**INTRODUCTION**

Rapidly changing climates and landscape modification are impacting ecosystems at all micro- and macroecological levels, incurring substantial economic and environmental costs (World Health Organization 2005, United Nations 2015, Kennedy et al. 2019). There is broad consensus among scientists, parliamentarians and applied decision-makers that anticipating probable future states is vital to mitigate impacts of environmental change on ecosystems (Schmidt et al. 2010, Dietze et al. 2018, Intergovernmental Panel on Climate Change 2018). Yet two challenges impede the improvement and adoption of ecological forecasts: (1) natural systems are complex and driven by a network of interacting processes (Levin 1998, Choler et al. 2001, Massoud et al. 2018) and (2) ecological time series tend to be integer-valued random variables that exhibit complex features including zero-inflation, over-dispersion, bounding, missing values and uneven spacing of samples (Lindén and Mäntyniemi 2011, Simpson 2018, Warton 2018, Kowal and Canale 2020), making them difficult to analyse using common statistical time series models.

Generalised Additive Models (GAMs) have been a major success in many applied fields and are increasingly used in ecology (Guisan et al. 2002, Hughes et al. 2018, Simpson 2018, Pedersen et al. 2019). Outlined in detail previously (Hastie and Tibshirani 1990, Wood 2004), GAMs can briefly be described as modified Generalised Linear Models (GLMs) in which the linear predictor includes a sum of smooth functions representing functional relationships between covariates and the response:

eq. 1

where is the conditional expectation of a response assumed to be drawn from an exponential distribution, is the unknown intercept, the  are unknown smooth functions and  is an appropriate monotonic link function. Each smooth function  is composed of spline like basis expansions whose coefficients (), which must be estimated, control the smooth’s shape. The size of the basis expansion is directly linked to the smooth’s potential complexity, with a larger set of basis functions allowing greater flexibility. Several advantages of GAMs are that they can model a diversity of response families, including discrete distributions (i.e. Poisson or Negative Binomial) that accommodate ecological features such as zero-inflation, and that they can be formulated to include hierarchical smoothing for multivariate responses (Wood 2017, Pedersen et al. 2019).

Given the set of basis coefficients that comprise each smooth, a GAM can in principle be directly estimated as a GLM. However due to the incredible flexibility of smooth functions, GAMs will invariably overfit if left unconstrained (Hastie and Tibshirani 1990, Wood 2004, Marra and Wood 2011). Advances in penalised likelihood estimation avoid this overfitting by placing quadratic penalties on the basis coefficients, penalizing the function’s ‘wiggliness’ and controlling the trade-off between fit and smoothness (Wood 2004, Wood 2016). From a Bayesian perspective, the for a smooth can be drawn from a multivariate Gaussian distribution with the penalty acting on the prior precision matrix. Larger values for the penalty shrink the coefficient covariances, effectively forcing the smooth toward a straight line when the data do not justify a nonlinear function (Marra and Wood 2011, Wood 2016). GAMs are particularly sought after for modelling time series to both identify nonlinear covariate effects and to uncover periods of rapid change, though strong temporal autocorrelation can make it challenging to estimate key parameters (Yang et al. 2012, Knape 2016, Simpson 2018, Spooner et al. 2018, Camara et al. 2021).

For many ecological studies that employ GAMs, a primary objective beyond statistical inference is to obtain accurate predictions about possible future states (Ward et al. 2014, Kaplan et al. 2016, Clark et al. 2020, Malick et al. 2020, Koolhof et al. 2021). However, a lingering issue in using GAMs for forecasting is the way in which smooth functions predict outside the range of training data. Many of the smooths used in ecological GAMs have zero second derivatives at the boundaries, meaning they will linearly extrapolate beyond the last observation (Elith et al. 2010, Zurell et al. 2012). This projection of a straight line indefinitely into the future can produce very unrealistic forecasts, particularly if the estimated function wiggles slightly near the boundary (**Figure 1 top panel**). There are technical solutions to help with this problem, for example by extending the smoothing penalty into the range of values that we wish to forecast (i.e. weeks or years ahead of the training data) or by forcing the smooth to use the last observed value when forecasting with a first derivative penalty (**Figure 1 bottom panel**). However, these modifications are insufficient to generate robust ecological forecasts with appropriate probabilistic uncertainties.



Figure 1: Estimated trends and forecasts from two GAMs applied to a discrete time series. In the top panel, a thin plate regression spline with a penalised second derivative is used for the trend, leading to a smooth function (top left) and linear extrapolation when forecasting (top right). In the bottom panel, the trend penalty is placed on the first derivative, resulting in flat extrapolation when forecasting. Trend shading shows 95% confidence intervals, while forecast shading shows 95% and 68% credible intervals. Both models were fitted to a simulated seasonal discrete time series in R using the *mgcv* package with the general formula: *y ~ s(year, k = 9) + s(season, bs = ‘cc’, k = 10) + ti(season, year), family = nb()*).

In this paper we outline a Bayesian dynamic GAM that provides a general solution to the problem of estimating smooth functions while generating reliable forecasts for discrete time series. The approach is simple: for univariate series we augment the GAM linear predictor with include a latent dynamic trend to capture the series’ temporal evolution process (either as a random walk or an autoregressive process up to order 3). For modelling a collection of multivariate time series, we accommodate dependencies among series’ temporal trends in a parsimonious way using a dynamic latent factor model. We introduce our associated R package *mvgam* (https://github.com/nicholasjclark/mvgam), which provides the following key functions:

* Estimate parameters of a dynamic GAM in a Bayesian Markov Chain Monte Carlo framework via the Gibbs sampling software JAGS (Plummer 2003, Wood 2016)
* Plot estimated smooths and posterior predictions, along with their associated probabilistic uncertainties
* Calculate time series correlations
* Perform model selection using rolling window forecasts
* Perform residual diagnostic checks using randomised quantile (Dunn-Smyth) residuals (Dunn and Smyth 1996)
* Update forecasts online via a Sequential Monte Carlo particle filter

We begin by presenting our model, including background material for the dynamic factor process. We then illustrate our package’s utility for ecologists interested in forecasting discrete time series using both simulations and real-world case studies. An introduction to *mvgam* via reproducible examples is provided in the Appendix S1 (Supporting Information).

**DYNAMIC GENERALISED ADDITIVE MODELS**

**Univariate models for a single ecological time series**

A Bayesian framework involves defining a joint probability distribution over all observable and unobservable quantities in a statistical model that aligns with expert beliefs about the data generating process (Gelman et al. 2017). A dynamic GAM is naturally viewed from a Bayesian perspective, where prior beliefs about the nonlinearity of a function can be elicited to inform the complexity and penalisation of the smooth (Wood 2013, Pedersen et al. 2019), while unobserved trends are consistent with the expectation that time series evolve as serially autocorrelated dynamic processes (Hyndman and Athanasopoulos 2018). In its basic form, the dynamic GAM is written as:

eq. 2

eq. 3

where is the conditional expectation of the response at time *t* and is the estimated value for the trend at time *t*. Here the trend evolves as a random walk with possible drift, where in eq. 3 is the drift parameter and the residual error is drawn from a zero-centred Gaussian distribution. The trend model in eq. 3 is easily modified to include AR parameters. For example, the following specifies an AR2 latent trend model:

eq. 4

Note that while it is possible to model residual autocorrelation similarly in *mgcv* with a correlation structure in the *gamm* function (Wood 2017), there is no straightforward way to include this autocorrelation process in out of sample predictions. The model is coded in the JAGS probabilistic programming language using the function *mvjagam*, which relies on the *jagam* helper function from the R package *mgcv* (Wood 2016) to automatically generate a skeleton JAGS model file, the smooth penalty matrices and starting values for the unknown penalty and *ß* parameters. The model file is modified to include the latent trend and to update any prior distributions specified by the user, while any additional data objects needed for modelling are created automatically. Employing the JAGS software through the R interface *rjags* (Plummer 2003), the model is conditioned on the observed data using Markov Chain Monte Carlo (MCMC) simulation via Gibbs samplers to calculate the posterior probability distribution of the unobserved parameters.

It is notable that our design conveniently allows any formula that would be specified in *mgcv* to be used for the GAM component of the linear predictor, providing a user-friendly way to explore dynamic ecological models that encompass nonlinear smooths. This task would be extremely challenging using competing state space software packages, particularly for discrete outcomes (Petris 2010, Auger‐Méthé et al. 2021). Other major advantages of our framework are that (1) missing values are allowed for the responses; (2) upper bounds can be accommodated using truncated likelihood functions; and (3) the latent trends can easily be extended into the future via their recursive equations, providing robust probabilistic forecast distributions. We accommodate both the Poisson and Negative Binomial distributions in *mvgam* with a log link. Given the log scale of the latent trend, weakly informative Gaussian priors are specified for the AR parameters () with mean = 0 and variance = 0.1 by default. The precision of the zero-centred Gaussian prior for the trend errors is drawn from a gamma distribution (shape = 0.1, rate = 0.001). Following Wood (2016), zero-centred multivariate Gaussian distributions are used as priors for each smooth’s *ß* parameters and vague gamma priors (updated automatically by the *jagam* function) are used for the smoothing penalties.

**Dynamic factor models for a set of multivariate ecological time series**

Here we describe how we modify our dynamic GAM into a joint statistical model for collections of discrete time series. Dynamic factor models are closely aligned with static latent factor models, which are commonly used in quantitative ecology to jointly model the distributions of species (Warton et al. 2015, Thorson et al. 2016, Ovaskainen et al. 2017, Ward et al. 2021). A latent factor model is a function of unmeasured random predictors (factors) that induce correlations between responses via factor loading coefficients. Often species do demonstrate correlated responses to gradients that have not been adequately measured by the study design, meaning that a smaller set of factors than the total number of species can adequately capture the main axes of covariation (Letten et al. 2015, Warton et al. 2015). A dynamic factor model extends this reasoning by assuming the factors evolve as dynamic autoregressive processes and that each series’ latent trend is composed of a linear combination of these common factors:

eq. 5

where is the expected response for time series *j* at time *t*, the ’s are the predicted values of the *K* unobserved factors at time *t* and the ’s are factor loadings. As in the univariate case, the dynamic factors can evolve either as random walks with drift or as autoregressive processes up to order 3.

A challenge with any factor model is the need to pre-specify the number of factors *K* (Thorson et al. 2016, Tobler et al. 2019). Setting *K* too small prevents the temporal dependencies from being adequately modelled, leading to poor convergence, difficulty estimating smooth parameters and possibly inferior forecasts. By contrast, setting *K* too large leads to unnecessary computation. We approach this problem by imposing sparsity on the factor loading matrix with regularised horseshoe priors. Following Piironen and Vehtari (2017), we estimate both a global hyperprior and an individual penalty for each factor (both drawn from half Cauchy distributions), which act multiplicatively to shrink the variance of the zero-centred Gaussian distribution from which each factor’s loadings is drawn. As each factor has its own penalty, some protection against specifying too large a *K* is given by shrinking the loadings for unneeded factors toward zero. We caution however that it is worth checking whether inferences or forecasts are sensitive to *K*, perhaps using the guidelines outlined by Tobler *et al* (2019). Additional constraints are imposed on the factor loadings to preserve identifiability by setting the upper triangle of the loading matrix to zero and ensuring the diagonals are non-negative (Hui 2016, Tobler et al. 2019).

**SIMULATIONS**

**CASE STUDIES**

**DISCUSSION**

Auger‐Méthé, M., K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos‐Barajas, J. Mills Flemming, A. Nielsen, and G. Petris. 2021. A guide to state–space modeling of ecological time series. Ecological Monographs **91**:e01470.

Camara, A. J. A., G. C. Franco, V. A. Reisen, and P. Bondon. 2021. Generalized additive model for count time series: An application to quantify the impact of air pollutants on human health. Pesquisa Operacional **41**.

Choler, P., R. Michalet, and R. M. Callaway. 2001. Facilitation and competition on gradients in alpine plant communities. Ecology **82**:3295-3308.

Clark, N. J., J. T. Kerry, and C. I. Fraser. 2020. Rapid winter warming could disrupt coastal marine fish community structure. Nature Climate Change:DOI: 10.1038/s41558-41020-40838-41555.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, M. A. Kenney, C. M. Laney, and L. G. Larsen. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proceedings of the National Academy of Sciences **115**:1424-1432.

Dunn, P. K., and G. K. Smyth. 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics **5**:236-244.

Elith, J., M. Kearney, and S. Phillips. 2010. The art of modelling range‐shifting species. Methods in Ecology and Evolution **1**:330-342.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. B. Rubin. 2017. Bayesian Data Analysis. Third edition. CRC Press, Boca Raton.

Guisan, A., T. C. Edwards Jr, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling **157**:89-100.

Hastie, T. J., and R. J. Tibshirani. 1990. Generalized additive models. Taylor & Francis, New York.

Hughes, T. P., J. T. Kerry, A. H. Baird, S. R. Connolly, A. Dietzel, C. M. Eakin, S. F. Heron, A. S. Hoey, M. O. Hoogenboom, G. Liu, M. J. McWilliam, R. J. Pears, M. S. Pratchett, W. J. Skirving, J. S. Stella, and G. Torda. 2018. Global warming transforms coral reef assemblages. Nature **556**:492-496.

Hui, F. K. 2016. boral–Bayesian ordination and regression analysis of multivariate abundance data in R. Methods in Ecology and Evolution **7**:744-750.

Hyndman, R. J., and G. Athanasopoulos. 2018. Forecasting: principles and practice. OTexts.

Intergovernmental Panel on Climate Change. 2018. Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.

Kaplan, I. C., G. D. Williams, N. A. Bond, A. J. Hermann, and S. A. Siedlecki. 2016. Cloudy with a chance of sardines: forecasting sardine distributions using regional climate models. Fisheries Oceanography **25**:15-27.

Kennedy, C. M., J. R. Oakleaf, D. M. Theobald, S. Baruch‐Mordo, and J. Kiesecker. 2019. Managing the middle: A shift in conservation priorities based on the global human modification gradient. Global Change Biology **25**:811-826.

Knape, J. 2016. Decomposing trends in Swedish bird populations using generalized additive mixed models. Journal of Applied Ecology **53**:1852-1861.

Koolhof, I. S., S. M. Firestone, S. Bettiol, M. Charleston, K. B. Gibney, P. J. Neville, A. Jardine, and S. Carver. 2021. Optimising predictive modelling of Ross River virus using meteorological variables. PLoS Neglected Tropical Diseases **15**:e0009252.

Kowal, D. R., and A. Canale. 2020. Simultaneous transformation and rounding (STAR) models for integer-valued data. Electronic Journal of Statistics **14**:1744-1772.

Letten, A. D., D. A. Keith, M. G. Tozer, and F. K. Hui. 2015. Fine‐scale hydrological niche differentiation through the lens of multi‐species co‐occurrence models. Journal of Ecology **103**:1264-1275.

Levin, S. A. 1998. Ecosystems and the biosphere as complex adaptive systems. Ecosystems **1**:431-436.

Lindén, A., and S. Mäntyniemi. 2011. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology **92**:1414-1421.

Malick, M. J., S. A. Siedlecki, E. L. Norton, I. C. Kaplan, M. A. Haltuch, M. E. Hunsicker, S. L. Parker-Stetter, K. N. Marshall, A. M. Berger, and A. J. Hermann. 2020. Environmentally driven seasonal forecasts of Pacific hake distribution. Frontiers in Marine Science **7**:844.

Marra, G., and S. N. Wood. 2011. Practical variable selection for generalized additive models. Computational Statistics & Data Analysis **55**:2372-2387.

Massoud, E. C., J. Huisman, E. Benincà, M. C. Dietze, W. Bouten, and J. A. Vrugt. 2018. Probing the limits of predictability: data assimilation of chaotic dynamics in complex food webs. Ecology Letters **21**:93-103.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters **20**:561-576.

Pedersen, E. J., D. L. Miller, G. L. Simpson, and N. Ross. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ **7**:e6876.

Petris, G. 2010. An R package for dynamic linear models. Journal of Statistical Software **36**:1-16.

Piironen, J., and A. Vehtari. 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics **11**:5018-5051.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Page 125 *in* Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Technische Universit at Wien Wien, Austria.

Schmidt, K. A., S. R. Dall, and J. A. Van Gils. 2010. The ecology of information: an overview on the ecological significance of making informed decisions. Oikos **119**:304-316.

Simpson, G. L. 2018. Modelling palaeoecological time series using generalised additive models. Frontiers in Ecology and Evolution **6**:149.

Spooner, F. E., R. G. Pearson, and R. Freeman. 2018. Rapid warming is associated with population decline among terrestrial birds and mammals globally. Global Change Biology **24**:4521-4531.

Thorson, J. T., J. N. Ianelli, E. A. Larsen, L. Ries, M. D. Scheuerell, C. Szuwalski, and E. F. Zipkin. 2016. Joint dynamic species distribution models: a tool for community ordination and spatio‐temporal monitoring. Global Ecology and Biogeography **25**:1144-1158.

Tobler, M. W., M. Kéry, F. K. Hui, G. Guillera‐Arroita, P. Knaus, and T. Sattler. 2019. Joint species distribution models with species correlations and imperfect detection. Ecology **100**:e02754.

United Nations. 2015. Transforming our world: The 2030 agenda for sustainable development. UN Publishing, New York.

Ward, E. J., S. C. Anderson, M. E. Hunsicker, and M. A. Litzow. 2021. Smoothed dynamic factor analysis for identifying trends in multivariate time series. Methods in Ecology and Evolution.

Ward, E. J., E. E. Holmes, J. T. Thorson, and B. Collen. 2014. Complexity is costly: a meta‐analysis of parametric and non‐parametric methods for short‐term population forecasting. Oikos **123**:652-661.

Warton, D. I. 2018. Why you cannot transform your way out of trouble for small counts. Biometrics **74**:362-368.

Warton, D. I., F. G. Blanchet, R. B. O’Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. Hui. 2015. So many variables: joint modeling in community ecology. Trends in Ecology & Evolution **30**:766-779.

Wood, S. 2016. Just Another Gibbs Additive Modeller: Interfacing JAGS and mgcv. Journal of Statistical Software **75**:1-15.

Wood, S. 2017. Generalized additive models: an introduction with R. Second edition. CRC Press, Boco Raton.

Wood, S. N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association **99**:673-686.

Wood, S. N. 2013. On p-values for smooth components of an extended generalized additive model. Biometrika **100**:221-228.

World Health Organization. 2005. Using climate to predict infectious disease epidemics. Geneva, Switzerland.

Yang, L., G. Qin, N. Zhao, C. Wang, and G. Song. 2012. Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. BMC Medical Research Methodology **12**:1-13.

Zurell, D., J. Elith, and B. Schröder. 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. Diversity and Distributions **18**:628-634.