P-splines and GAMLSS: a powerful combination, with an application to zero-adjusted distributions

Dimitrios M. Stasinopoulos [1], Robert A. Rigby [1], Gillian Z. Heller [2] and Fernanda De Bastiani [3]

[1] School of Computing and Mathematical Sciences, University of Greenwich, London, UK

[2] NHMRC Clinical Trials Centre, University of Sydney, Australia

[3] Department of Statistics, Federal University of Pernambuco, Recife, Pernambuco, Brazil

---

Address for correspondence: Dimitrios M. Stasinopoulos, "School of Computing and Mathematical Sciences, University of Greenwich, London, UK"

E-mail: d.stasinopoulos@londonmet.ac.uk.

Phone: (+420) 221 913 282.

Fax: (+420) 222 323 316.

Abstract: P-splines are a versatile statistical modelling tool, dealing with nonlinear relationships between the response and explanatory variable(s). GAMLSS is a distributional regression framework which allows modelling of a response variable using any parametric distribution. The combination of the two methodologies provides one of the most powerful tools in modern regression analysis. This paper discusses the application of the two techniques when the response variable is zero-adjusted (or semi-continuous), which combines a point probability at zero with a positive continuous distribution.

Key words: P-splines; gamlss; Box-Cox t distribution; zero-adjusted distribution, zero-heavy distribution; semi-continuous distribution

1 Introduction

This paper discusses two powerful ideas in modern statistical modelling: the P-splines approach to smoothing and Generalized Additive Models for Location, Scale, and Shape (GAMLSS). P-splines, introduced by Eilers and Marx (1996), is a very flexible smoothing tool for modelling the relationship between a response variable and one or more explanatory variables, and is based on penalized regression techniques. Its simplicity and flexibility allows it to be used in a variety of practical applications. A good review of P-splines and a more detailed comprehensive account of the topic is given by Eilers and Marx (2021).

GAMLSS was introduced by Rigby and Stasinopoulos (2005) as a way of overcoming some of the limitations associated with Generalized Linear Models (GLM, Nelder

and Wedderburn, 1972) and Generalized Additive Models (GAM, Hastie and Tibshirani, 1990). More specifically, the problems which GAMLSS aimed to overcome were heterogeneity and/or skewness and kurtosis in the distribution of the response. GAMLSS is a distributional regression model (that is, a proper parametric distribution is assumed for the response variable), useful when the distribution of the response variable does not belong to the exponential family and/or when the researcher is not only interested in how the explanatory variables shift the location of the distribution, but also how they affect its shape (e.g. variance (volatility), skewness, kurtosis). The idea that the two methodologies can work together nicely arose at a GAMLSS short course given at Utrecht University in 2008, by Mikis Stasinopoulos and Robert Rigby, as part of the International Workshop on Statistical Modelling. At the end of the course Brian Marx and Paul Eilers joined a very interesting conversation where new ideas about merging the two techniques emerged.

The implementation of GAMLSS at the time was rather rigid, relying on classical cubic smoothing splines for modeling nonlinear relationships. The estimation of the smoothing parameters in the cubic splines was not automatic, but relied on the specification of the degrees of freedom. This made the modelling of several explanatory variables rather difficult. The simplicity of P-splines, and the fact that they could be easily implemented in **R**, made the automatic selection of smoothing parameters far easier. This was the first achievement of the merger of the two methods. Monotonic, circular and other types of smoothers came later.

This article celebrates the merger of the two techniques by using an application in which the response variable has a *zero-adjusted* (or semi-continuous) distribution, i.e. a mixed distribution, comprising a probability mass at zero and a continuous

positive distribution (Rigby et al., 2019). These distributions have received attention in diverse application areas, such as medical (Liu et al, 2019), environmental (Popuri et al, 2015), actuarial (Heller et al, 2006), sports betting (Houghton et al, 2019) and agriculture (Belasco et al, 2012). Modelling approaches are generally either two-part models in which the zero probability and the continuous component are modelled independently, or a single model as in Tobit or Tweedie models. Here we adopt the two-part model approach, accommodated within the GAMLSS framework.

2 P-splines

P-splines result from combining the B-splines basis, which is an orthogonal basis of a piecewise polynomial, with a penalty applied to the coefficients of the model. The idea of P-splines proposed by Eilers and Marx (1996) is based on the work of O'Sullivan (1986) and is derived as the analytic solution of the following penalized least squares problem: minimize Equation (2.1) with respect to $\boldsymbol{\gamma}$:

$$\boldsymbol{S} = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma})^\top \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^\top \boldsymbol{P} \boldsymbol{\gamma}, \tag{2.1}$$

where $\boldsymbol{B}$ is a $n \times r$ regression design matrix that contains the B-spline basis function evaluations (based on a specific explanatory variable) as its columns; $\boldsymbol{P} = \boldsymbol{D}^\top \boldsymbol{D}$ is a $r \times r$ penalty matrix; $\boldsymbol{D}$ is a matrix such that $\boldsymbol{D}\boldsymbol{\gamma}$ forms dth-order differences of $\boldsymbol{\gamma}$; $\boldsymbol{\gamma}$ is the corresponding vector of coefficients of length r; $\boldsymbol{W}$ is an $n \times n$ diagonal matrix containing weights w_i for $i = 1, 2, \ldots, n$; and λ is a smoothing parameter. The solution to the penalized least squares problem is given by

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{B}^\top \boldsymbol{W} \boldsymbol{B} + \lambda \boldsymbol{D}^\top \boldsymbol{D}\right)^{-1} \boldsymbol{W} \boldsymbol{B}^\top \boldsymbol{y} \ , \tag{2.2}$$

and the fitted values are given by $\hat{\boldsymbol{\mu}} = \boldsymbol{B}\hat{\boldsymbol{\gamma}}$. λ is the 'explicit' smoothing parameter (or hyperparameter) which determines the amount of smoothing in (2.2): the greater the value of λ, the greater the influence of the penalty and therefore the smoother the fitted function is. There are other parameters 'implicit' in the fit of a P-spline: the *number of knots* and the *degree* of the piecewise polynomial in the definition of the basis $\boldsymbol{B}$, and the *difference* defined in the penalty matrix $\boldsymbol{D}$. These parameters do affect the model fitting but not as much as the smoothing parameter λ. The number of knots for B-splines could be set at between 20 and 50. Using the degree of the piecewise polynomial as three, as default, guarantees that the first and second derivatives of the fitted smooth function will be continuous, resulting in a smoother curve. Finally, the default choice for the order of the differences in the penalty is two. This choice corresponds to an assumption that the $\boldsymbol{\gamma}$ coefficients behave as a second order random walk, (while order zero would correspond to white noise (random effect) and order one to a random walk). In our experience these default recommendations work well in practice. Surprisingly the number of B-splines knots can be larger than the number of observations and the method still works.

The beauty of P-splines lies in its simplicity and its flexibility. In addition it has a random effects interpretation. Given the basis design matrix $\boldsymbol{B}$, equation (2.2) is easy to program and calculate in any computer language with reasonable matrix implementation. For more than one additive explanatory term, one can expand the $\boldsymbol{B}$ basis or rely on a (modified) backfitting algorithm; see Hastie and Tibshirani (1990) or Stasinopoulos et al. (2017, p. 68). For smoothing in more than one dimension, the tensor product of P-splines can be used, see Rodríguez-Álvarez and Oviedo de la Fuente (2021). Also by modifying the basis $\boldsymbol{B}$ or the penalty matrix $\boldsymbol{P}$ one can create different types of effects. Stasinopoulos et al. (2017, pp. 265-296) shows

several modifications of P-splines. For continuous terms, this includes: *monotonic* and *cyclical* P-splines and ridge and LASSO regression P-splines. (The latter two rely on modifying $\boldsymbol{P}$.) For categorical terms (factors), there exist: (i) *random effects*, (ii) *Gaussian Markov random fields* P-splines, applied to neighbouring (spatial) areas and (iii) *categorical* P-splines designed to reduce the number of levels of a factor. Note that the weights in Equation (2.2) are crucial in the implementation of P-splines within GAMLSS because they are used as iterative weights within the current GAMLSS algorithms, see Stasinopoulos et al. (2017, Chapter 3). The iterative weights are functions of the parameters of the response distribution.

When using P-splines, it is important to be aware of the domain on which the B-splines are defined. This is usually the region from the observed minimum to the observed maximum minimum of the explanatory axis, or some values quite close. It is essential that the domain includes all observed values of the explanatory variable. More details about P-splines are given by the creators of P-splines: Eilers and Marx (2010) and Eilers and Marx (2021), the latter being a comprehensive view of P-splines.

3 GAMLSS

The mathematical and stochastic modelling of the relationship between a response variable and one or more predictor variables was elegantly formulated as the classical linear model, with its accompanying least squares solution, dating back to Gauss and Legendre in the early 1800s. The "normal model", a normal distribution regression model, prevailed as an analysis framework until "Theoretical and applied statistics were both convulsed by the publication of the GLM paper by Nelder and

Wedderburn (1972)." (Aitkin, 2018). The relaxation of the restrictive requirements of the normal model progressed from the Generalized Linear Model (GLM), in which the response distribution was extended from the normal to the exponential family distribution and the link function was introduced; to Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990), which added smooth functions of the covariates to the linear predictor for the mean; to mean and dispersion modelling, in which modelling of the dispersion parameter of the response distribution was introduced (Aitkin, 1987; Smyth, 1989); to Generalized Additive Models for Location, Scale and Shape (GAMLSS) (Rigby and Stasinopoulos, 2005) which arguably represents the biggest leap forward of the regression methodology. The GAMLSS model extends the choice of response distribution to any computable parametric distribution; all distribution parameters may be modelled; and predictors may include linear terms, smooth functions of the covariates, and random effects. The GAMLSS model is

$$\begin{aligned} \boldsymbol{y} &\overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \\ g_k(\boldsymbol{\theta}_k) &= \boldsymbol{X}_k\boldsymbol{\beta}_k + s_{k1}(\boldsymbol{x}_{k1}) + \ldots + s_{kJ_k}(\boldsymbol{x}_{kJ_k}) \qquad \text{for } k = 1, \ldots, K \end{aligned} \tag{3.1}$$

where $\mathcal{D}(\theta_1, \ldots, \theta_K)$ denotes any computable K-parametric distribution; $\theta_1, \ldots, \theta_K$ are the distribution parameters; $g_k(\cdot)$ is the link function for θ_k; $\boldsymbol{X}_k$ is the fixed effects design matrix for θ_k and $\boldsymbol{\beta}_k$ is the corresponding vector of coefficients; $s_{kj}(\cdot)$ is the jth nonparametric smooth function in the model equation for θ_k; and $\boldsymbol{x}_{kj}$ is the explanatory variable for the jth smooth function for θ_k. Note that here we use $s_{kj}(\cdot)$ as a generic notation for any smoother, which can include machine learning techniques such as regression trees and neural networks. For P-splines, $s_{kj}(\cdot)$ takes the linear form $\boldsymbol{B}_{kj}\boldsymbol{\gamma}_{kj}$ where the coefficients $\boldsymbol{\gamma}_{kj}$ are restricted by $\boldsymbol{\gamma}_{kj}^\top \boldsymbol{P}_{kj}\boldsymbol{\gamma}_{kj} \leq \lambda_{kj}$. A similar specification applies also to random effects (Kauermann, 2010).

The innovation of GAMLSS is that all the parameters of the distribution can be modelled as functions of the explanatory variables and therefore all aspects of the distribution for the response i.e. location, scale, skewness and kurtosis can be modelled. While in the original GAMLSS paper (Rigby and Stasinopoulos, 2005) the number of distribution parameters was general, the **R** implementation **gamlss** uses $K = 4$ and the notation $\theta_1 = \mu$, $\theta_2 = \sigma$, $\theta_3 = \nu$, $\theta_4 = \tau$. We will partly use this notation in the zero-adjusted example of Section 4.

The suite of **R gamlss** packages (Stasinopoulos and Rigby , 2008; Stasinopoulos et al., 2017) includes over 100 response distributions. These comprise: *continuous* distributions with support on: the continuous real line $(-\infty, \infty)$; the positive continuous real line $(0, \infty)$; and the unit interval $(0, 1)$; *discrete* distributions with support on: the non-negative integers $\{0, 1, 2, \ldots\}$; and bounded integers $\{0, 1, 2, \ldots, n\}$; *mixed* (or semi-continuous) distributions with support on, for example, $[0, 1]$ and $[0, \infty)$. The former is referred to as *zero- and/or one-inflated*, with probability masses at zero and/or one and a continuous component on $(0, 1)$; and the latter is referred to as *zero-adjusted*, and has a probability mass at zero and a continuous component on the positive real line. In addition, the above distributions may be truncated, transformed or censored, providing a very rich family of response distributions available for modelling. Details are given in Rigby et al. (2019). The P-splines representation of a GAMLSS model is:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{B}_{k1}\boldsymbol{\gamma}_{k1} + \ldots + \boldsymbol{B}_{kJ_k}\boldsymbol{\gamma}_{kJ_k} \qquad \text{for } k = 1, \ldots, K$$

subject to $\boldsymbol{\gamma}_{kj}^{\top}\boldsymbol{P}_{kj}\boldsymbol{\gamma}_{kj} \leq \lambda_{kj}$ for $j = 1, \ldots, J_k$. This results in a penalized log-likelihood of the form:

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \quad = \quad \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) - 0.5 \sum_{k=1}^{K} \sum_{j=1}^{J_k} \lambda_{jk} \boldsymbol{\gamma}_{jk}^{T} \boldsymbol{P}_{jk} \boldsymbol{\gamma}_{jk} \tag{3.2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$ are vectors containing all relevant coefficients and hyperparameters and $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_1^n \log f(y|\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the log-likelihood. The penalized log-likelihood of equation (3.2) can be maximized for fixed $\boldsymbol{\lambda}$ using either of the two algorithms described in Stasinopoulos et al. (2017, Chapter 2). $\boldsymbol{\lambda}$ can be estimated using a (local) REML algorithm described in Rigby and Stasinopoulos (2012).

4 Zero-adjusted distributions

Response variables which can either be zero or a continuous positive quantity, are commonly encountered in diverse application areas. Some examples are: microbiome composition (Liu et al, 2019), daily precipitation and streamflow (Van Ogtrop et al, 2011; Popuri et al, 2015), annual insurance claim amounts (Heller et al, 2006), daily alcohol consumption (Liu et al, 2019), annual health services expenditure (Liu et al, 2019; Neelon et al, 2016b), annual expenditure on sports betting and online gambling (Houghton et al, 2019) and cattle mortality rates (Belasco et al, 2012).

Zero-adjusted distributions (sometimes referred to as semi-continuous or zero-heavy distributions) are appropriate for fitting models with such response variables, i.e. when the support of the response is in $[0, \infty)$ (including zero). These are *mixed* distributions, i.e. a mixture of a discrete distribution (a probability at value zero, i.e. $y = 0$) and a continuous distribution (for $y > 0$). Liu et al (2019) and Neelon et al (2016a) review modelling approaches for such response variables: these comprise two-part models, in which the zero probability and the continuous component are modelled separately, which is the approach that we take below; and Tobit models (Tobin,

1958), which assume an underlying normal (or more generally any) distribution for the response, which is left-censored at zero. This approach could be more restrictive, as the zero probability cannot be modelled explicitly. Informally (ignoring strict mathematical considerations) the distribution

$Y \sim \mathcal{D}(\boldsymbol{\theta}, \pi_0)$ of a zero-adjusted response takes the form:

$$f_Y(y|\boldsymbol{\theta}, \pi_0) = \begin{cases} \pi_0 & \text{if } y = 0 \\ (1-\pi_0) f_{Y_1}(y|\boldsymbol{\theta}) & \text{if } y > 0 \end{cases}$$

where $0 < \pi_0 < 1$; $\pi_0 = P(Y = 0)$ is the point probability at zero; and $f_{Y_1}(y|\boldsymbol{\theta})$ is any probability density function with support on the positive real line, with parameters $\boldsymbol{\theta}^\top = (\theta_1, \theta_2, \ldots, \theta_K)$. The log-likelihood of any zero-adjusted distribution response variable can be split into two separate components: the first component is a binary model, in which the probability π_0 of observing zero is modelled against the explanatory terms; while the second component, in which the response is positive, is from a model with support on the positive real line. A general zero-adjusted GAMLSS model can be written as:

$$\boldsymbol{y} \overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \tag{4.1}$$

$$g_\pi(\boldsymbol{\pi}_0) = \boldsymbol{X}_\pi \boldsymbol{\beta}_\pi + s_{\pi 1}(\boldsymbol{x}_{\pi 1}) + \ldots + s_{\pi J_\pi}(\boldsymbol{x}_{\pi J_\pi}) \tag{4.2}$$

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k \boldsymbol{\beta}_k + s_{k1}(\boldsymbol{x}_{k1}) + \ldots + s_{kJ_k}(\boldsymbol{x}_{kJ_k}) \qquad \text{for } k = 1, \ldots, K \tag{4.3}$$

where equation (4.2) is the predictor for the model for the probability π_0 of whether zero occurred, while equation (4.3) is the predictor for a standard GAMLSS model for parameter θ_k of the distribution of the positive values. Because the parameter π_0 is informational orthogonal to the parameters $\boldsymbol{\theta}_k$[1], maximization of the log-likelihood

[1] i.e. the likelihood function can be split into two components: one containing π_0 and one containing $\boldsymbol{\theta}_k$.

for (4.1) can be achieved by fitting two separate models, one for (4.2) and one for (4.3), see Supplementary material . Combining the two components consists of summing the two deviances to obtain the overall model deviance; if (randomised) quantile residuals are required they are calculated using the zero-adjusted cdf $F_Y(y|\boldsymbol{\theta}) = \pi_0 + (1-\pi_0)F_{Y_1}(y|\boldsymbol{\theta})$, where $F_{Y_1}(y|\boldsymbol{\theta})$ is the cdf with support on the positive real line.

Model (4.1) assumes that the probability that $Y = 0$ (i.e. π_0) does not depend on the distribution of Y for $Y > 0$. This assumption may be valid in certain empirical situations and not in others. However, modelling a specific dependence may be overly restrictive and hence inappropriate. One dependence model is the generalized Tobit model (Rigby et al. (2019), p. 182) which assumes $Y = 0$ (if $Z < 0$) and $Y = Z$ (if $Z > 0$), where Z has any distribution on the real line. Hence the point probability $P(Y = 0) = P(Z < 0)$ depends on the parameters of the distribution of Z. An alternative dependence model is to assume that Y has a Tweedie distribution (Tweedie, 1984; Wood and Fasiolo, 2017), where the Tweedie power parameter lies between 1 and 2. However this is a highly restrictive model, since it is a reparametrization of a compound Poisson-Gamma distribution (the sum of i.i.d. gamma random variables, where the number of terms in the sum has a Poisson distribution). The resulting distribution has a point probability at $Y = 0$ and usually a multi-modal distribution for $Y > 0$. This is unlikely to be an appropriate distribution, although it may be suitable if only the first two moments of Y are the focus of interest. Note also it has an additional problem that the normalizing constant of the Tweedie density is computable only by summing an infinite series.

In the GAMLSS **R** implementation there are two explicitly defined zero-adjusted distributions: the zero-adjusted gamma (`ZAGA`) and the zero-adjusted inverse Gaussian

(`ZAIG`, Heller et al, 2006). Both gamma and inverse Gaussian distributions are members of the exponential family and therefore have the nice theoretical properties of exponential family distributions. The problem is that in practice, exponential family distributions may fail to capture extreme right tail behaviour commonly observed. There is therefore the need for more flexible positive real line distributions. The GAMLSS package **gamlss.dist** provides several positive real line distributions apart from the gamma and inverse Gaussian: the Weibull, Pareto type 2, inverse gamma, log-normal (two-parameter distributions), the generalized gamma, generalized inverse Gaussian and Box-Cox Cole and Green (three-parameter distributions), and the Box-Cox Cole power exponential, Box-Cox t and generalized beta type 2 (four-parameter distributions) (Chapter 19, Rigby et al., 2019). More positive real line distributions can be generated "implicitly" within the GAMLSS **R** packages by taking any explicit GAMLSS continuous distribution with support on the real line and (a) inverse *log* transform it (i.e. exponentiate it) to a distribution with a support $(0, \infty)$, e.g. using function gen.Family("JSU",type="log") to generate a logJSU distribution; (b) *truncate* it to $(0, \infty)$, e.g. using function gen.trun(0,"JSU",type="left") from the **gamlss.tr R** package to generate a JSU distribution left truncated at 0 to $(0, \infty)$. Any of the above positive distributions can be converted into a corresponding zero-adjusted distribution using the **gamlss.inf R** package.

The first four moments of a zero-adjusted distribution are given in Rigby et al. (2019, p. 179). Typical plots of zero-adjusted distributions are shown in Figure 1, where a zero-adjusted Box-Cox t (`BCT`) and a zero-adjusted log Johnson's S_u (`logJSU`) distribution are displayed. Next we demonstrate the power of combining P-splines with GAMLSS using an environmental data set concerning streamflow.

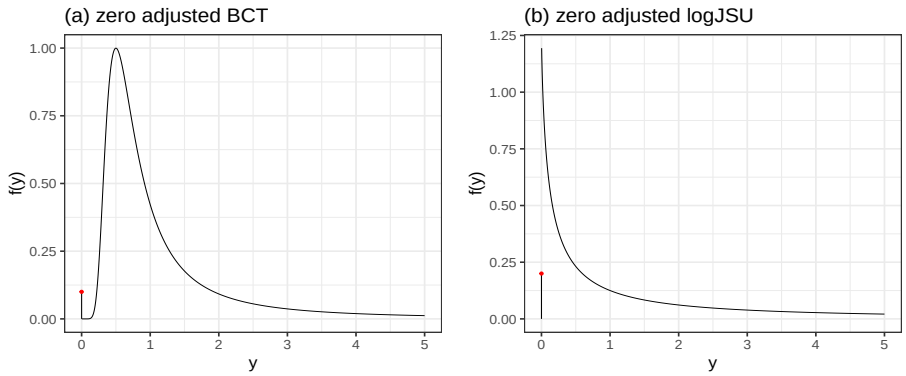

Figure 1: Plots of zero-adjusted distributions: (a) zero-adjusted Box-Cox t (`BCT`) distribution, with values $\mu = 1$, $\sigma = 1$, $\nu = -1$, $\tau = 10$ and $\pi_0 = 0.2$ (b) zero-adjusted log Johnson's S_u (`logJSU`) distribution with values $\mu = 0$, $\sigma = 1$, $\nu = -1$, $\tau = 10$ and $\pi_0 = 0.2$.

4.1 Streamflow data

Van Ogtrop et al (2011) analysed monthly streamflow data from the Balonne River, Queensland, Australia, recorded from 1951 to 2008. The data are a typical example of a zero-adjusted response variable.[2] Modelling streamflow is of environmental and economic interest. A total of 12.4% of the monthly streamflow observations have zero values; the remaining positive streamflows are strongly right-skewed, requiring a positive real line distribution with a long right tail. The reason for a right-skewed long-tailed distribution is evident in Figure 2, where the positive values of the streamflow are shown together with their 50% and 90% quantile values.

The explanatory variables in the data are (i) *time*, (ii) *month*, (iii) lags of log streamflow, $F_{t-i} = \log(\text{streamflow}_{t-i} + 1)$ for $i = 1, 2, 3, 4, 5$ and (iv) lags of the binary re-

[2]In the current analysis, because of data copyright issues, we used similar but slightly amended data.

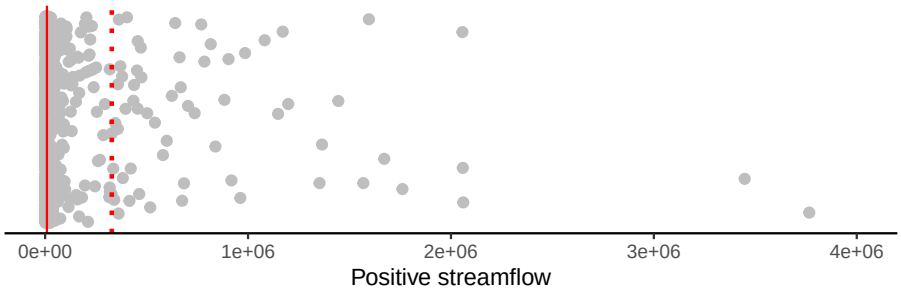

Figure 2: Dotplot of the positive monthly streamflows, 1951–2008; the red vertical lines are the 50% (solid) and 90% (dotted) quantiles of the positive values. The values are jittered on the vertical axis for display.

sponse $P_{t-i} = \mathbb{1}(y_{t-i} > 0)$[3], for $i = 1, 2, 3, 4, 5$, to account for the time series nature of the data; the climatological features: (v) *soi*, the southern oscillation index (monthly air pressure difference between Darwin and Tahiti), (vi) *eof1* and (vii) *eof2*, first and second empirical orthogonal functions of sea surface temperature, respectively; and (viii) *nino*, average sea surface temperatures in the region.

The zero-adjusted distribution model of equations (4.1) – (4.3) is fitted using the **R** package **gamlss.inf**. Since the package does not have a term selection method, we use the fact that the zero-adjusted response variable can be modelled using two separate models, one binary (for all data) and one for the positive responses. Terms are selected for the two models separately using the 'step-GAIC' methodology of the **gamlss** package (Stasinopoulos et al., 2017, Chapter 11), which minimises the Generalised Akaike Criterion where $\ell()$ is the fitted log-likelihood function, df the model degrees of freedom and k a constant. (Setting $k = 2$ we have the standard Akaike Information Criterion (AIC), and $k = \log n$ yields the Bayesian Information

[3]We use the notation $\mathbb{1}(\cdot)$ as the indicator function: $\mathbb{1}(x) = 1$ if x is true, $\mathbb{1}(x) = 0$ otherwise.

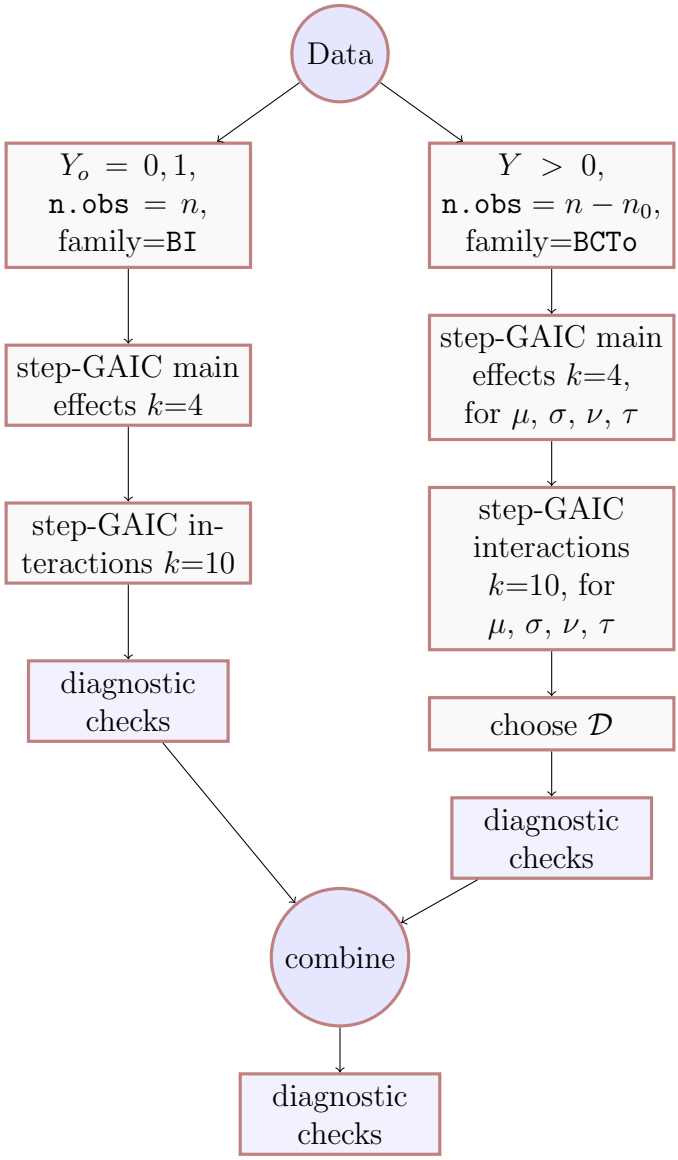

Figure 3: Model selection procedure used for the streamflow data. Two data sets are created: (i) a binary response Y_0 with n observations (where $Y_0 = 0$ if $Y = 0$ and $Y_0 = 1$ if $Y > 0$) and (ii) a positive response for $Y > 0$ with $n - n_0$ observations (n_0 is the number of zeros). A stepwise procedure (using the distributions `BI` and initially `BCTo` respectively) is first applied to the main effects with penalty $k = 4$ and then to the first order interactions with penalty $k = 10$. Note that if interactions were chosen the main effects are refitted (this part is not shown in the diagram). Then for the positive response data a distribution is chosen using the models for μ, σ, ν and τ, followed by model diagnostics to check the distribution adequacy. Finally the two models are combined and diagnostic checks are performed for the combined model. (Details of the procedure are given in Supplementary material .)

Criterion (BIC).) The two models can then be combined into a single zero-adjusted model for Y using the function `gamlssZAdj()` of package **gamlss.inf**. Figure 4.1 is a diagram of the procedure used to select the terms for each distribution parameter. This strategy has been designed in order to detect interactions; a detailed account is given in Supplementary material . [We believe that a similar strategy could be applied to any zero-adjusted regression model, assuming a reasonable number of observations ($n = 696$ in our case) and a relatively small number of feature variables (18 in our case). For positive response variables, (and similarly for response variables on the real line), only steps 3, 4 and 5 in Supplementary material (the right branch in Figure 4.1) are needed.]

4.1.1 Model for π_0

A logit regression model was used for modelling the zero probability π_0, using the methodology of selecting terms described on the left branch of Figure 4.1. The following model was selected:

$$\log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right) = \underset{(18.8)}{-11.021} - \underset{(0.055)}{0.402}\, F_{t-1} + \underset{(0.055)}{0.108}\, F_{t-2} + s(t) - \underset{(0.094)}{0.269}\, soi \qquad (4.4)$$

where t is time in months (from the from the first obervatrion). The standard errors of the linear coefficients were obtained using a nonparametric bootstrap sampling of size $B = 100$; these do not incorporate the uncertainty of the model selection procedure and are therefore optimistic. The function $s(\cdot)$ is fitted using P-splines. Note that while the selection strategy was designed for the systematic evaluation of interaction terms, no interaction terms were selected. The selected lagged values suggest a delay before changes in the explanatory variable have an impact in the binary model. In particular the negative coefficient of F_{t-1} shows that low riverflow

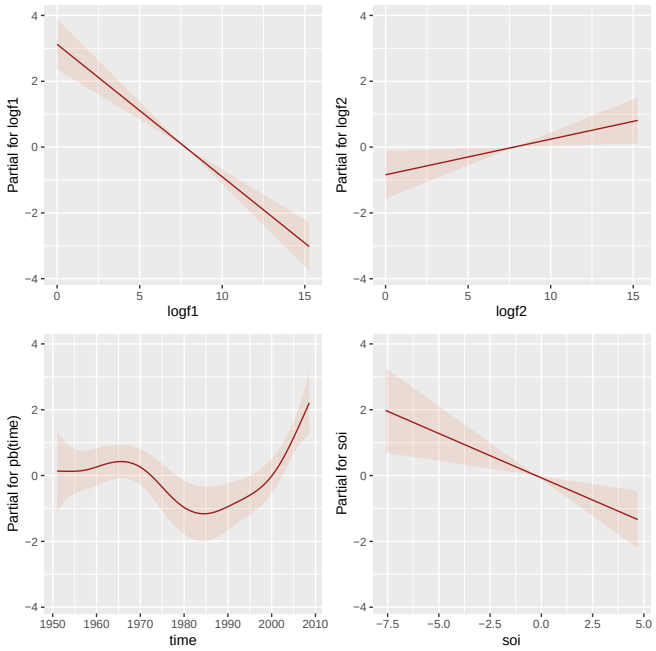

Figure 4: The fitted predictor terms for the π_0 model: top row: F_{t-1} and F_{t-2}, bottom row: *time* and *soi*.

last month increases the probability of zero riverflow this month.

The effects of the covariates on the log-odds of a zero streamflow are shown in Figure 4, which shows the partial effect from each term given that rest of the terms are fixed at their mean values. The P-spline function for *time* indicates that the probability of zero streamflow is stable till 1970, then dips till the early 1980s and then increases sharply till the end of the period of observation. The increasing frequency of zero streamflows after 1980 was also observed by Van Ogtrop et al (2011), who suggested that "increased water extraction occurred post 1980 upstream of the gauging station". Both *soi* and streamflow lag 1 have negative effects, while streamflow lag 2 has a positive effect on the probability of zero streamflow.

4.1.2 Models for the parameters $\boldsymbol{\theta}$ of the positive part

We used STEP 3 of the procedure described in Supplementary material to choose models for μ, σ, ν and τ for an initial Box-Cox t (`BCTo`) distribution, Rigby and Stasinopoulos (2006). The `BCTo` distribution was retained as the best fitting distribution in STEP 4 of the procedure. The resulting fitted coefficients for the `BCTo`(μ, σ, ν, τ) model are given below.

$$\ln \hat{\mu} = \underset{(10.28)}{57.37} + \underset{(0.042)}{0.52}\, F_{t-1} + s_c(\texttt{month}) - \underset{(0.005)}{0.026}t + \underset{(0.72)}{5.08}\, P_{t-1} + \underset{(0.051)}{0.25}\, soi \quad (4.5)$$

$$\ln \hat{\sigma} = \underset{(5.21)}{-9.81} - \underset{(0.019)}{0.11}\, F_{t-1} + s_c(\texttt{month}) + \underset{(0.003)}{0.005}\, t - \underset{(0.305)}{1.009}\, P_{t-1} + \underset{(0.14)}{0.29}\, P_{t-2} \quad (4.6)$$

$$\hat{\nu} = \underset{(0.077)}{0.18} + s_c(\texttt{month}) - \underset{(0.009)}{0.017}\, F_{t-1} \quad (4.7)$$

$$\ln \hat{\tau} = \underset{(11.92)}{2.50} \quad (4.8)$$

Standard errors, obtained using nonparametric bootstrapping ($B = 100$), are printed below the coefficients. As noted for the π_0 model, the bootstrap standard errors do not take between-model variation into account; a possibility at this stage is to switch to a fully Bayesian model using the **R** package **bamlss** (Umlauf et al., 2021), which provides variation measures using MCMC automatically. The term $s_c(\cdot)$ is a cyclical P-spline term. Cyclical smoothers are ideal for monthly effects since they ensure that the December and January effects connect smoothly, with a visually appealing smooth transition between months. The monthly effect as a factor with 12 levels was also available for model selection, but the smoother was chosen.

Similar to the selection strategy for the π_0 model, the selection strategy for μ, σ, ν and τ checked for interaction terms, but none was detected. The selection of an appropriate distribution, given the x-terms selected for each distribution parameter, described in STEP 4 of the Supplementary material , confirmed that the `BCTo` dis-

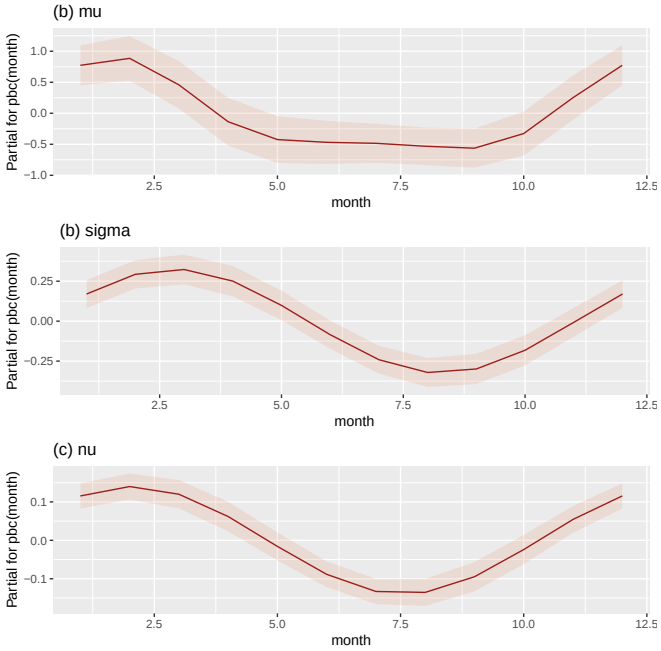

Figure 5: The cyclical P-spline smoothing terms for `month` for models for μ, σ and ν, from the model for the positive values of streamflow, with 95% confidence regions.

tribution was best for the data. The selected lagged variables for $\ln \mu$, $\ln \sigma$ and ν, similar to model for π_0, suggest a delay before changes in the explanatory variables have an impact on the positive streamflow distribution, and this delay is longer for $\ln \sigma$. The μ, σ and ν models each have a cyclical smoother for `month`, indicating that location, variation and skewness are all affected by the month of the year; these fitted terms are shown in Figure 5. Note that of the environmental explanatory variables (*soi*, *eof1* and *eof2*), only *soi* appears (in the μ equation). Also note the acf and pacf functions of the residual of the model (not shown here) do not indicate a residual auto-correlation pattern. In the next section we combine the two components of the model to create the full zero-adjusted model.

4.1.3 The full zero-adjusted model

Equations (4.4)–(4.8) define the full zero-adjusted `BCTo` model. Using STEP 6 (a) and (b) of the procedure described in Supplementary material , we fit the final chosen model. Figure 6 shows the worm plot (a detrended QQ-plot, Van Buuren and Fredriks , 2001) and the bucket plot (De Bastiani et al., 2022) from the residuals of the full zero-adjusted `BCTo` model. Note that the worm plot is good for checking departures from the distributional assumptions in general, while the bucket plot more specifically detects skewness and kurtosis. In the worm plot, one expects approximately 95% of the points (the worm) to fall within the 95% pointwise confidence interval (gray area) in an adequate model, which is the case for our zero-adjusted `BCTo` model.

The bucket plot provides a visual tool for checking whether the moment-based skewness and kurtosis are present in the residuals of a fitted model. The presence of skewness and/or kurtosis in the residuals usually reflects the fact that the assumed response distribution is inadequate. The model point (`mbct`) falls within the shaded elliptical region at the centre of the plot, which represents a 95% confidence region for a Jarque-Bera test, indicating that there is no evidence that the moment skewness and moment excess kurtosis of the residuals are different from zero. There is therefore no evidence that the zero-adjusted `BCTo` model is inadequate. (While the worm and bucket plots that we have checked show no violation of the model assumptions, for a thorough residual analysis multiple worm and multiple bucket plots should also be checked.) In conclusion it seems that the zero-adjusted `BCTo` model of Equations (4.4)–(4.8) provides an adequate fit for our data.

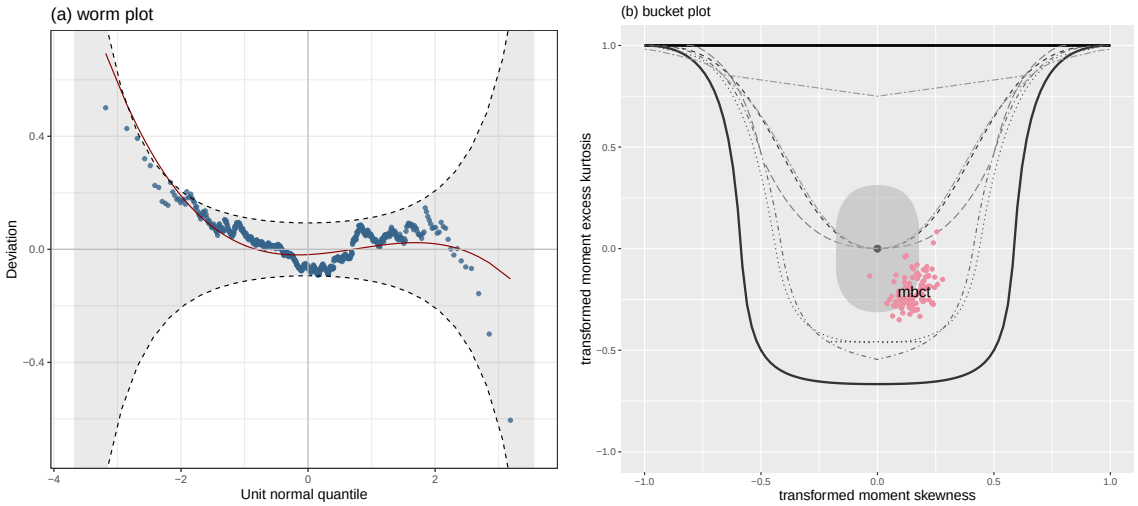

Figure 6: (a) Worm plot and (b) bucket plot from the residuals of the zero-adjusted model for streamflow.

5 Conclusions

GAMLSS is a "beyond mean regression" model (Kneib, 2013), while P-splines is a "beyond linear models" methodology. Combining the two methodologies creates a very powerful tool for the exploration and modelling of complex relationships. The simplicity and the versatility of P-splines also allows prior knowledge about the relationships to be accommodated (e.g. cyclical behaviour), which simple amendments to the original P-spline code allow.

Within GAMLSS this versatility is extended into studying other characteristics of the distribution rather than just the mean. We have demonstrated the power of the combination of these tools in a zero-adjusted GAMLSS model, with a strongly skewed distribution for the positive component (the Box-Cox t distribution), and predictors for the zero probability parameter and three of the four `BCTo` distribution parameters all contained P-spline terms, including cyclical P-splines. The advantage of any zero-

adjusted model fitted using GAMLSS and P-splines is that it simultaneously provides both an explanation of the generating data process and probabilistic predictions for future values. The procedure suggested in Supplementary material for choosing the model is rather slow, but it ensures that the appropriate variables and their interactions are selected. There is great scope for future work to improving the procedure. In our application, the modelling could be used in decision making by farmers and water authorities. Similar types of data are very common in environmental, medical and econometric studies and there is great scope for future work building on this approach for prediction.

We are indebted to Brian Marx for his contribution to statistical modelling. We are proud that we met him and that he considered us his friends. The comment he made on the eve of the short course given by Alan Agresti at the IWSM in Linz, "seeing Alan is like seeing the Rolling Stones", still makes us smile.

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, **36**, 332–339.

Aitkin, M. (2018). A History of the GLIM Statistical Package. *International Statistical Review*, **86**(2), 275–299.

Belasco, E. J., and Ghosh, S. K. (2012). Modelling semi-continuous data using mixture regression models with an application to cattle production yields. *The Journal of Agricultural Science*, **150**(1), 109-121.

De Bastiani, F., Stasinopoulos D. M., Rigby R. A., Heller G. Z. and Silva L. A.

(2022). Bucket plot: A visual tool for skewness and kurtosis comparisons. *Brazilian Journal of Probability and Statistics*, **36**(3), 421–440.

Eilers, P.H.C. and Marx, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*, **11**, 89–102.

Eilers, P.H.C. and Marx, B. D. (2010). Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 637–653.

Eilers, P.H.C. and Marx, B. D. (2021). *Practical Smoothing. The Joys of P-splines.* Cambridge University Press, United Kingdom.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* Chapman and Hall.

Heller G.Z., Stasinopoulos D.M. and Rigby R.A. (2006). The zero-adjusted inverse Gaussian distribution as a model for insurance data. *Proceedings of the 21st International Workshop on Statistical Modelling*, Galway, Ireland, 226–233.

Houghton, D. M., Nowlin, E. L., and Walker, D. (2019). From Fantasy to Reality: The Role of Fantasy Sports in Sports Betting and Online Gambling. *Journal of Public Policy and Marketing*, **38**(3), 332–353.

Kauermann, G. (2010). Penalized Splines, Mixed Models and Bayesian Ideas. In: Kneib, T. and Tutz, G. (eds) *Statistical Modelling and Regression Structures.* Physica-Verlag HD.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, **13**(4), 275–3036.

Liu, L., Shih, Y.C.T., Strawderman, R.L., Zhang, D., Johnson, B.A. and Chai, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: a review. *Statistical Science*, **34**(2), 253–279.

Neelon, B., O'Malley, A. J., and Smith, V. A. (2016a). Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview. *Statistics in Medicine*, **35**: 5070—5093.

Neelon, B., O'Malley, A. J., and Smith, V. A. (2016b) Modeling zero-modified count and semicontinuous data in health services research part 2: case studies. *Statistics in Medicine*, **35**: 5094—5112.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, **1**, 505–527.

Popuri, S.K., Neerchal, N.K., Mehta, A. (2015). Comparison of Linear and Tobit Modeling of Downscaled Daily Precipitation over the Missouri River Basin Using MIROC5. In: Lakshmanan, V. et al (eds) *Machine Learning and Data Mining Approaches to Climate Science*. Springer.

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

Rigby, R.A. and Stasinopoulos, D.M. (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statisitcal Modelling*, **6**, 209–229.

Rigby R.A. and Stasinopoulos D.M. (2012) Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research*. **23**(4), 318–332.

Rigby, R.A., Stasinopoulos D.M, Heller, G.Z., and De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Chapman & Hall/CRC, Boca Raton.

Rodríguez-Álvarez M. X. and Oviedo de la Fuente M. (2021). SOP: Generalised Additive P-Spline Regression Models Estimation. R package version 1.0.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B*, **51**, 47–60.

Stasinopoulos, D. M. and Rigby, R. A. (2008). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1–46.

Stasinopoulos, D.M., Rigby, R.A., Heller, G.Z., Voudouris, V. and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman & Hall/CRC, Boca Raton.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, **26**, 24–36.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In Ghosh, J.K.; Roy, J (eds.). *Statistics: Applications and New Directions: Proc. Indian Statistical Institute Golden Jubilee International Conference*, 579–604.

Umlauf N., Klein N., Simon T., and Zeileis A. (2021) bamlss: A Lego Toolbox for Flexible Bayesian Regression (and Beyond), *Journal of Statistical Software*, **100**(4), 1-53. https://10.18637/jss.v100.i04.

Van Buuren, S. and Fredriks, M. (2001). Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**,1259–1277.

Van Ogtrop, F.F., Vervoort, R.W., Heller, G.Z., Stasinopoulos, D.M. and Rigby, R.A. (2011). Long-range forecasting of intermittent streamflow. *Hydrology and Earth System Sciences*, **15**, 3343–3354.

Wood, S.N. and Fasiolo, M. (2017) A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, **73**(4), 1071-1081.