Extreme value models for environmental data

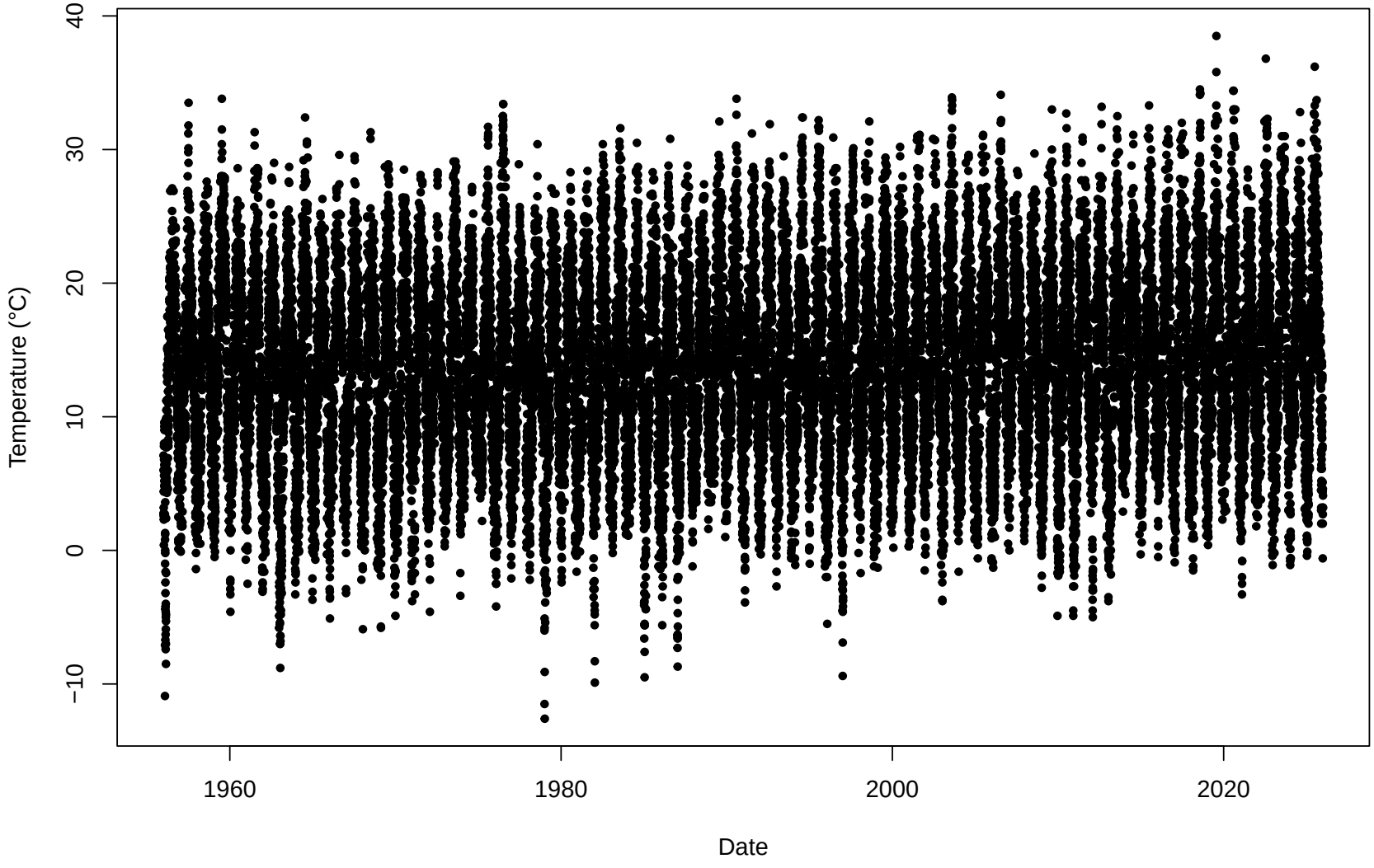
Ben Youngman
University of Exeter
b.youngman@exeter.ac.uk

Learning Spatio-Temporal Climate Extremes
UCLouvain, May 27 2026

Overview

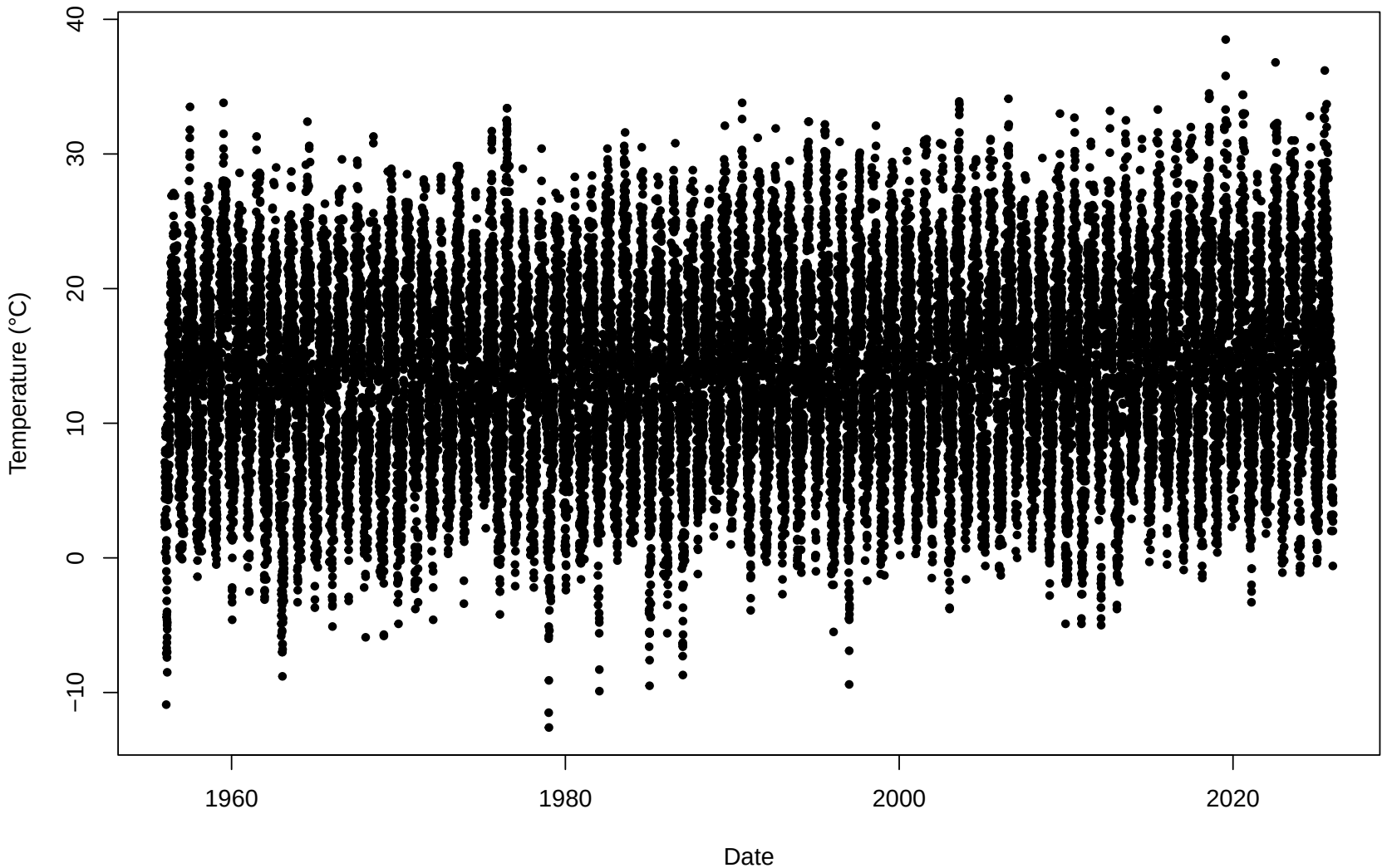
- Extreme value models for environmental data
- Options to overcome a lack of data
- Generalised additive modelling (GAMs)

Daily maximum temperatures in Brussels



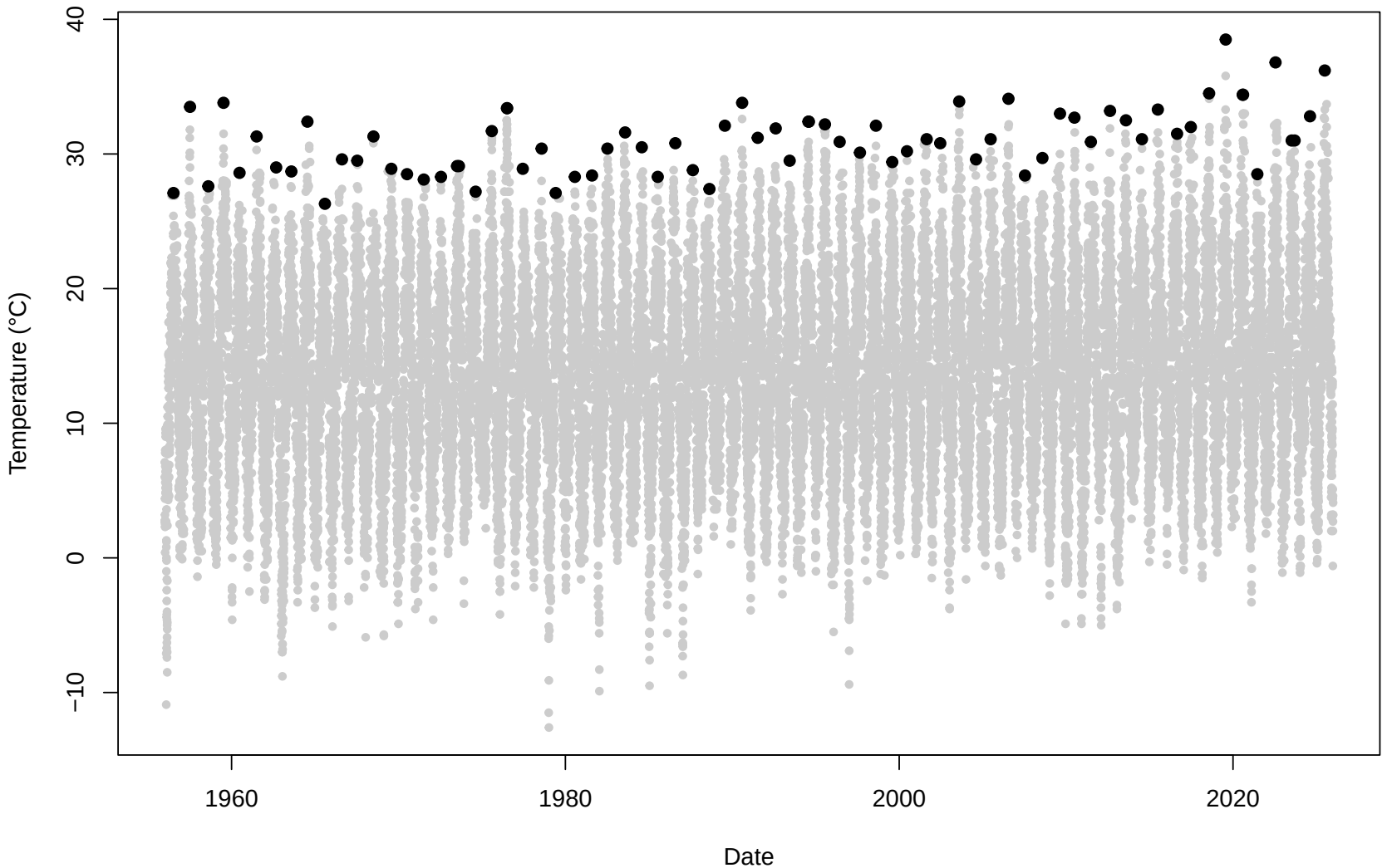
Statistical modelling of extremes

- One approach to modelling extremes is isolating block maxima (e.g. annual maxima) and fitting a generalised extreme value (GEV) distribution to them



Statistical modelling of extremes

- One approach to modelling extremes is isolating block maxima (e.g. annual maxima) and fitting a generalised extreme value (GEV) distribution to them



Statistical modelling of extremes

- Mathematically, we have annual maxima $Y_1, Y_2, \ldots, Y_T$ and assume that

$$Y_t \sim GEV(\mu, \psi, \xi)$$

  for $t = 1, \ldots, T$, where μ, ψ and ξ are the GEV distribution's location, scale and shape parameters, respectively

- However, with this approach we have
  - assumed the distribution doesn't change with time
  - discarded 364 out of every 365 data points

Statistical modelling of extremes

- We can achieve a changing GEV distribution by letting its parameters change with time, i.e. assuming that

$$Y_t \sim GEV\left(\mu(t), \psi(t), \xi(t)\right)$$

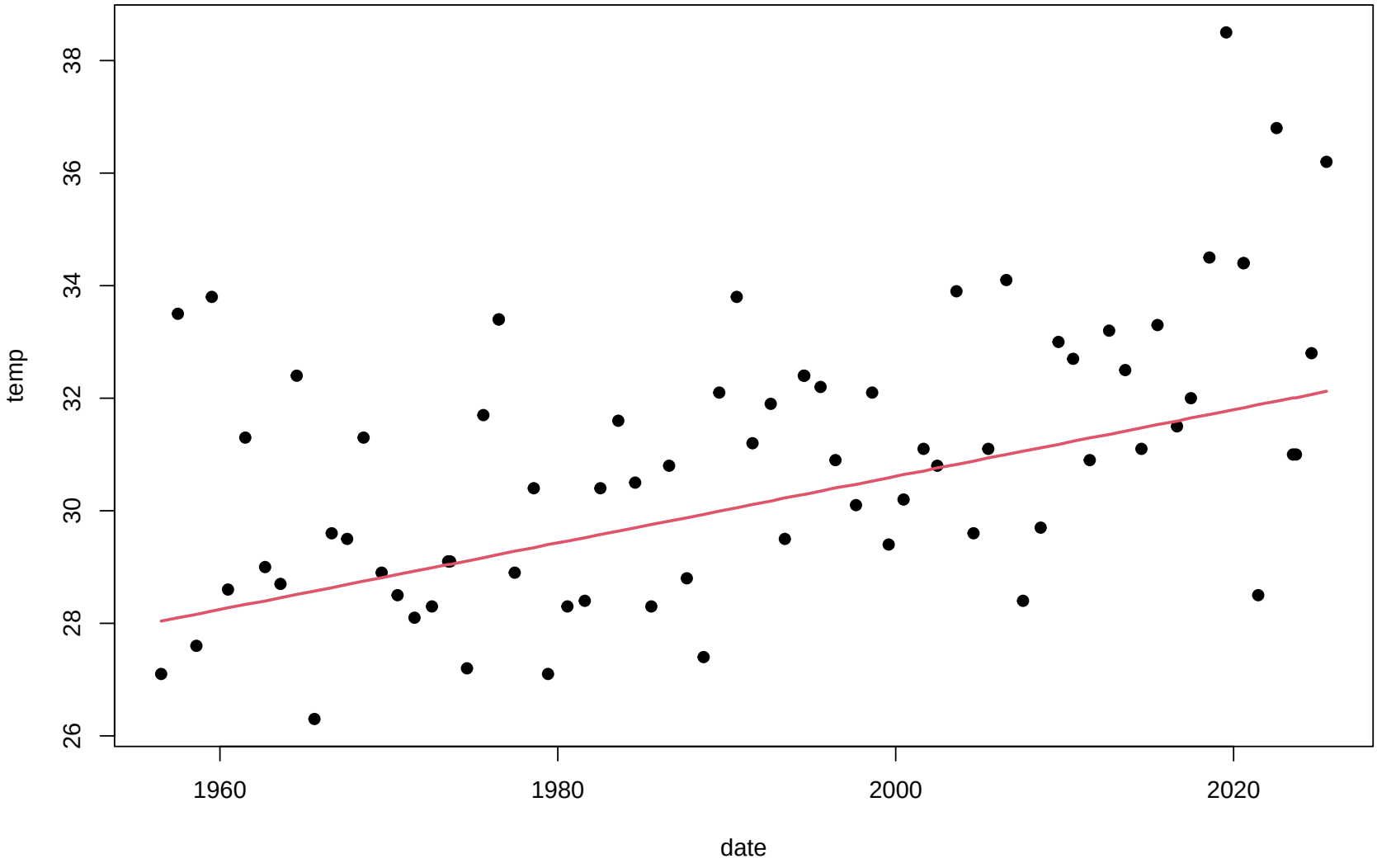
- For example, a simple linear trend in the location parameter is

$$\mu(t) = \mu_0 + \mu_1 t$$

where t might be measured in years

Statistical modelling of extremes

- A linearly-varying location parameter has the following estimate



Splines, Smooths and Generalised Additive Models

- Linear or polynomial assumptions could be considered quite restrictive
  - the structure that they impose might contradict physical knowledge
- A more flexible approach is to assume some smooth function, and to let the data decide how smooth it should be
- With a generalised additive model (GAM) we can assume that

$$\mu(t) = \mu_0 + g_1(t) + g_2(t) + \dots$$

where $g_1, g_2, \dots$ are different smooth functions

Fitting GAMs for extremes in R

- The evgam package in R lets us fit extreme value distributions with parameters that vary according to GAMs
- Its use is

```
> fitted_model <- evgam(formula, data, family)
```

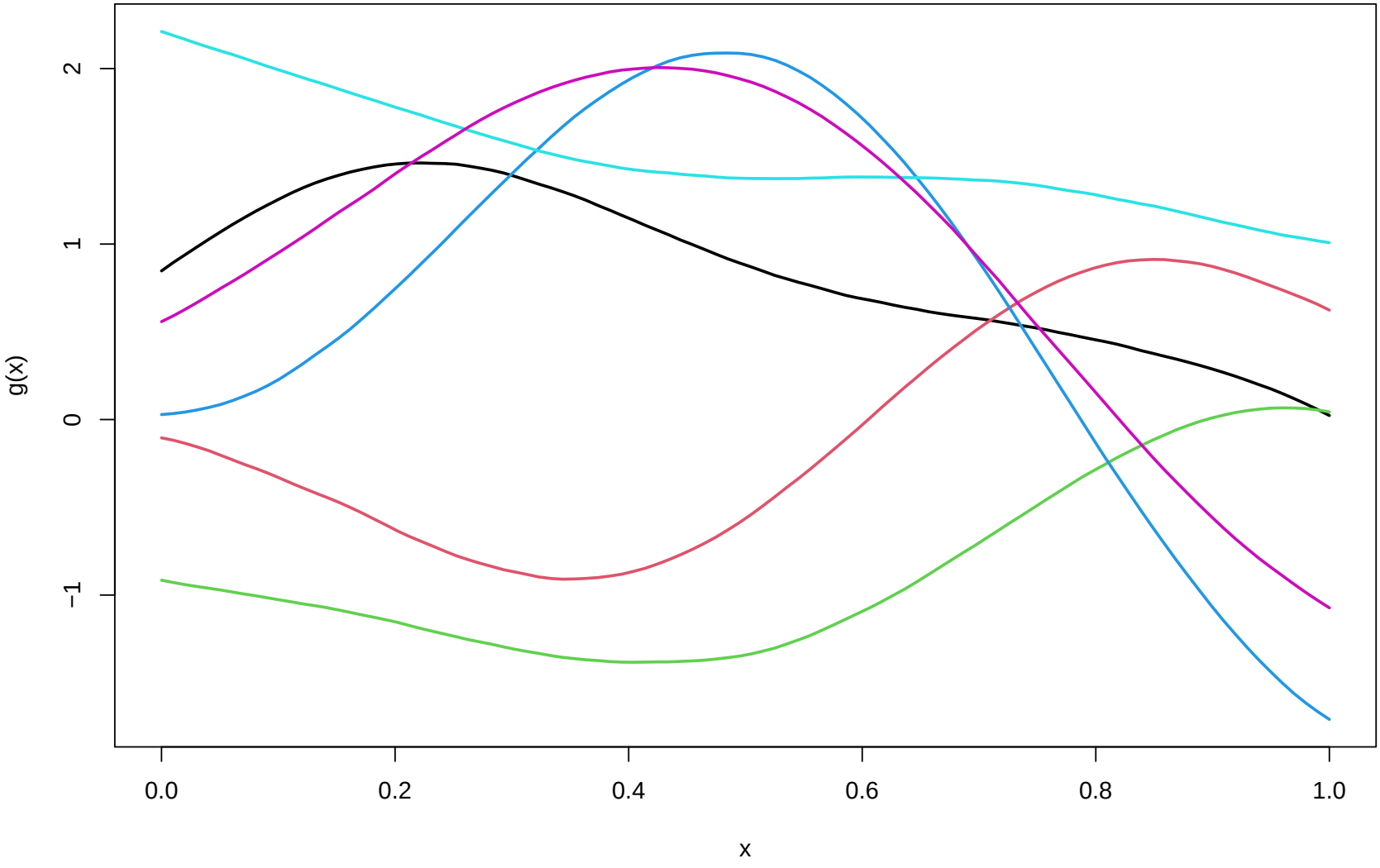
where formula defines the splines, data is a data.frame and family defines the extreme value distribution (e.g. 'gev', 'gpd', ...)

- For the GEV we can specify splines for its location, scale and shape parameters with

```
formula <- list(response ~ s(...) + s(...) + ...,
                         ~ s(...) + s(...) + ...,
                         ~ s(...) + s(...) + ...)
```

Splines, Smooths and Generalised Additive Models

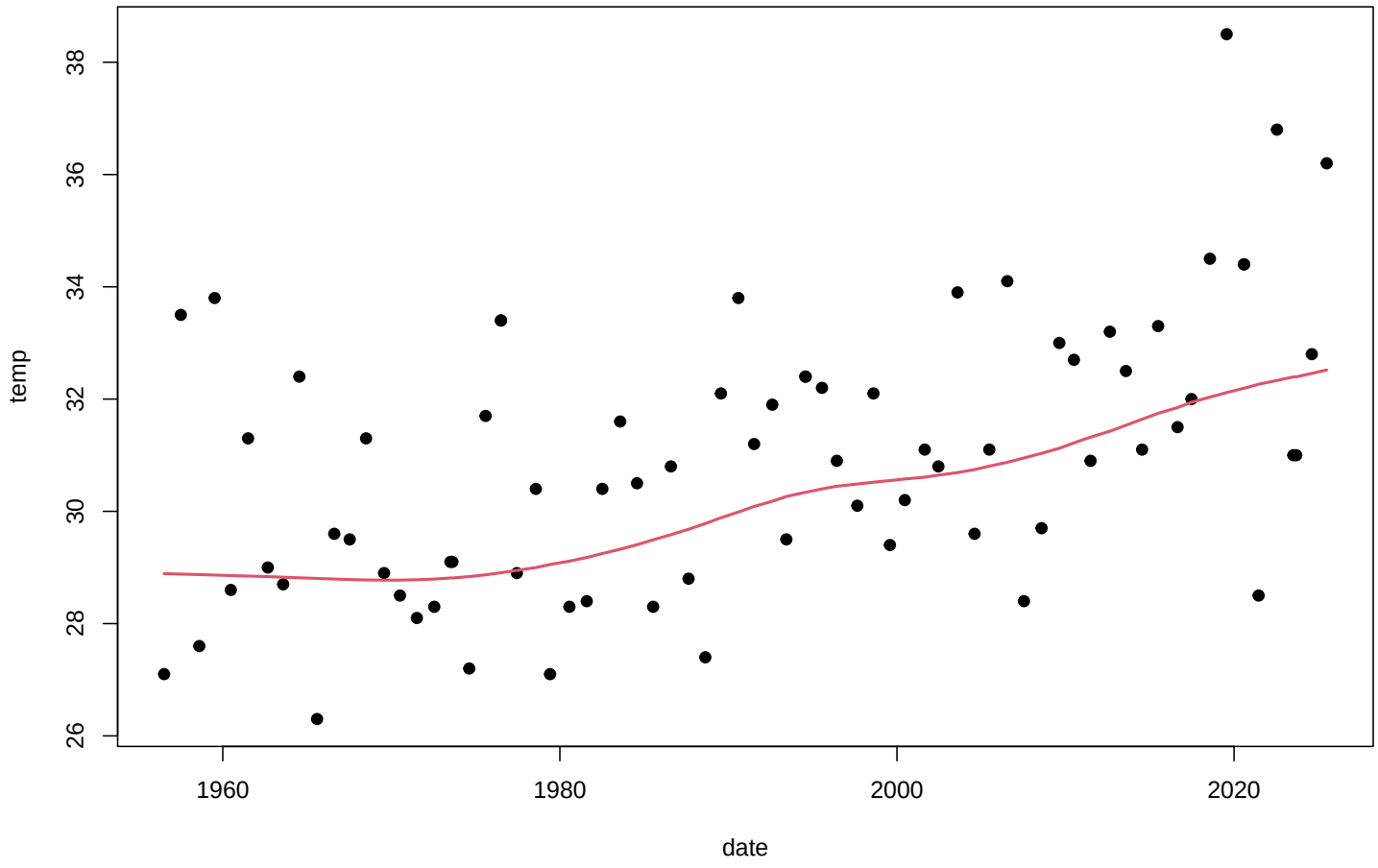
- A flexible one-dimensional form for g is the cubic regression spline



Statistical modelling of extremes

- We can assume a cubic regression splines for the GEV's location parameter

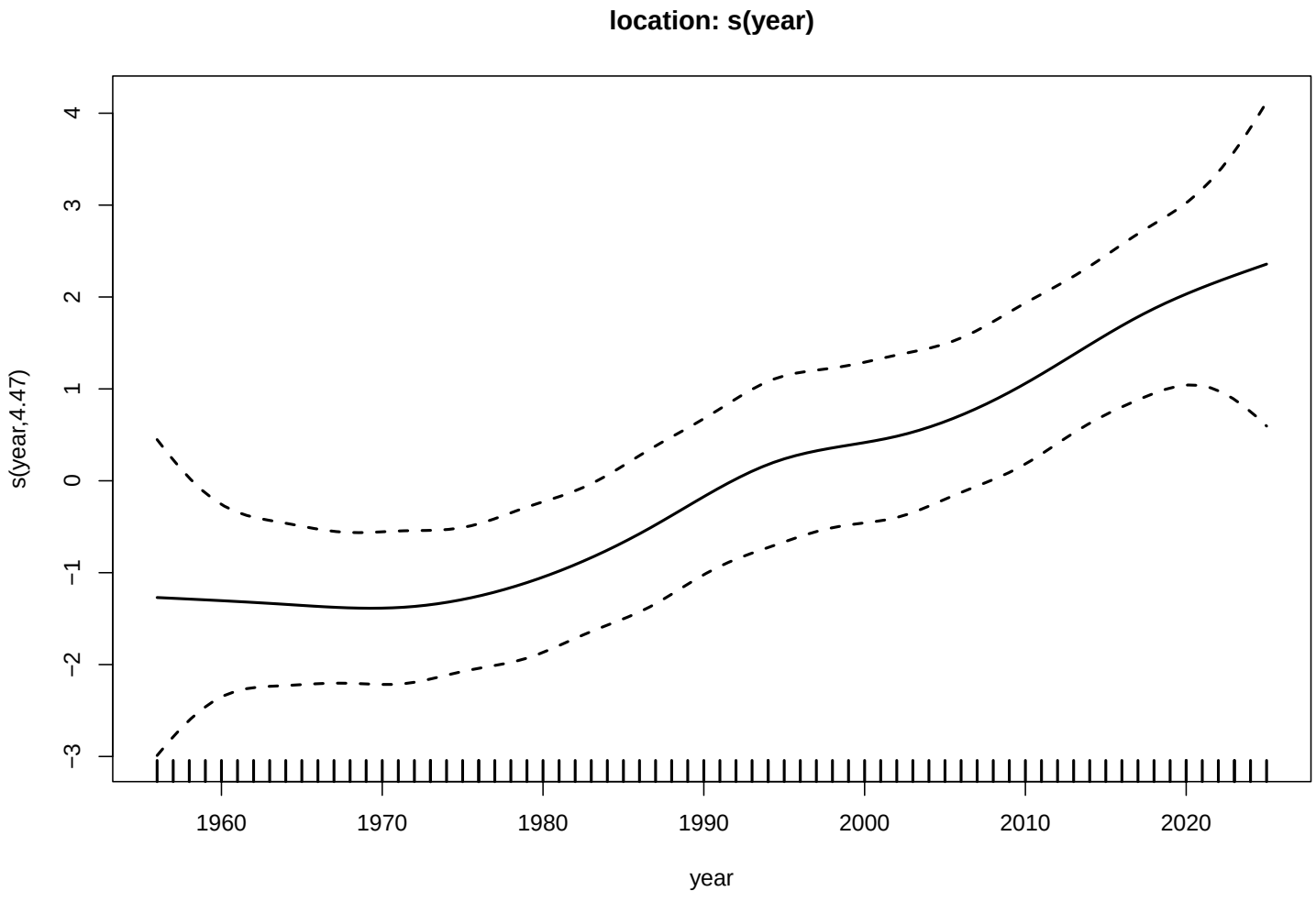
```
> formula <- list(tmax ~ s(year, bs = 'cr'), ~ 1, ~ 1)
```



Statistical modelling of extremes

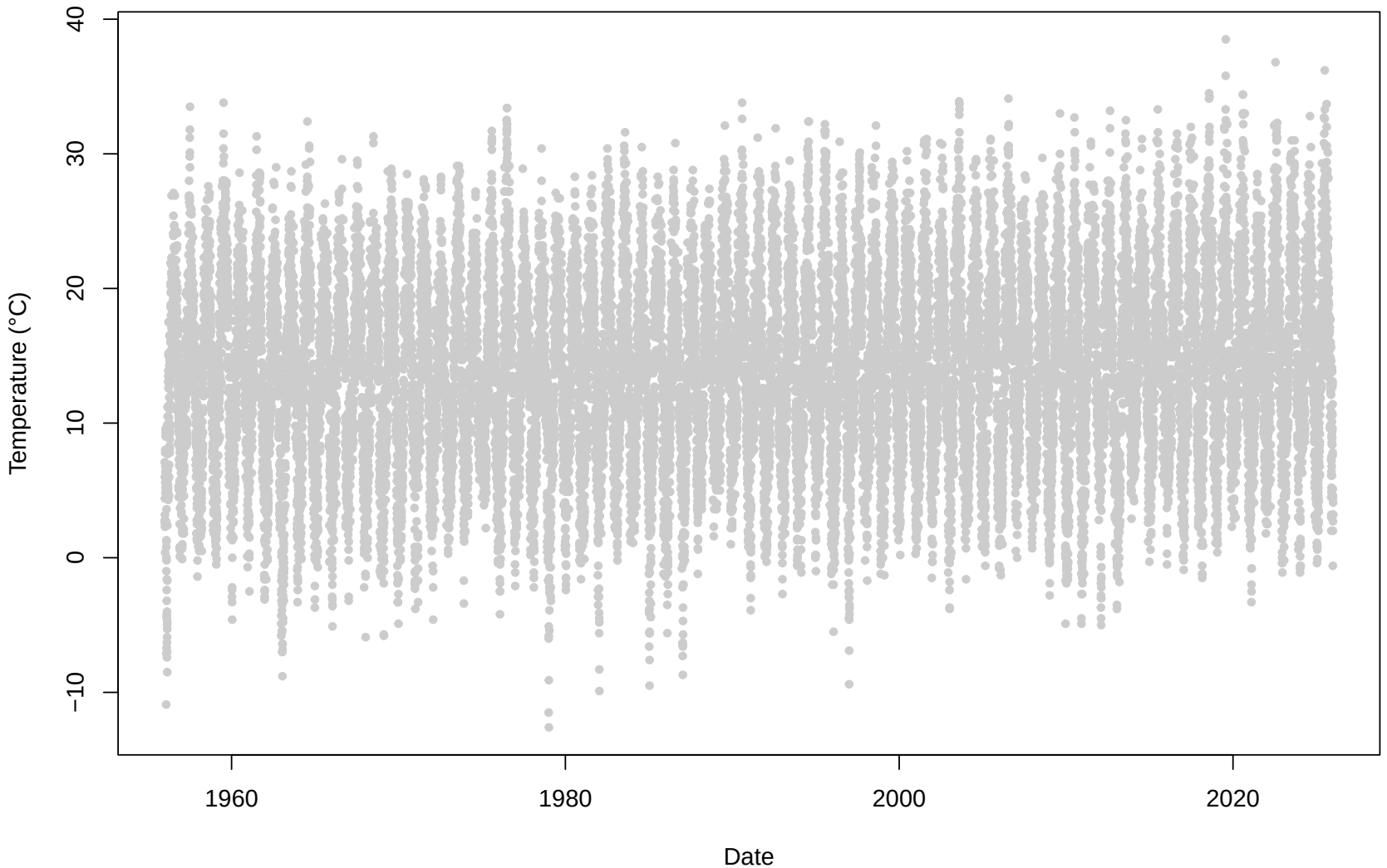
- We can isolate and view the smooth term that's been estimated

```
> plot(fitted_model)
```



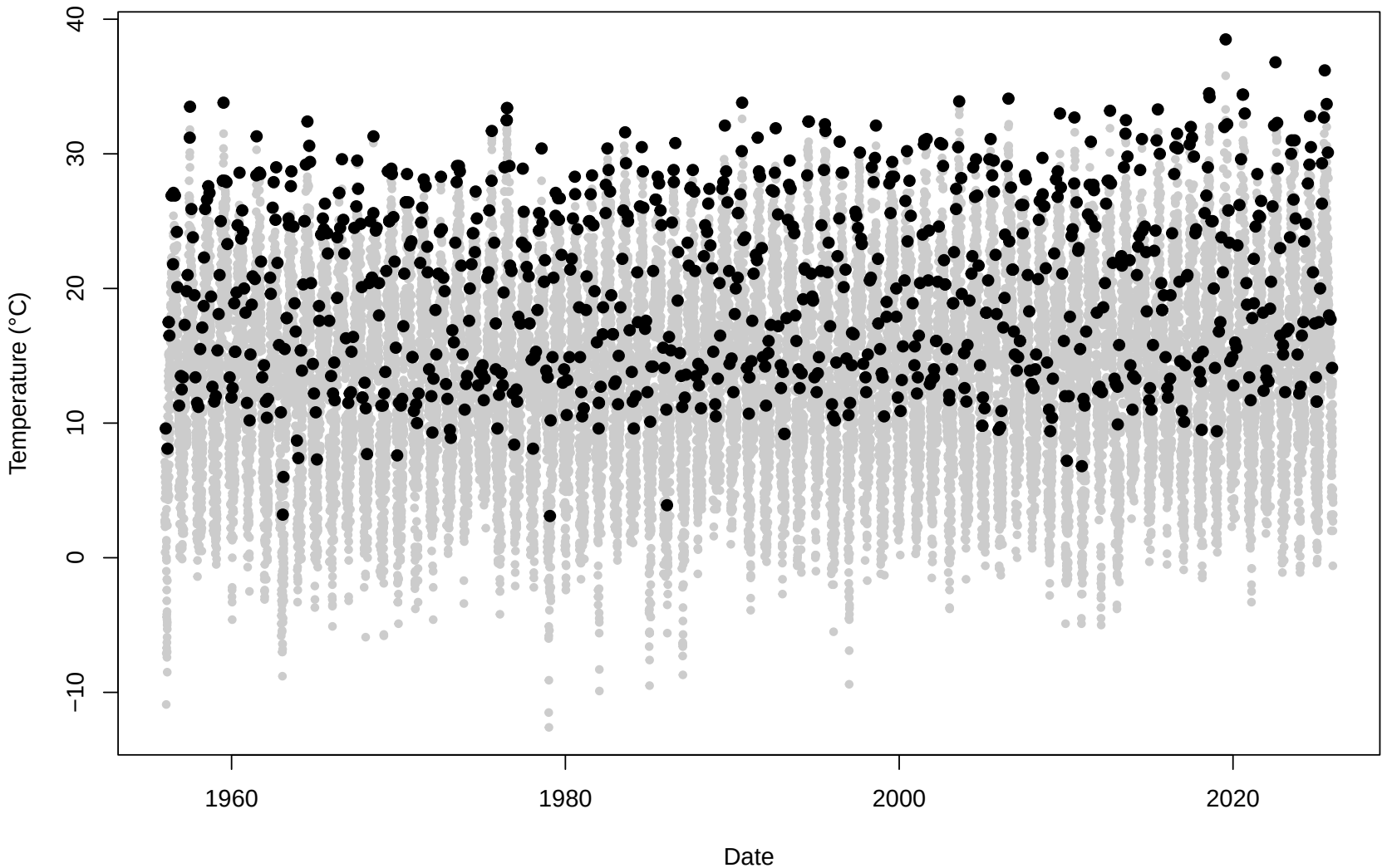
A different choice of maxima

- Retaining monthly maxima gives us much more data
  - and is necessary if we need to understand extremes on a monthly basis



A different choice of maxima

- Retaining monthly maxima gives us much more data
  - and is necessary if we need to understand extremes on a monthly basis



A different choice of maxima

- Retaining monthly maxima gives us much more data
  - and is necessary if we need understand extremes on a monthly basis
- We can estimate a GEV for each month, $F_{Jan}, F_{Feb}, \ldots, F_{Dec}$
- The distribution of the *annual maximum* is then

$$F_{annual}(x) = F_{Jan}(x)F_{Feb}(x)\ldots F_{Dec}(x)$$

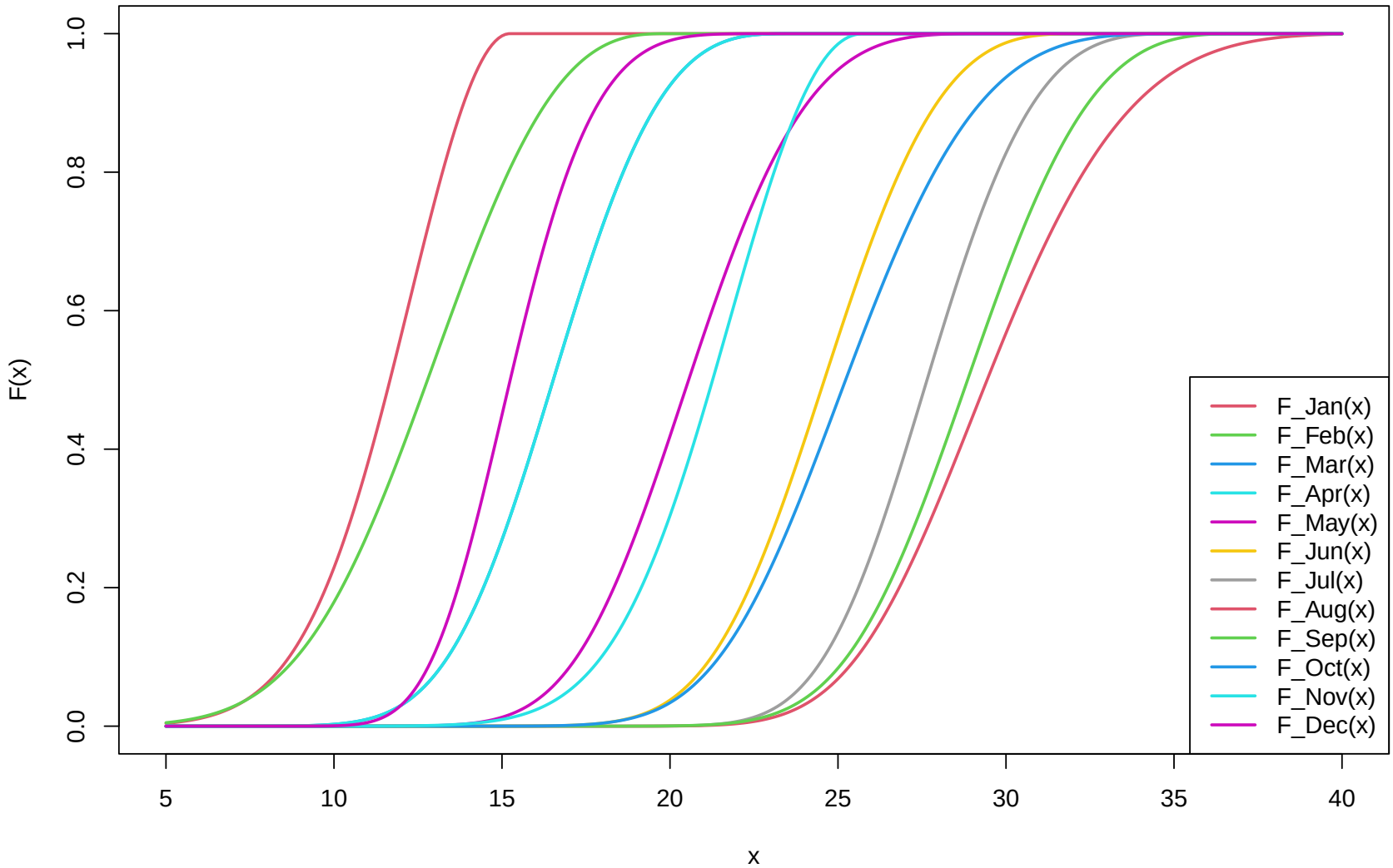
or more precisely

$$F_{annual}(x) = \left\{ \left[F_{Jan}(x)\right]^{31} \left[F_{Feb}(x)\right]^{28.25} \ldots \left[F_{Dec}(x)\right]^{31} \right\}^{12/365.25}$$

A different choice of maxima

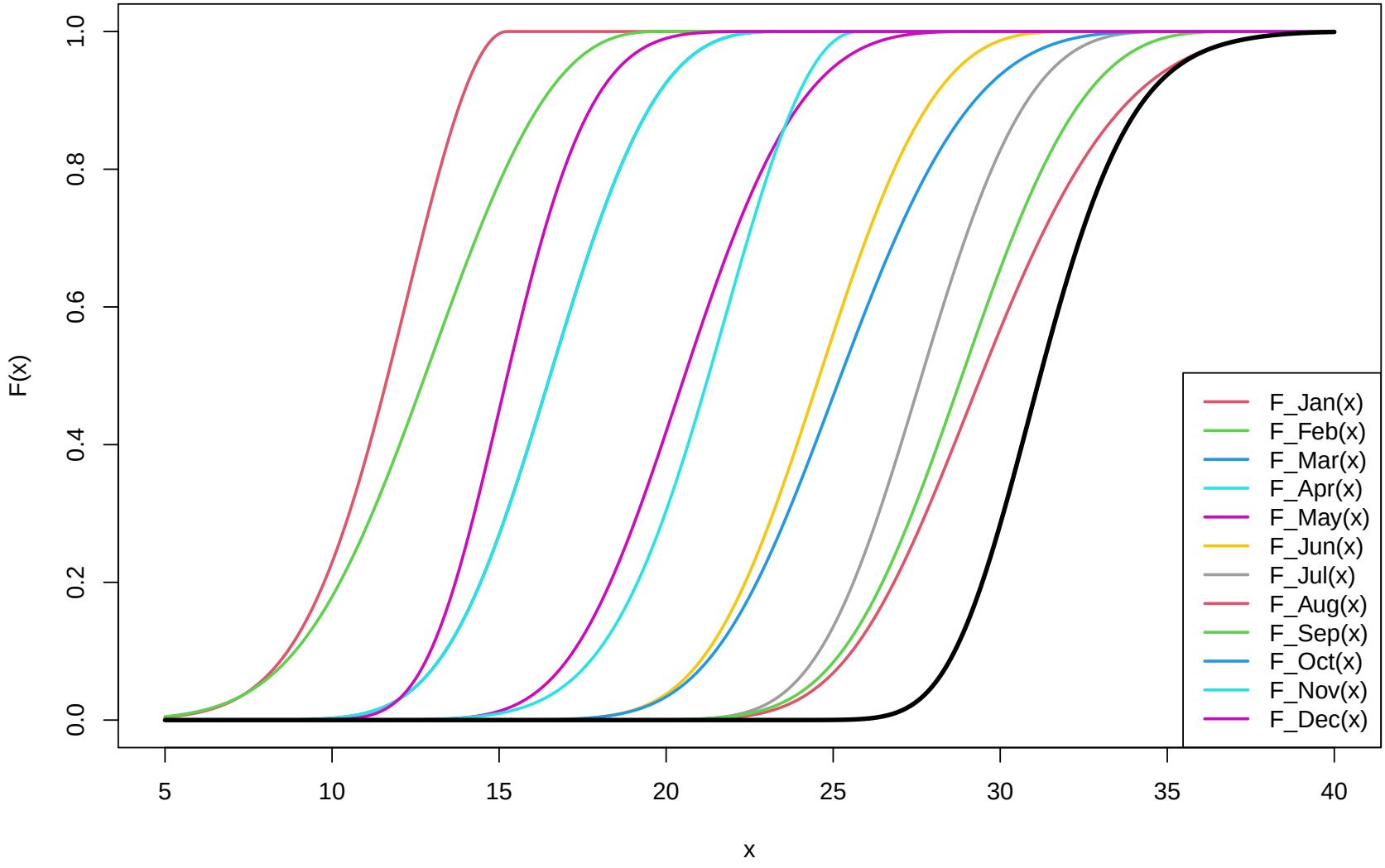
- Each month then has its own GEV estimate, and the annual maximum's distribution can be derived

```
> ~ as.factor(month)
```



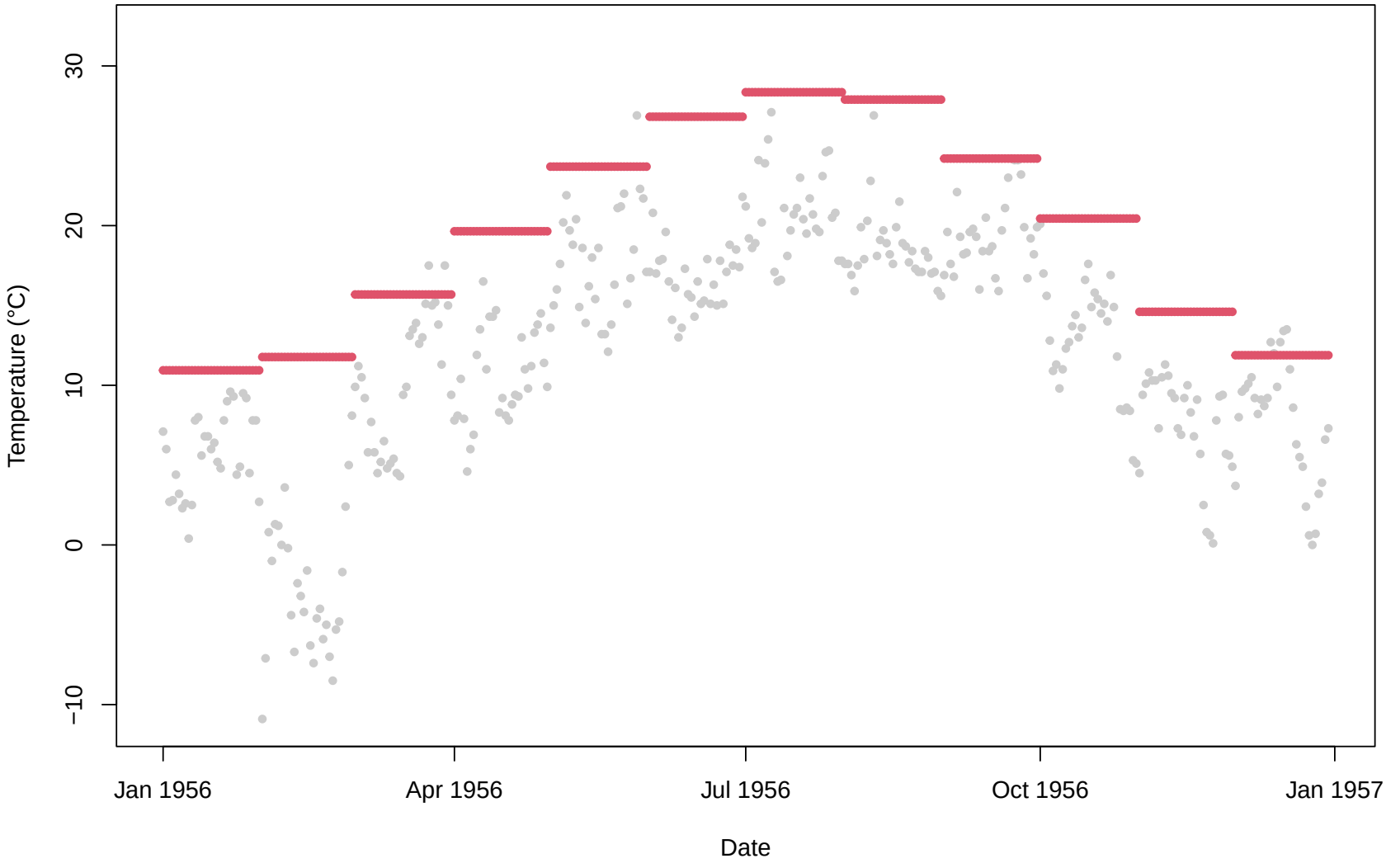
A different choice of maxima

- Each month then has its own GEV estimate, and the annual maximum's distribution can be derived



A different choice of maxima

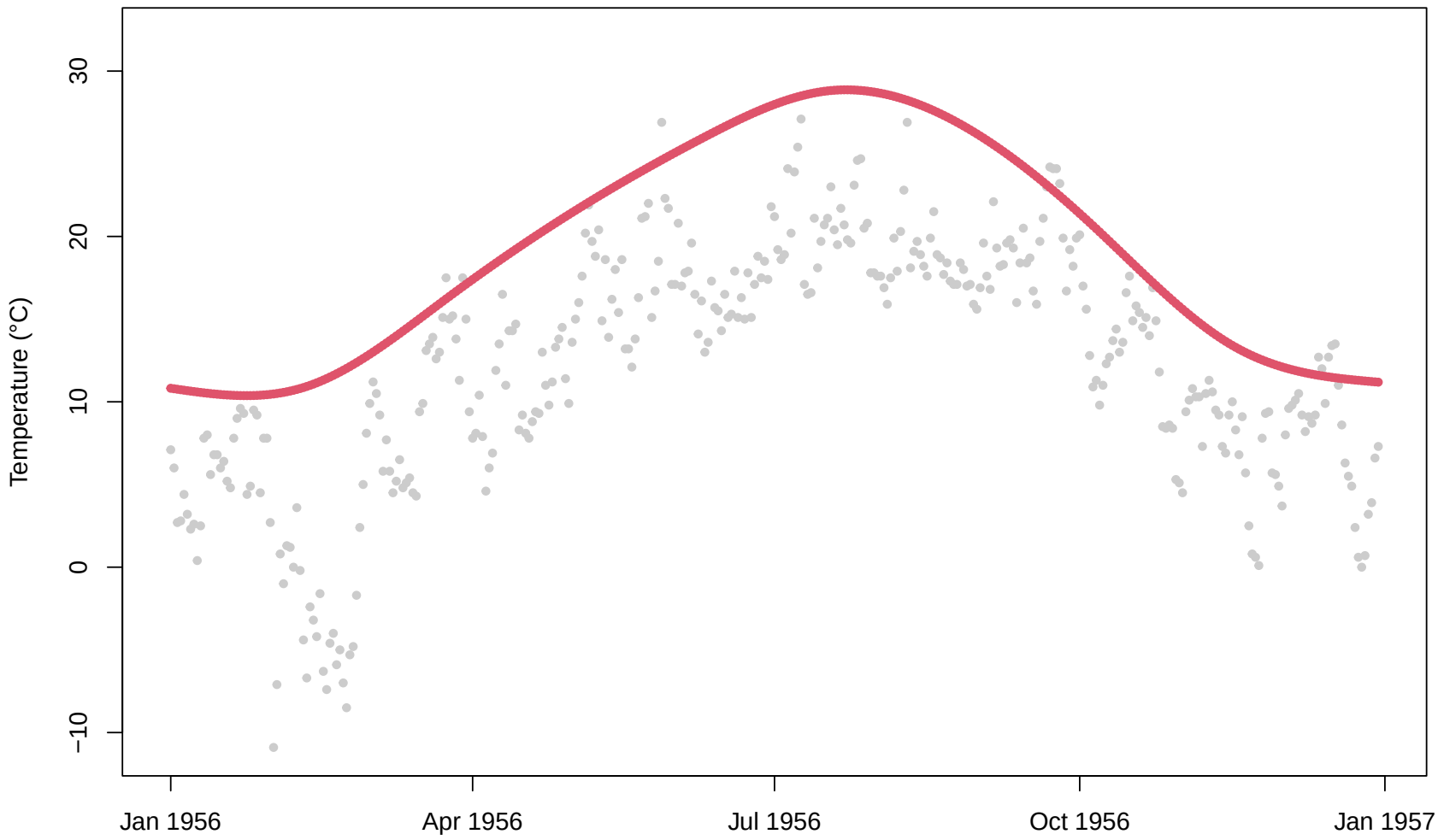
- . . . but this gives disjoint location parameter estimates



A different choice of maxima

- A cyclic cubic regression spline gives a different location parameter for each day of the year using monthly maxima
  - and it's continuous from 31 Dec to 1 Jan

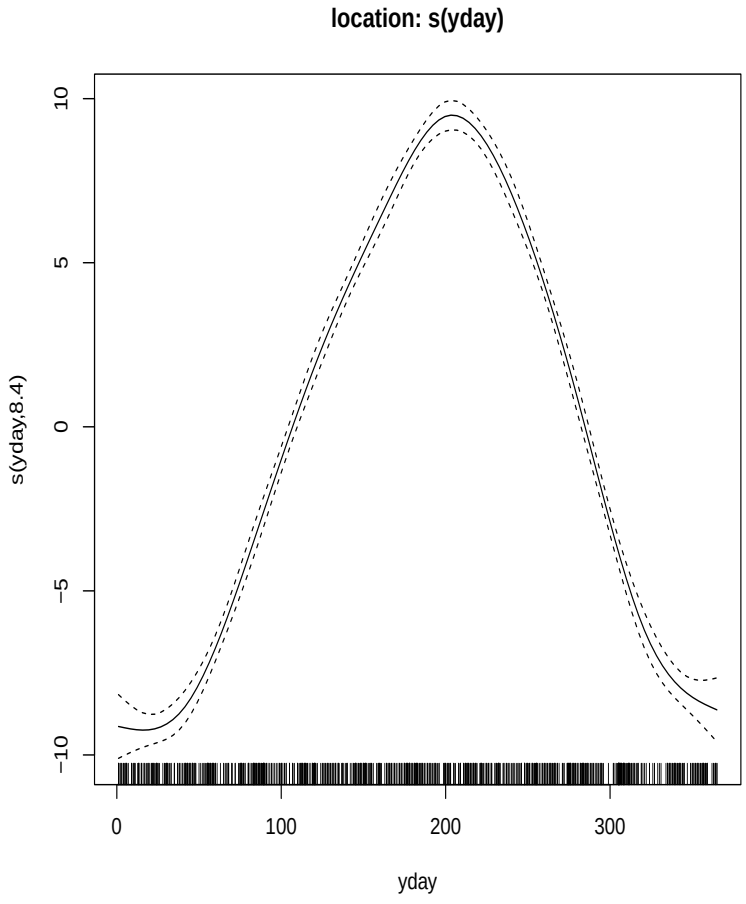
```
> ~ s(day_of_year, bs = 'cc')
```

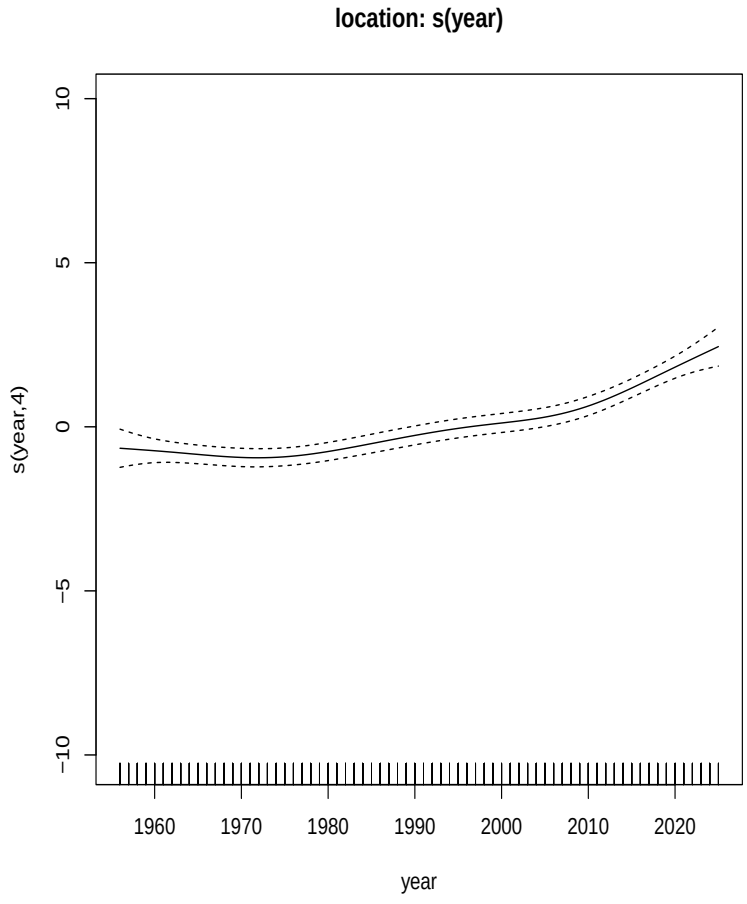


A different choice of maxima

- More than one smooth can be used

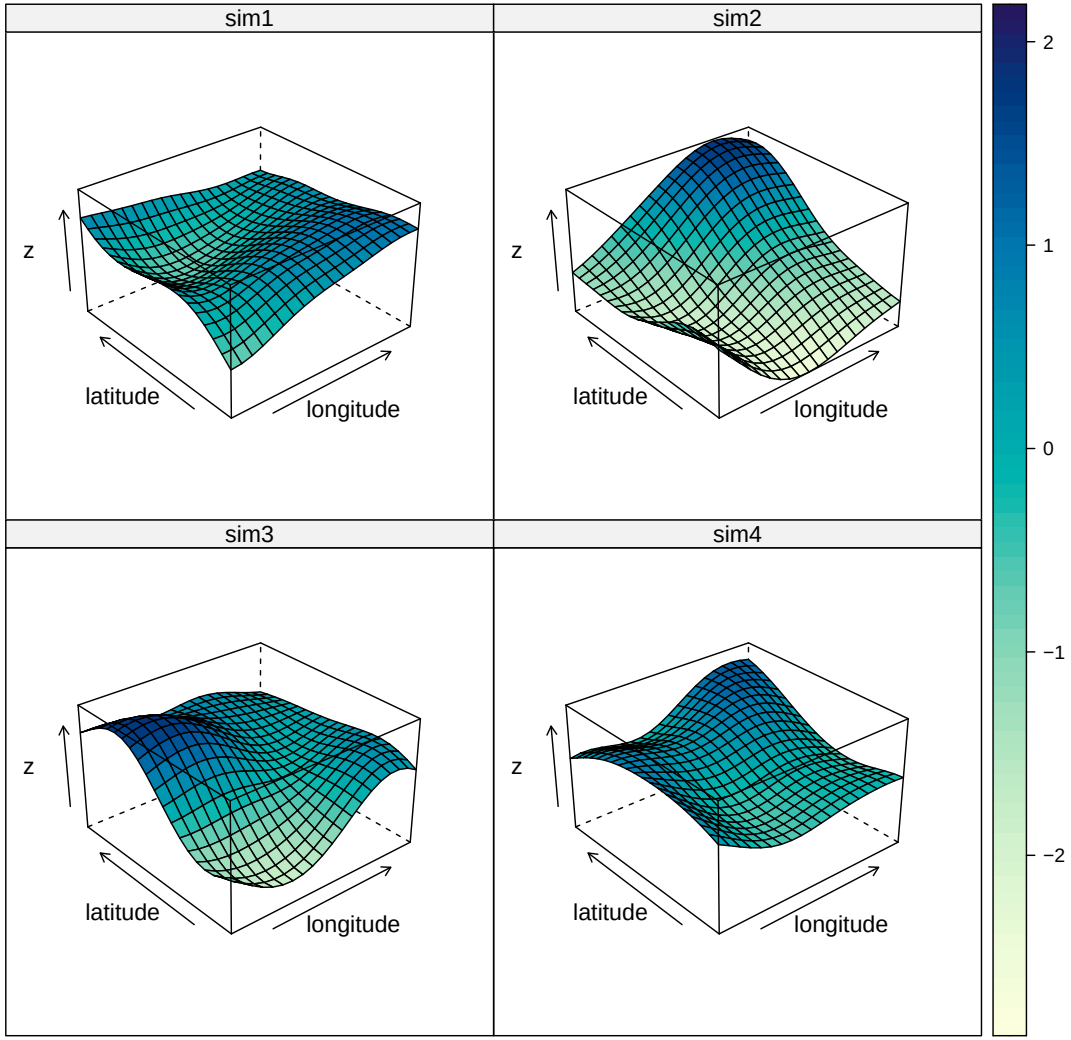
```
> ~ s(year, bs = 'cr') + s(day_of_year, bs = 'cc')
```





Spatial variation

- Splines can also be used to capture smooth variation over space
- Particularly useful is the thin-plate regression spline



Spatial variation

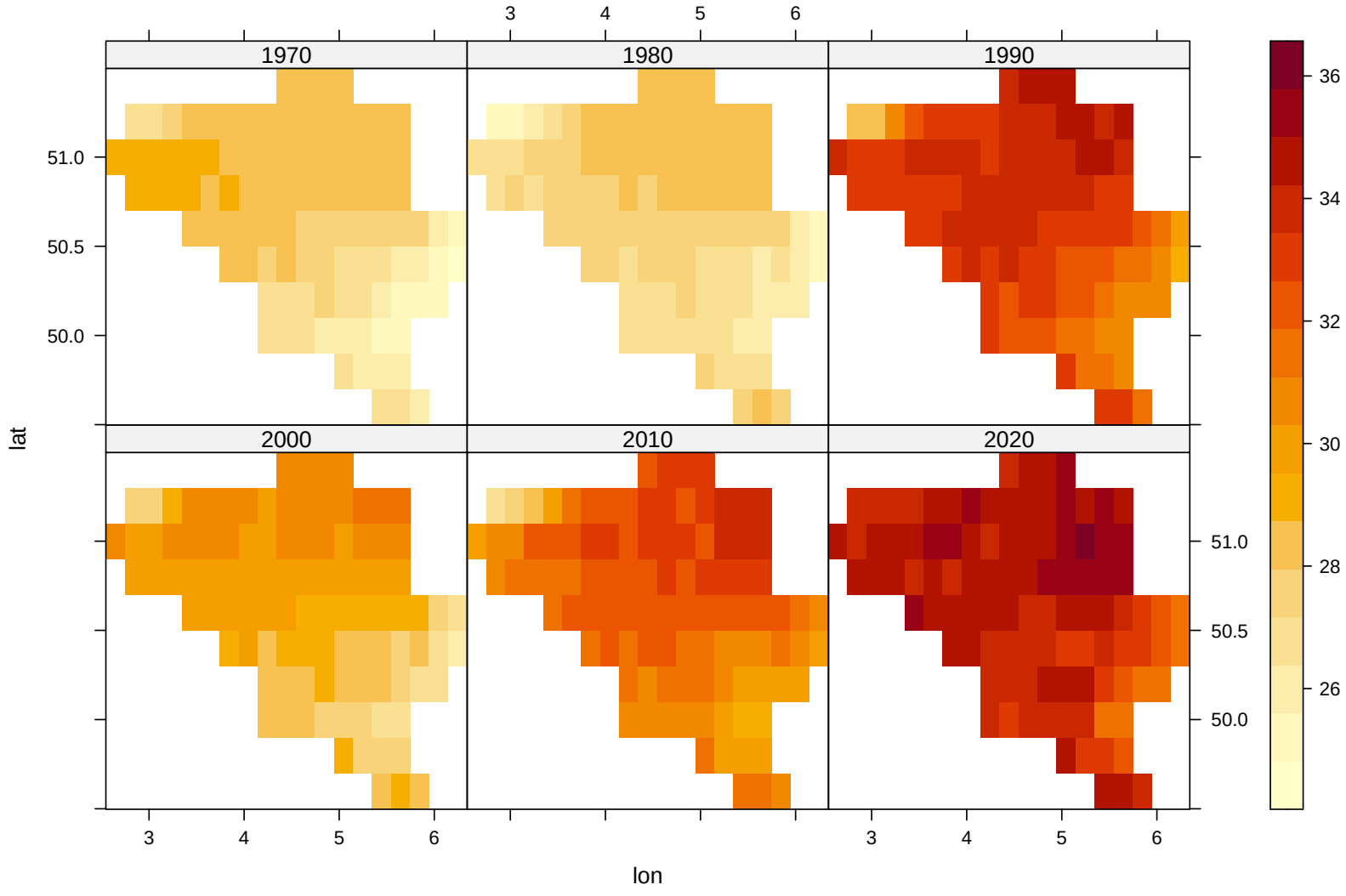
- To achieve spatial variation, we use higher-dimensional splines, which for the GEV's location parameter might correspond to

$$\mu(lon, lat) = \mu_0 + g_1(lon, lat)$$

  where g_1 is a thin-plate regression spline
- These can be achieved with `s(lon, lat, bs = 'tp')`

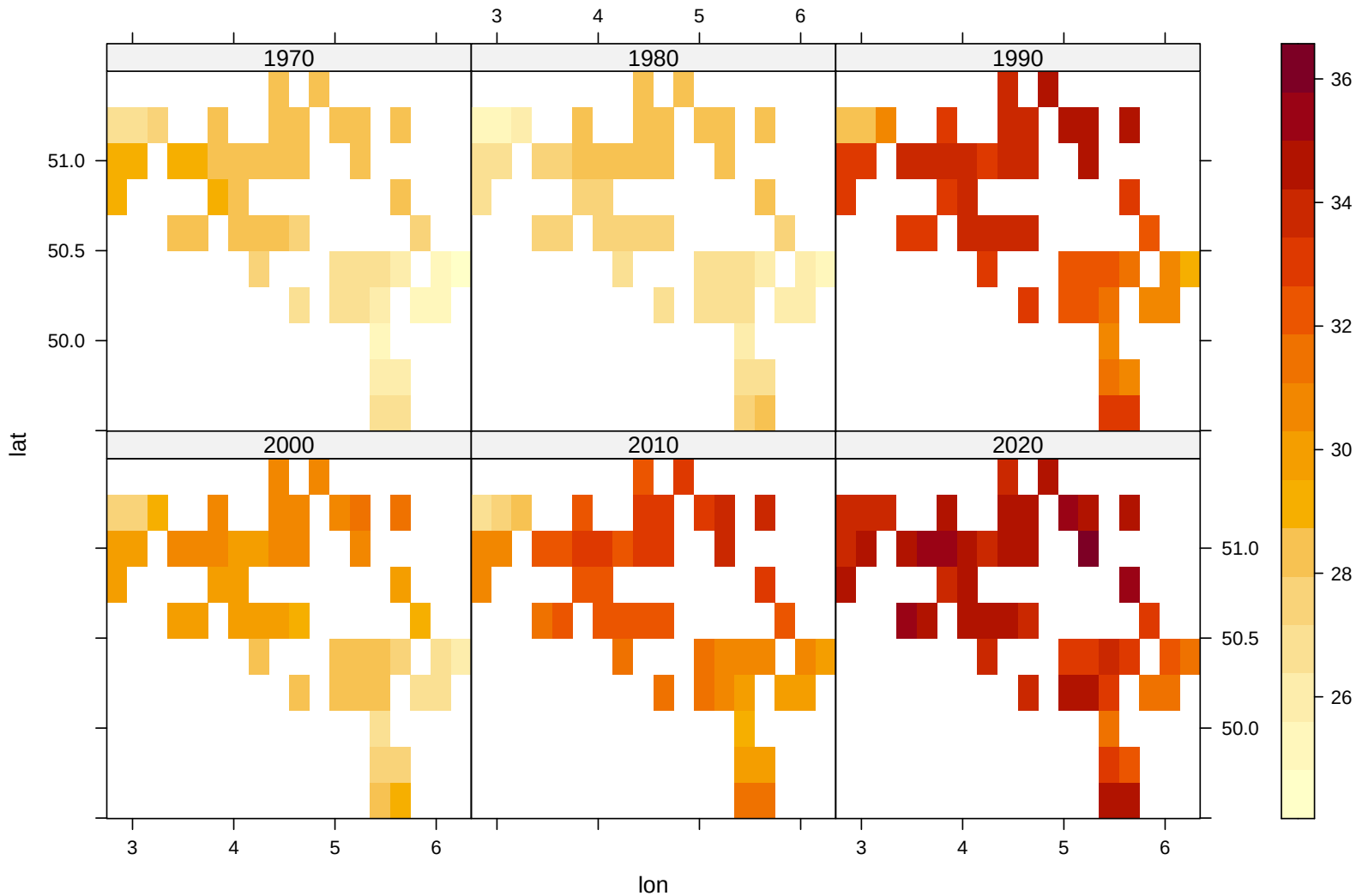
Belgian annual maximum temperatures

- We'll model annual maximum daily temperatures over Belgium



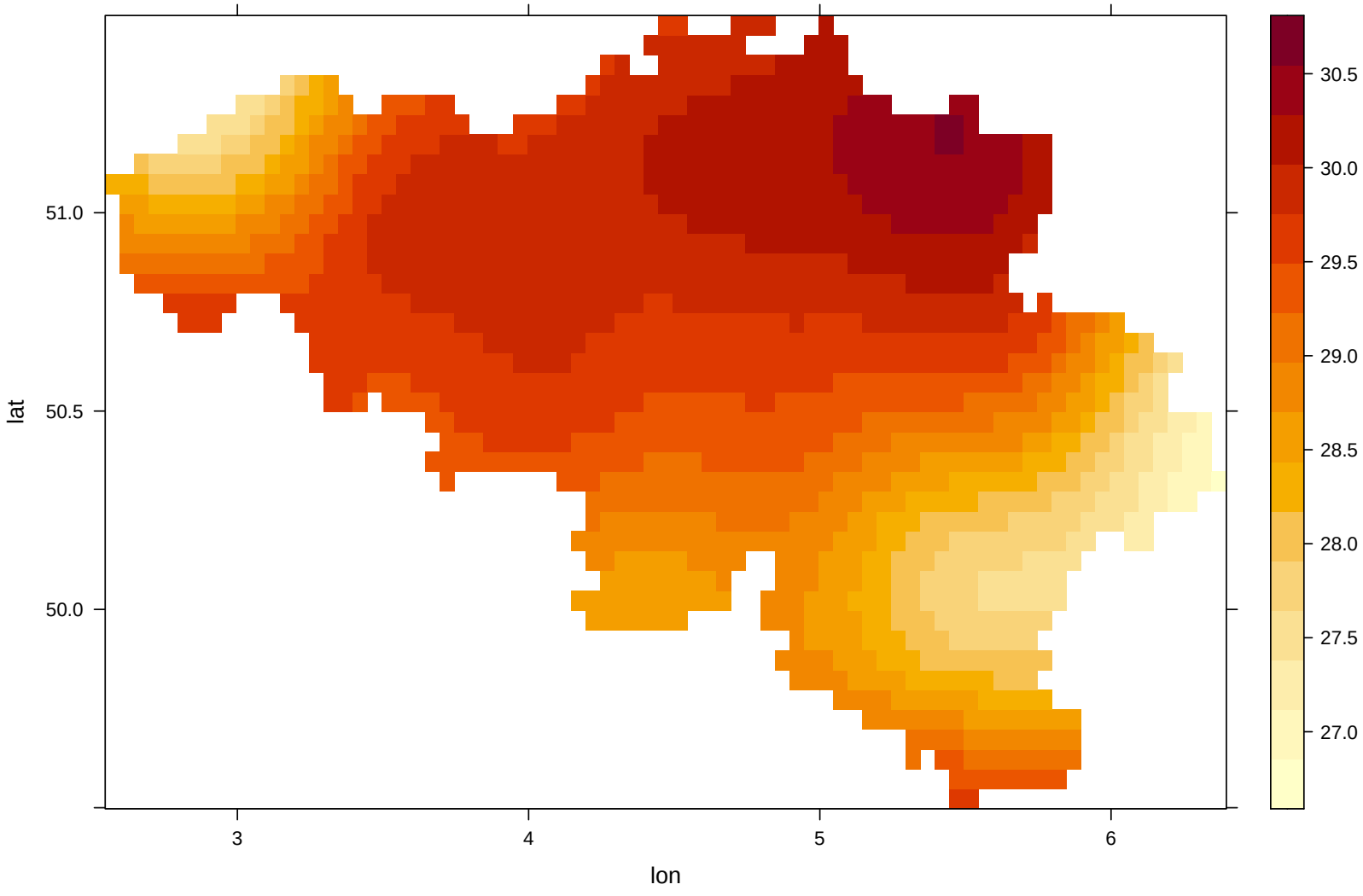
Belgian annual maximum temperatures

- Although the data are gridded (ERA5-Land), we can obtain a complete spatial estimate from point-referenced or irregularly-spread data



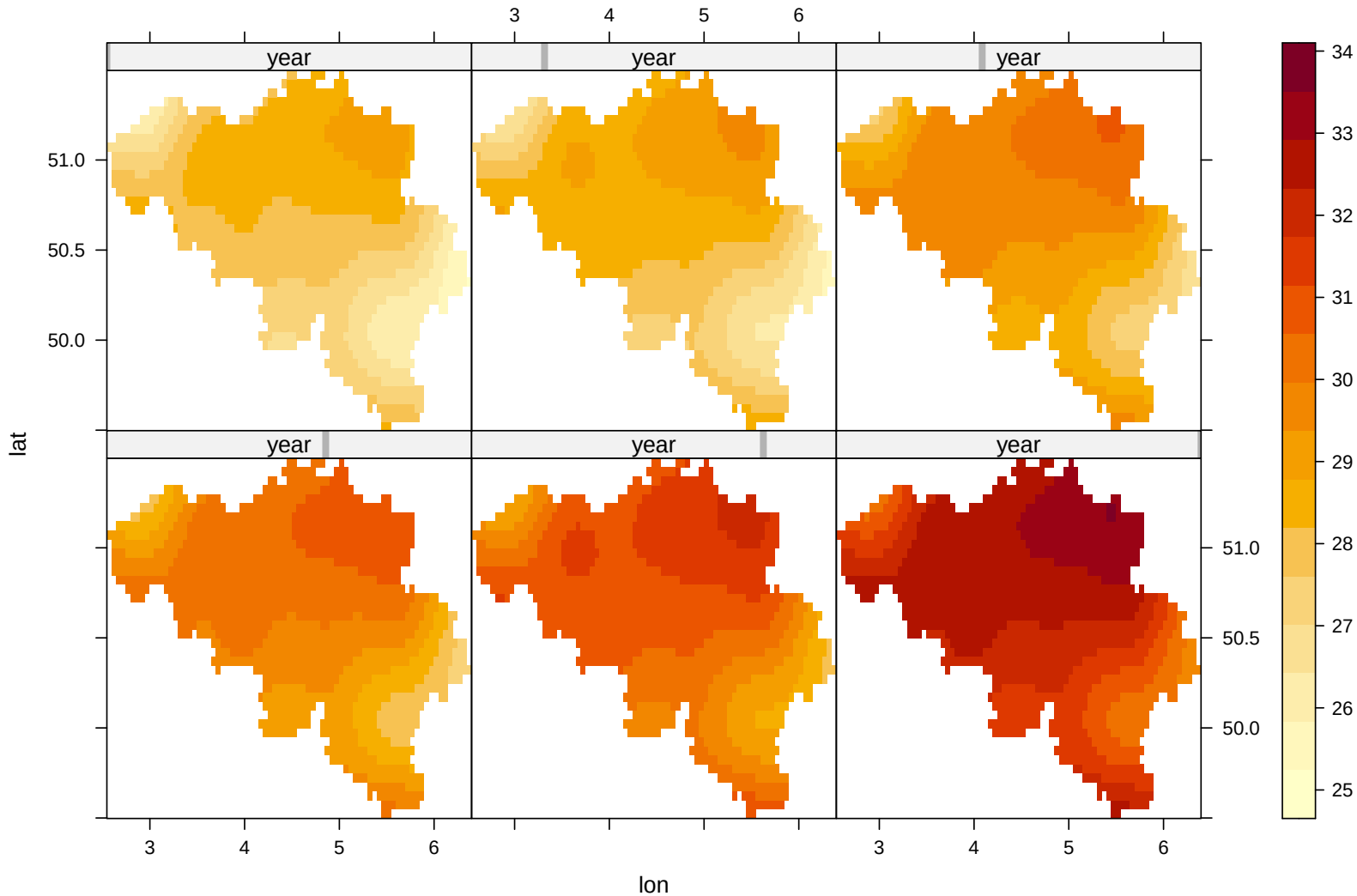
Belgian annual maximum temperatures

- Our extreme value model will assume that the GEV's location parameter takes the form of a thin-plate regression spline



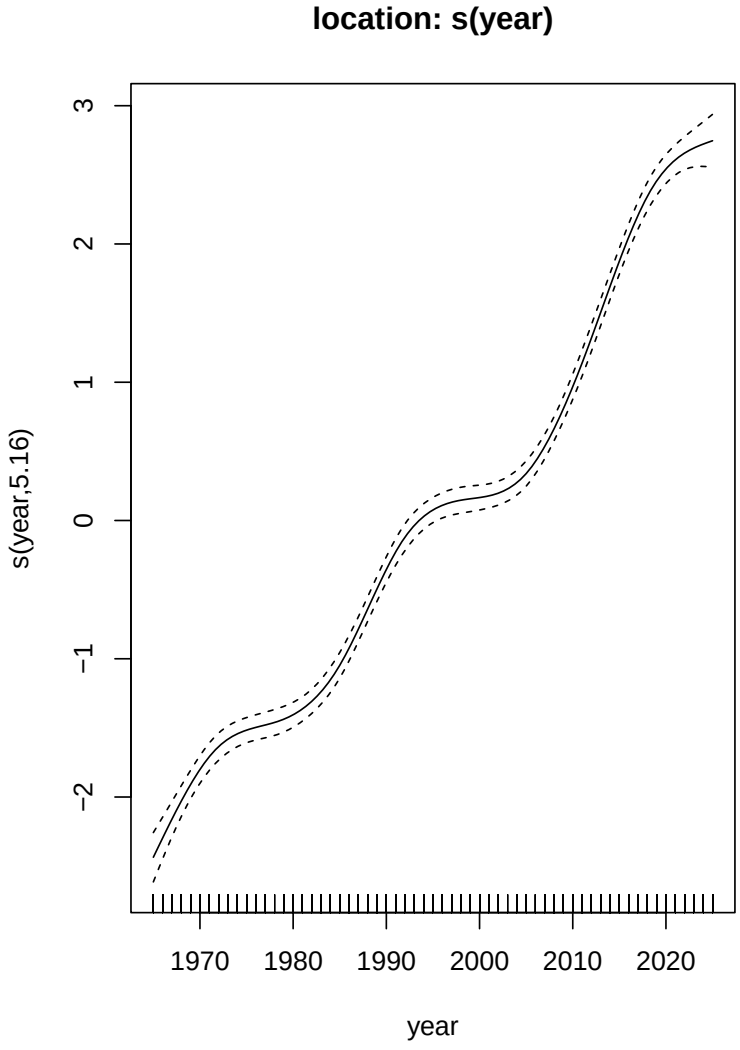
Spatio-temporal variation

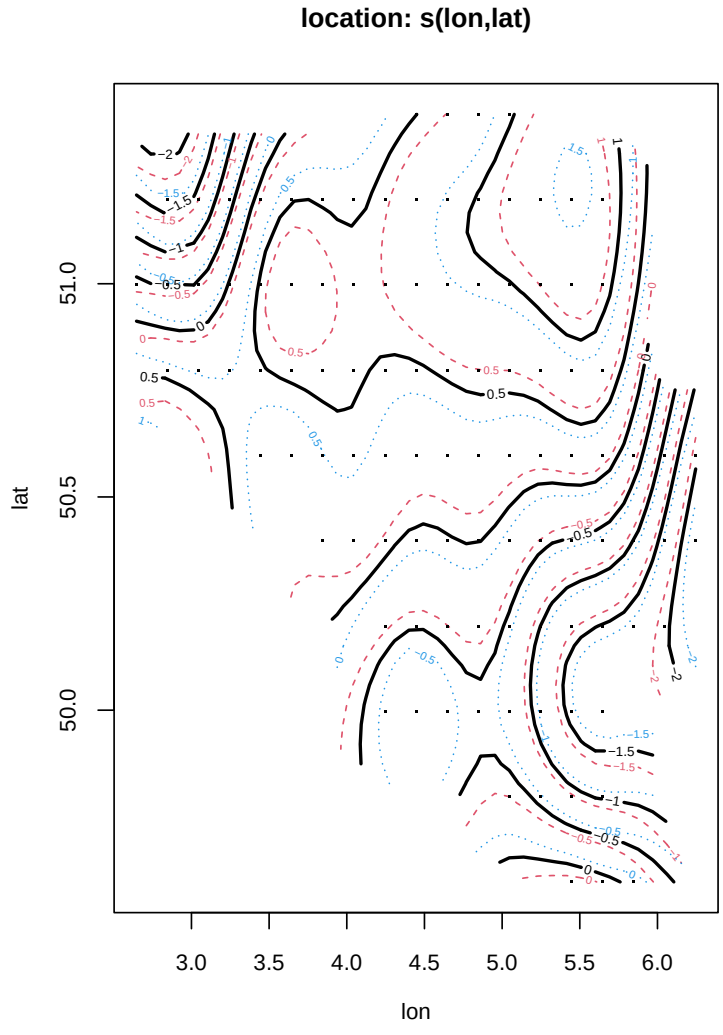
- Mixing splines that capture temporal and spatial variation gives spatio temporal variation
- Consider $\mu(t, lon, lat) = \mu_0 + g_1(t) + g_2(lon, lat)$



Spatio-temporal variation

- The forms of temporal and spatial variation can be isolated





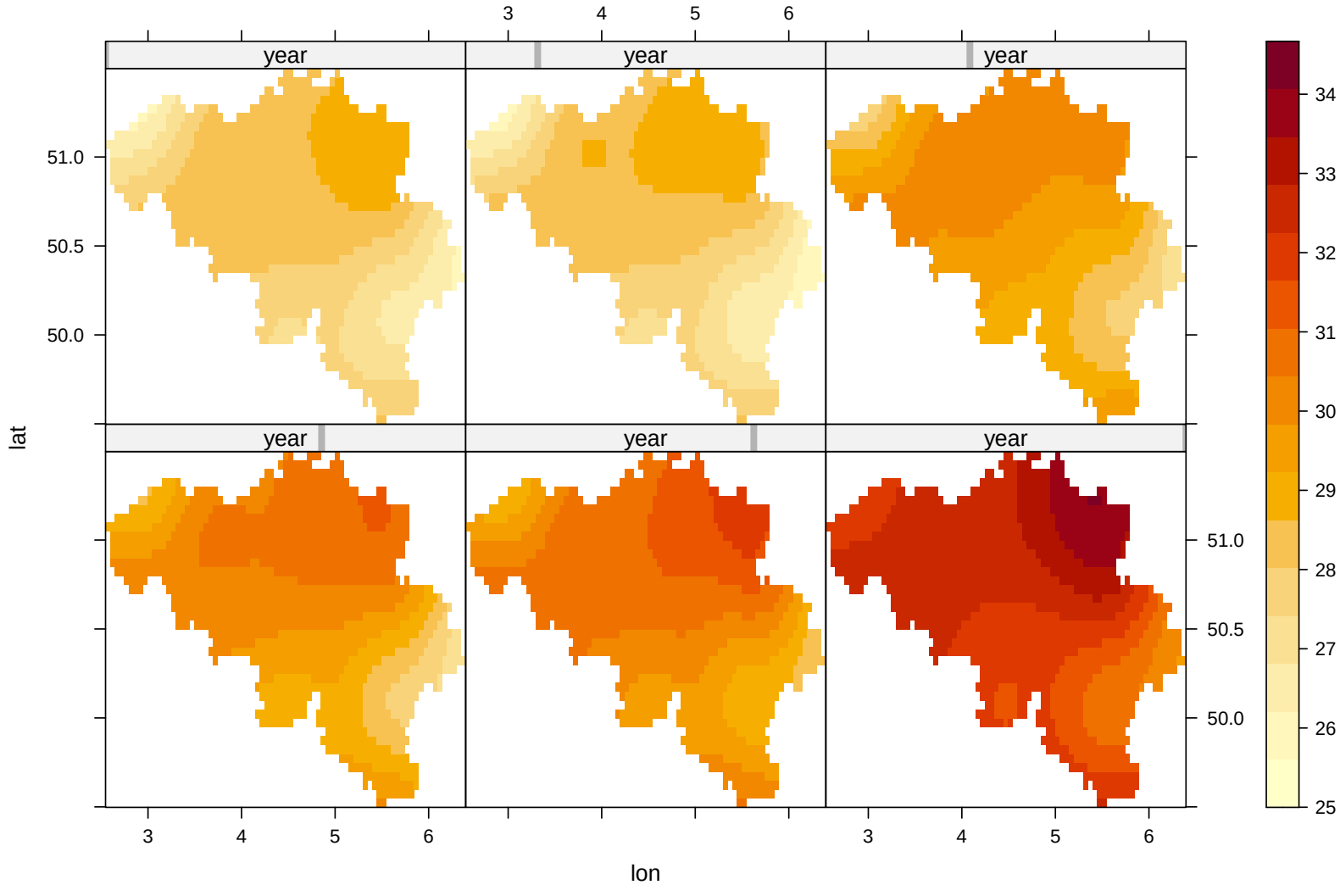
Spatio-temporal variation

- Finally, we might want to assume that the estimated form of spatial variation varies with time
- This can be achieved by taking the tensor product of splines

Spatio-temporal variation

- Consider $\mu(t, lon, lat) = \mu_0 + tensor(g_1(t), g_2(lon, lat))$

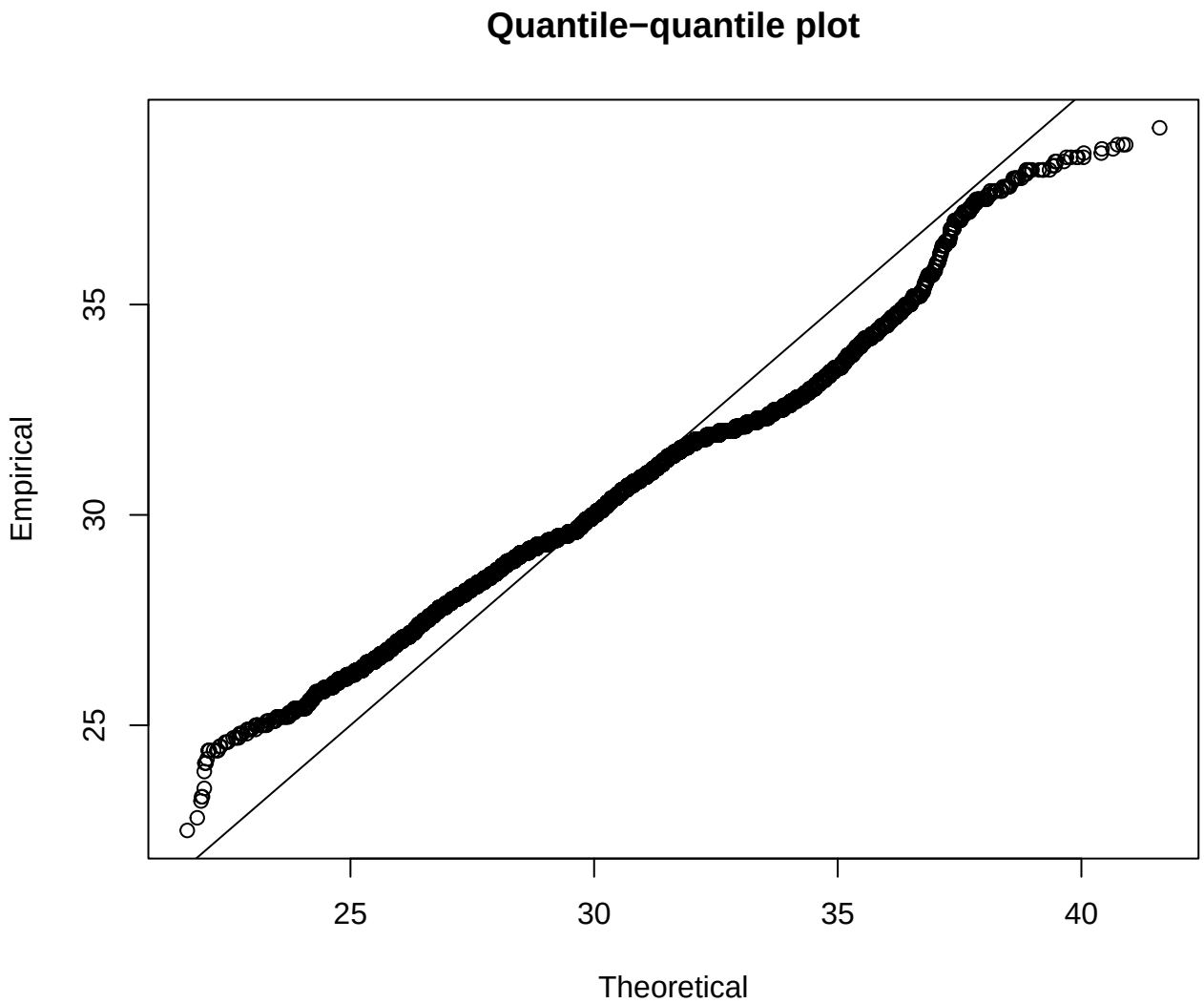
```
> ~ te(year, lon, lat, bs = c('cr', 'tp'), d = c(1, 2))
```



Model checks

- The quantile-quantile plot is a useful model check
  - here the model isn't capturing the data as well as we'd like

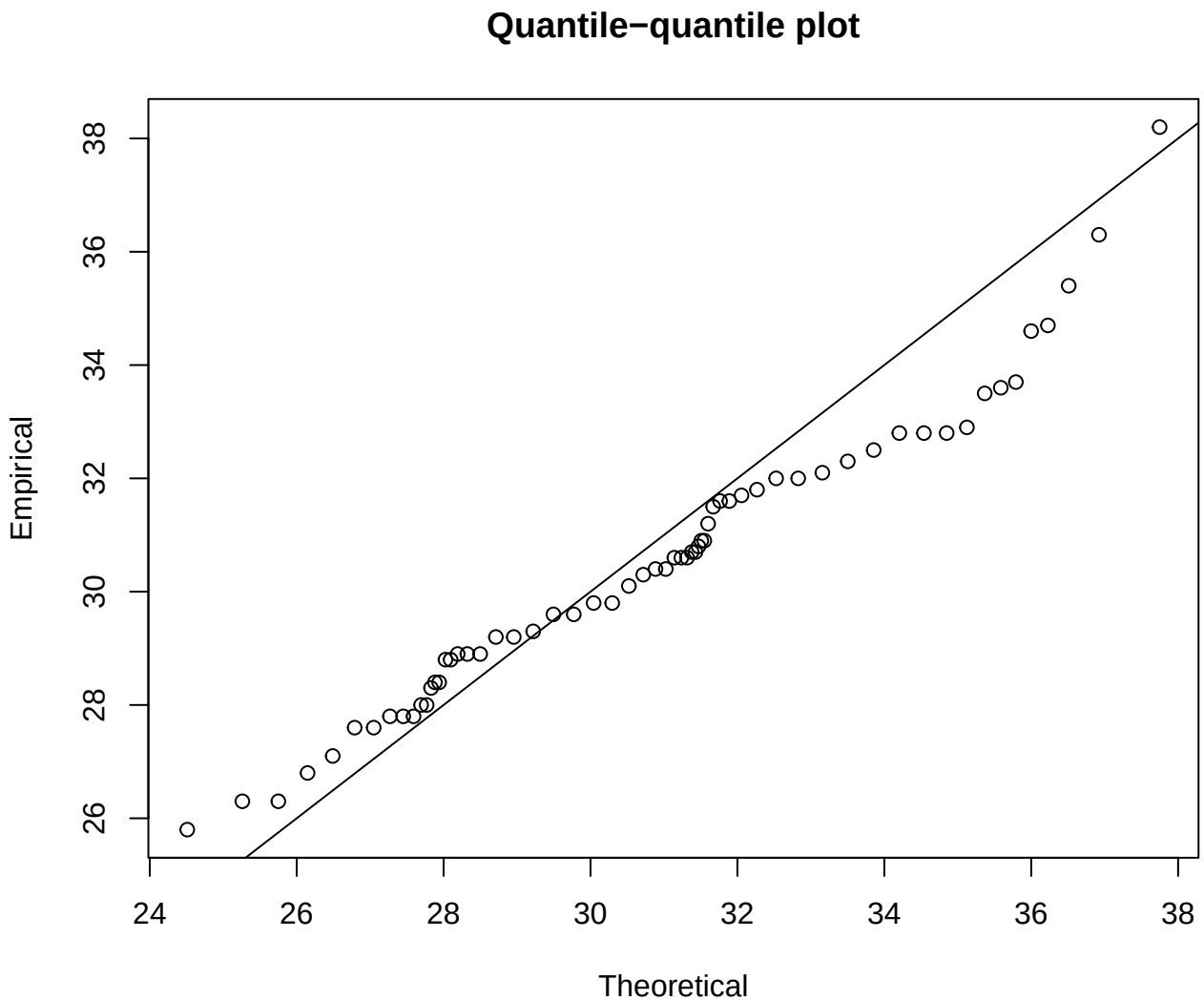
```
> predict(fitted_model, type = 'qqplot')
```



Model checks

- Sometimes it's useful to look a subset of the data to identify model inadequacies

```
> predict(fitted_model, newdata = brussels, type = 'qqplot')
```



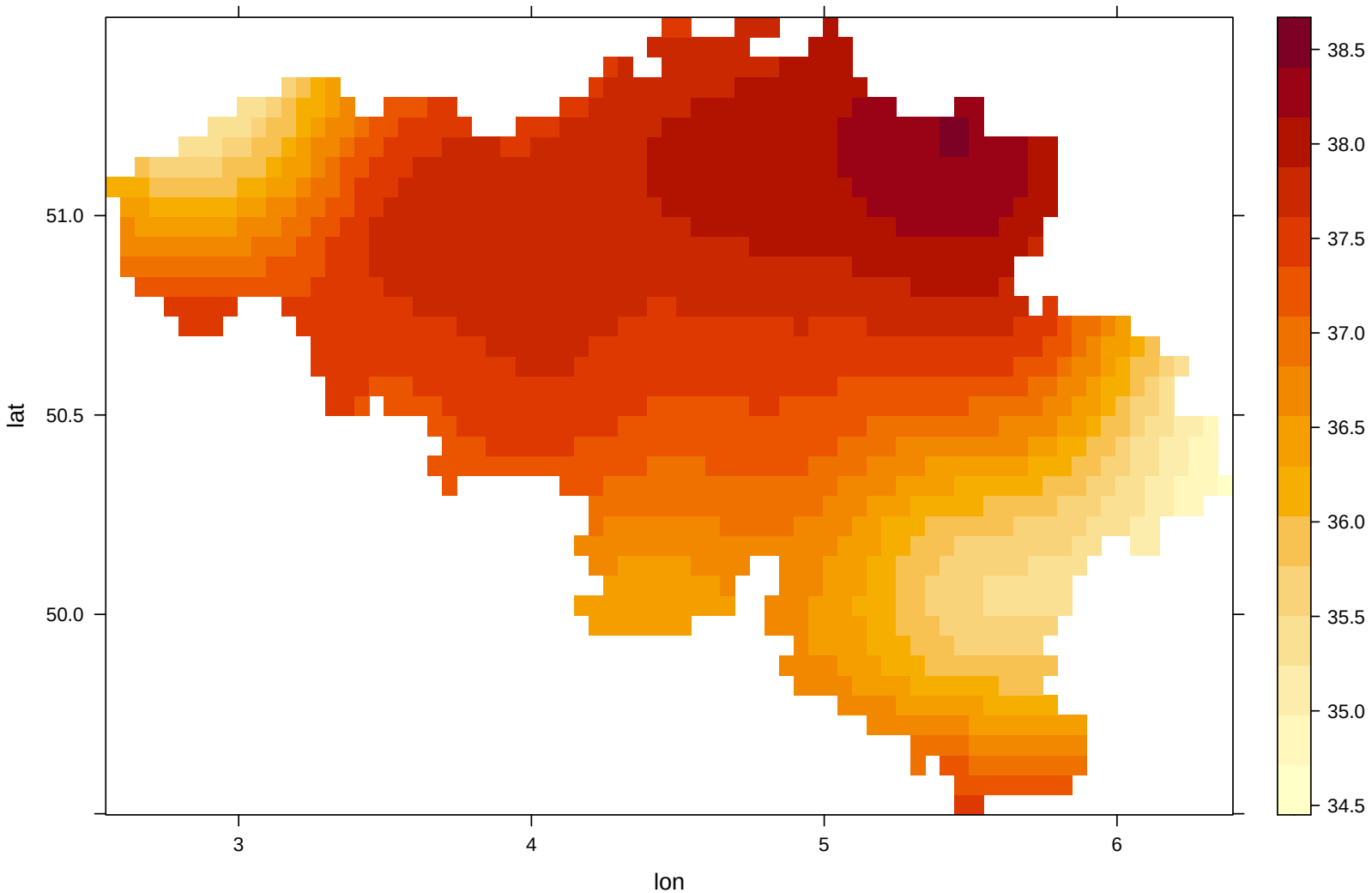
Return level estimation

- We're often interested in quantiles of our estimated extreme value distribution
  - e.g. the 0.99 quantile can have a 100-year return level interpretation

Return level estimation

- Quantiles are calculated via $z_q = \mu - \frac{\psi}{\xi}\left\{1 - \left[-\log(-q)\right]^{-\xi}\right\}$

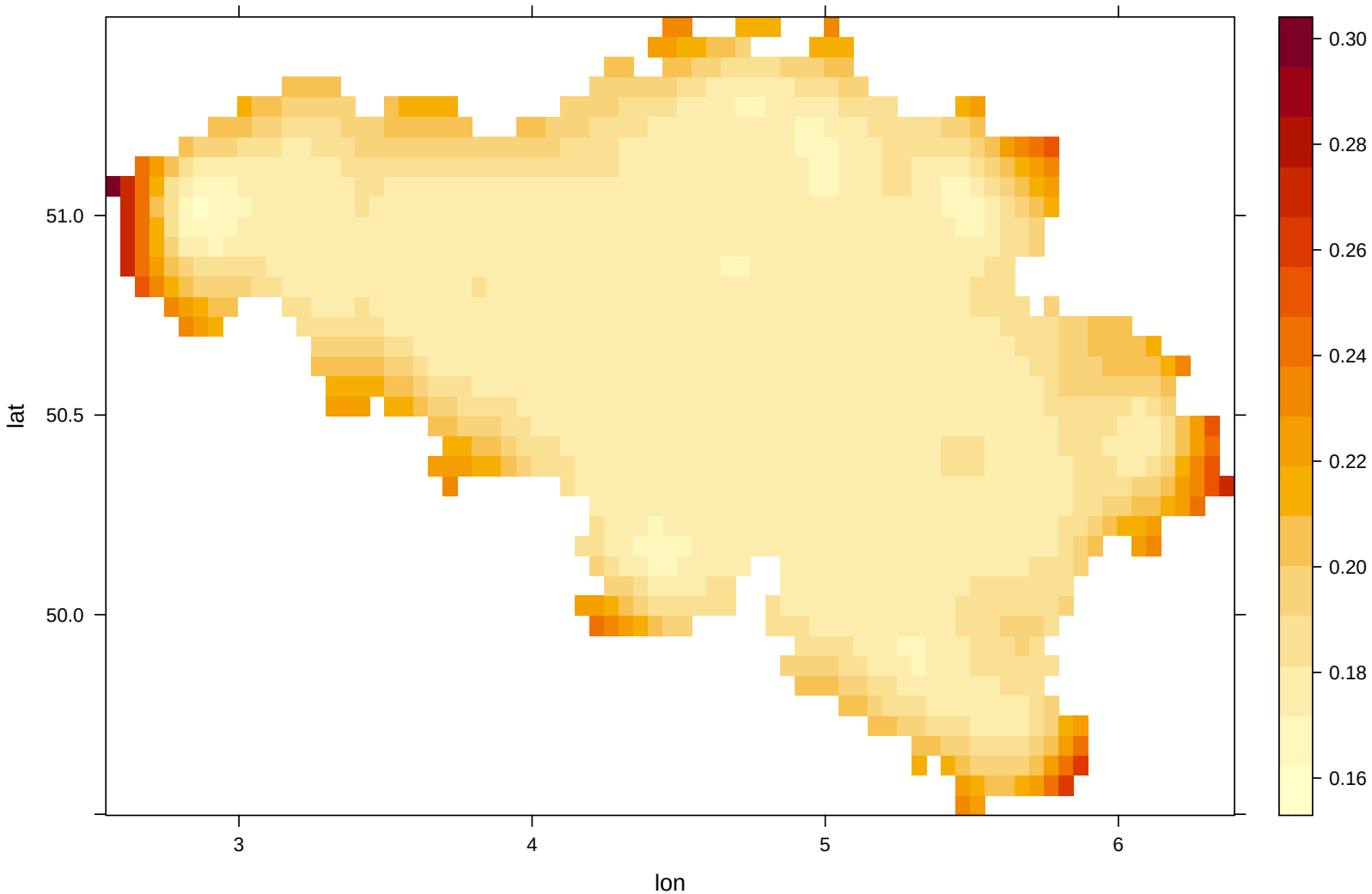
```
> predict(fitted_model, prob = 0.99)
```



Uncertainty quantification

- It's important to report the accuracy of any extreme value model estimates

```
> predict(fitted_model, prob = 0.99, se.fit = TRUE)
```



Summary

- Splines are a robust way of allowing spatio-temporal variation in extreme value models, especially for environmental phenomena
- The `R` package `evgam` allows the fitting of extreme value distributions that vary using GAM forms

References

Chavez-Demoulin, V. and A. C. Davison (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society C 54*(1), 207–222.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag.

Wood, S. N. (2010). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Wood, S. N., N. Pya, and B. Säfken (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association 111*(516), 1548–1563.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *Journal of the American Statistical Association 114*(528), 1865–1879.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software 103*(3), 1–26.