Received: 5 June 2021 | Accepted: 24 November 2021

DOI: 10.1111/2041-210X.13788

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Smoothed dynamic factor analysis for identifying trends in multivariate time series

Eric J. Ward[1] | Sean C. Anderson[2] | Mary E. Hunsicker[3] | Michael A. Litzow[4]

[1]Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, USA

[2]Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, BC, Canada

[3]Fish Ecology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, USA

[4]Shellfish Assessment Program, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Kodiak, AK, USA

Correspondence
Eric J. Ward
Email: eric.ward@noaa.gov

Handling Editor: José Miguel Ponciano

Abstract

1. Ecological processes are rarely directly observable, and inference often relies on estimating hidden or latent processes. State-space models have become widely used for this task because of their ability to simultaneously estimate the multiple sources of variation (natural variability and variance attributed to observation errors). For multivariate time series, a second aim is often dimension reduction, or estimating a number of latent processes that are smaller than the number of observed time series. Dynamic factor analysis (DFA) has been used for performing time-series dimension reduction, where latent processes are modelled as random walks. Whereas this may be suitable for some situations, random walks may be too flexible for other cases.
2. Here, we introduce a new class of models, where latent processes are modelled as smooth functions (basis splines, penalized splines or Gaussian process models). We implement these models in our BAYESDFA R package, which uses the RSTAN package for fitting. After evaluating model performance with simulated data, we apply conventional models and our smooth trend models to two long-term datasets from the west coast of the United States: (a) a 35-year dataset of pelagic juvenile rockfishes and (b) a 39-year dataset of fisheries catches.
3. Our simulations demonstrate that models matching the underlying trend smoothness make better out-of-sample predictions, but this advantage diminishes with increasing levels of observation error. For both case studies, the best smooth trend models had higher predictive accuracy, and yielded more precise predictions, compared to the conventional approach.
4. The smooth trend factor models introduced here offer a new approach for state-space dimension reduction of multivariate time series. These flexible Bayesian models may be particularly useful for data that are clumped in time, for data with high signal to noise ratios and generally for data where the underlying trend is assumed to be relatively smooth.

KEYWORDS
Bayesian modelling, B-spline, dynamic factor analysis, Gaussian process, smooth spline, Stan

1 | INTRODUCTION

Ecological data can be characterized by multiple sources of variability, including stochastic natural variation, and errors associated with data collection (observation, sampling and measurement errors). Disentangling these sources of variability is often challenging and necessitates the use of statistical methods, such as state-space models. These approaches have become ubiquitous in ecology, particularly for time-series data (Auger-Méthé et al., 2021)—in part because these models allow researchers to make inferences about ecological processes that are not directly observable. Applications of these models include estimating population change over time (Clark & Bjørnstad, 2004), movement dynamics (Patterson et al., 2008) and understanding spatio-temporal variation (Anderson & Ward, 2019).

Estimating the multiple sources of variation in state-space models is numerically complex and can be constrained explicitly or implicitly in ecological models via model assumptions. For example, discrete time state-space models of population trajectories generally assume latent population size n_t at time t can be approximated by an autoregressive process in log-space, $x_{t+1} = f(x_t) + \epsilon_t$, where $f()$ represents some function, $x_t = \log(n_t)$ and ϵ_t are normally distributed process deviations representing stochastic variability of the natural system (Dennis et al., 2006). Without the autoregressive constraint, the variance of the stochastic noise ϵ_t is not estimable in the presence of an observation or data model. Separating these sources of variability is critical to generate the unbiased estimates of population trends or density dependence (Knape, 2008). If inference is not dependent on the parameters of ecological interest (e.g. growth rates, density dependence), a wide range of alternative semi-parametric approaches exist that can be used to model the trajectory of x_t, including generalized additive models (GAMs, Wood, 2011) and Gaussian process models (Roberts et al., 2013). Because these models are not autoregressive with discrete time steps, the flexibility or 'wiggliness' of the model can be adjusted as part of the model fitting. In addition to their flexibility, these semi-parametric models may be better suited for situations when data are patchily distributed in time or unequally spaced, making estimation of process and observation errors more difficult.

Challenges posed by univariate time-series models also apply to multivariate models, with the additional complexity that the number of latent time series may be variable, $k = 1, \ldots, m$, where m is the number of time series observed. At one extreme, $k = m$, and each time series corresponds to a unique latent process. Motivating questions in analysing these models include estimating correlated latent processes or trends, or estimating effects of environmental covariates (Hovel et al., 2017). At the other extreme, $k = 1$, where each time series represents repeated measurements of the same process, with optional offsets included for each time series (e.g. offsets allowing for differing detectability). Applications focused on estimating a single trend from multivariate data include the development of ecological indicators. Models with intermediate numbers of latent states $1 < k < m$ require mapping of time series to latent trends. These may be specified *a priori* (Ward et al., 2010) or estimated within the modelling framework using dimension reduction techniques.

Many statistical approaches have been proposed in recent years for clustering or estimating common signals in multivariate time series (Liao, 2005). Examples include clustering based on similarities among time-series features (Sardá-Espinosa, 2019), identifying common patterns in the frequency domain (Holan & Ravishanker, 2018) and clustering based on neural networks (Cherif et al., 2011). Application of these methods to ecological data has been limited, in part because many of these approaches identify clusters from raw data and ignore observation error. An alternative approach that has been used in ecology to map the collections of multivariate time series to latent processes, while accounting for observation error, is dynamic factor analysis (DFA) (Zuur, Fryer, et al., 2003; Zuur, Tuck, et al., 2003). DFA is an extension of factor analysis for time-series data, and estimates a small number of unobserved processes ('trends'), that can describe observed data. Mapping of time series to trends is done via estimated factor loadings—these allow each time series to be modelled as a mixture of estimated latent trends, rather than assigning each time series to a single trend.

To date, applications of DFA models in ecology and other fields have assumed that underlying trends are modelled as a random walk, $x_{t+1} = x_t + \epsilon_t$. The objective of this paper is to introduce a new class of DFA models based on smooth functions, instead of autoregressive processes. Recent work has highlighted the application of hierarchical GAMs for multiple data sources (Pedersen et al., 2019). These approaches are flexible and likely to provide similar inference to DFA for a single latent trend; however, these methods have not been extended to include more than one process. We illustrate two options for modelling smooth functions for latent trends: splines ('B-splines' or penalized 'P-splines') and Gaussian process models. We compare both approaches to conventional autoregressive DFA models using simulated data and, as real-world applications, using two marine fish datasets from the west coast of the United States. All data and code for replicating our analysis are available on Github (https://github.com/fate-ewi/gpdfa) and Zenodo (Ward & Anderson, 2021), and in our existing R package 'BAYESDFA' (Ward et al., 2019).

2 | MATERIALS AND METHODS

2.1 | Dynamic factor model

The basic DFA model can be written as a multivariate state-space model, consisting of a latent process model and observation or data model. The process model for a DFA with k trends is expressed as a random walk, $\mathbf{x_{t+1}} = \mathbf{x_t} + \mathbf{w_t}$, where $\mathbf{x_t}$ is a k-element vector and $\mathbf{w_t} \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$. For identifiability constraints, the covariance matrix $\mathbf{Q}$ is generally constrained to be an identity matrix (Holmes et al., 2012; Zuur, Tuck, et al., 2003). Additional features may be incorporated into the process model including autoregressive or moving average coefficients, covariates or deviations

that are more extreme than that of the normal distribution (Ward et al., 2019). The observation model in a DFA is expressed as a linear combination of trends $\mathbf{x}_t$ and a matrix of loadings coefficients $\mathbf{Z}$, $\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{B}\mathbf{d}_t + \mathbf{e}_t$. In addition to the trends and loadings, time-varying covariates $\mathbf{d}_t$ may be optionally included and linked to the observations through estimated coefficients $\mathbf{B}$. The vector $\mathbf{e}_t$ represents residual observation error, which is typically modelled as a diagonal matrix, $\mathbf{e}_t \sim \text{MVN}(0, \mathbf{R})$, although off-diagonal elements may be estimated (Holmes et al., 2020). Further details of the Bayesian implementation of the DFA model and extensions are provided in Ward et al. (2019).

2.2 | Modelling trends as Gaussian processes

Conventional DFA models with trends modelled as random walks are flexible, but for some datasets, these models may be inappropriate. If data generating processes are not well approximated by a random walk, other models may be more suitable. As a first alternative to the random walk model, we treat the trends as a Gaussian process (Roberts et al., 2013). A discrete time Gaussian process model of trends treats the vector representing the kth trend as a stochastic process, where x_k is drawn from a multivariate normal distribution. As data in a DFA are generally standardized (mean 0, standard deviation 1), we can assume the mean of each trend to be 0, and all inference about the Gaussian process centres around the covariance matrix, $x_k \sim \text{MVN}(0, \boldsymbol{\Sigma})$. Rather than estimate each element of $\boldsymbol{\Sigma}$ independently, smooth covariance functions or 'kernels' are chosen to represent the covariance between points in time (typical choices include the exponential, Gaussian and Matérn functions). For the purpose of our DFA modelling, we adopt a Gaussian kernel so that the covariance between points i and j at times t_i and t_j on trend k can be expressed as $\text{cov}\left(x_{i,k}, x_{j,k}\right) = \sigma_k^2 \exp\left(\frac{-(t_i - t_j)^2}{2\theta_k^2}\right)$, where σ_k controls the magnitude of variation, and θ_k controls how smoothly correlation decreases as time points become further apart. We allow each trend to have its own covariance parameters (θ_k, σ_k), allowing each to have differing degrees of smoothness. Because of potential computation issues in high dimensionality problems such as spatial models (Anderson & Ward, 2019; Latimer et al., 2009), we also allow this Gaussian process model to be expressed as a Gaussian predictive process model. The difference between the predictive process approach and the full Gaussian process model is that instead of modelling the x_t themselves as random variables, random variables are modelled at a subset of locations $\mathbf{x}_k^*$ (referred to as 'knots') and projected to the locations of the data $\mathbf{x}_k$. If we assume $\mathbf{x}_k^* \sim \text{MVN}(0, \boldsymbol{\Sigma}^*)$, then this projection can be done as $\boldsymbol{x}_k = \Sigma'_{k,k^*} \boldsymbol{\Sigma}^{*-1} \mathbf{x}_k^*$, where the matrix Σ'_{k,k^*} is the transpose of the matrix describing the covariance between $\boldsymbol{x}_k$ and $\boldsymbol{x}_k^*$. The location of k^* can be spaced equally or depend on data; we assume that the k^* are equally spaced within each time series (with the endpoints also acting as knots).

2.3 | Modelling trends as splines

As an alternative model of latent trends in a DFA, we use a series of smoothing functions, known as basis splines ('B-splines'), or penalized basis splines ('P-splines'). These models can be thought of as a special case of Gaussian process models (Kimeldorf & Wahba, 1970) and offer flexibility similar to the more familiar generalized additive models (Wood, 2011). Splines are represented as a series of piecewise polynomial functions, where higher order polynomials result in more flexible curves (Hastie, 1992). A common choice of the order of these polynomials is a cubic or third degree, and will be the focus of our implementation for DFA. An additional input to splines is the locations of the control points (knots) between polynomial segments—more knots translate into a more flexible function, but also one with more parameters to estimate. We assume knots to be uniformly distributed over the time series. Uniform knot vectors may be appropriate for data collected at regular intervals, but for observations more patchily distributed in time, defining knots based on quantiles or other metrics may be warranted. Mathematically, modelling the trends in a DFA with B-splines can be expressed as a linear combination of the recursive B-spline weights $\mathbf{B}$ and estimated coefficients $\mathbf{a}$, $\mathbf{x}_k = \mathbf{a}\mathbf{B}$. The matrix $\mathbf{B}$ is generated from the raw data prior to estimation (R Core Team, 2020). In the DFA setting, $\mathbf{B}$ is shared across trends, but for trend-specific variability, we allow the coefficients $\mathbf{a}$ to have a trend-specific variance, $\mathbf{a}_k \sim \text{Normal}\left(0, \sigma_k^2\right)$. P-splines represent an extension of B-splines, with an added penalty for extra wiggliness (Crainiceanu et al., 2005; Eilers & Marx, 1996); this penalty reduces the impact of the number of B-spline basis functions on model fit (Wood, 2017). For our implementation in BAYESDFA, we use a linear penalty with second-order difference (Eilers & Marx, 1996).

2.4 | Simulations to compare model performance

To examine the relative performance of our proposed smooth trend models versus conventional approaches, we conducted a series of simulations to investigate sensitivity to (a) departures from random walks and (b) magnitude of observation error variance. We generated sets of simulated data consisting of 20 time steps and three time series. Each dataset was assumed to be generated from a single trend, which we modelled either as a random walk or as a smooth trend with a B-spline. Observed time series were generated from these trends by multiplying random loadings $Z_i \sim \text{Normal}\left(1, 0.1^2\right)$ and then adding observation error (we used three levels of the observation error standard deviation: $\sigma_{\text{obs}} = 0.25, 0.5, 1$). Each set of simulated time series was then fit with the same estimation models: a conventional DFA model estimating the trend as a random walk, smooth trends approximated with a B-spline (7, 13 and 20 knots), a P-spline (13 knots) and a full-rank Gaussian process (20 knots). For each combination of trend model and observation error, we generated a total of 100 simulated datasets.

Estimation was done in a Bayesian framework using our BAYESDFA R package (Ward et al., 2019). For the spline models, we assigned priors on the weights $\mathbf{a} \sim \text{Normal}(0, 1)$. Similarly, we assigned standard half-normal priors for the Gaussian process variances $\sigma_k \sim \text{Normal}(0, 1)$, and inverse Gamma priors for the scale $\theta_k \sim \text{IG}(3, 1)$. Bayesian estimation in the BAYESDFA package is done using Stan and the R package RSTAN (Stan Development Team, 2016), which implements Markov chain Monte Carlo (MCMC) using the No-U Turn Sampling (NUTS) algorithm (Carpenter et al., 2017; Hoffman & Gelman, 2014). The relative performance of each estimation model was done using out-of-sample predictive ability. For each of the simulated time series described above, we randomly held out 10% (two of every 20 observations) as a test set. Because of the large number of models (600 simulated datasets, 4,200 estimation models), we only ran one MCMC chain (3,000 iterations, discarding the first half as warm up) and generated posterior predictions for the test data. The normal log density of the test set was calculated for each MCMC iteration, and the expected log pointwise predictive density (ELPD) was used to summarize these values across draws (Vehtari et al., 2017).

2.5 | Application: one-trend models of juvenile fish dynamics

As a first application of smooth DFA models, we analysed time-series data of pelagic juvenile rockfishes collected in Southern California (USA). The California Cooperative Oceanic Fisheries Investigations (CalCOFI) programme has been conducting quarterly research vessel surveys to collect physical and biological data since 1949, to monitor changes to the California Current Ecosystem (Bograd et al., 2003). The CalCOFI data have been incorporated into models used to assess population status (MacCall, 2003), and numerous publications have used these time series as indicators of ecosystem state (McClatchie et al., 2008). These types of motivating questions also present an opportunity to apply DFA with both conventional and smoothed trends to generate ecosystem state indices. For this application, we focused on the dynamics of three co-occurring species of juvenile rockfishes: aurora rockfish *Sebastes aurora*, shortbelly rockfish *S. jordani* and bocaccio rockfish *S. paucispinis*. We restricted the time series to data collected since 1985, when sampling has been consistent in space and time on fixed sampling lines (Moser et al., 2001). Although CalCOFI cruises are done throughout the year, we were primarily interested in estimating interannual trends, and further restricted our analysis to considering spring cruises from 1 April to 22 May when densities of most rockfish species are highest (Mosek et al., 2000). All data were retrieved using the software R (R Core Team, 2020) and the RERDDAP package (Chamberlain, 2020).

With only three time series, we focused on DFA models with one trend and a single observation error variance shared across species. Other types of models, including hierarchical GAMs (Pedersen et al., 2019) or models allowing estimated offsets, may also be useful in this type of application. Where the DFA model differs is that unlike models with random intercepts or additive terms, the DFA factor loadings $\mathbf{Z}$ are multiplicative and may be close to zero. These cases may arise when a particular time series has a low signal to noise ratio, or if there is low correspondence with the latent trends estimated among all other time series. In addition to estimating a conventional one-trend DFA model with a latent autoregressive process, we evaluated smooth one-trend models (trend estimated with a B-spline, P-spline or Gaussian process). Because we had no *a priori* hypotheses about the complexity of these smoothed factor models, we evaluated a range of models for each (Table 1), using equally spaced knots.

2.6 | Application: two-trend models of commercial fisheries catches

As a slightly more complex example of the smooth factor analysis model, we examined the performance of two-trend models, using a dataset of commercial fisheries catches (landings) from the west coast of the United States. This dataset consists of 13 species or groups reported annually from multiple fisheries over a 39-year period (1981–2019) (PFMC, 2020). Landings off the US West Coast are dominated by Pacific hake (also Pacific whiting, *Merluccius productus*),

TABLE 1 Leave One Out Information Criterion (LOOIC) and Expected Log Posterior Density (ELPD) with standard errors in parentheses for each of the models applied to our case studies (CalCOFI time series of juvenile rockfishes, and the time series of commercial groundfish landings from the west coast of the United States). For each model, knots (or locations of control points) are assumed to be uniformly spaced over the time series. To aid in interpretation, the minimum LOOIC value has been subtracted from each case study and ELPD values have been subtracted from the maximum (0 for each metric reflects the most supported model)

Trend model	Knots	CalCOFI LOOIC	CalCOFI ELPD	Landings LOOIC	Landings ELPD
Random walk	NA	2.44 (12.16)	1.22 (6.08)	28.05 (49.21)	14.02 (24.61)
B-spline	6	4.67 (12.45)	2.33 (6.23)	2.14 (54.78)	1.07 (27.39)
B-spline	18	3.3 (12.25)	1.65 (6.12)	22.39 (48.45)	11.19 (24.23)
B-spline	30	0 (12.64)	0 (6.32)	64.32 (46.98)	32.16 (23.49)
P-spline	6	3.22 (12.63)	1.61 (6.32)	9.16 (55.73)	4.58 (27.86)
P-spline	18	3.12 (12.72)	1.56 (6.36)	2.53 (54.93)	1.26 (27.46)
P-spline	30	2.68 (12.5)	1.34 (6.25)	0 (54.63)	0 (27.31)
GP	6	2.54 (12.43)	1.27 (6.21)	3.79 (53.64)	1.89 (26.82)
GP	18	2.72 (12.33)	1.36 (6.17)	6.73 (53.31)	3.36 (26.66)
GP	30	1.56 (12.5)	0.78 (6.25)	7.24 (53.26)	3.62 (26.63)
GP	Full rank	0.51 (12.44)	0.26 (6.22)	4.97 (53.42)	2.49 (26.71)

but also include substantial catches of rockfishes (*Sebastes* spp.) and flatfishes (e.g. Dover sole, *Solea solea*). Over the course of the last four decades, these species have experienced variability associated with population dynamics and the environment, but the patterns of landings also reflect a dynamic fisheries management process. Examples of changes include temporarily closing areas to fishing to protect species of conservation concern, and implementing catch share programmes. These processes, combined with environmental conditions that have been positive for many species, have resulted in many increasing populations (Warlick et al., 2018). Given these various management and ecological changes, it is important to summarize the patterns of landings, and identify common trends as indicators for management and ecosystem status (Harvey et al., 2018).

As with our previous example, we compared conventional DFA models to those modelling the trends with smooth functions. Preliminary model comparisons with one-trend models suggested that two-trend models were most supported by the data, and thus will be the focus of our analysis. In addition to modelling the two-trend model with conventional DFA, we evaluated spline and Gaussian process models with equally spaced knots (Table 1). All models included a single observation error variance, shared across time series.

2.7 | Estimation and model selection

For each model considered in our applications, we ran three parallel MCMC chains for 4,000 iterations each, discarding the first 50% of the samples. We assessed convergence using split-$\hat{R}$ and effective sample size (Gelman et al., 2013) along with trace plots. We used the LOO package to calculate the approximate ELPD (Vehtari et al., 2017), and the Leave One Out Information Criterion (LOOIC, Vehtari et al., 2017, 2020) as a model selection tool (Ward et al., 2019), which approximates leave-one-out cross-validation. Preliminary model checks using LOOIC for the models included in our analysis indicated that many models had one to four data points that had high Pareto-*k* statistics (possibly because of model misspecification or model flexibility, Vehtari et al. (2017)). To avoid refitting these models, we implemented moment matching in the LOO package (Paananen et al., 2021; Vehtari et al., 2020).

3 | RESULTS

3.1 | Simulations to compare model performance

Our simulations were designed to explore the relative performance of DFA models that estimate trends as random walks versus our proposed smooth trends, when trends depart from random walks and are corrupted by observation error. Our results suggested that when observation error is relatively high, there is little difference between the smooth trend and random walk DFA models (Figure 1). As observation error decreases, ELPD favours smooth trend models when the underlying trend is smooth and random walk models when the underlying trend is random walk (Figure 1). The largest ELPD weight for the smooth trend model occurred when observation error was low (0.25) and the number of knots in the estimation model was closest to that of the simulation model ('BS7' Figure 1 left panel). The P-spline and Gaussian process models provided weak support to the true data generating model with low observation error, but models were indistinguishable with higher levels of observation errors. Results across knots for the B-spline estimation model demonstrate that flexibility increases as more knots are added, and the smooth trend approach becomes similar to the random walk (Figure 1 left panel).

3.2 | Application: one-trend models of juvenile fish dynamics

For our application of smooth dynamic factor models to the CalCOFI juvenile rockfish dataset, we found that the full-rank Gaussian process DFA model and B-spline model with 30 knots had slightly lower LOOIC values compared to alternative models (Table 1), and these high-dimensional models performed slightly better than the conventional DFA model. Varying the number of knots for the P-spline models and Gaussian process models resulted in qualitatively similar data support (Table 1), while the predictive accuracy of the B-spline model increased with more knots. This greater flexibility allowed more complex models to better capture recent variability in rockfish densities (Figure 2). Trend 1 can be seen as largely capturing the variability in the time series of aurora rockfish, which had the loading that was largest in magnitude (−0.11, 90% credible interval = −1.21–1.09). Bocaccio rockfish also loaded positively on trend 1, though the effect was weaker (−0.14, 90% credible interval = −1.36–1.36). The loading for shortbelly rockfish was smallest in magnitude (−0.07, 90% credible interval = −0.96–0.97).

3.3 | Application: two-trend models of commercial fisheries catches

When DFA models were applied to commercial fisheries landings data from the west coast of the United States, the model with the lowest LOOIC was the P-spline model with 30 knots (second was the B-spline model with six knots). The first trend exhibited nearly linear change from 1981 to 2001 and was relatively stationary from 2001 to 2019 (Figure 3). The second trend represented change from the early 1990s, with the strongest change occurring 2010–present. Estimates of the loadings from the best model indicated many species or species groups loaded negatively on trend 1 (lingcod, sablefish, rockfishes), but arrowtooth flounder and Pacific whiting had opposite loadings (Figure 3). Trend 2 from this model appeared to contrast species with relatively stationary catches before declining in 2010 (e.g. arrowtooth flounder, *Atheresthes stomas*) versus Petrale sole *Eopsetta jordani*—one of the only non-whiting species that has

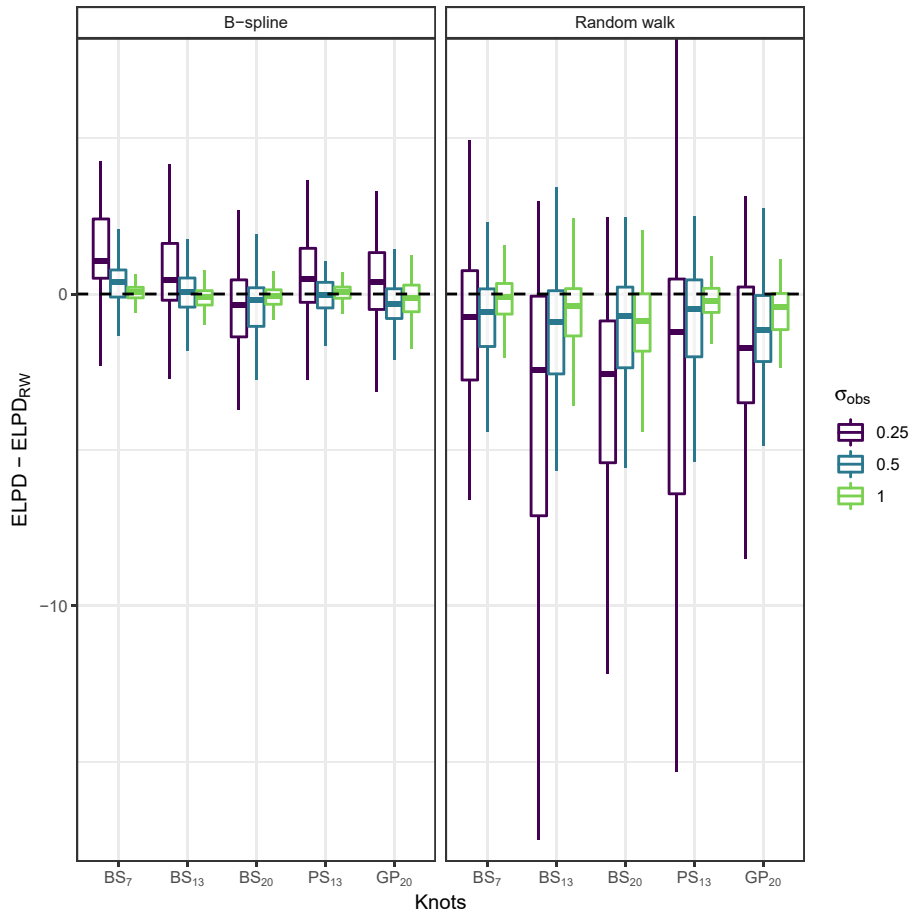

FIGURE 1 Simulation results, showing the difference in expected log pointwise predictive density (ELPD) between each model and the conventional dynamic factor analysis model with trends estimated as random walks. Data were generated from either a B-spline with seven knots or a random walk, and three levels of observation error were explored (0.25, 0.5, 1). Results are shown for each combination of observation error and estimation model (B-spline with 7, 13 or 20 knots; P-spline with 13 knots; Gaussian process with 20 knots, random walk). Each boxplot corresponds to 100 ELPD point estimates, and each facet represents a different data generating model (B-spline or random walk)

experienced positive catches since 2010. Predictions across all models appeared to characterize the trends of most species, and trends from the best model generated more precise predictions relative to the random walk (Figure 4), although neither model was able to capture the variability in Pacific whiting catches since 2000.

While low-dimensional Gaussian process and spline models performed similarly (Table 1), comparing higher order models demonstrates the contrast between approaches. As more knots were added to spline models, the wiggliness of the estimated trends generally increased for the B-spline models but remained smooth for the P-spline approach (Figure 5). Like the P-spline models, trends from the Gaussian process models did not become more wiggly as more knots were added, though the credible intervals of estimated trends were wider than either of the spline approaches (Figure 5). Estimates of θ_k for this Gaussian process model were relatively large (8.13, 4.78), allowing correlation between neighbouring points to decrease slowly and neighbouring points further away to have a larger effect. In contrast, the full-rank Gaussian process model was most supported for the CalCOFI data—this model had a relatively small value of $\theta_k = 1.15$, allowing correlation between adjacent points to decrease rapidly, translating into greater flexibility.

4 | DISCUSSION

Dynamic factor analysis represents a flexible approach for using state-space models to capture latent processes in multivariate time series (Zuur, Fryer, et al., 2003; Zuur, Tuck, et al., 2003). For some ecological processes—particularly those with high variability—random walks may be too constraining, while for others, using a random walk may be overly complex. Examples of cases where random walks may overfit trends may exist when there are large temporal gaps between observations, or data are collected from systems with high signal to noise ratios. As alternatives to the conventional random walk, we illustrate how DFA trends may be modelled using Gaussian process models smooth functions (B-splines, P-splines). The smoothness of spline approaches may be specified a priori by the user, and compared via model selection. As the variability of

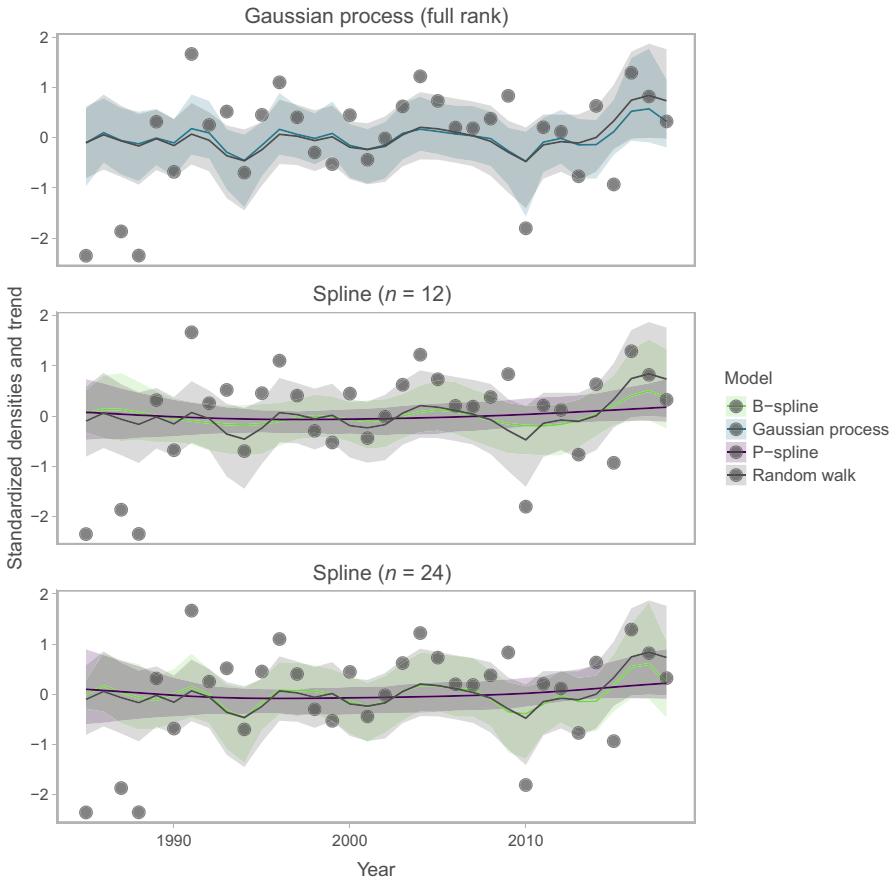

FIGURE 2 Standardized densities of juvenile shortbelly rockfish *Sebastes jordani* collected in the CalCOFI survey, and estimates of latent trends for three candidate models, representing a range of flexibility in splines compared to the conventional random walk. In addition to the conventional dynamic factor analysis model with a latent random walk (included in all panels for reference), predictions from a full-rank Gaussian process model and B-spline model with 12 knots and 24 knots are shown. The posterior mean from each model is shown as a solid line, and 90% credible intervals are shown with ribbons

latent trends is nearly always fixed in a conventional DFA for identifiability (Holmes et al., 2012; Zuur, Tuck, et al., 2003), adopting an alternative model of the trend does not limit inference or change the meaning of other parameters (e.g. loadings). Based on our application of these approaches to simulated data, smooth trend DFA models may be better supported in situations where the data generating process is more smooth than a random walk; examples included processes that are highly autocorrelated or have large amounts of environmental forcing.

In both of our case studies, comparing smooth DFA models to conventional ones, we found that using smooth functions to model DFA trends resulted in models with higher predictive ability (as measured with LOOIC). Our two case studies contrast two datasets with different degrees of variability. The CalCOFI dataset on juvenile rockfish abundance represents data with relatively high variability—both because of the sampling process, and because the nature of fish recruitment in space and time is stochastic. Our second example consisted of applying DFA models to time series of fisheries catches; these data are generally less variable than the CalCOFI data because catches are aggregated across a large spatial area and individual vessels. Like the CalCOFI example, we found that smooth trend DFA models were better supported over the conventional random walk; however, the models receiving the most support were lower dimension models (e.g. P-spline with 30 knots, B-spline with 6 knots). For both of our case studies, knot locations were assigned uniformly, and these results would be expected to change slightly if the knot locations were adjusted. For models with missing data, or datasets with unevenly distributed replicate samples, it may be important to consider non-uniform knot locations.

Our case studies also highlighted that predictions from smooth trend models that use splines or Gaussian processes may be nearly identical, raising the question of which approach may be better to use in practice. Spline models can give equivalent predictions to Gaussian process models with the same kernel used in our models (Kimeldorf & Wahba, 1970); however, the smoothing approaches differ slightly between these models. Analysts using these methods with DFA may be more interested in applying the Gaussian process model if inference about covariance parameters is of interest, while the B-spline or P-spline models may be computationally faster in many other applications. P-splines offer the additional advantage of being less sensitive to the number of knots.

Because of their flexibility, applications of LOOIC or related model selection tools to state-space models, including the DFA models in our analysis, may result in poor diagnostics (e.g. high Pareto-*k* statistics). Alternative approaches for evaluating predictive performance may be used, including the ELPD obtained via *k*-fold cross-validation (Vehtari et al., 2017, 2020). Rather than performing parameter estimation once per model, as was done in our analysis using the LOO package, calculating ELPD is more computationally challenging because with cross-validation, a model must be fit once per fold. Re-fitting the model multiple times also allows alternative cross-validation methods to be more easily applied. Commonly used

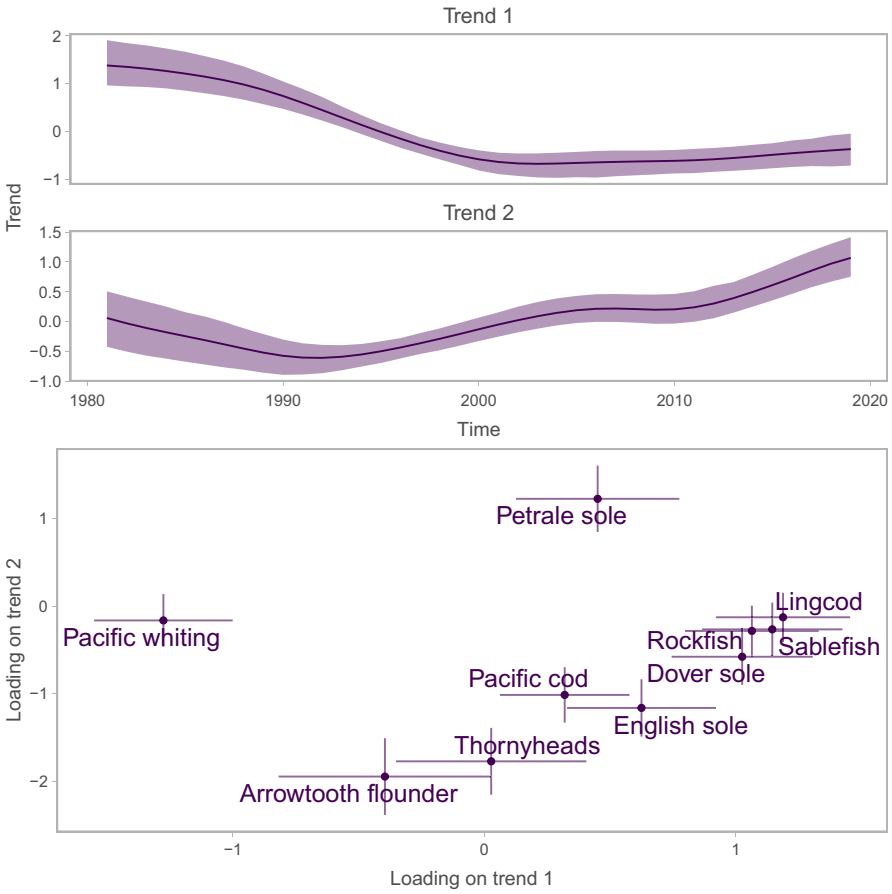

FIGURE 3 Estimated trends and loadings from the two-trend dynamic factor analysis model applied to commercial groundfish landings off the west coast of the United States. The model results with lowest Leave One Out Information Criterion are shown, a model that allows trends to be approximated with P-splines (30 knots). The posterior mean for each trend is shown, with ribbons representing 90% credible intervals. The loadings of each species on each trend are shown as points, with lines representing 90% credible intervals

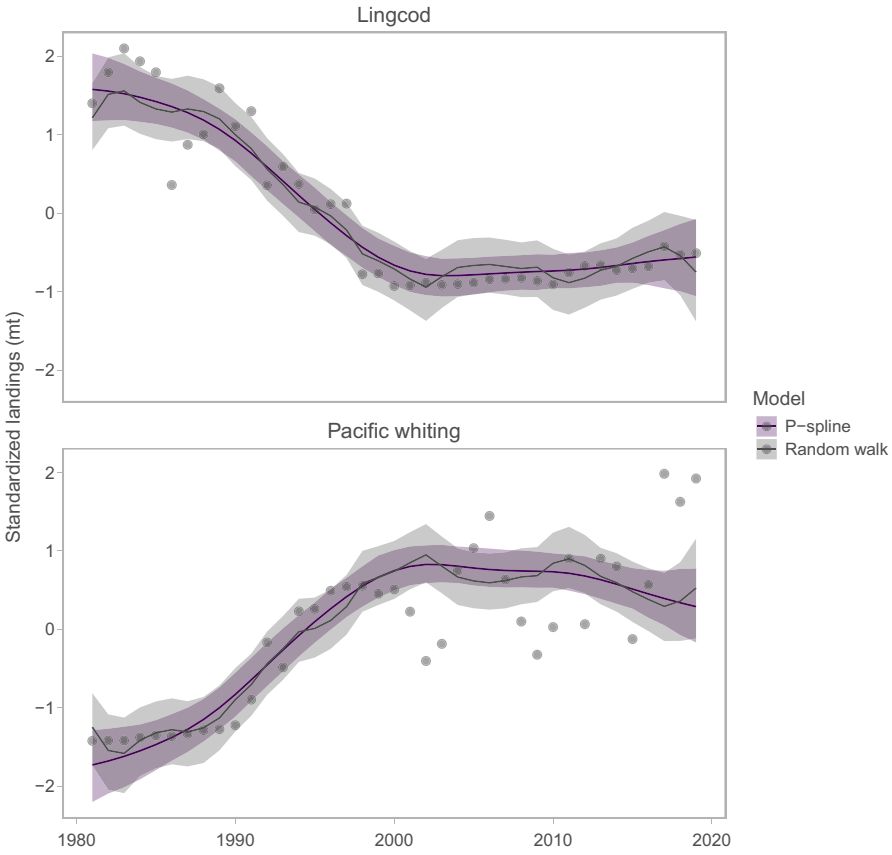

FIGURE 4 Estimated landings for two species included in our analysis, with contrasting trends (lingcod, Pacific whiting). Posterior means and 90% credible intervals (ribbons) for two candidate models are shown: a P-spline trend model with 30 knots, and a random walk model representing the conventional dynamic factor analysis

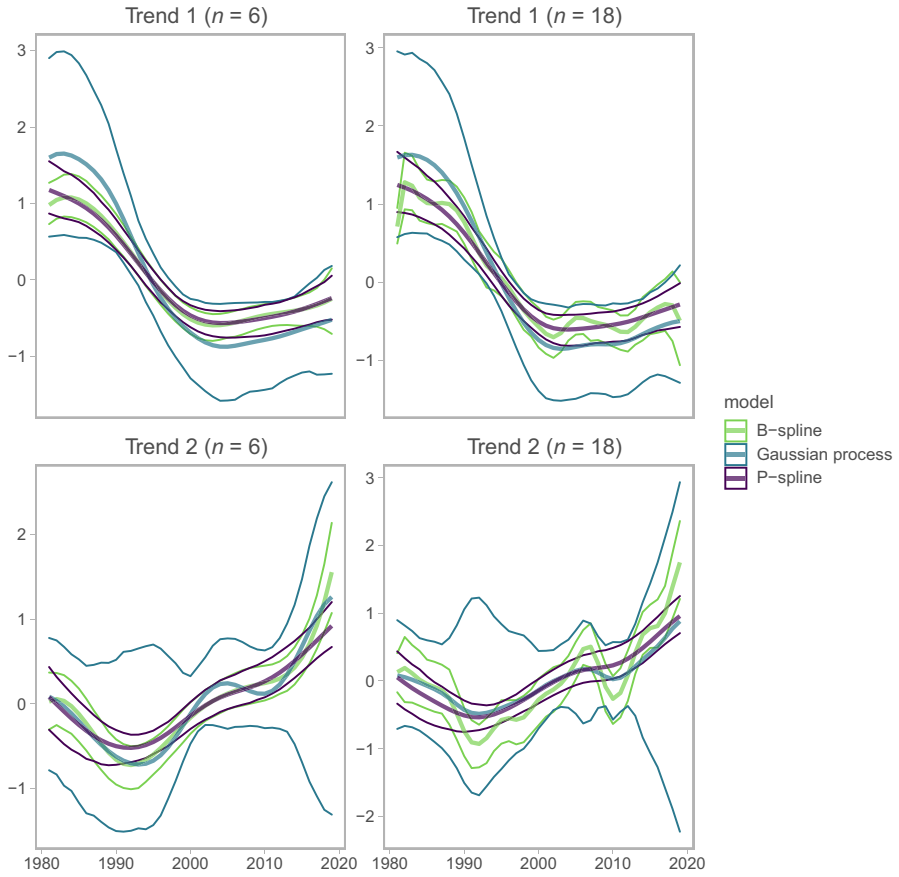

FIGURE 5 Estimated trends for the two-trend model of fisheries landings off the west coast of the United States. Shown are results for the B-spline and Gaussian process models with 6 and 18 knots (or control points). Solid lines represent the posterior means and 90% credible intervals are shown as ribbons

approximations like LOOIC represent an approximation to leave-one-out cross-validation where each data point is held out in turn. An alternative approach for time-series data is that the observations in each time step can be treated as a fold, and held out in turn. Extensions of this time-series approach include leave-future-out cross-validation, where data points are only used to predict future observations, not historical ones (Bürkner et al., 2020).

There are a number of possible extensions to the smooth function DFA models described in this paper. One extension would be to further constrain the wiggliness defined by the Gaussian process rate of correlation decay (θ) via a prior such as the penalized complexity (PC) prior (Simpson et al., 2017). Such a prior which would allow one to more easily impart prior beliefs about the parameter scale. Second, the smooth trends could themselves be hierarchical: The trends could share their wiggliness, draw smoothing parameters from a shared distribution or share a global smoother combined with group-specific smoothers (Pedersen et al., 2019). DFA represents a powerful and underutilized tool for dimension reduction of multivariate time series. Our extensions of conventional methods to implement smooth trends enhance the flexibility of this tool for estimating latent processes, and offer a robust approach for DFA that may also be useful in hindcasting or forecasting scenarios.

ACKNOWLEDGEMENTS

We thank the scientific staff who have collected and maintained the CalCOFI cruises (including California State Department of Fish and Wildlife, National Marine Fisheries Service, and Scripps Institution of Oceanography) and Pacific Fishery Management Council for providing coastwide catch data. We also thank Dr. Eric Pederson, 2 anonymous reviewers, and the editor (Dr. José Ponciano) who all provided feedback that improved our manuscript.

CONFLICT OF INTEREST

The authors have no conflicts to declare.

AUTHORS' CONTRIBUTIONS

M.E.H. and M.A.L. secured funding for the initial development of the BAYESDFA package and related publications; E.J.W. and S.C.A. did the majority of the code development, though all authors were involved in testing and simulation; E.J.W. and S.C.A. wrote the initial draft of the manuscript, with all authors providing critical edits and comments on figures. All authors reviewed the final draft and gave approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13788.

DATA AVAILABILITY STATEMENT

All data and code to reproduce our analyses are in our Github repository (https://github.com/fate-ewi/gpdfa) and Zenodo https://doi.org/10.5281/zenodo.5571171 (Ward & Anderson, 2021).

ORCID

Eric J. Ward https://orcid.org/0000-0002-4359-0296
Sean C. Anderson https://orcid.org/0000-0001-9563-1937
Michael A. Litzow https://orcid.org/0000-0003-1611-4881

REFERENCES

Anderson, S. C., & Ward, E. J. (2019). Black swans in space: Modeling spatiotemporal processes with extremes. *Ecology*, *100*, e02403. https://doi.org/10.1002/ecy.2403

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A. A., Leos-Barajas, V., Mills Flemming, J., Nielsen, A., Petris, G., & Thomas, L. (2021). A guide to state-space modeling of ecological time series. *Ecological Monographs*, *91*(4), e01470. https://doi.org/10.1002/ecm.1470

Bograd, S. J., Checkley, D. A., & Wooster, W. S. (2003). CalCOFI: A half century of physical, chemical, and biological research in the California Current system. *Deep Sea Research Part II: Topical Studies in Oceanography*, *50*, 2349–2353.

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, *90*, 2499–2523. https://doi.org/10.1080/00949655.2020.1783262

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

Chamberlain, S. (2020). *Rerddap: General purpose client for 'ERDDAP' servers*.

Cherif, A., Cardot, H., & Boné, R. (2011). SOM time series clustering and prediction with recurrent neural networks. *Neurocomputing*, *74*, 1936–1944. https://doi.org/10.1016/j.neucom.2010.11.026

Clark, J. S., & Bjørnstad, O. N. (2004). Population time series: Process variability, observation errors, missing values, lags, and hidden states. *Ecology*, *85*, 3140–3150. https://doi.org/10.1890/03-0520

Crainiceanu, C. M., Ruppert, D., & Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, *14*, 1–24.

Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L., & Staples, D. F. (2006). Estimating density dependence, process noise, and observation error. *Ecological Monographs*, *76*, 323–341. https://doi.org/10.1890/0012-9615(2006)76[323:EDDPNA]2.0.CO;2

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89–102. https://doi.org/10.1214/ss/1038425655

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Harvey, C., Garfield, N., Williams, G., Tolimieri, N., Schroeder, I., Hazen, E., Andrews, K., Barnas, K., Bograd, S., Brodeur, R., Burke, B., Cope, J., deWitt, L., Field, J., Fisher, J., Good, T., Greene, C., Holland, D., Hunsicker, M., & Zador, S. (2018). *Ecosystem status report of the California Current for 2018: A summary of ecosystem indicators compiled by the California Current Integrated Ecosystem Assessment Team (CCIEA).*

Hastie, T. J. (1992). *Statistical models in S*. J. M. Chambers & T. J. Hastie (Eds.). Wadsworth & Brooks/Cole.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Holan, S. H., & Ravishanker, N. (2018). Time series clustering and classification via frequency domain methods. *Wires Computational Statistics*, *10*, e1444. https://doi.org/10.1002/wics.1444

Holmes, E. E., Ward, E. J., & Scheuerell, M. D. (2020). *Analysis of multivariate time-series using the MARSS package*. Retrieved from https://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf

Holmes, E. E., Ward, E. J., & Wills, K. (2012). MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal*, *4*, 11–19. https://doi.org/10.32614/RJ-2012-002

Hovel, R. A., Carlson, S. M., & Quinn, T. P. (2017). Climate change alters the reproductive phenology and investment of a lacustrine fish, the three-spine stickleback. *Global Change Biology*, *23*, 2308–2320. https://doi.org/10.1111/gcb.13531

Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, *41*, 495–502. https://doi.org/10.1214/aoms/1177697089

Knape, J. (2008). Estimability of density dependence in models of time series data. *Ecology*, *89*, 2994–3000. https://doi.org/10.1890/08-0071.1

Latimer, A. M., Banerjee, S., Sang Jr, H., Mosher, E. S., & Silander Jr, J. A. (2009). Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern United States. *Ecology Letters*, *12*, 144–154. https://doi.org/10.1111/j.1461-0248.2008.01270.x

Liao, T. W. (2005). Clustering of time series data—A survey. *Pattern Recognition*, *38*, 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

MacCall, A. D. (2003). *Status of Bocaccio off California in 2003. In appendix to the status of the Pacific coast groundfish fishery through 2003: Stock assessment and fishery evaluation*. Pacific Fishery Management Council.

McClatchie, S., Goericke, R., Koslow, J., Schwing, F., Bograd, S., Charter, R., Watson, W., Lo, N., Hill, K., Gottschalck, J., L'Heureux, M., Xue, Y., Peterson, W., Emmett, R. T., Collins, C., Gaxiola-Castro, G., Durazo, R., Kahru, M., Mitchell, B., & Bjorkstedt, E. (2008). The state of the California Current, 2007–2008: La Niña conditions and their effects on the ecosystem. *California Cooperative Oceanic Fisheries Investigations Reports*, *49*, 39–76.

Mosek, H., Charter, L., Watson, W., Ambrose, I., Shakon, N., Charter, K., Saniiknoi, E., Fisheiies, S., Center, S., Fi, M., & Service, H. (2000). Abundance and distribution of rockfish (Sebastes) larvae in the Southern California Bight in relation to environmental conditions and fishery exploitation. *Abundance and distribution of rockfish larvae CalCOFI Report*, 41.

Moser, H. G., Charter, R. L., Watson, W., Amurose, A., Smith, P. E., Sani, E. M., & Charter, S. R. (2001). The CalCOFI ichthyoplankton time series: Potential contributions to the management of rocky-shore fishes. *CALCOFI Reports*, *42*, 17.

Paananen, T., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, *31*, 16. https://doi.org/10.1007/s11222-020-09982-2

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology & Evolution*, *23*, 87–94. https://doi.org/10.1016/j.tree.2007.10.009

Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, *7*, e6876. https://doi.org/10.7717/peerj.6876

PFMC. (2020). *Status of the pacific coast groundfish fishery: Stock assessment and fishery evaluation*. The Pacific Fishery Management Council.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*, 20110550.

Sardá-Espinosa, A. (2019). Time-series clustering in R using the dtwclust package. *The R Journal*, *11*, 22–43. https://doi.org/10.32614/RJ-2019-023

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

Stan Development Team. (2016). *RStan: The R interface to Stan*. Stan Development Team.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). *Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Ward, E. J., & Anderson, S. C. (2021). Code for Methods in Ecology and Evolution paper describing smooth trend dynamic factor analysis (v1.1). *Zenodo*, https://doi.org/10.5281/zenodo.5571171

Ward, E. J., Anderson, S. C., Damiano, L. A., Hunsicker, M. E., & Litzow, M. A. (2019). Modeling regimes with extremes: The bayesdfa package for identifying and forecasting common trends and anomalies in multivariate time-series data. *The R Journal*, *11*, 46. https://doi.org/10.32614/RJ-2019-007

Ward, E. J., Chirakkal, H., González-Suárez, M., Aurioles-Gamboa, D., Holmes, E. E., & Gerber, L. (2010). Inferring spatial structure from time-series data: Using multivariate state-space models to detect metapopulation structure of California sea lions in the Gulf of California, Mexico. *Journal of Applied Ecology*, *47*, 47–56. https://doi.org/10.1111/j.1365-2664.2009.01745.x

Warlick, A., Steiner, E., & Guldin, M. (2018). History of the West Coast groundfish trawl fishery tracking socioeconomic characteristics across different management policies in a multispecies fishery. *Marine Policy*, *93*, 9–21. https://doi.org/10.1016/j.marpol.2018.03.014

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

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

Zuur, A. F., Fryer, R. J., Jolliffe, I. T., Dekker, R., & Beukema, J. J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, *14*, 665–685. https://doi.org/10.1002/env.611

Zuur, A. F., Tuck, I. D., & Bailey, N. (2003). Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences*, *60*, 542–552. https://doi.org/10.1139/f03-030

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ward, E. J., Anderson, S. C., Hunsicker, M. E., & Litzow, M. A. (2022). Smoothed dynamic factor analysis for identifying trends in multivariate time series. *Methods in Ecology and Evolution*, 13, 908–918. https://doi.org/10.1111/2041-210X.13788