Learning when, where, and by how much, things change

Gavin Simpson

New York R Meetup • June 22 2020

Use statistics to learn from data in presence of noise

Learning from data...?

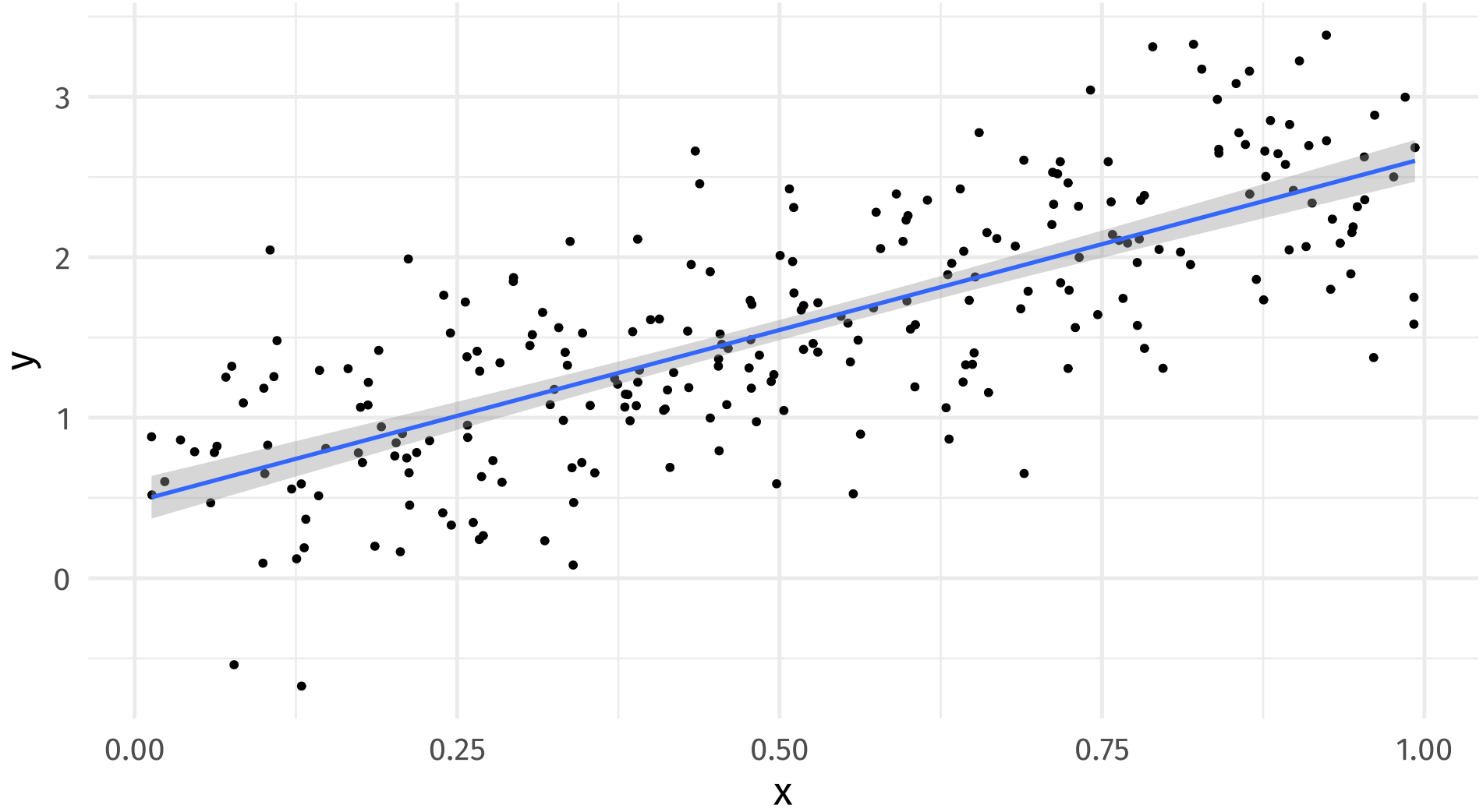
Estimate parameters for a theoretical model

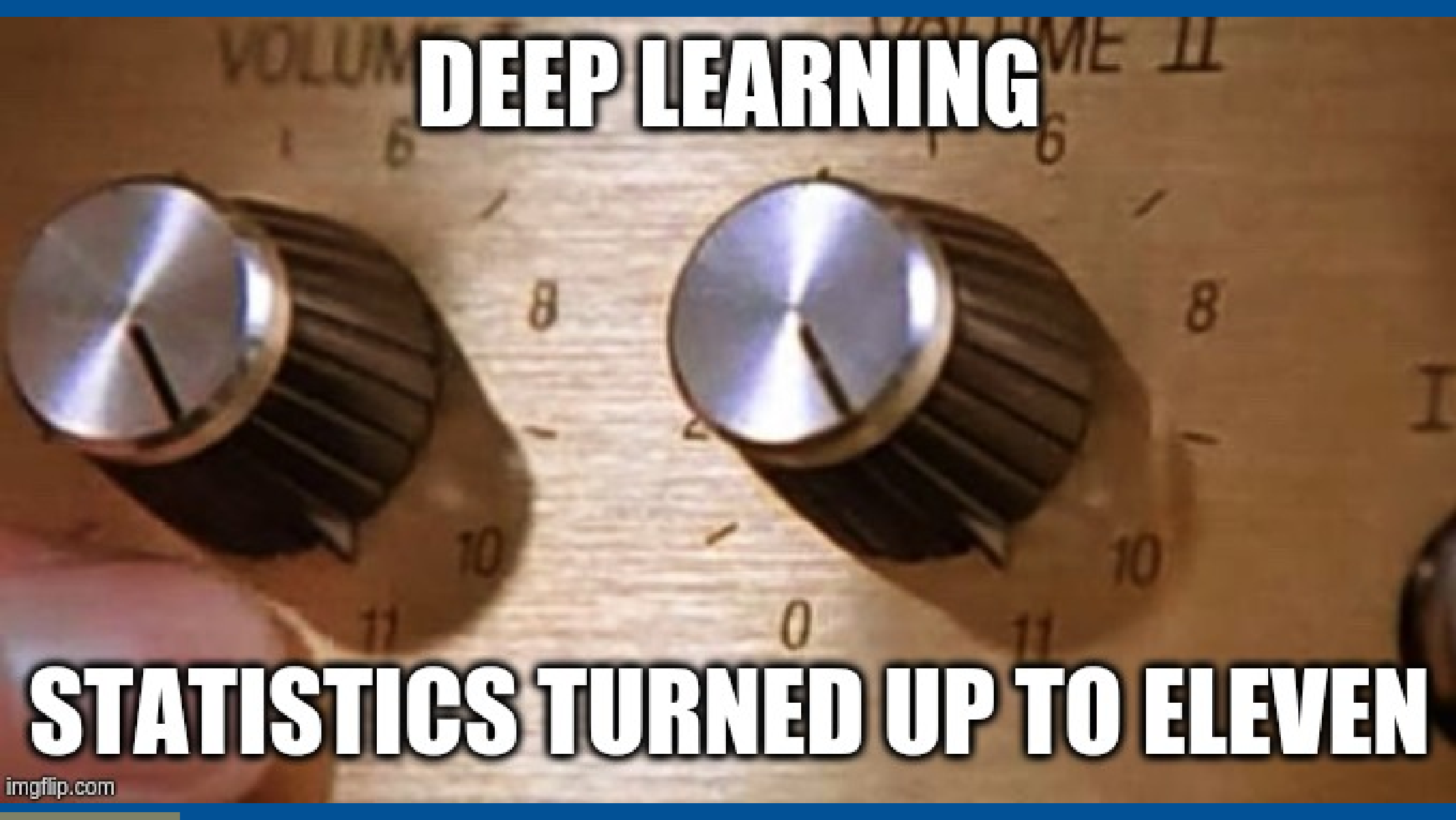
Compare theory with observation

Progress with little or no theory

Franki Chamaki

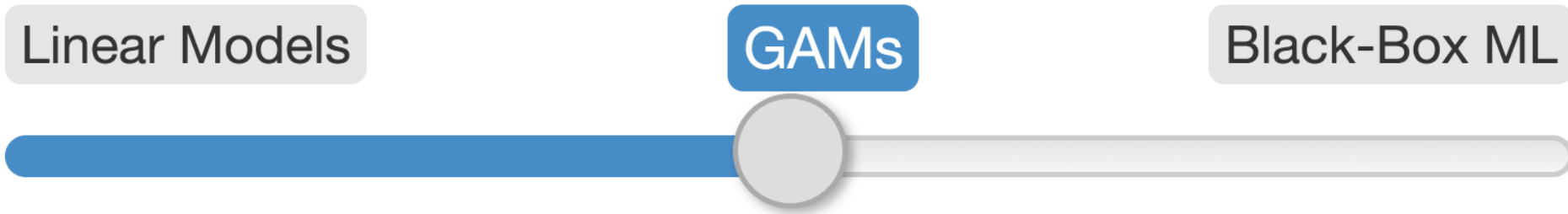
Learning from data




DEEP LEARNING
STATISTICS TURNED UP TO ELEVEN
imgflip.com

Learning involves trade-offs

Generalized Additive Models

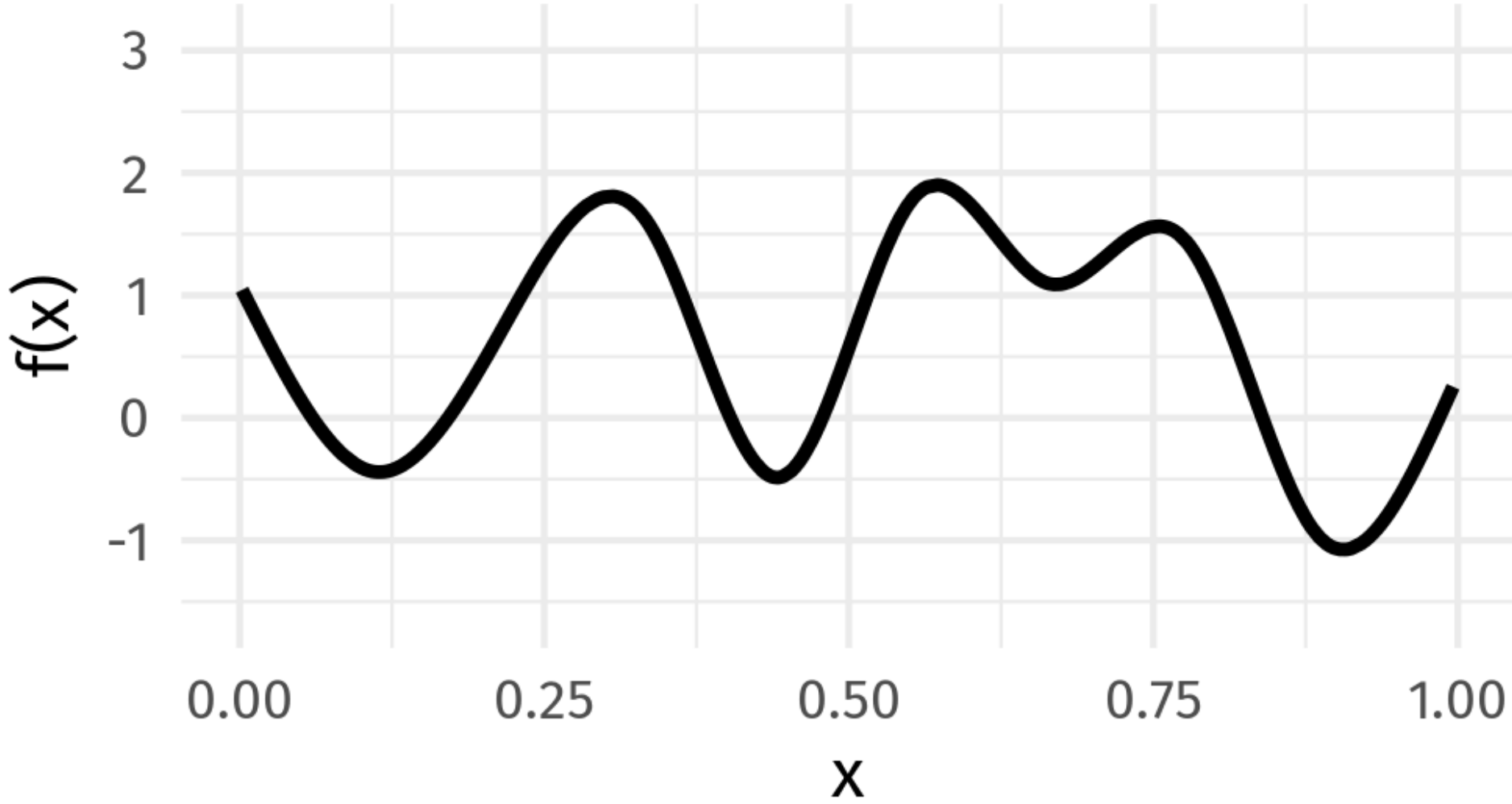


Source: GAMs in R by Noam Ross

GAMs fit wiggly functions


WIGGLY
THINGS

Wiggly things



GAMs

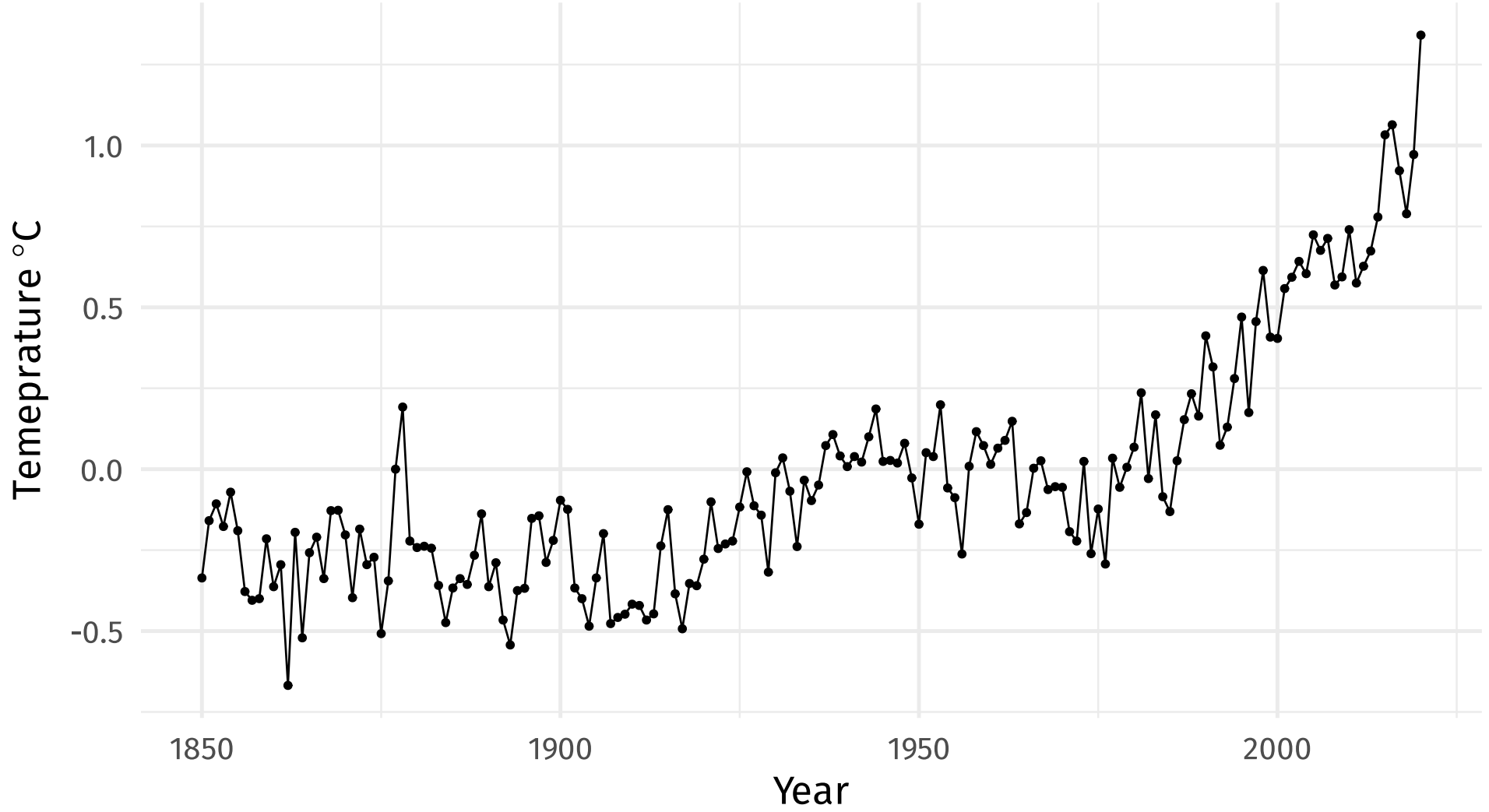
GAMs are not magical

Rob Potter

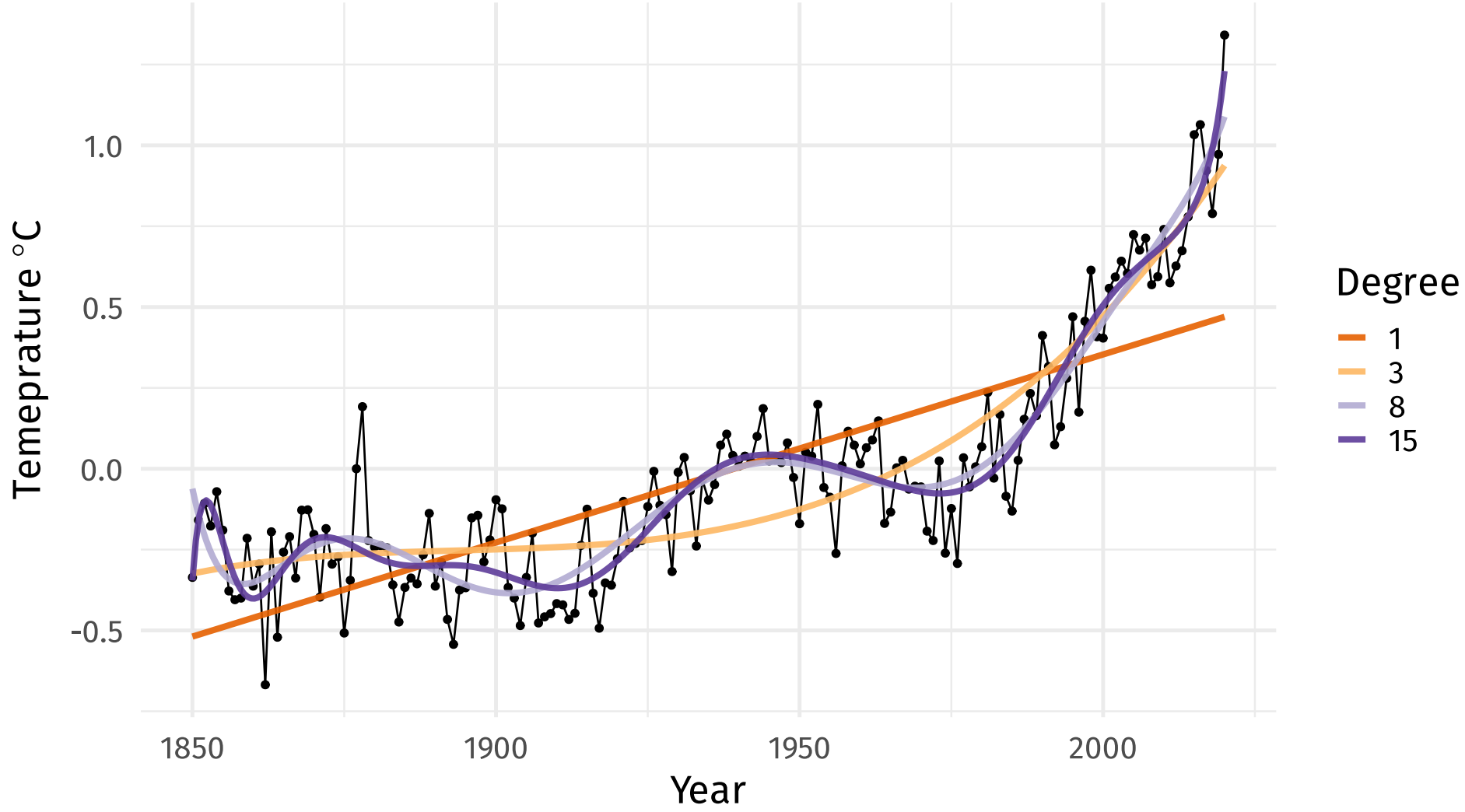
Basis Expansions

Example

HadCRUT4 time series

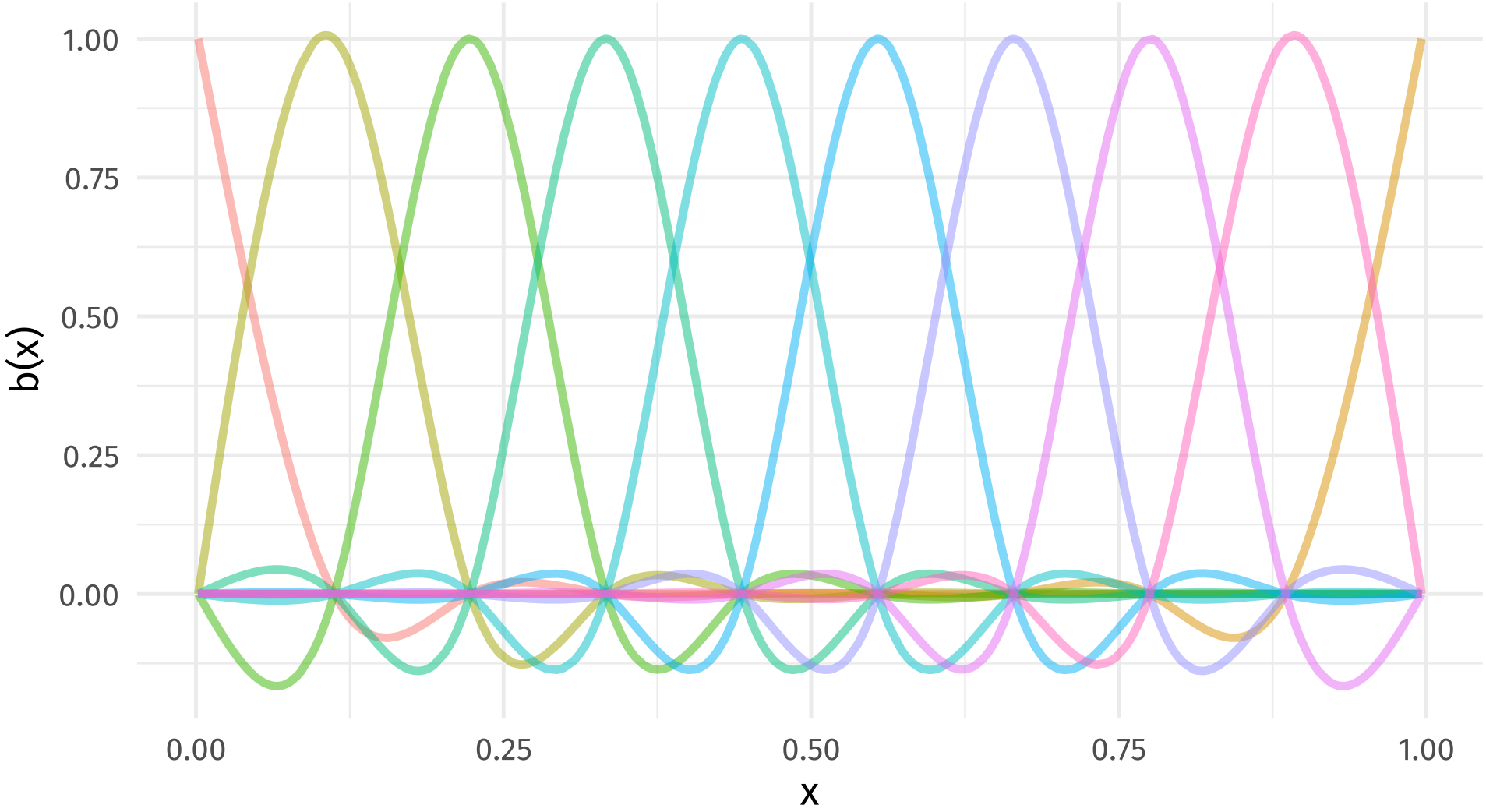


Polynomials

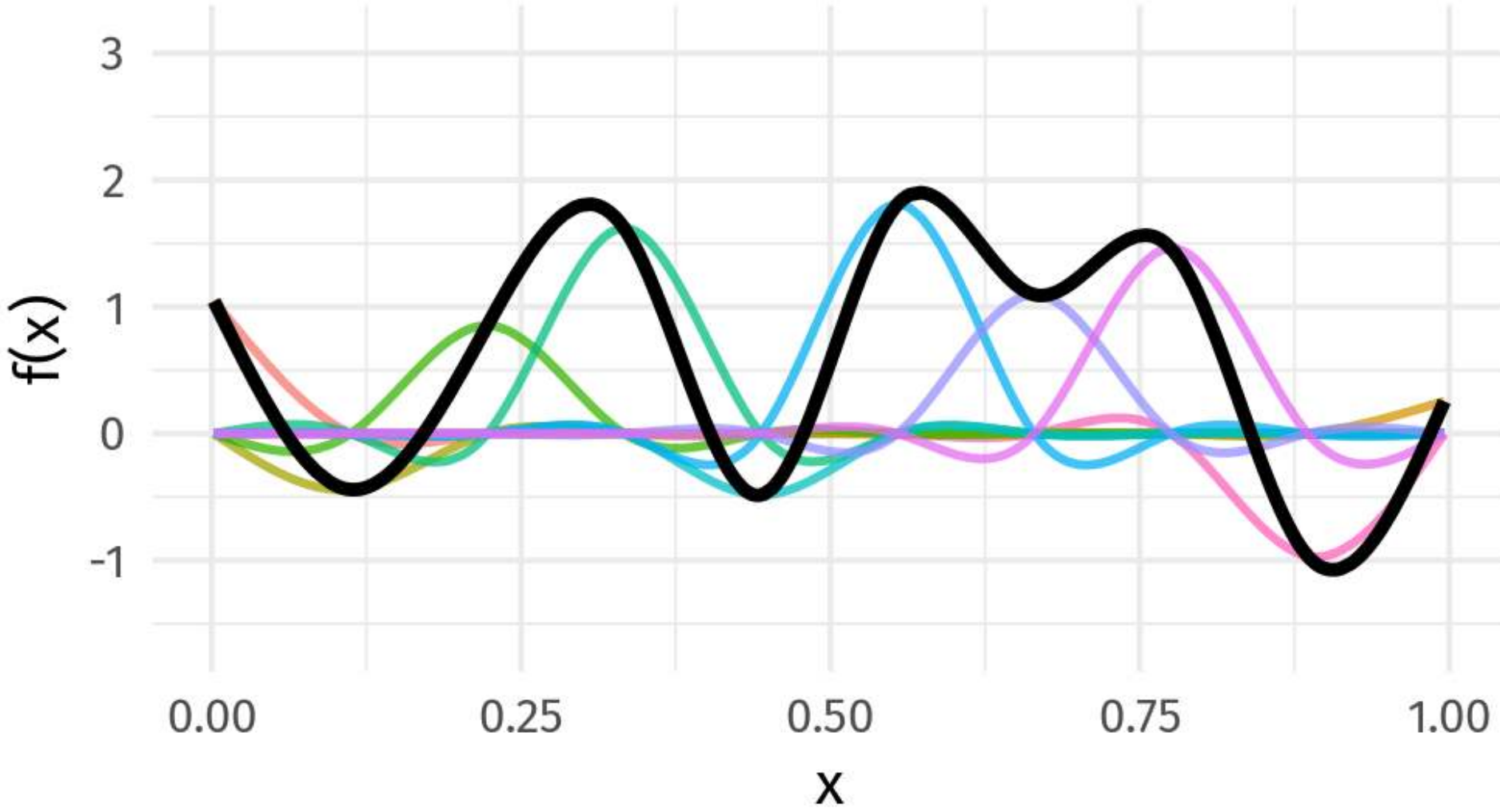


Not that basis expansion

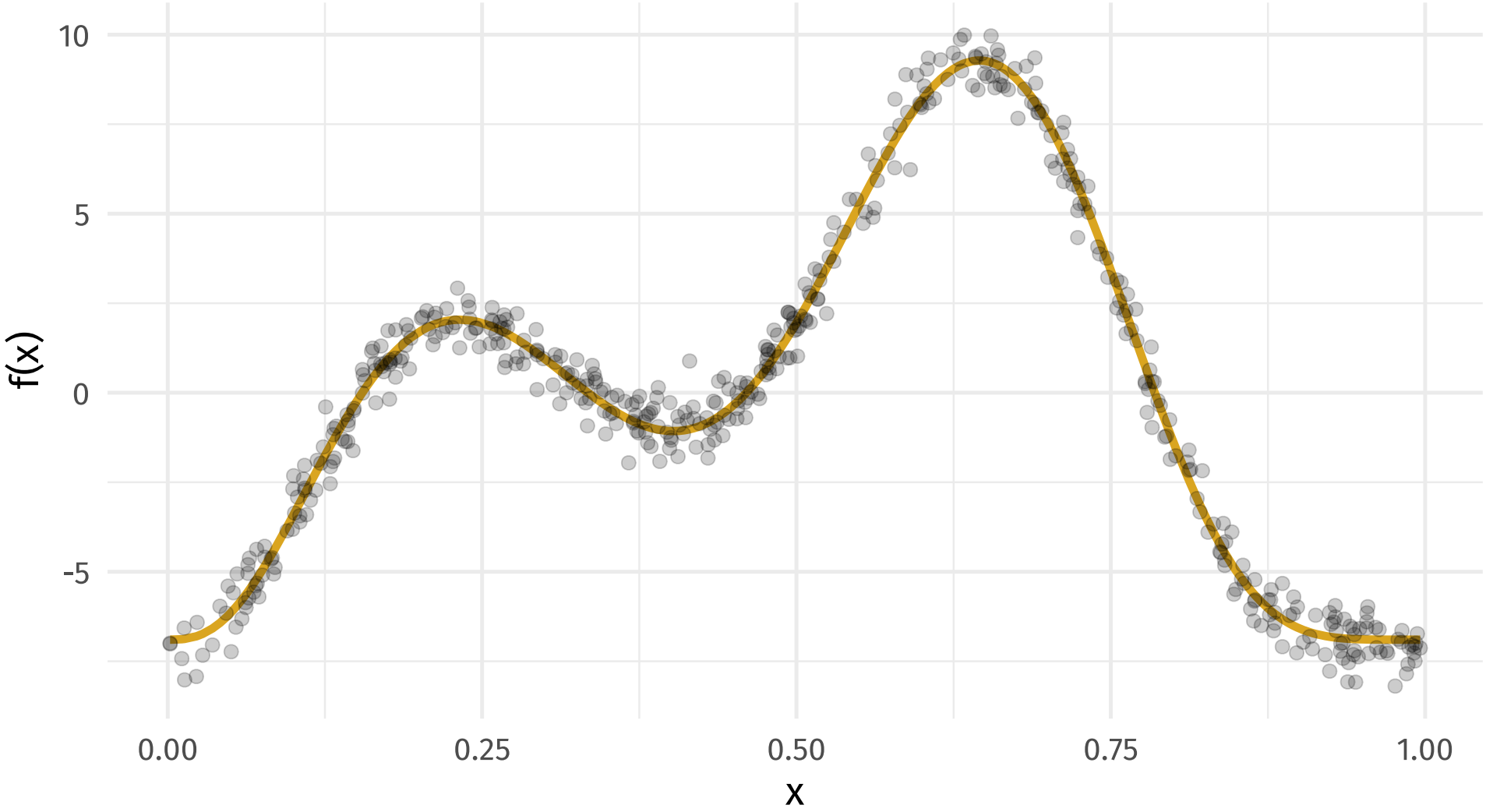
Splines formed from basis functions



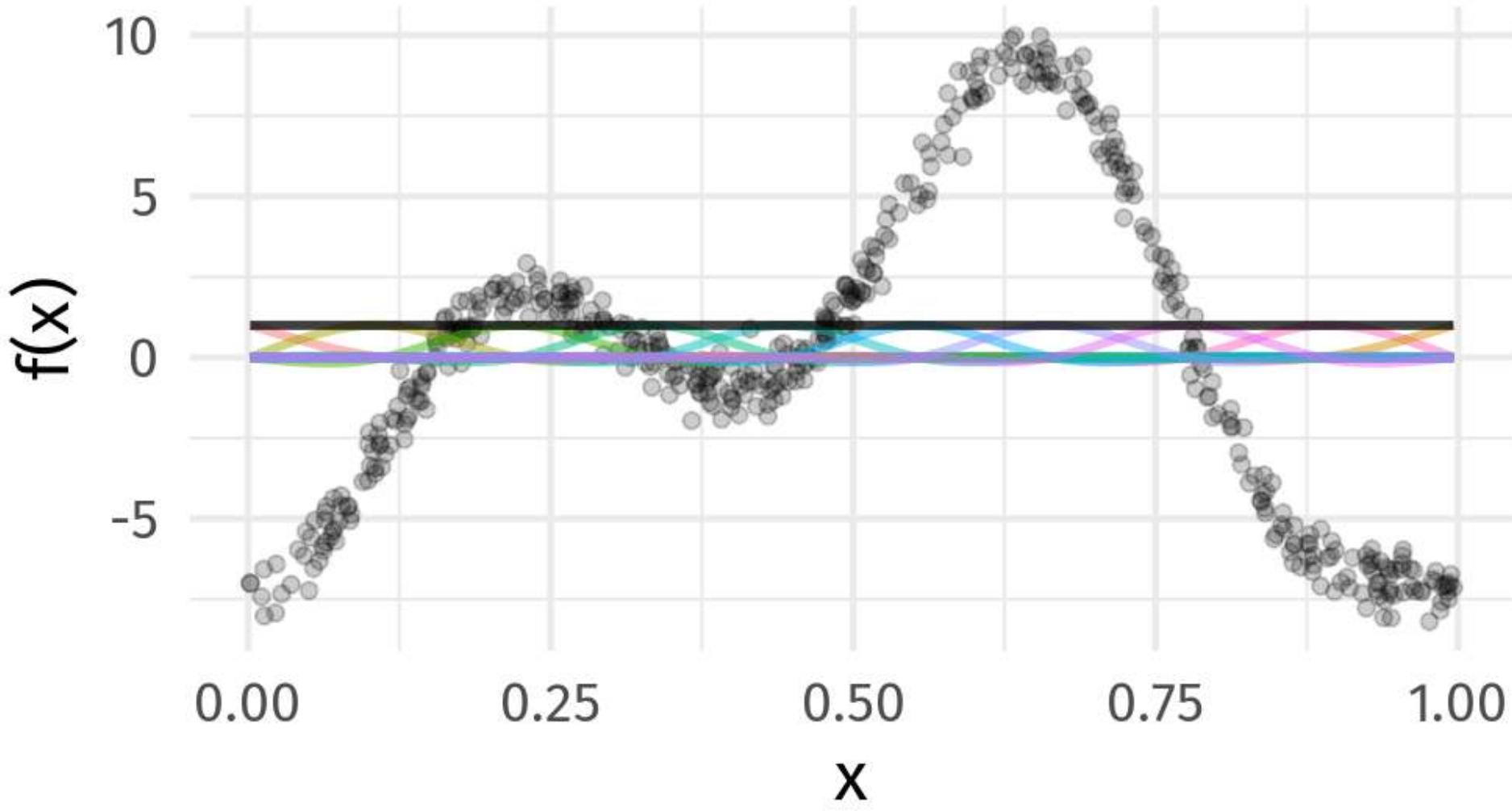
Weight basis functions ⇨ spline



How do GAMs learn from data?



Maximise penalised log-likelihood ⇨ β

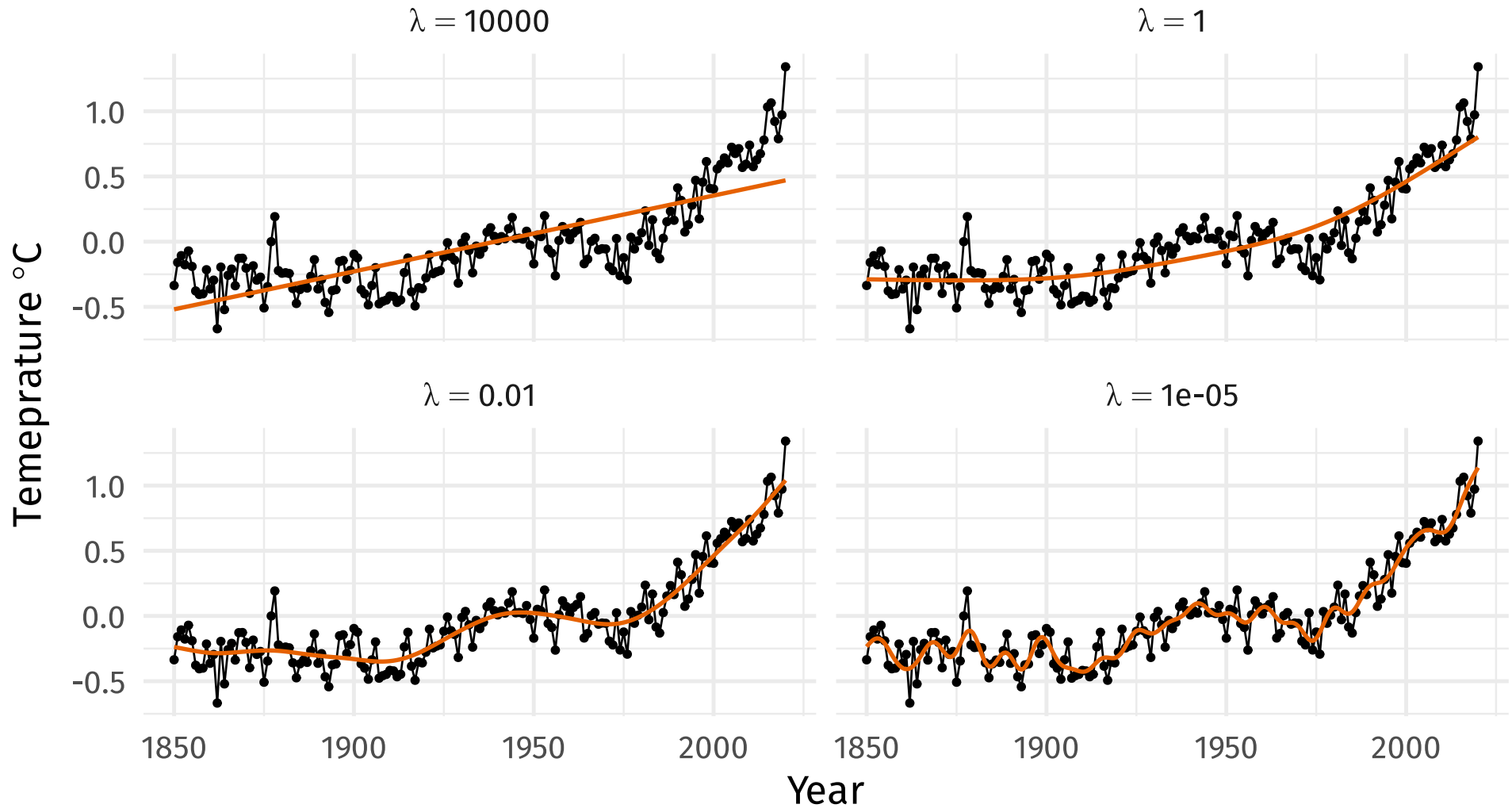


Avoid overfitting our sample

Use a wiggliness penalty — avoid fitting too wiggly models

Example

HadCRUT4 time series



OK some math

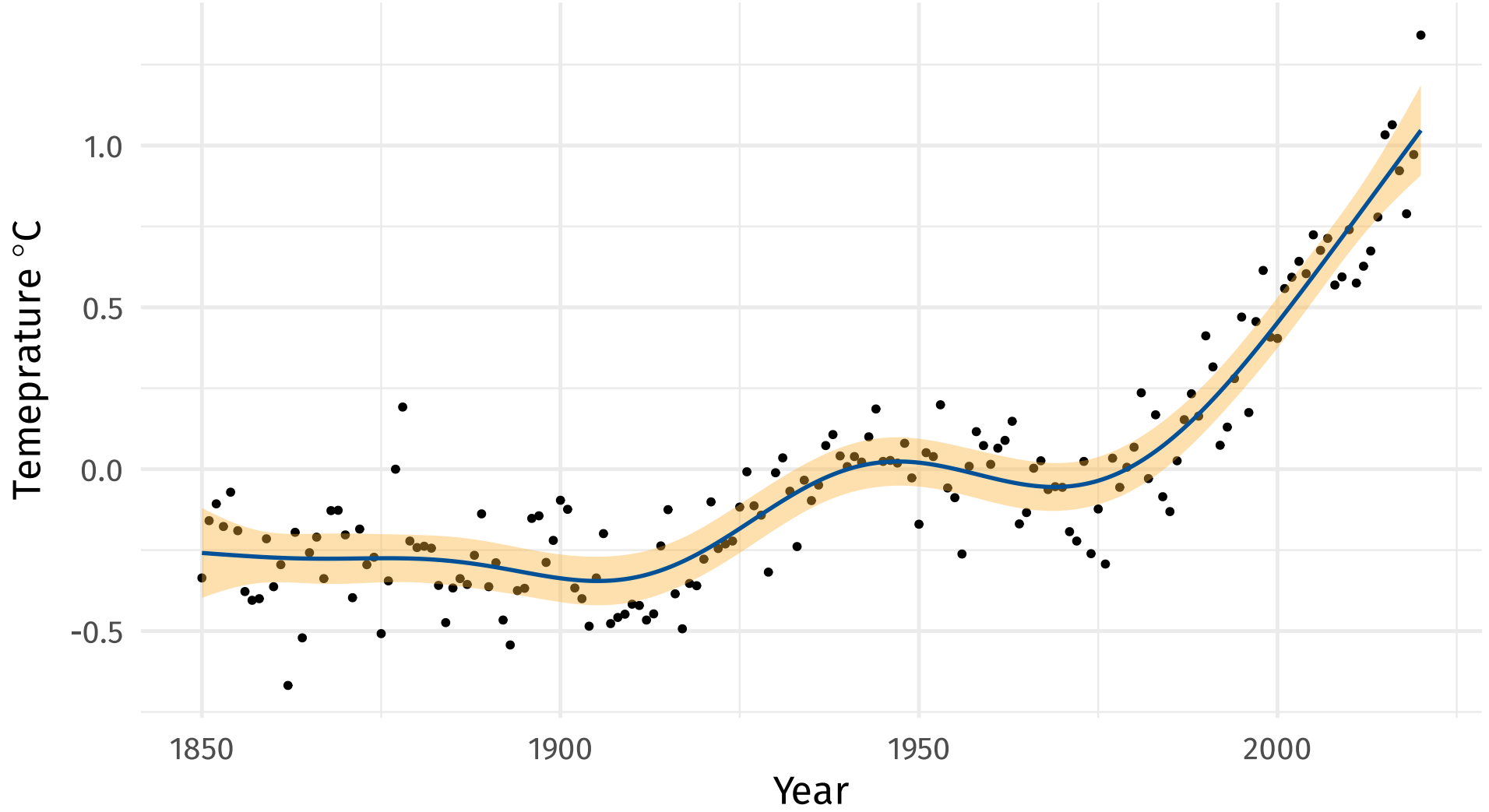
How wiggly?

$$\int_{\mathbb{R}} [f'']^2 dx = \boldsymbol{\beta}^\mathsf{T} \mathbf{S} \boldsymbol{\beta}$$

Penalised fit

$$\mathcal{L}_p(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - \frac{1}{2}\lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$$

Fitted GAM




mgcv

Fitting GAMs in *mgcv*

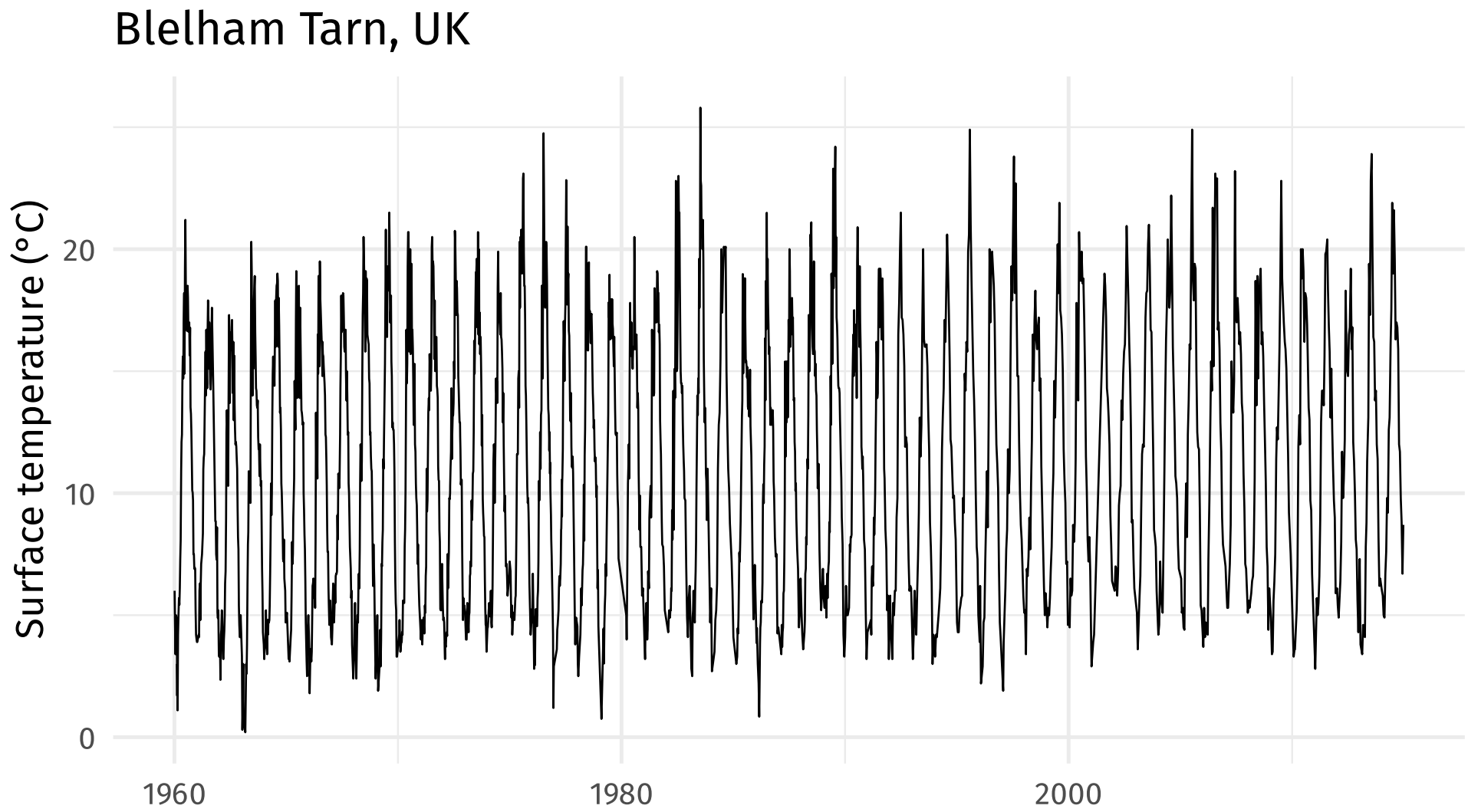
Wrap a variable in `s( )` to get a smooth

```
m ← gam(Temperature ~ s(Year), data = gtemp, method = "REML")
```

Fit using REML or ML (`method = "ML"`) smoothness selection

GCV can undersmooth but it's the default!

Climate change affecting lake temperatures?



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)

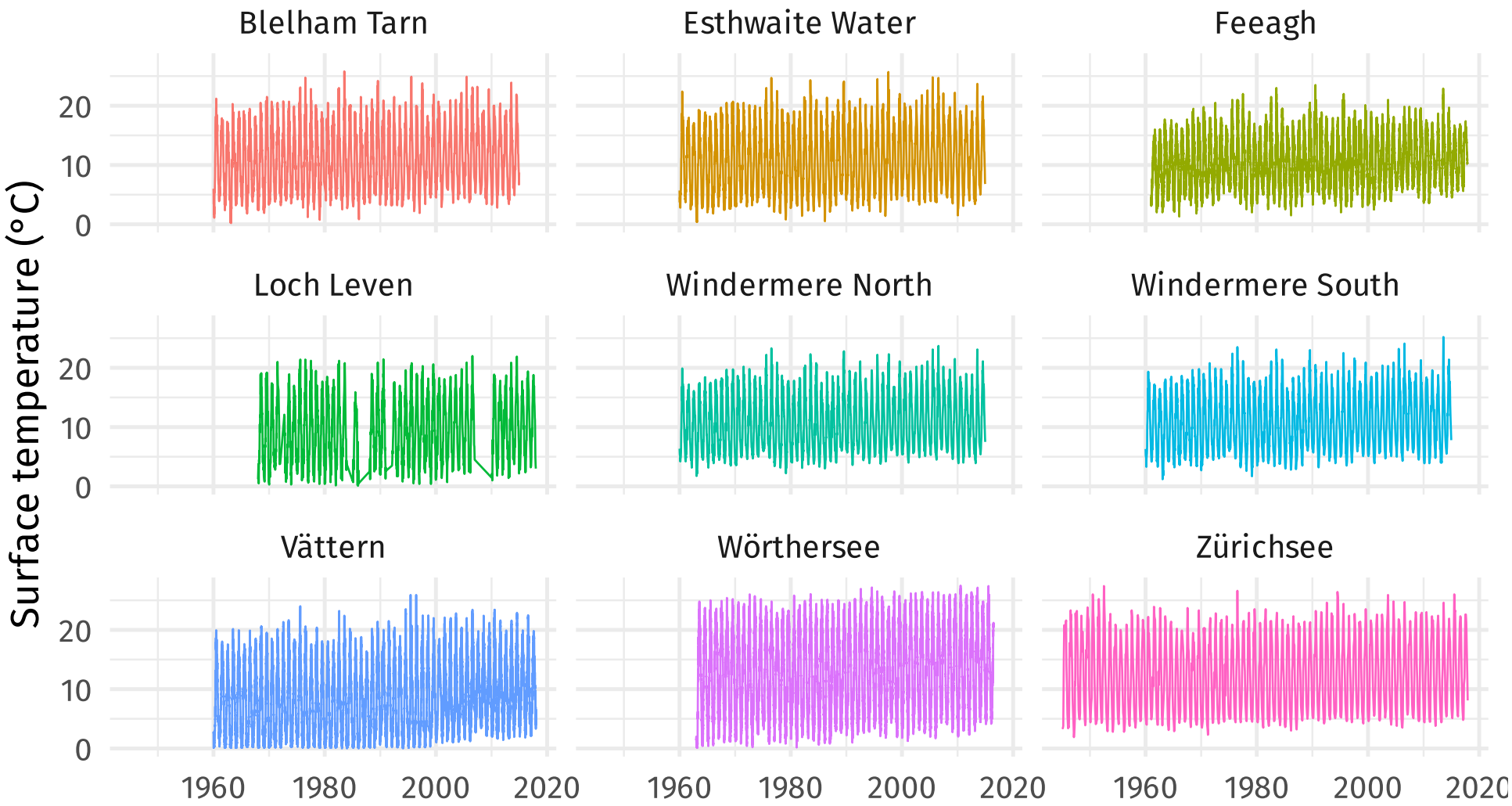
Reduction in habitat or refugia for cold-adapted species

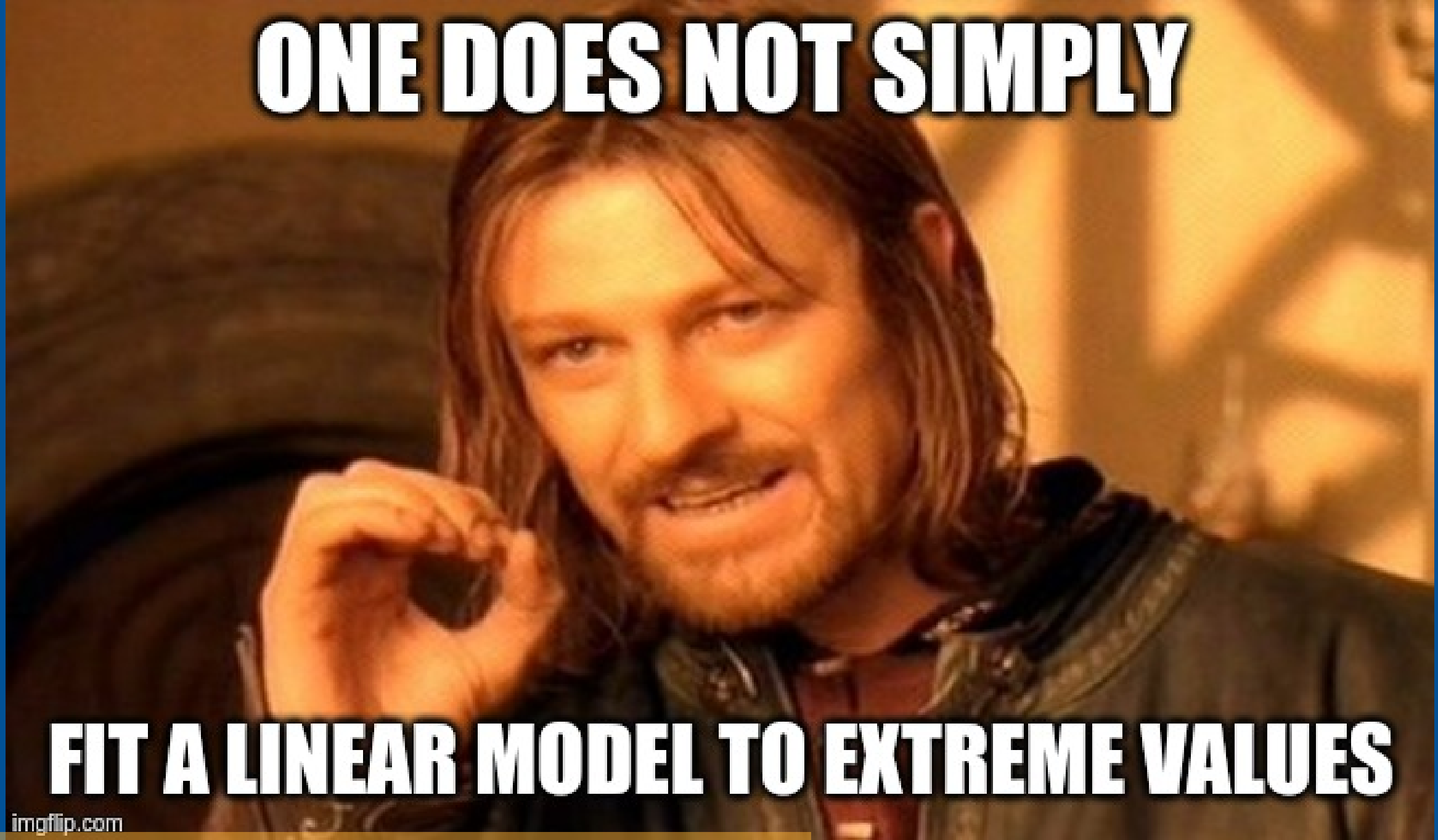
- Arctic charr (*Salvelinus alpinus*)
- Opossum shrimp (*Mysis salemaai*)

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

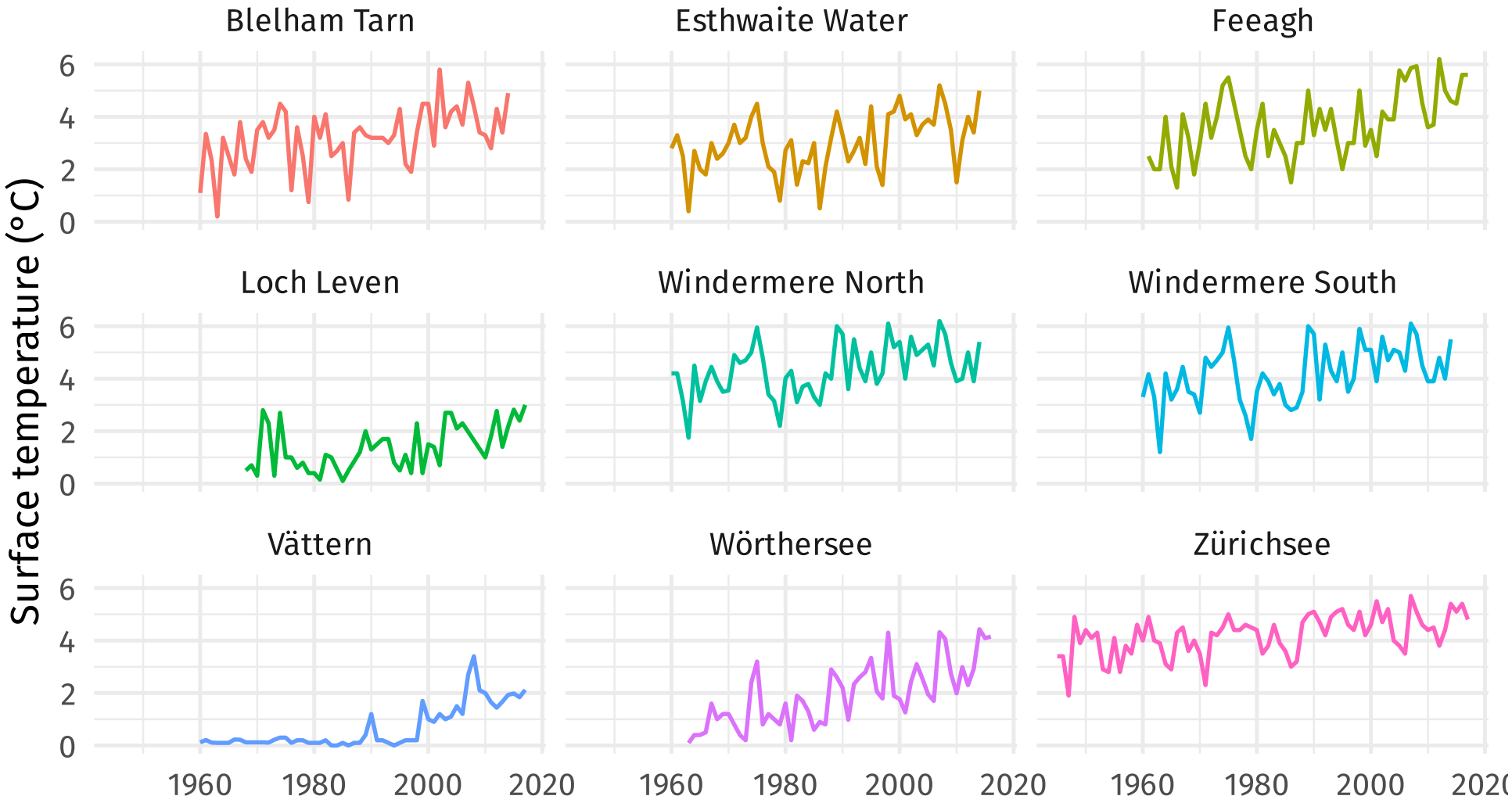
Multiple time series ⇨ HGAM




ONE DOES NOT SIMPLY
FIT A LINEAR MODEL TO EXTREME VALUES
imgflip.com

Central Limit Theorem

Annual minimum temperature



Block Minima

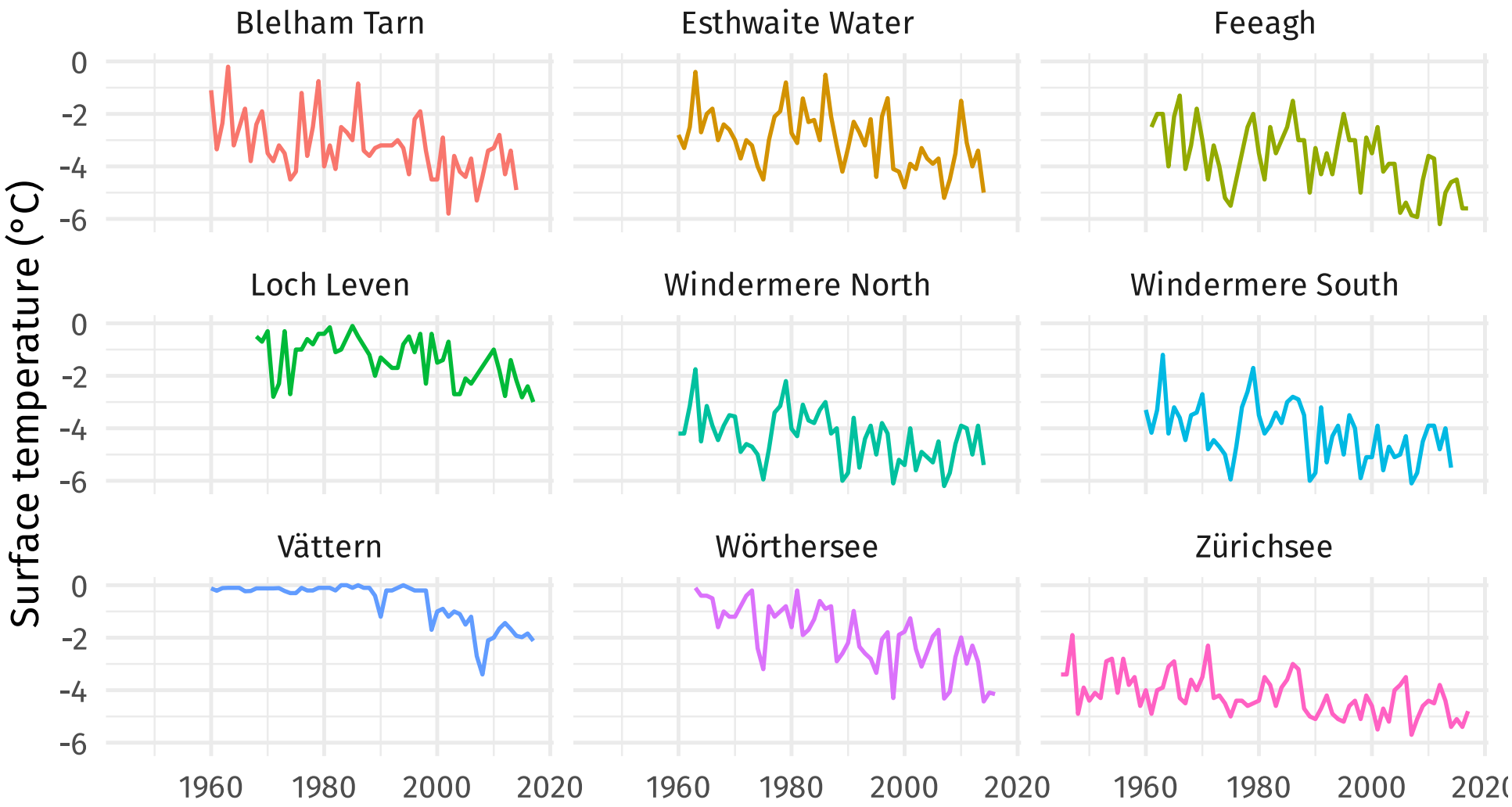
Fisher–Tippett–Gnedenko theorem

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.

Block Minima...?

Highly Technical Fix

Negate the minima ⇨ maxima



plus some jiggery-pokery after model fitting

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

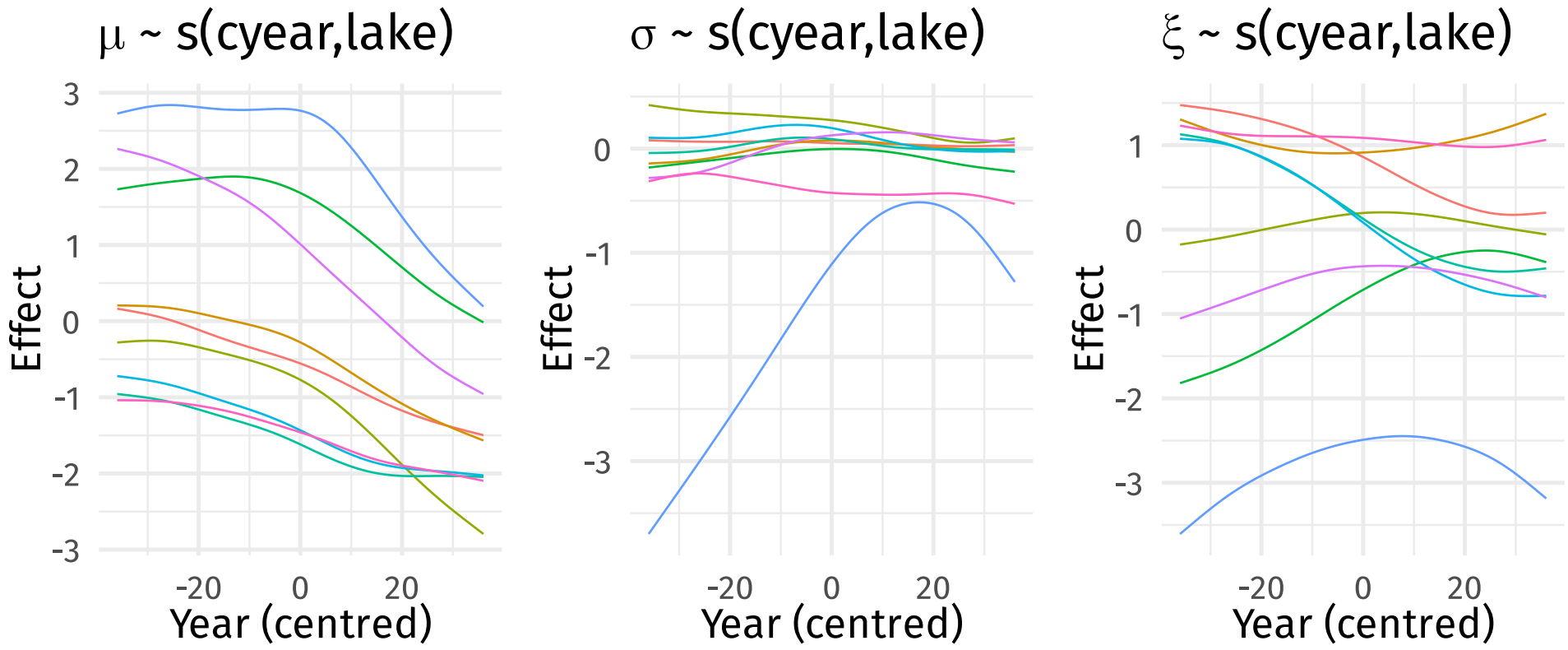
- Gumbel distribution when $\xi = 0$,
- Fréchet distribution when $\xi > 0$, &
- Weibull distribution when $\xi < 0$

Fit HGAMLSS using GEV for response

HGAMLSS...?

Model μ, σ, ξ with smooths of Year

Estimated smooths



Model code

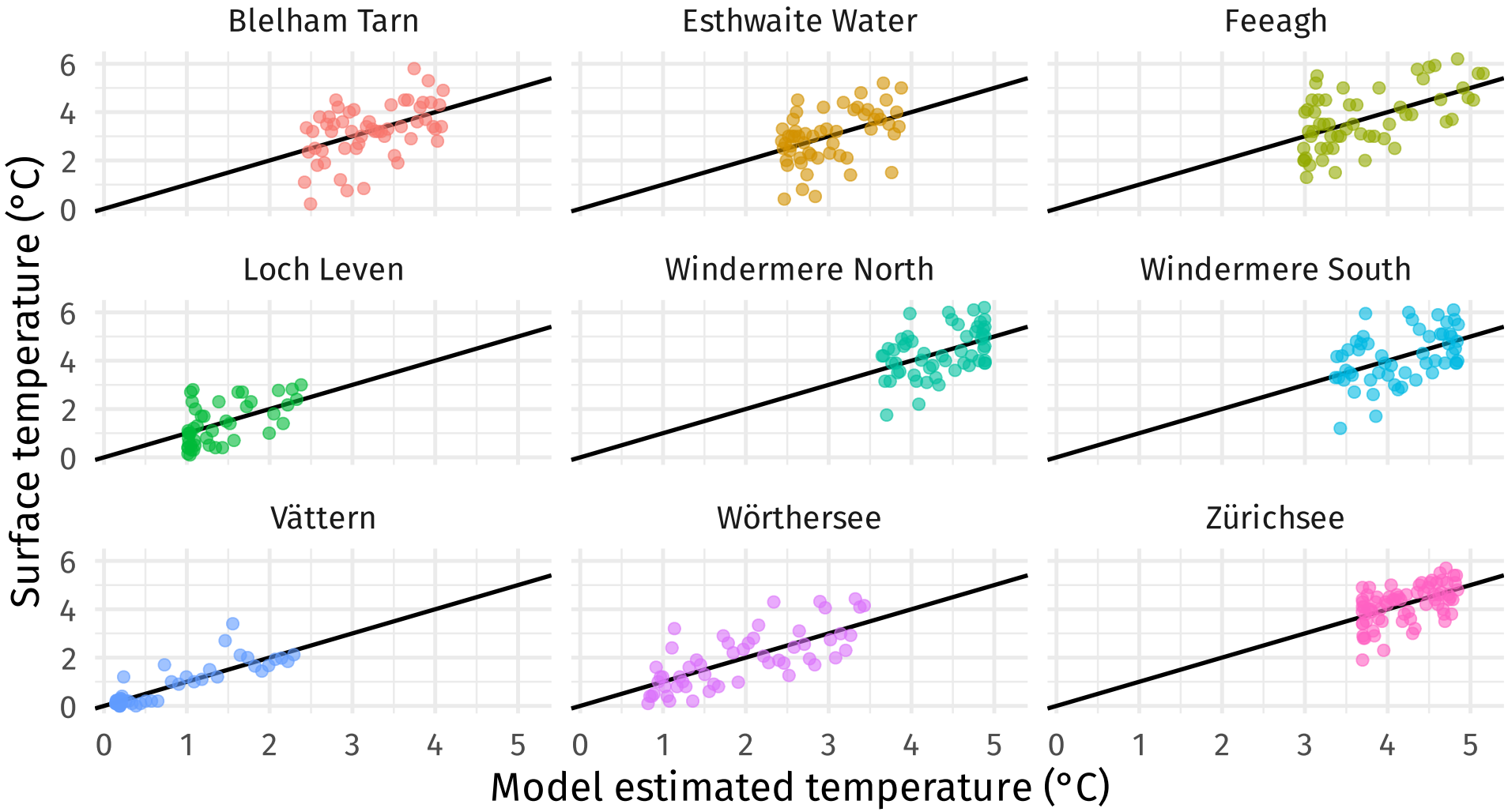
Provide a list of formulas

```
m1 ← gam(list(neg_min ~ s(cyear, lake, bs = 'fs'),
              ~ s(cyear, lake, bs = 'fs'),
              ~ s(cyear, lake, bs = 'fs')),
         data = minima, method = 'REML',
         family = gevlss(link = list('identity', 'identity', 'logit')),
         control = ctrl, optimizer = 'efs')
```

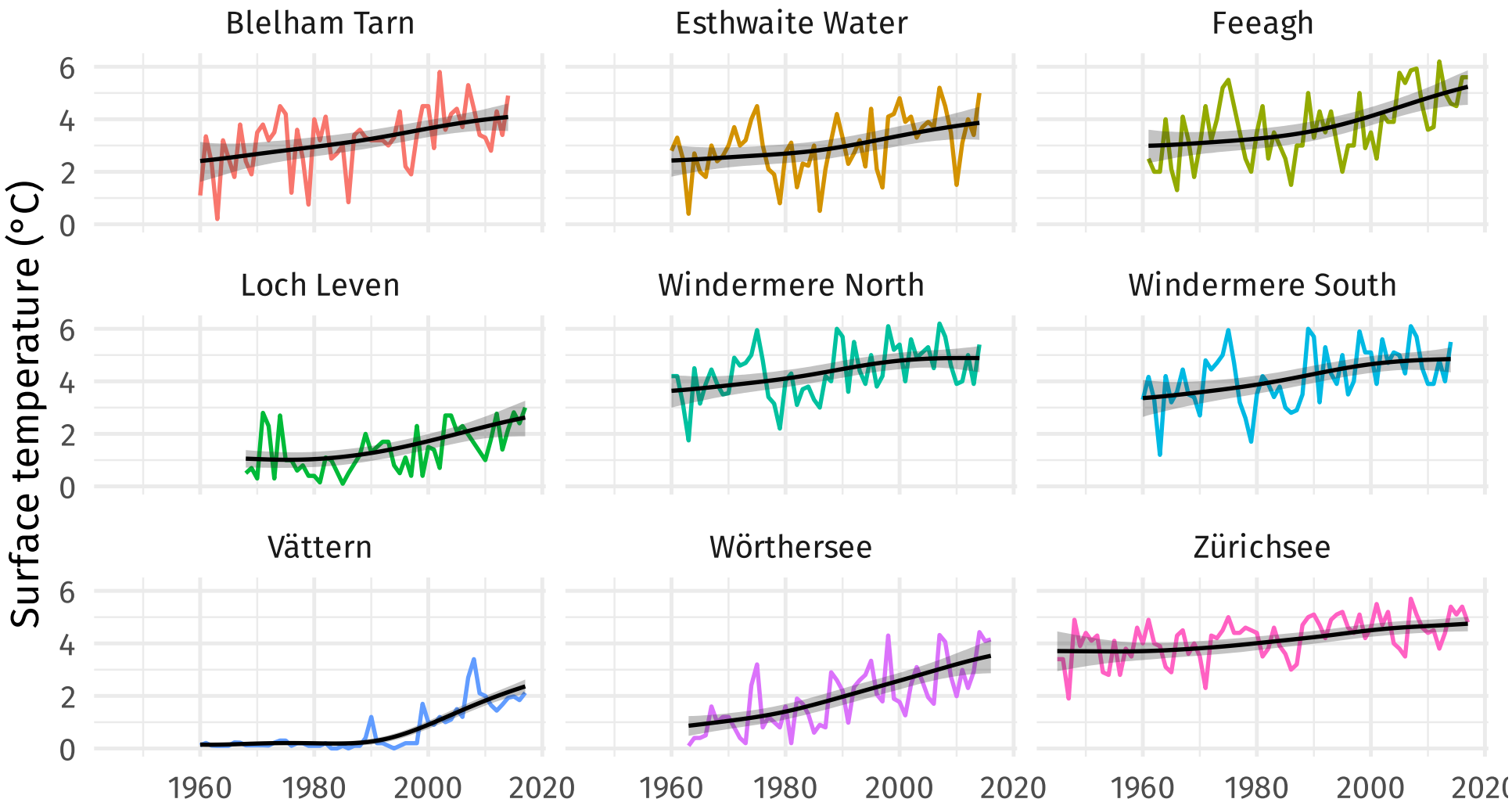
`bs = "fs"` is a factor-smooth interaction

- like a ransom slope & intercept but for a spline
- one spline per `lake`
- one smooth parameter

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
- HGAMLSS with the GEV distribution are a good way of modelling common trends in environmental extremes


brms

Fully Bayes

mgcv fits empirical Bayesian models with REML or ML smoothness selection

Improper Gaussian priors — we don't penalise the linear bits of the basis

We can fit fully Bayesian models using *brms* with (almost) all the smooths from *mgcv*

Can't use `te( )` or `ti( )` for tensor product smooths (smooth interactions)

Can use `t2( )` though

Microcystin

Microcystin

A liver toxin produced by cyanobacteria

Frequent cause of negative human health effects, kills dogs, etc

Cyanobacteria can bloom under the right conditions — HABs

Increases in HABs globally driven by nutrient pollution & climate change

11 years of bi-weekly data from Qu'Appelle Valley in Saskatchewan, Canada

Hayes, N.M. *et al*, 2020. *Limnol. Oceanogr. Let.* 58, 1736. https://doi.org/10.1002/lol2.10164

Non-detects

Fitting GAMs in brms

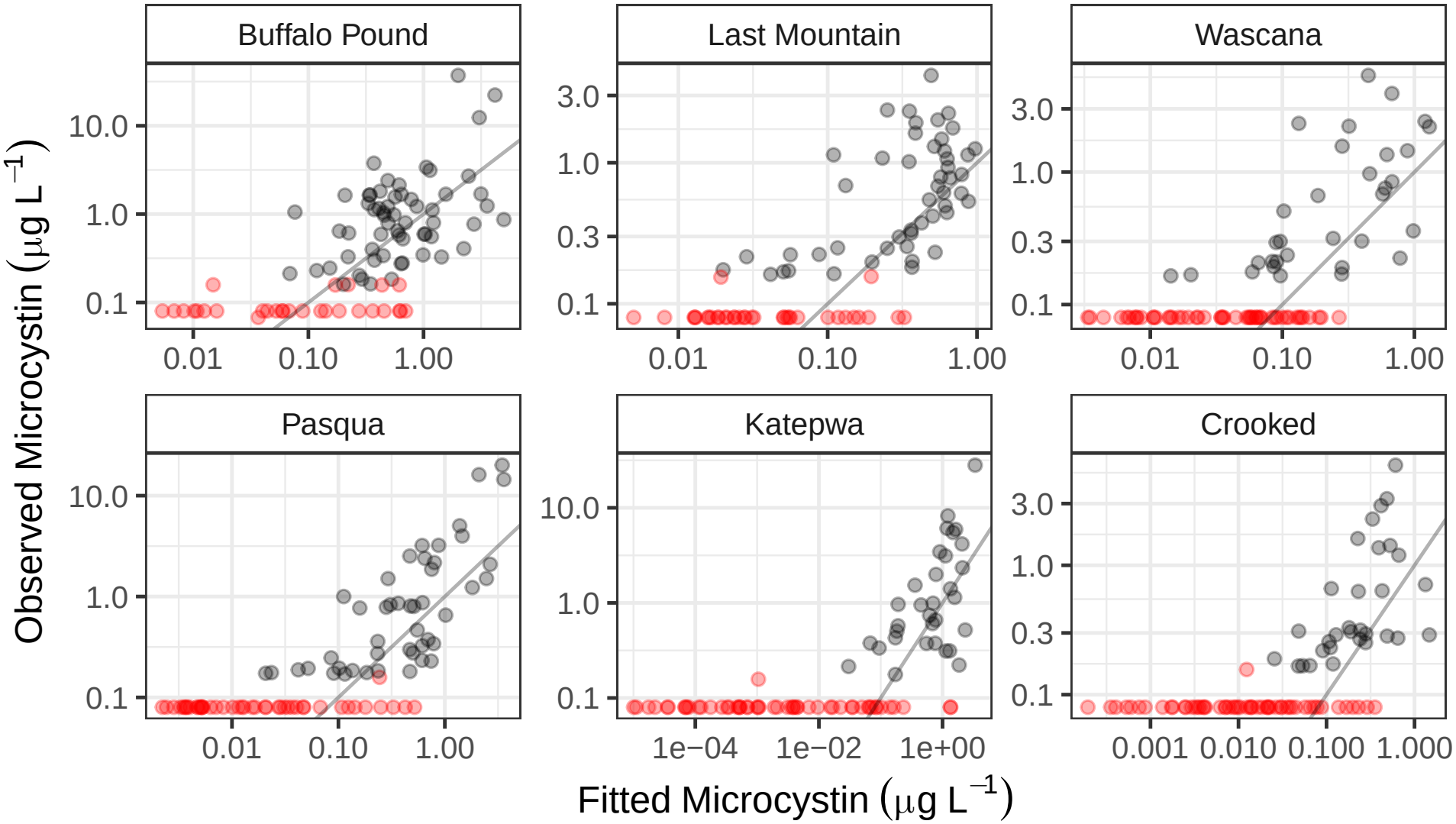
non-censored version in *mgcv*

```
mgcv_mod ← gam(micro_censored ~ lake + te(DOY, cYear, by = lake),
               data = dfd, family = Gamma(link = "log"),
               method = "REML"
               control = ctrl)
```

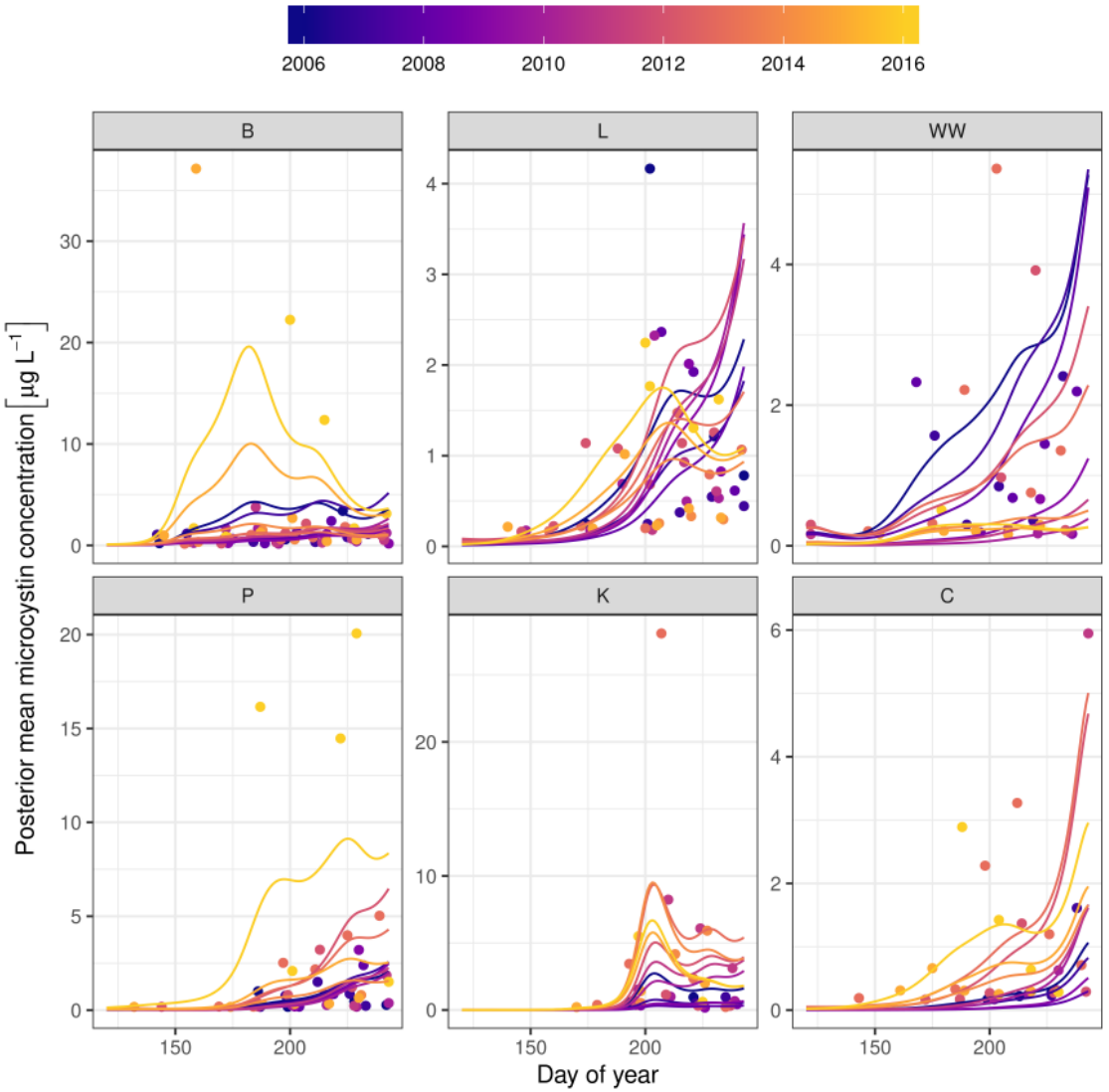
censored version in *brms*

```
brms_mod ← brm(microcystin | cens(censored) ~ lake + t2(DOY, cYear, by =
lake),
              data = dfd, family = Gamma(link = "log"),
              warmup = 1000, iter = 3000, chains = 4, cores = CORES,
              seed = 8354, control = list(adapt_delta = 0.99))
```

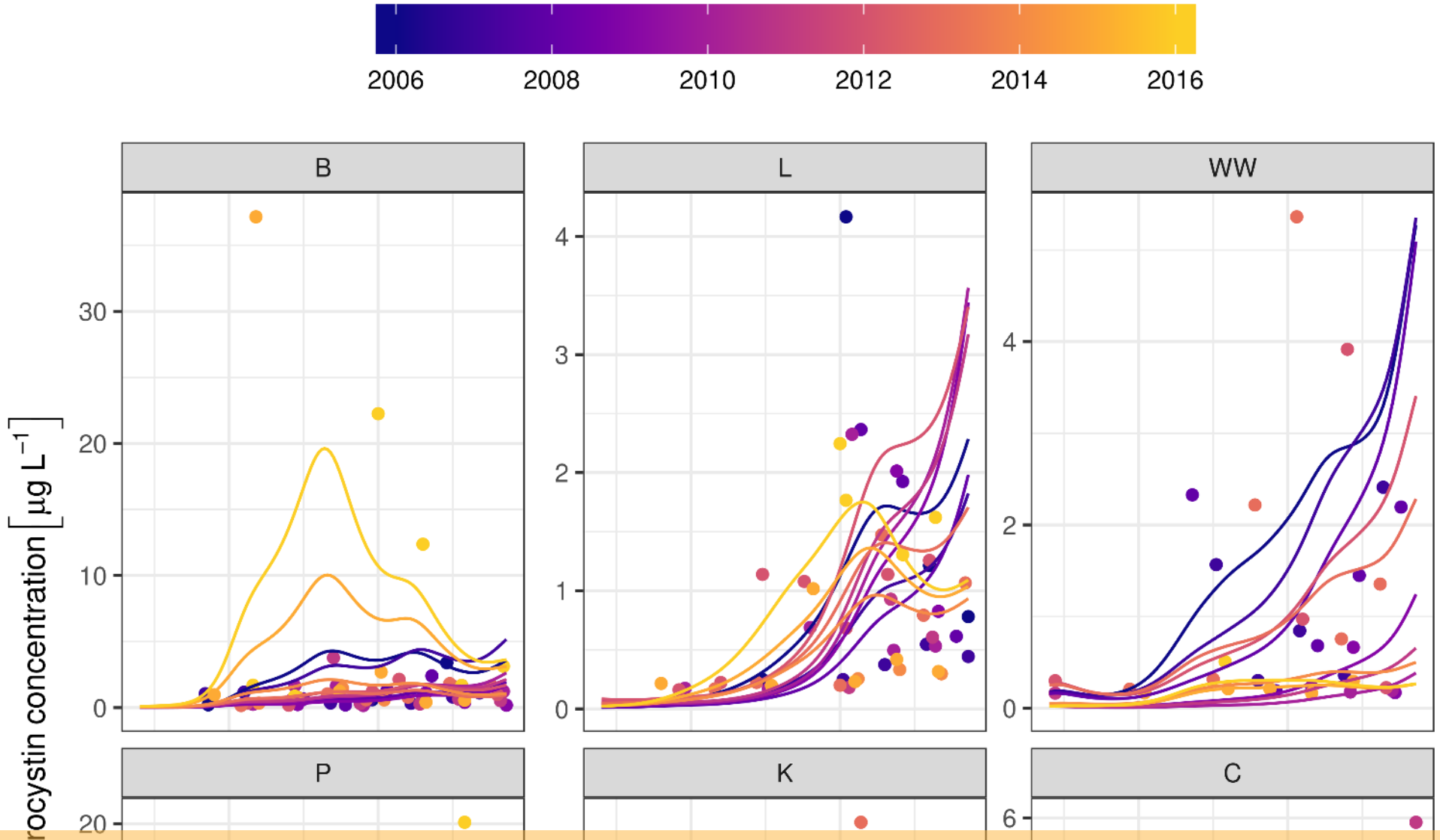
Fitted Microcystin



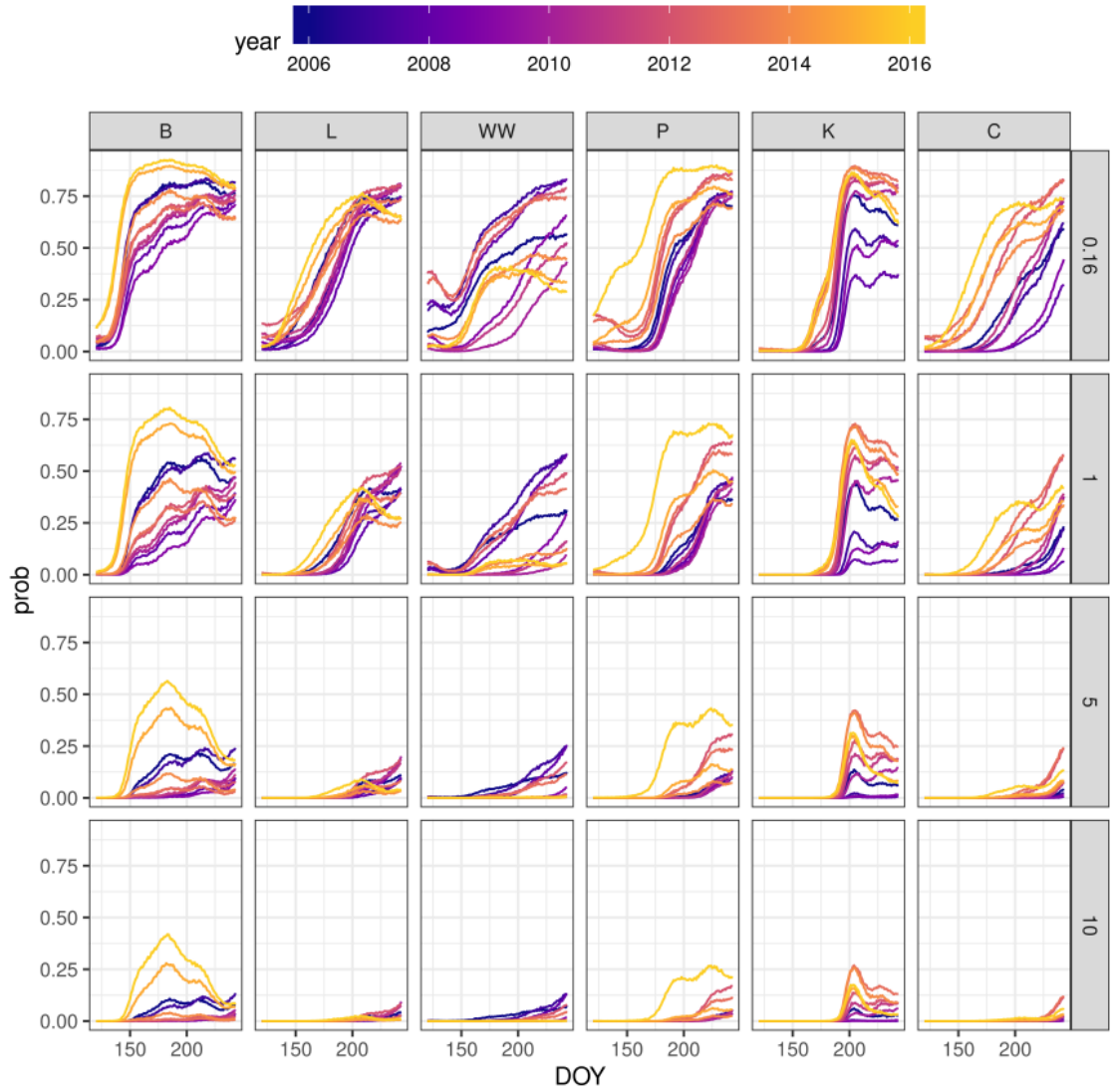
Fitted Microcystin



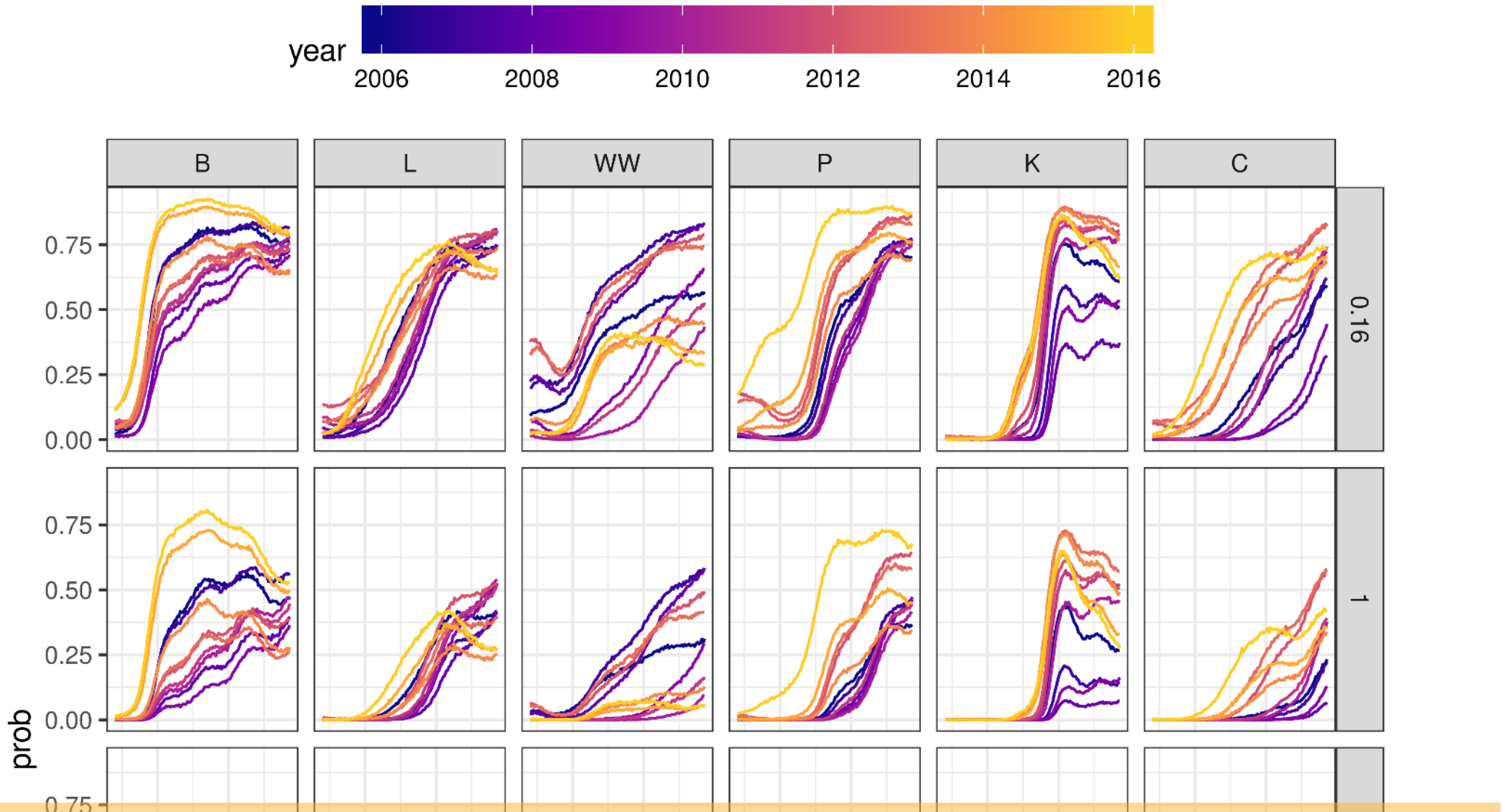
Fitted Microcystin



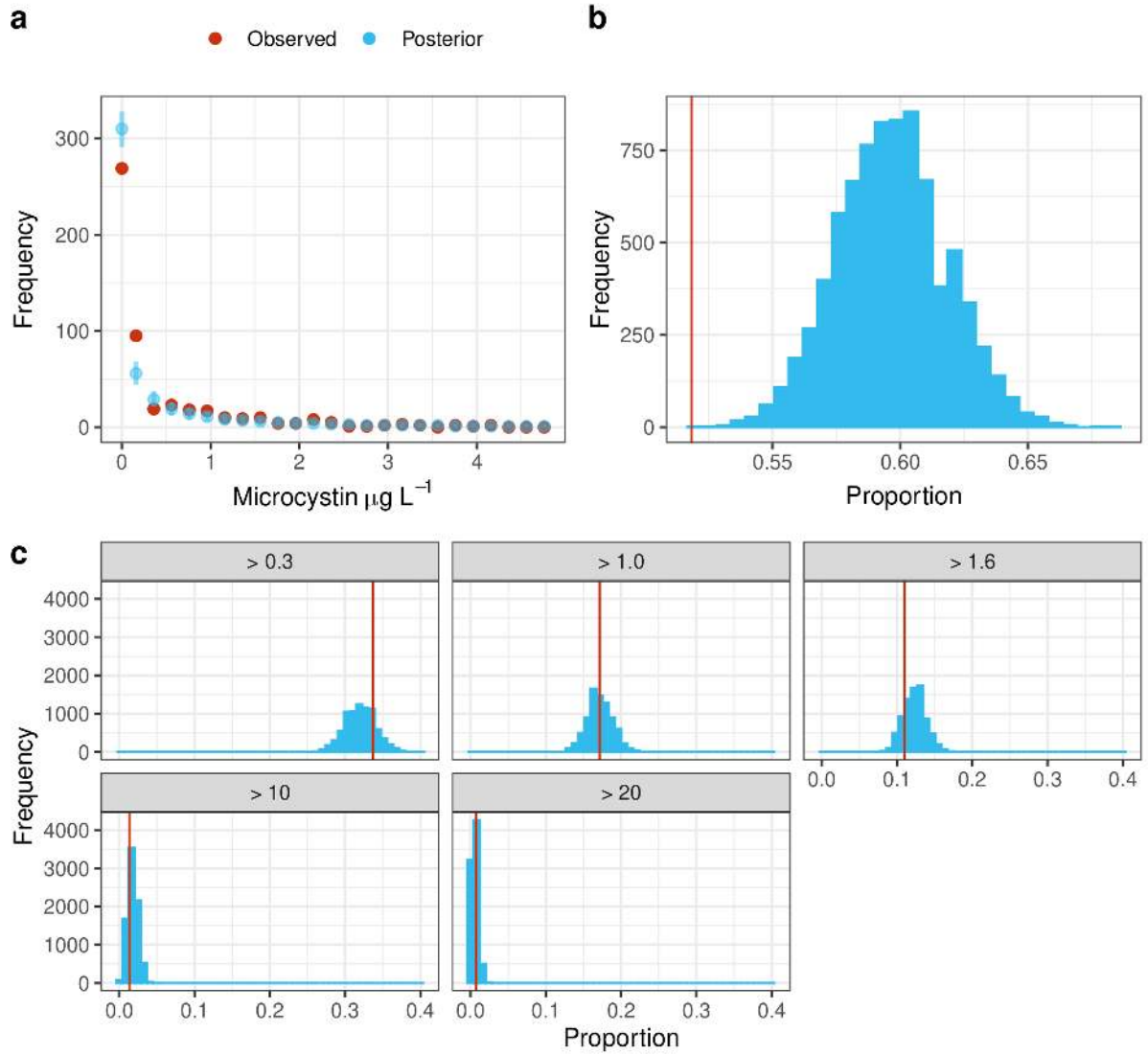
Probability of exceeding thresholds



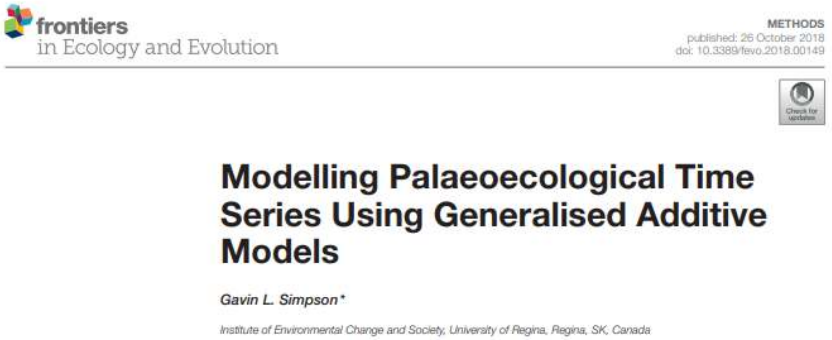
Probability of exceeding thresholds



Posterior predictive checks

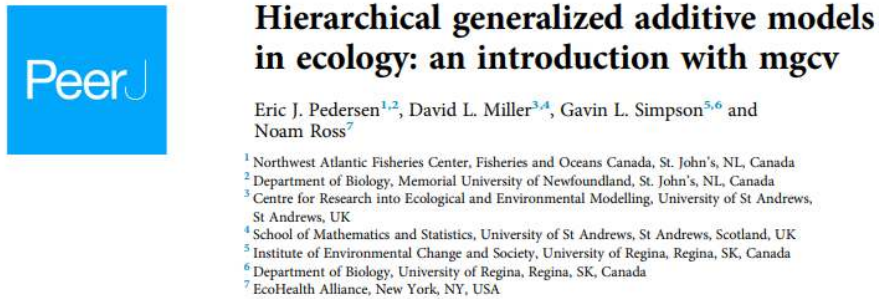


Papers



Simpson (2018) *Frontiers in Ecology & Evolution*

doi: 10/gfrc4p



Pedersen *et al* (2019) *PeerJ*

doi: 10/c6wz

Acknowledgements

Funding

NSERC CRSNG

FACULTY OF GRADUATE STUDIES & RESEARCH

Data

- Microcystin data from QULTER Peter Leavitt (U Regina)
- Iestyn Woolway and colleagues for archiving the lake surface water data

Slides

- HTML Slide deck bit.ly/nyr-gam
- RMarkdown Source