

# Hierarchical Generalized Additive Models in ecology: an introduction with mgcv

Eric J. Pedersen<sup>1,2,\*</sup>, David L. Miller<sup>3,4</sup>, Gavin L. Simpson<sup>5</sup>, and Noam Ross<sup>6</sup>

<sup>1</sup>Northwest Atlantic Fisheries Center, Fisheries and Oceans Canada, St. John's, NL, Canada

<sup>2</sup>Department of Biology, Memorial University, St. John's, NL, Canada

<sup>3</sup>Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, UK

<sup>4</sup>School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

<sup>5</sup>Institute of Environmental Change and Society, University of Regina, Regina, SK, Canada

<sup>6</sup>Ecohealth Alliance, New York, NY, USA

\*Corresponding author. Email: eric.j.pedersen@gmail.com

## Abstract

In this paper, we discuss an extension to two popular approaches to modelling complex structures in ecological data: the generalized additive model (GAM) and the hierarchical model (HGLM). The hierarchical GAM (HGAM), allows modelling of nonlinear functional relationships between covariates and outcomes where the shape of the function itself varies between different grouping levels. We describe the theoretical connection between these models, HGLMs and GAMs, explain how to model different assumptions about the degree of inter-group variability in functional response, and show how HGAMs can be readily fitted using existing GAM software, the mgcv package in R. We also discuss computational and statistical issues with fitting these models, and demonstrate how to fit HGAMs on example data.

## I: Introduction

Two of the most popular and powerful modelling techniques currently in use by ecologists are generalized additive models (GAMs; Wood, 2017a) for modelling flexible regression functions,

and generalized linear mixed models (“hierarchical generalized linear models” (HGLMs) or simply “hierarchical models”; Bolker et al., 2009; Gelman et al., 2013) for modelling between-group variability in regression relationships.

At first glance, GAMs and HGLMs are very different tools used to solve different problems. GAMs are used to estimate smooth functional relationships between predictor variables and the response. HGLMs, on the other hand, are used to estimate linear relationships between predictor variables and response (although nonlinear relationships can also be modeled through quadratic terms or other transformations of the predictor variables), but impose a structure where predictors are organized into groups (often referred to as “blocks”) and the relationships between predictor and response may vary across groups. Either the slope or intercept, or both, may be subject to grouping. A typical example of HGLM use might be to include site-specific effects in a model of population counts, or to model individual level heterogeneity in a study with repeated observations of multiple individuals.

However, the connection between HGLMs and GAMs is quite deep, both conceptually and mathematically (Verbyla et al., 1999). HGLMs and GAMs fit highly variable models by “pooling” parameter estimates towards one another, by penalizing squared deviations from some simpler model. In an HGLM, this occurs as group-level effects are pulled towards global effects (penalizing the squared differences between each group-level parameter estimate and the global effect). In a GAM, this occurs in the enforcement of a smoothness criterion on the variability of a functional relationship, pulling parameters towards some function that is assumed to be totally smooth (such as a straight line) by penalizing squared deviations from that totally smooth function.

Given this connection, a natural extension to the standard GAM framework is to allow smooth functional relationships between predictor and response to vary between groups, but in such a way that the different functions are in some sense pooled toward a common shape. We often want to know both how functional relationships vary between groups, and if a relationship holds across groups. We will refer to this type of model as a *hierarchical GAM*, or HGAM.

There are many potential uses for HGAMs. For example, we can use HGAMs to estimate how the maximum size of different fish species varies along a common temperature gradient (figure 1). Each species will typically have its own response function, but since the species overlap in range, they should have similar responses over at least some of the temperature gradient; figure 1 shows all three species reach their largest maximum sizes in the centre of the temperature gradient. Estimating a separate function for each species throws away a lot of shared information and could result in highly noisy function estimates if there were only a few data points for each species. Estimating a single average relationship could result in a function that did not predict any specific group well. In our example, using a single global temperature-size relationship would miss that the three species have distinct temperature optima, and that the orange species is significantly smaller at all temperatures than the other two (figure 1). We prefer a hierarchical model that includes a global temperature-size curve plus species-specific curves that were penalized to be close to the mean function.

This paper discusses several approaches to group-level smoothing, and corresponding trade-offs. We focus on fitting HGAMs with the popular **mgcv** package for the R statistical programming

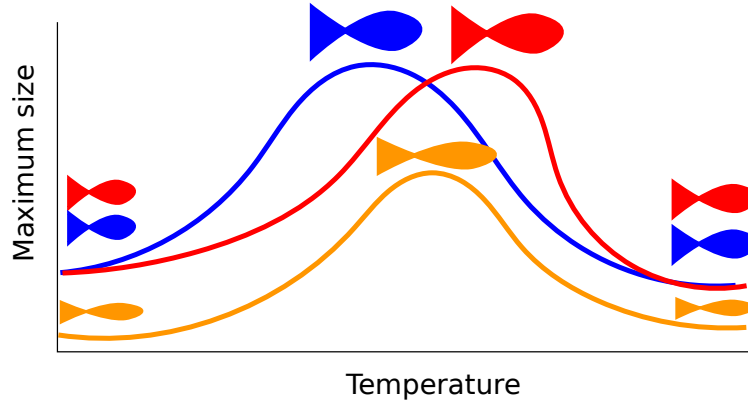


Figure 1: Hypothetical example of functional variability between different group levels. Each line indicates how the abundance for different species of fish in a community might vary as a function of average water temperature. The orange species shows lower abundance at all temperatures, and the red and blue species differ at which temperature they can achieve the maximum possible size. However, all three curves are similarly smooth and peak close to one another relative to the entire range of tested temperatures.

language, which allows for a variety of HGAM model structures and fitting strategies. We discuss options available to the modeller and practical and theoretical reasons for choosing them. We demonstrate the different approaches across a range of case studies.

This paper is divided into five sections. Part II is a brief review of how GAMs work and their relation to hierarchical models. In part III, we discuss different HGAM formulations, what assumptions each model makes about how information is shared between groups, and different ways of specifying these models in **mgcv**. In part IV, we work through example analyses using this approach, to demonstrate the modelling process and how HGAMs can be incorporated into the ecologist’s quantitative toolbox. Finally, in part V, we discuss some of the computational and statistical issues involved in fitting HGAMs in **mgcv**. We have also included all the code needed to reproduce the results in this manuscript in supplemental code (online), and on the GitHub repository associated with this paper [github.com/noamross/mixed-effect-gams](https://github.com/noamross/mixed-effect-gams).

## II: A review of Generalized Additive Models

The generalized linear model (GLM; McCullagh & Nelder, 1989) relates the mean of a response ( $y$ ) to a linear combination of explanatory variables. The response is assumed to be conditionally distributed according to some exponential family distribution (e.g., binomial, Poisson or Gamma distributions for trial, count or strictly positive real response, respectively). The generalized additive model (GAM; Hastie & Tibshirani, 1990; Ruppert, Wand & Carroll, 2003; Wood, 2017a) allows the relationships between the explanatory variables (henceforth covariates) and the response to be described by smooth curves (usually *splines* (de Boor,

1978), but potentially other structures). In general we have models of the form:

$$\mathbb{E}(Y) = g^{-1} \left( \beta_0 + \sum_{j=1}^J f_j(x_j) \right),$$

where  $\mathbb{E}(Y)$  is the expected value of the response  $Y$  (with an appropriate distribution and link function  $g$ ),  $f_j$  is a smooth function of the covariate  $x_j$ ,  $\beta_0$  is an intercept term and  $g^{-1}$  is the inverse link function. Hereafter, we will refer to these smooth functions as *smoothers*. In the example equation above, there are  $J$  smoothers and each is a function of only one covariate, though it is possible to construct smoothers of multiple variables.

Each smoother  $f_j$  is represented by a sum of  $K$  simpler, fixed *basis functions* ( $b_{j,k}$ ) multiplied by corresponding coefficients ( $\beta_{j,k}$ ), which need to be estimated:

$$f_j(x_j) = \sum_{k=1}^K \beta_{j,k} b_{j,k}(x_j).$$

$K$ , referred to as “basis size”, “basis complexity” or “basis richness”, determines the maximum complexity of each smoother.

It would seem that large basis size could lead to overfitting, but this counteracted by a *smoothing penalty* that influences basis function coefficients so as to prevent excess wiggleness and ensure that appropriate complexity of each smoother. For each smoother, one or more *penalty matrices* ( $\mathbf{S}$ ), specific to the form of the basis functions, is pre- and post-multiplied by the parameter vector  $\boldsymbol{\beta}$  to calculate the penalty ( $\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$ ). A penalty term is then added to the model log-likelihood  $L$ , controlling the trade-off via a *smoothing parameter* ( $\lambda$ ). The penalized log-likelihood used to fit the model is thus:

$$L - \lambda \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$$

Figure 2 shows an example of how different choices of the smoothing parameter ( $\lambda$ ) affect the shape of the resulting smoother. Data (points) were generated from the blue function and noise added to them. In the left plot  $\lambda$  was selected using Restricted Maximum Likelihood (REML) to give a good fit to the data, in the middle plot  $\lambda$  was set to zero, so the penalty has no effect and the function interpolates the data, the right plot shows when  $\lambda$  is set to a very large value, so the penalty removes all terms that have any wiggleness, giving a straight line.

To measure the complexity of a penalized smooth terms we use the *effective degrees of freedom* (EDF), which at a maximum is the number of coefficients to be estimated in the model, minus any constraints. The EDF can take non-integer values and larger values indicate more wiggly terms (see Wood (2017a, Section 6.1.2) for further details). The number of basis functions,  $k$  sets a maximum for the EDF, as a smoother cannot have more than  $k$  EDF. When the EDF is well below  $k$ , increasing  $k$  generally has very little effect on the shape of the function. In general,  $k$  should be set large enough to allow for potential variation in the smoother while still staying low enough to keep computation time low (see section V for more on this). In **mgcv**, the function `mgcv::check.gam` can be used to determine if  $k$  has been set too low.

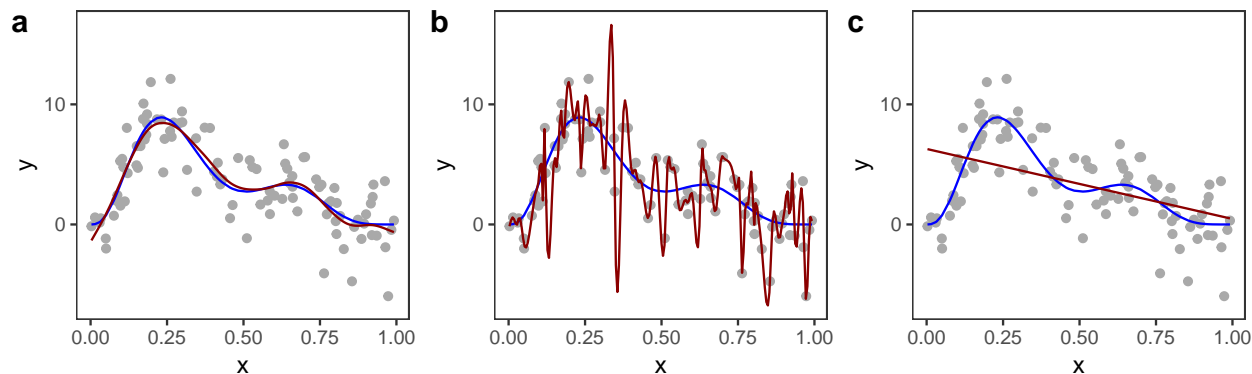


Figure 2: Effect of different choices of smoothing parameter ( $\lambda$ ) on the shape of the resulting smoother. Left:  $\lambda$  estimated using REML; middle:  $\lambda$  set to zero (no smoothing); Right:  $\lambda$  is set to a very large value.

Random effects are also “smooths” in this framework. In this case, the penalty matrix is the inverse of the correlation matrix of the basis function coefficients (Kimeldorf & Wahba, 1970; Wood, 2017a). To include a simple single-level random effect to account for variation in group means (intercepts) there will be one basis function for each level of the grouping variable, that takes a value of 1 for any observation in that group and 0 for any observation not in the group. The penalty matrix for these terms is a  $n_g$  by  $n_g$  identity matrix, where  $n_g$  is the number of groups. This means that each group-level coefficient will be penalized in proportion to its squared deviation from zero. This is equivalent to how random effects are estimated in standard mixed effect models. The penalty term is then proportional to the inverse of the variance of the fixed effect estimated by standard hierarchical model software (Verbyla et al., 1999).

This connection between random effects and splines extends beyond the varying-intercept case. Any single-penalty basis-function representation of a smooth can be transformed so that it can be represented as a combination of a random effect with an associated variance, and possibly one or more fixed effects. See Verbyla et al. (1999) or Wood, Scheipl & Faraway (2013) for a more detailed discussion on the connections between these approaches.

## Basis types and penalty matrices

Different types of smoothers are useful for different needs, and have different associated penalty matrices for their basis function coefficients. In the examples in this paper, we will use three types of smoothers: thin plate regression splines, cyclic cubic smoothers, and random effects.

Thin plate regression splines (TPRS; Wood, 2003), are a general purpose spline basis which can be used for problems in any number of dimensions, provided one can assume that the amount of smoothing in any of the covariates is the same (so called isotropy or rotational invariance). TPRS, like many splines, use a penalty matrix made up of terms based on the

the integral of the squared derivatives of basis functions across their range (see Wood (2017a) page 216 for details on this penalty). Models that overfit the data will tend to have large derivatives, so this penalization reduces wiggleness. We will refer to the order of penalized derivatives as  $m$ . Typically, TPRS are second-order ( $m = 2$ ), meaning that the penalty is proportionate to the integral of the squared second derivative. However, TPRS may be of lower order ( $m = 1$ , penalizing squared first derivatives), or higher order ( $m > 2$ , penalizing squared higher order derivatives). We will see in section III how lower-order TPRS smoothers are useful in fitting HGAMs. Example basis functions and penalty matrix  $\mathbf{S}$  for a  $m = 2$  TPRS with six basis functions for evenly spaced data are shown in figure 3.

Cyclic cubic smoothers are another smoother that penalizes the squared second derivative of the smooth across the function. In these, though, start and end of the smoother are constrained to match in value and first derivative. These are useful for fitting models with cyclic components such as seasonal effects. We will use these smoothers to demonstrate how to fit HGAMs to cyclic data.

## Smoothing penalties vs. shrinkage penalties

Penalties can have two effects on how well a model fits: they can penalize how wiggly a given term is (smoothing) and they can penalize the absolute size of the function (shrinkage). The penalty can only affect the components of the smoother that have derivatives (the *range space*), not the other parts (the *null space*). For 1-dimensional thin plate regression splines (when  $m = 2$ ), this means that there is a linear term left in the model, even when the penalty is in full force (as  $\lambda \rightarrow \infty$ ), as shown in figure 3 (this is also why figure 2c resulted in a linear, rather than flat, fit to the data). The random effects smoother we discussed earlier is an example of a pure shrinkage penalty; it penalizes all deviations away from zero, no matter the pattern of those deviations. This will be useful later in section III, where we use random effect smoothers as one of the components of a HGAM.

## Interactions between smooth terms

It is also possible to create interactions between covariates with different smoothers (or degrees of smoothness) assumed for each covariate, using *tensor products*. For instance, if one wanted to estimate the interacting effects of temperature and time (in seconds) on some outcome, it would not make sense to use a two-dimensional TPRS smoother, as that would assume that a one degree change in temperature would equate to a one second change in time. Instead, a tensor product allows us to create a new set of basis functions that allow for each marginal function (here temperature and time) to have its own marginal smoothness penalty. A different basis can be used in each marginal smooth, as required for the data at hand.

There are two approaches used in **mgcv** for generating tensor products. The first approach (Wood, 2006a) essentially creates an interaction of each pair of basis functions for each marginal term, and a penalty for each marginal term that penalizes the average wiggleness in that term; in **mgcv**, these are created using the `te()` function. The second approach (Wood,

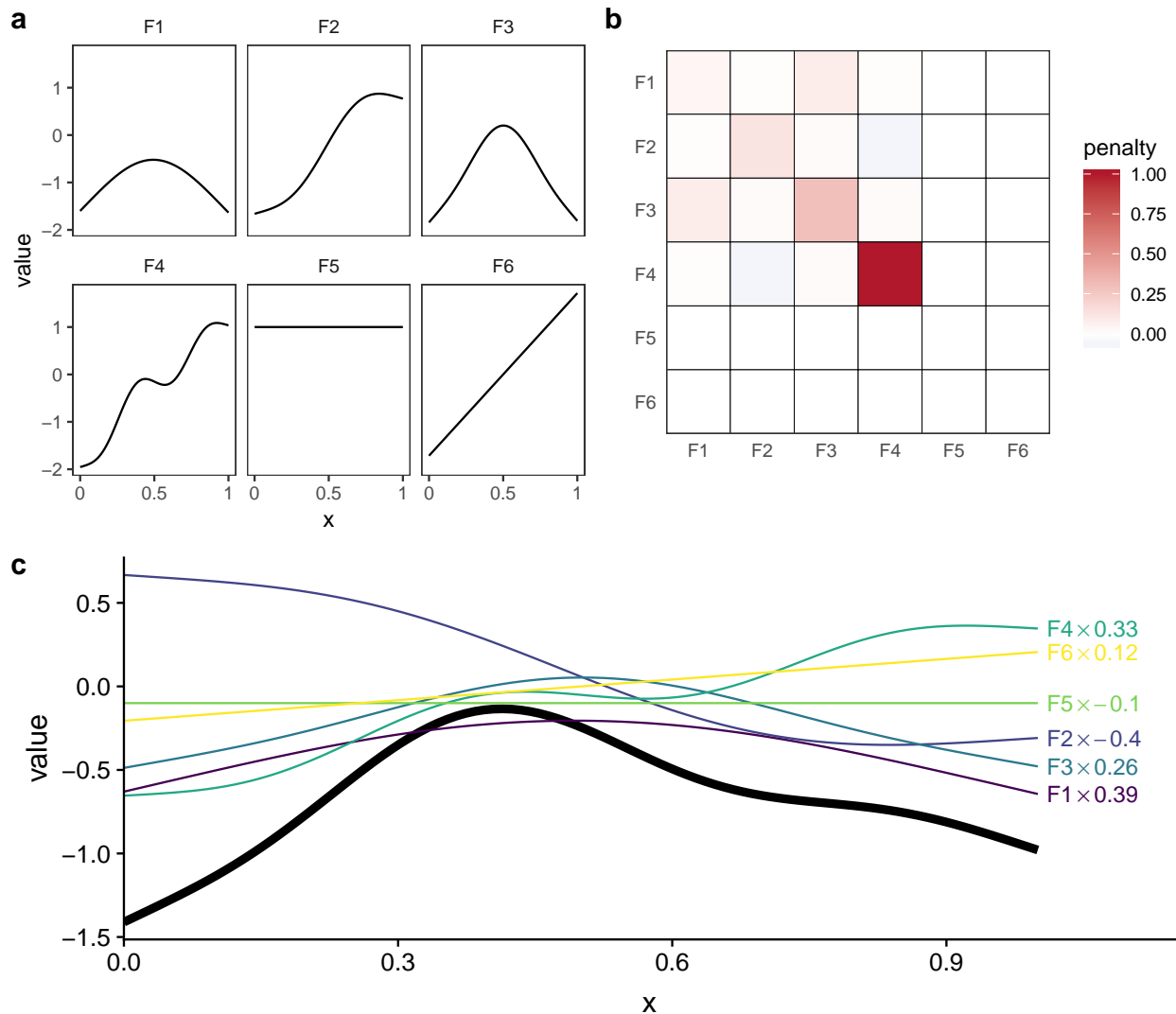


Figure 3: a) Examples of the basis functions associated with a six basis function thin plate spline ( $m=2$ ), calculated for data,  $x$ , spread evenly between  $x = 0$  and  $x = 1$ . Each line represents a single basis function. b) The smoothing penalty matrix for the thin plate smoother. Red entries indicate positive values and blue indicate negative values. For example, functions  $F3$  and  $F4$  would have the greatest proportionate effect on the total penalty (as they have the largest values on the diagonal), whereas function  $F5$  and  $F6$  would not contribute to the wiggleness penalty at all (all the values in the 5th and 6th row and column of the penalty matrix are zero). This means these functions are in the null space of the penalty matrix, and are treated as completely smooth. c) An example of how the basis functions add up to create a single smooth function. Thin coloured lines represent each basis function multiplied by a coefficient, and the solid black line is the sum of those basis functions.

Scheipl & Faraway, 2013) separates each penalty into penalized (range space) and unpenalized components (null space; components that don't have derivatives, such as intercept and linear terms in a one-dimensional cubic spline), then creates new basis functions and penalties for all pair-wise combinations of penalized and unpenalized components between all pairs of marginal bases; in **mgcv**, these are created using the **t2()** function. The advantage of the first method is that it requires fewer smoothing parameters, so is faster to estimate in most cases. The advantage of the second method is that the tensor products created this way only have a single penalty associated with each marginal basis (unlike the **te()** approach, where each penalty applies to all basis functions), so it can be fitted using standard mixed effect software such as **lme4** (Bates et al., 2015).

## Comparison to hierarchical linear models

Hierarchical generalized linear models (Gelman, 2006; HGLMs; also referred to as generalized linear mixed effect models, multilevel models etc; e.g., Bolker et al., 2009) are an extension of regression modelling that allows the inclusion of terms in the model that account for structure in the data — the structure is usually of the form of a nesting of the observations. For example, in an empirical study, individuals may be nested within sample sites, sites are nested within forests, and forests within provinces. The depth of the nesting is limited by the fitting procedure and number of parameters to estimate.

HGLMs are a highly flexible way to think about grouping in ecological data; the groupings used in models often refer to the spatial or temporal scale of the data (McMahon & Diez, 2007) though can be based on any useful grouping.

We would like to be able to think about the groupings in our data in a similar way, even when the covariates in our model are related to the response in a smooth way. The next section investigates the extension of the smoothers we showed above to the case where observations are grouped and we model group-level smoothers.

## III: What are hierarchical GAMs?

### What do we mean by hierarchical smoothers?

In this section, we will describe how to model inter-group variability using smooth curves and how to fit these models using **mgcv**. Model structure is key in this framework, so we start with three model choices:

1. Should each group have its own smoother, or will a common smoother suffice?
2. Do all of the group-specific smoothers have the same wiggleness, or should each group have its own smoothing parameter?
3. Will the smoothers for each group have a similar shape to one another — a shared global smoother?



These three choices result in five possible models (figure 4):

1. A single common smoother for all observations.
2. A global smoother plus group-level smoothers that have the same wiggleness.
3. A global smoother plus group-level smoothers with differing wiggleness.
4. Group-specific smoothers without a global smoother, but with all smoothers having the same wiggleness.
5. Group-specific smoothers with different wiggleness.

It is important to note that “similar wiggleness” and “similar shape” are two distinct concepts; functions can have very similar wiggleness but very different shapes. Wiggleness measures how quickly a function changes across its range, and it is easy to construct two functions that differ in shape but have the same wiggleness. For this paper, we consider two functions to have similar shape if the average squared distance between the functions is small (assuming the functions have been scaled to have a mean value of zero across their ranges). This definition is somewhat restricted; for instance, a cyclic function would not be considered to have the same shape as a phase-shifted version of the same function, nor would two normal distributions with the same mean but different standard deviations. The benefit of this definition of shape, however, is that it is straightforward to translate into penalties akin to those described in section II. Figure 4, model 4 illustrates the case where models have different shapes. Similarly, two curves could have very similar overall shape, but differ in their wiggleness. For instance, one function could be equal to another plus a high-frequency oscillation term. Figure 4, model 3 illustrates this.

We will discuss the trade-offs between different models and guidelines about when each of these models is appropriate in section V. The remainder of this section will focus on how to specify each of these five models using **mgcv**.

## Coding hierarchical GAMs in R

Each of the models in Figure 4 can be coded straightforwardly in **mgcv**. We will use two example datasets to demonstrate how to code these models (see the supplemental code to reproduce these examples):

A. The **CO2** dataset, available in R via the **datasets** package. This data is from an experimental study by Potvin, Lechowicz & Tardif (1990) of CO<sub>2</sub> uptake in grasses under varying concentrations of CO<sub>2</sub>, measuring how concentration-uptake functions varied between plants from two locations (Mississippi and Quebec) and two temperature treatments (chilled and warm). Twelve plants were used and CO<sub>2</sub> uptake measured at 7 CO<sub>2</sub> concentrations for each plant (figure 5a). Here we will focus on how to use HGAMs to estimate inter-plant variation in functional responses. This data set has been modified from the default version available with R, to recode the **Plant** variable as an unordered factor **Plant\_uo**<sup>1</sup>.

---

<sup>1</sup>Note that **mgcv** requires that grouping or categorical variables be coded as factors in R; it will raise an error message if passed data coded as characters. It is also important to know whether the factor is coded as ordered or unordered (see **?factor** for more details on this). This matters when fitting groupwise smoothers using the **by=** argument (as is used for fitting models 3 and 5, shown below). If the factor is unordered, **mgcv**

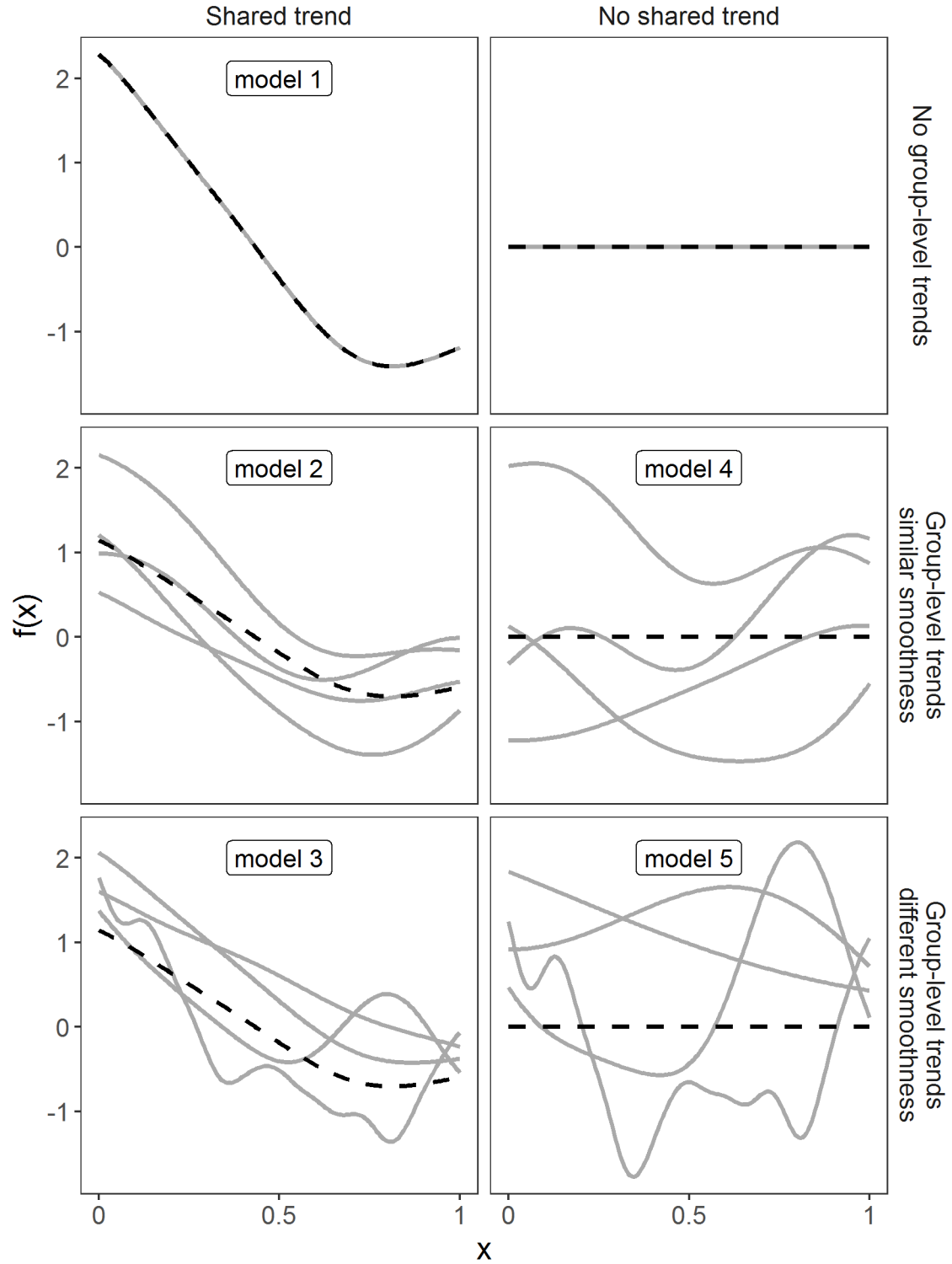


Figure 4: Alternate types of functional variation  $f(x)$  that can be fitted with HGAMs. The dashed line indicates the average function value for all groups, and each solid line indicates the functional value at a given predictor value for an individual group level. The null model (of no functional relationship between the covariate and outcome, top right), is not explicitly assigned a model number.

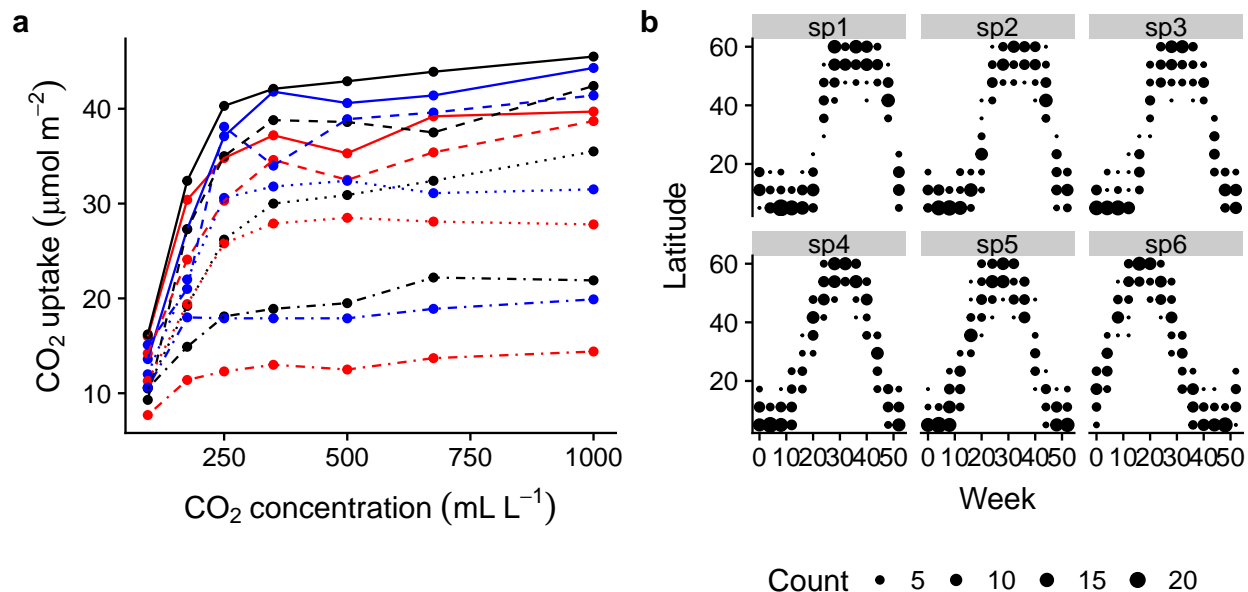


Figure 5: Example data sets used throughout section III. a) Grass CO<sub>2</sub> uptake versus CO<sub>2</sub> concentration for 12 individual plants. Color and linetype included to distinguish individual plant trends. b) Simulated data set of bird migration, with point size corresponding to weekly counts of 6 species along a latitudinal gradient (zeros excluded for clarity).

B. Data generated from a hypothetical study of bird movement along a migration corridor, sampled throughout the year (see supplemental code). This dataset consists of simulated sample records of numbers of observed locations of 100 tagged individuals each from six species of bird, at ten locations along a latitudinal gradient, with one observation taken every four weeks. Counts were simulated randomly for each species in each location and week by creating a species-specific migration curve that gave the probability of finding an individual of a given species in a given location, then simulated the distribution of individuals across sites using a multinomial distribution, and subsampling that using a binomial distribution to simulation observation error (i.e. not every bird present at a location would be detected). The data set (`bird_move`) consists of the variables `count`, `latitude`, `week` and `species` (figure 5b). This example allows us to demonstrate how to fit these models with interactions and with non-normal (count) data. The true model used to generate this data was model 2: a single global function plus species-specific deviations around that global function.

Throughout the examples we use Restricted Maximum Likelihood (REML) to estimate model coefficients and smoothing parameters. We strongly recommend using either REML or marginal likelihood (ML) rather than the default GCV criteria when fitting GAMs, for the

will set up a model with one smoother for each grouping level. If the factor is ordered, `mgcv` will set any basis functions for the first grouping level to zero. In model 3 the ungrouped smoother will then correspond to the first grouping level, rather than the average functional response, and the group-specific smoothers will correspond to deviations from the first group. In model 5, using an ordered factor will result in the first group not having a smoother associated with it at all.

reasons outlined in Wood (2011). In each case some data processing and manipulation has been done to obtain the graphics and results below. See supplemental code for details on data processing steps.

## A single common smoother for all observations (Model 1)

We start with the simplest model we can in our framework and include many details here to ensure that readers are comfortable with the terminology and R functions we are going to use later.

For our `CO2` data set, we will model  $\log_e(\text{uptake})$  as a function of two smoothers: a thin plate regression spline of  $\log_e$ -concentration, and a random effect for plant to model plant-specific intercepts. Mathematically:

$$\log_e(\text{uptake}_i) = f(\log_e(\text{conc}_i)) + \zeta_{\text{Plant\_uo}} + \varepsilon_i$$

where  $\zeta_{\text{Plant\_uo}}$  is the random effect for plant and  $\varepsilon_i$  is a Gaussian error term. Here we assume that  $\log_e(\text{uptake}_i)$  is normally distributed.

In R we can write our model as:

```
CO2_mod1 <- gam(log(uptake) ~ s(log(conc), k=5, bs="tp") +
                 s(Plant_uo, k=12, bs="re"),
                 data=CO2, method="REML", family="gaussian")
```

This is a common GAM structure, with a single smooth term for each variable. Specifying the model is similar to specifying a GLM in R via `glm()`, with the addition of `s()` terms to include one-dimensional or isotropic multidimensional smoothers. The first argument to `s()` are the terms to be smoothed, the type of smoother to be used for the term is specified by the `bs` argument, and the maximum number of basis functions is specified by `k`. There are different defaults in **mgcv** for `k`, depending on the type of smoother chosen; here we use a `tp` smoother (`bs="tp"`) for the concentration smoother, and set `k=5` as there are only 7 separate values of concentration measured, so the default `k=10` (for `tp`) would be too high; further, setting `k=5` saves on computational time (see section V). The random effect smoother (`bs="re"`) that we used for the `Plant_uo` factor has a default `k` equal to the number of levels in the grouping variable (here, 12). We specified `k=12` just to make this connection apparent.

Figure 6 illustrates **mgcv**'s default plotting out for `CO2_mod1`: the left panel shows the estimated smoother for concentration, and the right shows a quantile-quantile plot of the estimated random effects vs Gaussian quantiles, which can be used to check our model.

Looking at the effects by term is useful, but we are often interested in fitted values or predictions from our models. Using the built in prediction functions with **mgcv**, we can estimate what the fitted function (and uncertainty around it) should look like for each level, as shown in Figure 7 (see supplemental code for more details on how to generate these predictions).

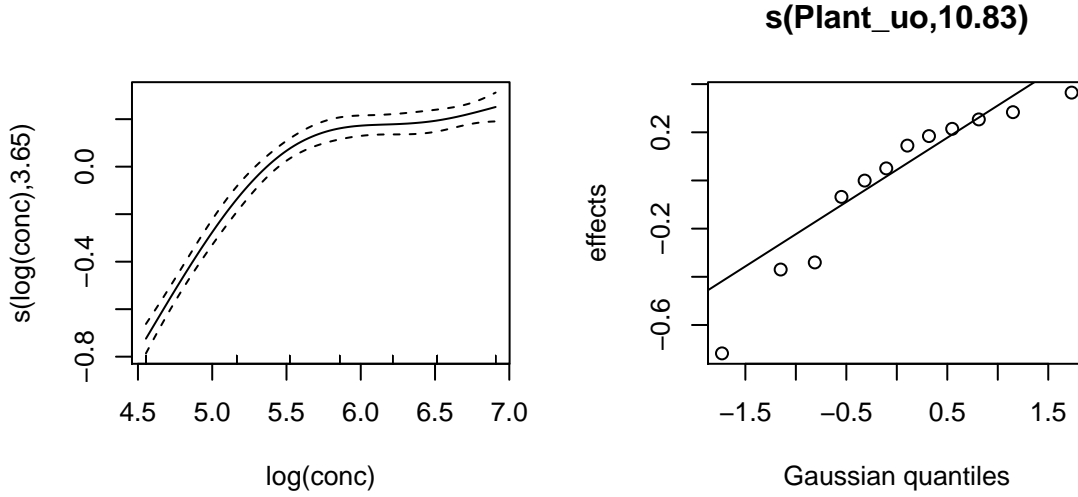


Figure 6: **mgcv** plotting output for model 1 applied to the CO2 dataset. Left shows the smoother of  $\log_e$  concentration and right plot shows a quantile-quantile plot of the random effects against Gaussian quantiles, used to check the appropriateness of the normal random effect assumption. Numbers in the vertical axis of the left figure and the title of the right give the effective degrees of freedom of the smoothers.

Examining these plots, we see that while functional responses among plants are similar, some patterns are not captured by this model. For instance, for plant Qc2, the model clearly underestimates CO2 uptake. A model including individual differences in functional responses may better explain variation.

For our bird example, we model the count of birds as a function of location and time, including their interaction. For this we structure the model as:

$$\mathbb{E}(\text{count}_i) = \exp(f(\text{week}_i, \text{latitude}_i))$$

where we assume that  $\text{count}_i \sim \text{Poisson}$ . For the smooth term,  $f$ , we employ a tensor product of **latitude** and **week**, using a thin plate regression spline (TPRS) for the marginal latitude effect, and a cyclic cubic regression spline for the marginal week effect to account for the cyclic nature of weekly effects (we expect week 1 and week 52 to have very similar values)<sup>2</sup>, both splines had basis complexity (**k**) of 10.

<sup>2</sup>The cyclic smoother requires that the start and end points of the cyclic variable are specified, via the **knots** argument; the smoother will have the exact same value at the start and end. In the absence of a specified start and end point, **gam** will assume the end points are the smallest and largest observed levels of the covariate (see `mgcv::smooth.construct.cc.smooth.spec` for more details). Note that in **bird\_mod1** we have specified week 0 and week 52 as the endpoints, as the first (week 1) and last weeks (week 52) of the year should not have exactly the same expected value.

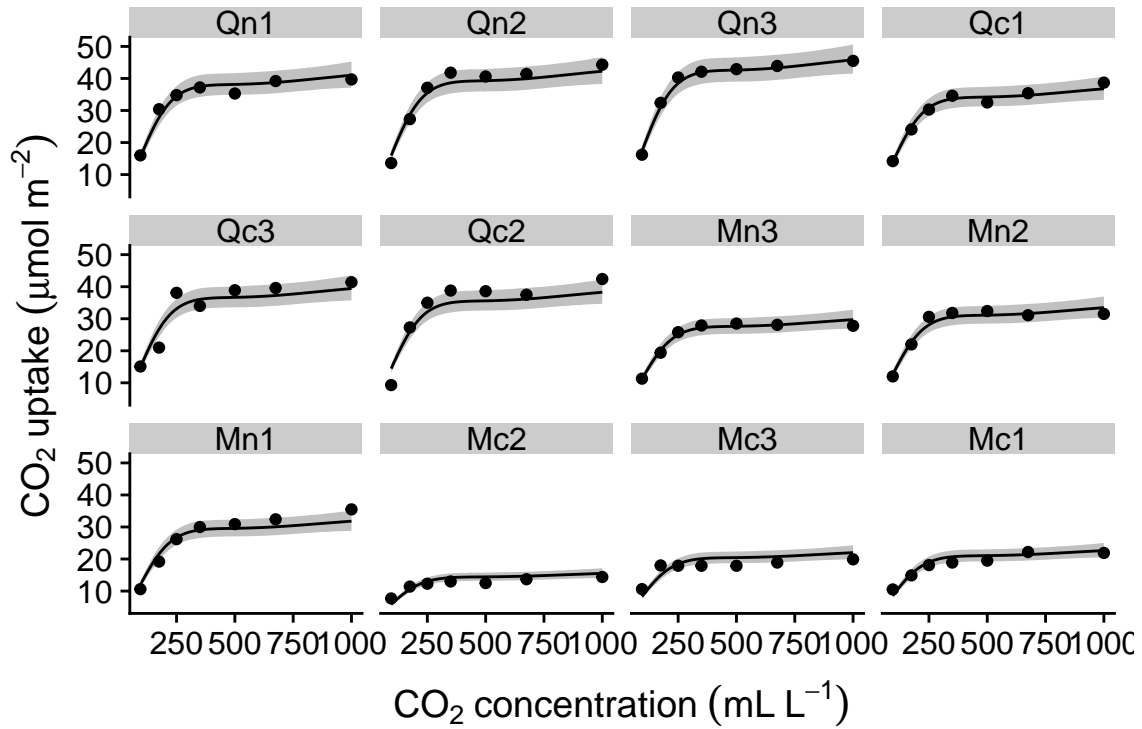


Figure 7: Predicted uptake function ( $\pm 2$  s.e.) for each plant, based on model 1 (a single global function for uptake plus a individual-level random effect intercept). Model predictions are for log-uptake, but are transformed here to show the fitted function on the original scale of the data.

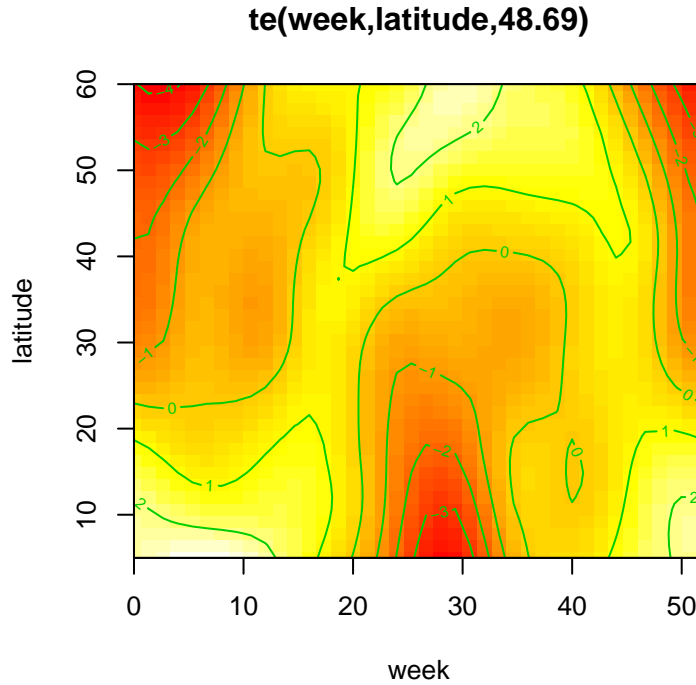


Figure 8: Plot illustrating the average log-abundance of all bird species at each latitude for each week, with yellow colours indicating more individuals and red colours fewer.

```
bird_mod1 <- gam(count ~ te(week, latitude, bs=c("cc", "tp"), k=c(10, 10)),
  data=bird_move, method="REML", family="poisson",
  knots = list(week = c(0, 52)))
```

Figure 8 shows the default plot (created by running `plot(bird_mod1, pages=1, scheme=2, rug=FALSE)`) for the week-by-latitude smoother. It shows birds starting at low latitudes in the winter then migrating to high latitudes from the 10th to 20th week, staying there for 15-20 weeks, then migrating back. However, the plot also indicates a large amount of variability in the timing of migration. The source of this variability is apparent when we look at the timing of migration of each species (cf. figure 5b).

All six species in figure 5b show relatively precise migration patterns, but they differ in the timing of when they leave their winter grounds and the amount of time they spend at their summer grounds. Averaging over all of this variation results in a relatively imprecise (diffuse) estimate of migration timing (figure 8), and viewing species-specific plots of observed versus predicted values (figure 9), it is apparent that the model fits some of the species better than others. This model could potentially be improved by adding inter-group variation in migration timing. The rest of this section will focus on how to model this type of variation.

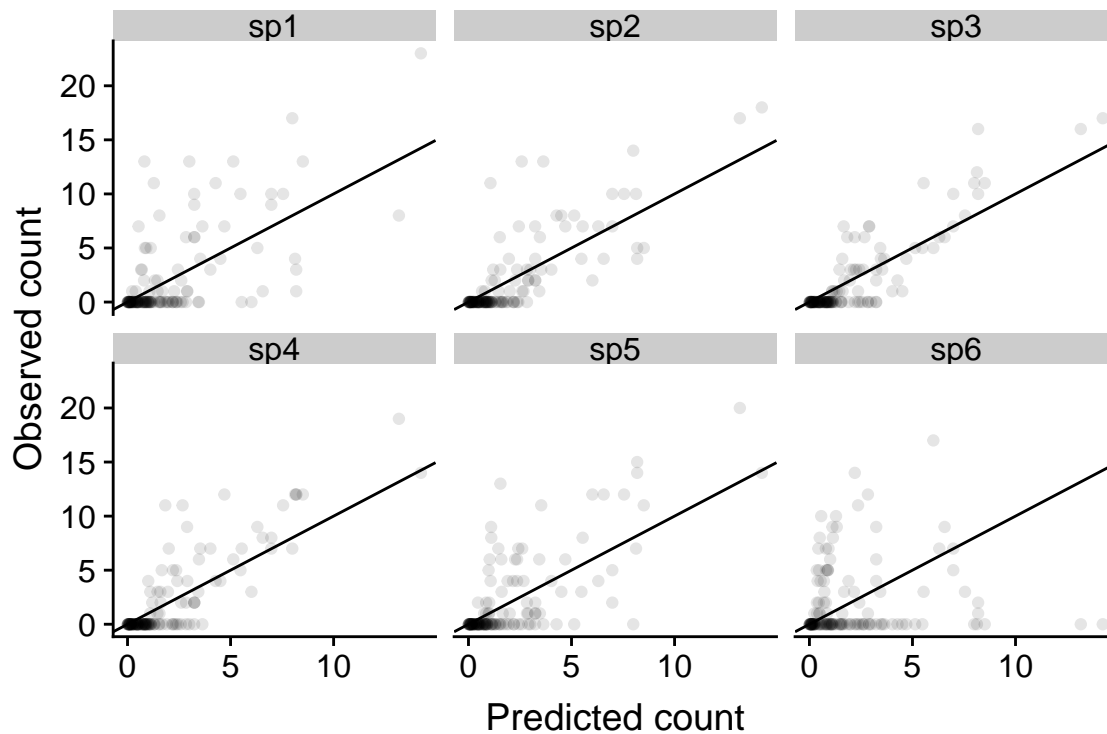


Figure 9: Observed counts by species versus predicted counts from `bird_mod1` (1-1 line added as reference). If our model fitted well we would expect that all species should show similar patterns of dispersion around the 1-1 line (and as we are assuming the data is Poisson, the variance around the mean should equal the mean). Instead we see that variance around the predicted value is much higher for species 1 and 6.



## A single common smoother plus group-level smoothers that have the same wiggliness (Model 2)

Model 2 is a close analogue to a GLMM with varying slopes: all groups have similar functional responses, but inter-group variation in responses is allowed. This approach works by allowing each grouping level to have its own functional response, but penalizing functions that are too far from the average.

This can be coded in **mgcv** by explicitly specifying one term for the global smoother (as in model 1 above) then adding a second smooth term specifying the group-level smooth terms, using a penalty term that tends to draw these group-level smoothers to zero. For one-dimensional smoothers, **mgcv** provides an explicit basis type to do this, the factor-smoother interaction or "fs" basis (see `?mgcv::factor.smooth.interaction` for details). This smoother creates a copy of each set of basis functions for each level of the grouping variable, but only estimates one smoothing parameter for all groups. To ensure that all parts of the smoother can be shrunk towards zero effect, each component of the penalty null space is given its own penalty<sup>3</sup>.

We modify the previous CO<sub>2</sub> model to incorporate group-level smoothers as follows:

$$\log_e(\text{uptake}_i) = f(\log_e(\text{conc}_i)) + f_{\text{Plant\_uo}_i}(\log_e(\text{conc}_i)) + \varepsilon_i$$

where  $f_{\text{Plant\_uo}_i}(\log_e(\text{conc}_i))$  is the smoother for concentration for the given plant. In R we then have:

```
C02_mod2 <- gam(log(uptake) ~ s(log(conc), k=5, m=2) +
                 s(log(conc), Plant_uo, k=5, bs="fs", m=2),
                 data=C02, method="REML")
```

Figure 10 shows the fitted smoothers for C02\_mod2. The plots of group-specific smoothers indicate that plants differ not only in average log-uptake (which would correspond to each plant having a straight line at different levels for the group-level smoother), but differ slightly in the shape of their functional responses. Figure 11 shows how the global and group-specific smoothers combine to predict uptake rates for individual plants. We see that, unlike in the single global smoother case above, none of the curves deviate from the data systematically.

The factor-smoother interaction-based approach mentioned above does not work for higher-dimensional tensor product smoothers. Instead, the group-specific term can be specified with a tensor product of the continuous smoothers and a random effect for the grouping parameter<sup>4</sup>. e.g.: `y ~ te(x1, x2, bs="tp", m=2) + t2(x1, x2, fac, bs=c("tp", "tp", "re"), m=2, full=TRUE)`. We illustrate this approach below on the bird migration data.

<sup>3</sup>As part of the penalty construction, each group will also have its own intercept (part of the penalized null space), so there is no need to add a separate term for group specific intercepts as we did in model 1.

<sup>4</sup>As mentioned in section II, these terms can be specified either with `te()` or `t2()` terms. Using `t2` as above (with `full=TRUE`) is essentially a multivariate equivalent of the factor-smoother interaction; it requires more smooth terms than `te()`, but can be fit using other mixed effects software such as **lme4**, which is useful when fitting models with a large number of group levels (see Section V on computational issues for details).

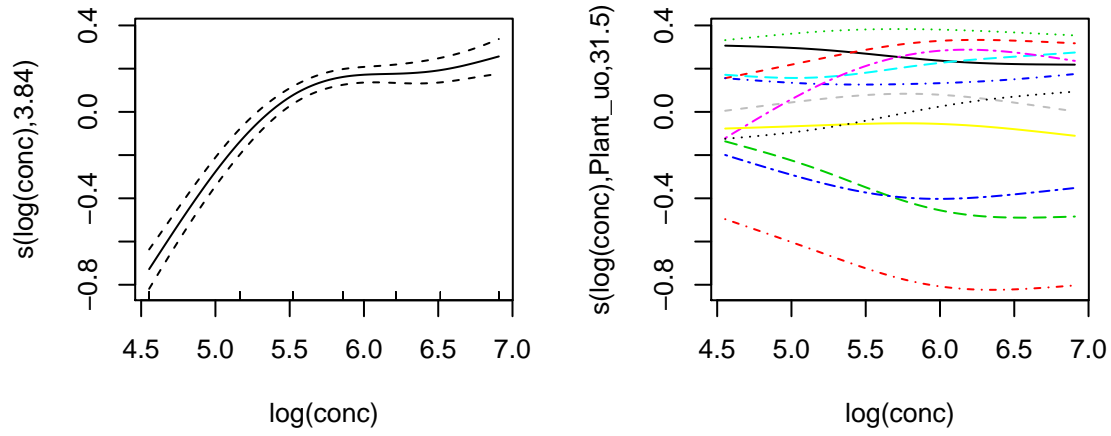


Figure 10: Global function (left) and group-specific deviations from the global function (right) for CO2\_mod2

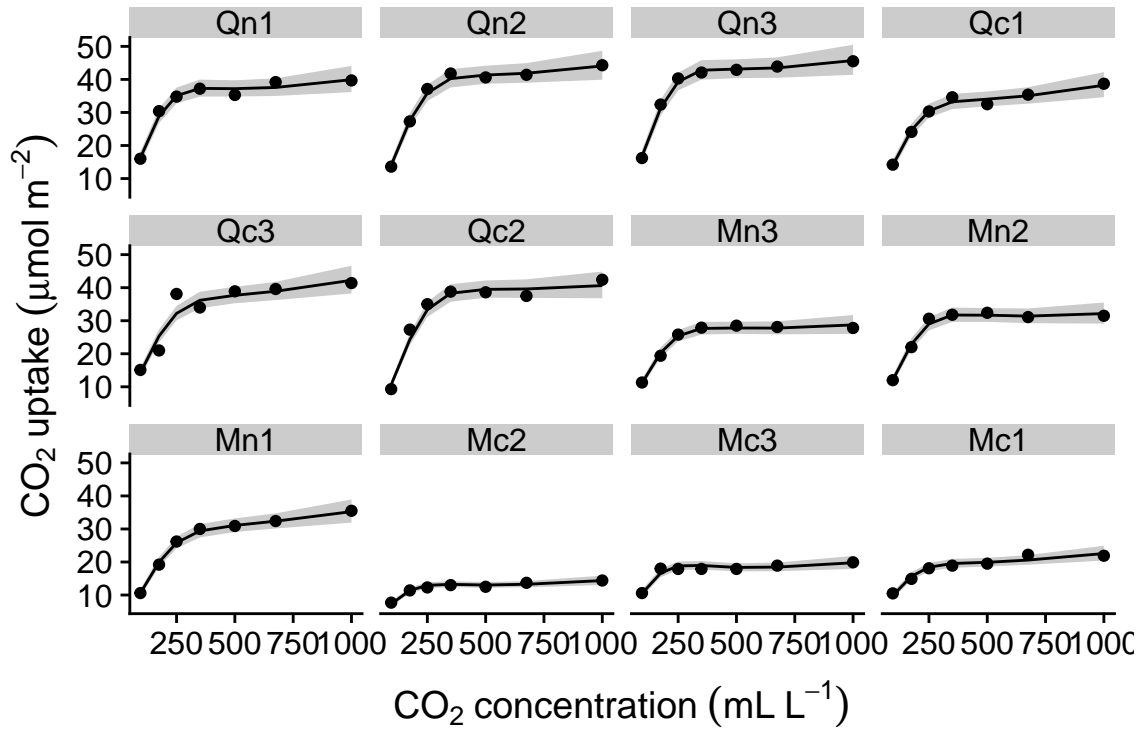


Figure 11: Predicted uptake values (lines) versus observed uptake for each plant, based on model 2.

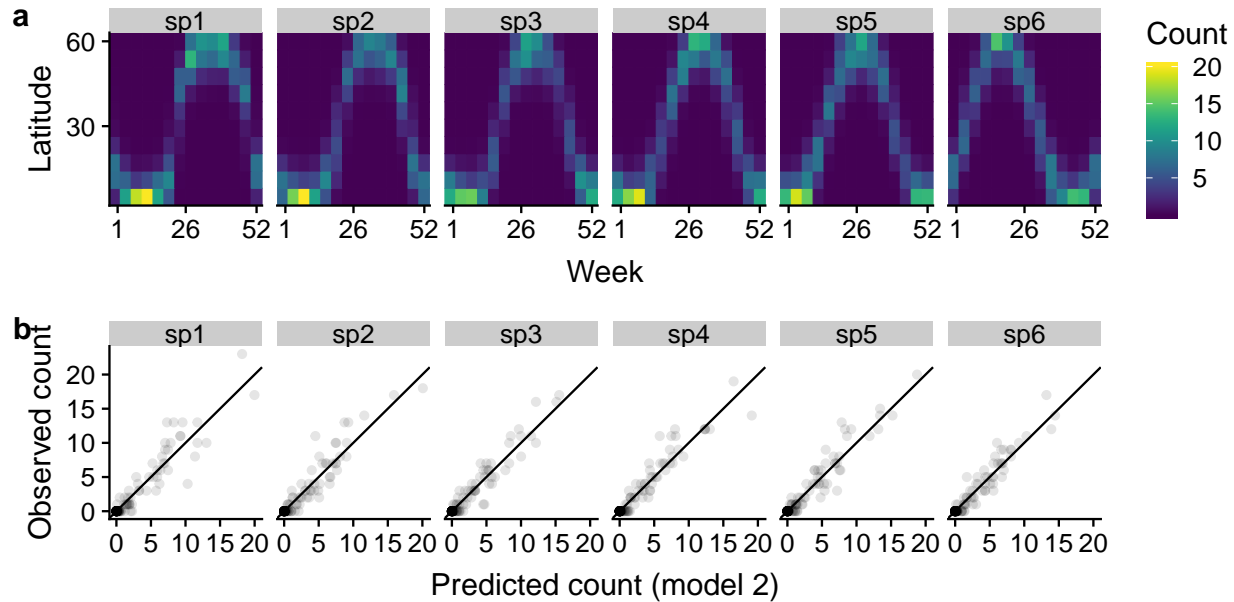


Figure 12: a) Predicted migration paths for each species based on `bird_mod2`, with lighter colors corresponding to higher predicted counts. b) Observed counts versus predictions from `bird_mod2`.

```
bird_mod2 <- gam(count ~ te(week, latitude, bs=c("cc", "tp"),
                        k=c(10, 10), m=2) +
                t2(week, latitude, species, bs=c("cc", "tp", "re"),
                  k=c(10, 10, 6), m=2, full=TRUE),
                data=bird_move, method="REML", family="poisson",
                knots = list(week = c(0, 52)))
```

Model 2 is able to effectively capture the observed patterns of interspecific variation in migration behaviour (figure 12a). It shows a much tighter fit between observed and predicted values, as well as less evidence of over-dispersion in some species compared to model 1 (figure 12b).

### A single common smoother plus group-level smoothers with differing wigginess (Model 3)

This model class is very similar to model 2, but we now allow each group-specific smoother to have its own smoothing parameter and hence its own level of wigginess. This increases the computational cost of the model (as there are more smoothing parameters to estimate), and means that the only information shared between groups is through the global smoother, the common error term, and through the random effect for group-level intercepts (if used). This is useful if different groups differ substantially in how wiggly they are.

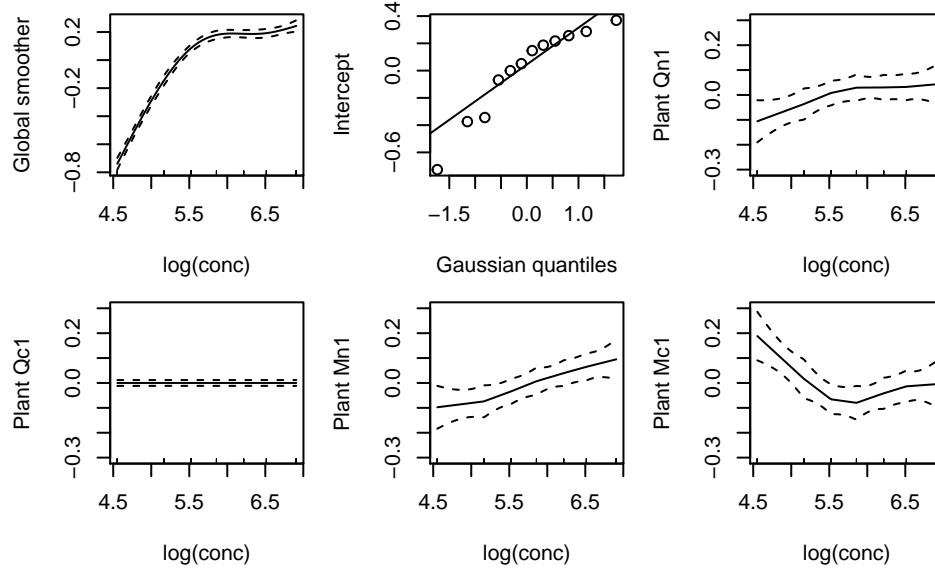


Figure 13: Functional relationships for the CO2 data estimated for model 3. Top left: the global smoother; Top middle: species-specific random effect intercepts. The remaining plots are a selected subset of the plant-specific smoothers, indicating how the functional response of that plant differs from the global smoother.

Fitting a separate smoother (with its own penalties) can be done in **mgcv** by using the **by** argument in the **s()** and **te()** (and related) functions. Therefore, we can code the formula for this model as:

```
y ~ s(x, bs="tp") + s(x, by=fac, m=1, bs="tp") + s(fac, bs="re")
```

Note two major differences here from how model 2 was specified:

1. We explicitly include a random effect for the intercept (the **bs="re"** term), as group-specific intercepts are not incorporated into factor **by** variable smoothers (as would be the case with a factor smoother or a tensor product random effect).
2. We specify **m=1** instead of **m=2** for the groupwise smoothers, which means the marginal TPRS basis for this term will penalize the squared 1st derivative of the function, rather than the second derivative. This, also, reduces co-linearity between the global smoother and the group-specific terms which occasionally leads to high uncertainty around the global smoother (see section V for more details). TPRS with **m=1** have a more restricted null space than **m=2** smoothers, so should not be as collinear with the global smoother (Wieling et al., 2016; Baayen et al., 2018). We have observed that this is much more of an issue when fitting model 3 compared to model 2.

We modify the C02 model to follow this approach like so:

```
C02_mod3 <- gam(log(uptake) ~ s(log(conc), k=5, m=2, bs="tp") +
  s(log(conc), by=Plant_uo, k=5, m=1, bs="tp") +
  s(Plant_uo, bs="re", k=12),
  data=C02, method="REML")
```

Figure 13 shows a subsample of the group-specific smoothers from this model. It is apparent from this that some groups (e.g. `Qc1`) have very similar shapes to the global smoother (differing only in intercept), others do differ from the global trend, with higher uptake at low concentrations and lower uptake at higher concentrations (e.g. `Mc1`, `Qn1`), or the reverse pattern (e.g. `Mn1`).

Using model 3 with higher-dimensional data is also straightforward; by terms work just as well in tensor-product smoothers as they do with isotropic smoothers. We can see this with our bird model:

```
bird_mod3 <- gam(count ~ species +
  te(week, latitude, bs=c("cc", "tp"),
    k=c(10, 10), m=2) +
  te(week, latitude, by=species, bs= c("cc", "tp"),
    k=c(10, 10), m=1),
  data=bird_move, method="REML", family="poisson",
  knots = list(week = c(0, 52)))
```

As above, we have set (`m=1`) for the latitude marginal effect to avoid issues of collinearity between the global and groupwise smoother. Note that switching `m=1` to `m=2` does not have any effect on the marginal basis for `week`, where we are using a cyclic smoother instead of a `tp`s.

The fitted model for `bird_mod3` is visually indistinguishable from `bird_mod2` (figure 12) so we do not illustrate it here.

## Models without global smoothers (models 4 and 5)

We can modify the above models to exclude the global term (which is generally faster; see section V). When we do not model the global term, we are allowing each factor to be different, though there may be some similarities in the shape of the functions.

### Model 4:

Model 4 (shared smoothers) is simply model 2 without the global smoother; this type of model takes the form: `y~s(x, fac, bs="fs")` or `y~te(x1, x2, fac, bs=c("tp", "tp", "re"))` in `mgcv`. This model assumes all groups have the same smoothness, but that the individual shapes of the smooth terms are not related. Here we just show how to code these models; plotting them works in the same way as for models 1-3 above, the plots for these datasets are very similar to the plots for model 2. This will not always be the case; if in a given study there are very few data points in each grouping level (relative to the strength of the functional relationship of interest), estimates from model 4 will typically be much more variable than from model 2, as there is no way for the model to share information on function shape between grouping levels without the global smoother. See section V on computational issues for more on how to choose between different models.

```

C02_mod4 <- gam(log(uptake) ~ s(log(conc), Plant_uo, k=5, bs="fs", m=2),
               data=C02, method="REML")

bird_mod4 <- gam(count ~ t2(week, latitude, species, bs=c("cc", "tp", "re"),
                          k=c(10, 10, 6), m=c(2, 2, 2)),
               data=bird_move, method="REML", family="poisson",
               knots = list(week = c(0, 52)))

```

## Model 5:

Model 5 is simply model 3 without the first term:  $y \sim \text{fac} + s(x, \text{by}=\text{fac})$  or  $y \sim \text{fac} + \text{te}(x_1, x_2, \text{by}=\text{fac})$  (as above, plots are very similar to model 3).

```

C02_mod5 <- gam(log(uptake) ~ s(log(conc), by=Plant_uo, k=5, bs="tp", m=2) +
               s(Plant_uo, bs="re", k=12),
               data= C02, method="REML")

bird_mod5 <- gam(count ~ species +
               te(week, latitude, by=species, bs= c("cc", "tp"),
                 k=c(10, 10), m=2),
               data=bird_move, method="REML", family="poisson",
               knots = list(week = c(0, 52)))

```

## Comparing different HGAM specifications

These models can be compared using standard model comparison tools. Model 2 and model 3 will generally be nested in model 1 (depending on how each model is specified) so comparisons using generalized likelihood ratio tests (GLRTs) may be used to test if groupwise smoothers are necessary. However, we do not currently recommend this method. There is not sufficient theory on how accurate parametric p-values are for comparing these models; there is uncertainty about what degrees of freedom to assign to models with varying smoothness, and slightly different model specifications may not result in nested models. (See Wood (2017a) Section 6.12.4 and `?mgcv::anova.gam` for more discussion on using GLRTs to compare GAMs.)

Comparing models based on AIC is a more robust approach to comparing the different model structures. There is well-developed theory of how to include effects of penalization and smoothing parameter uncertainty when estimating the model complexity penalty for AIC (Wood, Pya & Säfken, 2016). We demonstrate this approach in Table 1. Using AIC, there is strong support for including among-group functional variability for both the CO2 dataset and the `bird_move` dataset (compare models 1 versus models 2-5). For the CO2 dataset (Table 1A), there is relatively strong evidence that there is more inter-group variability in smoothness than model 2 allows, and weaker evidence that model 4 or 5 (separate smoothers for all plants) show the best fit.

Table 1: AIC table comparing model fits for example datasets

Model	df	AIC	deltaAIC
<b>A. CO2 models</b>			
CO2_mod1	17	-119	101
CO2_mod2	39	-199	22
CO2_mod3	42	-216	4
CO2_mod4	53	-219	1
CO2_mod5	56	-220	0
<b>B. bird_move models</b>			
bird_mod1	51	3374	1820
bird_mod2	140	1554	0
bird_mod3	208	1682	128
bird_mod4	130	1555	1
bird_mod5	200	1634	80

For the `bird_move` dataset (Table 1B), model 2 (global smoother plus group-level smoothers with a shared penalty) gives the best fit for all models including a global smooth (which is good as we simulated the data from a model with this structure!). However, model 4 (without a global term) still fits this data better than model 2 based on AIC. This highlights an issue with AIC for selecting between models with and without a global smooth: as it is possible to fully recreate the global term by just allowing each group-level smoother to have a similar shape to one another (that is, the global term is totally concave with the group-level smoothers; see section V) model selection criteria such as AIC may indicate that the extra parameters required to fit the global smoother are unnecessary.

Given this issue with selecting global terms, we strongly recommend not selecting models based purely on AIC. Instead, model selection should be based on expert subject knowledge about the system, computational time, and most importantly, the inferential goals of the study. Table 1A indicates that models 4 and 5 (which do not have a global function) fit the CO2 data better than models with a global function, and that model 4 fits the `bird_move` data better than model 2. However, it is the shape of the global function that we are actually interested in here, as models 4 and 5 cannot be used to predict the concentration-uptake relationship for plants that are not part of the training set, or the average migration path for birds. The same consideration holds when choosing between model 2 and 3: while model 3 fits the CO2 data better than model 2 (as measured by AIC), model 2 can be used to simulate functional variation for unobserved group levels, whereas this is not possible within the framework of model 3. The next section works through two examples to show how to choose between different models, and section V discusses these and other model fitting issues in more depth.

It also is important to recognize that AIC, like any function of the data, is a random variable and should be expected to have some sampling error (Forster & Sober, 2011). In cases when the goal is to select the model that has the best predictive ability, we recommend holding

some fraction of the data out prior to the analysis and comparing how well different models fit that data or using  $k$ -fold cross validation as a more accurate guide to how well a given model may predict out of sample. Predictive accuracy may also be substantially improved by averaging over multiple models (Dormann et al., 2018).

## IV: Examples

We now demonstrate two worked examples on one data set to highlight how to use HGAMs in practice, and to illustrate how to fit, test, and visualize each model. We will demonstrate how to use these models to fit community data, to show when using a global trend may or may not be justified, and to illustrate how to use these models to fit seasonal time series.

For these examples, data are from a long-term study in seasonal dynamics of zooplankton, collected by the Richard Lathrop. The data were collected from a chain of lakes in Wisconsin (Mendota, Monona, Kegonsa, and Waubesa) approximately bi-weekly from 1976 to 1994. They consist of samples of the zooplankton communities, taken from the deepest point of each lake via vertical tow. The data are provided by the Wisconsin Department of Natural Resources and their collection and processing are fully described in Lathrop (2000).

Zooplankton in temperate lakes often undergo seasonal cycles, where the abundance of each species fluctuates up and down across the course of the year, with each species typically showing a distinct pattern of seasonal cycles. The inferential aims of these examples are to (i) estimate variability in seasonality among species in the community in a single lake (Mendota), and (ii) estimate among-lake variability for the most abundant taxon in the sample (*Daphnia mendotae*) across the four lakes. To enable evaluation of out-of-sample performance, we split the data into testing and training sets. As there are multiple years of data, we used data from the even years to fit (train) models, and the odd years to test the fit.

Each record consists of counts of a given zooplankton taxon taken from a subsample from a single vertical net tow, which was then scaled to account for the relative volume of subsample versus the whole net sample and the area of the net tow and rounded to the nearest 1000 to give estimated population density per  $m^2$  for each taxon at each point in time in each sampled lake. As this meant that data were not counts, and observed densities spanned four orders of magnitude, we modelled density using a Gamma distribution with a log-link. For any net tow sample where a given taxon was not observed, we set that taxon's density to 1000 (the minimum possible sample size)<sup>5</sup>.

First, we demonstrate how to model community-level variability in seasonality, by regressing scaled density on day of year, with species-specific curves. As we are not interested here in average seasonal dynamics, we will focus on models 4 and 5 (if we wanted to estimate

---

<sup>5</sup>A more appropriate model for this data would be to assume that density is *left censored*, where 1000 is treated as a threshold which the data may lie below, but it is not possible to measure lower than this. However, **mgcv** does not currently have a left-censored family. The **brms** package, for Bayesian model fitting, can fit a left-censored Gamma distribution, so it would be possible to fit this model using that software. We discuss using HGAMs in **brms** in section V.



the seasonal dynamics for rarer species, adding a global smooth term might be useful, so we could borrow information from the more common species). As the data are seasonal, we use cyclic smoothers as the basis for seasonal dynamics. Therefore we need to specify start and end points for our cycles using the `knots` argument to `gam`, as well as specify that this is smoother type to the factor-smooth interaction term using the `xt` argument (the `xt` argument is how any extra information that a smoother might need is supplied; see `?mgcv:s` for more information). Note that we also include a random effect smoother for both `taxon` and `taxon:year_f`, where `year_f` is just `year` transformed into a factor variable, to deal with the fact that average zooplankton densities can show large year-to-year variation. The argument `drop.unused.levels=FALSE` is also included so the `gam` function does not drop the year factor levels corresponding to those in the held-out test data set.

#### Model 4:

```
zoo_comm_mod4 <- gam(density_adj ~ s(day, taxon,
                                   bs="fs",
                                   k=10,
                                   xt=list(bs="cc"))+
                   s(taxon, year_f, bs="re"),
  data=zoo_train,
  knots = list(day =c(0, 365)),
  family = Gamma(link = "log"),
  method = "REML",
  drop.unused.levels = FALSE)
```

#### Model 5:

```
# Note that s(taxon, bs="re") has to be explicitly included here, as the
# day by taxon smoother does not include an intercept
zoo_comm_mod5 <- gam(density_adj ~ s(day, by=taxon,
                                   k=10, bs="cc") +
                   s(taxon, bs="re") +
                   s(taxon, year_f, bs="re"),
  data=zoo_train,
  knots = list(day =c(0, 365)),
  family = Gamma(link = "log"),
  method = "REML",
  drop.unused.levels = FALSE)
```

At this stage of the analysis (prior to model comparisons), it is useful to determine if any of the fitted models adequately describe patterns in the data (i.e. goodness of fit testing). The `mgcv` package provides tools to facilitate this process, using the `gam.check` function. This

Table 2: Results from running `k.check` on `zoo_comm_mod5`.

	<code>k'</code>	<code>edf</code>	<code>k-index</code>	<code>p-value</code>
<code>s(day):taxonC. sphaericus</code>	8	4.78	0.89	0.43
<code>s(day):taxonCalanoid copepods</code>	8	6.66	0.89	0.47
<code>s(day):taxonCyclopoid copepods</code>	8	5.31	0.89	0.45
<code>s(day):taxonD. mendotae</code>	8	6.95	0.89	0.48
<code>s(day):taxonD. thomasi</code>	8	6.57	0.89	0.46
<code>s(day):taxonK. cochlearis</code>	8	5.92	0.89	0.49
<code>s(day):taxonL. siciloides</code>	8	0.52	0.89	0.48
<code>s(day):taxonM. edax</code>	8	4.69	0.89	0.46
<code>s(taxon)</code>	8	6.26	NA	NA
<code>s(taxon,year_f)</code>	152	51.73	NA	NA

function creates a set of standard diagnostic plots: a QQ plot, a plot of response versus fitted values, a histogram of residuals, and a plot of residuals versus fitted values. It also conducts a test for each smooth term to determine if the number of degrees of freedom (`k`) for each smooth is adequate (see `?mgcv::gam.check` for details on how that test works). The code for checking model 4 and 5 for the community zooplankton model is given as:

```
gam.check(zoo_comm_mod4)
gam.check(zoo_comm_mod5)
```

For the sake of brevity, and as fitted versus response plots are generally less useful for non-normally distributed data (as it can be difficult to visually assess if the observed data shows more heteroskedasticity than expected for a specified distribution) we have plotted QQ plots and fitted-versus residual plots for model 5; the results for model 4 are virtually indistinguishable to the naked eye. These plots (Figure 14) indicate that the Gamma distribution seems to fit the observed data well except at low values, where the deviance residuals are larger than predicted by the theoretical quantiles (Figure 14 left). There also does not seem to be a pattern in the residual versus fitted values (Figure 14 right), except for a line of residuals at the lowest values, which correspond to all of those observations where a given taxon was absent from the sample. The `k.check` test shows that the default maximum degrees of freedom for the smoothers used in model 5 are sufficient for all species, as the effective degrees of freedom (`edf`) for all estimated smoothers are well below their maximum possible value (`k'`), and the `p-value` for the observed `k-index` (which measures pattern in the residuals) is not significant:

In this table, each row corresponds to a single smooth term, `k'` corresponds to the number of basis functions used for that smoother in the fitted model (smaller than the specified `k` in the model itself, as some basis functions are automatically dropped to ensure the model is identifiable). The column `edf` is the estimated Effective Degrees of Freedom for that smoother, the `k-index` is a measure of the remaining pattern in the residuals, and the `p-value` is calculated based on the distribution of the `k-index` after randomizing the order of the

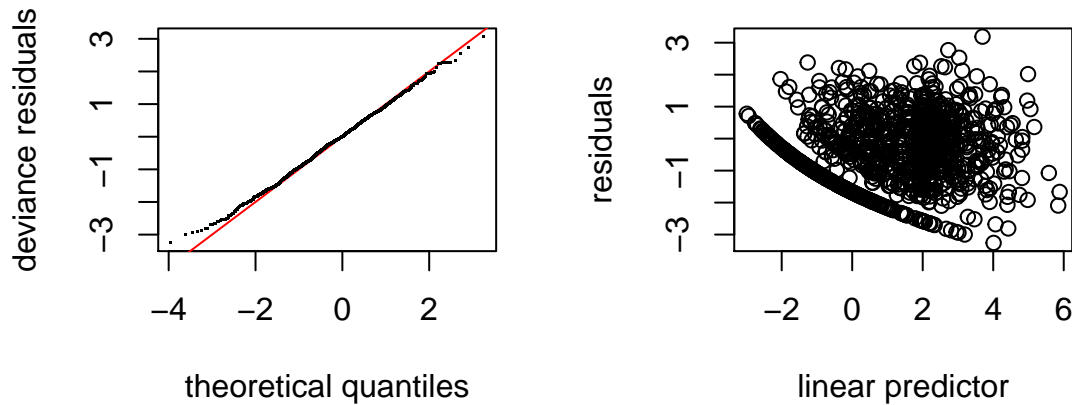


Figure 14: Diagnostic plots for model 5 fitted to zooplankton community data in Lake Mendota. Top row: QQ plots plus the 1-1 line. Bottom row: deviance residuals versus fitted values (on the link scale).

residuals. Note that there is no p-value for the random effects smoothers `s(taxon)` and `s(taxon, year_f)` as the p-value is calculated from simulation-based tests for autocorrelation of the residuals. As `taxon` and `year_f` are treated as simple random effects with no natural ordering, there is no meaningful way of checking for autocorrelation.

Differences between models 4 (shared smoothness between taxa) and 5 (different smoothness for each taxa) seem to be driven by the low seasonality of *L. siciloides* relative to the other species, and how this is captured by the more flexible model 5 (Figure 15). Still, both models show very similar fits to the training data. Model 4 is slightly better at predicting out of sample fits for all taxa except *M. edax* (Table 3). Even though *L. siciloides* showed weak seasonal dynamics, Model 4 and 5 still did a better job of predicting the density of this species out of sample compared to a simple model with only a species-specific intercept (Table 3).

Next, we look at how to fit inter-lake variability in dynamics for just *Daphnia mendotae*. Here, we will compare models 1, 2, and 3 to determine if a single global function is appropriate for all four lakes, or if we can more effectively model variation between lakes with a shared smoother and lake-specific smoothers.

### Model 1:

```
zoo_daph_mod1 <- gam(density_adj ~ s(day, bs="cc", k=10)+
  s(lake, bs="re") +
  s(lake, year_f, bs="re"),
  data=daphnia_train,
```

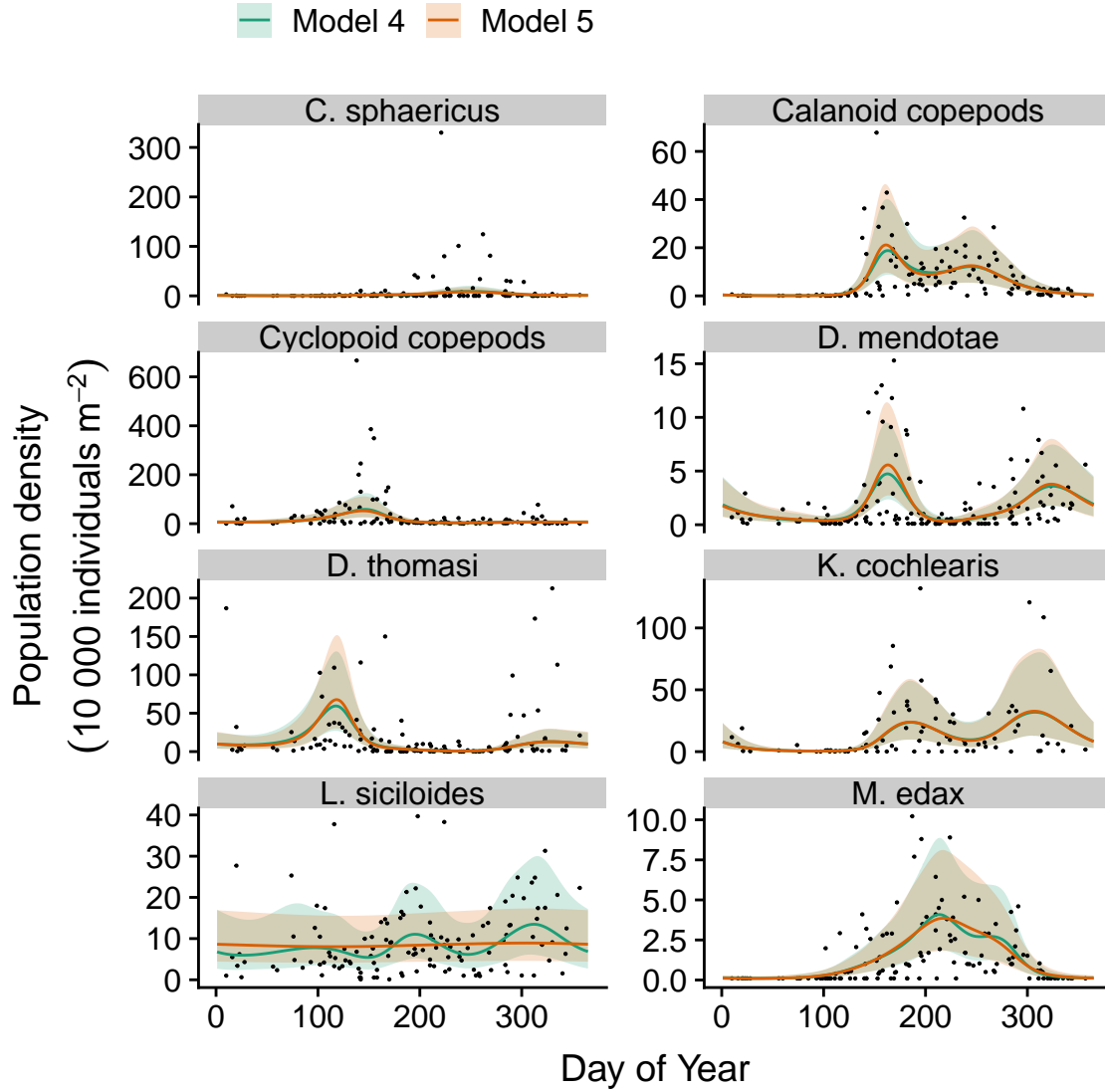


Figure 15: Species-specific seasonal dynamics for the eight zooplankton species tracked in Lake Mendota. Black points indicate individual plankton observations. Lines indicate predicted average values for model 4 (black) and model 5 (red). Ribbons indicate  $\pm 2$  standard errors around the mean.

Table 3: Out-of-sample predictive ability for model 4 and 5 applied to the zooplankton community dataset. RMSE values represent the square root of the average squared difference between model predictions and observations for test data. Intercept only results are for a null model with only year and year-by taxon random effect intercepts included.

taxon	Total RMSE of held out data (10 000 individuals·m <sup>-2</sup> )		
	Intercept only	Model 4	Model 5
<i>C. sphaericus</i>	20.9	20	20.1
Calanoid copepods	12.1	10.9	10.9
Cyclopoid copepods	109	104	105
<i>D. mendotae</i>	3.03	2.79	2.82
<i>D. thomasi</i>	29.5	30.8	31.8
<i>K. cochlearis</i>	94.8	95.7	95.6
<i>L. siciloides</i>	6.12	5.5	5.86
<i>M. edax</i>	2.27	1.95	1.94

```
knots=list(day =c(0, 365)),
family=Gamma(link ="log"),
method="REML",
drop.unused.levels = FALSE)
```

## 554 Model 2:

```
zoo_daph_mod2 <-
  gam(density_adj ~ s(day, bs="cc", k=10) +
    s(day, lake, k=10, bs="fs", xt=list(bs="cc")) +
    s(lake, year_f, bs="re"),
    data=daphnia_train,
    knots=list(day=c(0, 365)),
    family=Gamma(link ="log"),
    drop.unused.levels = FALSE,
    method="REML")
```

## 555 Model 3:

```
zoo_daph_mod3 <- gam(density_adj~s(day, bs="cc", k=10) +
  s(day, by=lake, k=10, bs="cc")+
  s(lake, bs="re") +
```

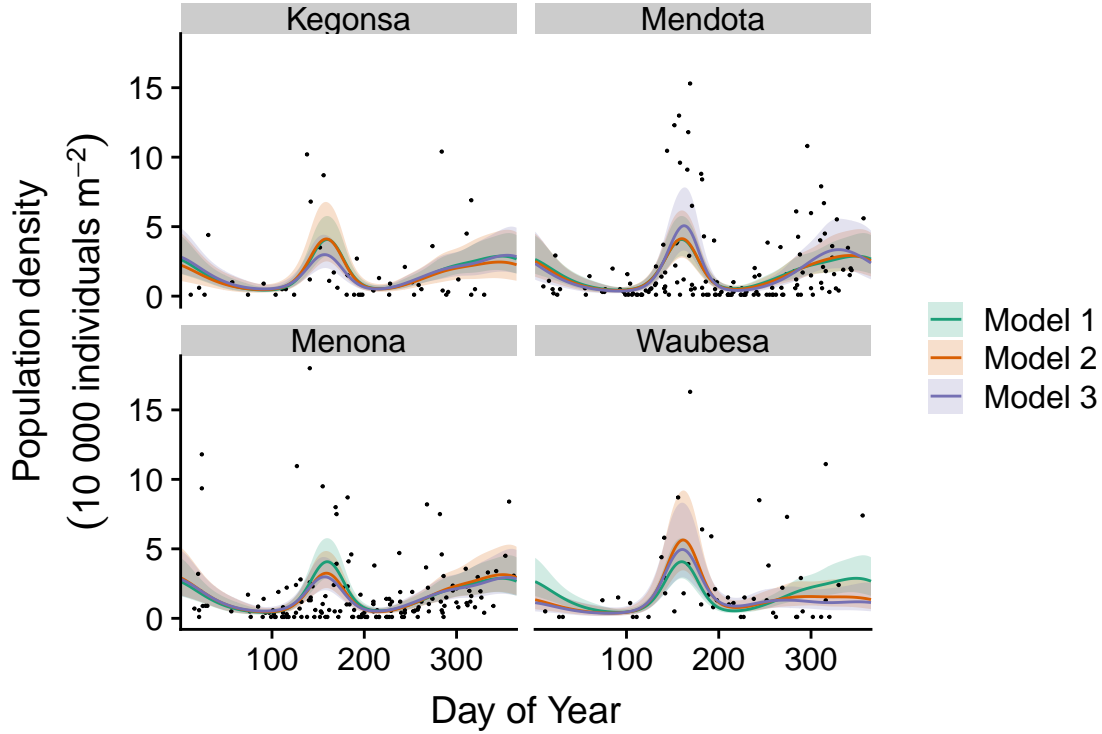


Figure 16: Raw data (points) and fitted models (lines) for *D. mendota* data. Green: model 1 (no inter-lake variation in dynamics); orange: model 2 (interlake variation with similar smoothness); purple: model 3 (varying smoothness among lakes). Shaded bands are drawn at  $\pm 2$  standard errors around each model.

```
s(lake, year_f, bs="re"),
data=daphnia_train,
knots=list(day =c(0, 365)),
family=Gamma(link = "log"),
method="REML",
drop.unused.levels = FALSE)
```

We will exclude the `gam.check` diagnostic plots and results, as they do not indicate any issues with model fit. The AIC values indicate that both model 2 (1093.71) and 3 (1085.7) are better fits than model 1 (1097.62), but models 3 fits somewhat better than model 2.<sup>6</sup> There does not seem to be a large amount of inter-lake variability (the effective degrees of freedom per lake are low in models 2 & 3). Plots for all three models (Figure 16) show that Mendota, Monona, and Kegonsa lakes are very close to the average and to one another for both models, but Waubesa shows evidence of a more pronounced spring bloom and lower winter abundances.

Model 3 is able to predict as well as model 1 or 2 (Table 4), indicating that allowing for

<sup>6</sup>When comparing models via AIC, we use the standard rule of thumb from Burnham & Anderson (1998), where models that differ by 2 units or less from the lowest AIC model have substantial support, and those differing by more than 4 units having less support.

Table 4: Out-of-sample predictive ability for model 1-3 applied to the *D. mendotae* dataset. RMSE values represent the average squared difference between model predictions and observations for held-out data.

Lake	Total RMSE of held out data (10 000 individuals·m <sup>-2</sup> )			
	Intercept only	Model 1	Model 2	Model 3
Kegonsa	2.1	2.1	2.1	2
Mendota	3	2.7	2.7	2.7
Menona	3	3	2.9	2.9
Waubesa	7	7.2	7.2	7.1

inter-lake variation in seasonal dynamics improved model prediction. None of the models did well in terms of predicting Lake Kegonsa or Lake Waubesa dynamics out of sample compared to a simple model with only a lake-specific intercept and no intra-annual variability (Table 4).

## V: Computational and statistical issues when fitting HGAMs

Which of the five model formulations should you choose for a given data set? There are two major trade-offs to consider. The first is the bias-variance trade-off: more complex models can account for more fluctuations in the data, but also tend to give more variable predictions, and can overfit. The second trade-off is model complexity versus computational cost: more complex models can include more potential sources of variation and give more information about a given data set, but will generally take more time and computational resources to fit and debug. We discuss both of these trade-offs in this section. We also discuss how to extend the HGAM framework to fit more complex models.

### Bias-variance trade-offs

The bias-variance trade-off is a fundamental concept in statistics. When trying to estimate any relationship (in the case of GAMs, a smooth relationship between predictors and data) bias measures how far, on average, an estimate is from the true value. The variance of an estimator corresponds to how much that estimator would fluctuate if applied to multiple different samples of the same size taken from the same population. These two properties tend to be traded off when fitting models. For instance, rather than estimating a population mean from data, we could simply use a predetermined fixed value regardless of the observed data<sup>7</sup>. This estimate would have no variance (as it is always the same regardless of what the data

<sup>7</sup>While this example may seem contrived, this is exactly what happens when we assume a given regression coefficient is equal to zero (and thus exclude it from a model).

look like) but would have high bias unless the true population mean happened to equal the fixed value we chose. Penalization is useful because using a penalty term slightly increases model bias, but can substantially decrease variance (Efron & Morris, 1977).

In GAMs, the bias-variance trade-off is managed by the terms of the penalty matrix, and equivalently random effect variances in HGLMs. Larger penalties correspond to lower variance, as the estimated function is unable to wiggle a great deal, but also correspond to higher bias unless the true function is close to the null space for a given smoother (e.g., a straight line for thin plate splines with 2nd derivative penalties, or zero for a random effect). The computational machinery used by **mgcv** to fit smooth terms is designed to find penalty terms that best trade-off bias for variance to find a smoother that can effectively predict new data.

The bias-variance trade-off comes into play with HGAMs when choosing whether to fit separate penalties for each group level or assign a common penalty for all group levels (i.e., deciding between models 2 & 3 or models 4 & 5). If the functional relationships we are trying to estimate for different group levels actually vary in how wiggly they are, setting the penalty for all group-level smoothers equal (models 2 & 4) will either lead to overly variable estimates for the least variable group levels, over-smoothed (biased) estimates for the most wiggly terms, or a mixture of these two, depending on the fitting criteria.

We developed a simple numerical experiment to determine whether **mgcv**'s fitting criteria tend to set estimated smoothness penalties high or low in the presence of among-group variability in smoothness when fitting model 2 or 4 HGAMs. We simulated data from five different groups, with all groups having the same levels of the covariate  $x$ , ranging from 0 to  $2\pi$ . For each group, the true function relating  $x$  to the response,  $y$ , was a sine wave, but the frequency varied from 0.25 (equal to half a cycle across the range of  $x$ ) to 4 (corresponding to 4 full cycles across the range). We added normally distributed error to all  $y$ -values, with a standard deviation of 0.2. We then fit both model 4 (where all curves were assumed to be equally smooth) and model 5 (with varying smoothness) to the entire data set, using REML criteria to estimate penalties. For this example (Fig. 17a), requiring equal smoothness for all group levels resulted in **mgcv** underestimating the penalty for the lowest frequency (most smooth) terms, but accurately estimating the true smoothness of the highest frequency terms as measured by the squared second derivative of the smooth fit versus that of the true function (Fig. 17b). This implies that assuming equal smoothness will result in underestimating the true smoothness of low-variability terms, and thus lead to more variable estimates of these terms. If this is a potential issue, we recommend fitting both models 4 and 5 and using standard model evaluation criteria (e.g., AIC) to determine if there is evidence for among-group variability in smoothness. For instance, the AIC for model 4 fit to this data is -178, whereas it is -211 for model 5, implying a improvement in fit by allowing smoothness to vary. However, it may be the case that there are too few data points per group to estimate separate smoothness levels, in which case model 2 or model 4 may still be the better option even in the face of varying smoothness.

The ideal case would be to assume that among-group penalties follow their own distribution (estimated from the data), to allow variation in smoothness while still getting the benefit of pooling information on smoothness between groups. This is currently not implemented in **mgcv**. It is possible to set up this type of varying penalty model in flexible Bayesian



modelling software such as *Stan* (see below for a discussion of how to fit HGAMs using these tools), but how to fit this type of model has not been well studied.

It may seem there is also a bias-variance trade-off between choosing to use a single global smoother (model 1) or a global smoother plus group-level terms (models 2 and 3). In model 1, all the data is used to estimate a single smooth term, and thus should have lower variance than models 2 and 3, but higher bias for any given group in the presence of inter-group functional variability. However, in practice, this trade-off will be handled via penalization; if there are no average differences between functional responses, **mgcv** will penalize the group-specific functions toward zero, and thus toward the global model. The choice between using model 1 versus models 2 and 3 should generally be driven by computational costs. Model 1 is typically much faster to fit than models 2 and 3, even in the absence of among-group differences. If there is no need to estimate inter-group variability, model 1 will typically be more efficient.

A similar issue exists when choosing between models 2 and 3 and models 4 and 5. If all group levels have very different functional shapes, the global term will get penalized toward zero in models 2 and 3, so they will reduce to models 4 and 5. The choice to include a global term should be made based on scientific considerations (is the global term of interest?) and computational considerations.

## Complexity-computation trade-offs

The more flexible a model is, the larger an effective parameter space any fitting software has to search. It can be surprisingly easy to use massive computational resources trying to fit models to even small datasets. While we typically want to select models based on their fit and our inferential goals, computing resources can often act as an effective upper bound on model complexity. For a given data set, assuming a fixed family and link function, the time taken to estimate an HGAM will depend (roughly) on four factors: (i) the number of coefficients to be estimated (and thus the number of basis functions chosen), (ii) the number of smoothing parameters to be estimated, (iii) whether the model needs to estimate both a global smoother and groupwise smoothers, and (iv) the algorithm and fitting criteria used to estimate parameters.

The most straightforward factor that will affect the amount of computational resources is the number of parameters in the model. Adding group-level smoothers (moving from model 1 to 2-5) means that there will be more regression parameters to estimate. For a dataset with  $g$  different groups and  $n$  data points, fitting a model with just a global smoother,  $y \sim s(x, k=k)$  will require  $k$  coefficients, and takes  $\mathcal{O}(nk^2)$  operations to evaluate. Fitting the same data using a group-level smoother (model 4,  $y \sim s(x, fac, bs="fs", k=k)$ ) will require  $\mathcal{O}(nk^2g^2)$  operations to evaluate. In effect, adding a group-level smoother will increase computational cost by an order of the number of groups squared. The effect of this is visible in the examples we fit in section III. Table 5 compares the relative time it takes to compute model 1 versus the other models.

One way to deal with this issue would be to reduce the number of basis functions used when fitting group-level smoothers when the number of groups is large, limiting the flexibility of

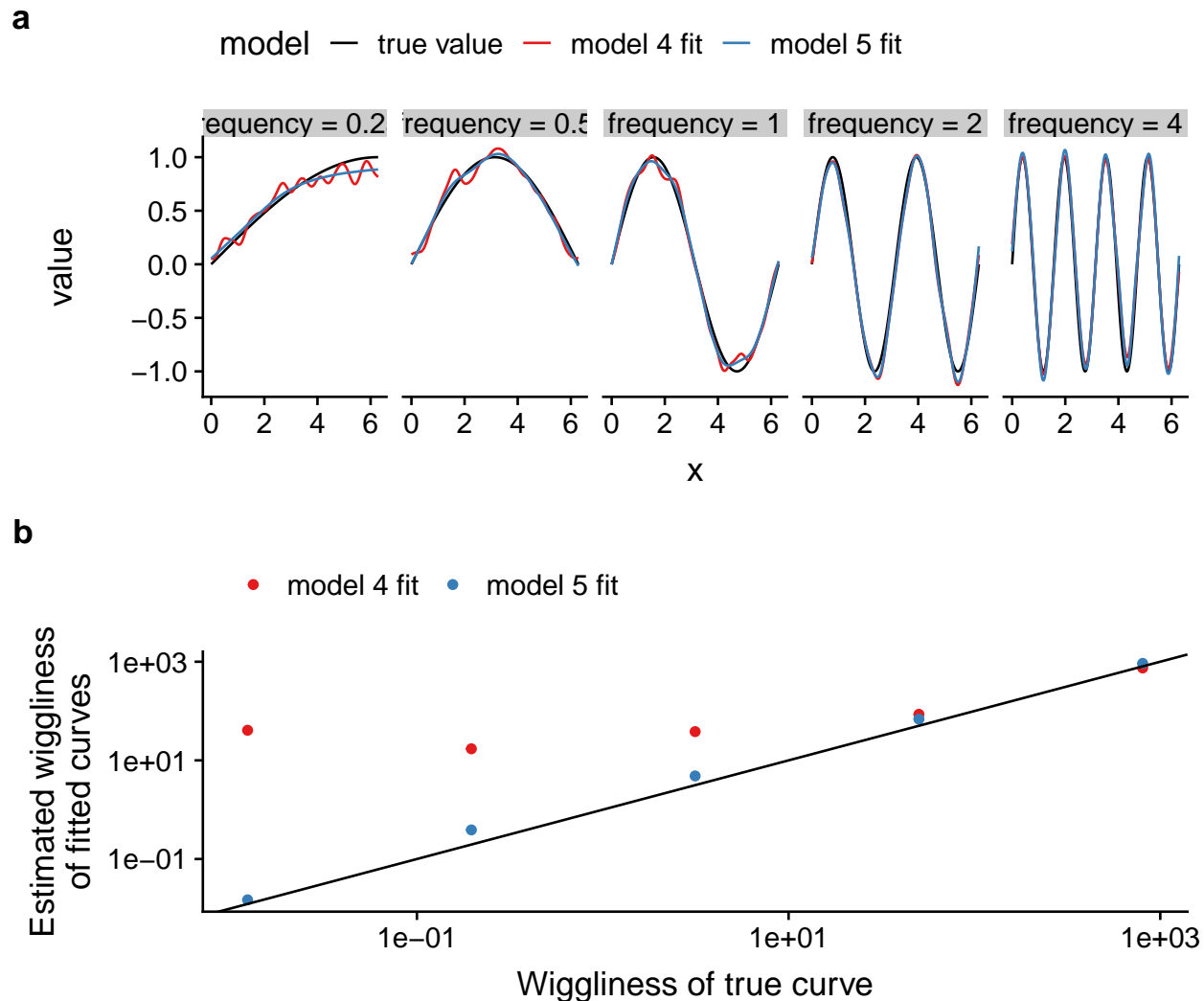


Figure 17: a) Illustration of bias that can arise from assuming equal smoothness for all group levels (model 4, red line) versus allowing for intergroup variation in smoothness (model 5, red line) when the true function (black line) shows substantial variation in smoothness between groups. b) Estimated wigginess (as measured by the integral of the squared 2nd derivative) of the true function for each group level versus that for the functions estimated by model 4 (red) and model 5 (blue), indicating substantial undersmoothing for low-variability curves by model 4.

Table 5: Relative computational time and model complexity for different HGAM formulations of the two example data sets from section III. All times are scaled relative to the length of time model 1 takes to fit to that data set. The number of coefficients measures the total number of model parameters (including intercepts). The number of smoothers is the total number of unique penalty values estimated by the model.

model	relative time	# of terms	
		coefficients	penalties
<b>A. CO2 data</b>			
1	1	17	2
2	7	65	3
3	15	65	14
4	6	61	3
5	17	61	13
<b>B. bird movement data</b>			
1	1	90	2
2	82	540	5
3	180	624	14
4	66	541	3
5	35	535	12

the model. It can also make sense to use more computationally-efficient basis functions when fitting large data sets, such as P-splines (Wood, 2017b) or cubic splines. Thin plate splines entail greater computational costs (Wood, 2017a).

Including a global smoother (models 2 and 3 compared to models 4 and 5) will not generally substantially affect the number of coefficients that need to be estimated (Table 5). Adding a global term will add at most  $k$  extra terms. It can be substantially less than that, as `mgcv` drops basis functions from co-linear smoothers to ensure that the model matrix is full rank.

Adding additional smoothing parameters (moving from model 2 to 3, or moving from model 4 to 5) is more costly than increasing the number of coefficients to estimate, as estimating smoothing parameters is computationally intensive (Wood, 2011). This means that models 2 and 4 will generally be substantially faster than 3 and 5 when the number of groups is large, as models 3 and 5 fit a separate set of penalties for each group level. The effect of this is visible in comparing the time it takes to fit model 2 to model 3 (which has a smoother for each group) or models 4 and 5 for the example data (Table 5). Note that this will not hold in all cases. For instance, model 5 takes less time to fit the bird movement data than model 4 does (Table 5B).

## Alternative formulations: `bam()`, `gamm()`, and `gamm4()`

When fitting models with large numbers of groups, it is often possible to speed up computation substantially by using one of the alternative fitting routines available through `mgcv`.

The first option is the function `bam()`, this requires the least changes to existing code written using the `gam()` function. `bam()` is designed to improve performance when fitting large data sets via two mechanisms. First, it saves on memory needed to compute a given model by using a random subset of the data to calculate the basis functions. It then blocks the data and updates model fit within each block (Wood, Goude & Shaw, 2015). While this is primarily designed to reduce memory usage, it can also substantially reduce computation time. Second, when using `bam()`'s default fREML ("Fast REML") method, you can use the `discrete=TRUE` option: this first bins continuous covariates into a smaller number of discrete values before estimating the model, substantially reducing the amount of computation needed (Wood et al. (2017); see `?mgcv:bam` for more details). Setting up models 1-5 in `bam()` uses the same code as we have previously covered; the only difference is that you use the `bam()` instead of `gam()` function, and have the additional option of discretizing your covariates.

`bam()` has a larger computational overhead than `gam()`, so for small numbers of groups, it can be slower than `gam()` (Figure 18). As the number of groups increases, computational time for `bam()` increases more slowly than for `gam()`; in our simulation tests, when the number of groups is greater than 16, `bam()` can be upward of an order of magnitude faster (Figure 18). Note that `bam()` can be somewhat less computationally stable when estimating these models (i.e., less likely to converge).

The second option is to fit models using one of two dedicated mixed effect model estimation packages, `nlme` and `lme4`. The `mgcv` package includes the function `gamm()`, which uses the `nlme` package to estimate the GAM, automatically handling the transformation of smooth terms into random effects (and back into basis function representations for plotting and other statistical analyses). The `gamm4()` function, in the separate `gamm4` package, uses `lme4` in a similar way. Using `gamm()` or `gamm4()` to fit models rather than `gam()` can substantially speed up computation when the number of groups is large, as both `nlme` and `lme4` take advantage of the sparse structure of the random effects, where most basis functions will be zero for most groups (i.e., any group-specific basis function will only take a non-zero value for observations in that group level). As with `bam()`, `gamm()` and `gamm4()` are generally slower than `gam()` for fitting HGAMs when the number of group levels is small (in our simulations, <8 group levels), however they do show substantial speed improvements even with a moderate number of groups, and were as fast as or faster to calculate than `bam()` for all numbers of grouping levels we tested (Figure 18)<sup>8</sup>.

---

<sup>8</sup>It is also possible to speed up both `gam()` and `bam()` by using multiple processors in parallel, whereas this is not currently possible for `gamm()` and `gamm4()`. For large numbers of grouping levels, this should speed up computation as well, at the cost of using more memory. However, computation time will likely not decline linearly with the number of cores used, since not all model fitting sets are parallelizable, and performance of cores can vary. As parallel processing can be complicated and dependent on the type of computer you are using to configure, we do not go into how to use these methods here. The help file `?mgcv:mgcv.parallel` explains how to use parallel computations for `gam()` and `bam()` in detail.

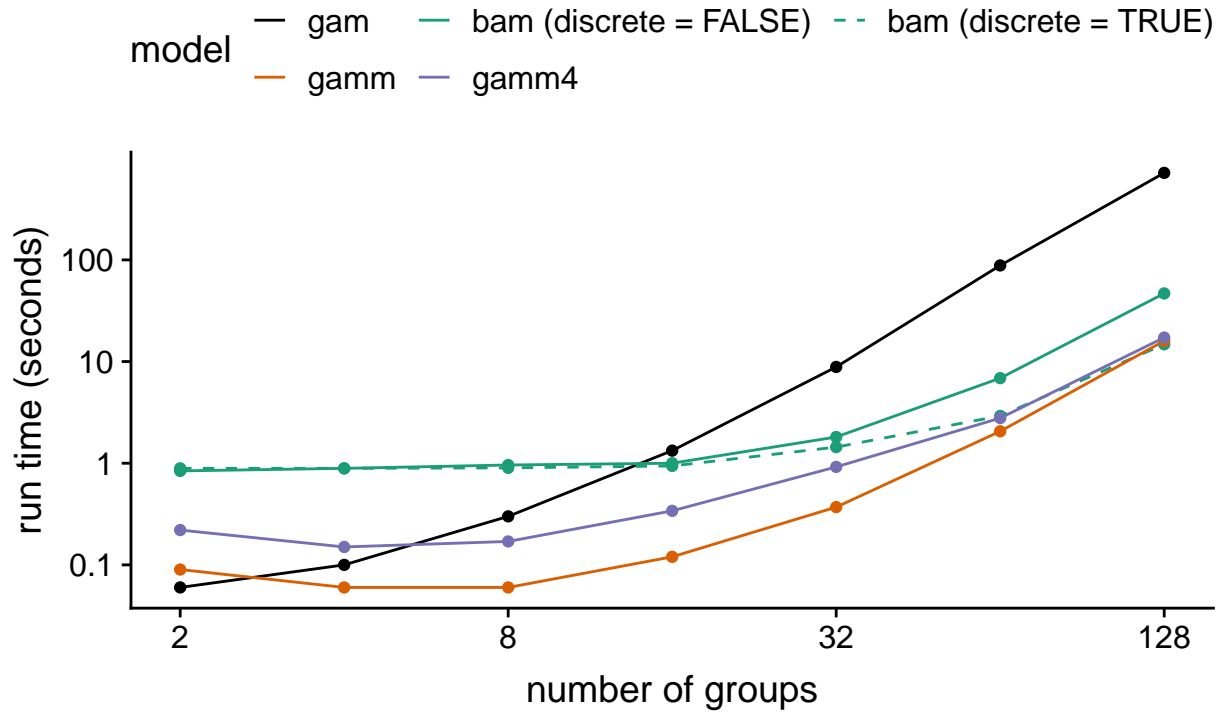


Figure 18: Elapsed time to estimate the same model using each of the four approaches. Each data set was generated with 20 observations per group using a unimodal global function and random group-specific functions consisting of an intercept, a quadratic term, and logistic trend for each group. Observation error was normally distributed. Models were fit using model 2:  $y \sim s(x, k=10, bs="cp") + s(x, fac, k=10, bs="fs", xt=list(bs="cp"), m=1)$ . All models were run on a single core.

Both `gamm()` and `gamm4()` require a few changes to model code. First, there are a few limitations on how you are able to specify models 1-5 in both frameworks. Factor-smoother interaction (`bs="fs"`) basis setup works in both `gamm()` and `gamm4()`. However, as the `nlme` package does not support crossed random effects, it is not possible to have two factor-smoother interaction terms for the same grouping variable in `gamm()` models (e.g., `y~s(x1, grp, bs="fs")+s(x2, grp, bs="fs")`). These type of crossed random effects are allowed in `gamm4`. The use of `te()` terms are not possible in `gamm4`, due to issues with how random effects are specified in the `lme4` package, making it impossible to code models where multiple penalties apply to a single basis function. Instead, for multidimensional group-level smoothers, the alternate function `t2()` needs to be used to generate these terms, as it creates tensor products with only a single penalty for each basis function (see `?mgcv::t2` for details on these smoothers, and Wood, Scheipl & Faraway (2013) for the theoretical basis behind this type of tensor product). For instance, model 2 for the bird movement data we discussed in section III would need to be coded as:

```
bird_mod4_gamm4 <-
  gamm4(count ~ t2(week, latitude, species, bs=c("cc", "tp", "re"),
                 k=c(10, 10, 6), m=2),
        data=bird_move, family="poisson")
```

These packages also do not support the same range of families for the dependent variable; `gamm()` only supports non-Gaussian families by using a fitting method called penalized quasi-likelihood (PQL) that is slower and not as numerically stable as the methods used in `gam()`, `bam()`, and `gamm4()`. Non-Gaussian families are well supported by `lme4` (and thus `gamm4`), but can only fit them using marginal likelihood (ML) rather than REML, so may tend to over-smooth relative to `gam()` using REML estimation. Further, neither `gamm()` nor `gamm4()` supports several of the extended families available through `mgcv`, such as zero-inflated, negative binomial, or ordered categorical and multinomial distributions.

## Estimation issues when fitting both global and groupwise smoothers

When fitting models with separate global and groupwise smoothers (models 2 and 3), one issue to be aware of is concurvity between the global smoother and groupwise terms. Concurvity measures how well one smooth term can be approximated by some combination of the other smooth terms in the model (see `?mgcv::concurvity` for details). For models 2 and 3, the global term is entirely concurved with the groupwise smoothers. This is because, in the absence of the global smooth term, it would be possible to recreate that average effect by shifting all the groupwise smoothers so they were centered around the global mean. In practical terms, this has the consequence of increasing uncertainty around the global mean relative to a model with only a global smoother. In some cases, it can result in the estimated global smoother being close to flat, even in simulated examples with a known strong global effect. This concurvity issue may also increase the time it takes to fit these models (for example, compare the time it takes to fit models 3 and 5 in Table 5). These models can still be estimated because of penalty terms; all of the methods we have discussed for fitting model 2 (factor-smoother terms or random effect tensor products) automatically create a penalty for

the null space of the group-level terms, so that only the global term has its own unpenalized null space. Both the REML and ML criteria work to balance penalties between nested smooth terms (this is why nested random effects can be fitted). We have observed that **mgcv** still occasionally finds solutions with simulated data where the global term is over-smoothed.

To avoid this issue, we recommend both careful choice of basis and setting model degrees of freedom so that groupwise terms are either slightly less flexible than the global term or have a smaller null space. In the examples in section III, we used smoothers with an unpenalized null space (standard thin plate splines) for the global smoother and ones with no null space for the groupwise terms<sup>9</sup>. When using thin plate splines, it may also help to use splines with a lower order of derivative penalized in the groupwise smoothers than the global smoothers, as lower-order “tp” splines have fewer basis functions in the null space. For example, we used **m=2** (penalizing squared second derivatives) for the global smoother, and **m=1** (penalizing squared first derivatives) for groupwise smoothers in models 2 and 3. Another option is to use a lower number of basis functions (**k**) for groupwise relative to global terms. This will reduce the maximum flexibility possible in the groupwise terms. We do caution that these are just rules of thumb. In cases where an accurately estimated global smoother is essential, we recommend either fitting model 1 or using specialized functional regression software such as the **fors** function in the *refund* package (Scheipl, Staicu & Greven, 2014), which allows the user to enforce constraints on the groupwise smoothers so that they always sum to zero at any given point (avoiding the collinearity issue). Also, see below for more information on functional regression.

## A brief foray into the land of Bayes

As mentioned in section II, the penalty matrix can also be treated as the inverse of a prior covariance matrix for model parameters  $\beta$ . Intuitively, the basis functions and penalty we use form a prior (in the informal sense) on how we’d like our model term to behave. REML gives an empirical Bayes estimate of the smooth model (Laird & Ware, 1982), where terms in the null space of the smoother have improper, flat priors (i.e., any value for these terms are considered equally likely), any terms in the range space are treated as having a multivariate normal distribution, and the penalty terms are treated as having an improper flat prior (see Wood (2017a) Section 5.8 for more details on this connection). The large-sample approximation of the posterior Bayesian covariance matrix (Wood, 2006b) for model parameters can be extracted from any fitted **gam()** or **bam()** model with **vcov(model)**. This can in turn be used to generate samples from the posterior distribution of the model, as the approximate Bayesian covariance matrix already incorporates the uncertainty from having to estimate the covariance matrix into it (the standard confidence intervals used in **mgcv** are in fact Bayesian posterior credible intervals, which happen to have good frequentist properties; Wood, 2006b; Marra & Wood, 2012). Viewing our GAM as Bayesian is a somewhat unavoidable consequence of the

---

<sup>9</sup>For model 2 both the factor-smoother, and tensor products of random effect (“re”) and other smooth terms do not have a penalized nullspace by construction (they are full rank), as noted above. For model 3 groupwise terms, we used basis types that had a penalty added to the nullspace, so called “shrinkage” methods: **bs="ts"**, **"cs"**, or **"ps"** have this property.

equivalence of random effects and splines — if we think that there is some true smoother that we wish to estimate, we must take a Bayesian view of our random effects (splines) as we do not think that the true smoother changes each time we collect data (Wood, 2017a, Section 5.8).

This also means that HGAMs can be included as components in a more complex fully Bayesian model. The **mgcv** package includes a function `jagam()` that can take a specified model formula and automatically convert it into code for the JAGS (or BUGS) Bayesian statistical packages, which can be adapted by the user to their own needs.

Similarly, the **brms** package (Bürkner, 2017), which can fit complex statistical models using the Bayesian software **Stan** (Carpenter et al., 2017) allows for the inclusion of **mgcv**-style smooth terms as part of the model specification. The **brms** package does not currently support `te()` tensor products, but does support factor-smooth interactions and `t2()`-style tensor products, which means all of the models fitted in this paper can be fit by **brms**.

## Beyond HGAMs: functional regression

The HGAMs we have discussed are actually a type of *functional regression*, which is an extension of standard regression models to cases where the outcome variable  $y_i$  and/or the predictor variables  $x_i$  for a given outcome are functions, rather than single variables (Ramsay & Silverman, 2005). HGAMs as we have described them are a form of function-on-scalar regression (Ramsay & Silverman, 2005; Reiss, Huang & Mennes, 2010), where we are trying to estimate a smooth function that varies between grouping levels. Here the “scalar” refers to the grouping level, and the function is the smooth term that varies between levels; in contrast, a standard GAM is a type of scalar-on-scalar regression, as the goal is to use a set of single values (scalars) to estimate each (scalar) response.

We have deliberately focused our paper on these simpler classes of functional regression model, and chosen to use the term HGAM rather than functional regression, as we believe that this more clearly connects these models to modelling approaches already familiar to ecologists. Further, we consider the unit of analysis to still be individual observations, as compared to functional regression where the unit of analysis is whole functions. For instance, we are interested in applications such as species distribution modelling, where the presence of a given species may be predicted from a sum of several species-specific functions of different environmental variables.

However, there is an extensive literature dedicated to the estimation of more complex functional regression models for any interested reader (see Morris (2015) and Greven & Scheipl (2017) for a good introduction and overview of more recent work in this field). The **refund** package (Greven & Scheipl, 2017) uses the statistical machinery of **mgcv** to fit these models, and should be usable by anyone familiar with **mgcv** modelling syntax. Functional regression is also a major area of study in Bayesian statistics (e.g., Kaufman, Sain & others (2010)).



## Conclusion

HGAMs are a powerful tool to model intergroup variability, and we have attempted to illustrate some of the range and possibilities that these models are capable of, how to fit them, and some issues that may arise during model fitting and testing. Specifying these models and techniques for fitting them are active areas statistical research, so this paper should be viewed as a jumping-off point for these models, rather than an end-point; we refer the reader to the rich literature on GAMs (e.g. Wood, 2017a) and functional regression (Ramsay & Silverman, 2005; Kaufman, Sain & others, 2010; Scheipl, Staicu & Greven, 2014) for more on these ideas.

## Acknowledgements

The authors would like to thank Carly Ziter, Tiago Marques, Jake Walsh, Geoff Evans, Paul Regular, and Laura Wheeland for their thoughtful feedback on earlier versions of this manuscript, and the Ecological Society of America for hosting the **mgcv** workshops that this work started from. EJP was funded by National Science and Engineering Research Council of Canada (NSERC) and Fisheries and Oceans Canada. GLS is funded by a Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2014-04032). DLM was partly funded by OPNAV N45 and the SURTASS LFA Settlement Agreement, managed by the U.S. Navy’s Living Marine Resources program under Contract No. N39430-17-C-1982. NMR was partially funded by the USAID PREDICT-2 program.

All authors contributed to developing the initial idea for this paper, and to writing and editing the manuscript. Author order after the first author was chosen using the code:

```
set.seed(11)
sample(c('Miller', 'Ross', 'Simpson'))
```

All code used to generate this paper, as well as prior versions of this manuscript, are available at: [github.com/noamross/mixed-effect-gams](https://github.com/noamross/mixed-effect-gams).

## Bibliography

- Baayen RH., Rij J van., Cat C de., Wood S. 2018. Autocorrelated errors in experimental data in the language sciences: some solutions offered by Generalized Additive Mixed Models. In: *Mixed-Effects Regression Models in Linguistics*. Springer, 49–69.
- Bates D., Mächler M., Bolker B., Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.
- Bolker BM., Brooks ME., Clark CJ., Geange SW., Poulsen JR., Stevens MHH., White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends*

- 865 *in Ecology & Evolution* 24:127–135.
- 866 Burnham KP., Anderson DR. 1998. *Model selection and inference: A practical information-*  
867 *theoretic approach*. New York, NY: Springer Science & Business Media.
- 868 Bürkner P-C. 2017. **brms**: An R package for Bayesian multilevel models using Stan. *Journal*  
869 *of Statistical Software* 80:1–28.
- 870 Carpenter B., Gelman A., Hoffman MD., Lee D., Goodrich B., Betancourt M., Brubaker M.,  
871 Guo J., Li P., Riddell A. 2017. Stan: A probabilistic programming language. *Journal of*  
872 *statistical software* 76.
- 873 de Boor C. 1978. *A Practical Guide to Splines*. Springer.
- 874 Dormann CF., Calabrese JM., Guillera-Arroita G., Matechou E., Bahn V., Bartoń K., Beale  
875 CM., Ciuti S., Elith J., Gerstner K., Guelat J., Keil P., Lahoz-Monfort JJ., Pollock LJ.,  
876 Reineking B., Roberts DR., Schröder B., Thuiller W., Warton DI., Wintle BA., Wood SN.,  
877 Wüest RO., Hartig F. 2018. Model averaging in ecology: A review of Bayesian, information-  
878 theoretic, and tactical approaches for predictive inference. *Ecological Monographs* 88:485–504.  
879 DOI: 10.1002/ecm.1309.
- 880 Efron B., Morris C. 1977. Stein’s paradox in statistics. *Scientific American* 236:119–127.
- 881 Forster M., Sober E. 2011. AIC scores as evidence: A Bayesian interpretation. In: Bandy-  
882 opadhyay PS, Forster MR eds. *Philosophy of Statistics*. Handbook of the Philosophy of  
883 Science. Boston, MA: Elsevier B.V., 535–549.
- 884 Gelman A. 2006. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*  
885 48:432–435.
- 886 Gelman A., Carlin J., Stern H., Dunson D., Vehtari A., Rubin D. 2013. *Bayesian Data*  
887 *Analysis, third edition*. Taylor & Francis.
- 888 Greven S., Scheipl F. 2017. A general framework for functional regression modelling. *Statistical*  
889 *Modelling* 17:1–35. DOI: 10.1177/1471082X16681317.
- 890 Hastie TJ., Tibshirani RJ. 1990. *Generalized Additive Models*. Taylor & Francis.
- 891 Kaufman CG., Sain SR., others. 2010. Bayesian functional ANOVA modeling using gaussian  
892 process prior distributions. *Bayesian Analysis* 5:123–149.
- 893 Kimeldorf GS., Wahba G. 1970. A correspondence between Bayesian estimation on stochastic  
894 processes and smoothing by splines. *The Annals of Mathematical Statistics* 41:495–502. DOI:  
895 10.1214/aoms/1177697089.
- 896 Laird NM., Ware JH. 1982. Random-effects models for longitudinal data. *Biometrics*  
897 38:963–974.
- 898 Lathrop RC. 2000. Madison Wisconsin Lakes Zooplankton 1976 - 1994. *Environmental Data*  
899 *Initiative*.
- 900 Marra G., Wood SN. 2012. Coverage properties of confidence intervals for generalized additive

- model components. *Scandinavian Journal of Statistics* 39:53–74. DOI: 10.1111/j.1467-9469.2011.00760.x.
- McCullagh P., Nelder JA. 1989. *Generalized Linear Models, Second Edition*. CRC Press.
- McMahon SM., Diez JM. 2007. Scales of association: hierarchical linear models and the measurement of ecological systems. *Ecology Letters* 10:437–452.
- Morris JS. 2015. Functional regression. *Annual Review of Statistics and Its Application* 2:321–359. DOI: 10.1146/annurev-statistics-010814-020413.
- Potvin C., Lechowicz M., Tardif S. 1990. The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. *Ecology*:1389–1400.
- Ramsay J., Silverman B. 2005. *Functional Data Analysis*. New York, NY: Springer Science+Business Media, Inc.
- Reiss PT., Huang L., Mennes M. 2010. Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics* 6. DOI: 10.2202/1557-4679.1246.
- Ruppert D., Wand MP., Carroll RJ. 2003. *Semiparametric Regression*. Cambridge University Press.
- Scheipl F., Staicu A-M., Greven S. 2014. Functional additive mixed models. *Journal of Computational and Graphical Statistics* 24:477–501.
- Verbyla AP., Cullis BR., Kenward MG., Welham SJ. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48:269–311. DOI: 10.1111/1467-9876.00154.
- Wieling M., Tomaschek F., Arnold D., Tiede M., Bröker F., Thiele S., Wood SN., Baayen RH. 2016. Investigating dialectal differences using articulography. *Journal of Phonetics* 59:122–143.
- Wood SN. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65:95–114.
- Wood SN. 2006a. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* 62:1025–1036. DOI: 10.1111/j.1541-0420.2006.00574.x.
- Wood SN. 2006b. On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics* 48:445–464.
- Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73:3–36. DOI: 10.1111/j.1467-9868.2010.00749.x.
- Wood SN. 2017a. *Generalized Additive Models: An Introduction with R, 2nd Edition*. Boca Raton, FL: CRC Press.
- Wood SN. 2017b. P-splines with derivative based penalties and tensor product smoothing of

936 unevenly distributed data. *Statistics and Computing* 27:985–989. DOI: 10.1007/s11222-016-  
937 9666-x.

938 Wood SN., Goude Y., Shaw S. 2015. Generalized additive models for large data sets.  
939 *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64:139–155. DOI:  
940 10.1111/rssc.12068.

941 Wood SN., Li Z., Shaddick G., Augustin NH. 2017. Generalized additive models for gigadata:  
942 Modeling the U.K. black smoke network daily data. *Journal of the American Statistical*  
943 *Association* 112:1199–1210. DOI: 10.1080/01621459.2016.1195744.

944 Wood SN., Pya N., Säfken B. 2016. Smoothing parameter and model selection for gen-  
945 eral smooth models. *Journal of the American Statistical Association* 111:1548–1563. DOI:  
946 10.1080/01621459.2016.1180986.

947 Wood SN., Scheipl F., Faraway JJ. 2013. Straightforward intermediate rank tensor product  
948 smoothing in mixed models. *Statistics and Computing* 23:341–360. DOI: 10.1007/s11222-012-  
949 9314-z.