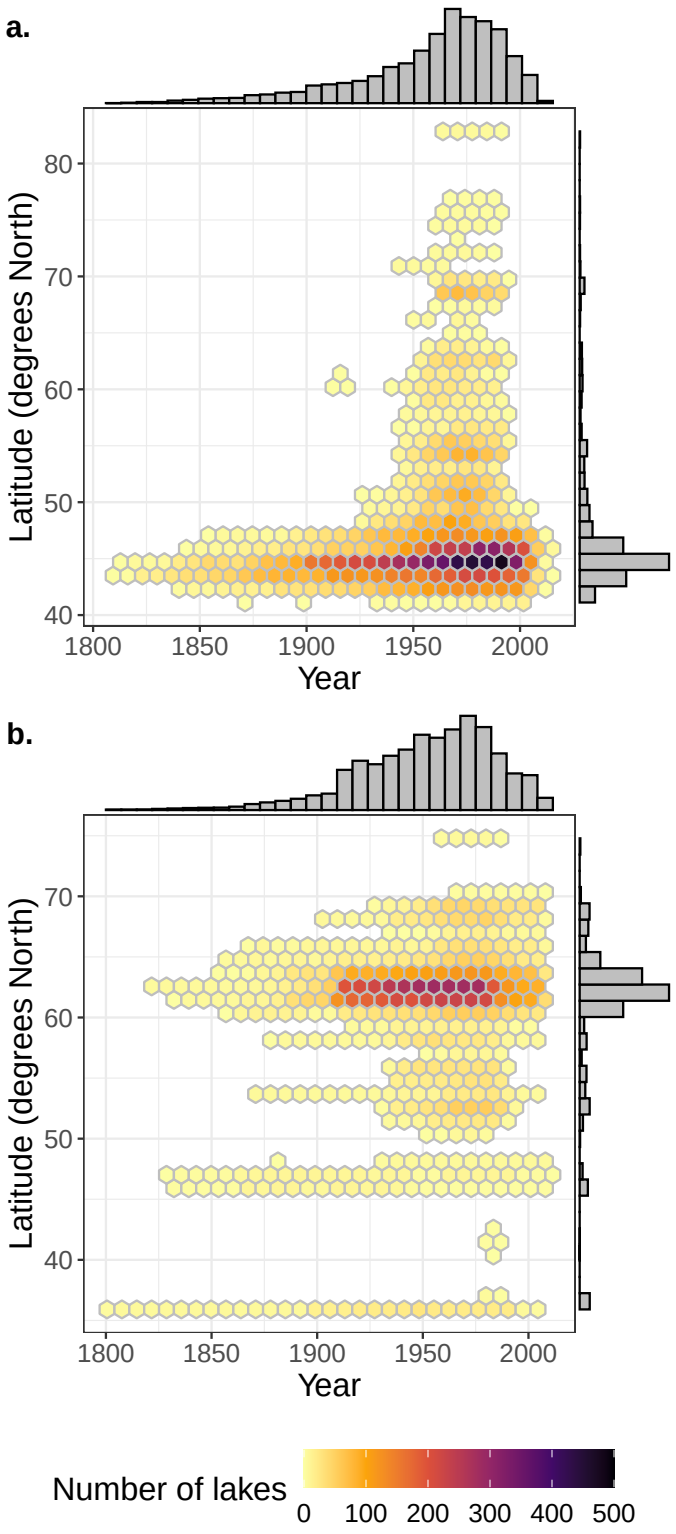


Figure 9: Number of observations in the final North American (a) and Eurasian (b) datasets for a given latitude and year; the marginal histograms are relative to the axis they are opposite to.

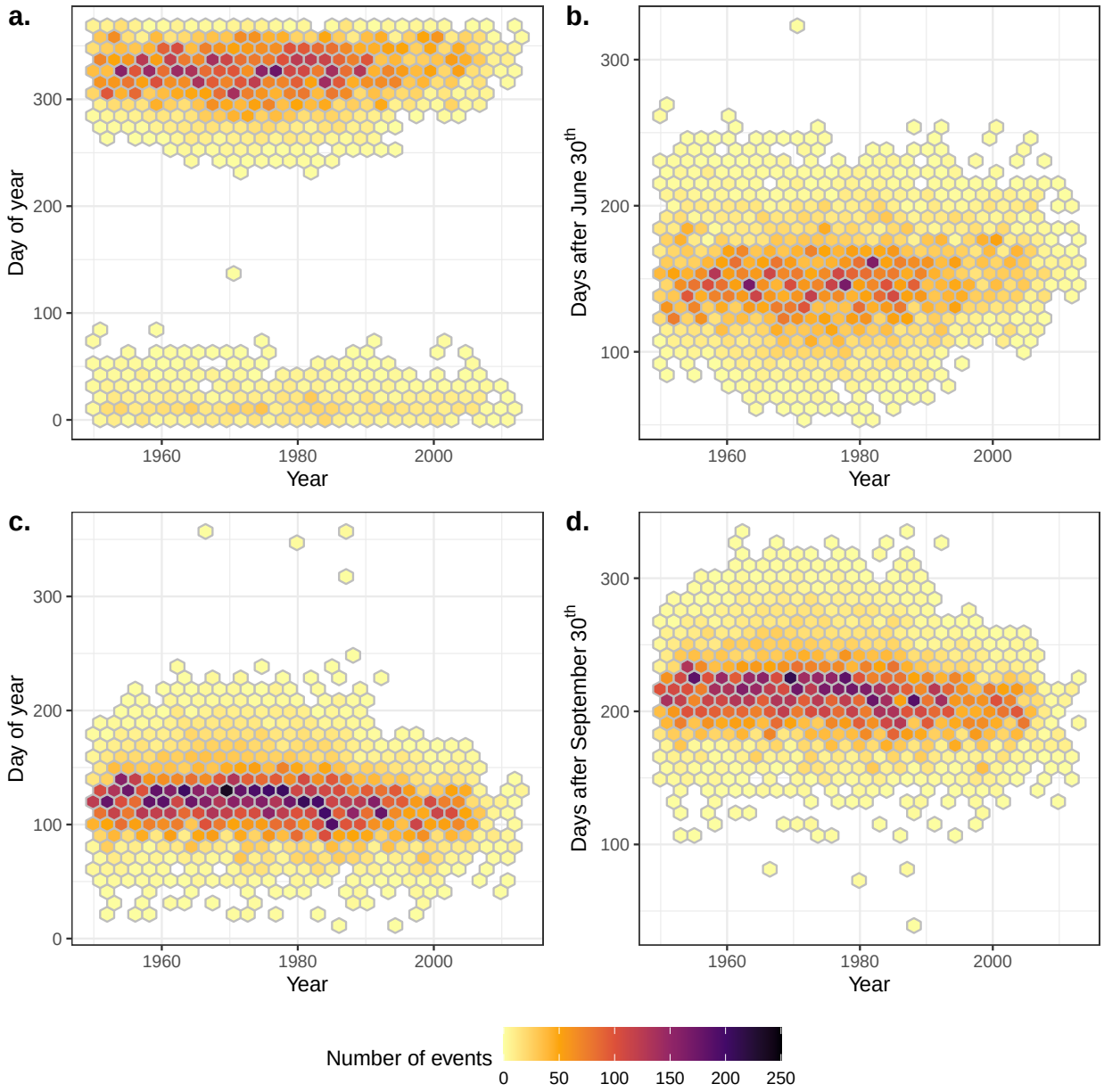


Figure 10: Number of freezing (a, b) and thawing (c, d) events on a given day. Panels (a) and (c) indicate the day of year of the event, such that January $1^{st}$ is 1. Plot (b) indicates the days of freezing as the number of days after June $30^{th}$, such that July $1^{st}$ is 1. Plot (d) indicates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