Estimating Changes in Lake Ice Occurrence in the Northern Hemisphere Using Generalized Additive Models and a Time-to-Event Approach

Stefano Mezzini Gavin L. Simpson

Abstract

Despite recently being recognized as an important control on lake ecosystem functionality, under-ice lake ecology has been relatively understudied by limnologists. Due to recent changes in climate, the frequency and duration of ice formation on lakes has decreased, and while some studies have been done on the recent loss of lake ice, the majority of these studies use data from single sites, which prevents the estimation of spatial and long-term trends at larger scales. The work presented here demonstrates how the occurrence of lake ice can be analyzed using a time-to-event approach. This project uses piecewise-exponential additive models and generalized additive models to estimate spatio-temporal changes in lake ice onset, offset, and duration in the past 70 years with a large dataset of 568 lakes and 628 distinct observation stations from the northern hemisphere. The results presented here demonstrate a widespread and severe decrease in the hazard of freezing and the duration of lake ice post 1995. Considerations are made of the effects of climate change on lake ice phenology and 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 (Hampton *et al.*, 2017), but the main drivers of this seasonality are hard to identify. 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 local communities depend on winter ice cover for economic, cultural, and spiritual activities (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. Many communities also rely on ice fishing as a mean of sustenance during winter.

The annual freezing of lakes also 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 in duration by regressing on the estimated coefficents, as in the case of 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 trends between lakes while incorporating the variance that exists between lakes. 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 lakes (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 instantaneous probability of the occurrence of an event. For instance, an ice-free lake has a specific (unknown) probability of freezing at a given moment in time, t. In contrast, the probability of a lake 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 "risk" of an event occurring in a given window of time Δt, 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 instantaneous hazard of an event happening by dividing $P(t < T \leq t + \Delta t)$ by Δt, so that we can compare hazards from time periods of different Δts. If we let Δt approach 0, we can estimate the instantaneous hazard of the event happening at any time 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 time between observations, $\log(t_i - t_{i-1})$.

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 $\log(\Delta t_i)$, and a binary indicator variable 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 a hierarchical Bayesian approach to estimate the the spatial component of the model and the unaccounted variation between lakes in the model (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 2). Although the analysis could have been performed for the entire dataset, the dataset was reduced to decrease model fitting time and potential sampling bias. 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 3 and 4). 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 1, and SI Figure 5).

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 if to gtoup observation stations (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 evnets and thaw events (given that the lake had frozen in the previous year). The PED data sets had

a structure of the form:

tstart	tend	interval	offset	ped_status	lake	Year	long	lat	altitude
0	72	(0,72]	4.28	0	SUWA	1950	138.08	36.05	759
72	81	(72,81]	2.20	0	SUWA	1950	138.08	36.05	759
81	84	(81,84]	1.10	0	SUWA	1950	138.08	36.05	759
84	88	(84,88]	1.39	0	SUWA	1950	138.08	36.05	759
88	90	(88,90]	0.69	0	SUWA	1950	138.08	36.05	759
90	91	(90,91]	0.00	0	SUWA	1950	138.08	36.05	759

The columns `tstart` and `tend` are the beginning and end of intervals (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.

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. PAMs for freeze and thaw dates were fit using the `pammtools` package (version 0.2.2 or higher; Bender & Scheipl, 2018; Bender *et al.*, 2018), while GAMs for the duration of ice

cover were fit using `mgcv` (version 1.8-31; Wood, 2017). All plots were generated using `ggplot2` (version 3.3.0; Wickham, 2016), `gratia` (version 0.3.1; Simpson & Singmann, 2020), `cowplot` (version 1.0.0; Wilke, 2019). All plots use a palette with colors that are distinguishable by most color-vision -deficient people, when necessary.

2.3 Model structure

The North American and Eurasian HPAMs for freeze and thaw dates accounted for the change in hazard within years (i.e. with `tend`) and between years (i.e. with `year`), 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 intercation 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. Using `ti` terms we can account for changes in the effect of `tend` and `year` over space to test if Arctic lakes are losing ice cover faster than temperate lakes.

The `bs` arguments whether the basis type that each smooth uses. Cubic regression splines (`cr`) are fast-fitting and one-dimensional splines composed of cubic splines. 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 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 (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 `offset = data$offset` need to be specified.)

3 Supplemental Information

Appendix: 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

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.

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

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*.

R Foundation for Statistical Computing, Vienna, Austria.

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

Simpson G.L. & Singmann H. (2020). Gratia: Graceful 'ggplot'-Based Graphics and Other Functions for GAMs Fitted Using 'mgcv'

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.

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'

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.

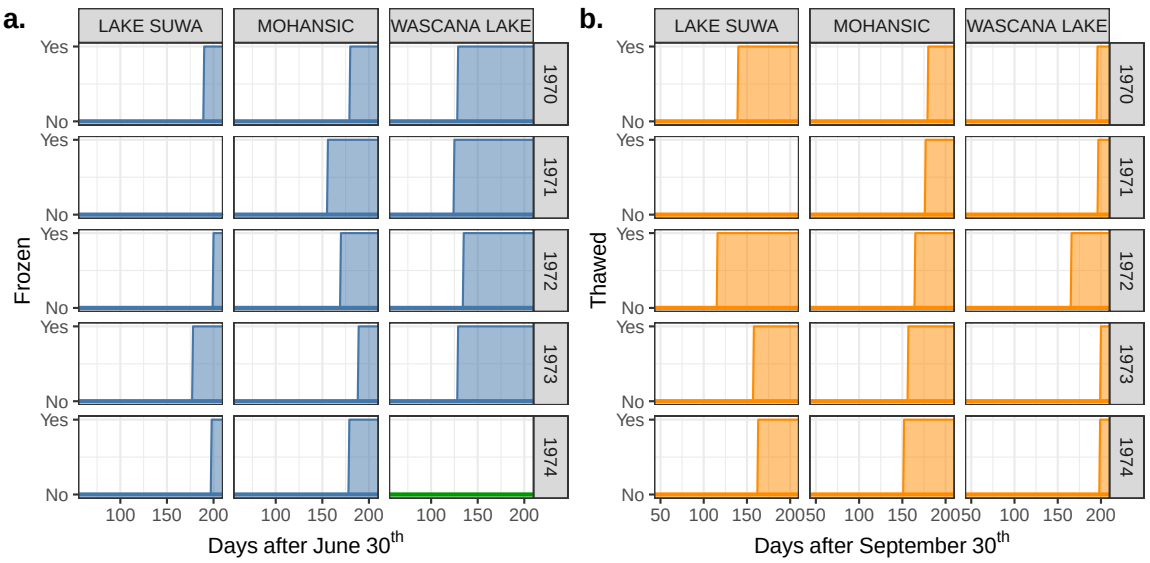

Figure 1: Freeze (a) and thaw (b) events for three lakes in the final dataset for years 1970-1979. Shaded areas indicate when the lake was frozen (a) or ice-free (b); the green baseline for Wascana lake in 1974 indicates that the lake froze, but the date of freezing is unknown. Note that lake Suwa did not freeze in 1971 and 1978.

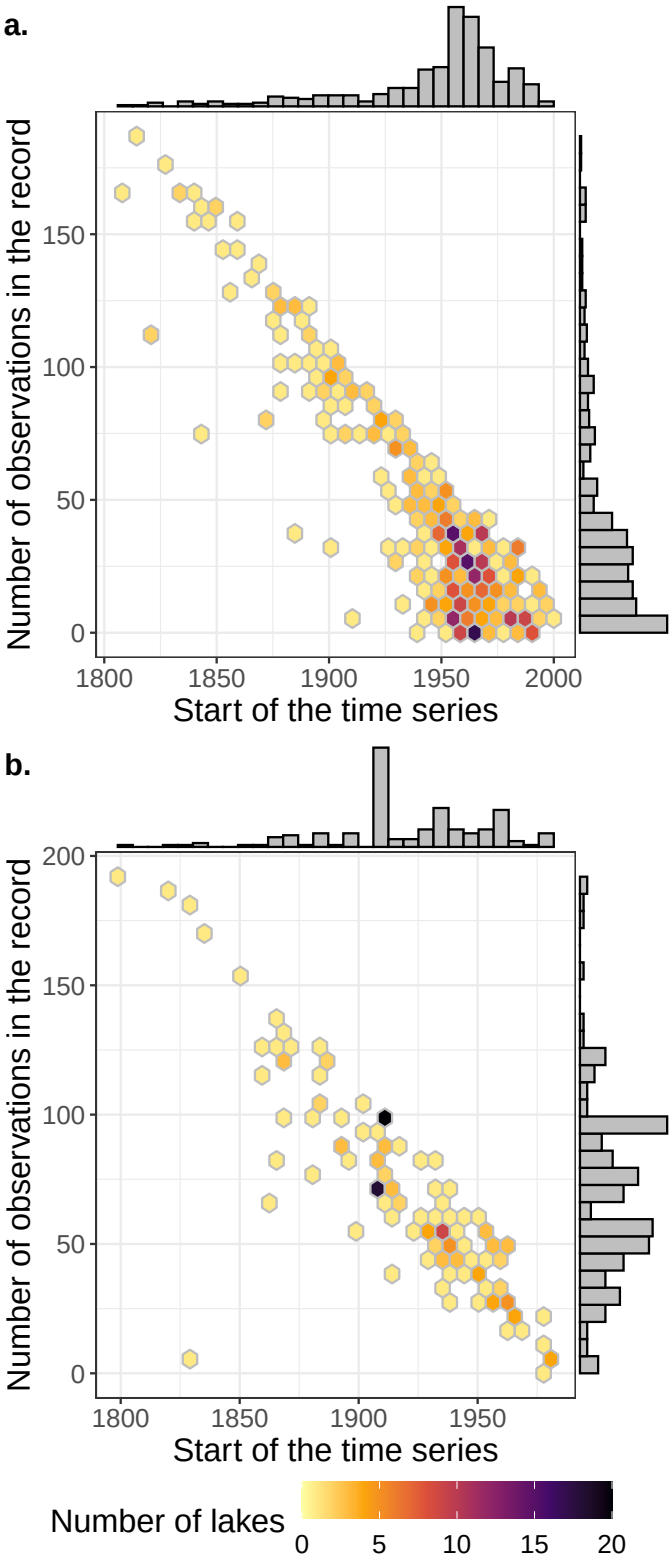

Figure 2: Number of lakes in the final North American (a) and Eurasian (b) datasets for a given record length and starting year; the marginal histograms are relative to the axis they are opposite to.

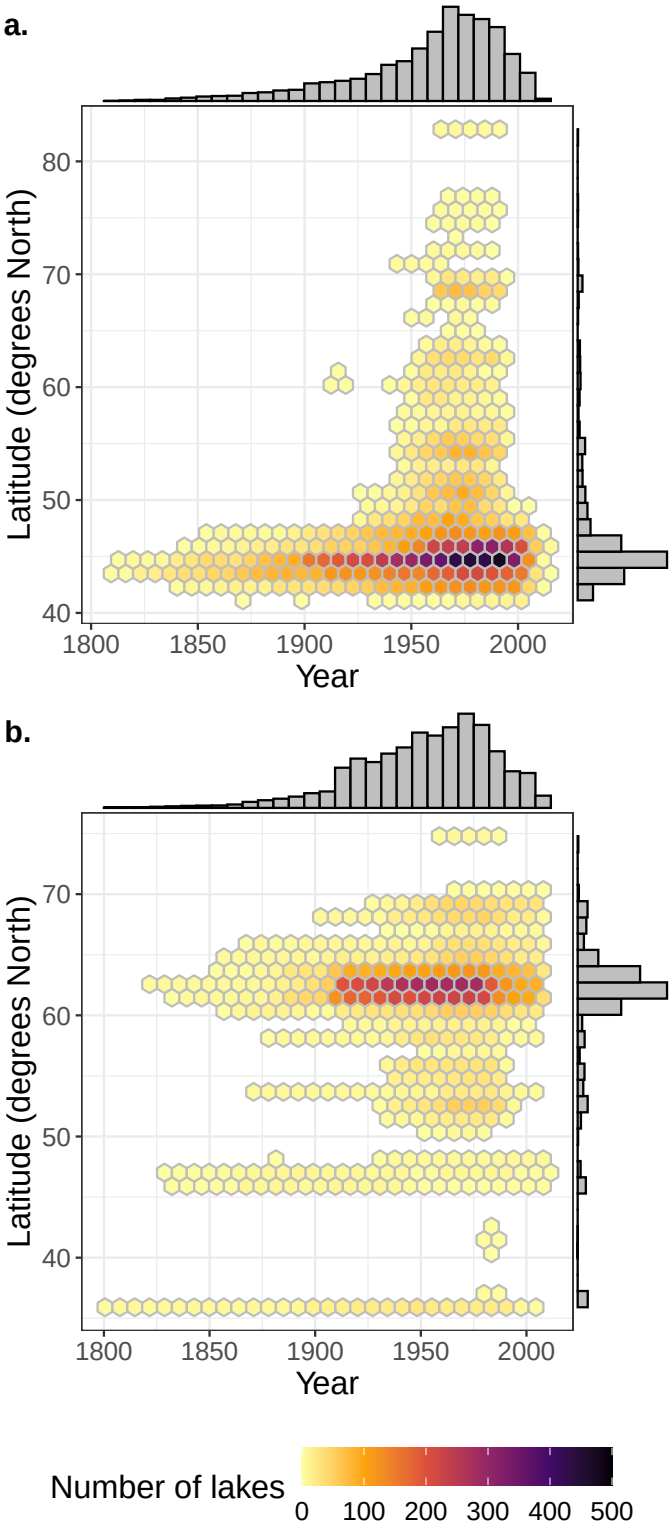

Figure 3: Number of observations in the final North American (a) and Eurasian (b) datasets for a given latitude and year; the marginal histograms are relative to the axis they are opposite to.

Figure 4: Polar projection of the lakes that in the final dataset. The dashed lines indicate the 45° N and 62° N parallels.

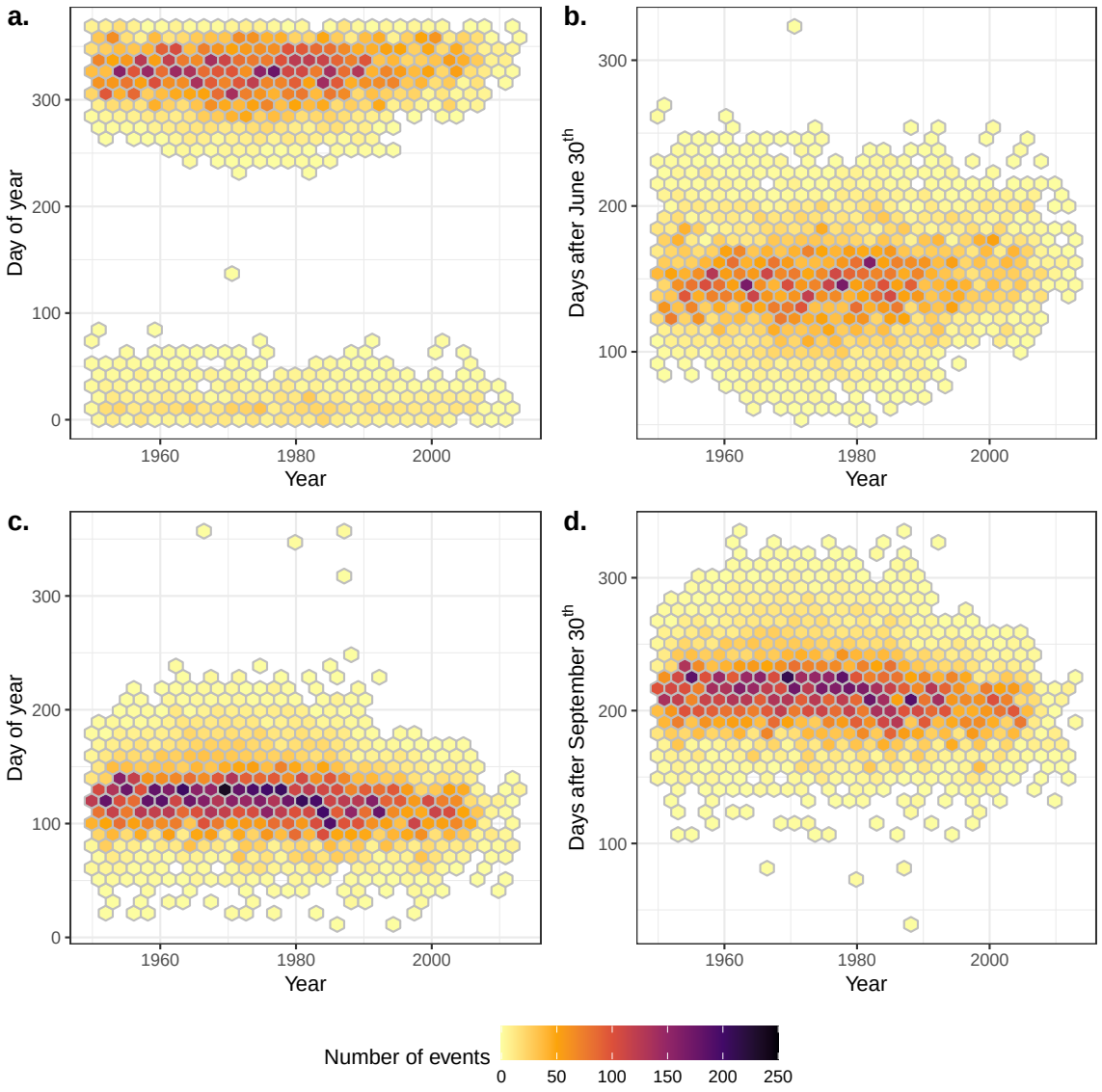

Figure 5: Number of freezing (a, b) and thawing (c, d) events on a given day. Panels (a) and (c) indicate the day of year of the event, such that January 1^{st} is 1. Plot (b) indicates the days of freezing as the number of days after June 30^{th}, such that July 1^{st} is 1. Plot (d) indicates the days of thawing as the number of days after September 31^{st}, such that October 1^{st} is 1.