Associate Editor

Generalized additive models (GAMs) while commonly used to describe nonlinear relationships between response and predictor variables present challenges when forecasting outside of the range of observed data. The authors present a new R package, mvgam, that fits dynamic GAMs in a Bayesian framework that leverages information from a latent temporal process in the spirit of improving forecasts from such models. Simulations and an empirical case study using tick data from the National Ecological Observatory Network improve near-term ecological forecasts. The paper is well written and the documentation (code and vignette) are phenomenal. Very well done.

\*Response:

Thank you for the positive feedback on our work

I have two major points to improve the manuscript and some minor ones below. These points should be addressed along with both excellent reviews provided. First, it would be helpful to provide include an introductory table or paragraph that give more context for how mvgam extends existing software and models. For instance, how does mvgam extend the BayesX package (Brezger et al. 2005) and others?

\*Response: We appreciate this suggestion and have added a paragraph to compare and contrast our software with existing options for fitting time GAMs:

“mvgam extends functions available in existing software packages in several ways. First, while fully Bayesian GAMs can be estimated using a variety of software including brms (Bürkner 2017), BayesX (Brezger et al. 2005), and bamlss (Umlauf et al. 2018), mvgam is the only software we are aware of that can simultaneously estimate any smooth function available in mgcv together with latent dynamic trends (bamlss and BayesX can estimate a diversity of smooth functions but to our knowledge dynamic latent processes cannot be jointly estimated; brms offers more flexibility for time series and can accommodate dynamic processes, including AR and ARMA processes, but we are not aware of extensions to dynamic factors). Second, our software can employ Hamiltonian Monte Carlo through the Stan interface for much more efficient and unbiased MCMC sampling compared to Gibbs samplers (BayesX uses its own custom Gibbs samplers, while bamlss does not employ full MCMC). Finally, our package is designed for analysing and forecasting sets of discrete time series, and as such the additional utilities we offer for working with time series (including options to compare models using rolling forecast evaluation as well as routines to assimilate new observations ‘online’ for automatic forecast updating; Appendix 1) make our software attractive for a range of applied forecasting tasks.” (Lines 301-316)

Second, it would be helpful to replace the air passenger example in the vignette described in Appendix 1 with a more ecological example. I found it confusing that species were referred to in this example and think that it would be clearer to the readers of MEE to have an ecological example instead.

\*Response: Thank you for this suggestion. We have updated the Appendix 1 example to now include both a simulation component (to illustrate the disadvantages of an autoregressive observation model for overdispersed discrete series) and an empirical example where the task is to forecast landings of commercial fishes. We agree this set of examples is more approachable to the readers of MEE

Third, in the case study (line 305), I am not sure if having a start of tick season of June 1 for all sites is appropriate as the predictor. It would be very helpful to have a map of the NEON sites used in this analysis. For instance, I wondered if the June 1 start of tick season was appropriate but couldn’t tell without knowing what sites were in the analysis while reading just the main text. My understanding is that the onset of sampling for ticks at NEON sites is based on green up at each site, so I’m not sure if choosing June 1 as the start of tick season for the predictor is appropriate for all sites. It may be that this needs to be calculated based on the timing of tick sampling starting at each individual site.

\*Response: We have changed the NEON analysis considering reviewer and editor comments. We now estimate the seasonal smooth function across the entire year (52 epidemiological weeks), which avoids the need to arbitrarily choose when the ‘start’ of the season should be and showcases how flexible our Bayesian framework is for handling missing observations. Note how the uncertainties during the winter periods in our estimated seasonal smooth functions increase in Figures 5 and 6 (as they should, consider that no sampling has taken place during those periods). We have also added further descriptions of the NEON sampling protocol and refer to the primary reference that includes detailed maps of the sampling locations.

Minor points:

- Line 25: Could a term other than bounding be used? This came across as jargon.

\*Response: We have replaced this with the term ‘truncation’

- Line 33: Here and throughout the authors refer to smooths. Would it be more appropriate/formal to refer to smooths as smoothing functions? I found the term ‘smooths’ to be awkward.

\*Response: We have replaced this term with the term ‘smooth functions’ throughout

- Line 73: the authors refer to contextual information. Please elaborate here as it is not clear what contextual information could refer to.

\*Response: We have updated this line to read:

“Moreover, ecological observations are almost always multivariate when contextual information, such as data from environmental predictors or observations of non-target species, is considered.”

- Line 93: What is meant by, ‘the size of the basis expansion…’? Does this mean the number of basis functions? Is there a way to rephrase this to be clearer to the reader?

\*Response: We have replaced that phrase to now read:

“the total number of basis functions”

- Line 120: Is it necessary to include, ‘…have zero second derivatives at the boundaries…’? I thought the description following that was clearer, “… they will linearly extrapolate beyond the last observation.’

\*Response: We consider this information necessary as it provides the technical explanation for why a smooth function with a penalty on the second derivative will extrapolate linearly

- Lines 143-153: I think it would be very valuable to readers if you could list parenthetically the names of functions in mgcv that perform these calculations.

\*Response: Thank you for the suggestion. We have added the relevant function names to each bullet point in that passage

- Line 179: How realistic is it ecologically that the standard deviation would be time-invariant? A bit of discussion here about this assumption, why it is made, and its limitations would be helpful.

\*Response: This is a simplifying assumption that will often be necessary as it will become very challenging to simultaneously estimate non-constant volatility and overdispersion processes for discrete series. However, we recognize that users may wish to opt for more complex temporal processes, and so we have added the following to this passage:

“The assumption of a fixed standard deviation for the temporal process error could potentially be a limitation if the series of interest displays non-constant volatility with perturbations that may be evidence of responses to ‘shocks’. While the mvgam package does not currently have an option to include stochastic volatility or moving average trends, these processes could be added by the user at any time (the package can be used to generate all model files, data objects and initial values so that a model can be easily modified for conditioning outside of mvgam, i.e. with rstan, CmdstanR or rjags directly).” (Line 200)

- Line 221: Check noun-verb agreement on loadings. It should be loadings need rather than loadings needs.

\*Response: Changed

- Lin 260: The difference between moderate and strong temporal dynamics is not clear to me. Is there another way to describe those dynamics, or an example or picture you could provide to clarify?

\*Response: We have added a new Figure S1 to demonstrate two series with the same seasonal dynamics but with different strengths of trend dynamics to clarify this description

- Line 275: check spelling of implementation

\*Response: Changed

- Lines 277-278: How was model convergence assessed?

\*Response: We have the following to describe convergence checks:

“Convergence of chains was checked with the Gelman-Rubin diagnostic (Gelman and Rubin 1992) and by visual inspection of posterior chains.”

- Line 341: What do you mean by strong dynamics? Is there another way to describe strong dynamics?

\*Response: We explain more clearly what we mean by strength of dynamics, and we provide a new supplementary figure to visualize two series with identical seasonalities but different relative strengths of underlying trend:

“… magnitude of the temporal component relative to seasonality (0.3, for moderate dynamics, or 0.7 for strong dynamics; see Figure S1 for an example of two series with the same seasonality but different strengths of trend).”

- Lines 415-418: What were the run times of the examples given? What are some of the issues you have come across with posterior estimation? These details would be very useful to readers considering using the package.

\*Response: This is a very good suggestion. We have added a paragraph in the Discussion entitled “Challenges in estimating DGAM parameters” to highlight some of the difficulties that can arise, how we set up priors to adapt to series on different observation scales, observed run times to reach suitable effective sample sizes and challenges with joint estimation of overdispersion and autocorrelation processes.

- Lines 449-452: What is the DOI of the NEON data used in these analyses?

\*Response: We have added the correct citation for the NEON data. This line now reads:

“We fit species-specific DGAMs to four years of data (2015 – 2018) for 17 A. americanum plots (nested in 7 NEON sites) and for eight I. scapularis plots (nested in three sites) using the most recent release of the NEON tick drag sampling product (National Ecological Observatory Network 2022)”

Citation: National Ecological Observatory Network. 2022. Ticks sampled using drag cloths (DP1.10093.001). National Ecological Observatory Network (NEON). Dataset accessed from https://data.neonscience.org on Feb 1, 2022.

- Figure 1, 3, and 4: Some of the y-axis labels that have s() in them are not very intuitive. Can these axis labels be renamed to be more descriptive?

\*Response: We originally used the s() notation of the *mgcv* package so that users who are familiar with that package could readily interpret them. However we acknowledge that this notation is not as descriptive as it could be, so we have elected to follow the examples in Gavin Simpson’s *gratia* package (https://gavinsimpson.github.io/gratia/) and now use the y-axis label ‘Partial effect’

Reviewer: 1

The authors introduced Dynamic Generalized Additive Models (DGAMs), taking advantage of the Bayesian latent factor approach to improving regular Generalized Additive Models (GAMs). Compared to the GAMs (eq. 1), the DGAMs’ innovation lies in the additional z term in the latent state (eq. 2), where z can take a random walk (eq. 3), an autoregressive process (eq. 4), or a combination of factors (eq. 5). The main benefit is to allow structural inference on time, thus capturing the dynamic nature of time series data and offering better forecasting capabilities. Computationally, DGAMs are constructed as Bayesian hierarchical models that can be implemented using Markov chain Monte Carlo, particularly the Gibbs sampling software JAGS. The authors developed an R package that translates model specifications into JAGS code, with additional model fitting and diagnostics tools. They illustrated DGAMs with both simulation data and actual data. Overall, the method is clearly described, and the manuscript is well written. It is conducive for the authors to provide the tutorial for the package. I only have some minor comments.

First, at the high level, it might be helpful to situate DGAMs in different goals of time series analysis. Time-series approaches can be used for multiple purposes, such as filtering, forecasting, and smoothing. It appears that DGAMs primarily help the forecasting. Do they have co-benefits or trade-offs with other goals?

\*Response: We have added text to describe how GAMs and DGAMs can in fact be used to approach each of the major goals in time series analysis:

“GAMs are particularly sought after for modelling time series to identify nonlinear or time-varying covariate effects, perform smoothing of historical time series and 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).” (Line 114)

“Because the z\_t are unobserved latent variables they will continue to evolve, even when an observation Y\_t is missing, via dynamic equations that conveniently provide recursive expressions for h-step ahead prediction, historical filtering and updating of forecasts (Durbin and Koopman 2012).” (Line 177)

“Update forecasts online via a Sequential Monte Carlo particle filter using functions pfilter\_mvgam\_init() and pflter\_mvgam\_online()” (Line 295)

Second, the package can show the JAGS code, which is an important feature that makes the model transparent and allows users to modify it. It is only introduced in the supplements, and I suggest highlighting it in the main text. Also, a graphical model visualization can be an added feature that further aid communication.

\*Response: We do highlight that ability to produce annotated model code and all necessary data structures / initial value functions to condition the model outside of *mvgam* in the Discussion. We have also added a new Appendix\_S5 which shows a model file for one of the NEON empirical models:

“Notably, JAGS or Stan model files, together with all data necessary to condition the model, are made available to the user in mvgam, allowing an enormous diversity of bespoke models to be implemented through addition of other stochastic or hierarchical elements.” (Line 527)

“an example JAGS model file complete with automatic descriptions of required data structures is shown in Appendix S5” (Line 533)

We would have to see some examples of how such model configurations can be visualized before deciding whether they are worth pursuing as a feature. In our experience, mathematical notation of models (like we display in eq. 5 for example) is the most useful means of communicating model configurations using a universal language. In addition, Stan code is not designed to construct a graphical model per se, so this utility may not apply consistently to our package. Perhaps the reviewer could point out some existing examples to illustrate the point more clearly?

Third, the two examples work well with DGAMs, but the authors could discuss the limitations or the applicabilities of DGAMs. Under what circumstances do they not work well? What ecologists should explore new data to decide whether DGAMs is a good choice? The package provides some diagnostics tools–it would be helpful to show how to use those tools to evaluate models, especially the data that is not suitable for DGAMs.

\*Response: We are careful throughout to state that our package is designed for working with discrete, integer-valued time series, i.e.:

“We have introduced an R package for fitting Bayesian Dynamic GAMs (DGAM) that incorporate the flexibility of the widely popular penalised smoothing functions in mgcv with latent dynamic components for analysing and forecasting discrete time series” (Line 516)

We also highlight in some of our Appendix examples how our diagnostic functions can be used to highlight problems, such as autocorrelation of residuals or poor behaviour of Quantile-Quantile plots. See https://rpubs.com/NickClark47/mvgam3 for examples.

However, it is difficult to make heuristic recommendations about what models to pursue or not pursue, as choosing a modelling approach is a very bespoke task that will vary from analysis to analysis. Certainly the construction of some general checking functions, such as those that we illustrate with our residual diagnostics and posterior predictive check examples, can help modelers think about what missing systematic heterogeneities need to be included in their model, but invested users should be open to devising their own checking and diagnostic functions to ensure the model is generating data that aligns with their specific goals and objectives.

Last, the manuscript focuses almost exclusively on the latent state (E(y)) and somehow skims over the data model (y ~ distribution). Practically, ecologists will have the first question to choose a data model, with the main choices of Poisson and negative binomial. What are the implications of different data models? Are they the only choices? It would be interesting to explore or point the readers to other references.

\*Response: While it is not in our scope to discuss advantages and disadvantages of different observation models in this work, we agree that choice of observation process is important for working with discrete time series. We have added a small section in our new Discussion paragraph to highlight this point:

“One situation that we have encountered is the difficulty in jointly estimating a latent trend and overdispersion parameters such as in the Negative Binomial or Tweedie distributions. This is because both processes (overdispersion and autocorrelation) may be able to explain the dispersion around the mean, particularly when using Random Walk or AR trends that can jump around easily. Users will need to use theory and judgement to decide how to tackle these challenges, for example by assuming there is overdispersion in the observation process (with consultation from appropriate references; i.e. Lindén and Mäntyniemi 2011) but that the trend is smooth, in which case a latent Gaussian Process with suitable length scale would be appropriate.” (Line 569)

Specific comments

L68-69: ecosystem time series data are not integer-valued.

\*Response: Could the reviewer please be more specific? We recognize that not all ecological time series are integer valued, but it is not clear what an ‘ecosystem time series’ is

L303: NOAA.

\*Response: Changed

L314 and many occurrences: it is understandable to use R notations for model formulas, but it may be more beneficial to use mathematical notations to make it generalizable.

\*Response: The standard notation for each of our fitted models would be virtually identical to equation 2, though the exact covariates that comprise each smooth function s() will differ. We feel that writing out all the formulas would be redundant, but are happy to take advice from editors and technical editors on this point

L328: negative binomial distribution has two parameters–does this parameterization only include the expectation?

\*Response: We have updated this line to explain that we use the overdispersion parameterization (two parameter distribution) for the Negative Binomial observations:

“We used random walk dynamic factor models (M = 4 for *Ixodes* and M for *Amblyomma*) for the temporal evolution and assumed a Negative Binomial distribution for the observations, with complexity-penalising priors placed on the Negative Binomial overdispersion parameters (Simpson et al. 2017).”

L363: explain PIT histograms.

\*Response: We have changed this line to be clearer about what a PIT histogram is, and we have added a reference for interested readers to view:

“Inspection of Probability Integral Transform (PIT) histograms, which should be uniform if predictions are evenly distributed about the truth (Simonis et al. 2021), revealed that all models tended to underpredict to some degree (left-skewed PIT histograms; Figure S5).”

Reviewer: 2

Apologies for the delay in returning my review; I had intended to ask for an extension prior to accepting the review but having been asked to do two reviews in a short window I confused them and only ended up asking for one extension --- unfortunately not yours. Also, the document provided for review was 150+ pages (even if some of that was the preprint) and going through it all took longer then the typical paper. The manuscript by Clark & Wells describes a new approach to modelling and forecasting ecological data, which combines elements of generalized additive models (especially penalised splines) with latent stochastic trend components, to provide for better out-of-sample (or n-ahead) forecasts than are possible via extrapolation of trends modelled as a penalized spline. The main conceptual advance here is in the specification of latent stochastic trends in place of a penalized spline for the long-term trend in the model. Penalized splines are still used for seasonal or other (quasi-) cyclic variation in the data. The stochastic trend components available in the accompanying software are a random walk or low-order autoregressive processes, and fitting is performed within the JAGS probabilistic programming language and thus in typically fully Bayesian. The manuscript is well-written, and nicely illustrated with plenty of relevant examples that show of the potential of the DGAM approach. The topic and software are of interest to the readers of MEE and I believe that the work represents a novel and interesting contribution to the field, which especially allows quite flexible models for forecasting, an areas where GAMs have traditionally been less-widely used because of the obvious problems of extrapolating beyond the data from a flexible spline. The manuscript provides comparisons with models that can be fitted with {mgcv}, arguably the most widely-used and capable software for fitting GAMs used by ecologists. It is with these comparisons that I find most to disagree or argue about and where I feel the manuscript and paper could be improved. I also have a few concerns about the models fitted to the tick data.

I really enjoyed reading the manuscript; partly why it's taken me longer than I expected to review it is that it has got me thinking about the kinds of models envisaged by the authors, how we might do similar things with \*mgcv\* or extend the authors approach to a more functional setting. Good papers stimulate discussion and new ideas and this manuscript has certainly done that.

\*Response: Thank you for the positive feedback on our work

Another area where it shines is the wealth of illustrative material supplied as appendices that are essentially vignettes or workbooks complete with rendered code and outputs that are easy to follow and show off many of the aspects of the software that for obvious reasons couldn't be shown in the main manuscript. These will be of significant importance for those readers of the paper that will inevitably begin to use the models described here by Clark & Wells. If I had one quibble, sometimes the authors make statements or decisions in these vignettes that are not well / even / clearly explained, to this reader at least. It might help if the authors have a friendly colleague or two they could run these past to identify the places where what is done or why it is done is not clear. I apologise for not doing this myself; I read the appendices on paper and didn't go through and markup where I saw these issues and it would be too time consuming to reread all of this now and mark up where I encountered issues. That said, these are only quibbles, not a glaring problem that must be fixed.

\*Response: We appreciate the critiques. As these are open-access tutorials, we plan to continually improve them to both showcase the package’s utility and to maintain a comprehensive set of tests. In doing so, we expect the explanations to improve over time.

It is unclear at times if the authors are describing/discussing the general idea of a DGAM or their software \*mvgam\*. I see no theoretical reasons that the stochastic components are limited to a random walk or AR(p) processes of order $p \in {1, 2, 3}$ for example. This tension is there, lingering, throughout the manuscript. I would suggest trying to separate as best as possible the concept of a GAM with a latent stochastic process and \*mvgam\*.

\*Response: Thank you for the insight. This is a good point and we have used it to motivate a restructuring of the manuscript. We now introduce DGAMs separately from our software, highlighting how they can be estimated and their advantages of autoregressive observation models. We then introduce mvgam as a software tool for fitting DGAMs, which reads far better than the previous version.

What is the advantage of the latent stochastic process over say inclusion of a stochastic trend via inclusion of autoregressive terms directly into the linear predictor, via addition of $y\_{t-1}$, $y\_{t-2}$ etc? I can see that perhaps latent stochastic processes might have extra utility for multivariate models, where (here I'm assuming you mean "multivariate" to mean "multiple series") --- though even then I could see how to fit those within mgcv without needing the latent processes --- but what about the univariate case? In proposing a new, much more complex modelling approach, I feel that there is an onus in those proposing the methodology to demonstrate clearly why their method is better and this is area where the manuscript currently doesn't do as good a job as it could.

\*Response: This is a good point and it required more detailed explanation. We have added a paragraph to describe the benefits of the latent trend (state-space) approach:

“where E(Y\_t) is the conditional expectation of the response at time t and z\_t is the (latent) dynamic process estimate at time t. Readers familiar with state-space models will recognise the benefits of separating the temporal and observation processes (Auger‐Méthé et al. 2021, Heilman et al. 2022), but it is worth clarifying these advantages explicitly. First, estimating the trend as a dynamic random variable avoids problems that can occur in competing autoregressive observation models where measurement error or outliers can have large influences on estimated AR parameters and cause highly unstable forecasts (see an example in Appendix 1). Second, it is far easier to handle missing or irregularly sampled observation using latent processes. Because the z\_t are unobserved latent variables they will continue to evolve, even when an observation Y\_t is missing, via dynamic equations that conveniently provide recursive expressions for h-step ahead prediction, historical filtering and updating of forecasts (Durbin and Koopman 2012). In contrast, a missing observation in an AR3 observation model will result in NAs for four rows of the design matrix (one missing Y\_t and three missing AR predictors) that can make parameter estimation difficult for software that automatically excludes rows with missing values (such as commonly used linear modelling methods in R, for example). Other advantages of a state-space form are that trend dynamics provide a probabilistic model for the temporal evolution of a process, which can often be more useful than a smoothed trend (such as a penalised spline) in that it facilitates simulation and comparison with other processes, that new observations can be assimilated to adapt a forecast distribution via recursive Kalman or particle filtering (Massoud et al. 2018) and that multiple observation processes can depend on shared latent processes (Ward et al. 2021).” (Lines 169-190)

The comparisons are with models where the trend is fitted as a spline --- I am guilty of fitting models like this and of publishing or writing about examples where we make all continuous terms in a model be smooth functions. However much of my work is not centred on forecasting, and as a result making everything smooth is a reasonable approach to take. If a principle aim of a study was forecasting, I would likely look at models with stochastic trends. I appreciate you might want to highlight that forecasting with splines is not a good idea, but this point has been made many times in the literature and as such it represents a weak argument to focus the manuscript around.

\*Response: We appreciate there is a sizeable literature to demonstrate the poor extrapolation behaviour of splines, but it is our experience that many ecologists using splines are not aware of these limitations. Indeed, even scientists working at the World Health Organisation were recently forced to publicly revise death toll forecasts when they realized that their models (which relied on extrapolating a thin plate regression spline that wiggled excessively near the boundary) were providing nonsensical forecasts: “WHO group used a mathematical function called a thin-plate spline to estimate expected deaths for 2020–21. Unfortunately, commentators on Twitter noted, this function seemed to be too sensitive to a slight dip in Germany’s deaths in 2019; it predicted a decline in deaths in 2020 and 2021, as well (see ‘The German puzzle’)” (<https://www.nature.com/articles/d41586-022-01526-0>). However, we have taken the reviewer’s comments on board to provide a more thorough comparison of DGAMs to other types of time series models that users may be tempted to fit to discrete ecological series (please see detailed responses below), and we feel these additions have improved the manuscript.

So, I would like to see a comparison with a model of the form gam(y ~ s(season) + y\_lag1, data = df). for example. (I appreciate that getting \*n\*-ahead forecasts for \*n\* > 1 is going to be more problematic now as you will have to iteratively generate the predictions plugging in $\hat{y}\_{T+n}$ for `y\_lag1` when predicting for $\hat{y}\_{T+n+1}$). This could be extended to higher order lagged effects for example.

\*Response: We appreciate the suggestion of a more apples-to-apples comparison. We have added an autoregressive GAM as an additional model configuration for comparison in the simulations, using log(Yt-1) as a parametric predictor in place of the smooth trend. We have also added a paragraph to describe the potential disadvantages of this formulation (see our response above) and we have included a more detailed simulation experiment where we provide reproducible code for fitting and forecasting from AR3 observation models using mgcv in our new version of Appendix 1 (also found here: https://rpubs.com/NickClark47/mvgam). Therein we find that using raw versions of lagged observations leads to highly unstable forecasts when the observation model includes overdispersion, which is not surprising given the high sensitivity of AR parameters to measurement error / outliers. When using logged versions of lagged observations, the forecast behaviour is much more sensible but still inferior to DGAMs with latent trends. Our new simulations in the manuscript confirm this pattern, with DGAMs continuing to provide better forecast scores and 90% interval coverage across nearly all comparisons. We also encountered problems in simulations that highlight another drawback of the autoregressive observation model, the issue of missing AR predictors when an observation is missing. As we state in the revised manuscript:

“Note however that because each missing observation results in additional missing rows in the design matrix (due to missing values in AR predictors), we were unable to fit this model for the simulations where 50% of observations were missing.” (Line 357)

In the simulations section of the paper, the authors fit a GAM with s(year, by = series, m = 1, k = 3, bs = "gp") and I have several issues with this as a good model for the long term trend in the data. Firstly, why `k = 3`? The data simulation seems to allow for far more complex long-term trends than could possibly be fitted using a 3 DF spline. Why a Gaussian process spline? This is probably one of the least used smooths in \*mgcv\*'s arsenal of basis types and the convergence of GPs with penalized smooths only exists if parameters of the GP smooth are fixed, so it is of much more limited use in practice than using a thin plate regression spline for example. In most examples there is little to choose between a TPRS and a properly fitted GP smooth \*in \*\*mgcv\*\*\* (and I'll argue below I don't think this smooth has been properly fitted). Without justification, this comes across as arbitrary. Additionally, this formulation of a GP smooth requires the modeller to specify the length scale of the correlation functions that are the basis functions. This is required because, as mentioned above, the convergence of GPs with penalised splines happens when the basis functions do not depend on other parameters of the model. This can only happen here is $\rho$, the length scale parameter of the correlation functions (used as basis functions) is fixed and specified \*a priori\*. With `m = 1`, the authors are specifying that the correlations functions of the GP smooth are spherical correlations functions, but because they do not specify the length scale here as element `m[2]`, the defaults of the GP smooth kick in. These defaults are not suitable for time series; the value of $\rho$ used is that suggested by Kammann & Wand (1003), namely, $\rho = \text{max}\_{ij} ||\mathbf{x}\_i - \mathbf{x}\_j||$, where $\rho$ is taking the distance between the maximally-separated pair of observations. In a time series this means the first and last time point, and thence you are forcing the smooth to model long-term autocorrelation. I've rarely encountered a situation using this kind of smooth where using the default is a useful thing to do for these smooths. Ideally, in this situation, one would choose $\rho$ by profiling the REML or ML score over a grid of values for $\rho$ by fitting the model at these values of $\rho$ and finding where the REML or ML score is minimised. Given the above, there needs to be some more thought put into the \*mgcv\* model being fitted --- or, if all of this really was intentional, some justification for \*why\* this approach was done. Its hard to see past deficiencies in the model fitted when critiquing the GAM approach in general.

\*Response: Thank you for that information, we were not fully aware of how the GP basis worked in mgcv when the last version of the manuscript was written but have a deeper understanding of the mechanisms now. As you will see in our response to (8) below, we have replaced the GP smooth trends in the simulations with hierarchical TRPS smooth functions to address this point and the point raised about the potential advantages of hierarchical trends. Results have not changed substantially but we feel the simulations are fairer and more rigorous than they were previously.

For the NEON tick abundance, is a cyclic spline appropriate? Is the assumption here that the counts are exactly 0 from week 40 in year \*y\* to week 15 in year \*y+1\*? Otherwise you are forcing the abundance in week 15 to be equal to that of week 40? I'll admit to not fully understanding how you can possibly create something like figure 5 (lower left panel) with no gaps in the predicted time series when you only fit the seasonal component to just 26 of the possible weeks of a calendar year - the axes imply the standard interpretation of a 52 week year. At the very least there is some slight of hand happening here that I find hard to justify. Likewise, why should the seasonal smooths need 26 degrees of freedom, 1 per epidemiological week! What is the biological justification for allowing something so complex as the estimated seasonal effects where there seem to be periods of increase (decrease) in abundance punctuated by short periods of stasis or slower change. Is this biologically plausible? Why would abundance in week 14 on average decline only to ion average increase again in week 15? Likewise the more pronounced minimum around week 20 that is followed by a subsequent increase in abundance that is then followed by a further decline. Absence anything to the contrary this seasonal smooth looks over-fitted to me. The same features are observed in Figure 6. Are you just fitting short-term autocorrelation here? Is that a good thing? How is the over-specification of complexity (IMHO) on these seasonal smooths affecting estimates of the long-term trends. It's hard to ignore the possibility that these smooths aren't effectively interpolating the weekly observations here. These modelling decisions need to be better justified, and if they can't be they should be changed.

\*Response: We have combined two comments that both centered on our choice of data selection / smooth function for seasonality in the NEON data. We agree that arbitrarily setting threshold weeks was not necessary and was confusing. We have now included the full time series for NEON examples (all 52 epidemiological weeks) and used a less complex smooth function for cyclic seasonality (k = 12) to force the seasonal pattern to take on the kind of smoothness that is compatible with our prior knowledge on tick population dynamics:

“we assumed the global seasonal pattern was moderately nonlinear and flexible enough to capture the characteristic double peaks commonly seen in hard tick nymph abundance survey time series (Wallace et al. 2019), and we assumed the seasonal function was cyclic with equal values between the end of December and the beginning of January.” (Line 427)

Note how the uncertainties during the winter periods in our estimated seasonal smooth functions increase in Figures 5 and 6 (as they should, considering that no sampling has taken place during those periods). However, we also caution in the manuscript that our goal was to use these empirical NEON data to highlight the package and how it could be used to facilitate model selection, not to carry out a rigorous analysis of these data:

“For each species we fit four models representing different hypothetical dynamics, though we caution that our goal here was not to carry out a rigorous analysis but to highlight how DGAMs could be used to facilitate model selection and scrutiny” (Line 415)

Why `k = 3` for the ``cum\_gdd` smooth; given the sum-to-zero constraint on the smooth this means this term is fitting an essentially quadratic effect. Even if this is the right effect, why fix `k` to be slow low; we typically advocate setting `k` to some value somewhat higher than the actual value you might think or hypothesise to be the correct wigliness.

\*Response: We have increased k to 5 for these smooth functions and have added a justification for our decision:

“Null: there is no seasonality, rather the latent factors / random site-level effects of cum\_gdd fully influence the dynamics for the plot-level series. We hypothesised that the site-specific partial effects of cum\_gdd could be mildly nonlinear, so we set k = 5 for this smooth function. Formula in R syntax: y ~ s(site, bs = ‘re’) + s(cum\_gdd, site, k = 5, bs = ‘fs’) + Z”

## Minor comments. The following are more minor comments as I came to them in the paper

1. Throughout, the method is typically known as a "generalized additive model" (note the "z" not the "s" in the spelling). Consider going with convention and spell GAM that way and change your DGAM to match.

\*Response: Changed throughout

2. L 91 "spline like"? Why? These either are or are not spline basis functions.

\*Response: We removed this confusing terminology to now read:

” is composed of basis functions whose coefficients…”

3. L 125 extending the penalty into the period which one wants to forecast only provides a solution to the clearly-too-narrow credible intervals on the forecasts. As my blog post showed, for a B spline, the extrapolation behaviour for the smooth (B spline with 2nd order derivative penalty) is linear whether the penalty is extended into the forecast period or not. What is different is that when the penalty is extended into the forecast window, the uncertainty about the estimate is larger. But even then it is not noticeably so, and the uncertainty of the TPRS (the default smooth in mgcv) is also quite wide in the forecasting period. Putting the penalty forecasts a constant value for the TPRS smooth; for the B spline we get slightly more complex behaviour than simply repeating the last observation. And combining multiple penalties with the B spline produces even more complex behaviour again. Fore example, combining 0th, 1st, and 2nd order derivative penalties for a B spline "trend" produces reasonable extrapolation behaviour \*for the example data\* in the absence of new information even if the forecasts for the example aren't great, most of the forecasted observations are within the credible interval of the estimated smooth. This sentence needs to be tightened to better reflect behaviour in practice or perhaps some rephrasing and rewording is all that is required.

\*Response: Thank you for the suggestion. We have reworded that line in question to now read:

“There are technical solutions to help with this problem, for example by extending the evaluation of the penalty into the range of values that we wish to forecast (i.e. weeks or years ahead of the training data) to ensure model uncertainty grows in a more realistic fashion out of sample, or by forcing the smooth to use the last observed value (with fixed uncertainty) when forecasting by imposing a first derivative penalty (Figure 1 bottom).”

4. L 137 I presume the limitation here is simply an implementation one and not any constraint on the method itself?

\*Response: Yes this is correct. The method itself (separating the temporal and observation processes) has parallels in state-space modelling and is advantageous for time series modelling.

5. L 147 what are "multivariate time series correlations"? Among-series correlations?

\*Response: We have changed this line to improve clarity:

“Compute correlations among latent trends for multivariate sets of series using function lv\_correlations()”

6. L 263 "hierarchical gam" -> "hierarchical GAM"

\*Response: Changed

7. L 265 using `m = 2` should be fine in the `fs` smooth because it is fully penalized

\*Response: We tested both m=1 and m=2 and results did not change, so we elect to keep m=1 to demonstrate how the derivative penalty can be specified by the user

8. L 266 why are you being \*random\* when it come to the seasonal smooths but "fixed" for the trend terms? It's a bit counter intuitive to me to treat these differently? Was the aim to allow for different smoothing parameters for the trends for each series but assume that the seasonal terms would share a smoothing parameter across series because these seasonal effects should be expected to be of about the same wiggliness? Do you expect --- or observe --- that the trends for each series simulated have different wigglinesses? I can imagine the opposite, then on average each trend has about the same wigglines given how you simulated them, but that they have very different shapes. You could achieve this by using the `fs` basis for the trends too. Did you try this? Does it make any difference if you do?

\*Response: Thank you for the suggestion. We have modified our simulations to both include an autoregressive observation model and a model that uses hierarchical trends. This does not change conclusions very much but provides a more complete comparison of different modelling configurations

9. L 284 "coverages" -> "coverage"

\*Response: Changed

10. L 298 consider "irregular sample," -> "irregular sampling in time,"

\*Response: Changed

11. L 360 Am I misreading the plots? I see 0.89, 0.97, 0.97, and 0.98?

\*Response: We have updated these numbers after re-running analyses and the text now matches the figure

12. Appendices; why do the outputs from your package functions always render differently to the output from other functions. You must be doing something different in your `print()` methods that is causing this...

\*Response: We cannot see any difference. Do these discrepancies show up in the html versions as well?