Smooth operator — using generalised additive models to quantify trends in noisy spatiotemporal data

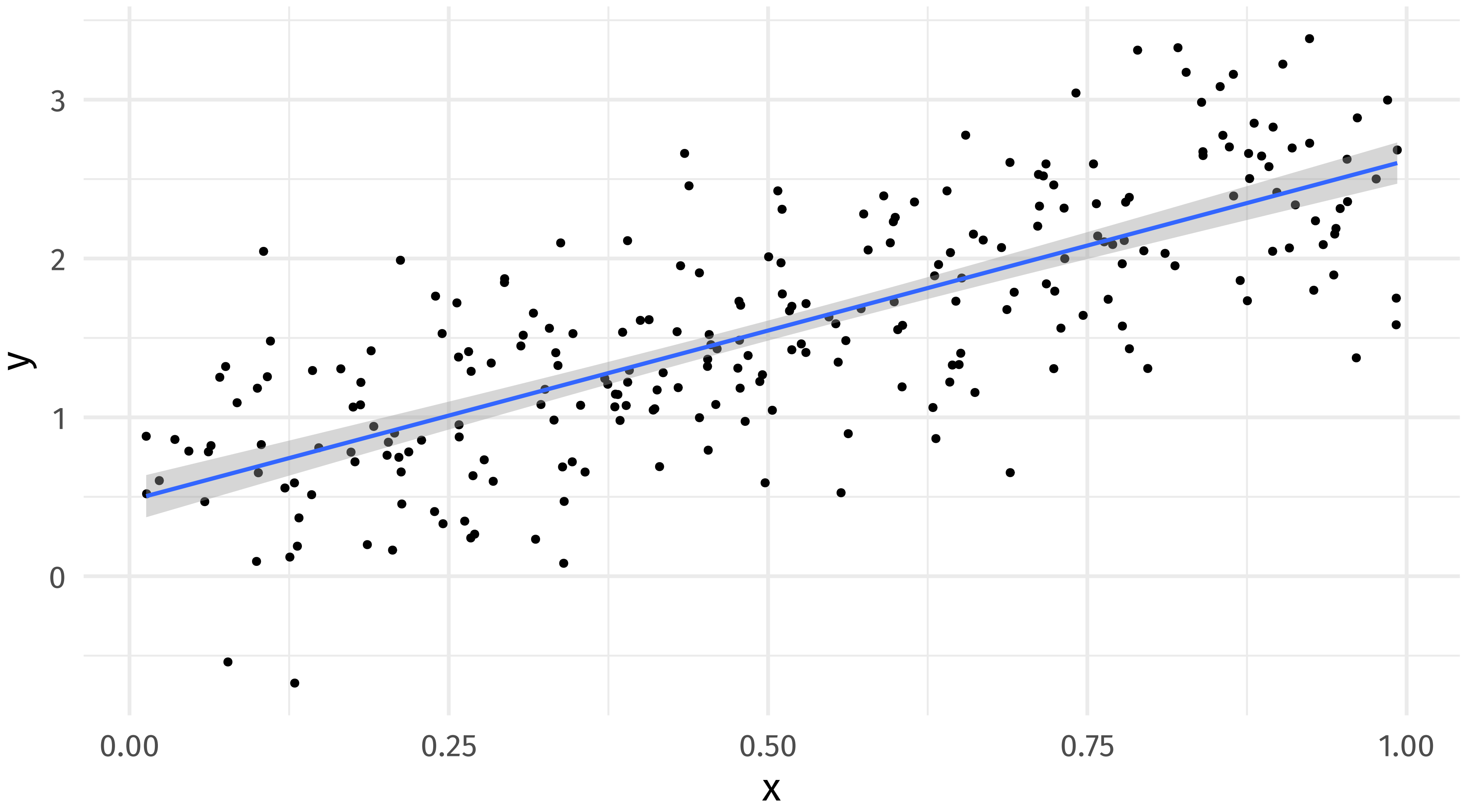
Gavin Simpson

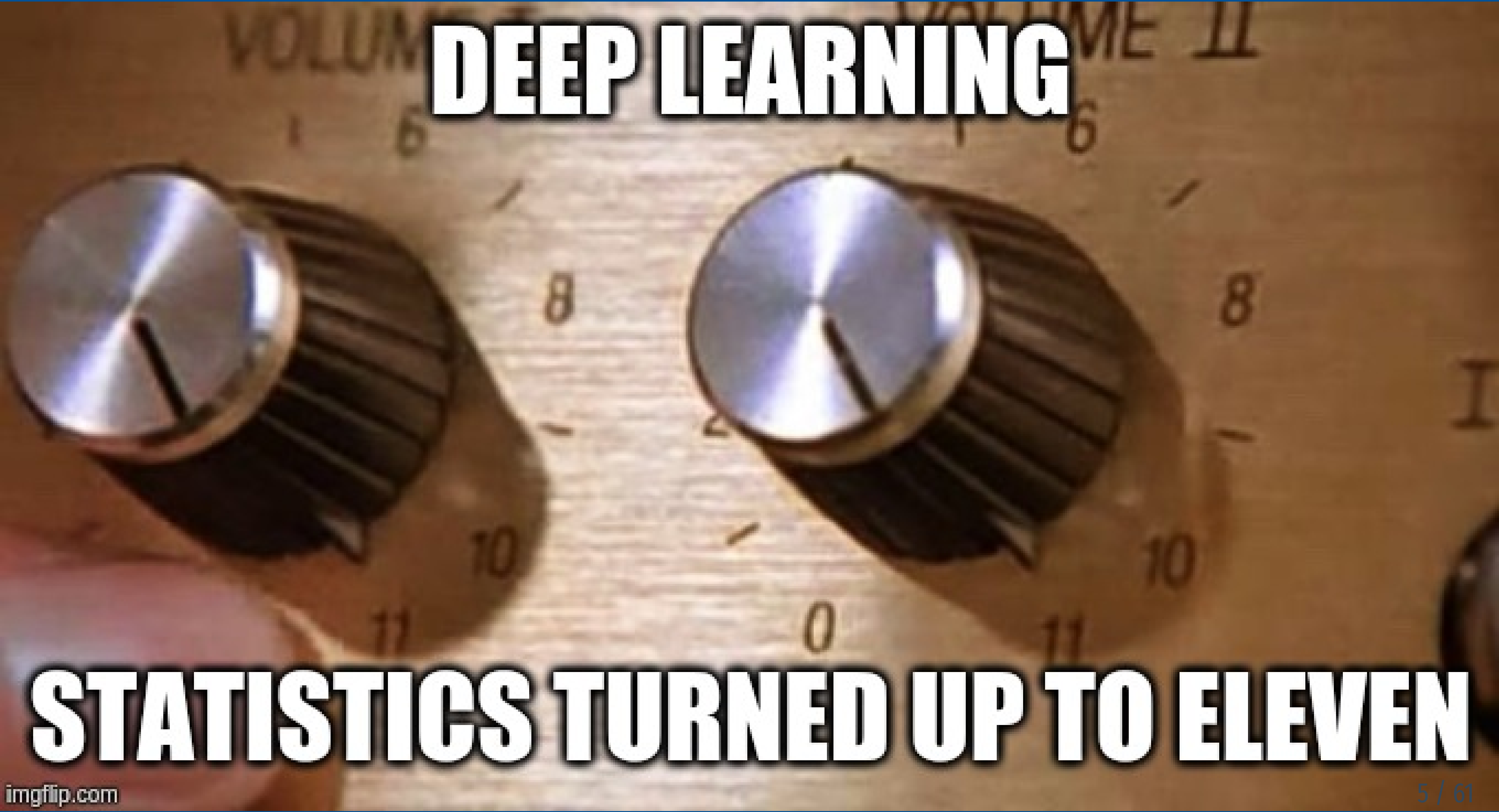
HIFMB · University of Oldenburg · November 3, 2020

Use statistics to learn from data in presence of noise

Franki Chamaki

Learning from data



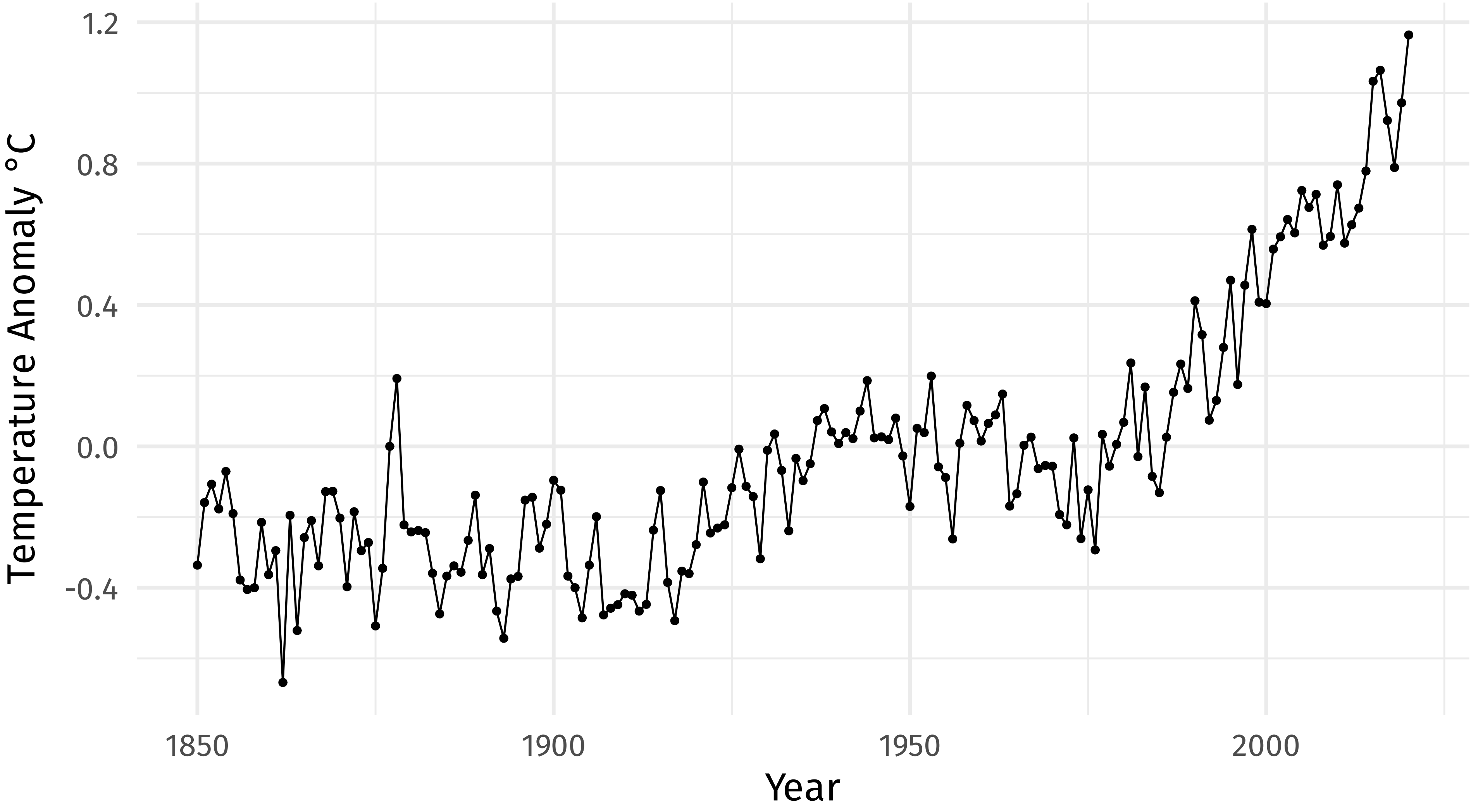

DEEP LEARNING
STATISTICS TURNED UP TO ELEVEN
imgflip.com

Learning involves trade-offs

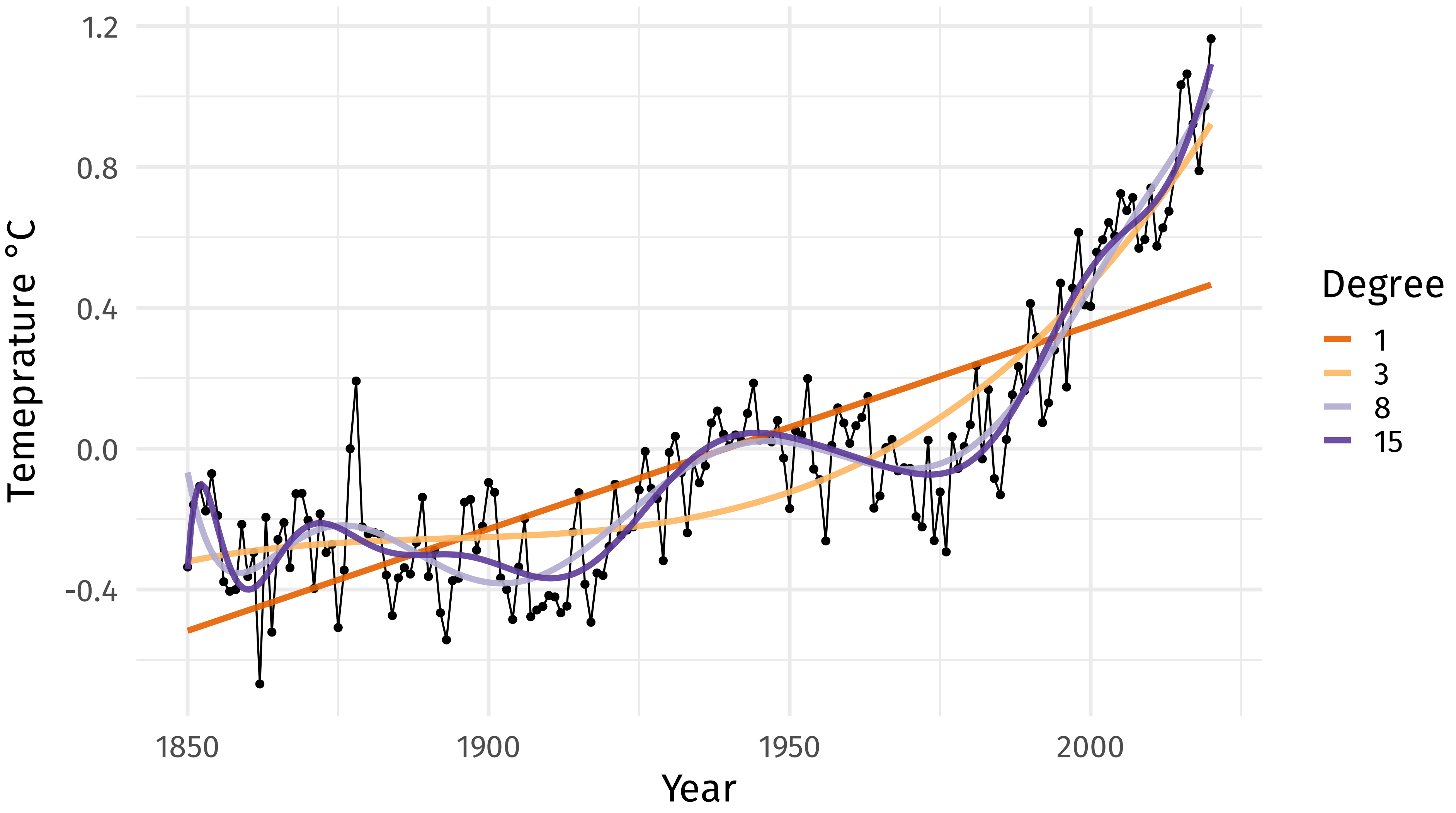
Generalized Additive Models

Source: GAMs in R by Noam Ross

HadCRUT4 time series

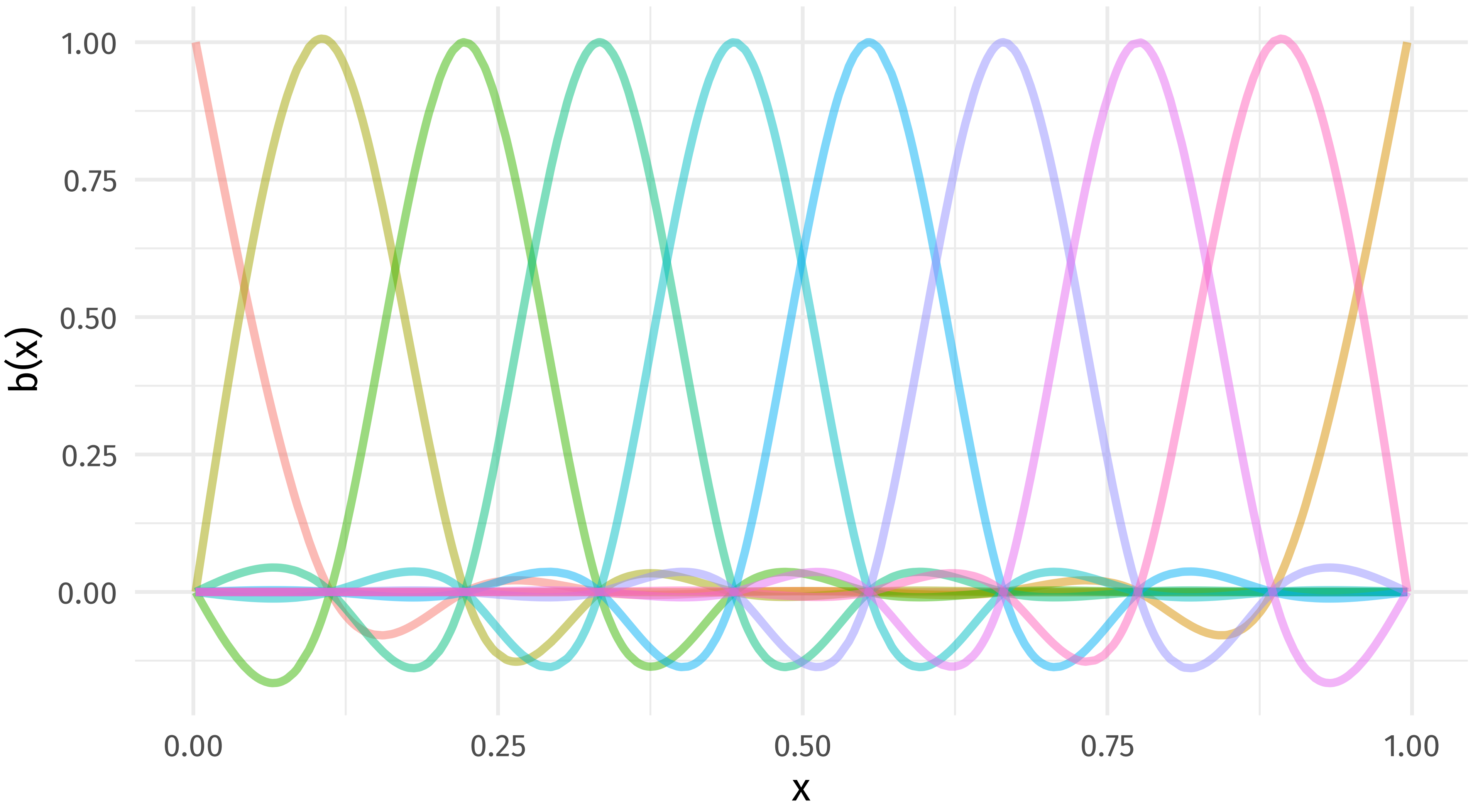


Polynomials

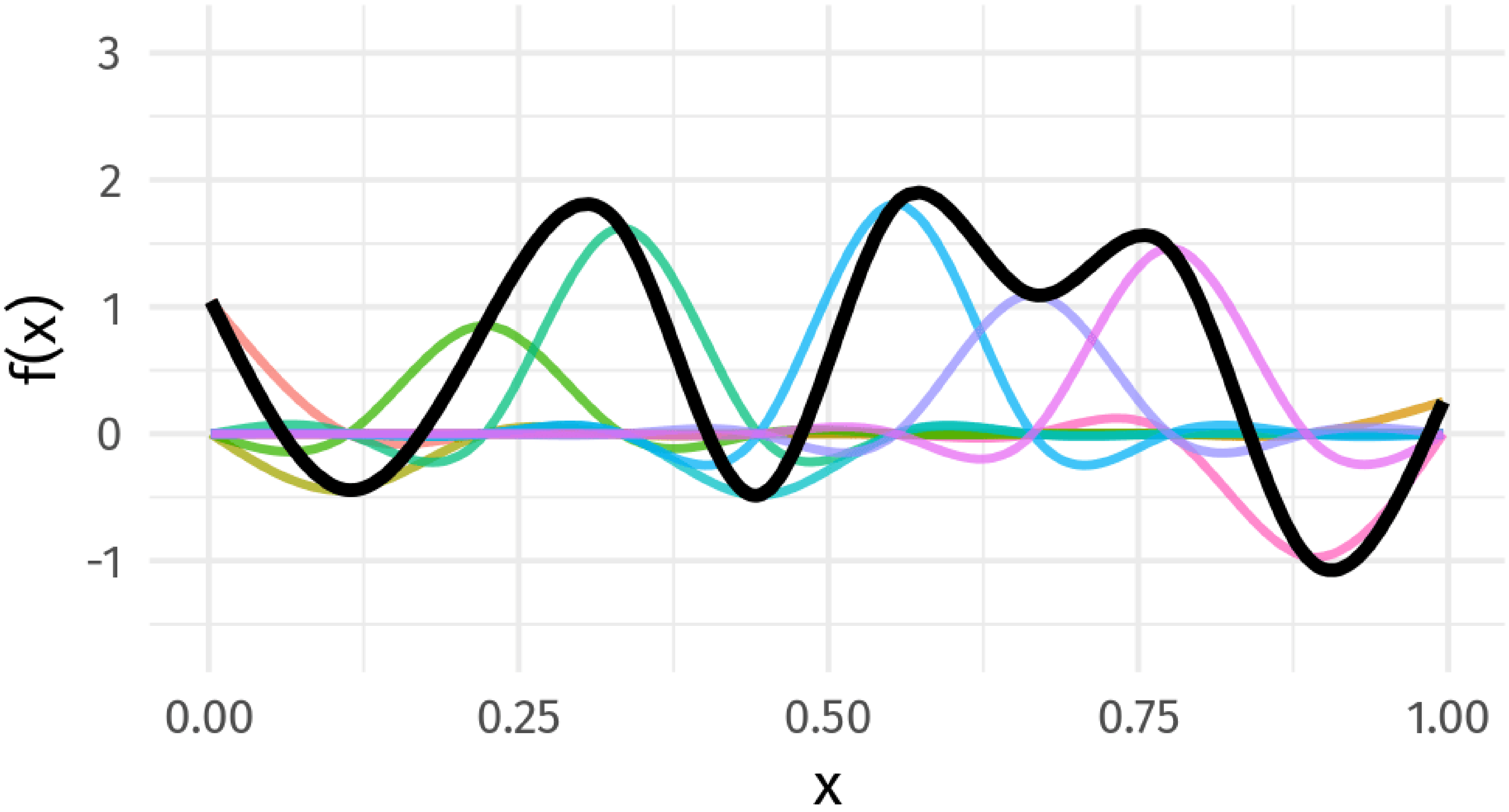


GAMs use splines

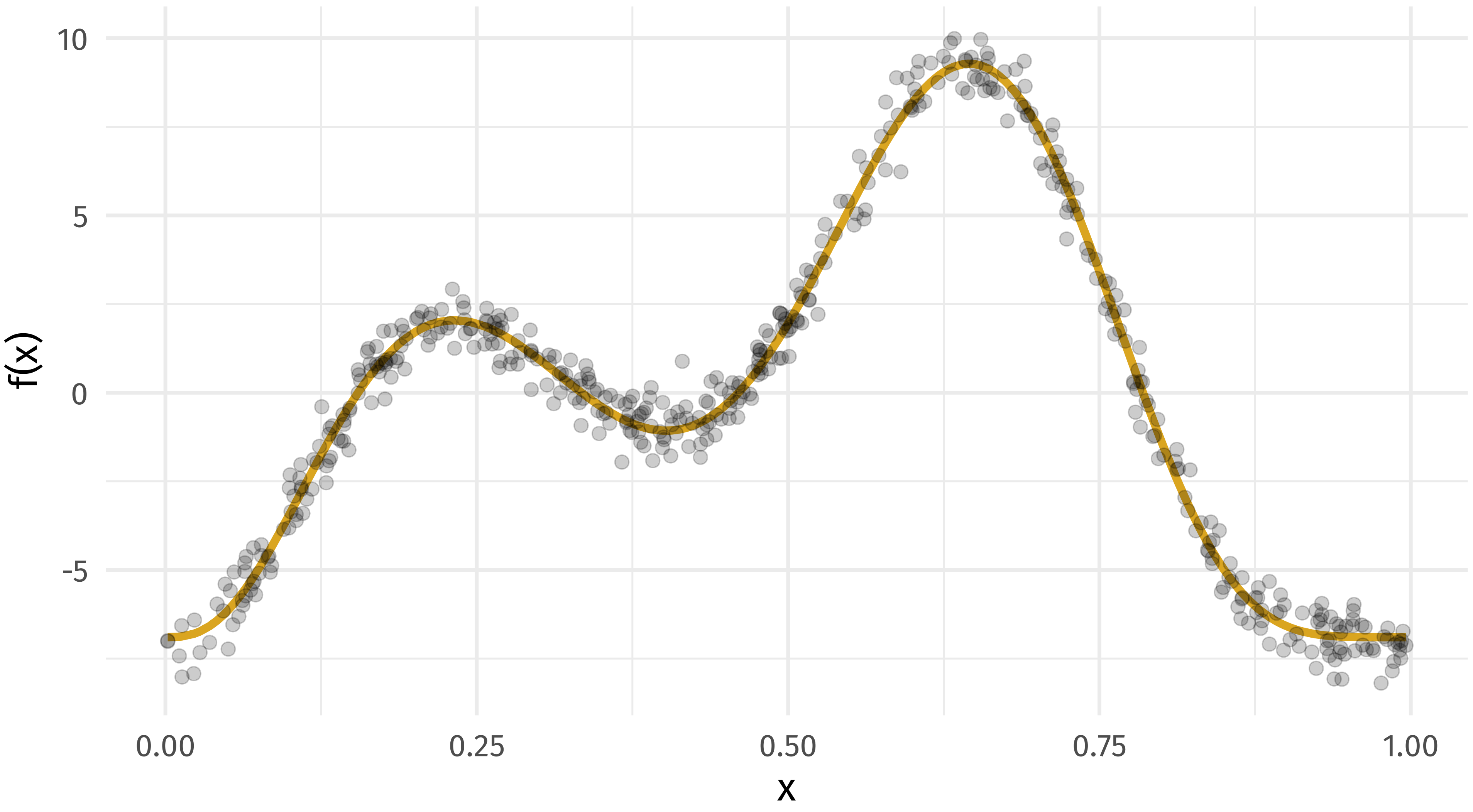
Splines formed from basis functions



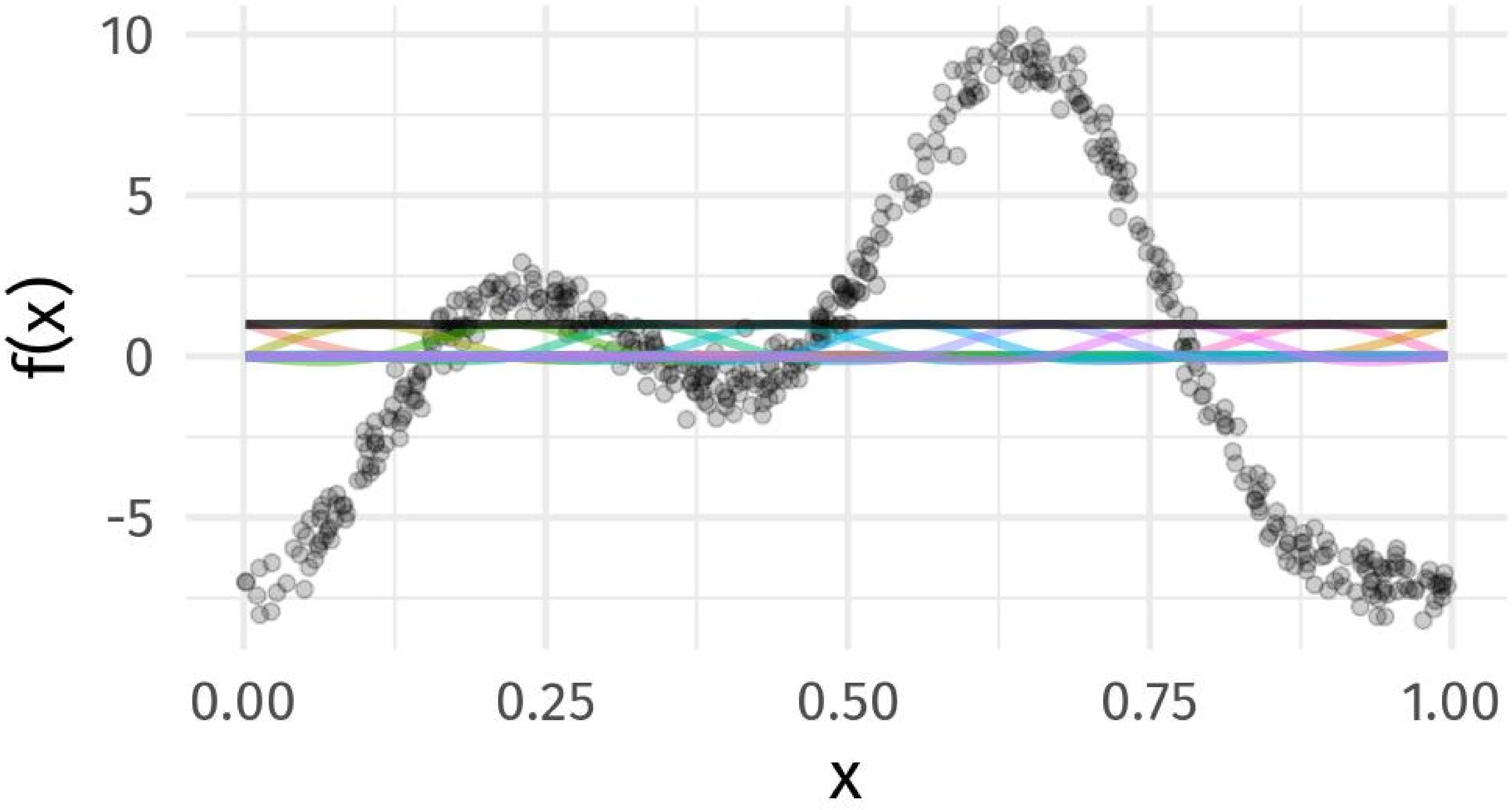
Weight basis functions ⇒ spline



How do GAMs learn from data?



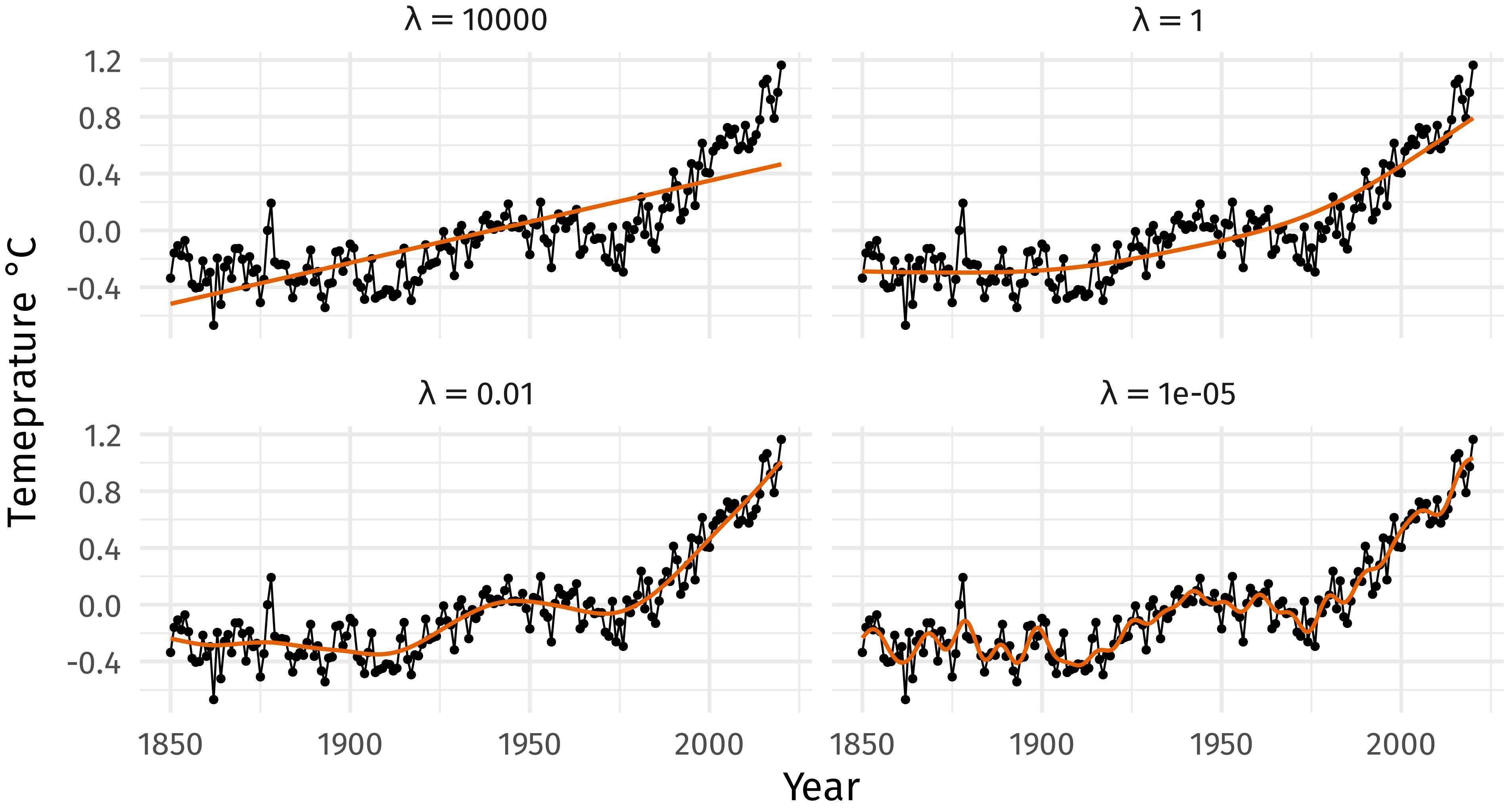
Maximise penalised log-likelihood ⇒ β



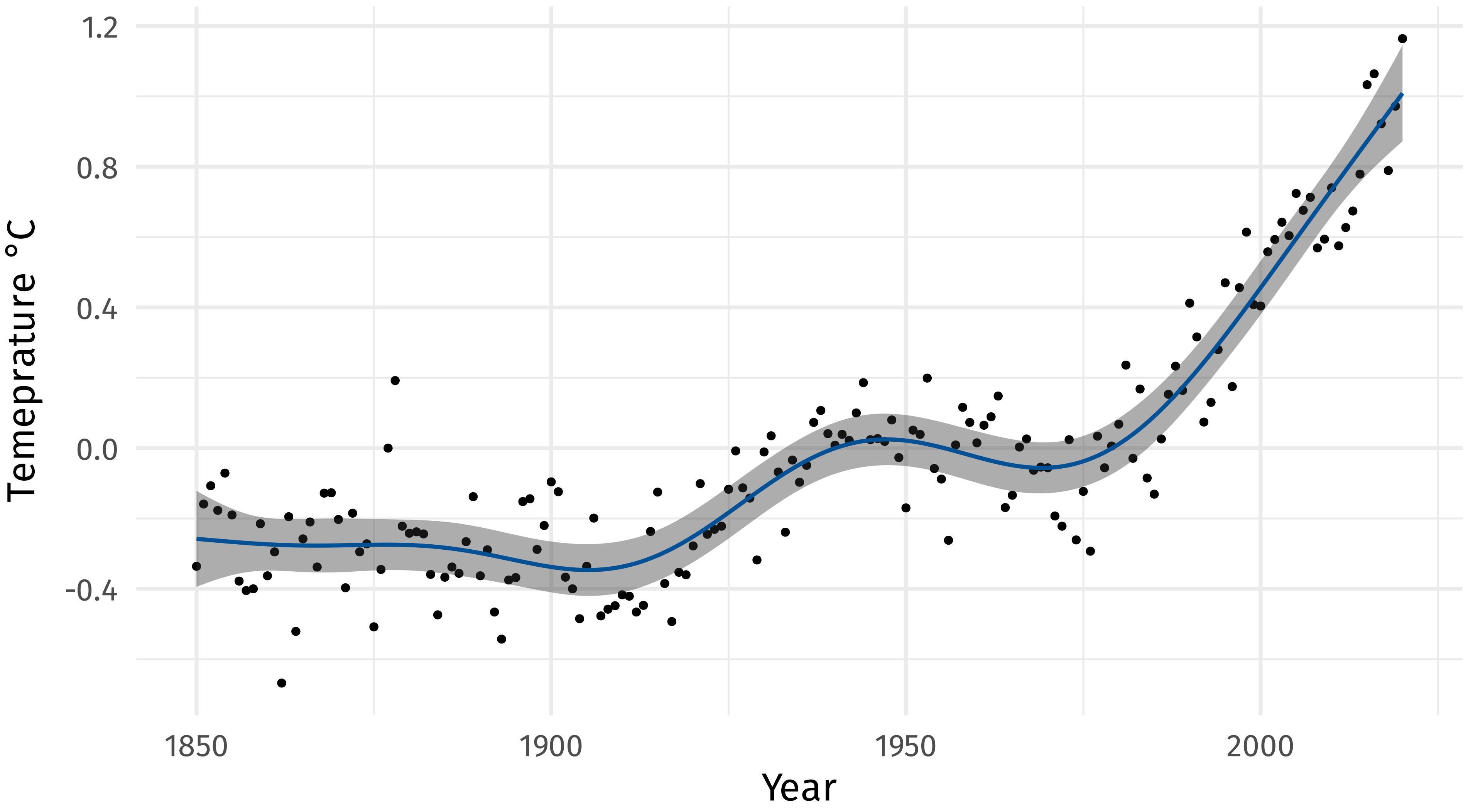
Avoid overfitting our sample

Use a wiggliness penalty — avoid fitting too wiggly models

HadCRUT4 time series



Fitted GAM



Vignettes

Vignette 1

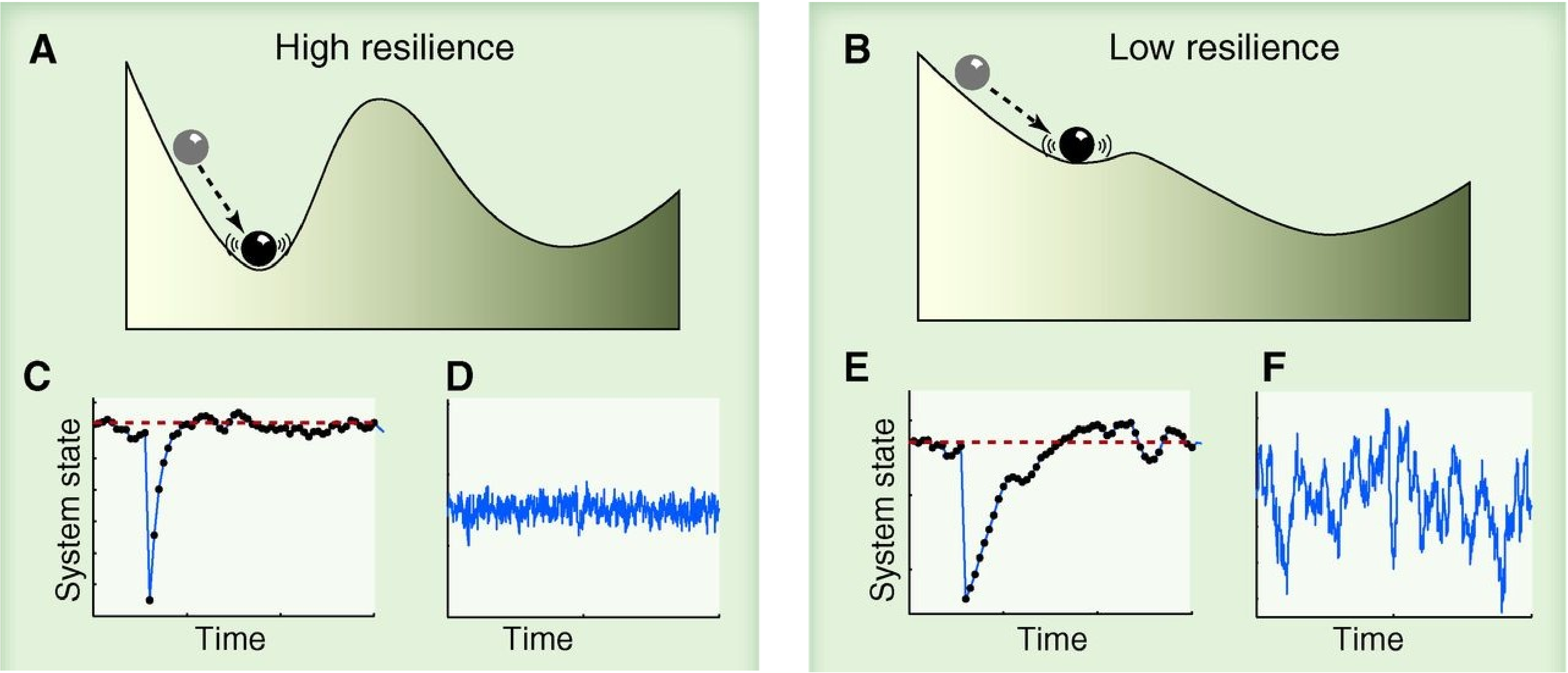
Statistical estimation of ecosystem variance

Mean

Variance

Variance

Intuitive & key descriptor of ecosystem state

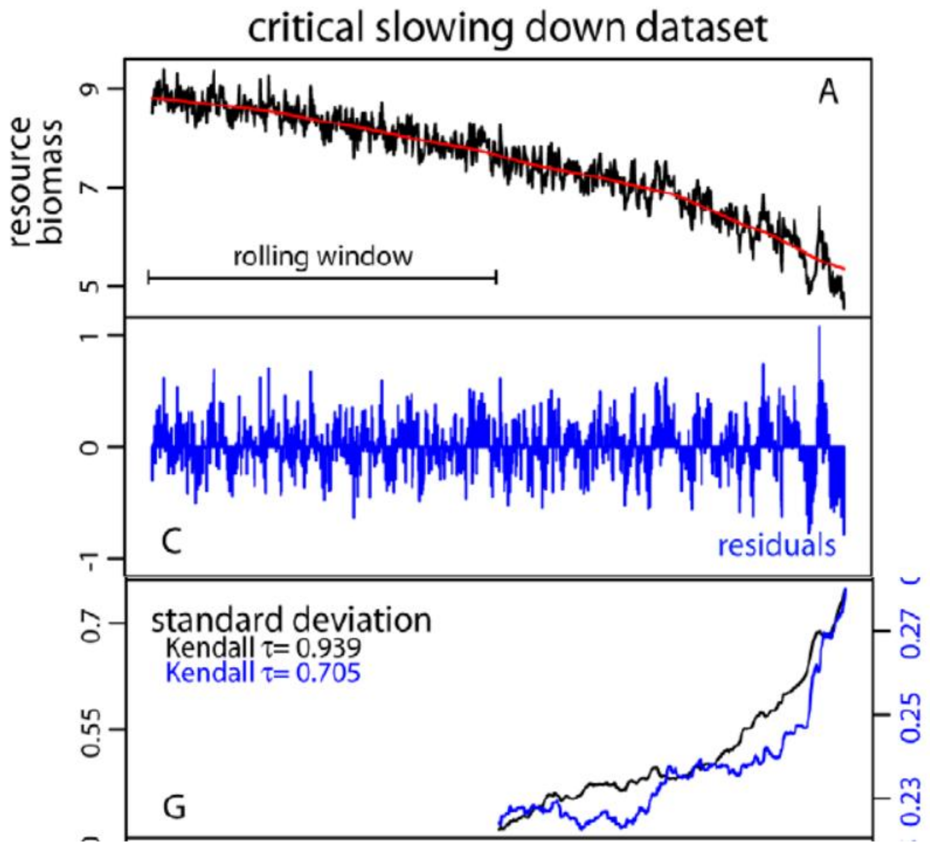


Source: Scheffer *et al* Science (2012)

Moving windows — don't do this!

Moving window approach:

1. detrend
2. choose window size
3. estimate variance
4. move window 1 time step
5. repeat 3. & 4. until done
6. look for trend



Modified from Dakos *et al* (2012) PLOS One

Problems

Ad hoc

- How to detrend (method, complexity of trend, ...)
- What window width?

Ideally regularly spaced data

- What to do about gaps, missing data, irregularly spaced data?

Statistical testing hard

- Kendall's τ assumes independence
- Surrogate time series assumes regular sampling, sensitive to choice of ARMA

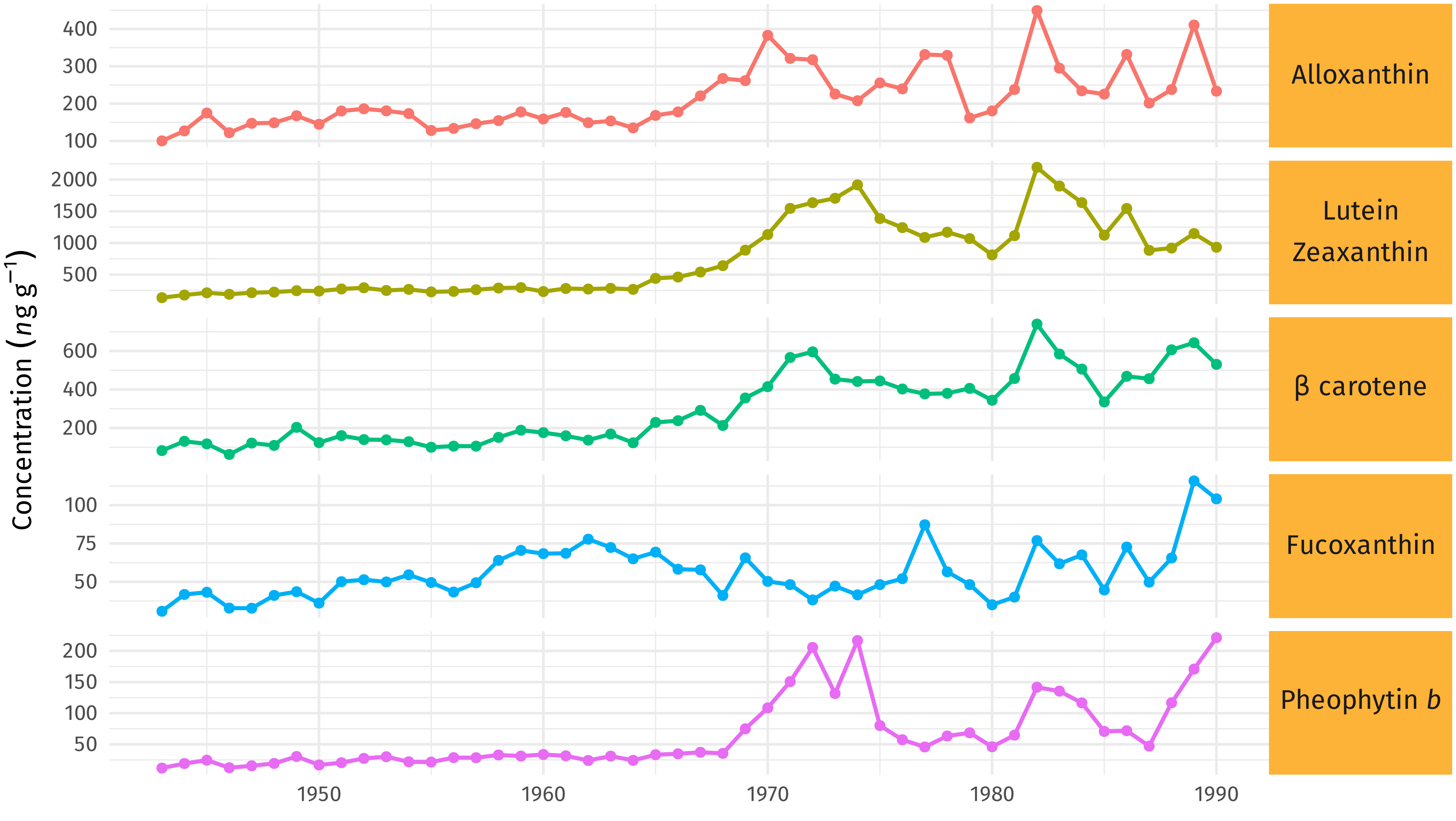
Lake 227

Lake 227

- Experimentally manipulated
- Annual sediment samples 1943–1990
- Analyzed for fossil pigments

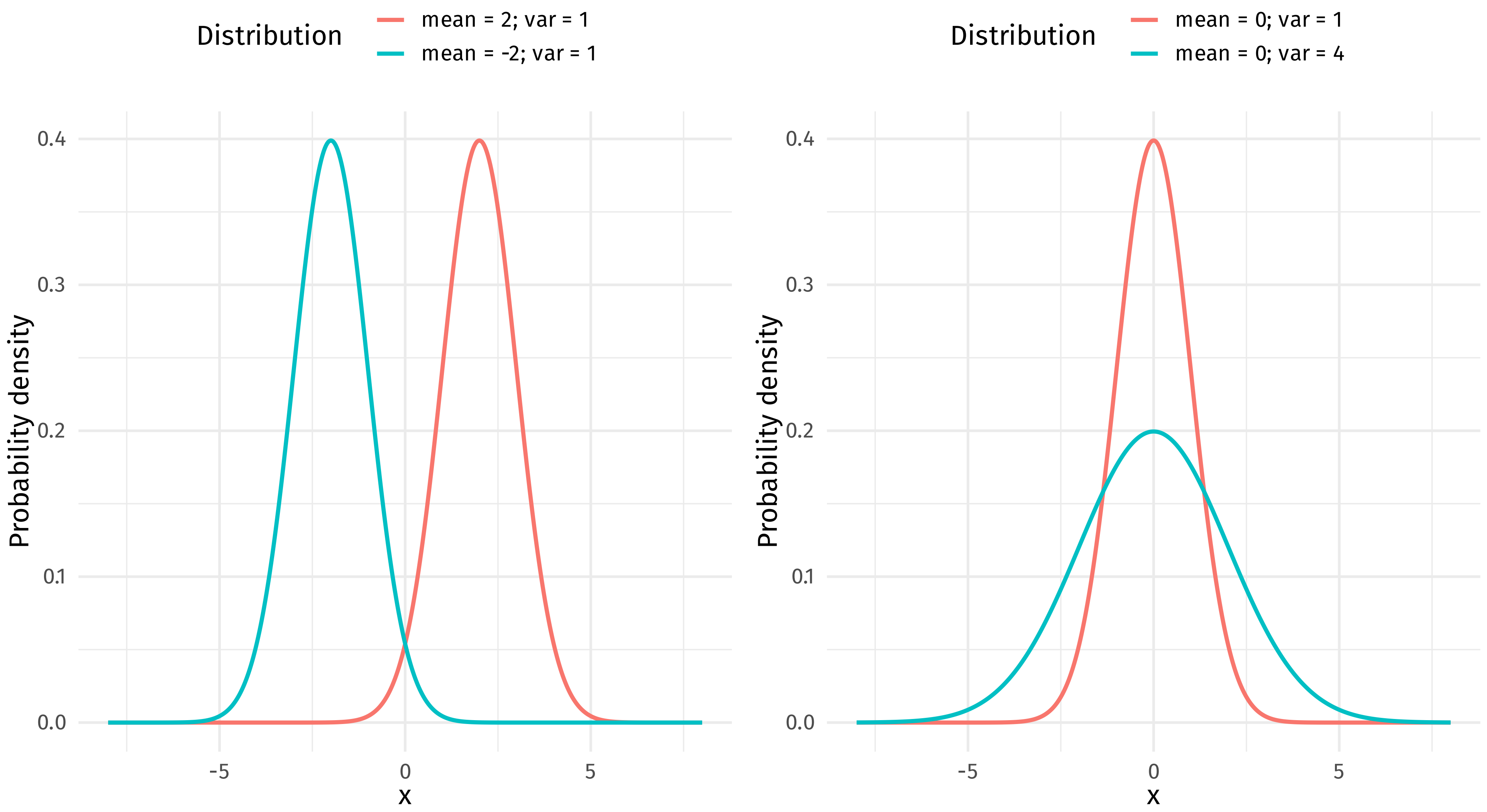
Cottingham *et al* (2000) **Ecology Letters** showed via a Levene's test that **variances** pre- & post-intervention were different

Lake 227


Concentration $(ng\ g^{-1})$
400
300
200
100
2000
1500
1000
500
600
400
200
100
75
50
200
150
100
50
1950
1960
1970
1980
1990
Alloxanthin
Lutein
Zeaxanthin
β carotene
Fucoxanthin
Pheophytin *b*

Challenge

Parameters beyond the mean



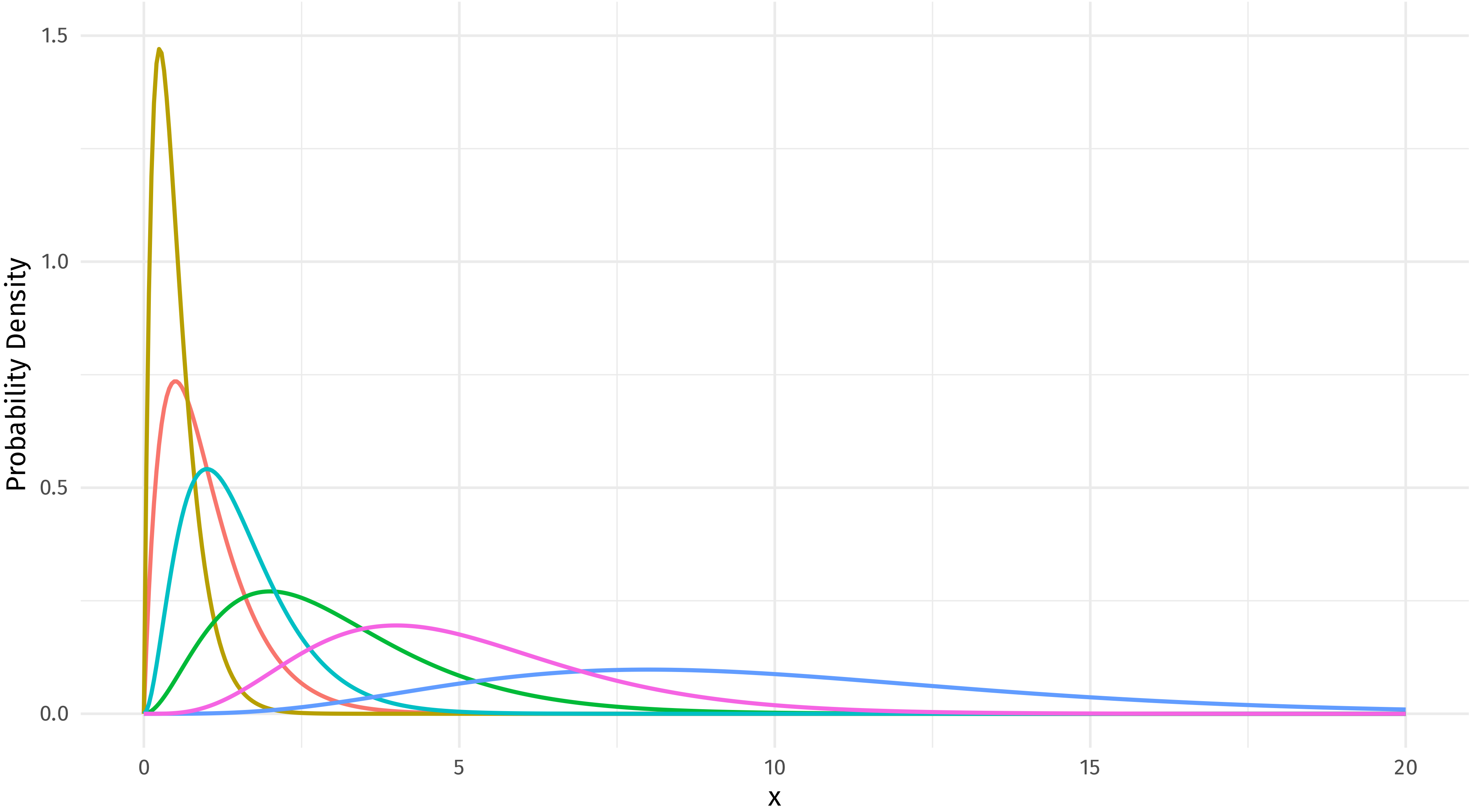
Distributional models

$$y_i | \boldsymbol{x}_i \sim \mathcal{D}(\vartheta_1(\boldsymbol{x}_i), \ldots, \vartheta_K(\boldsymbol{x}_i))$$

For the Gaussian distribution

- $\vartheta_1(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$
- $\vartheta_2(\boldsymbol{x}_i) = \sigma(\boldsymbol{x}_i)$

The Gamma distribution

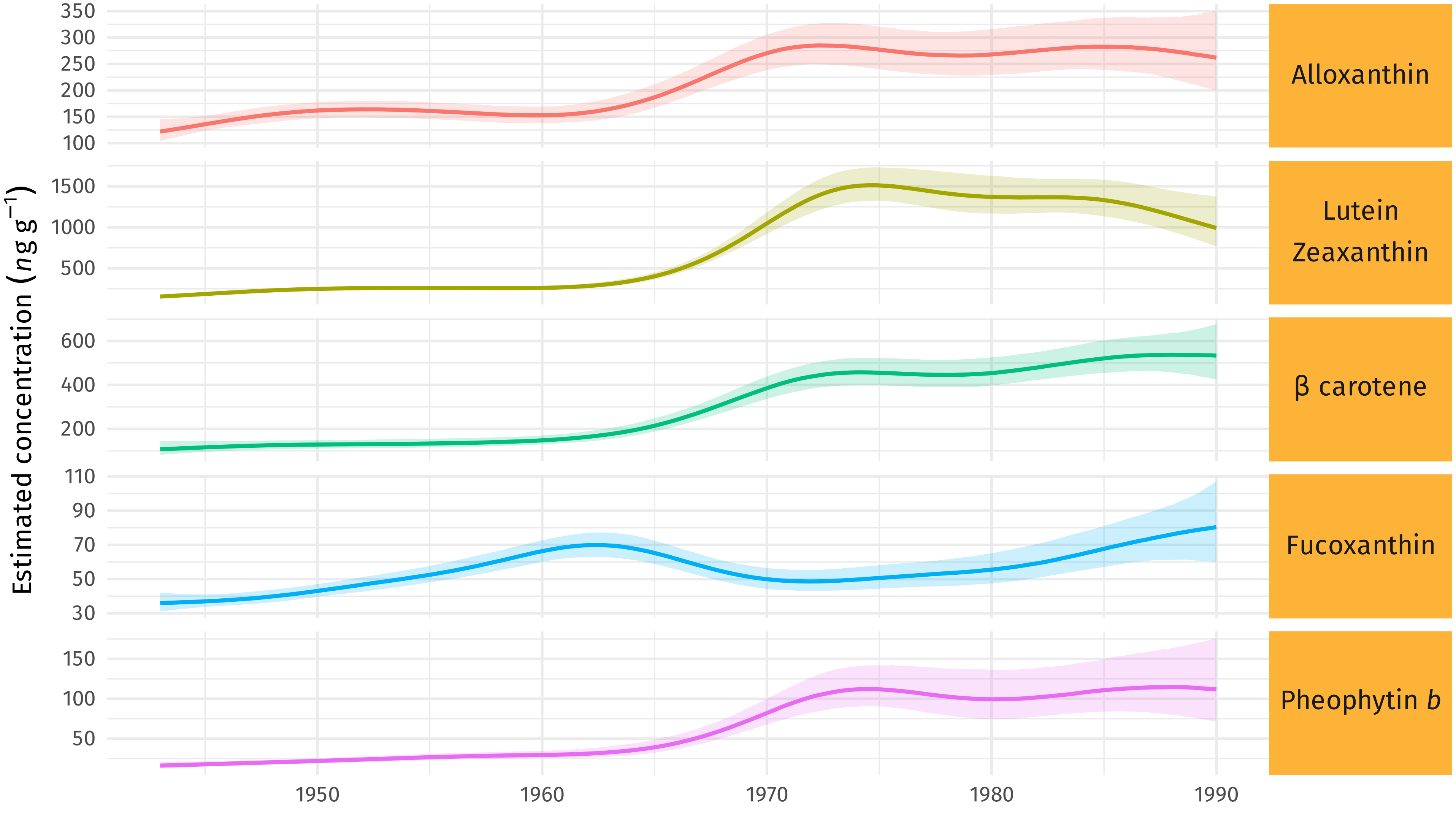


Lake 227 sedimentary pigments

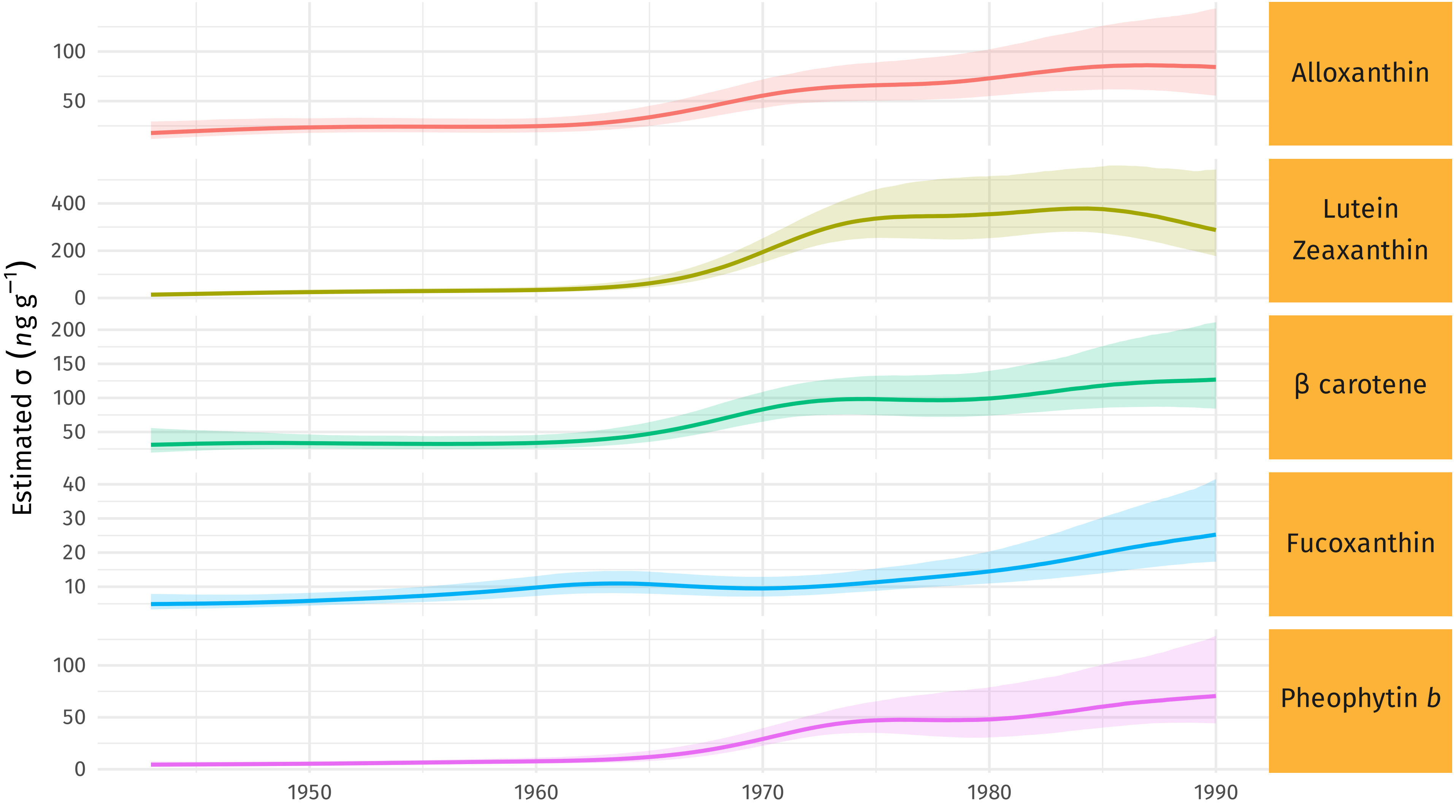
Fitted gamma distributional model using **brms** and **Stan**

$$
\begin{aligned}
y_t &\sim \mathcal{G}(\mu = \eta_{1,t}, \alpha = \eta_{2,t}) \\
\eta_{1,t} &= \exp(f_{1,\text{pigment}}(\text{Year}_t)) \\
\eta_{2,t} &= \exp(f_{2,\text{pigment}}(\text{Year}_t))
\end{aligned}
$$

Results — pigment means



Results — pigment variances



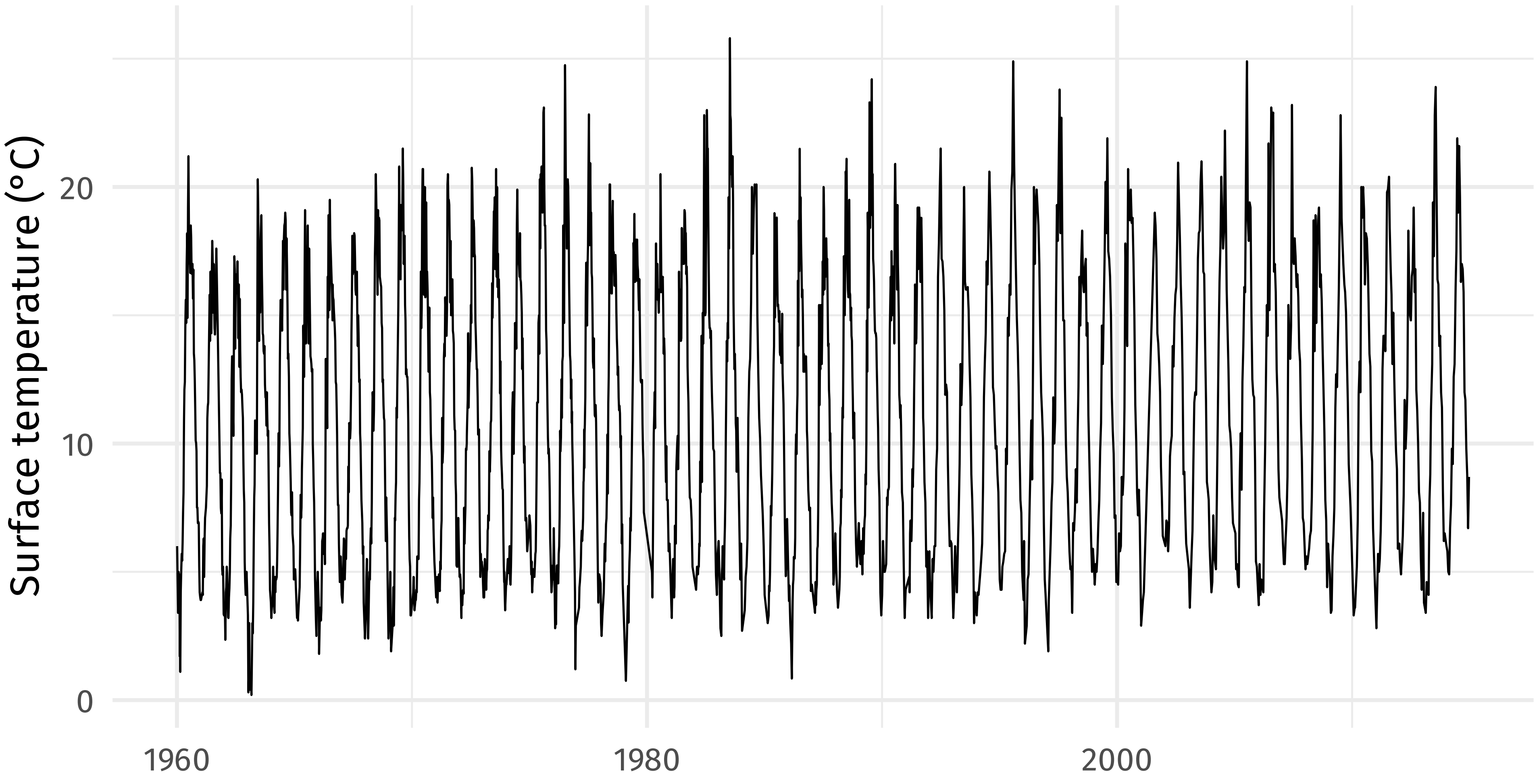
Summary

- Using distributional GAMs we now have a means to robustly estimate the variance of an system and how that variance changes over time
- Use models such as these to begin to quantify an important ecological metric from a principled statsitical view point

Vignette 2

Climate change affecting lake temperatures?

Blelham Tarn, UK



Data: Woolway *et al* (2019) *Climate Change* **155**, 81–94 doi: 10/c7z9

Why worry about minimum temperatures?

Why worry about minimum temperatures?

Annual minimum temperature is a strong control on many in-lake processes (eg Hampton *et al* 2017)

Extreme events can have long-lasting effects on lake ecology — mild winter in Europe 2006–7 (eg Straile *et al* 2010)

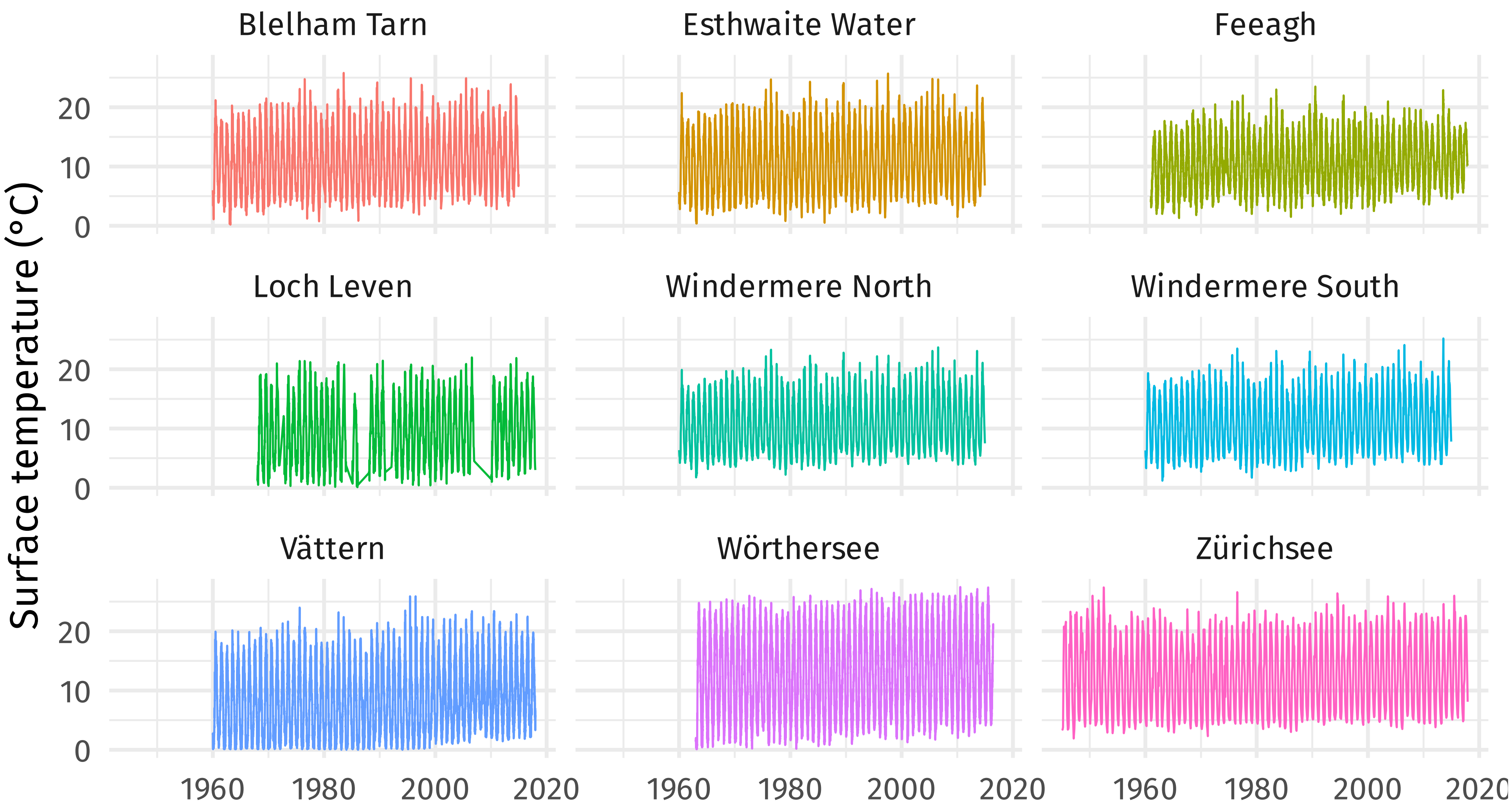
Reduction in habitat or refugia for cold-adapted species

- Arctic charr (*Salvelinus alpinus*)
- Opossum shrimp (*Mysis salemaai*)

Hampton *et al* (2017). Ecology under lake ice. *Ecology Letters* **20**, 98–111. doi: 10/f3tpzh

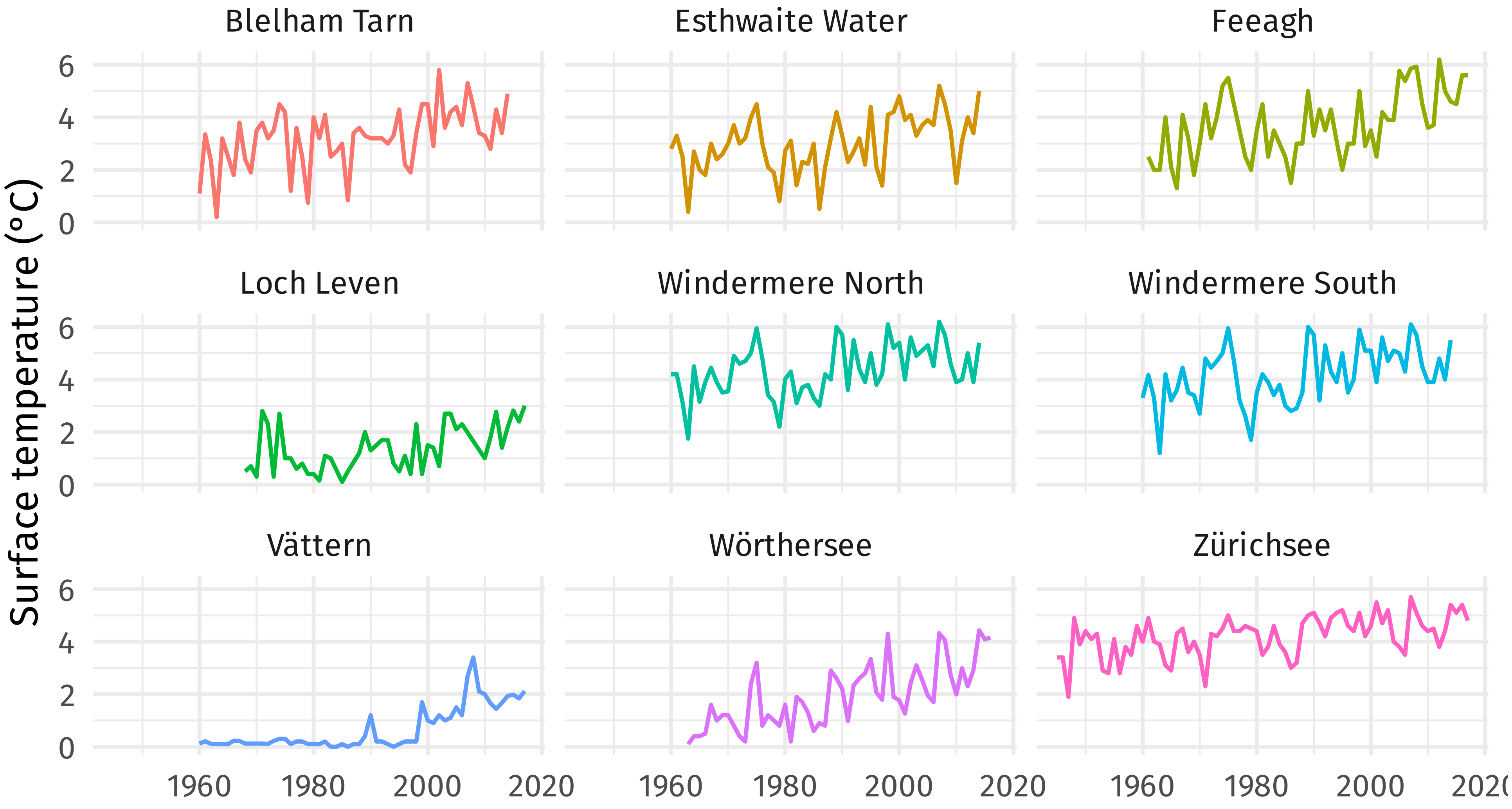
Straile *et al* (2010). Effects of a half a millennium winter on a deep lake — a shape of things to come? *Global Change Biology* **16**, 2844–2856. doi: 10/bx6t4d

Multiple time series ⇒ HGAM



Central Limit Theorem

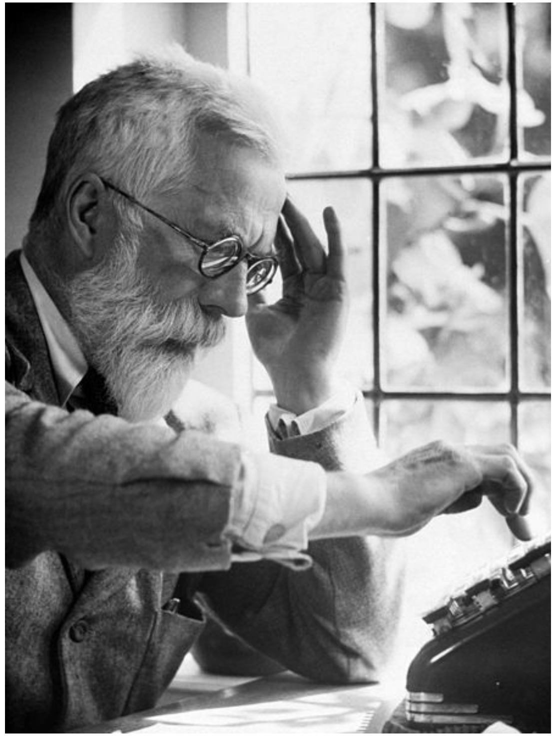
Annual minimum temperature



Block Minima

Fisher–Tippett–Gnedenko theorem

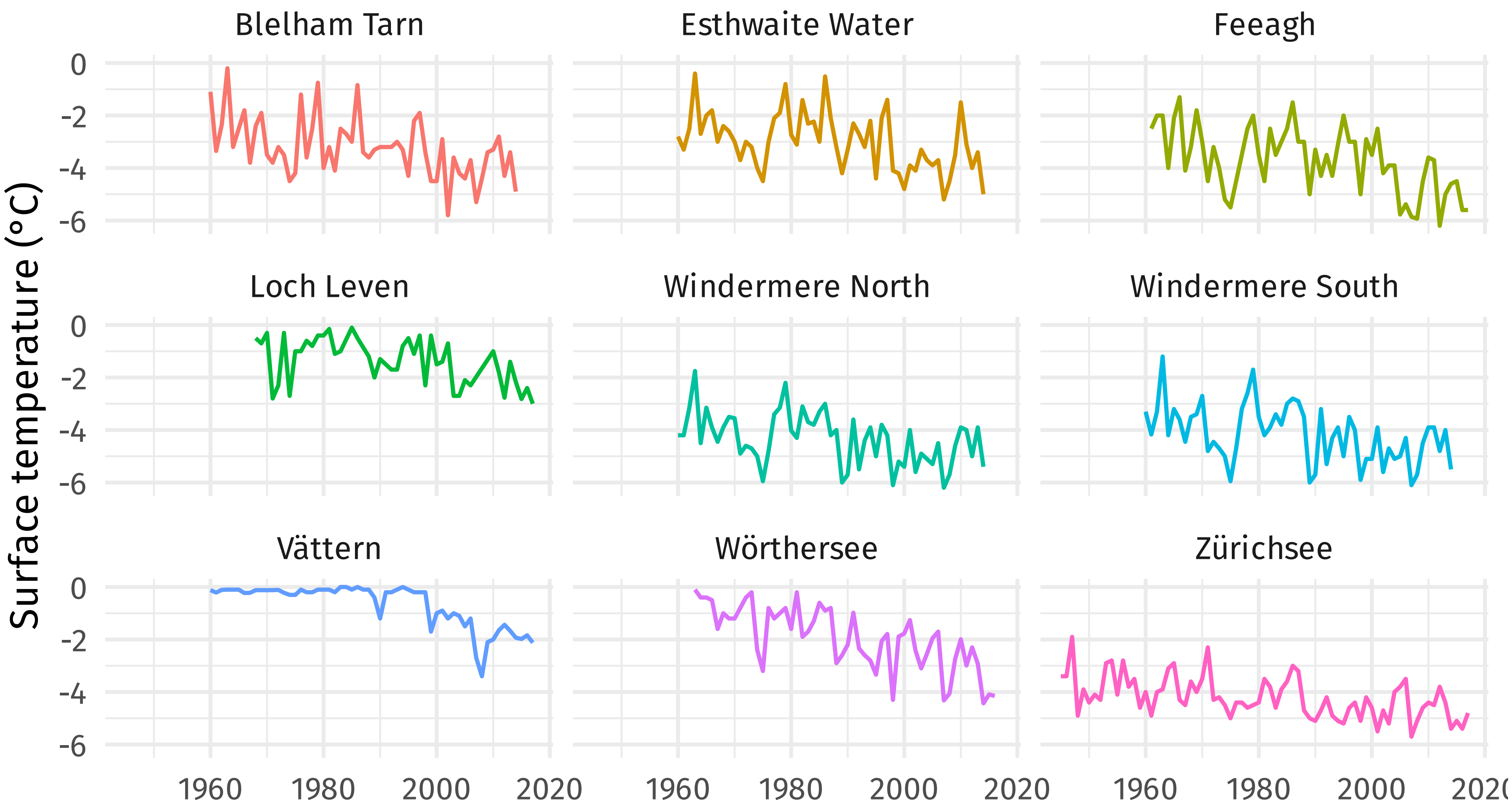
The **maximum** of a sample of *iid* random variables after proper renormalization can only converge in distribution to one of three possible distributions; the *Gumbel* distribution, the *Fréchet* distribution, or the *Weibull* distribution.



Source: Wikimedia Commons

Block Minima...?

Negate the minima ⇒ maxima



Three Distributions...?

Generalised extreme value distribution

In 1978 Daniel McFadden demonstrated the common functional form for all three distributions — the GEVD

$$G(y) = \exp\left\{-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_{+}^{-1/\xi}\right\}$$

Three parameters to estimate

- location μ,
- scale σ, and
- shape ξ

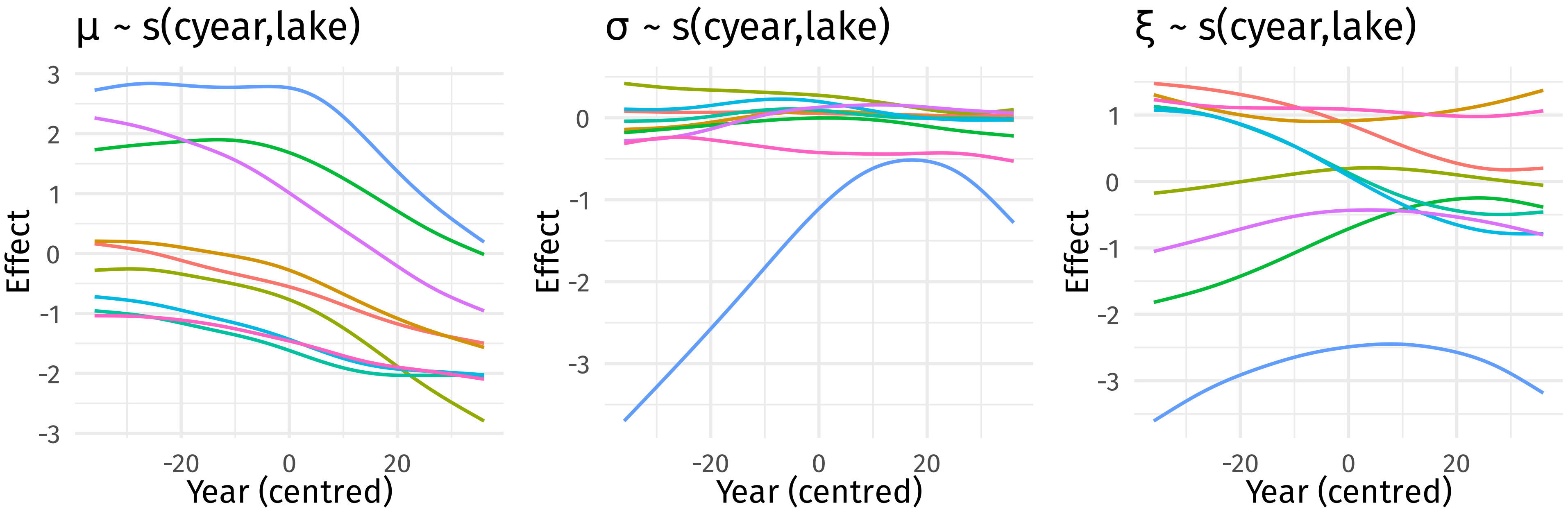
Three distributions

- Gumbel distribution when $\xi = 0$,
- Fréchet distribution when $\xi > 0$, &
- Weibull distribution when $\xi < 0$

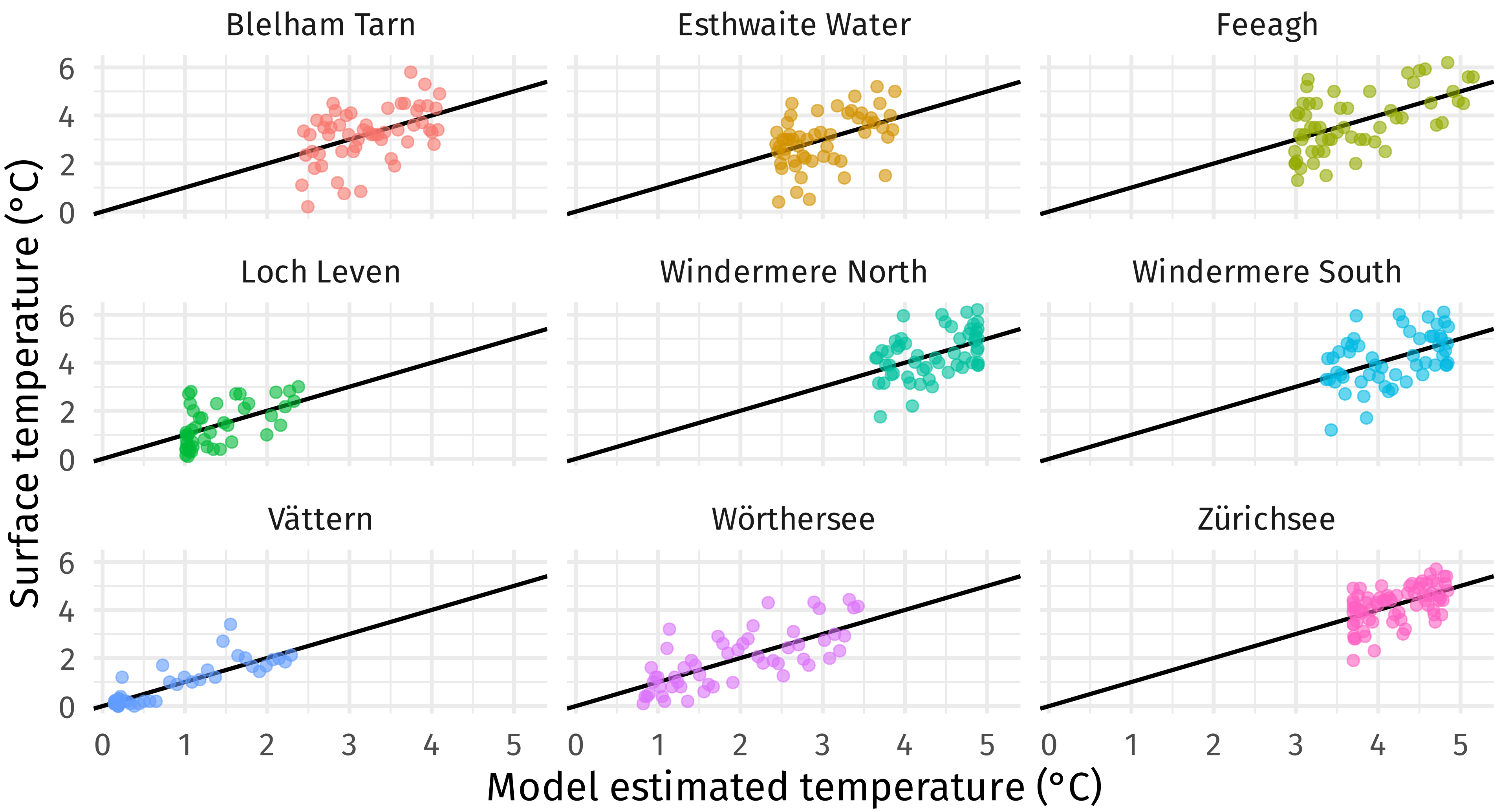
Fit Distributional GAM using GEV for response

Model μ, σ, ξ with smooths of Year

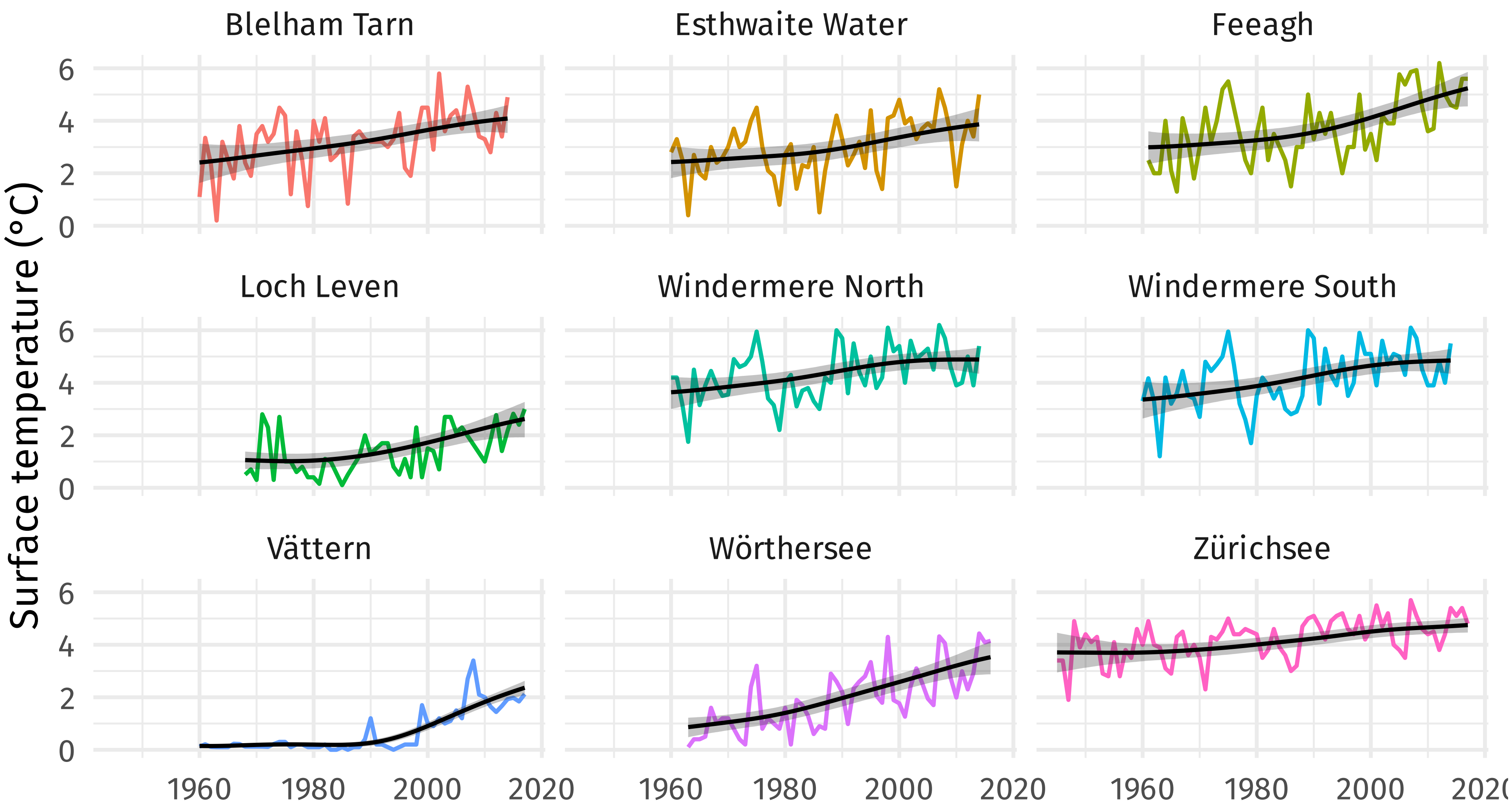
Estimated smooths



Observed vs fitted



Estimated minimum temperature

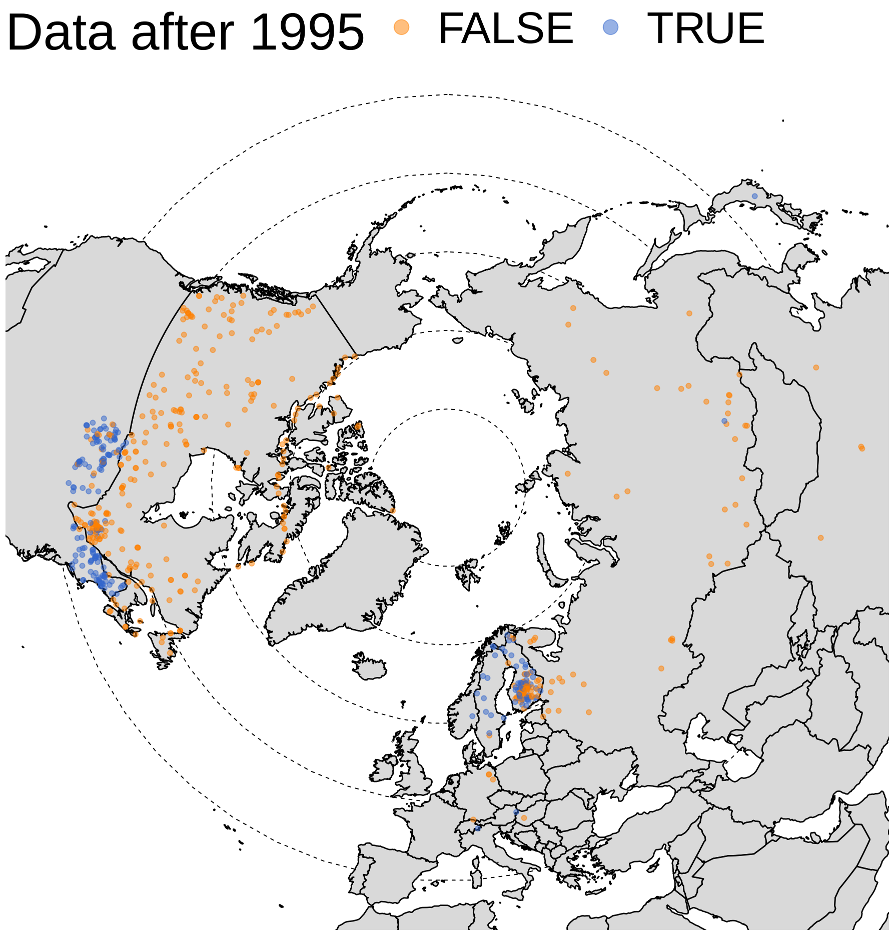


Summary

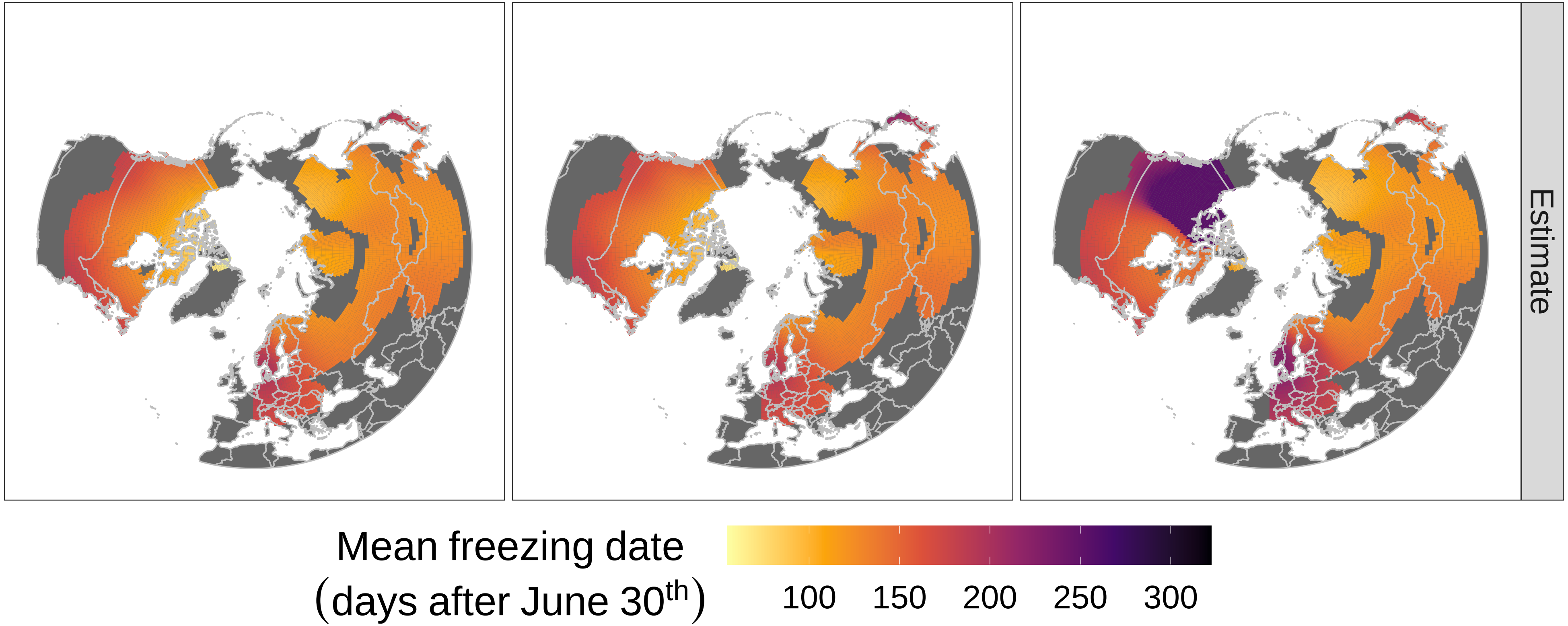
- Lake **minimum** surface water temperatures have increased by on the order of 1–3 degrees over the last 60 years
- Evidence that the distribution of annual minima has changed in many lakes — implications for future extreme events which have long-term knock-on effects
- Distributional GAMs with the GEV distribution are an excellent way of modelling common trends in environmental extremes

Current research

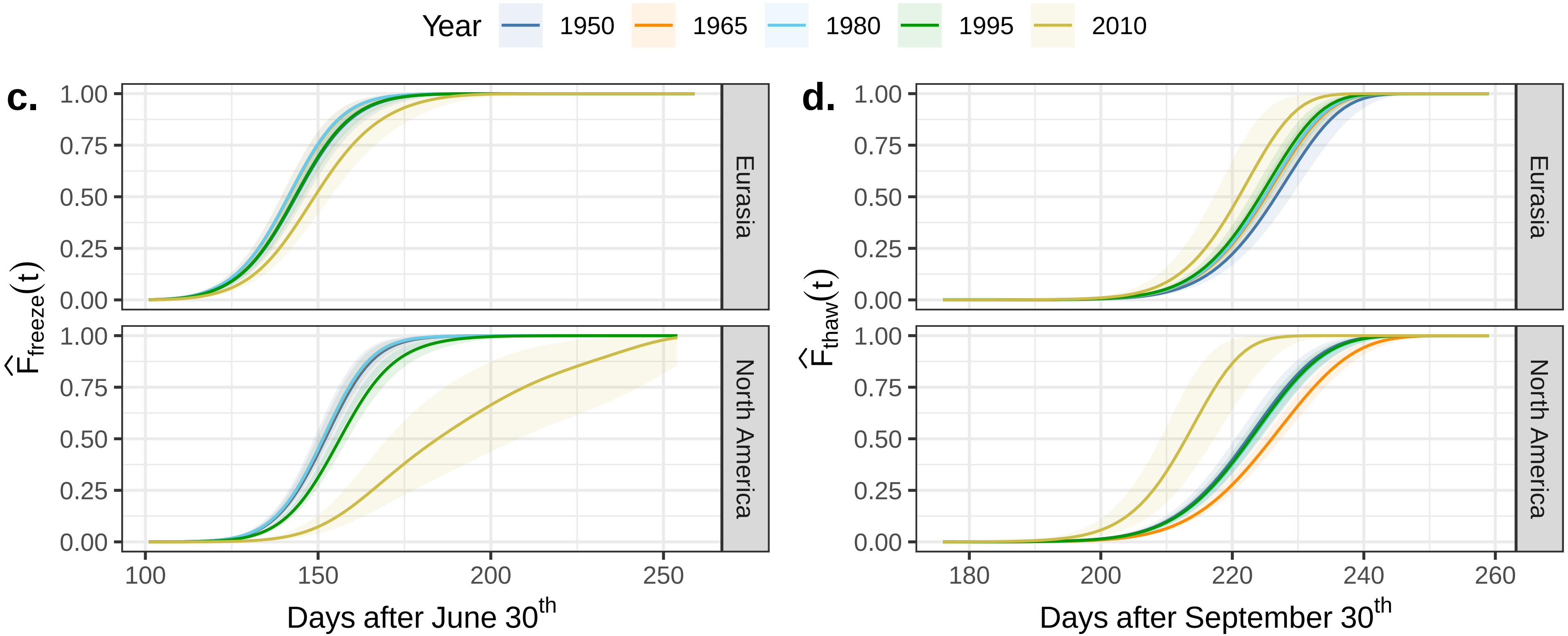
1. The California Cooperative Oceanic Fisheries Investigations (CalCOFI) – Natalya Gallo (UCSD)
2. Changing Ice-Phenology in north hemisphere lakes – Stefano Mezzini (UofR)



Ice Phenology



Ice Phenology



Acknowledgements

Funding

NSERC CRSNG

FACULTY OF GRADUATE STUDIES & RESEARCH

Data

- Peter Leavitt (URegina) for the Lake 227 Data
- Iestyn Woolway and colleagues for archiving the lake surface water data
- Lake ice data from Global Lake and River Ice Phenology Database (GLRIPD, http://nsidc.org/data/G01377.html)

Slides

- HTML Slide deck bit.ly/hifmb-talk