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

Honors Thesis

Stefano Mezzini

April 2020

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 data set 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.

Acknowledgements

I would like to thank my supervisor, Dr. Gavin Simpson, for suggesting I pursue an Honors degree. This project has allowed me to learn a great amount of information and cutting-edge statistical methods. I look forward to continuing this work.

I would like to thank Dr. Chris Somers, Dr. Kerri Finlay, and Dr. Andrei Volodin for agreeing to be on my Honors committee and for their suggestions and guidance during the project, including the corrections to my thesis. I would also like to thank Dr. Chris Somers for taking the role of Honors Coordinator.

I would like to thank Dr. Michael Kozdron for his help with visualizing how to compare piecewise-exponential additive models to segmented regression models.

Dedication

Clarissa, you kept me going through some of the toughest times in my life. You are the center of my world and you keep it spinning!

Kim, thank you for being there to talk and understand the struggles and frustration that come with coding, modelling, and writing in Rmarkdown at the same time!

Contents

List of Figures

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 (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 may 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). The lake is well known due to the sinusoidal ice ridge that forms with the daily expansion and contraction of the ice due to fluctuations in temperature (Arakawa, 1954). The ice formation is commonly referred to as *omiwatari*, which can be be translated as "The Passage of Gods" or "The God's Crossing". The *omiwatari* is believed to form from the footsteps of the male God Takeminakata when the ice is thick enough for Him to walk on it and

visit the goddess Yasakatome. People consider the *omiwatari* to be a sign that the ice is thick enough to be safely walked on and perform purification rituals (Sharma *et al.*, 2016).

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.

One such example is the use of a regression tree to estimate the mean air temperature at which ice cover would become intermittent, i.e. it would no longer freeze every year (Burger, 2018). Sharma *et al.* (2019) estimated this value to be 8.4°C (Figure 1). However, whether a lake freezes or not is not dictated by the mean annual air temperature. Instead, it is more likely to be determined by the frequency and length of periods where temperatures are below 0°C as well as the severity of such temperatures. In fact, a lake can have intermittent ice cover even with a mean annual temperature below 0°C (as in the case of Lake Clark, US, where the average is −0.56°C; see Figure 1) or can freeze every year with a mean air temperature above 10°C (as in the case of Mohansic lake, US, where the average is 11.02°C; see Figure 1). Given the lack of a hard threshold between lake types in Figure 1, it seems rather unlikely that 8.4°C is an important threshold in determining whether a lake will freeze each year or not.

A second example is the analysis of freeze and thaw dates via linear models (e.g. Shuter, Minns & Fung, 2013) or segmented (linear) regression models (SRMs, referred to as continuous segmented regression in Sharma *et al.*, 2016). Although these models can estimate the change in dates over time, their estimates may be unreliable

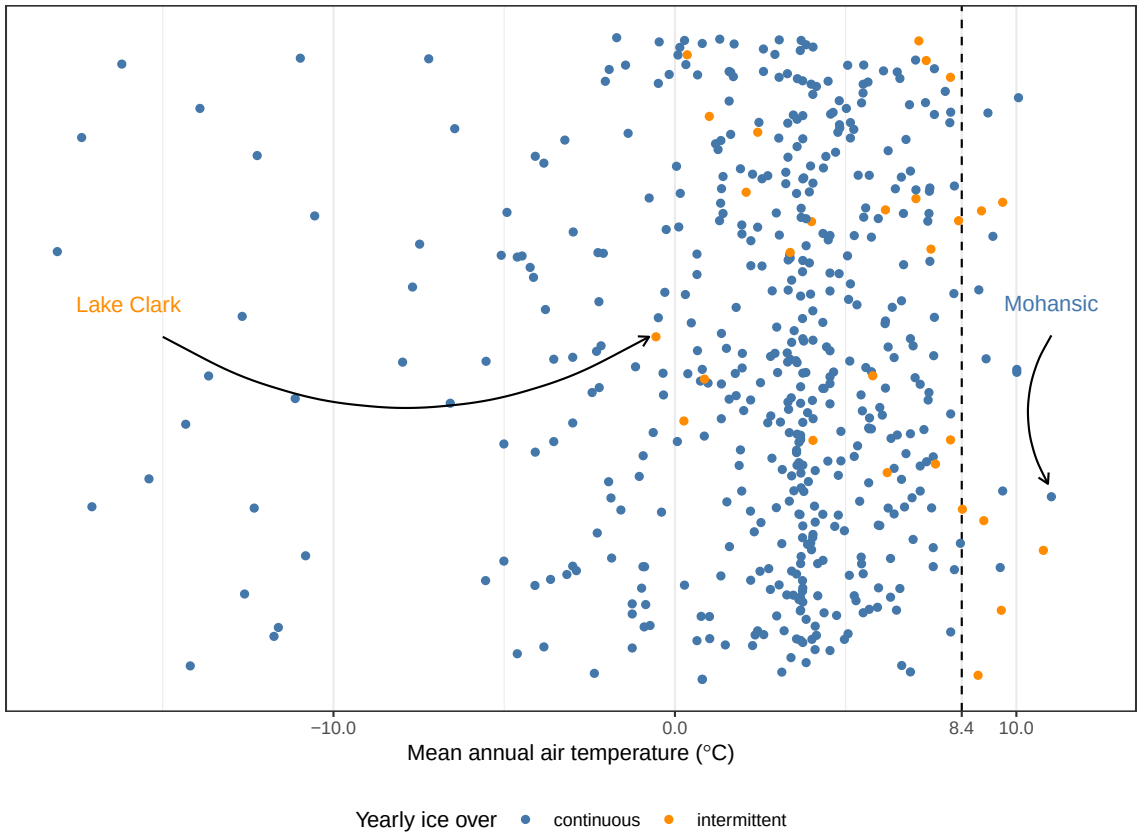

Figure 1: Reconstruction of the first branch of the regression tree fit by Sharma *et al.* (2019). The tree estimates that mean annual air temperature explains the largest proportion of the distinction between lakes that freeze every year (blue) and lakes that do not freeze regularly (orange). The dashed line indicates the split point. Of the lakes with continuous ice cover, Mohansic lake had the highest mean annual temperature (11.02° C), while Lake Clark had the lowest mean annual temperature among lakes without yearly ice cover (−0.56°C).

since the response (day of year) is cyclical and thus cannot be normally distributed. Furthermore, the relationships between the response and the covariates are often non-linear, so smooth (rather than linear) functions of covariates should produce better results, whether they be nonparametric (e.g. LOWESS) or semi-parametric (e.g. Generalized Additive Models).

In this project, I estimate the change in lake ice occurrence and duration since 1950. More specifically, I am interested in estimating the spatio-temporal changes in three main aspects of lake ice phenology: (1) the first day of complete freezing in each year (if any), (2) the date of the last thaw in each year, and (3) the duration of the period of ice coverage. I analyzed the freeze and thaw dates using a time-to-event approach, which allows me to estimate the probability of a lake freezing (thawing) on a given day. Freeze and thaw dates could also be analyzed with more common regression methods, but choosing an appropriate distribution for the response (i.e. day of year) would be more difficult, since the distribution would have to be circular, such that the day after the last day of each year coincides with January 1^{st} of the following year. A time-to-event approach allows us to perform a regression on the hazard of an event using more commonly-used and accessible methods (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. However, the probability of a lake being frozen at time t is equal to the probability 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 risk 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 then compare hazards from time periods of different lengths Δt. 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. $P(T \leq t) = 0.5$) when the hazard is equal to $\lambda(t) = 1.177$ (and thus $\log[\lambda(t)] = 0.1633$, see Figure 2).

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 at tiemes t_i and t_{i+1} but not necessarily constant in the interval $(t_i, t_{i+2}]$, i.e. they assume that the hazard rate is piecewise-constant between 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 (see Bender *et al.*, 2018). PAMs use an offset term equal to the log-transformed time between observations to standardize $\log\left[\hat{\lambda}(t)\right]$ to a common time interval Δt. From a survival analysis perspective, Δt might be the periods between doctor's visits to test whether a patient has recovered or not. It is more likely that the patient recovered between visits rather than during the doctor's visit, and a patient's chance of healing is likely affected by the length of Δt.

Before fitting a PAM, it is important to convert the data set 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 an offset and a binary indicator variable which is equal to 0 if the event has not yet happened and is equal to 1 if the event occurred in the interval $t_i + \Delta t_i$ (Bender *et al.*, 2018). An example of a PED structure is given in section 2.3.

To estimate the change in $\log\left[\lambda_{freeze}(t)\right]$ and $\log\left[\lambda_{thaw}(t)\right]$ since 1950 through-

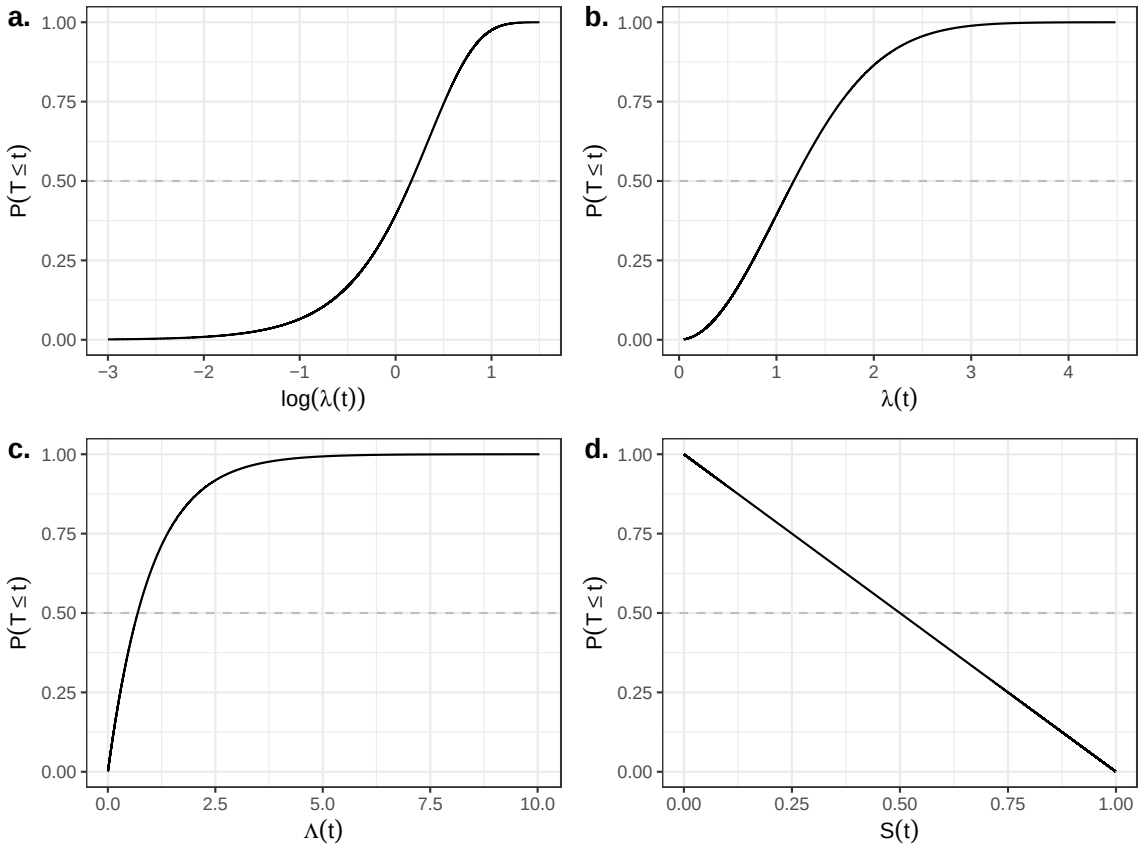

Figure 2: Cumulative probability of an event occurring before time t for given values of log-hazard $\log[\lambda(t)]$, hazard $\lambda(t)$, cumulative hazard $\Lambda(t)$, and survival $S(t)$. Note that $P(T \leq t) = 0.5$ when $\log[\lambda(t)] = 0.163$, $\lambda(t) = 1.177$, $\Lambda(t) = 0.693$, and $S(t) = 0.5$.

out the Northern hemisphere, I fit PAMs to a large ice phenology data set, and I accounted for spatio-temporal trends using covariates of geographic location (space) and year (time). In addition, I used a hierarchical Bayesian approach (Pedersen *et al.*, 2019) to incorporate a part of the unaccounted variation between lakes in the model. Hierarchical Piecewise-exponential Additive models (HPAMs) allowed me to model the change in $\lambda(t)$ for multiple lakes at once using a total of four models for $\lambda_{freeze}(t)$ and $\lambda_{thaw}(t)$ in North America and Eurasia. The aim of this thesis is to introduce a new time-to-event approach for analyzing ice phenology using statistical methods that have become more accessible in the last few years but are still not commonly used.

2 Methods

2.1 Lake ice data sets

The data for Buffalo Pound lake were previously analyzed by Finlay *et al.* (2019), while the northern hemisphere lake 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 split into two parts: North American lakes and Eurasian lakes. The GLRIPD was filtered to only include lakes with available coordinates and observations starting after 1950. Although the data set spans over 500 years, only two lakes have records prior to the year 1800, while the majority (51.4%) of the records started after 1950 and have fewer than 50 observations (Figure 3). Although the following analysis can be performed for the entire data set, the choice to only analyze data from 1950 onwards was made to decrease model fitting time and potential sampling bias, since the spatial distribution of the available lakes changes substantially over time and may confound trends. 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, especially for North America (Figures 4 and 5). All observations in the data set 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. Thus, while December 31^{st} and January 1^{st} are the 365^{th} and 1^{st} days of the year (in non-leap years), they (always) occur 184 and 185 days after June 30^{th} and 92 and 93 days after September 30^{th}, respectively (Figures 6, and 7). The converted values ranged from 55 days to 322 days for freezing and 44 days to 337 days for thawing (Figure 7).

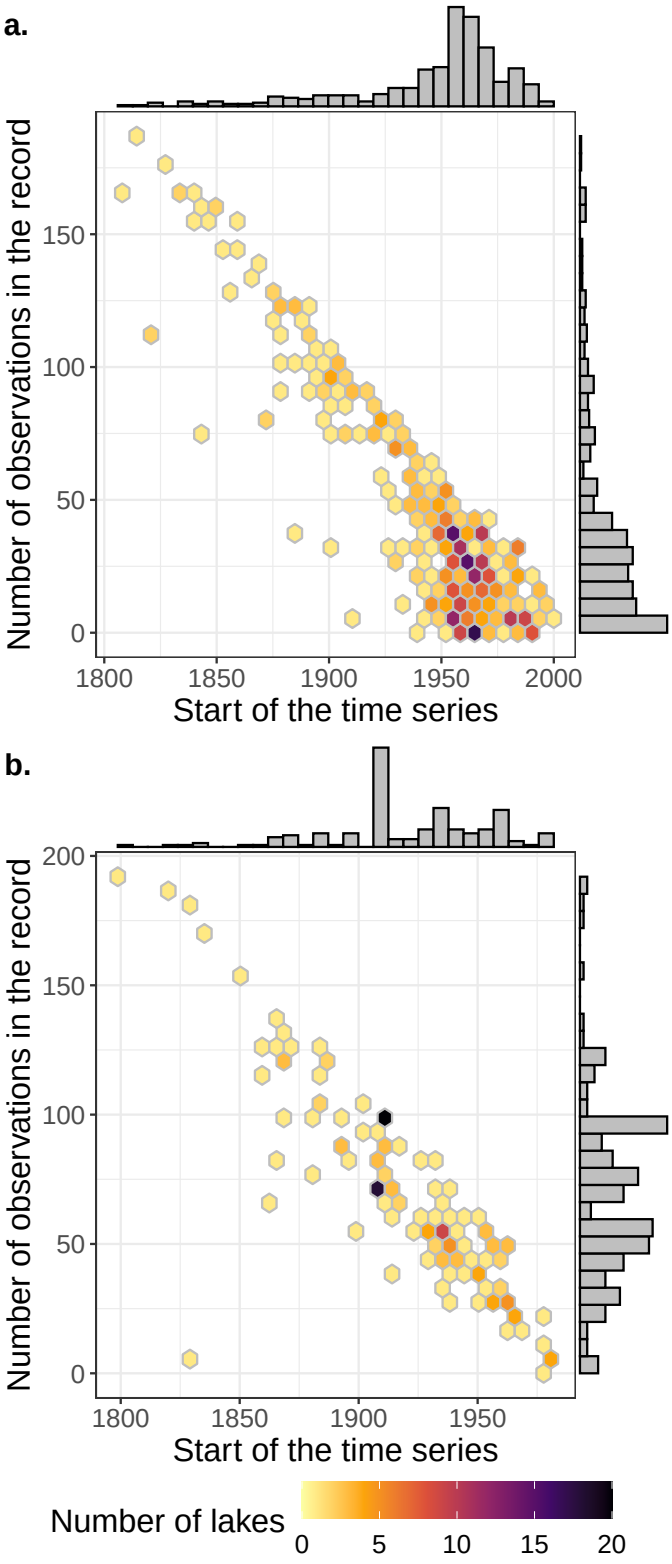

Figure 3: Number of lakes in the final North American (a) and Eurasian (b) data sets for a given record length and starting year; the marginal histograms are relative to the axis they are opposite to.

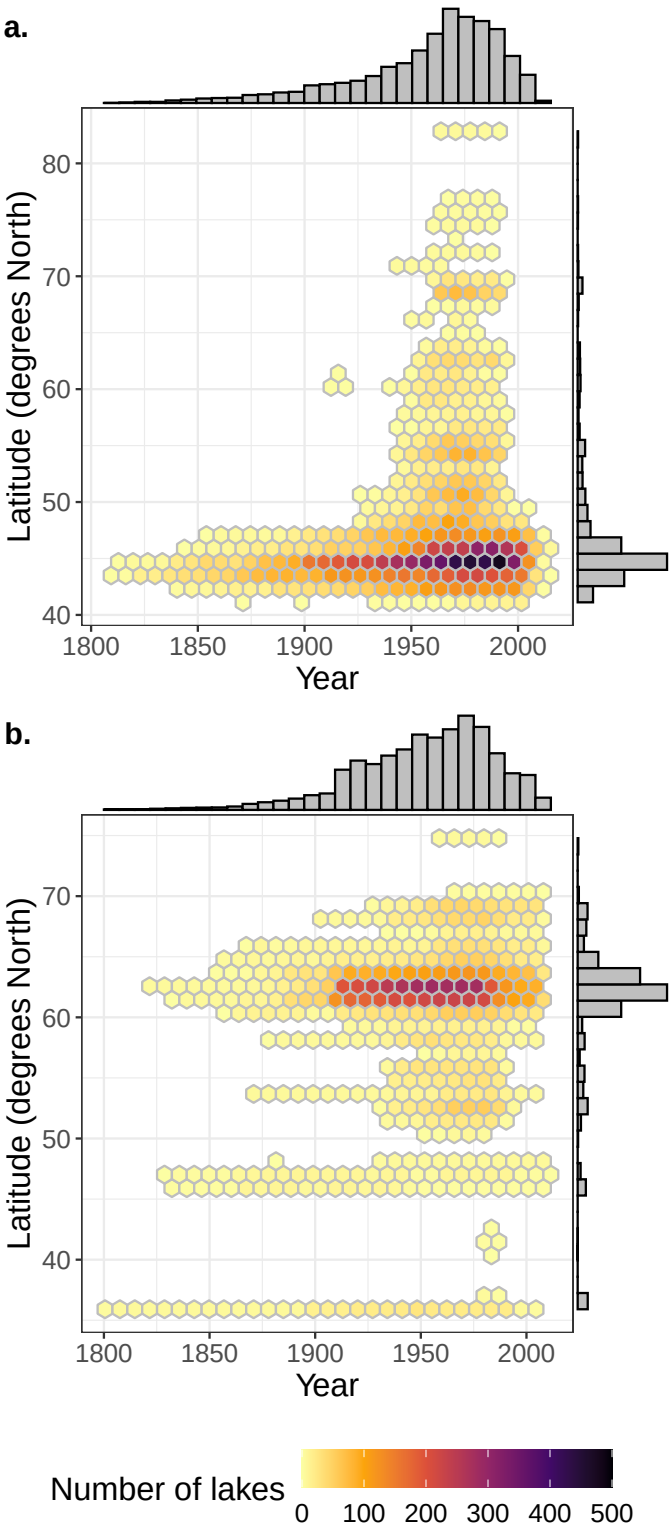

Figure 4: Number of observations in the final North American (a) and Eurasian (b) data sets for a given latitude and year; the marginal histograms are relative to the axis they are opposite to.

Figure 5: Polar projection of the lakes that in the final data set. The dashed lines indicate the 45° N and 62° N parallels.

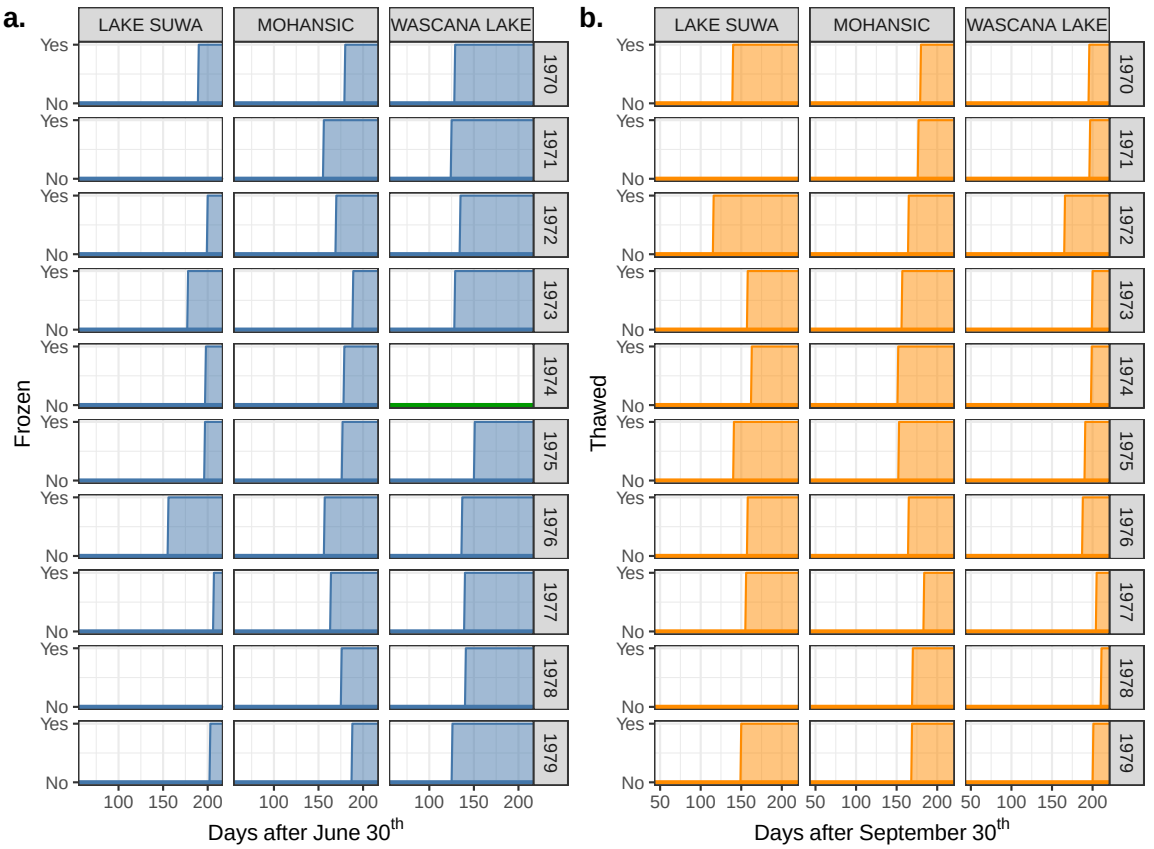

Figure 6: Freeze (a) and thaw (b) events for three lakes in the final data set for years 1970-1979. Shaded areas indicate when the lake was frozen (a) or ice-free (b); the green baseline for Wascana lake in 1974 indicates that the lake froze, but the date of freezing is unknown. Note that lake Suwa did not freeze in 1971 and 1978.

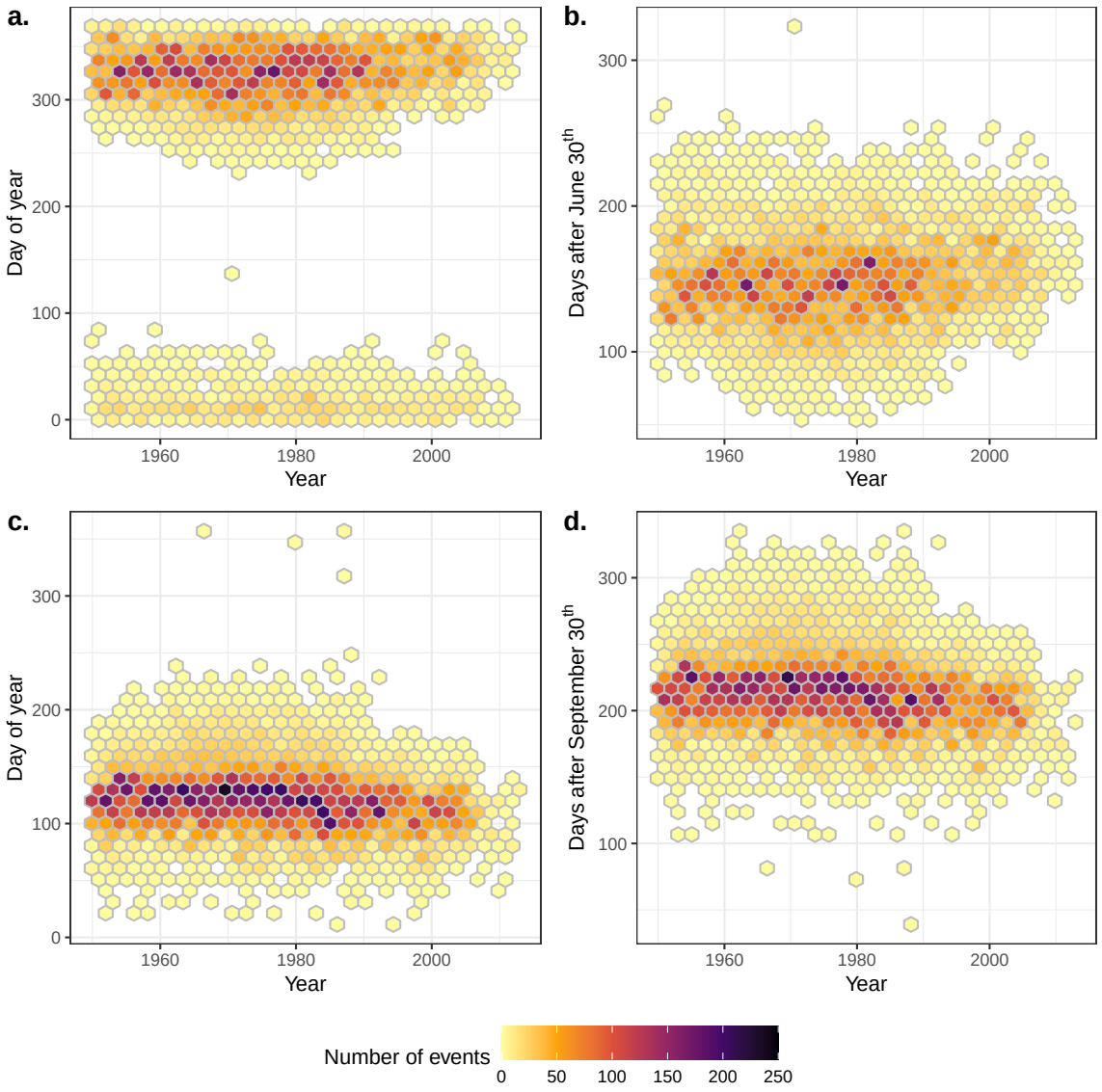

Figure 7: Number of freezing (a, b) and thawing (c, d) events on a given day. Panels (a) and (c) indicate the day of year of the event, such that January 1^{st} is 1. Plot (b) indicates the days of freezing as the number of days after June 30^{th}, such that July 1^{st} is 1. Plot (d) indicates the days of thawing as the number of days after September 31^{st}, such that October 1^{st} is 1.

The coordinates of the lakes were then checked using Google Maps. The coordinates of a lake were modified 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. I changed the names of lakes if there were multiple sets of observations that belonged to the same lake but had different names (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).

All of the data processing was performed on in `R` (R Core Team, 2020), and the script is available in the GitHub repository as `data/freezng-dates.R` (see Appendix ii). The final data set 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 ii). Years in which a lake did not freeze were excluded from the thaw data set (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; 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) and `brms` (version 2.12.0; Bürkner, 2017). The SRMs for the freezing dates of lake Suwa were fit using the `segmented` package (version 1.1-0; Muggeo, 2003). All plots were generated using `ggplot2` (version 3.3.0; Wickham, 2016) and `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 Single-lake models: Buffalo Pound Lake

To illustrate the models used in this project, I initially explore fitting PAMs to the dates of ice-on and ice-off for Buffalo Pound Lake (Canada). Buffalo Pound lake is a natural, shallow, polymictic lake in southern Saskatchewan, Canada, with rare, weak thermal stratification (Finlay *et al.*, 2019). Observations of ice coverage were taken on a weekly basis starting in 1979 until 2014. The PED data set had a structure of the form:

id	tstart	tend	interval	offset	ped_status	Year	on.date	season
1	0	120	(0,120]	4.79	0	1979	1979-12-01	1979-1980
1	120	121	(120,121]	0.00	0	1979	1979-12-01	1979-1980
1	121	122	(121,122]	0.00	0	1979	1979-12-01	1979-1980
1	122	126	(122,126]	1.39	0	1979	1979-12-01	1979-1980
1	126	127	(126,127]	0.00	0	1979	1979-12-01	1979-1980
1	127	129	(127,129]	0.69	0	1979	1979-12-01	1979-1980

The columns `tstart` and `tend` are the beginning and end of intervals for which the hazard function is estimated; the intervals are indicated in the `interval` column and are recorded as the number of days after June 30^{th}, such that July 1^{st} is 1. Like with the GLRIPD, the dates were converted to the number of days after June 30^{th} to avoid the discontinuity that would have occurred if using the day of year (Figure 8). The `offset` variable is the log-transformed number of days within each `interval`; PAMs use this term to normalize the hazard of waiting periods longer than the smallest waiting time. Since the smallest interval in this data set is of one day, the PAM will estimate the hazard of freezing at a daily level, so all one-day intervals have an offset of $\log(1) = 0$. In contrast, longer intervals have non-zero offsets since the hazard of freezing within these periods is higher than that of freezing on a single

day. The `ped_status` column indicates whether the lake is frozen (1) or not (0). `Year` is a covariate indicating the reference year, while the `id` column indicates which observations come from the same sampling station and year. Finally, the `on.date` column indicates the date on which the the lake froze, while `season` indicates the year starting on July 1^{st}, similarly to the definition of an academic year.

2.3.1 Piecewise additive models (for freeze/thaw)

The general PAM has the following structure:

$$\log[\mathbb{E}(Y)] = f_1(tend) + f_2(year) + f_{1,2}(tend, year), \tag{7}$$

where Y is the status of the lake (i.e. 0 if the event has not occurred and 1 if the event has occurred, whether freezing or thawing), $f_1(tend)$ is the smooth of the number of days after the reference date (June 30^{th} for freezing and September 30^{th} for thawing), $f_2(year)$ is the smooth of year, and $f_{1,2}(tend, year)$ is the two-dimensional tensor product interaction smooth between the smooths of *tend* and *year*. The smooth of *tend* allows the model to estimate the change in hazard within each year, the *year* smooth allows the model to estimate the change in hazard between years, and the tensor product interaction term allows the *tend* term to vary over the years (and vice-versa, although it is harder to visualize). In `R` the model was coded as:

```
gam(ped_status ~
      s(tend, bs = 'cr', k = 10) +
      s(Year, bs = 'cr', k = 10) +
      ti(tend, Year, bs = 'cr', k = 5),
    data = bp.ice.freeze,
    family = poisson(link = 'log'),
    offset = offset,
    method = 'REML')
```

(Since the `pamm` function from the `pammtools` package is a wrapper function

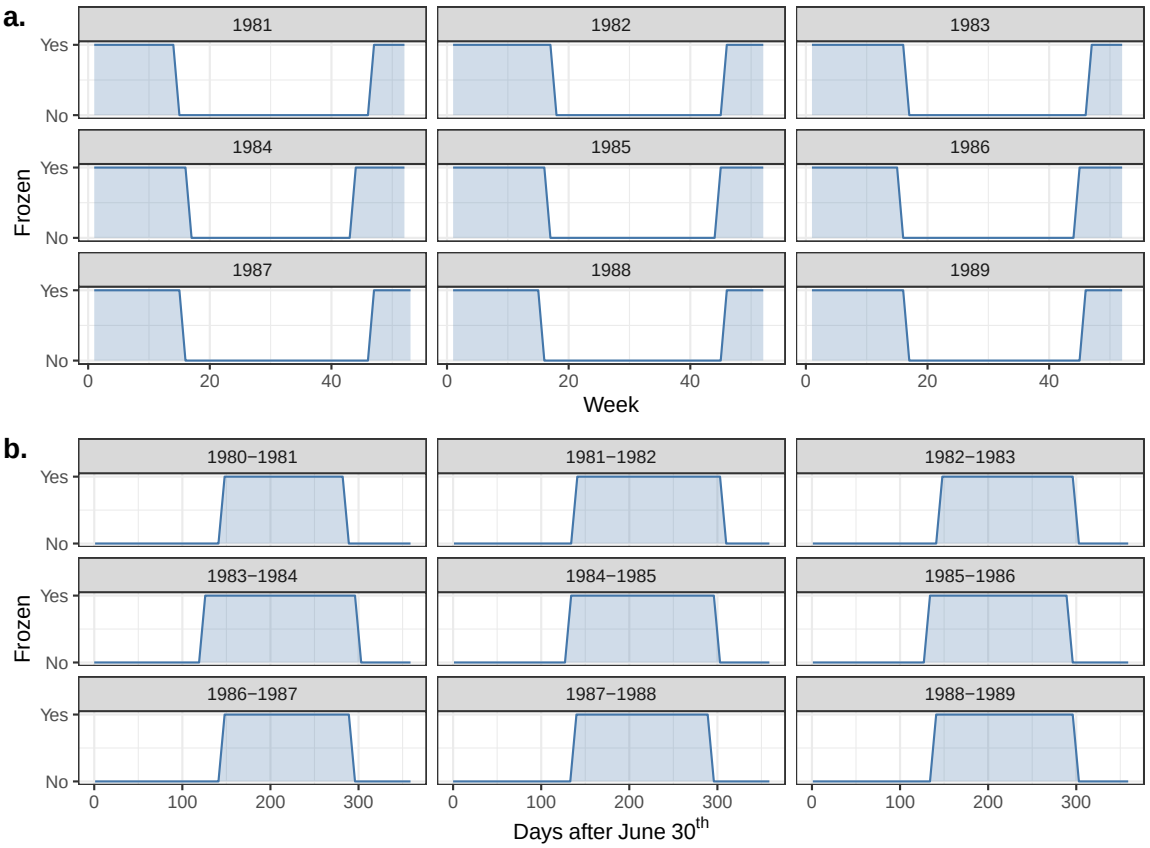

Figure 8: Periods of ice cover on Buffalo Pound lake (Canada), plotted against week (a) and days afer June 30^{th} (b). The shaded area indicates when the lake was frozen.

for the `gam` function from the `mgcv` package, one could use `pammtools::pamm` instead of `mgcv::gam`. `pammtools::pamm` pre-specifies `family = poisson()`, `offset = data$offset` and `method = 'REML'`. If one was interested in fitting a GAM to a large data set, the argument `engine = 'bam'` can be specified to use the `mgcv::bam` function instead.)

The arguments `bs = 'cr'` specify that the smooth terms are cubic regression splines, while the `k` argument indicates the dimension of the basis of the smooth term (i.e. the maximum complexity allowed for the term). Note that `k = 5` in the `ti` term indicates that each term in the tensor product can have up to $k-1$ degrees of freedom, so the maximum effective degrees of freedom for the tensor product with `k=5` is $(5-1)^2 = 16$. The `family` argument indicates that the response `ped_status` is conditionally Poisson-distributed given the data and the model, with a log link function. The `offset` argument is the $\log(\Delta t)$ offset used by PAMs to normalize $\hat{\lambda}_{freeze}(t)$ 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. A similar model was fit to the freezing dates of Lake Suwa (Japan) to compare PAMs to the segmented regression method used in Sharma *et al.* (2016).

2.3.2 Duration models

The duration of ice cover for Buffalo Pound lake was estimated using a GAM with a Gamma distribution and a log link function (Simpson, 2018). The model had the simple form

$$\log\left[\mathbb{E}(duration)\right] = f(year), \tag{8}$$

where the smooth of year $f(year)$ allows the model to estimate the change in

duration of ice cover over the years. In R the model was defined as:

```
gam(duration ~ s(Year, bs = 'cr', k = 10),
    data = bp.ice,
    family = Gamma('log'),
    method = 'REML')
```

2.4 Duration models: Lake Stechlin

Although GAMs with a Gamma distribution can estimate the change in ice cover duration over time, this is only the case for lakes that freeze annually, since the distribution only supports strictly positive numbers (i.e. $\{x \in \mathbb{R} : x > 0\}$. If the number of years with no ice cover in the data set is small, a GAM with a Tweedie distribution may lead to acceptable results, provided that the average duration is not too far from zero. This is not the case for the record from Lake Stechlin (Germany). The record has a low number of observations (35), and the majority of these (51.4%) are zeros, which resulted in a poor fit of the Tweedie GAM. Figure 9 shows that the assumption of normality of the residuals is violated and that the model estimates are biased towards over-estimating the duration.

In cases such as this where the proportion of zeros is high, a hurdle Gamma model (also known as a zero-altered Gamma model) might fit the data better. A hurdle gamma model assumes that the duration of ice cover is modelled by a gamma process that is dependent on (but distinct from) whether or not the lake froze, which is modelled by a binomial process. Since hurdle gamma models cannot (currently) be fit with `mgcv` package, the model was fit using the `brms` package (Bürkner, 2017). The `brms` package uses a fully Bayesian approach to model fitting, since it fits models using the `Stan` probabilistic programming language (Stan Development Team, 2015). `brms` fits models using the No-U-Turn Sampler algorithm (commonly referred to as NUTS, see Hoffman & Gelman, 2014), which is a special case of a Hamiltonian Monte

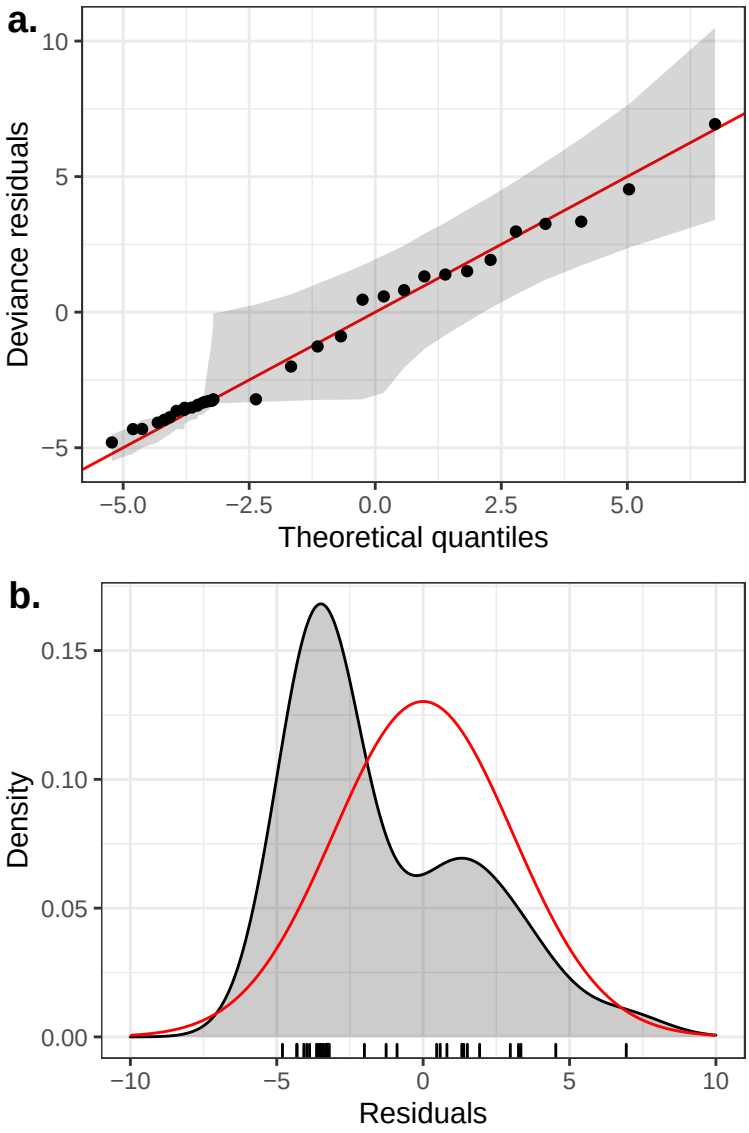

Figure 9: Diagnostics plots for the Tweedie GAM for the duration of ice cover for Lake Stechlin: quantile-quantile plot of the residuals of the model (a), with 95% credible intervals obtained via 1000 simulation runs, density of the model residuals (grey) with a zero-mean normal distribution with the same standard deviation (b). To meet the assumptions of normality, the residuals in (a) should be equally spaced and within the grey area, while the density function in (b) should be close to the normal density.

Carlo algorithm, a type of Markov Chain Monte Carlo (MCMC) algorithm. As such it is necessary to specify the number of MCMC chains and the number of iterations per chain that should be used to estimate the posterior distribution of the response and model parameters. The duration model for Lake Stechlin had a single predictor of `Year` and a constant hurdle probability, such that the probability of freezing was assumed to be constant over the years, and the duration of ice cover was Gamma-distributed conditionally on the data and the model, and given that the lake had frozen. The model had 4 chains with 4000 iterations each, and half of the iterations were used for warm-up. In `R`, the model had the following structure:

```
brm(bf(duration ~ s(Year, bs = 'cr', k = 10),
       hu ~ 1),
    family = hurdle_gamma(),
    data = stechlinsee,
    chains = 4,
    iter = 4000,
    control = list(adapt_delta = 0.99999, max_treedepth = 20))
```

Similarly to the definition of models in `mgcv`, the `s(Year)` function specifies the smooth of year with cubic regression splines and a maximum of $k - 1 = 9$ degrees of freedom. The `hu ~ 1` section of the formula indicates that the hurdle probability is constant. The `chains` and `iter` arguments specify the number of chains and iterations, respectively. By default, half of the iterations are used for the warm-up. The list passed to the `control` contains arguments for the NUTS algorithm (Hoffman & Gelman, 2014).

2.5 Segmented regression

In order to compare PAMs to the segmented regression method used by Sharma *et al.* (2016), a SRM was fit to freezing dates of lake Suwa (Japan). It should be noted that while Sharma *et al.* (2016) did not include some of the data due to changes

in the Japanese calendar, the dates were recently corrected in the GLRIPD, so the entire data set was analyzed here. The model was fit using the `segmented` package (Muggeo, 2008) using a single predictor of `Year` and a single estimated breakpoint for the change in slope. The model was fit in `R` using the following code:

```
lm.freeze <- lm(On.DOY.jul ~ Year, data = suwa)    # fit a LM
segm.freeze <- segmented(lm.freeze, seg.Z = ~ Year) # segment the LM
```

The prediction intervals for the SRM were obtained using the `predict.segmented` function from the `segmented` package. The prediction intervals for the PAM to which the SRM was compared could not be obtained with a single function because the PAM does not directly estimate the day on which the lake will freeze. Therefore, the estiamtes for the PAM were made by using the estimated probability of freezing for various days and years. The average freezing dates were estimated to be the days where the probability of the lake being frozen was equal to 0.5. Similarly, the 95% prediction intervals were determined using the days on which the probability of being frozen was approximately equal to 0.025 or 0.975 (Figure 10).

2.6 Hierarchical models

2.6.1 Hierarchical piecewise additive models

The hierarchical PAMs for freezing and thawing dates for North America and Eurasia had a similar structure to the individual PAMs, with the addition of two factor smooth interactions for seasonal trends and yearly trends for each lake, a spatial term, and tensor product interaction terms between each of the global smooths (i.e. all terms but the factor smooth). In `R` the models have a structure similar to the following:

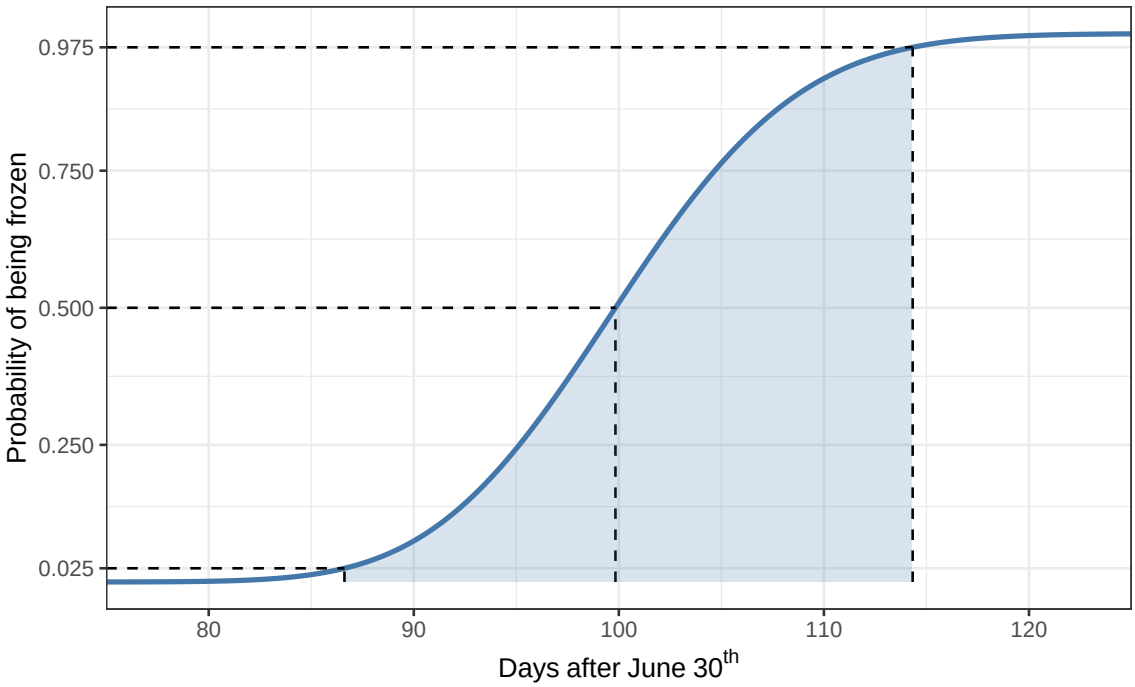

Figure 10: Fictitious example of how the expected freezing and thaw dates and relative 95% credible intervals were estimated from the PAMs and HPAMs. In this example the mean would be 100 since $P(T < 100) = 0.5$, and the 95% prediction intervals would be $[86.62044, 114.3264]$.

```
bam(ped_status ~
      s(tend, bs = 'cr', k = 10) +
      s(Year, bs = 'cr', k = 20) +
      s(tend, lake, bs = 'fs', k = 10) +
      s(Year, lake, bs = 'fs', k = 10) +
      s(lat, long, bs = 'ds', k = 20) +
      ti(tend, Year, bs = 'cr', k = c(4, 4)) +
      ti(tend, long, lat, bs = c('cr','ds'), d = c(1,2), k = c(4,5))+
      ti(Year, long, lat, bs = c('cr','ds'), d = c(1,2), k = c(4,5)),
    data = freeze.na,
    family = poisson('log'),
    offset = offset,
    method = 'REML')
```

The basis specification `bs = 'fs'` indicates that the smooth terms are factor-smooth interactions, which are a type of predictor that includes a smooth of the first variable (i.e. `tend` or `Year`) for each `lake` in the data. Note that the predictors of `tend` and `Year` which use a cubic regression spline basis (i.e. `bs = 'cr'`) estimate the common trends of within- and between-year variation, respectively, while the factor smooth interactions estimate the deviations of each lake from the global patterns of `tend` and `Year`.

The spatial smooth `s(long, lat)` accounts for the effect of geographic location. The argument `bs = 'ds'` specifies that the term uses Duchon splines, which 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). Finally, the tensor product interaction terms allow the model to account for the change in the effect of one predictor as the value of another varies. For instance, the `ti(tend, Year)` tensor product interaction models the change in the within-year trends over time. Similarly, the `ti(tend, long, lat)` and `ti(Year, long, lat)` terms allow us to estimate the change in within-year and between-year trends in hazard over space. Using `ti` terms we can, for example, estimate the change in freeze and thaw dates over the years, or if Arctic lakes are losing ice cover faster than temperate lakes.

Note that tensor product *interaction* terms, `ti`, are different from tensor product smooths, which are indicated with `t2`. While a `t2` smooth estimates the mean effect of two or more predictors and their interaction on the response, the `ti` terms only contain information on the interaction alone, so `ti` terms are the interaction between smooths where the main effect of each smooths has been removed. Thus, with `t2` terms we can jointly estimate the effect of multiple terms and their interactions on the response, while `ti` terms only estimate how the effect of one smooth on the response changes for different values of other smooths. In the case of the model above, the term `t2(tend, Year)` would estimate the trends in hazard within- and between-years. In contrast, `s(tend)` and `s(Year)` would estimate the *average* change in hazard within and between years, respectively. Finally, the term `ti(tend, Year)` would only estimate the *change* in between-year patterns over the years. Thus, fitting a model with `s(tend) + s(Year) + ti(tend, Year)` is approximately equivalent to fitting a model with the single term `t2(tend, Year)`.

2.6.2 Hierarchical duration models

The duration of ice cover was analyzed using three similar models of increasing complexity. Similarly to the HPAMs, each of these models included the same smooth predictors, namely a global year smooth, a year factor-smooth interaction term for each lake, a two-dimensional spatial smooth with Duchon splines, and a tensor product interaction term between year and geographic location.

Since the observations with zero days of ice cover were few for both North America (0.9%) and Eurasia (1.7%), a Tweedie GAM was fit to each data set. However, these models were unable to explain a substantial part of the variance in the duration, as shown by the Q-Q plots in Figure 11.

To account for the changes in variance over time and space, Tweedie Location, Shape, and Scale (TWLSS) models were also fit to the data sets. These models allow

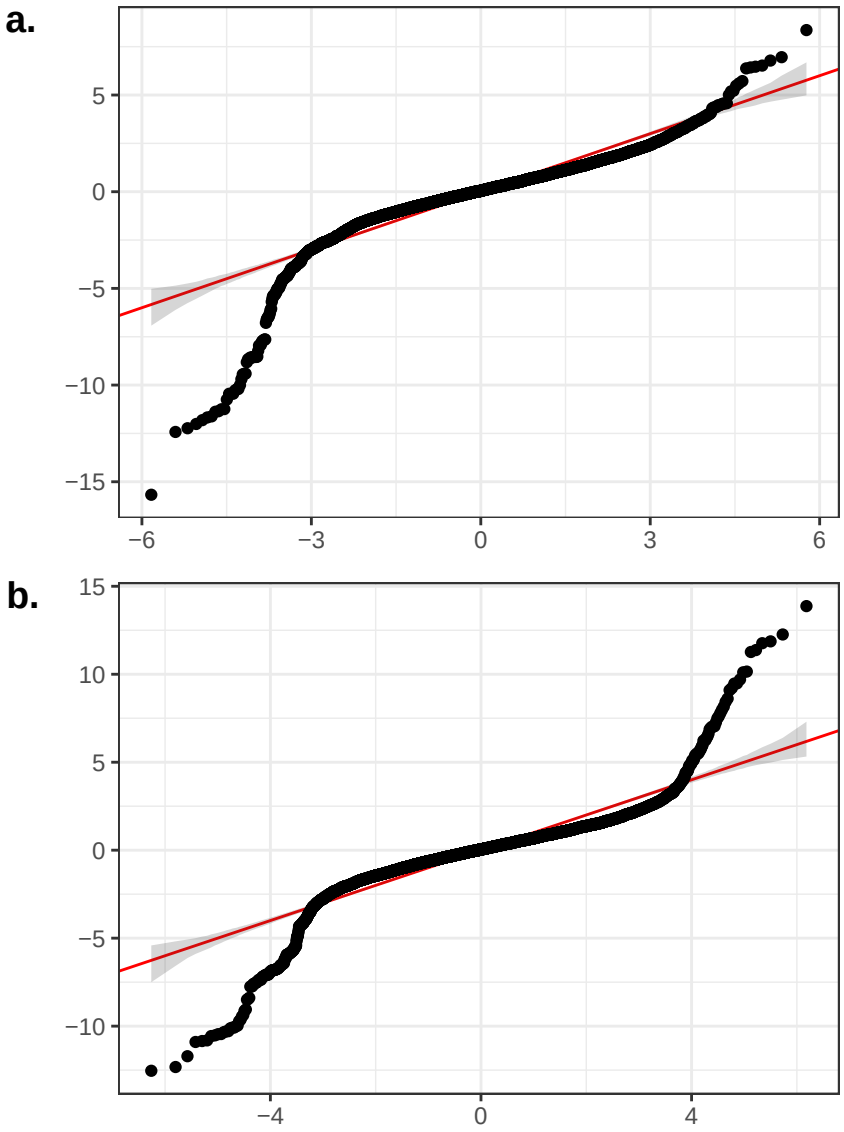

Figure 11: Quantile-quantile plots of the Tweedie GAMs fit to the duration of North America (a) and Eurasia (b). The shaded areas indicate the with 90% confidence intervals obtained via 1000 simulated draws from the distribution estimated by the model. Each simulated data sets had the same number of observations as the data set to which the model was fit.

the power and scale parameters of the Tweedie distribution to vary as functions of the covariates, thus allowing the models to account for changes in variance and the mean-variance relationship in the full conditional distribution of the response. The models were similar to the simpler Tweedie models, but also had a global year term, a two-dimensional spatial term, and a random-effects lake term for the power parameter and the scale parameter. Thus, both the variance and the mean-variance relationship were allowed to vary between lakes and over time and space. The models have the following structure in R:

```
gam(list(duration ~
           # formula for the mean
           s(Year, bs = 'cr', k = 20) +
           s(Year, lake, bs = 'fs', k = 10) +
           s(long, lat, k = 20, bs = 'ds') +
           ti(Year, lat, long, bs = c('cr', 'ds'),
              d = 1:2, k = c(10, 10)),

         ~ # formula for the power
           s(Year, bs = 'cr', k = 5) +
           s(long, lat, k = 5, bs = 'ds') +
           s(lake, bs = 're'),

         ~ # formula for the scale
           s(Year, bs = 'cr', k = 5) +
           s(long, lat, k = 5, bs = 'ds') +
           s(lake, bs = 're')),
    data = ice.eura,
    family = twlss(),
    method = 'REML')
```

3 Results

3.1 Single-lake models: Buffalo Pound Lake

The results for the PAMs fit to the Buffalo Pound Lake time series are shown in Figure 12. Although both the hazard of freezing and the hazard of thawing increased exponentially within each year, the hazard of freezing increased substantially faster than the hazard of thawing. Both hazards changed between years, but the change was not monotonic (Figures 12a-d). As one might expect, years with a flatter hazard curve had slower increases in the cumulative probability functions (e.g. $\hat{\lambda}_{thaw}(t)$ and $\widehat{F}_{thaw}(t)$ for 2014 in Figures 12b and 12f). The variances in freezing and thawing dates were $s^2_{freeze} = 77.54$ and $s^2_{thaw} = 95.81$. Overall, the average duration of ice cover in Buffalo Pound increased by approximately 9 days within the observation period ($p = 0.258$) but the duration model could only explain a very small portion of the variance ($R^2_{adj} = 0.00831$, see Figure 12g).

3.2 Duration models: Lake Stechlin

The Tweedie model for the duration of ice cover on lake Stechlin estimated that the duration decreased significantly prior to 1970 and remained short before increasing in length after 1980. In contrast, the gamma hurdle model shows that the duration of ice cover did not change, given that the lake had frozen (Figure 13). The slight concurvity of the smooths is in part due to the lower density of data near the tails of the models, which also causes the widening of the credible intervals.

3.3 Segmented regression

Figure 14 compares the average freezing dates of Lake Suwa estimated by the SRM (blue) with those estimated by the PAM (orange). There appears to be little disagreement between the estimates of the two models. Although there is some dif-

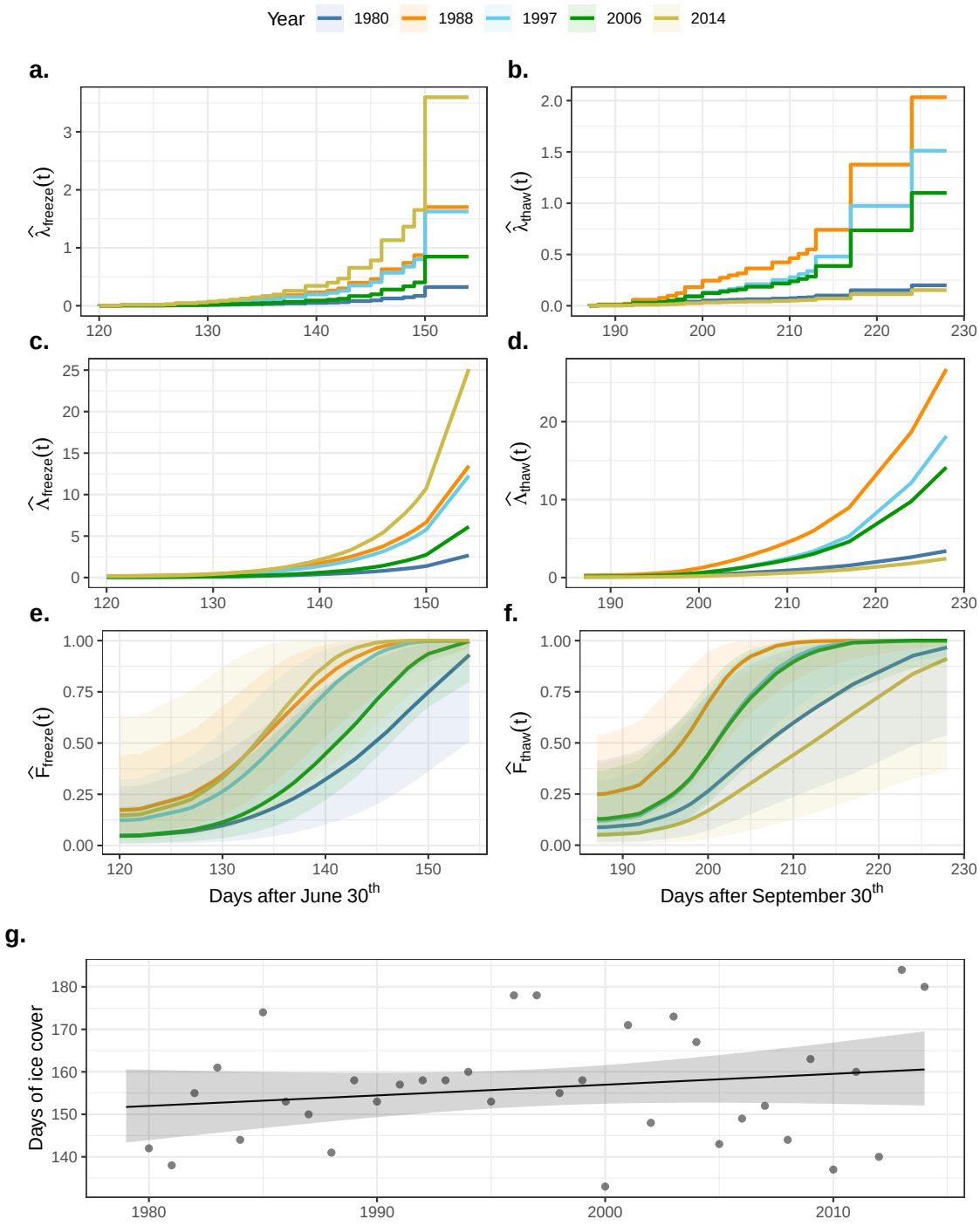

Figure 12: Estimated cumulative stepwise hazard (a, b), cumulative segmented hazard (c, d), and probability (e, f) of freezing $F_{freeze}(t)$ and thawing $F_{freeze}(t)$, respectively, for Buffalo Pound lake during various years. Average duration of ice cover on Buffalo Pound lake (g).

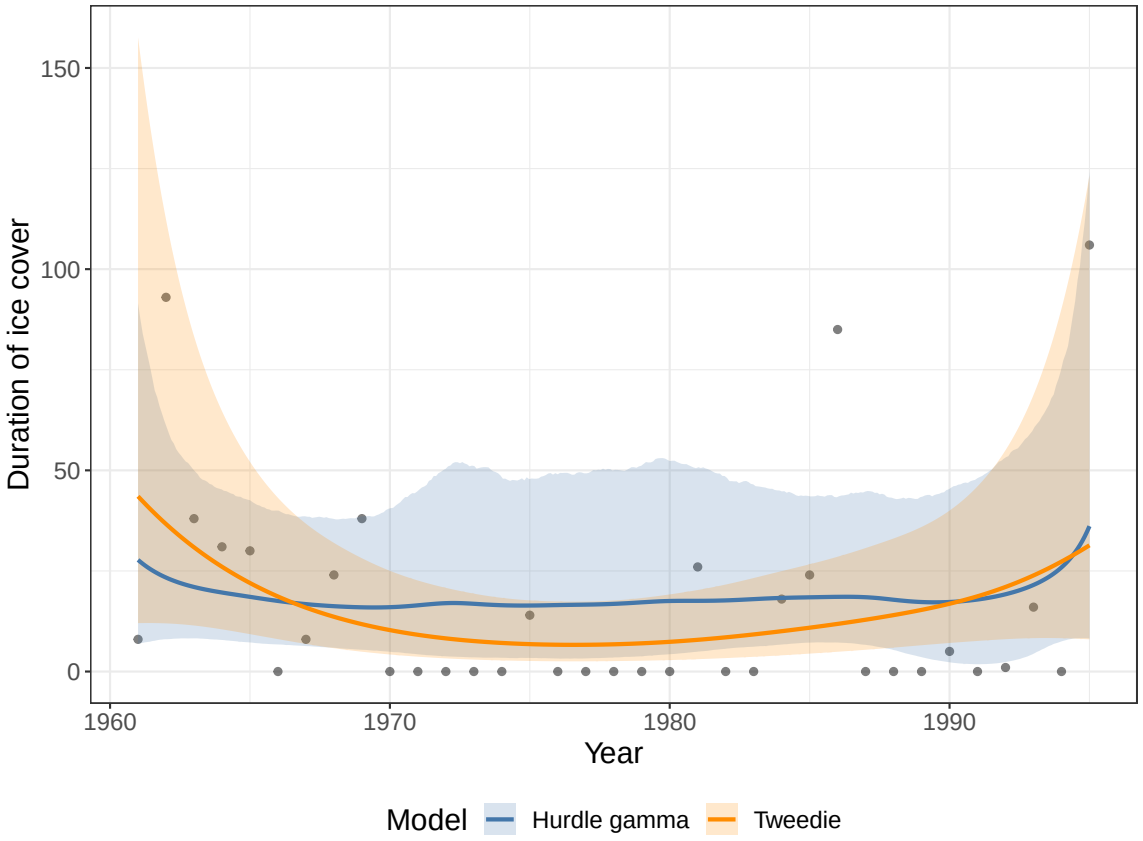

Figure 13: Tweedie (blue) and hurdle gamma (orange) models fit to the duration of ice cover for Lake Stechlin (Germany). The lines indicate the expected duration of ice cover, while the shaded areas are 95% credible intervals of the mean.

ference in the prediction intervals, this is mostly due to the differences in flexibility of the models, since the SRM assumes the effect of year to be (piecewise) linear, while the PAM allows the effects of year to be smoothly varying (Figure 14).

3.4 Hierarchical models

3.4.1 Hierarchical piecewise additive models

The results from the HPAMs fit to the North American and Eurasian data sets are shown in Figures 15 though 18. The cumulative hazard and cumulative probability of freezing for Eurasian have decreased since 1950. On average, Eurasian lakes in 2010 were expected to freeze 5.5 days later than in 1950. While $\widehat{\Lambda}_{freeze}(t)$ and $\widehat{F}_{freeze}(t)$ also decreased for North American lakes, the decrease after 1995 was much larger. On average, North American lakes in 2010 were expected to freeze 33 days later than in 1950. In addition, the estimated change in $\widehat{\Lambda}_{freeze}(t)$ and $\widehat{F}_{freeze}(t)$ in North America during 2010 is much slower than in the years 1950-1995, which indicates that on average the hazard of freezing was much lower farther into the year than in previous decades (Figure 15c).

The cumulative hazard and cumulative probability of thawing for Eurasian lakes has increased since 1950, such that the average thaw date for Eurasian lakes in 2010 is 5.5 days earlier than in 1950. In contrast, North American lakes in 2010 thawed approximately 10 days earlier than in 1950, on average (Figure 15).

The small change in the spatial smooths of the HPAMs over the years (not shown) indicate lakes in Eurasia are following the global trend closely, while the North American HPAMs indicate that more northern lakes are freezing later and thawing earlier than more southern lakes, when compared to dates in 1950. However, the absence of lakes at high latitudes in the data set in the last two decades prevents a confident estimate. Similar considerations could be made for the factor smooth interactions, although the narrow spread and the high smoothness of the `fs` terms indicate that

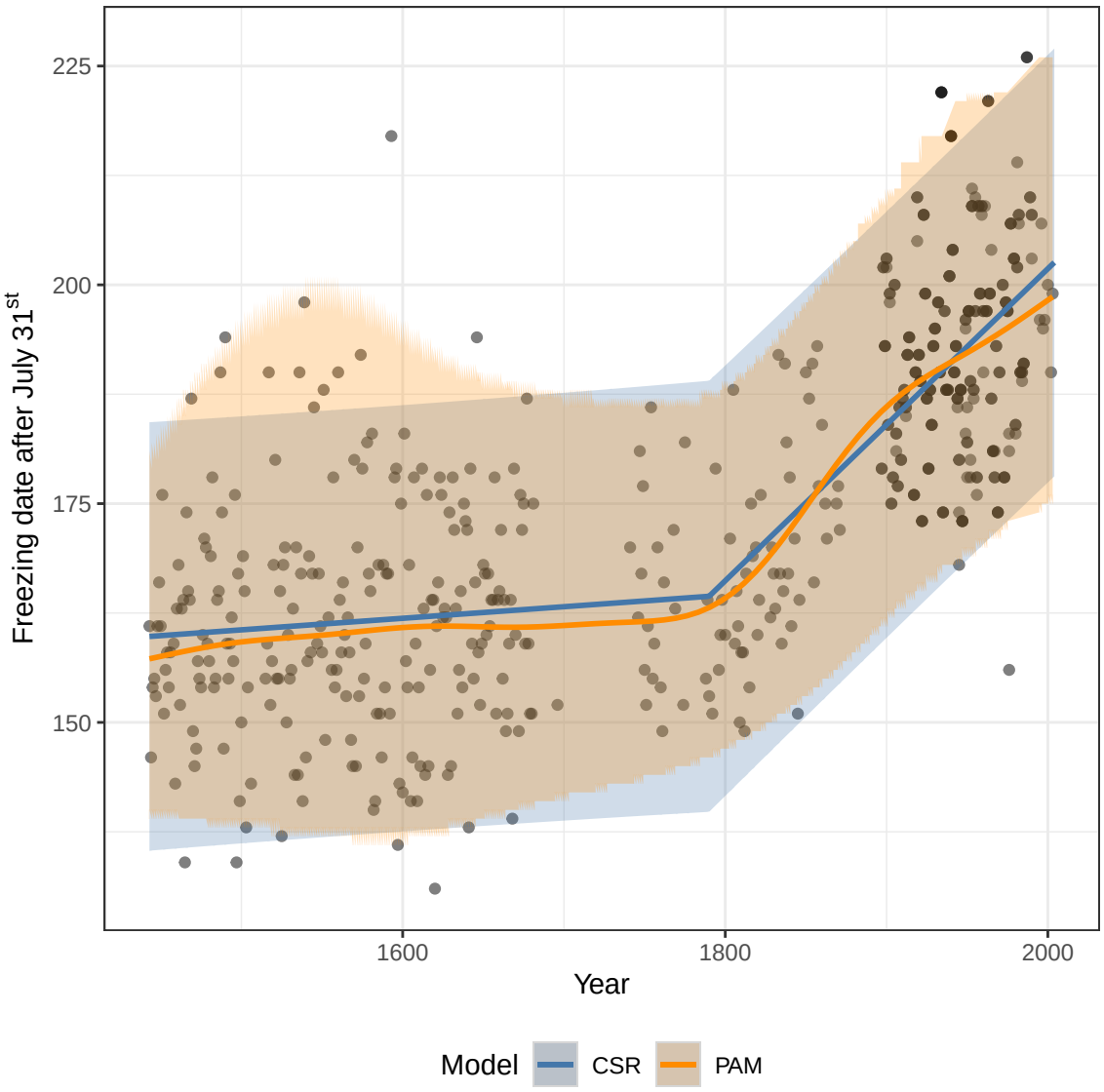

Figure 14: Estimated average freezing date for lake Suwa, Japan via a SRM (blue) and a PAM (orange), with 95% prediction intervals. The values from the PAM are estimated, since the response in the PAM is the hazard of freezing and not the day of freezing.

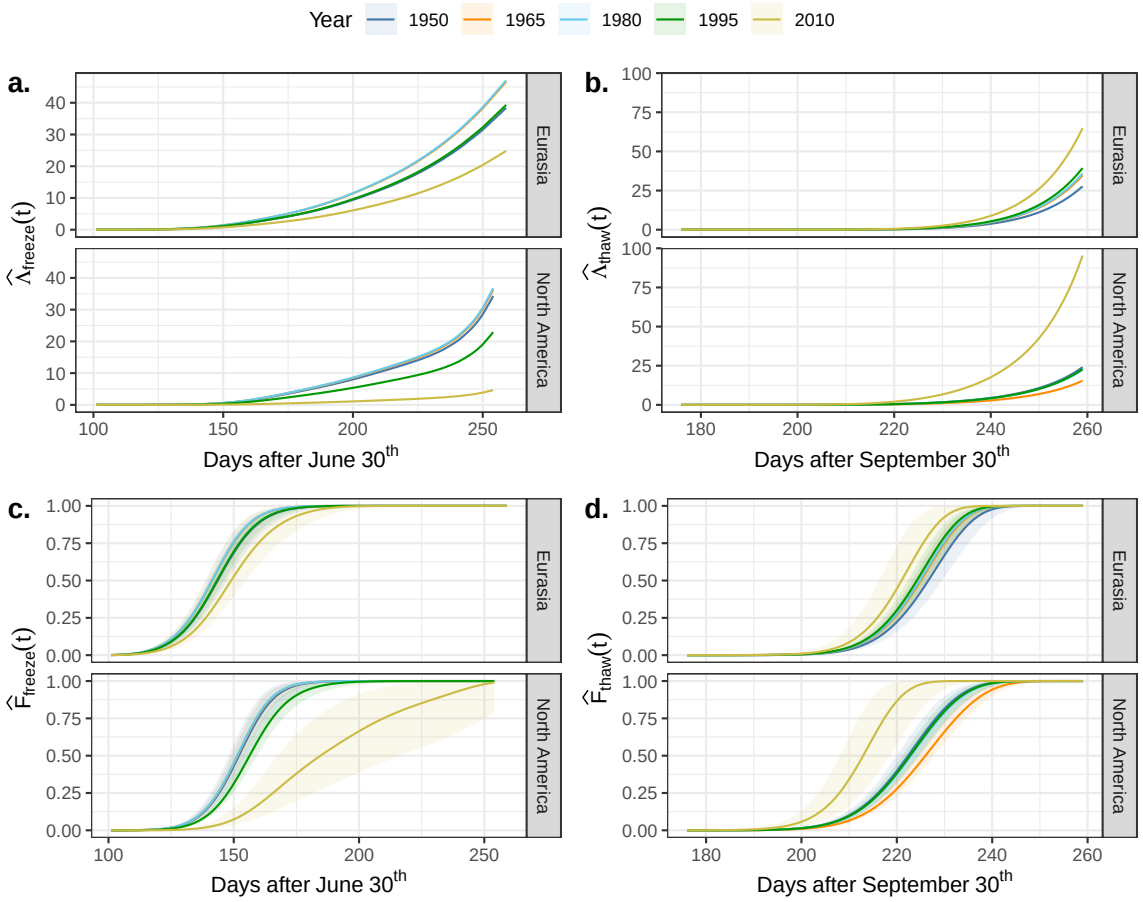

Figure 15: Estimated cumulative hazard (a, b) and cumulative probability (c, d) of lakes being frozen (left) or thawed (right).

there was very little deviation from the global `Year` term in the four HPAMs (Figure 16).

Due to the lower amount of data following 2005, predictions for the change in spatial patters over the years were only produced for areas with sufficient data, namely the the Great Lakes area (Figure 17) and Northern Europe (Figure 18). Overall, the two regions had a shift towards later freezing dates by 10-20 days, but the thaw dates only changed substantially in the southern and more continental areas.

3.4.2 Hierarchical duration models

As one might expect, the duration of ice cover was longer for lakes at higher latitudes and locations with continental cold climates. Lakes closer to the coast (e.g. Japanese and central European lakes) tended to have a shorter duration, while the most northern lakes were frozen for the majority of the year, if not continuously. The change in duration of ice cover over the years was substantially different between the two continents. In North America, between 1950 and 2010, the duration of ice cover on lakes south of 65°N decreased by 15 days on average, while it increased by 5 days on average for lakes north of 65°N. The duration decreased for central and western lakes, with greater decreases ($\sim$ 30 days) occurring for the more western lakes, while it increased by 10-25 days on average for lakes east of 75°W.

In Eurasia, the duration of ice cover decreased severely throughout the entire continent, with changes ranging from -50 days at lower latitudes (~30°N) and as high as approximately -140 days in the most northern lakes. The change in duration also varied from longitudinally, with a mean decrease of approximately 50-60 days in Europe and as high as 100-110 days throughout central Asia (60-130°E). The lowest decrease in duration occurred in areas with low average duration in 2010 was estimated to occur in Japan and south-eastern Europe (Figure 19). Overall, the mean duration of ice cover decreased from a range of $[27.5, 366]$ days and a variance of 4557

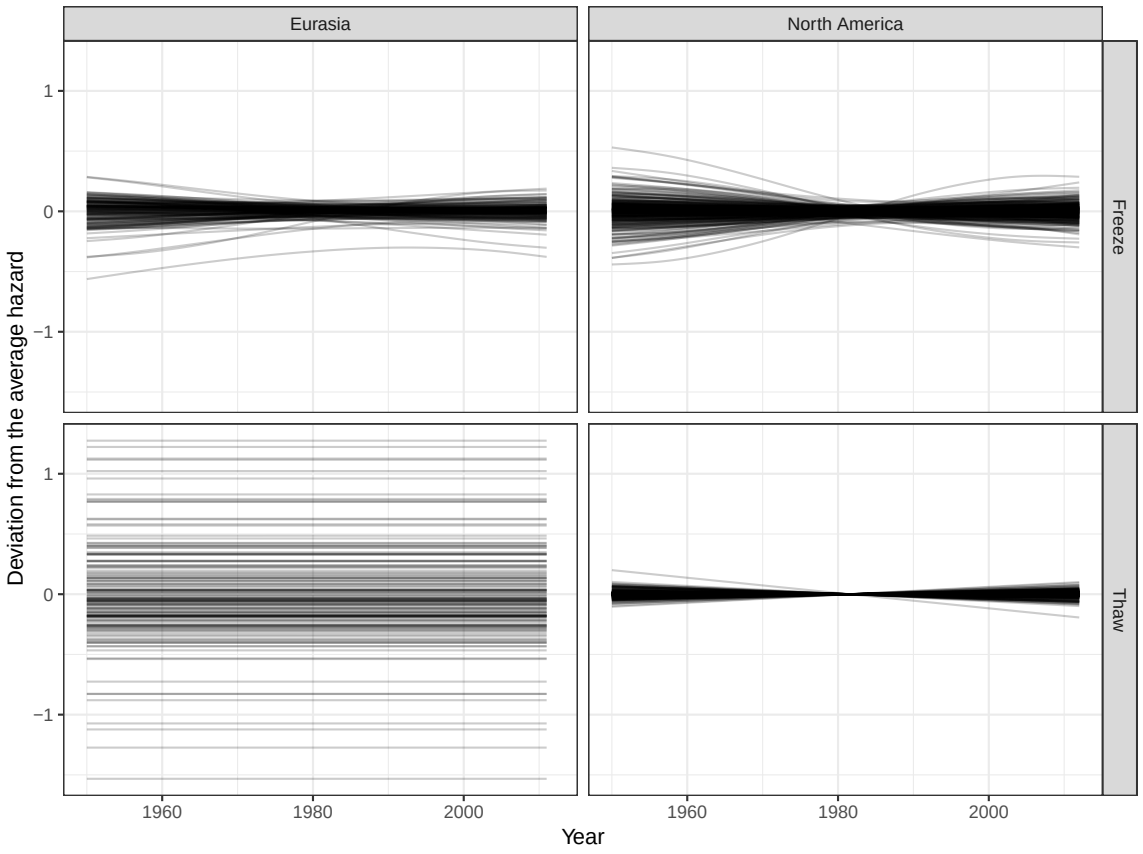

Figure 16: factor smooth interactions from the HPAMs for the hazard of freezing for lakes in North America and Eurasia, and the hazard of thawing for lakes in North America and Eurasia. The y axis indicates how much the hazard of each lake deviated from the mean trend.

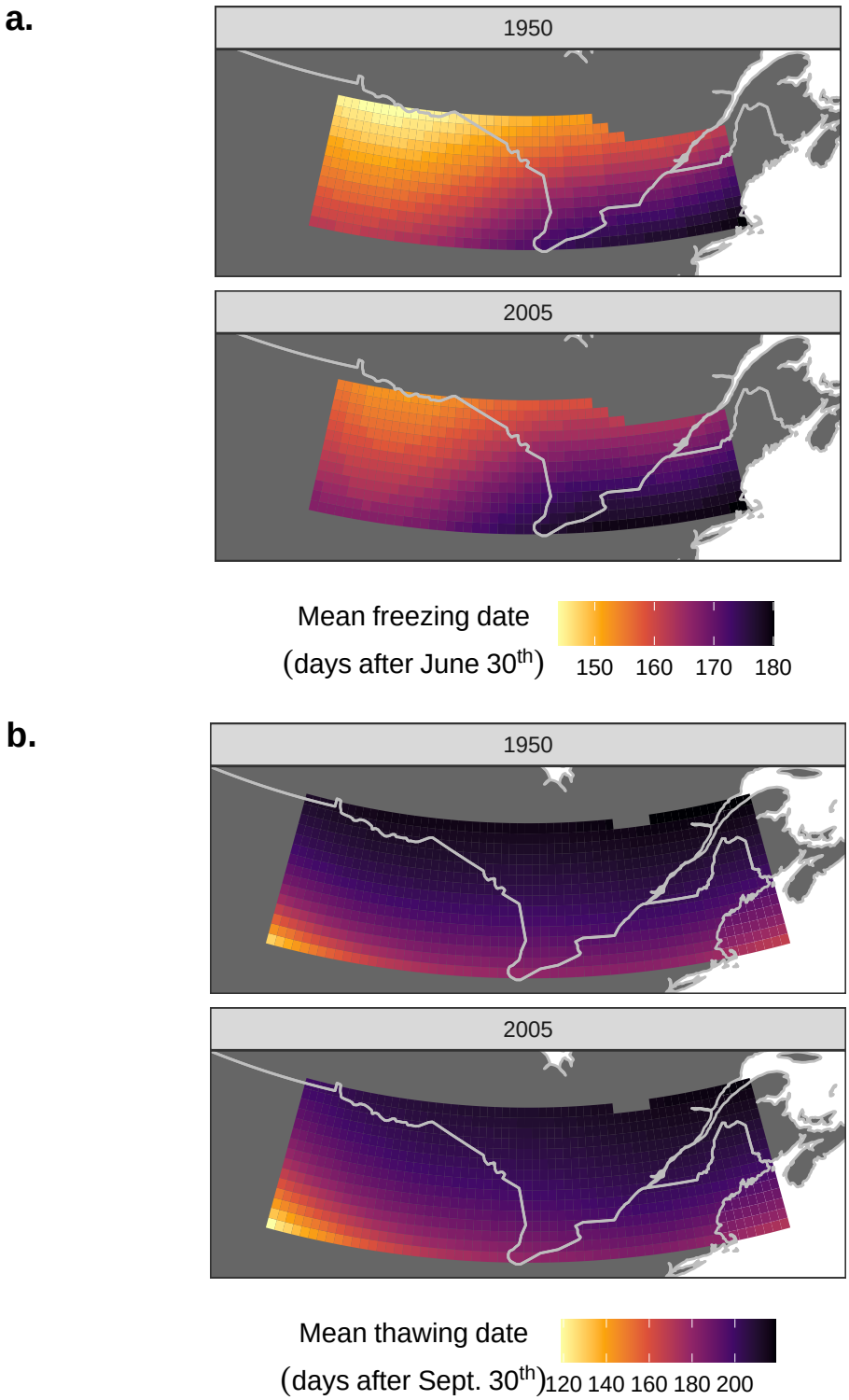

Figure 17: Change in the average dates of freezing and thawing for lakes in the Great Lakes area (North America) between the years 1950-2010.

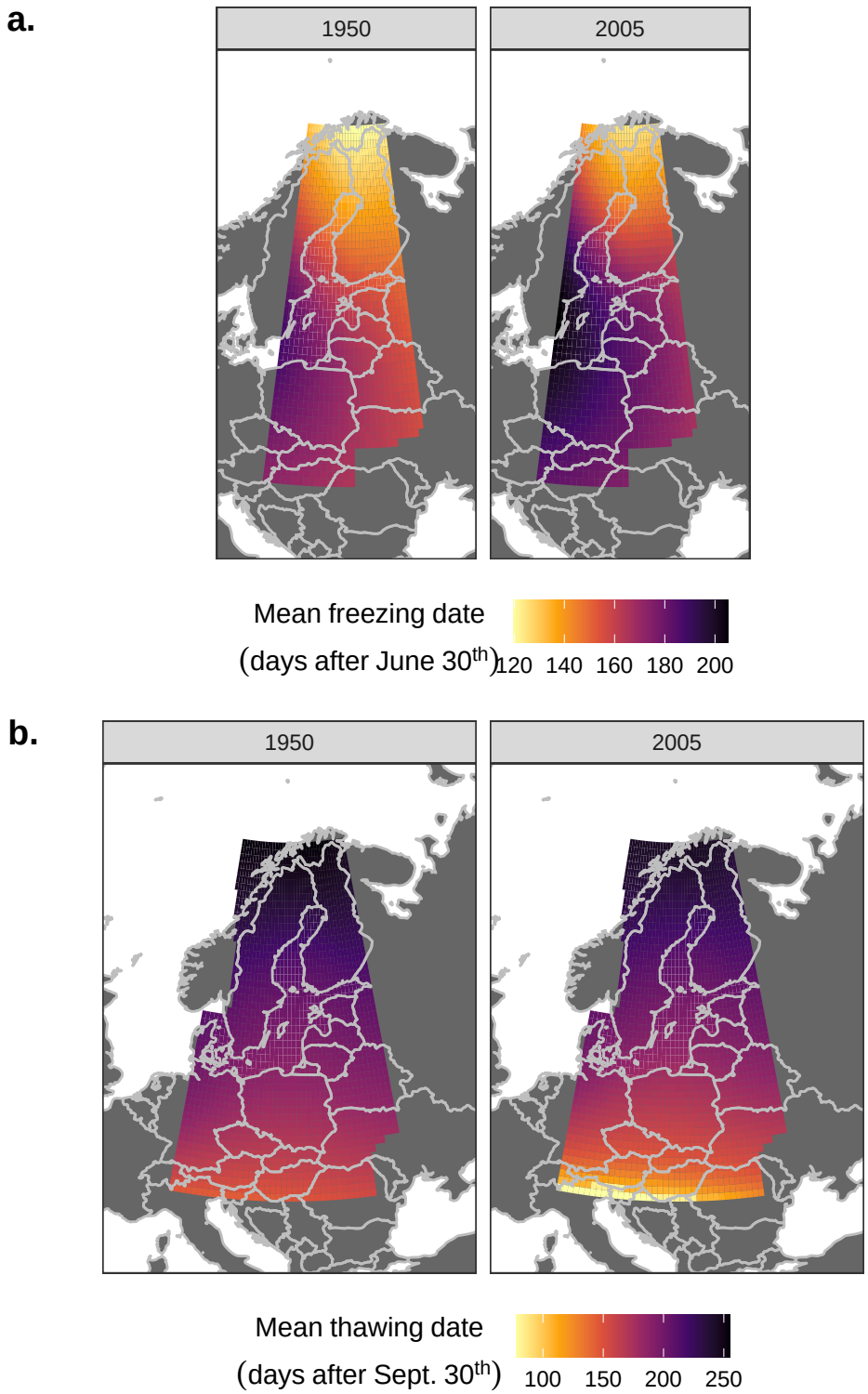

Figure 18: Change in the average dates of freezing and thawing for lakes in Northern Europe between the years 1950-2010

in 1950 to a range of $[12, 185.8]$ and a variance of 1294. Thus, many lakes that used to be frozen year-round in 1950 may now only freeze for approximately half a year, and the spatial variance within the continent has decreased significantly within the last 70 years.

While the HGAMs were able to account for the effects colder climates in more continental and northern locations, as evidenced by the longer estimated duration of ice cover in central Asia (e.g. central Russia, China, Mongolia), they tended to predict more accurately for lakes that did froze annually and had longer periods of ice cover (Figure 20).

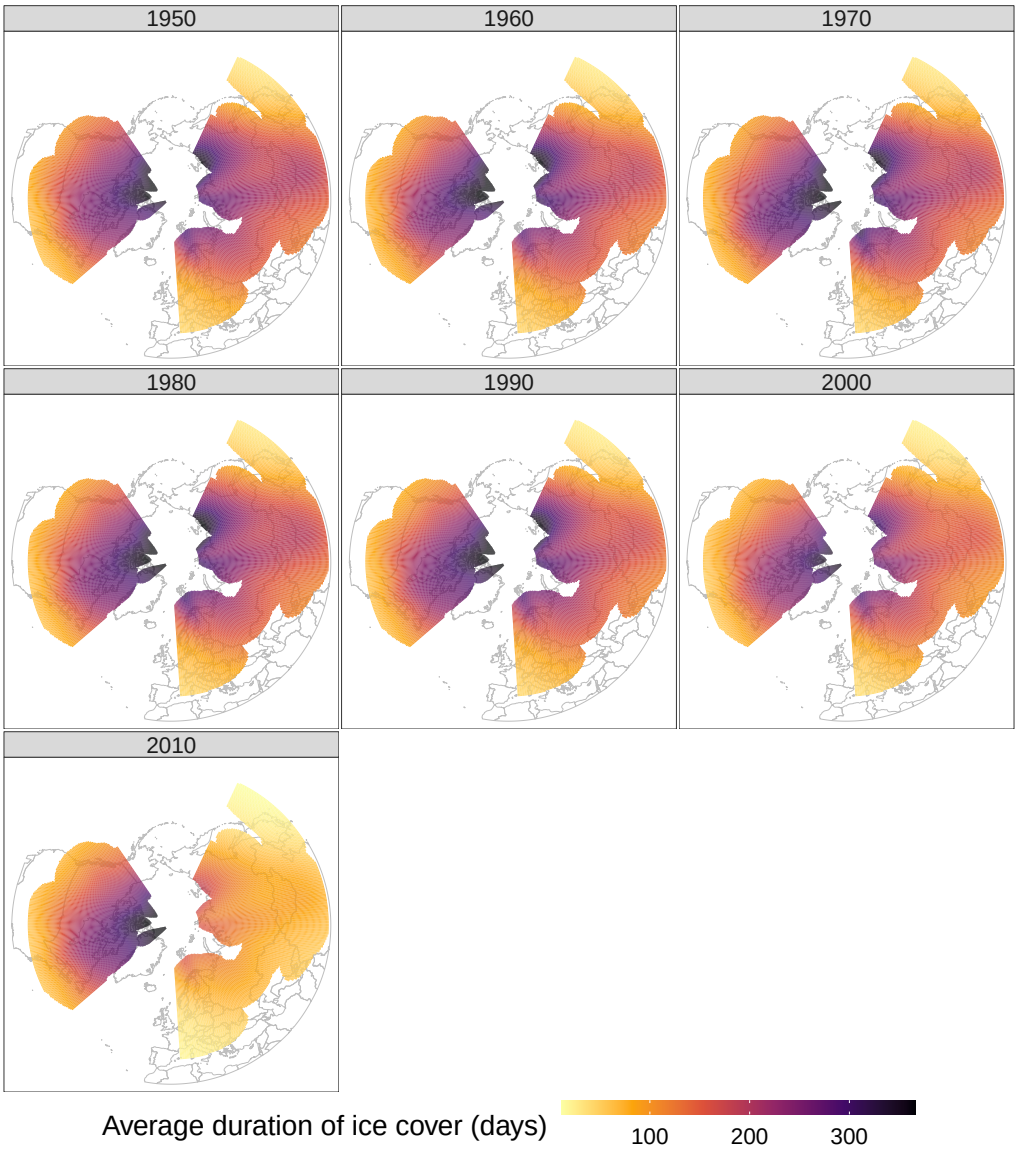

Figure 19: Change in the estimated average duration of lake ice cover in the northern hemisphere during the years 1950-2010.

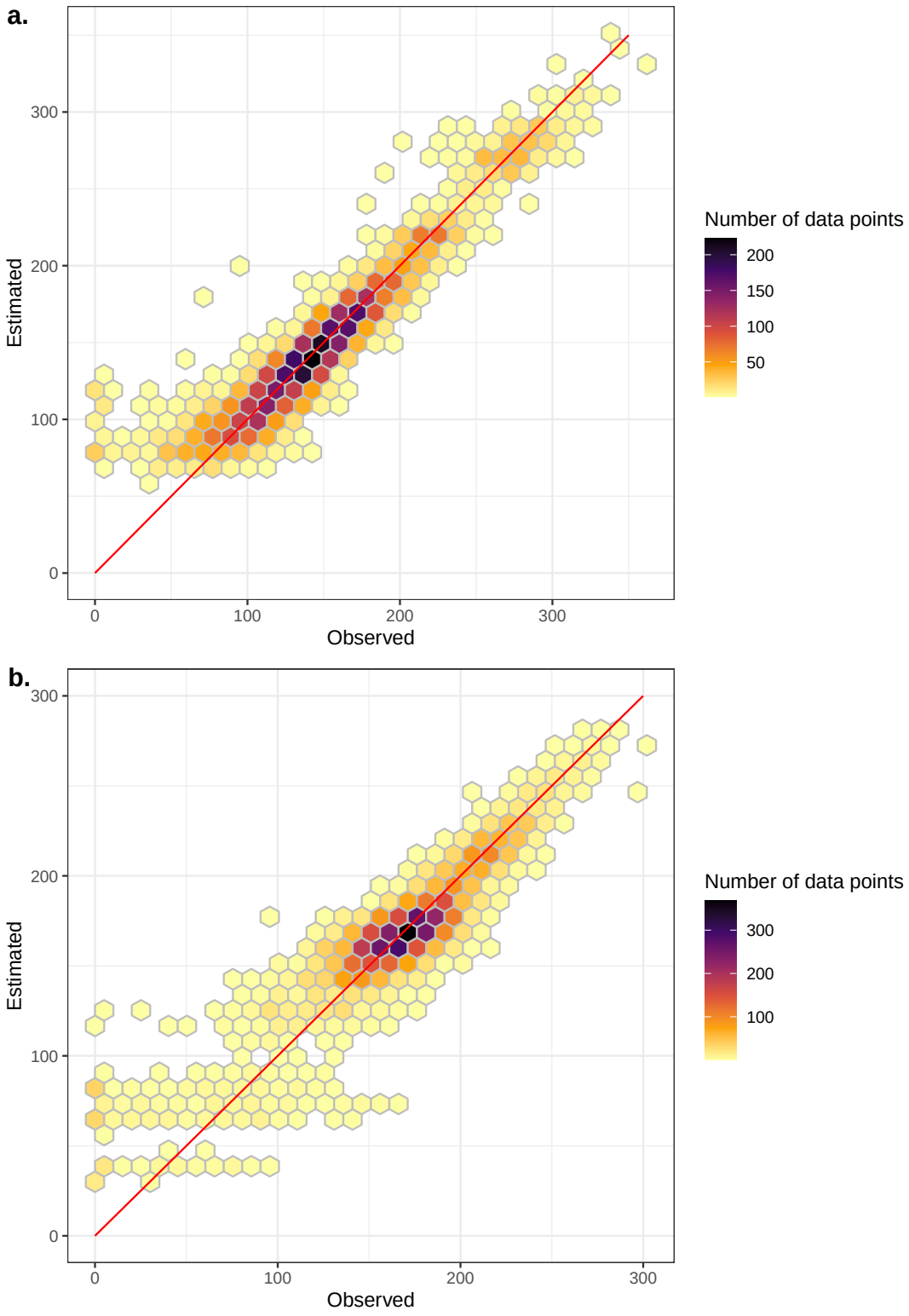

Figure 20: Prediction accuracy of the two HGAMs for the duration of ice cover on lakes. Obervations along the red 1:1 line indicate high prediction accuracy.

4 Discussion

4.1 Single-lake models: Buffalo Pound Lake

The exponential increases in $\hat{\lambda}_{freeze}(t)$ and $\hat{\lambda}_{thaw}(t)$ within years indicate that the probability of Buffalo Pound freezing or thawing increases rapidly within a year. However, the faster increase in $\hat{\lambda}_{freeze}(t)$ than in $\hat{\lambda}_{thaw}(t)$ suggests that the freezing of the lake occurs within a smaller time period that the thawing of the lake. This is also supported by the fact that the variance in day of year of freezing was lower than the variance in the day of year of thawing. A part of the inter-annual variance is likely due to differences in air temperature between years, but the small change in hazard over the years and the small size of the data set prevent us from estimating any long-term inter-annual climatic cycles confidently.

Furthermore, although the average duration of ice cover increased by approximately a day every four years, the low R^2_{adj} (0.00831) indicates that the duration has a high proportion of variance that cannot be predicted by the day of year alone, so the drivers of ice cover may by highly stochastic and not have a strong systematic trend. Still, the R^2_{adj} and the statistical significance of the `Year` smooth might be higher if the observations were taken with higher temporal precision (e.g. at a daily scale rather than weekly). The addition of more predictors to the model may also allow us to estimate these cycles with higher confidence. Potential additions to the model include covariates for daily wind speed, daily air temperature, and the air temperature during the days antecedent the event.

4.2 Duration models: Lake Stechlin

The difference between the estimates of the two models is mostly due to the difference in what each model estimates, since the Tweedie model estimates the overall average duration of ice cover, while the hurdle gamma model estimates the average

duration *given that the lake has frozen*. As mentioned in section 2.4, the left skew in the distribution of the residuals of the Tweedie model indicates that the model likely overestimates the duration of ice cover (Figure 9b).

Since the hurdle gamma model does not show a significant change in lake ice duration, it might be useful to investigate the change in the probability of freezing, i.e. the hurdle parameter in the model. Unfortunately, this kind of analysis would require estimate, since the NUTS algorithm was unable to converge with more complex zero-hurdle models.

Note that although the models show an increase in ice cover in the last years, the models should not be used to conclude that the duration increased post 1990, since the estimated duration near the tails is highly uncertain, as indicated by the wider credible intervals.

4.3 Segmented regression

The high degree of similarity between the SRM and the PAM estimated mean freezing dates indicates that segmented regression can produce acceptable estimates if the assumptions of linearity and sudden change are met. This is unlikely to be the case for cases in which the response has a more wiggly and nonlinear relationship with the predictors. In addition, since the freezing date is necessarily bounded between 1 and 366, models fit to day of freezing cannot be tested using commonly used techniques for linear models such as ANOVA. Such methods assume the residuals are normally distributed (i.e. they can range from -∞ to ∞) and thus are biased towards producing excessively small p-values, particularly if the range in freezing dates is large. Furthermore, while the slope of the first segment can be tested using t and z tests, the uncertainties in the slopes of any other segments need to include the uncertainty in the breakpoints, so t and z tests are not appropriate. Instead Muggeo (2008) suggests using the Davies' test. Alternatively, changes in slope for all con-

tinuous functions can be estimated using the uncertainty around the first derivative (i.e. the slope) of the functions, provided that the estimated variance of the functions is sufficiently unbiased.

4.4 Hierarchical models

The severe shift towards later freezing earlier thawing of North American lakes has greatly changed the lives of many many people, particularly Indigenous People who live in Northern Canada (Golden, Audet & Smith, 2015). Many First Nations report a reduction in food and energy security as well as major changes to traditional activities. A large portion of Indigenous communities in northern Canada rely on (lake) ice roads as a means of travel and transportation between communities and to obtain food and resources (Golden *et al.*, 2015). Many of these people 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 can occur in coastal and sloped locations if 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 that occurs 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), 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). The results presented in this project imply that many people will have to face (and have faced) severe changes in their lifestyle and their relationship with the local environment. The shift towards later freezing and with an increased uncertainty around the expected freezing date has resulted in great stress and changes in the lifestyle of

many people, especially Indigenous people in Northern North America. Our limited knowledge of under-ice ecology will likely prevent us from knowing what we have lost before it is gone (Hampton *et al.*, 2017), and youth, particularly Indigenous youth, have already experience stress due to changes in climate and loss of ice (Petrasek MacDonald *et al.*, 2013).

The lack of substantial lake-specific deviations from the average trends over the years in North America indicates that the shift to later freezing and earlier thawing are common trends between North American lakes, once the spatial effect is accounted for. The estimates presented here do not match previous evidence which indicates that the polar regions are likely to warm at significantly faster rates than temperate and equatorial regions (Holland & Bitz, 2003). This may be due to the low number of arctic lakes in the last decade of the data set, which prevents estimates with high certainty and does not contain any changes that might have occurred recently. As evidenced by the change in the hazard of freezing in North America following 1995, freeze and thaw dates may change greatly within 10-15 years. Since North American polar lakes have historically been frozen for the majority of the year (Figure 19), even a short increase in mean air temperature above 0°C may result in long periods of ice loss.

Overall, the widening of the period of time when $\widehat{F}_{freeze}(t)$ is between 0 and 1 may be due to multiple factors. Two possible explanations are an increase in variance of the ice-on dates within each lake, or a slower freezing process. Either cause is likely to decrease the stability and predictability of lake ice formation, which will likely cause severe changes in the ability to use lake ice for transportation or leisure. Still, the lack of a uniform change in the duration of ice cover in North America indicates that the changes within the continent are not homogeneous, and demonstrates the importance of large-scale spatial models when estimating changes in climate and ice phenology.

Although the change in average freeze and thaw dates in Eurasia was considerably less than that in North America, the significant decrease in the duration of ice cover indicates that lakes freeze and thaw multiple times within a year. The decrease in ice cover in central Europe and Eastern Asia indicates that lakes in these regions are less likely to freeze annually (if at all) in the coming years. The shift towards intermittent ice cover within a year may lead to great unpredictability regarding when lakes will freeze since lakes that do not freeze yearly or have short periods of ice cover tend to have greater between-year variance in their duration of ice cover (Figure 20), potentially due to factors such as wind disturbance and daily fluctuations in temperature having a greater effect on whether a lake is able to freeze or not.

The drastic and widespread decrease in ice cover in the Eurasia has already had a strong impact on the lives of many of many 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). During the last century, the number of years in which an *omiwatari* did not form on Lake Suwa has increased significantly, with more than 67% of the unusually warm years occurring since 1989.

The 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 people (Orru *et al.*, 2014). Due to the stochasticity of the duration ice cover on the lake, the amount of fish that is caught can vary as much as by a factor of 10 from year to year (Orru *et al.*, 2014).

The unpredictability of when lakes and rivers will freeze has also caused many skating events and competitions in northern Europe to be cancelled, including the Swedish "Viking ride", a long-distance skating race that was held annually from 1999

until 2018, when it was cancelled due to the unpredictability of the lake ice (Knoll *et al.*, 2019).

4.5 Future work

Potential future work includes estimating the variance within lakes by nesting random effects for the observation stations at each lake into the lake random effects. However, due to the large number of lakes in the data set, nested random effects will require substantially more computing power and time than the models presented in this thesis. The variance between and within lakes could also be analyzed using location-scale models to estimate how the average $\hat{\lambda}_{freeze}(t)$ and $\hat{\lambda}_{thaw}(t)$ and their variance change over space and time, although this would prevent fitting the model as a Poisson HGAM, since the distribution of the response would be different, and these types of models cannot be currently fit using the `mgcv` package.

Predictions using data prior to 1950 could also be done for smaller regions with a large amount of long time series such as Finland or the Great Lakes Area (Figures 4 and 5).

Another option is the inclusion of air temperature as a predictor in the HPAMs to estimate the effect of air temperature on the freezing and thawing of lakes. Alternatively, the temperature record could be lagged by a number of days d to estimate the effect of antecedent air temperature on the freeze or thaw date. The most likely value for d could be estimated by using more traditional and less flexible models, so that the patterns in the residuals may be analyzed using common time series analysis techniques such as auto-regressive and moving-average processes (Cowpertwait & Metcalfe, 2009).

4.6 Conclusion

Although the average duration of ice cover on lakes and the average hazard of freezing and thawing did not change substantially between 1950 and 1995, there has been widespread and dramatic change in the hazard of freezing and thawing and duration of ice cover in the following decades. There is currently little understanding about how this has affected biological systems, but the damage to many people's lifestyle, traditions and economy are evident. Therefore, it is imperative that steps are taken to mitigate the damages to human and ecological communities, and that lake ice is studied further before we can no longer understand how it affects ecological systems.

Appendix i: Abbreviations used

Abbreviation	Term
GAM	Generalized Additive Model
HGAM	Hierarchical Generalized Additive Model
HPAM	Hierarchical Piecewise-exponential Additive Model
LM	Linear Model
MCMC	Markov Chain Monte Carlo
NUTS	No-U-Turn Sampler
PAM	Piecewise-exponential Additive Model
SRM	Segmented Regression Model
TWLSS	Tweedie Location, Shape, and Scale

Appendix ii: 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/stefano-honours. The repository contains separate folders for the data, code, custom functions, and plots used in the project, as well as the project proposal and the thesis document.

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

Burger S.V. (2018). *Introduction to machine learning with r: Rigorous mathematical analysis*, First edition. O'Reilly Media, Sebastopol, California.

Bürkner P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software* **80**. https://doi.org/10.18637/jss.v080.i01

Cowpertwait P.S.P. & Metcalfe A.V. (2009). *Introductory time series with r*. Springer, Dordrecht; New York.

Duchon J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: *Constructive theory of functions of several variables*. (Eds W. Schempp & K. Zeller), pp. 85–100. Springer Berlin Heidelberg, Berlin, Heidelberg.

Finlay K., Vogt R.J., Simpson G.L. & Leavitt P.R. (2019). Seasonality of pCO $_2$ in a hard-water lake of the northern great plains: The legacy effects of climate and limnological conditions over 36 years: Regulation of seasonal pCO $_2$. *Limnology and Oceanography* **64**, S118–S129. https://doi.org/10.1002/lno.11113

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.

Hoffman M. & Gelman A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15**, 1593–1623

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

Muggeo V. (2008). Segmented: An r package to fit regression models with broken-line relationships. *R News* **8**, 20–25

Muggeo V.M.R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine* **22**, 3055–3071. https://doi.org/10.1002/sim.1545

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.

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. (2018). Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution* **6**, 149. https://doi.org/10.

3389/fevo.2018.00149

Stan Development Team (2015). *Stan modeling language: User's guide and reference manual*, Version 2.9.0.

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.

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.