Received: 10 October 2024 | Accepted: 6 January 2025

DOI: 10.1111/2041-210X.14498

Methods in Ecology and Evolution

REVIEW

Bayesian views of generalized additive modelling

David L. Miller[1,2]

[1]Biomathematics and Statistics Scotland, Dundee, UK

[2]UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, UK

Correspondence
David L. Miller
Email: dave@ninepointeightone.net

Funding information
OPNAV N45 (US Navy Living Marine Resources Program), Grant/Award Number: N39430-17-C-1982

Handling Editor: Nicholas Clark

Abstract

1. Generalized additive models (GAMs) are a frequently used, flexible framework applied to many problems in statistical ecology. They are commonly used to incorporate smooth effects into models via splines, including spatial components in species distribution models.
2. GAMs are often considered to be a purely frequentist framework ('generalized linear models with wiggly bits'), however links between frequentist and Bayesian approaches to these models were highlighted early-on in the literature. From a practical perspective, Bayesian thinking underlies many parts of the implementation in the popular R package mgcv, so understanding these underpinnings can be informative during model building and assessment.
3. This article aims to highlight useful links (and differences) between Bayesian and frequentist approaches to smoothing, as detailed in the statistical literature, in accessible way, with a focus on the mgcv implementation. By harnessing these links we can expand the set of modelling tools we have at our disposal, as well as our understanding of how existing methods work.
4. Two important topics for quantitative ecologists are covered in detail: model term selection and uncertainty estimation. Taking Bayesian viewpoints for these problems makes them much more tractable in many applied settings. Examples are given using data from the NOAA Alaska Fisheries Science Center's groundfish assessment program.

KEYWORDS
basis-penalty smoothers, empirical Bayes, random effects, smoothers

1 | INTRODUCTION

Flexible modelling of responses for a variety of distributions (binary, count, bounded, continuous) is an indispensable tool for quantitative ecologists. Common applications include species distribution modelling (Golding & Purse, 2016), abundance estimation (Miller et al., 2022), dose–response modelling (Jacobson et al., 2022), movement (Aarts et al., 2012), ecosystem health (Augustin et al., 2009) and more. In each case what is important is incorporating the structure of the data and/or data collection process into the model (be that the form of relationships, spatial correlation, blocking effects, etc.). Informally this structure can be thought of as imposing some prior on how we would like the terms in the model to behave. In this article, I regurgitate some results from the statistical literature emphasizing this (Section 2) and then show how these tools can be used (or are already used) by those engaged in ecological modelling (Section 3).

Generalized additive models (GAMs; e.g. Wood, 2017) are often taught as an extension of the linear model: adding wiggles (via smoothers) to make a (G)LM more flexible, often as a more principled step forward from adding polynomial terms. "Smooth" is often a synonym for spline (DeBoor, 1978), but there are many possible model terms that can be specified as *basis functions* subject to *penalties*: so-called "basis-penalty smoothers". This class of models includes ranges from very simple structures (random effects), through to more structured penalties, obtaining more complex hierarchical random effects models. Having the penalty encode spatial information about a graphical structure gives (Gaussian) Markov random fields (Rue & Held, 2005) or multivariate spline models like thin-plate regression splines (Wood, 2003). We can also use tensor products of terms to construct multidimensional interaction-type effects (Wood, 2000), allowing for different units to be used for each covariate (i.e. anisotropy). Here I use the word smooth to include all these possible flexible model terms and generally denote them as *s*().

Though the term "GAM" has significant baggage regarding the fitting method and type of terms, it really just describes the form of the linear predictor in the model (terms add together) and the response distribution. For example, a model may look like:

$$g(\mu_i) = \boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{\theta} + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}, x_{4i}), \tag{1}$$

where $\mu_i \equiv \mathbb{E}(Y_i)$ and $Y_i \sim \mathrm{EF}(\mu_i, \phi)$, where Y_i $(i = 1, \ldots, n)$ is the response and $\mathrm{EF}(\mu_i, \phi)$ indicates an exponential family distribution with mean μ_i and scale parameter ϕ. $\boldsymbol{a}_i^{\mathsf{T}}$ is a vector of slopes and intercept covariates, where $\boldsymbol{\theta}$ are their associated coefficients. The s_j are "smooth" functions of one or more of the covariates $x_{1i}, x_{2i}, x_{3i}, x_{4i}$.

This definition can be adapted to generalized additive mixed models (GAMMs) and generalized linear mixed models (GLMMs), as we will see below.

The smooth terms are what makes GAMs an interesting and useful evolution of the generalized linear model. In a very general sense, they are constructed from sums of simple basis functions (e.g. DeBoor, 1978). We can construct a complicated function by summing smaller, less complicated *basis functions*. In general for some smooth *s* of covariate *x*, we have the following decomposition:

$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x), \tag{2}$$

where b_k are fixed basis functions (with maximum complexity or basis dimension *K*) and β_k are coefficients to be estimated. This basis function approach is extremely flexible, so to avoid overfitting we penalize the flexibility of each smooth term according to its *wiggliness*. This means that we can let *K* be relatively large, and let the penalty remove the extra flexibility. The fitted model has much smaller effective degrees of freedom (EDF); that is the degrees of freedom actually used by the model, once the penalty is taken into account (usually defined as the sum of the diagonal elements of the hat matrix; Wood, 2017, section 5.4.2). Figure 1 illustrates these concepts. Generally such a penalty will be an integral (sometimes a sum) of squared derivatives of *s* (since derivatives measure the changes in *s*). The penalty can be written in the form $\sum_{m=1}^{M} \lambda_m \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{S}_m \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of coefficients, $\boldsymbol{S}_m$ is a matrix of the fixed parts of the penalty (integrated, squared derivatives of the b_ks, which do not change) and λ_m are smoothing parameters to be estimated that

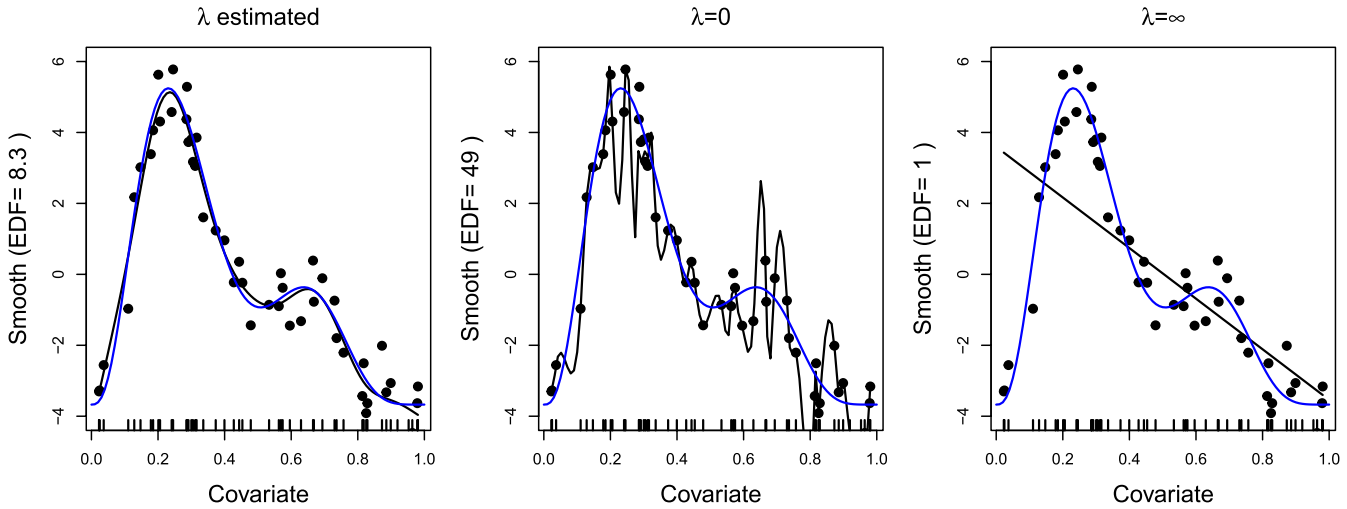

FIGURE 1 The effect of smoothing parameters on the effective degrees of freedom (EDF) of a smooth. In each case, data were simulated from the true, blue, function with normal noise (with zero mean and standard deviation of 0.5) added. The data are shown as points. In each of the three plots, a thin-plate regression spline was fitted to the data with differing smoothing parameters. In the right plot, the smoothing parameter, λ, was estimated from the data, giving an EDF of 8.3. In the middle plot, the smoothing parameter was set to zero, meaning the penalty has no effect, leading to a very wiggly fit (EDF is the maximum). In the left plot, the smoothing parameter was set to be (numerically) infinite, leading to a penalty that doesn't allow for any wiggles. This leads to an EDF of 1, leaving only a linear fit (since this has no second derivative, it lies in the nullspace of the penalty and is unpenalized; see Section 3.1).

control the influence of the penalty (where the $\boldsymbol{S}_m$s are padded with zeros so the sum forms a block matrix); see for example Wood (2017, section 4.2.2). Writing the penalty in this way means that we can compute the $\boldsymbol{S}_m$s once and during fitting the penalty is calculated by matrix multiplication only. Note that multiple penalty terms can correspond to a single smooth or multiple smooths may share a single smoothing parameter so M is not necessarily the number of smooth terms in the model.

We want to estimate model parameters that describe the data best, in the sense that we want to draw lines (or surfaces, etc.) that are close to the data but do not interpolate them. Formally, we can set this up as a penalized log-likelihood (Hastie & Tibshirani, 2000) to find

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ l(\boldsymbol{\beta}) - \sum_{m=1}^{M} \lambda_m \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{S}_m \boldsymbol{\beta} \right\}, \tag{3}$$

where l is the log-likelihood and there are M smoothing parameters to estimate. Here we are trading-off between fit (high log-likelihood giving smooths close to the data) and penalty (large penalty indicating smooths are too wiggly).

Conditional on the λ_ms, estimation of $\hat{\boldsymbol{\beta}}$ in (3) is relatively straightforward and the problem can be attacked with penalized iteratively re-weighted least squares (PIRLS) as for a GLM (Wood, 2017, section 6.1.1). Estimating both $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}$ is more complicated, as the smoothing parameters will constrain the values that $\hat{\boldsymbol{\beta}}$ can take. There have been various proposals for methods to fit such models in a frequentist framework and these fall into two categories: prediction error minimizing methods, like generalized cross-validation (GCV) or Akaike's information criterion (AIC), or likelihood based methods like restricted maximum likelihood and marginal likelihood (REML/ML) (see Wood, 2011, for a review). Prediction error minimizing methods have been shown to overfit (undersmooth) at finite sample sizes (for GCV; Reiss & Ogden, 2009), especially when errors are correlated (for AIC; Krivobokova & Kauermann, 2007), so REML/ML have become the preferred methods. REML and ML cast the smooth functions as random effects (Ruppert et al., 2003) and smoothing parameters as variance parameters, so we can think of $\boldsymbol{\beta}$ as being a multivariate normal random effect with a variance proportional to λ (with structure imposed by the penalty matrix/matrices).

When we talk about adding smooth functions to our model, we tend to concentrate on equations like (1), looking at the mean effects of including smooths rather than thinking about the penalty. We usually view the penalty as a way of constraining our fit, stopping it from being too wiggly and ensuring that our model does not overfit. We can also think of the basis-penalty as the consequence of the problem definition, we have chosen them due to what we know about the dependencies and structures in the data (or data collection process). In practice, for univariate smoothing, switching between basis functions does not tend to make a big impact on results unless there are clear features that need to be accounted for (such as cyclic phenomena, boundary issues, etc.); it is certainly not the case that one should spend time searching for an 'optimal' basis. In a loose sense selecting the basis is equivalent setting-up a prior on the kinds of functions we want to fit. The rest of this article investigates this idea further and explores some useful applications in ecology.

2 | BAYESIAN INTERPRETATIONS

We can quickly get to a convenient Bayesian formulation by exponentiating the objective function in (3) (Wood, 2017, section 5.8), which in the frequentist case gives us a penalized likelihood $\mathscr{L}_p$:

$$\mathscr{L}_p(\boldsymbol{\beta}, \lambda) = \mathscr{L}(\boldsymbol{\beta}) \exp\left(- \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_\lambda \boldsymbol{\beta} \right). \tag{4}$$

We recognize this as Bayes theorem: we might better write $\mathscr{L}_p(\boldsymbol{\beta}, \lambda)$ as $p(\boldsymbol{\beta} | \lambda, \mathbf{y})$ (the posterior for $\boldsymbol{\beta}$) and the likelihood $\mathscr{L}(\boldsymbol{\beta})$ as $p(\mathbf{y} | \lambda, \boldsymbol{\beta})$. Finally, $\exp\left(-\boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_\lambda \boldsymbol{\beta}\right)$ acts as a prior on $\boldsymbol{\beta}$, $p(\boldsymbol{\beta})$. This prior is proportional to a multivariate normal distribution with mean zero and think of $\mathbf{S}_\lambda$ (defined as $\mathbf{S}_\lambda = \sum_m \lambda_m \mathbf{S}_m$) as a prior precision matrix.

2.1 | Specifying priors

By using a smooth term for a given covariate in our model, we are specifying that observations which are close to each other (in some sense) in covariate space have similar values, the response varies smoothly (according to some measure of smoothness) and that the true function we seek to estimate is more likely to be smooth than wiggly (hence penalizing wigglyness). The Bayesian formulation allows us to be more explicit about these beliefs (Wood, 2017, section 5.8). In general, if we want to fit a model $y_i = s(x_i)$, there is no unique solution unless some restriction is put on the form of s (Watson, 1984).

Looking at (4), this says that $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{S}_\lambda^-)$, where $\boldsymbol{S}_\lambda^-$ is the pseudoinverse of $\boldsymbol{S}_\lambda$. Large penalty entries in $\mathbf{S}$ correspond to wiggly basis functions (we want to penalize those more strongly) which, when inverted, give small variances (our prior is that basis function's coefficient is close to zero): this makes explicit our belief that smoother functions are more likely than wiggly ones (Wood, 2006).

Since often some of the elements of $\boldsymbol{\beta}$ are not penalized (e.g. slope or intercept terms, which do not have derivatives), this leads to improper priors as there are no constraints on the value of the slope or intercept for those terms. In this case, the pseudoinverse of $\boldsymbol{S}_\lambda$ is required. Some basis-penalty smoothers do lead to proper priors for all elements of $\boldsymbol{\beta}$ (e.g. the P-spline approach of Lang & Brezger, 2004) and generally a proper prior can be found for any smooth by using the methods in Section 3.1. Identifiability constraints (Wood, 2017, section 5.4.1) that need to be imposed on the model (e.g. that there is only one intercept in the model) may also lead to proper priors (Marra & Wood, 2011).

Various different basis function-penalty combinations available in the literature express different priors on how we want our model terms to behave. For example, cyclic smoothers give us terms which 'match' up to a set number of derivatives at the start/end of the data and can be useful for temporal/seasonal effects. Many solutions

have been proposed to the issue of spatial smoothers in areas with complex coastlines (e.g. Miller & Wood, 2014 and references therein). Wood et al. (2008) propose the soap film smoother, which simultaneously estimates a boundary smooth while constraining values inside the boundary. These models can be fitted using normal GAM machinery, since effects are generated by transforms of the covariates (application of the basis functions) and the prior (penalty) on the corresponding coefficients.

REML and ML are referred to as *empirical Bayes* methods, as when we take the random effects interpretation of $\boldsymbol{\beta}$, we can think of this as a prior and our fit criterion assesses the likelihood of the data given the implied prior on $\boldsymbol{\beta}$, as in (4). The 'empirical' of the name indicates that there is no prior for the smoothing parameters (see Carlin & Louis, 2000 for an overview of empirical Bayes methodology). Taking a fully Bayesian (FB) approach, it is common to put a vague gamma prior on each element of λ or uniform priors on their logarithm (Wood, 2016). Specifying priors on smoothing/variance parameters can be tricky (Simpson et al., 2017), this is especially the case for smoothing parameters as the true values of the smoothing parameter(s) could be infinite if the true smooth is linear (right plot in Figure 1). It can also be hard to come-up with informed priors about smoothing parameters, as we often do not have a direct interpretation of their values.

When using splines we must also decide on knot placement/number and basis complexity/dimension [K in (2); these are usually linked]. Since $\boldsymbol{S}_{\lambda}^{-}$ involves basis functions (or at least their derivatives), the number of basis functions (and/or number of knots) and knot placement will affect the posterior. Effects of placement can be mitigated to some extent by over-specifying the number of knots/basis functions and allow wigglyness to be dictated by the smoothing parameter (Pya & Wood, 2016; Wood, 2017, section 5.9). Eigen-based approaches like thin-plate regression splines (Wood, 2003), make placement data-based in cases where regular grids are computationally taxing. Other related approaches include the use of triangulation-based techniques to optimize placement based on data locations (Lindgren et al., 2011).

2.2 | Obtaining posteriors

For a FB approach, we formulate a likelihood and attach priors to the smoothing parameters λ, as well as the model coefficients $\boldsymbol{\beta}$. We could then use MCMC to obtain a posterior. There are many software implementations which can achieve this, so here I only list R packages specifically tailored to GAMs: mgcv::jagam, which implements translation between mgcv and JAGS (Wood, 2016) or brms (Bürkner, 2017), which implements most mgcv models in Stan (Carpenter et al., 2017). Dedicated software packages such as BayesX (Brezger et al., 2005) can also be used. If one wishes to avoid MCMC, integrated nested Laplace approximations (INLA; often implemented via the R-INLA package) could be used instead (Rue et al., 2009; Wood, 2019). Packages that parameterize their multivariate normal distributions using precision matrices rather than variances allow us to side-step the pseudoinversion of the penalty discussed above.

As discussed above, if we take an empirical Bayes (EB) view of the world and do not put priors on λ, we can still obtain posteriors for $\boldsymbol{\beta}$, conditional on λ. For computational efficiency, the Laplace approximation is often used here. Both R-INLA (used in 'empirical Bayes mode') and mgcv use this approach.

Using either approach, we can get to the posterior marginal for $\boldsymbol{\beta}$: $\boldsymbol{\beta} \mid \boldsymbol{y}, \lambda \sim N\left(\hat{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$ where for the Gaussian likelihood case $\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{S}_{\lambda}\right)^{-1}\sigma^2$ and for the exponential family the expression is approximate and we have $\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S}_{\lambda}\right)^{-1}\phi$, where σ^2 is a variance parameter, ϕ is a scale parameter, and $\boldsymbol{W}$ is a weight matrix (Wood, 2017, section 6.10). For FB, we can obtain a posterior for λ and an unconditional posterior for $\boldsymbol{\beta}$. For EB, we only have information conditional on the value of the smoothing parameter(s). Wood et al. (2016) propose a correction to $\boldsymbol{V}_{\boldsymbol{\beta}}$ to account for the uncertainty in the smoothing parameter(s) using a Taylor expansion to approximate the extra variability in the smoothing parameter.

In practice, we can fit our models using EB methods (such as using REML/ML in mgcv) then sample from their posteriors. As noted, this is as straightforward as plugging the mean coefficient estimates and covariance matrix into a multivariate normal random number generator when using a Gaussian likelihood, though in the exponential family case one may have to use a Metropolis-Hastings sampler and proposing from a t-distribution to get reasonable results (such a sampler can be accessed in mgcv using the gam.mh function).

3 | SOME EXAMPLES

The Bayesian results above lead to some useful applications. Here, I highlight a couple of the more commonly used ones. To illustrate these techniques, data from the NOAA Alaska Fisheries Science Center's groundfish assessment program (https://www.afsc.noaa.gov/RACE/groundfish/survey_data/default.htm) was used. The survey consists of summer bottom trawls at set of stations from 1982 through to 2018 and are shown in Figure S1. Response was catch per unit effort (CPUE; measured as individuals per hectare, effectively a density). Location (recorded as latitude/longitude but projected for analysis), date, surface temperature, bottom temperature (both recorded during the trawl, in degrees Celsius) and bathymetry (recorded in metres) were available as covariates. See Stevenson and Lauth (2019) (and references therein) for further details of the survey. The examples below are not intended to be a serious analyses of the data. Data were downloaded from the NOAA AFSC website and processed for this analysis.

3.1 | Term selection

We begin by fitting a model to the CPUE data for walleye pollock (*Gadus chalcogrammus*) in the Eastern Bering sea for 2010 only. The model includes a bivariate smooth of location and then univariate

smooths of surface temperature, bottom temperature and bathymetry. Expected CPUE was modelled as

$$\mathbb{E}(\text{CPUE}_i) = \exp\left[\beta_0 + s(x_i, y_i) + s(\text{Surface}_i) + s(\text{Bottom}_i) + s(\text{Depth}_i)\right], \quad (5)$$

where i indexes the station-years. CPUE was assumed to follow a Tweedie distribution (see e.g. Shono, 2008, for previous applications in fisheries) with a log link. To model CPUE, we may not need all of the covariates: space (x, y), bottom depth (Depth_i), bottom temperature (Bottom_i) and surface temperature (Surface_i). Rather than using hypothesis testing for term selection, here I apply shrinkage/penalty-type methods to remove terms during model fitting, effectively putting different priors on how to deal with the slope and intercept in each smooth. Many approaches are possible (Marra & Wood, 2011) but here I focus on two approaches implemented in mgcv.

As described in Section 2.1, the prior placed on $\boldsymbol{\beta}$ can be improper due to rank deficiency in $\mathbf{S}$. This means that there are linear or intercept terms that are not penalized. We refer to these terms as being *in the nullspace* of the penalty (the rest of the terms being referred to as the *range space*). Figure 1 illustrates this. We can make our priors proper by simply adding an extra penalty term to the model for the nullspace components of each term (the double penalty approach of Marra & Wood, 2011). This is achieved by eigendecomposing the penalty matrix, $\mathbf{S} = \mathbf{U}\longrightarrow\mathbf{U}^\intercal$. We can then form the additional penalty matrix $\mathbf{S}^* = \mathbf{U}^*\mathbf{U}^{*\intercal}$ where $\mathbf{U}^*$ is a matrix of eigenvectors corresponding to the zero entries on the diagonal of $\longrightarrow$. Our original penalty $\mathbf{S}$ stays as-is, as the components in $\mathbf{S}^*$ do not have an effect (since their entries in $\longrightarrow$ are (almost) zero). This approach is implemented as the select = TRUE option in mgcv::gam, and includes one additional smoothing parameter for each smooth term in the model, corresponding to each term's nullspace. Alternatively one can form a basis where the terms that lie in the nullspace have a shrinkage penalty applied to them by simply adding a small value to their corresponding diagonal entries of $\longrightarrow$ so that the resulting penalty matrix is not rank-deficient (the shrinkage approach of Marra & Wood, 2011; implemented as the cs and ts bases in mgcv). One can think of this as adding a ridge regression penalty to the nullspace or, equivalently, as a regularization of the nullspace terms (see, e.g. Hooten & Hobbs, 2015 for further discussion of regularization in ecology).

These two approaches lead to rather different interpretations of how wigglyness should be penalized, or rather: the prior structure of the smooths. The shrinkage approach assumes that the terms in the nullspace should be penalized less than the other parts of the smooth (since their contributions are small), so as the smoothing parameter increases the model goes from very wiggly, to just the terms in the nullspace (e.g. back to a linear model), to having no effect (estimated as zero). This is appealing, as we can clearly see that increasing the smoothing parameter (decreasing the variance scaling) results in a less wiggly result, until the term is removed from the model. The double penalty approach treats the null and range spaces separately and makes no assumption about how much to smooth the nullspace components relative to the other parts of the smooth. This means that the nullspace components can be removed before the rest of the model, since there is a smoothing parameter for each part.

We can fit (5) in mgcv and see what the differences are between the results using these different prior specifications. Comparing the results from fitting all terms as thin-plate regression splines (no selection), using the double penalty approach and shrinkage revealed that the two term selection methods completely removed the surface temperature term from the model. When no selection method was used, the surface temperature term remained as a linear term in the model (though it was not different from 0 according to an F-test). Figure 2 compares the resulting smooth terms, though there are some minor differences the other smooths remain the same between the three models (though this is not guaranteed in general).

3.2 | Uncertainty around smooth terms

From Section 2.2, we could use the posterior of $\boldsymbol{\beta}$ to generate possible parameters then use these to generate possible smooths. From these simulated smooths, we could then consider pointwise intervals over the range of the covariate to build percentile confidence bands. Black lines in Figure 3 shows 1000 posterior samples of the smooth of depth for the shrinkage model in the previous section (black lines), and their 95% quantiles are the bounds of the blue band. We can take a shortcut and rather than simulating, we know that each smooth can be written as a linear combination ($s(x_i) = \mathbf{X}_i\boldsymbol{\beta}$ for a model with a single smooth in it). We can then use construct point-wise credible intervals as $\hat{s}(x_i) \pm z_{\alpha/2}\sqrt{v_i}$, where $\hat{s}$ is our estimated smooth, v_i is the variance of the smooth at point x_i and $z_{\alpha/2}$ is the usual appropriate value from a normal CDF. Justification for these intervals was developed in Nychka (1988) for normal responses and expanded to the generalized case in Marra and Wood (2012). These intervals have good frequentist across-the-function properties: that is a 95% credible interval has close to 95% coverage, when coverage is averaged over the whole function. There may be over and under coverage at the peaks and troughs of the function as we know less about the exact turning points than we do about the function on the way to that turning point (as by its nature we generally do not know if we have samples at exactly the corresponding covariate value at the turning point). The red band in Figure 3 shows these intervals. Since these intervals have good coverage and tell us about the whole function (by the across-the-function property), we can use them to test the hypothesis $H_0: s(x) = 0 \ \forall x$—whether a term should be dropped from the model because it has no effect (the p-values presented in output of mgcv::summary). See Wood (2013) for more detail on how p-values are calculated for this test.

We use the posterior samples in Figure 3 simply to calculate the blue band in the figure here but they can be useful beyond this. Simulating from the posterior of smooth terms (via simulation from the posterior of $\boldsymbol{\beta}$, conditional λ or incorporating uncertainty via the approximation described in Section 2.2) can potentially reveal interesting properties of the fitted smooth which are not reflected in the plotted bands.

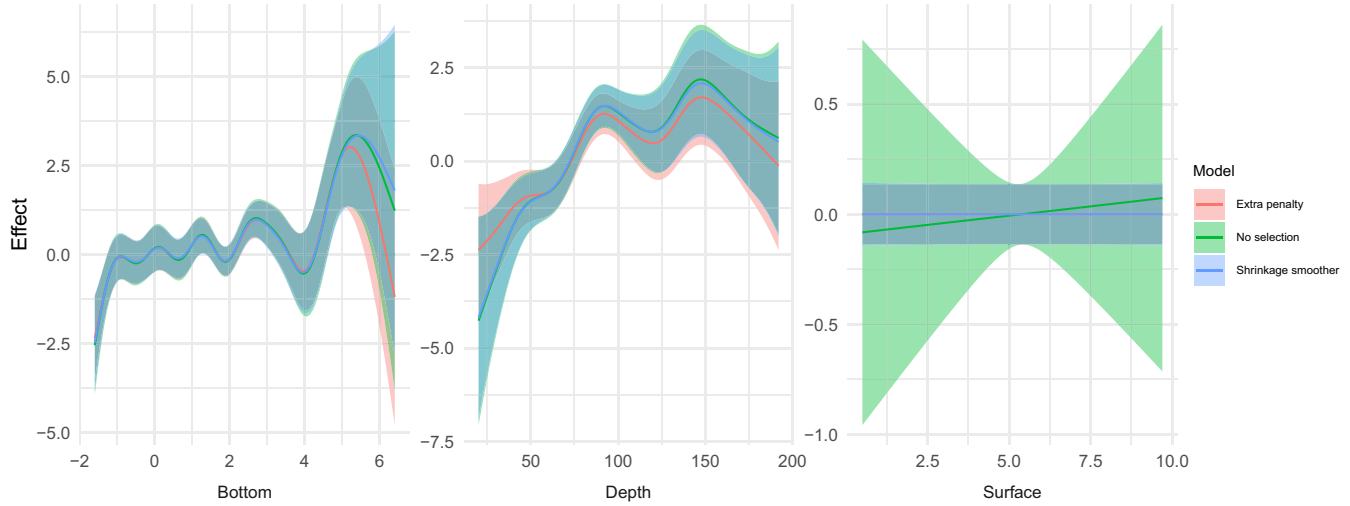

FIGURE 2 Comparison using the shrinkage and double penalty approaches for term selection, with no selection for reference. Models including bottom temperature, depth and surface temperature (left to right) and spatial terms (see Figure S2) were fitted to the walleye pollock CPUE data for 2010. Plots are on the linear predictor (log) scale. Both the shrinkage and double penalty approaches remove the surface temperature term (right), whereas the no selection method (thin-plate regression splines) leave a linear term. Bottom temperature uncertainty is estimated to be much smaller at the upper data range for the extra penalty method. Other terms have minimal differences. Note that confidence bands are generated including uncertainty in the intercept for the top row.

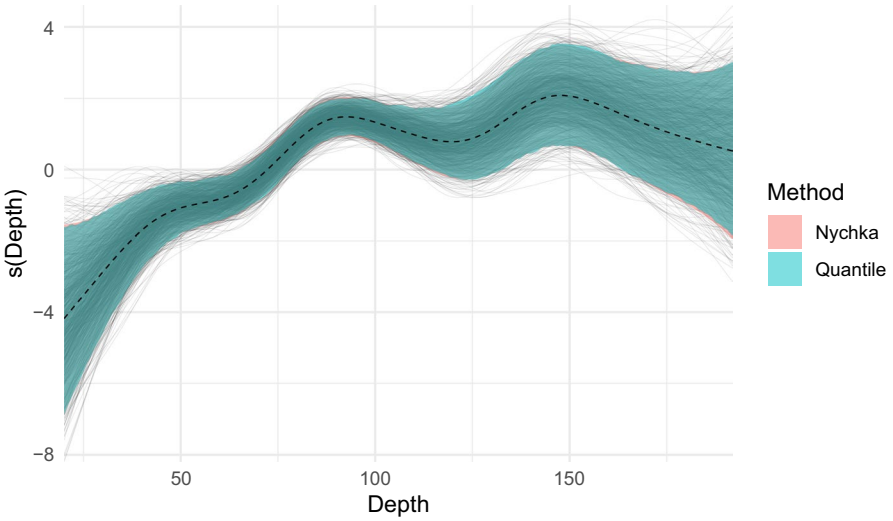

FIGURE 3 Comparison of posterior samples and Nychka-type credible intervals for the shrinkage model discussed in Section 3.1. Dashed black line gives the mean smooth. 1000 posterior samples were generated (black lines) using the algorithm given in Section 3.3, 95% pointwise quantiles of the black lines are given by the green ribbon. 95% (Nychka-type) credible interval is also shown (red ribbon) using the procedure in Section 3.2.

3.3 | Posterior simulation/parametric bootstrap

Sometimes we want more than just uncertainty around individual terms in the model, we want to know about uncertainty either in the model's predictions or summary statistics generated from predictions. Since we can simulate from the posterior of the model parameters, we can use those parameters to calculate functions of the simulated parameters. Calculating summary statistics on the results to obtain uncertainties about those quantities. This is particularly powerful as it allows us to calculate uncertainty about any function of the predictions (including transformations which are non-linear, such as when applying link functions, where this is necessary), avoiding potentially tricky derivations needed to obtain analytical expressions for the variance (see, e.g. the derivations in the appendix of Miller et al., 2022).

A general algorithm (Wood, 2017, section 7.2.7) is as follows:

1. Let B be the number of samples to generate.
2. Form $\mathbf{L}_p$, the matrix that maps the model covariates to the linear predictor (the prediction equivalent of the design matrix).
3. For b in $1, \ldots, B$:
   a. Simulate β_b from the (approximate) posterior of $\boldsymbol{\beta}$.
   b. Calculate the linear predictor $\boldsymbol{\eta}_b = \mathbf{L}_p \boldsymbol{\beta}_b$.
   c. Apply the inverse link function, g, so $\boldsymbol{\mu}_b = g^{-1}(\boldsymbol{\eta}_b)$.
   d. Calculate and store the required summary of $\boldsymbol{\mu}_b$.
4. Perform inference on the B summaries (e.g. calculating empirical variance, percentile intervals, etc).

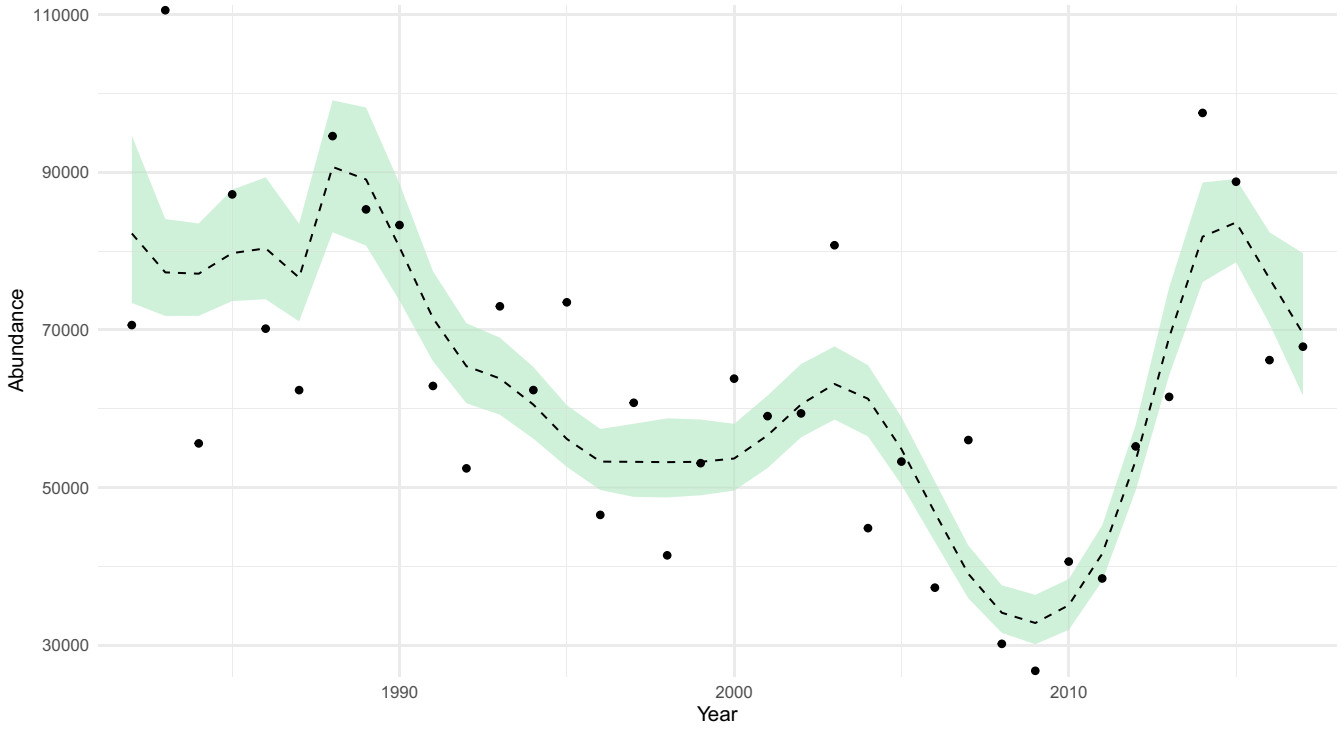

FIGURE 4 Per-year estimates of total abundance at stations for the walleye pollock data from a spatiotemporal model. Black dots indicate the observed values (summed per year). Black dashed line shows the mean of the samples from the posterior, summarized at the year level. The green band shows a point-wise 95% quantile interval. Note these are not very smooth, as predictions are only made at the yearly level.

As an example of where we need to take summaries of non-linear functions of the linear predictor, we can fit a spatiotemporal model to all years (1982–2017) of walleye pollock CPUE data. Our model is then $\mathbb{E}(\text{CPUE}_i) = \exp[s(x_i, y_i, t_i)]$ (where t_i indicates year). Now the smooth $s()$ is constructed as a tensor product of a two dimensional thin-plate regression spline smooth of x and y, and a one dimensional cubic spline smooth of year. We want to obtain a time series of total predicted abundance at the stations per year. So we need to predict at each year and sum over the stations (i.e. space) at 3. (d) in the above algorithm. Figure 4 shows the predictions for the model, made by summing the predictions over space for each year. For simplicity here abundance is calculated by summing over the grid of all trawl station locations, it might be more appropriate to sum over a finer spatial grid and since surveys are in the summer only, abundance estimates were only made once per year (hence the piecewise linear nature of Figure 4). Increasing the spatial or temporal resolution involves modifying $\mathbf{L}_p$ and recalculating steps 3. (c), 3. (d) and 4. (simulation from the posterior does not have to be repeated).

4 | DISCUSSION

This article has highlighted the Bayesian interpretation of GAMs (specifically as implemented in mgcv), which are often thought of as a frequentist method. The article has emphasized that "GAM" only describes a (very flexible) model structure and that there are alternative ways to fit and interpret these models. Taking a Bayesian interpretation gives us many ways in which these links can exploited in practice for applied statistical work: they are not only of mathematical interest.

Several topics have been excluded in this paper for reasons of brevity and clarity, but curious readers may be interested in follow-on topics. In (1) we only consider the case where we are interested in $\mathbb{E}(Y_i)$ where $Y_i \sim \text{EF}(\mu_i, \phi)$ but we need not restrict ourselves to these situations. There are several additional distributions available within mgcv which may be of use, including survival models (cox.ph), scaled t-distributions (scat) and ordered categorical response (ocat), as described in Wood et al. (2016) (see the ?family.mgcv manual page for a full description of all available distributions). We can also extend our models to GAMs for location, shape and scale (GAMLSS; per Rigby & Stasinopoulos, 2005), allowing for the specification of linear predictors for the shape and scale parameters for many distributions including: normal (gaulss), generalized extreme value (gevlss) and zero-inflated Poisson (ziplss). Of some potential interest in ecology, are shape-constrained splines which can be used to ensure that resulting smooths are, for example monotonically increasing/decreasing. These smoothers are implemented in the mgcv-adjacent scam package (Pya & Wood, 2015).

The Bayesian interpretations discussed here have been helpful to construct more reasonable estimates of uncertainty (including smoothing parameter uncertainty) and in order to understand how to construct confidence intervals that have good coverage

properties. In practice, Miller et al. (2022) use the posterior simulation approach outlined here to obtain uncertainty estimates for various aggregations of a complex spatiotemporal model of fin whale abundance (including time series within and between years and uncertainty maps). Since these uncertainty estimation schemes are constructed in simulation-based approach, they can be significantly easier to reason about and much easier to estimate uncertainty from data subsets than traditional analytic estimates. Fitting the GAM via REML/ML is fast (allowing for exploration), then uncertainty estimation procedures are constructed by replacing appropriate steps in from the simulation recipe given above.

Given the multivariate normal prior on the smoother parameters, $\hat{\beta}$, we can view a GAM as a Gaussian processes (GP; Rasmussen & Williams, 2006). Kimeldorf and Wahba (1970) give the general theory for the theoretical link between stochastic processes (such as GPs) and Kent and Mardia (1994) provide further details on links between thin-plate regression splines and one specific type of GP: kriging. Considering random effects as a specific type of basis function, Hefley et al. (2017) provide a more practical guide to this equivalence, specifically with regard to highly structured spatiotemporal data.

These links can surely be used further to develop other new methodology and enhance our understanding of the models that we fit. This approach has already been exploited to show that the stochastic partial differential equation approach proposed by Lindgren et al. (2011) can be viewed as a basis-penalty smoother and implemented in mgcv (Miller et al., 2019). It is a shame that these conceptual links have not been better recognized and exploited further; even a very popular textbook (Ruppert et al., 2003) describes the mixed model representation of the GAM as a "convenient fiction". Coming from the other direction, Fahrmeir et al. (2010) expand on the idea of Bayesian regularization and its interpretation, deriving corresponding priors for ridge regression, lasso, L_p regularization, elastic net, and so forth.

The jagam function (from mgcv) and the brms package allow ecologists to quickly build models using familiar syntax very similar to that for linear models, then transplant these into whatever FB computation system they prefer (see the recipe provided by Miller et al., 2019). The models fitted in Section 3 could be fitted in for example JAGS or Nimble, using jagam to create necessary code. The main difference between those models and the ones presented here would be the priors on the smoothing parameters, which are not terribly interesting in these cases. Where these ideas really shine are in allowing smooths to be included as linear predictors for parameters in for example FB occupancy or mark-recapture models. A general strategy that might be useful is using the GAM as a spatial distribution process for the study species, but building more complex observation processes (possibly from multiple data sources) in FB framework such as the one provided by Nimble. In this way, the complex spatial structure is automatically generated and custom code is only required to interface this part to the observation processes.

Moving beyond mere computational convenience and harnessing the broader Bayesian framework implicit in this modelling strategy can help increase understanding and synthesis, as well as providing further modelling extensions within a familiar framework.

ACKNOWLEDGEMENTS

This work was partly funded by OPNAV N45 and the SURTASS LFA Settlement Agreement, and being managed by the US Navy's Living Marine Resources program under Contract No. N39430-17-C-1982. Richard Glennie and Ian Durbach provided extremely useful feedback on an early draft. John Addy also provided extremely useful feedback later on. I also wish to thank two anonymous reviewers who gave helpful comments on the manuscript.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14498.

DATA AVAILABILITY STATEMENT

The Bering Sea data were downloaded from NOAA's Alaska Fisheries Science Center at https://www.afsc.noaa.gov/RACE/groundfish/survey_data/default.htm and was processed using scripts archived at https://doi.org/10.5281/zenodo.14605076 (miller-data-1). R code for the analyses presented here are available at https://doi.org/10.5281/zenodo.14605014 (miller-data-2).

ORCID

David L. Miller https://orcid.org/0000-0002-9640-6755

REFERENCES

Aarts, G., Fieberg, J., & Matthiopoulos, J. (2012). Comparative interpretation of count, presence-absence and point methods for species distribution models: Species distribution as spatial point process. *Methods in Ecology and Evolution*, *3*(1), 177–187. https://doi.org/10.1111/j.2041-210X.2011.00141.x

Augustin, N. H., Musio, M., von Wilpert, K., Kublin, E., Wood, S. N., & Schumacher, M. (2009). Modeling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association*, *104*(487), 899–911. https://doi.org/10.1198/jasa.2009.ap07058

Brezger, A., Kneib, T., & Lang, S. (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, *14*, 1–22. https://doi.org/10.18637/jss.v014.i11

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carlin, B. P., & Louis, T. A. (2000). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, *95*(452), 1286–1289.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

DeBoor, C. (1978). *A practical guide to splines*. Springer.

Fahrmeir, L., Kneib, T., & Konrath, S. (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage,

smoothing and predictor selection. *Statistics and Computing*, *20*(2), 203–219. https://doi.org/10.1007/s11222-009-9158-3

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*(5), 598–608. https://doi.org/10.1111/2041-210X.12523

Hastie, T. J., & Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science*, *15*(3), 196–223.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., & Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, *98*(3), 632–646. https://doi.org/10.1002/ecy.1674

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3–28.

Jacobson, E. K., Henderson, E. E., Miller, D. L., Oedekoven, C. S., Moretti, D. J., & Thomas, L. (2022). Quantifying the response of Blainville's beaked whales to U.S. naval sonar exercises in Hawaii. *Marine Mammal Science*, *38*(4), 1549–1565. https://doi.org/10.1111/mms.12944

Kent, J., & Mardia, K. (1994). The link between kriging and thin-plate splines. In F. P. Kelly (Ed.), *Probability, statistics and optimization* (pp. 325–339). Wiley.

Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, *41*(2), 495–502.

Krivobokova, T., & Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, *102*(480), 1328–1337. https://doi.org/10.1198/016214507000000978

Lang, S., & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, *13*(1), 183–212. https://doi.org/10.1198/1061860043010

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach: Link between Gaussian fields and Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, *55*(7), 2372–2387. https://doi.org/10.1016/j.csda.2011.02.004

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components: Coverage properties of GAM intervals. *Scandinavian Journal of Statistics*, *39*(1), 53–74. https://doi.org/10.1111/j.1467-9469.2011.00760.x

Miller, D. (2025a). *Bayesian views of generalized additive modelling: Supplementary material*. https://doi.org/10.5281/zenodo.14605014

Miller, D. (2025b). *Processing of NOAA AFSC RACE data for the eastern bering sea*. https://doi.org/10.5281/zenodo.14605076

Miller, D. L., Becker, E. A., Forney, K. A., Roberts, J. J., Cañadas, A., & Schick, R. S. (2022). Estimating uncertainty in density surface models. *PeerJ*, *10*, e13950. https://doi.org/10.7717/peerj.13950

Miller, D. L., Glennie, R., & Seaton, A. E. (2019). Understanding the stochastic partial differential equation approach to smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, *25*, 1–16. https://doi.org/10.1007/s13253-019-00377-z

Miller, D. L., & Wood, S. N. (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics*, *21*(4), 715–731. https://doi.org/10.1007/s10651-014-0277-4

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, *83*(404), 1134. https://doi.org/10.2307/2290146

Pya, N., & Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, *25*(3), 543–559. https://doi.org/10.1007/s11222-013-9448-7

Pya, N., & Wood, S. N. (2016). A note on basis dimension selection in generalized additive modelling. *arXiv preprint arXiv:1602.06696* https://doi.org/10.48550/arXiv.1602.06696

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Reiss, P. T., & Ogden, T. R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *71*(2), 505–523.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, *54*(3), 507–554.

Rue, H., & Held, L. (2005). Gaussian Markov random fields: Theory and applications. In *Monographs on statistics and applied probability* (Vol. *104*). Chapman & Hall, Monographs.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Ruppert, D., Wand, M., & Carroll, R. (2003). *Semiparametric regression. Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press.

Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, *93*(1–2), 154–162. https://doi.org/10.1016/j.fishres.2008.03.006

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28. https://doi.org/10.1214/16-STS576

Stevenson, D. E., & Lauth, R. R. (2019). Bottom trawl surveys in the northern Bering Sea indicate recent shifts in the distribution of marine species. *Polar Biology*, *42*(2), 407–421. https://doi.org/10.1007/s00300-018-2431-1

Watson, G. S. (1984). Smoothing and interpolation by kriging and with splines. *Journal of the International Association for Mathematical Geology*, *16*(6), 601–615. https://doi.org/10.1007/BF01029320

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *62*(2), 413–428.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *65*(1), 95–114.

Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, *48*(4), 445–464. https://doi.org/10.1111/j.1467-842X.2006.00450.x

Wood, S. N. (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*(1), 3–36.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*(1), 221–228. https://doi.org/10.1093/biomet/ass048

Wood, S. N. (2016). Just another Gibbs additive modeler: Interfacing JAGS and mgcv. *Journal of Statistical Software*, *75*(7), 1–15. https://doi.org/10.18637/jss.v075.i07

Wood, S. N. (2017). *Generalized additive models. An introduction with R* (2nd ed.). CRC Press.

Wood, S. N. (2019). *Simplified integrated nested Laplace approximation*. Biometrika. https://doi.org/10.1093/biomet/asz044

Wood, S. N., Bravington, M. V., & Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *70*(5), 931–955. https://doi.org/10.1111/j.1467-9868.2008.00665.x

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Plot of the raw CPUE in space per year for walleye pollock.

Figure S2. Comparison using the shrinkage and double penalty approaches for term selection, with no selection for reference.

How to cite this article: Miller, D. L. (2025). Bayesian views of generalized additive modelling. *Methods in Ecology and Evolution*, *00*, 1–10. https://doi.org/10.1111/2041-210X.14498