An introduction to state-space modeling of ecological time series

Marie Auger-Méthé[1,2,*], Ken Newman[3,4], Diana Cole[5], Fanny Empacher[6], Rowenna Gryba[1,2], Aaron A. King[7], Vianey Leos-Barajas[8,9], Joanna Mills Flemming[10], Anders Nielsen[11], Giovanni Petris[12], and Len Thomas[6]

[*]auger-methe@stat.ubc.ca

[1]Department of Statistics, University of British Columbia, Vancouver, V6T 1Z4, Canada

[2]Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, V6T 1Z4, Canada

[3]Biomathematics and Statistics Scotland, Edinburgh, Scotland, EH9 3FD, UK

[4]School of Mathematics, University of Edinburgh, Edinburgh, Scotland, EH9 3FD, UK

[5]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent, CT2 7FS, UK

[6]Centre for Research into Ecological and Environmental modeling, University of St Andrews, Scotland

[7]Center for the Study of Complex Systems and Departments of Ecology & Evolutionary Biology and Mathematics, University of Michigan, Ann Arbor, Michigan, 48109, USA

[8]Department of Statistics, Iowa State University, Ames, Iowa, 50011, USA

[9]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina, 27695, USA

[10]Department of Mathematics and Statistics, Dalhousie University, Halifax, B3H 4R2, Canada

[11]National Institute for Aquatic Resources, Technical University of Denmark, 2920 Charlottenlund, Denmark

[12]Department of Mathematical Sciences, University of Arkansas, Fayetteville, Arkansas, 72701, USA

Abstract

State-space models (SSMs) are an important modeling framework for analyzing ecological time series. These hierarchical models are commonly used to model population dynamics and animal movement, and are now increasingly being used to model other ecological processes. SSMs are popular because they are flexible and they model the natural variation in ecological processes separately from observation error. Their flexibility allows ecologists to model continuous, count, binary, and categorical data with linear or nonlinear processes that evolve in discrete or continuous time. Modeling the two sources of stochasticity separately allows researchers to differentiate between biological stochasticity (e.g., in birth processes) and imprecision in the sampling methodology, and generally provides better estimates of the ecological quantities of interest

than if only one source of stochasticity is directly modeled. Since the introduction of SSMs, a broad range of fitting procedures have been proposed. However, the variety and complexity of these procedures can limit the ability of ecologists to formulate and fit their own SSMs. In addition, many SSM users are unaware of the potential estimation problems they could encounter, and of the model selection and validation tools that can help them assess how well their models fit their data. In this paper, we present a review of SSMs that will provide a strong foundation to ecologists interested in learning about SSMs, introduce new tools to veteran SSM users, and highlight promising research directions for statisticians interested in ecological applications. The review is accompanied by an in-depth tutorial that demonstrates how SSMs models can be fitted and validated in `R`. Together, the review and tutorial present an introduction to SSMs that will help ecologists to formulate, fit, and validate their models.

1 Introduction

State-space models (SSMs) are a popular modeling framework for analyzing ecological time-series data. They are commonly used to model population dynamics (Newman et al., 2014), including metapopulation dynamics (Ward et al., 2010), they have a long history in fisheries stock assessment (Aeberhard et al., 2018), and have been recently proposed as a means of analyzing sparse biodiversity data (Kindsvater et al., 2018). Moreover, they have been a favored approach in movement ecology for more than a decade (Patterson et al., 2008), and are increasingly used with biologging data (Jonsen et al., 2013). They are also used in epidemiology (Dukic et al., 2012; Fasiolo et al., 2016) and disease ecology (Hobbs et al., 2015). These common uses of SSMs, and their many unique applications (e.g., investigating animal health from photographs, Schick et al., 2013), demonstrate their wide-spread importance in ecology.

SSMs are popular for time series in part because they model process variation and observation error separately. SSMs are a type of hierarchical model (Cressie et al., 2009) and their hierarchical structure accommodates the modeling of two time series: 1) a state, or process, time series that is unobserved and attempts to reflect the true, but hidden, state of nature; and 2) an observation time series that consists of observations of, or measurements related to, the state time series. For example, actual fish population size over time would be the state time series, while incomplete and imprecise counts of fish sampled in a survey, or caught in a fishery, would be the observation time series. Process variation represents the stochastic processes that changes the population size of a fish stock through time (e.g., the birth and death processes), while observation error reflects differences between the hidden state and the observed data due to randomness or imprecision in the sampling or survey methodology. These two stochastic components act at different levels of the model hierarchy, and the SSM framework allows them to be modeled separately. The assumptions that the hidden states are autocorrelated (e.g., that a large population in year t will likely lead to a large population in year $t + 1$), and that observations are independent once we account for their dependence on the states (Fig. 1A), allow SSMs to separate these two levels of stochasticity. When we fit a SSM to time series, we can often estimate the process and observation parameters, as well as the hidden states. These estimates of the hidden states generally reflect the true state of nature better than the original observations (Fig. 1B). For

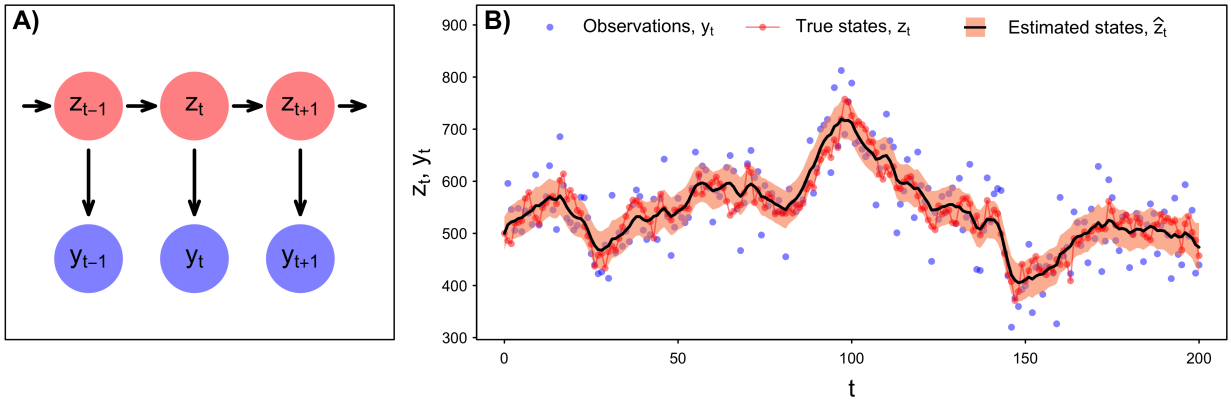

Figure 1: The dependence structure and evolution of the two time series comprising a simple univariate SSM. Panel A) represents the dependence relationships with arrows and demonstrates that once the dependence of the observations, y_t, on the states, z_t, is accounted for, the observations are assumed independent. Panel B) represents our toy model (Eqs. 1-2). The blue and red dots are the simulated observations and states respectively. The black line and red bands are the estimated state and associated 95% confidence intervals. The true states, but not the observations, usually fall in the 95% confidence intervals. This demonstrates that the state estimates can be a closer approximation of the truth than the observations.

example, the estimates of the hidden states will generally reflect the true fish population size better than the survey- or fisheries-based counts.

Since their initial development, there have been important advancements in SSMs, and in their application in ecology. The first SSMs, often referred as Normal Dynamic Linear Models (NDLMs), were a special case where the state and the observation time series are normally distributed with a linear mean structure. Two seminal papers on NDLMs, Kalman (1960) and Kalman and Bucy (1961), provided an algorithmic procedure, the now famous Kalman filter, for making inferences about the hidden states given imperfect observations. These papers led to developments that revolutionized aerospace engineering in the 1960s and allowed the Apollo mission to correct the trajectory of a spacecraft going to the moon, given inaccurate observations of its location through time (Grewal and Andrews, 2010). The earliest applications of SSMs to ecological data, which used NDLMs and the Kalman filter, were in the 1980-90s and focused primarily on fisheries (Mendelssohn, 1988; Sullivan, 1992) and animal movement (Anderson-Sprecher and Ledolter, 1991). These first animal movement SSMs were closely analogous to the original aerospace application in that they recreated the trajectory of an animal based on inaccurate observations. However, these ecological models required parameters to be estimated. Unlike a planned mission to the moon, we rarely have *a priori* knowledge of the intended speed and direction of an animal. Developments in the time-series literature made use of the Kalman filter to evaluate the likelihood function for unknown parameters, thus allowing calculation of maximum likelihood parameter estimates in addition to state estimates (Harvey, 1990). NDLMs, however, are a very restricted class of SSMs and their applicability to many ecological time series, which have nonlinear and

non-Gaussian structure, is thus limited. In the 1990s, the simultaneous popularization of Markov chain Monte Carlo methods (Gilks et al., 1995), including the freely available BUGS software (Lunn et al., 2009), and high speed desktop computing considerably expanded the diversity of possible SSMs to include non-Gaussian and nonlinear formulations (e.g., Meyer and Millar, 1999). Further developments have advanced fitting procedures in both Bayesian and frequentist frameworks (de Valpine, 2004; Ionides et al., 2015; Kristensen et al., 2016; Monnahan et al., 2017). These methods provide the means to fit increasingly complex SSMs with multiple data streams (e.g., Hobbs et al., 2015) and hierarchical levels (e.g., Jonsen et al., 2005).

While advancements in fitting SSMs have changed how we model time series in ecology, the computational burden required to fit some of these models is often high enough that comparisons between multiple SSMs can be difficult, and the complex structure of some SSMs complicates model validation and diagnostics. In the ecological literature, there has been a recent interest in model comparison and validation for hierarchical models or models for datasets with complex dependence structure (Hooten and Hobbs, 2015; Roberts et al., 2017; Conn et al., 2018). In line with this, new validation tools for SSMs are being developed (Thygesen et al., 2017).

While SSMs are powerful tools for modeling ecological time series, the fitting procedures may seem prohibitively complex to many practitioners. The variety of inference procedures and tools that can be used to fit SSMs (Harvey, 1990; Doucet et al., 2001; Durbin and Koopman, 2012; Ionides et al., 2015; Kristensen et al., 2016) may bewilder all but the most quantitative ecologists, limiting their capacity to formulate, fit, and evaluate their own SSMs. While there are some popular application-specific `R` (R Core Team, 2019) packages with functions to fit specialized SSMs (e.g., `MARSS` for multivariate NDLMs, Holmes et al., 2012; `bsam`, for animal movement, Jonsen et al., 2005), few ecologists are aware of the full range of SSMs that can be fitted with such packages. In addition, these packages may be inadequate for the data-at-hand, especially when using SSMs to answer novel questions or with new data types. A further complication with the application of SSMs is the potential for estimability issues where some states or parameters cannot be estimated well, or at all, given the available data (Auger-Méthé et al., 2016). For example, such estimability issues may arise because the formulation of a SSM is too complex for the data (e.g., the time resolution of the process model is too fine relative to the time resolution of the observations). While there has been some effort to provide a general, and easy to use, set of tools for ecologists to fit SSMs to their data (e.g., King et al., 2009; de Valpine et al., 2017), the available tools and array of choices may be overwhelming to those with little familiarity with SSMs. Given these challenges and the recent advancements in inference methods and model diagnostics, we believe the time is ripe to provide a review of these developments for scientists wanting to fit SSMs to ecological time series.

In this review, we first describe SSMs through a set of examples and discuss what types of data should lead ecologists to consider a SSM (Section 2). Next, we review the different inference methods that can be used to fit a SSM to data (Section 3). We then discuss how one can assess whether a SSM suffers from estimability or identifiability issues (Section 4). Lastly, we describe model selection procedures (Section 5) and diagnostic tools that can be used to verify whether a model is adequate (Section 6), crucial steps that are often ignored. This review is accompanied by an in-depth tutorial that provides examples

of how one can use R (R Core Team, 2019) to fit, and validate, SSMs with various inference methods. We believe this review will give a strong foundation to ecologists interested in learning about SSMs and hope it will provide new tools to veteran SSM users interested in inference methods and model validation techniques.

2 What is a SSM?

SSMs models are hierarchical models for time series that are particularly useful when we want to disentangle process variation from observation error. They model two time series: an observation and a state time series. The observation time series are the data (e.g., fish counts through time), while the state time series aims to represent the true, but unknown or hidden, states of nature (e.g., the associated true population size of the fish stock). The observations could be univariate, as when we have only one estimate of the total population size each year (e.g., see Section 2.2), or multivariate, as when we have annual counts for different age classes (e.g., see Section 2.3). The observations are imperfect measures of an underlying time series and we refer to the individual values of the hidden time series as states. While we generally collect data at discrete points in time, the underlying biological process can be conceptualized in discrete or continuous time, and SSMs can allow for either representation (McClintock et al., 2014).

2.1 A toy example: normal dynamic linear model

To formalize the description of SSMs, we start by describing a simple, toy example. It models a time series of univariate observations, denoted y_t, made at discrete and evenly-spaced points in time t ($t = 1, 2, \ldots, T$). The time series of states, denoted z_t, is defined at the same time points t as the observations. Our model is a simple normal dynamic linear model (NDLM), thus process variance and observation error are modeled with Gaussian distributions and both time series are modeled with linear equations.

To model these two time series, SSMs make two main assumptions. First, SSMs assume that the state time series evolves as a Markov process (Aeberhard et al., 2018). This Markov process, which is generally of first-order, is a relatively simple way to incorporate temporal dependence. For our toy model, this means that the state at time t, z_t, depends only on the state at the previous time step, z_{t-1}. Second, SSMs assume that the observations are independent of one another once we account for their dependence on the states. More formally, we say that, given the corresponding state z_t, each observation y_t is conditionally independent of all other observations, y_s, $s \neq t$. Thus, any marginal dependence between observations is the result of the dependence between hidden states (Aeberhard et al., 2018). For our toy model, this means that y_t is independent of y_{t-1}, and all other observations, once we account for the dependence of y_t on z_t (Fig. 1A). In a population dynamics context, this could be interpreted to mean that the values of observations are autocorrelated only because the true population size of the animal is autocorrelated through time and the discrepancy between the true population size and the observation is not correlated in time. We can see

this structure in the equations for our toy SSM:

$$z_t = \beta z_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}\left(0, \sigma_p^2\right), \tag{1}$$

$$y_t = \alpha z_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\left(0, \sigma_o^2\right). \tag{2}$$

The autocorrelation in the states is captured by the parameter β. The observations are a function of the states only and the parameter α allows the observation at time t to be a biased estimate of the state at time t. The process variation and observation error are both modeled with normal distributions but have different standard deviations (σ_p^2 and σ_o^2). We have not defined the state at time 0, z_0, and many authors will provide an additional equation, often referred as the initialization equation, which describes the probability of different values of z_0 (e.g., $z_0 \sim \mathrm{N}(0, \sigma_{z_0}^2)$). For our toy example, we view z_0 as a fixed and unknown parameter.

The terminology used to refer to the process and observation equations varies in the literature. A process equation can be referred as a process model, state equation, state model, or transition equation. An observation equation can be referred as an observation model, measurement equation, or measurement model. In this paper, we generally use the terms 'process equation' and 'observation equation' respectively, and we often describe SSMs with equations that combine a deterministic function with a stochastic component (e.g., Eqs. 1-2).

To further reveal the dependence structure and understand how to fit SSMs to data, it can help to additionally characterize a SSM in terms of probability distributions for the states and the observations, e.g.:

$$f(z_t|z_{t-1}, \boldsymbol{\theta}_p), \qquad t = 1, \ldots, T, \tag{3}$$

$$g(y_t|z_t, \boldsymbol{\theta}_o), \qquad t = 1, \ldots, T. \tag{4}$$

In the case of our toy model, f and g are two normal probability density functions, while $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_o$ are vectors of parameters associated with each equation (i.e., $\boldsymbol{\theta}_p = (\beta, \sigma_p^2)$, $\boldsymbol{\theta}_o = (\alpha, \sigma_o^2)$). This definition clearly demonstrates that states are random variables and thus that SSMs are a type of hierarchical model.

One of the goals of fitting a SSM to data is typically to estimate unknown parameters. Here, to contrast them with the states, we refer to these as the fixed parameters and denote them together as $\boldsymbol{\theta}$. For example, here $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_o, z_0)$, thus α, β, σ_p^2, σ_o^2, z_0 in Eqs. 1-2. A second important goal is to estimate the unobserved states, $\mathbf{z}_{1:T} = (z_1, z_2, ..., z_T)$, where T is the length of the time series. The notation $1{:}t$, which we use throughout, refers to the sequence 1, 2, ..., t. Fig. 1B shows how close estimates of the states can be to their true values.

SSMs can be fitted using frequentist or Bayesian approaches to statistical inference. When using a Bayesian approach, a third level is added to the model hierarchy: the prior distribution(s) for the fixed parameters denoted by the probability density function, $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ are known values called hyperparameters. While we refer to $\boldsymbol{\theta}$ as fixed parameters to differentiate them from the states, in Bayesian inference $\boldsymbol{\theta}$ is a vector of random variables. In Section 3, we discuss in detail how we can use these probabilistic descriptions of the model for inference.

This simple linear and normal model is a useful toy example that we use throughout the main text to explain the concepts associated with fitting and validating SSMs. We also

use this model in Appendix S1 to demonstrate how to use R to fit a SSM to data. While the simplicity of this toy example makes it a useful teaching tool, it is not a particularly useful model for ecology. We now turn to the description of a set of ecological SSMs.

2.2 Handling nonlinearity and nonnormality in population dynamics models

We use a set of simple univariate SSMs for population dynamics to demonstrate that even simple ecological models can rarely be blindly modeled as NDLMs. Jamieson and Brooks (2004) applied multiple SSMs to abundance estimates from North American ducks obtained through annual aerial counts. We start with one of their simplest models, for which the process equation is the stochastic logistic model of Dennis and Taper (1994). This model allows for density dependence, i.e., a change in growth rate dependent on the abundance in the previous year:

$$z_t = z_{t-1} \exp\left(\beta_0 + \beta_1 z_{t-1} + \epsilon_t\right), \qquad \epsilon_t \sim \mathrm{N}(0, \sigma_p^2). \tag{5}$$

As in the toy example above, z_t denotes the true hidden state, in this case the number of ducks in year t. The parameter $\beta_0 > 0$ determines the median rate of population growth when population size is 0. The parameter $\beta_1 \leq 0$ determines how much the growth rate decreases depending on the population size at time $t-1$, with $\beta_1 < 0$ indicating density dependence. The process variation, ϵ_t, is normally distributed and represents the random change in growth rate each year. The observations y_t are modeled as unbiased estimates of the true population size z_t with a normally distributed error:

$$y_t = z_t + \eta_t, \qquad \eta_t \sim \mathrm{N}(0, \sigma_o^2). \tag{6}$$

Even though the observation equation is linear with a Gaussian error, the SSM is not a NDLM because of the exponent in the process equation (Eq. 5). Jamieson and Brooks (2004) reparametrized the model by modeling the population size on a logarithmic scale, $w_t = \log(z_t)$, which resulted in the following formulation:

$$w_t = w_{t-1} + \beta_0 + \beta_1 \exp(w_{t-1}) + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \sigma_p^2), \tag{7}$$

$$y_t = \exp(w_t) + \eta_t, \qquad \eta_t \sim \mathrm{N}(0, \sigma_o^2). \tag{8}$$

While such reparametrization can sometimes linearize the model, in this case the model remains nonlinear. Jamieson and Brooks (2004) use a Bayesian framework to fit this model, see their original paper for the description of the priors.

The modeling of density dependence has been extensively debated in the literature, and Jamieson and Brooks (2004) also explored an alternative process equation, a stochastic Gompertz model:

$$z_t = z_{t-1} \exp\left(\beta_0 + \beta_1 \log z_{t-1} + \epsilon_t\right), \qquad \epsilon_t \sim \mathrm{N}(0, \sigma_p^2), \tag{9}$$

which assumes that the per-unit-abundance growth rate depends on the log abundance, $\log(z_{t-1})$, instead of the abundance, z_{t-1} (Dennis and Taper, 1994). Such a model is often

linearized as follows:

$$w_t = \beta_0 + (1 + \beta_1)w_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \sigma_p^2), \tag{10}$$

$$g_t = w_t + \eta_t, \qquad \eta_t \sim \mathrm{N}(0, \sigma_o^2), \tag{11}$$

where $w_t = \log(z_t)$ and $g_t = \log(y_t)$ are the logarithms of the states and observations, respectively (e.g., Dennis et al., 2006). The linearized version of this model is a NDLM that can be fitted with tools such as the Kalman filter (Dennis et al., 2006). This statistical convenience may have contributed to the uptake of the stochastic Gompertz SSM in the literature. However, it may not always be adequate to assume that the growth rate depends logarithmically on population density (Dennis and Taper, 1994).

Viljugrein et al. (2005) extended these models to incorporate external covariates. As the fluctuating availability of wetlands can influence the population growth of ducks, they included the number of ponds in year t, p_t, as a covariate in the process equation. They modified the Gompertz stochastic model as follows:

$$w_t = (1 + \beta_1)w_{t-1} + \beta_2 p_t + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \sigma_p^2). \tag{12}$$

They used the untransformed observation equation (Eq. 8) and fit the model with a Bayesian approach (see Viljugrein et al., 2005, for the description of priors). This set of examples shows how easy it is to adapt and extend models in the SSM framework. While even simple ecological models may only be linear with transformations and assumptions, Jamieson and Brooks (2004) and Viljugrein et al. (2005) showed that accounting for the observation errors improved the inference regardless of the process equation.

2.3 Joining multiple data streams in population dynamics models

SSMs are often used in fisheries stock assessments (Aeberhard et al., 2018). Here, to showcase the ability of SSMs to handle multiple data streams and multivariate data, we present a simplified version of a state-space stock assessment model described by Nielsen and Berg (2014). The first data stream comprises the catches at age, $C_{a,t}$, representing how many fish from each age class a are caught in the commercial fishery in each year t. The second data stream includes age-specific indices from distinct scientific surveys, $I_{a,t,s}$, where s identifies the survey. These distinct surveys can occur in different years and only capture some portion of the age classes.

The hidden state in each year t is a vector combining the log-transformed stock sizes, $N_{a,t}$, and fishing mortality rates, $F_{a,t}$, for each age class: $\boldsymbol{z}_t = (\log N_{1,t}, \ldots, \log N_{\mathrm{A}^+,t}, \log F_{1,t}, \ldots, \log F_{\mathrm{A}^+,t})'$, where A^+ represents the oldest age class, which is often a cumulative age class that accounts for all fish older than a certain age. Just as for the toy example, the process equations describe the state in year t as a function of the state in year $t-1$. However, unlike the toy model, we no longer have a single process equation. We have instead

a set of equations describing recruitment, survival, and mortality as follows:

$$\log(N_{1,t}) = \log(N_{1,t-1}) + \epsilon_{N_{1,t}}, \tag{13}$$

$$\log(N_{a,t}) = \log(N_{a-1,t-1}) - F_{a-1,t-1} - M_{a-1,t-1} + \epsilon_{N_{a,t}}, \qquad 2 \leq a < A^+, \tag{14}$$

$$\log(N_{A^+,t}) = \log(N_{A^+-1,t-1} \exp^{-F_{A^+-1,t-1}-M_{A^+-1,t-1}} + N_{A^+,t-1} \exp^{-F_{A^+,t-1}-M_{A^+,t-1}}) + \epsilon_{N_{A^+,t}}, \tag{15}$$

$$\log(F_{a,t}) = \log(F_{a,t-1}) + \epsilon_{F_{a,t}}, \qquad 1 \leq a \leq A^+. \tag{16}$$

While the age and year specific log fishing mortality rates, $\log F_{a,t}$, are considered states that evolve as a random walk through time, the equivalent natural mortality rate, $\log M_{a,t}$, is assumed known from outside sources. The process variation for all of these equations are assumed to be Gaussian with zero mean, but they differ in their variance and covariance parameters. For recruitment and survival, the variation is assumed to be uncorrelated, e.i., $\epsilon_{N_{1,t}} \sim \mathrm{N}(0, \sigma^2_{N_{a=1}})$, and $\epsilon_{N_{a,t}}$ and $\epsilon_{N_{A^+,t}} \sim \mathrm{N}(0, \sigma^2_{N_{a>1}})$. However, for fishing mortality, the yearly variation is assumed to be correlated across age classes, i.e., $\boldsymbol{\varepsilon}_{F_t} = (\epsilon_{F_1,t}, ..., \epsilon_{F_{A^+},t})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_F)$. The covariance matrix, $\boldsymbol{\Sigma}_F$, is assumed to have an auto-regressive order 1, AR(1), correlation structure, where each element $\Sigma_{a,\tilde{a}} = \rho^{|a-\tilde{a}|}\sigma_a\sigma_{\tilde{a}}$ and $\rho^{|a-\tilde{a}|} = \mathrm{cor}(\epsilon_{F_{a_t}}, \epsilon_{F_{\tilde{a}_t}})$.

The observation equations relate the time series of observed age-specific log-catches, $\log C_{a,t}$, and age-specific log-indices from scientific surveys, $\log I_{a,t,s}$, to the unobserved states, $\boldsymbol{z}_t$, as follows:

$$\log C_{a,t} = \log\left(\frac{F_{a,t}}{Z_{a,t}}(1 - e^{-Z_{a,t}})N_{a,t}\right) + \eta_{a,t,c}, \tag{17}$$

$$\log I_{a,t,s} = \log\left(Q_{a,s}e^{-Z_{a,t}\frac{D_s}{365}}N_{a,t}\right) + \eta_{a,t,s}, \tag{18}$$

where $Z_{a,t}$ is the total mortality rate of age class a in year t (i.e., $Z_{a,t} = M_{a,t} + F_{a,t}$), D_s is the number of days into the year when the survey s is conducted, and each $Q_{a,s}$ is a model parameter describing the catchability coefficient. The observation error terms, $\eta_{a,t,c}$ and $\eta_{a,t,s}$, are assumed to be Gaussian distributed and their variances are designed such that the catch data sources, and the two scientific surveys have their own covariance matrix. Accounting for this observation uncertainty is very important. Much effort goes into ensuring that stock assessment models assign the correct relative weightings to each of the different information sources (see Berg and Nielsen, 2016, where they compare different observation covariance structures for four ICES stocks). For example, the stock assessment model for the Faroe haddock stock assumes the catches are independent across ages, whereas each survey index is assumed to have an $AR(1)$ correlation structure across ages.

Because the data available for the older age class are often limited, as it is the case for the Faroe haddock stock, a few additional assumptions are often made. In addition to having a cumulative age class, A^+, the fishing mortality of fish of the oldest age class is often assumed to be equal that of the previous age class: $F_{A^+,t} = F_{A^+-1,t}$. Similarly, it is often assumed that a survey's catchability is the same for the two oldest age classes: $Q_{A^+,s} = Q_{A^+-1,s}$. Such assumptions tend to be biologically reasonable since older fish tend to be of similar

size and are thus likely to be caught by the same type of fishing gear. Fisheries scientists are often interested in calculating quantities like the spawning stock biomass. We can derive these quantities from this model by combining the state estimates with additional biological information, such as mean weight in stock and the proportion of mature females in each age class. The model described above is implemented in the R package `stockassessment` (available on GitHub at https://github.com/fishfollower/SAM).

2.4 Accounting for the complex structure of movement data

The first difference correlated random walk model (DCRW, Jonsen et al., 2005) was one of the earliest SSMs for animal movement and is still commonly used. It was developed for Argos doppler shift location data (Jonsen et al., 2005). Argos tags have been frequently used to track marine animals because they overcome some of the challenges associated with using conventional GPS units in an aquatic environment. Unlike GPS however, Argos data have large observation errors (mean error ranging from 0.5-36km, Costa et al., 2010), which has led marine ecologists to use SSMs to account for this added level of stochasticity.

The two time series modeled by the DCRW are the time series of N observed Argos locations at irregular time intervals i, $\mathbf{y}_i = \left[\begin{smallmatrix} y_{i,lon} \\ y_{i,lat} \end{smallmatrix}\right]$, and the time series of T true locations of the animal at regular time intervals t, $\mathbf{z}_t = \left[\begin{smallmatrix} z_{t,lon} \\ z_{t,lat} \end{smallmatrix}\right]$. A simplified version of the DCRW model is:

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \gamma(\mathbf{z}_{t-1} - \mathbf{z}_{t-2}) + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}), \qquad 1 \leq t \leq T, \tag{19}$$

$$\mathbf{y}_i = (1 - j_i)\mathbf{z}_{t-1} + j_i\mathbf{z}_t + \boldsymbol{\eta}_i, \qquad \boldsymbol{\eta}_i \sim \mathrm{T}(\boldsymbol{\Psi} \circ \mathbf{S}_i, \mathbf{D}_i), \qquad 1 \leq i \leq N, \tag{20}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{\epsilon,lon} & \rho\sigma_{\epsilon,lon}\sigma_{\epsilon,lat} \\ \rho\sigma_{\epsilon,lat}\sigma_{\epsilon,lon} & \sigma^2_{\epsilon,lat} \end{bmatrix}, \tag{21}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{lon} \\ \psi_{lat} \end{bmatrix}, \tag{22}$$

$$\mathbf{S}_i = \begin{bmatrix} s_{lon,q_i} \\ s_{lat,q_i} \end{bmatrix}, \tag{23}$$

$$\mathbf{D}_i = \begin{bmatrix} df_{lon,q_i} \\ df_{lat,q_i} \end{bmatrix}. \tag{24}$$

The process equation assumes that the animal's location at time t, $\mathbf{z}_t$, is not only dependent on the previous location, $\mathbf{z}_{t-1}$, but also on the animal's previous movement in each coordinate, $\mathbf{z}_{t-1} - \mathbf{z}_{t-2}$. The parameter γ can take values between 0 and 1 (i.e., $0 \leq \gamma \leq 1$), and controls the degree of correlation between steps. Values close to 0 mean that the movement only depends on the previous location and value close to 1 means that the animal has a tendency to move at the same speed and in the same direction as the previous step. The covariance matrix for the process variation, $\boldsymbol{\Sigma}$, allows for covariance between longitude and latitude, but in many instances it is simpler to assume that $\rho = 0$.

The observation equations accounts for various characteristics of the Argos data. Because data are taken at irregular time intervals, i, the true location of the animal is linearly

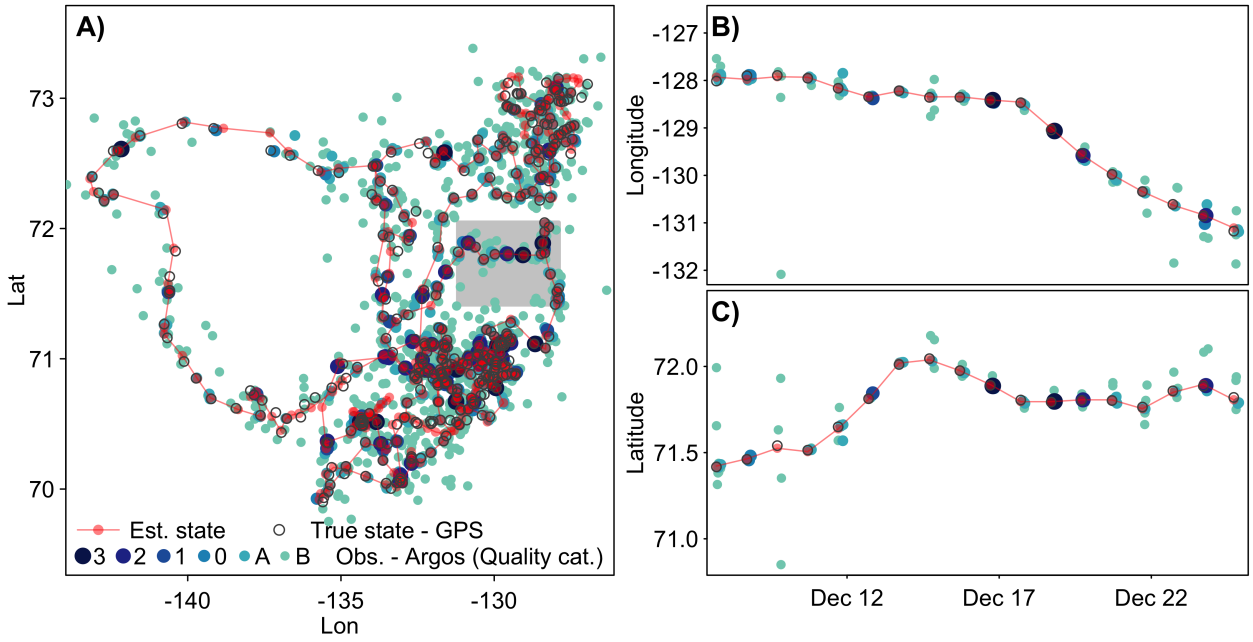

Figure 2: The DCRW (Eqs. 19-24) fitted to a polar bear Argos track and validated with GPS data. Panel A) maps the observed Argos data points in shades of blue and green (darker colors representing higher quality observations), the estimated true locations in red, and the true locations of the bear (GPS data) in dark grey circles. Panels B) and C) show the longitudes and latitudes of a small subset of the time series (indicated by a grey box in the map). These panels highlight the temporal clustering of observations, which likely helped the state estimation procedures.

interpolated to the time of the observation, with j_i representing the proportion of the regular time interval between $t-1$ and t when the observation $\mathbf{y}_i$ was made. Because the data often have outliers, the measurement errors are modeled with t-distributions, which have fat tails. Finally, each Argos location has a quality rating, which classifies the location into one of six categories, q_i. These categories differ in the size of their errors and are associated with different bivariate t-distributions: $\mathrm{T}(\mathbf{\Psi} \circ \mathbf{S}_i, \mathbf{D}_i)$. In particular, each category is associated with a unique scale parameter, s_{c,q_i}, and degrees of freedom, df_{c,q_i}, for each coordinate (i.e., $c = lon$ or lat). Instead of estimating these 12 parameters, many researchers fix them to known values derived from field experiments (e.g., Jonsen et al., 2005). To allow for discrepancies between the fixed values and the ones that may fit the data best, we can add correction factors, ψ_c. In Jonsen et al. (2005), a single correction factor is used for both coordinates. From experience, we have found that longitude and latitude often differ in the degree of correction they need. Thus, we use one correction factor for each coordinate c. Note that the Hadamard product, $\circ$, simply states that we perform entrywise multiplication of the correction factors to the scale parameters, i.e., $\psi_c s_{c,q_i}$, for $c = (lon, lat)$. Fig. 2 shows the DCRW fitted to a polar bear track, and Appendix S1 provides the code to fit the model.

2.5 Accommodating continuous-time processes in movement models

So far we have only described SSMs where the hidden state evolves in discrete time steps. However, in some cases, it may be more appropriate to assume that the biological process occurs in continuous time. For example, in the previous section we described a SSM that models movement in discrete time steps. However, animal movement generally occurs continuously through time and modeling it as such can simplify the way irregularly-timed observations are modeled (McClintock et al., 2014). Using the movement model of Johnson et al. (2008), we showcase in this section how SSMs can accommodate continuous-time process equations.

The SSM of Johnson et al. (2008) models the movement of an animal with a continuous-time correlated random walk (CTCRW). The process equation is formulated in terms of how changes in velocity v through time affect the location μ of an animal. While the model works in 2-dimensional space, the velocity processes in each coordinate are considered independent. For simplicity, we therefore only describe the process in one coordinate. Let velocity at time t be $v(t)$. Change in velocity over time is described using a type of diffusion model called an Ornstein-Uhlenbeck (OU) process. At time $t + \Delta$, velocity is:

$$v(t+\Delta) = \gamma + e^{-\beta\Delta}\left(v(t) - \gamma\right) + \zeta(\Delta), \qquad \zeta(\Delta) \sim \mathrm{N}\left(0, \sigma^2_{OU}(1 - e^{-2\beta\Delta})/2\beta\right), \tag{25}$$

where γ is the long-term average velocity (the "drift"), β (where $\beta > 0$) indexes how fast the velocity tends to this average, and $\zeta(\Delta)$ is a random perturbation with variability indexed by σ^2_{OU}. Usually, $\gamma = 0$ and we make that assumption here. The new velocity essentially consists of a weighted average of the drift and the previous velocity plus a random deviate whose variance increases with Δ. After time difference $\Delta = 3/\beta$, $v(t)$ and $v(t + \Delta)$ are almost independent with correlation approximately 0.05.

While the core of the process model describes changes in velocity, the observations are modeled in terms of locations and thus we want our state at time t, $z(t)$, to keep track of both the velocity and the location, $z(t) = \begin{bmatrix} \mu(t) \\ v(t) \end{bmatrix}$. Locations, $\mu(t)$, can be related to velocities, $v(t)$, by noting that

$$\mu(t+\Delta) = \mu(t) + \int_t^{t+\Delta} v(u)du. \tag{26}$$

By integrating over the velocity (Eq. 25), Johnson et al. (2008) showed that the change in location in time Δ is:

$$\mu(t+\Delta) = \mu(t) + v(t)\left(\frac{1 - e^{-\beta\Delta}}{\beta}\right) + \xi(\Delta), \qquad \xi(\Delta) \sim \mathrm{N}\left(0, \frac{\sigma^2_{OU}}{\beta^2}\right). \tag{27}$$

The covariance between the two components of the process variability is

$$Cov(\xi(\Delta), \zeta(\Delta)) = \frac{\sigma^2_{OU}}{2\beta^2}\left(1 - e^{-\beta\Delta}\right)^2. \tag{28}$$

Because Δ can take any non-negative value, we can keep track of the states at any time intervals, thus easily accommodating observations collected at irregular time intervals.

To simplify the notation, let t_i denote the times when the i^{th} observation, y_i, occurred. Let us also equate z_i with $z(t_i)$, and equally for μ_i and v_i, and finally let $\Delta_i = t_i - t_{i-1}$. If we combine the equations above, the final process equations in matrix notation form are:

$$\mathbf{z}_i = \begin{bmatrix}\mu_i \\ v_i\end{bmatrix} = \begin{bmatrix}1 & (1-e^{\beta\Delta_i})/\beta \\ 0 & e^{-\beta\Delta_i}\end{bmatrix}\begin{bmatrix}\mu_{i-1} \\ v_{i-1}\end{bmatrix} + \boldsymbol{\eta}_i, \qquad \boldsymbol{\eta}_i \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_p^2\right), \tag{29}$$

where

$$\boldsymbol{\Sigma}_p^2 = \begin{bmatrix}\frac{\sigma_{OU}^2}{\beta^2} & \frac{\sigma_{OU}^2}{2\beta^2}\left(1-e^{-\beta\Delta_i}\right)^2 \\ \frac{\sigma_{OU}^2}{2\beta^2}\left(1-e^{-\beta\Delta_i}\right)^2 & \sigma_{OU}^2(1-e^{-2\beta\Delta_i})/2\beta\end{bmatrix}. \tag{30}$$

The observation equation can be chosen as usual, for example as a normal distribution around the true location:

$$y_i = \mu_i + \epsilon_i, \qquad \epsilon_i \sim \mathrm{N}\left(0, \sigma_o^2\right). \tag{31}$$

The SSM defined by Eqs. 29-31 is a linear Gaussian SSM and can therefore be fitted with a Kalman filter.

This model allows various extensions to include different aspects of animal movement. For example, Johnson et al. (2008) show how for harbor seals, haul-out behavior of tagged animals can be incorporated using data on how long the tag has been dry. They model β as an increasing function of dry time, thereby decreasing velocity v and slowing down the animal when it hauls out. To account for the large outliers associated with Argos data one can use a t-distribution (see Section 2.4), in which case the Kalman filter will no longer be adequate and other fitting methods will be required (Albertsen et al., 2015).

2.6 Integrating count and categorical data streams in a multi-level process model

As shown above, the SSM framework provides the flexibility to create complex models. Here, using the model of Schick et al. (2013), we demonstrate how distinct types of observations, including count and categorical data, can be integrated in a single SSM for the health, monthly movement and survival of North Atlantic right whales.

Schick et al. (2013) extracted two types of data from photographic observations of individual whales. The first type, denoted $y_{i,t,k}$, is the number of sightings of individual i in geographic zone k and month t. The second type, denoted $H_{q,i,t}$, is the value for the q^{th} visual health metric for individual i in month t. The six visual health metrics (e.g., skin condition, entanglement status, whale lice infestation level) are on ordinal scales, each with two or three levels. In addition, information on the search effort in each geographic zone is known, and other ancillary data, such as age, are used.

Three process equations model the health, survival, and monthly movement of each individual whale. Whale i in month t is characterized by its age $a_{i,t}$, health status $h_{i,t}$ (defined on an arbitrary scale), and location $k_{i,t}$ (one of 9 geographic zones). Health status, $h_{i,t}$, is modeled as a function of previous health status and age:

$$h_{i,t} = \alpha_0 + \alpha_1 h_{i,t-1} + \alpha_2 a_{i,t-1} + \alpha_3 a_{i,t-1}^2 + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim \mathrm{N}(0, \sigma^2). \tag{32}$$

When $\alpha_2 > 0$ and $\alpha_3 < 0$, the quadratic age term allows for the fact that health status, and thus survival probability, initially increases but declines with advanced age. Survival from month t to $t+1$ is modeled as a Bernoulli process, with survival probability modeled with a logit link function:

$$\text{logit}(s_{i,k,t}) = \beta_{0,k} + \beta_1 h_{i,t}. \tag{33}$$

Here, $\beta_{0,k}$ denotes the fixed effect for zone k and β_1 the relationship with health. Hence, survival probability depends on health and on the occupied zone, potentially allowing researchers to identify the geographic zones associated with reduced survival.

While a whale is assumed to stay in a single zone during the month, it can move from one zone to another each month. The monthly location of each individual, $z_{i,t}$, is only known when the individual is sighted that month. The subscript t, throughout, represents the number of months since the beginning of the time series. For each month of the year (January, ..., December), denoted $t^{(u)}$, the movement between zones is modeled with a transition matrix, where each element, $m_{j,k,t^{(u)}}$, describes the probability of moving from zone j to zone k (i.e. $m_{j,k,t^{(u)}} = \Pr(z_{i,t^{(u)}+1} = k | z_{i,t^{(u)}} = j)$). As the complete geographic range of the whales is assumed to be covered by these zones, a living whale will be in one of the nine distinct zones at time $t+1$, $\sum_{k=1}^{9} m_{j,k,t^{(u)}} = 1$. The changes in transition probabilities between the months of the year, $t^{(u)}$, allow for the modeling of seasonal migration.

The model has two sets of observation equations. First, the number of sightings of whale i in location k and month t is modeled as a Poisson random variable.

$$y_{i,k,t} \sim \text{Pois}(\lambda_i E_{k,t}), \tag{34}$$

where $E_{k,t}$ denotes the search effort in zone k and month t and λ_i denotes the expected number of sightings of individual i per unit effort. Second, each visual health metric is modeled as coming from a multinomial logit distribution. The probability of being in each level of the q^{th} health metric depends on the true health status, $h_{i,t}$, and the model is structured so as to ensure that the ordinal aspect of the variables is respected (i.e., that lower values means lower health). For example, if the health metric $H_{q,i,t}$ has a three-level ordinal scale, the observation equations for this metric are:

$$H_{q,i,t} \sim \text{Multinom}(1, \boldsymbol{p}_{q,i,t}) \tag{35}$$

$$\text{logit}(p_{q,i,t,1}) = \log\left(\frac{p_{q,i,t,1}}{p_{q,i,t,2} + p_{q,i,t,3}}\right) = c_{q,0,1} + c_{q,1,1} h_{i,t} \tag{36}$$

$$\text{logit}(p_{q,i,t,1} + p_{q,i,t,2}) = \log\left(\frac{p_{q,i,t,1} + p_{q,i,t,2}}{p_{q,i,t,3}}\right) = c_{q,0,2} + c_{q,1,2} h_{i,t} \tag{37}$$

$$p_{q,i,t,3} = 1 - p_{q,i,t,1} - p_{q,i,t,2}. \tag{38}$$

The vector $\boldsymbol{p}_{q,i,t}$ contains the probabilities with which an individual with true health $h_{i,t}$ is assigned a specific health level. Forcing the parameters $c_{q,0,1} < c_{q,0,2}$ and $c_{q,1,1} < c_{q,1,2}$, and modeling cumulative probabilities (Eq. 37) ensure that the order of the levels is accounted for. The probabilistic nature of the model allows health metrics to depend on the true health status but to be observed with error.

This model allows inference about various aspects of North Atlantic right whales. For example, we can learn which visual health metrics show the strongest links to underlying health, whether geographic regions (and thereby human activity) have an impact on survival, and at which times of the year movement to certain zones occurs. While this model may seem complex at first sight, each of the individual hierarchical levels are relatively straightforward. SSMs provide the framework to link these levels.

2.7 Modeling discrete state values with hidden Markov models

There is an important class of SSMs that we largely ignore in this review: hidden Markov models (HMMs). HMMs are a special class of SSMs, where the states are discrete (generally categorical with a finite number of possible values) rather than continuous (Langrock et al., 2012). They have gained popularity in ecology, in particular in movement ecology, where they are used to model an animal that switches between distinct behavioral modes (Langrock et al., 2012).

The two main characteristics of HMMs are that 1) each observation is assumed to be generated by one of N distributions, and 2) that the hidden state sequence that determines which of the N distributions is chosen at time t is modeled as a Markov chain, where the probability of being in each mode at time t depends only on the state value at the previous time step (Langrock et al., 2012). For example, one can use an HMM to model the movement of an animal that switches between two behavioral modes (N=2). The HMM could model the time series of observed step lengths and turning angles with Weibull and wrapped Cauchy distributions. Each behavioral mode would be associated with a different parameter set for these distributions, and, thus, results in different movement patterns. The key advantage of HMMs is their mathematical simplicity: what would be a high-dimensional integration in a SSM with continuous state values (see Section 3) is now a simple sum. As such, having a finite number of possible state values (i.e., discrete states) significantly simplifies the analysis (Langrock et al., 2012). The mathematical simplicity of HMMs makes them highly attractive, and various tools have been developed to fit HMMs to data (Zucchini et al., 2016). While some of the topics we discuss in this paper are relevant to HMMs, HMMs are easier to fit to data than other SSMs. We refer readers interested in HMMs to the book of Zucchini et al. (2016).

Some SSMs will include both discrete and continuous states. For example, the model presented in Section 2.4 was originally developed to model the movement of animals tracked with Argos data that switched between two behavioral modes (Jonsen et al., 2005). Instead of having a single γ parameter that controls how correlated the steps are (e.g., Eq. 19), this model has two parameters, γ_{b_t}, where b_t is the behavioral mode at time t, which can take one of two values. The switch between the behavioral modes is modeled with a simple Markov chain, where the behavioral mode value at time t only depends on that of the previous time step (i.e., $P(b_t = j|b_{t-1} = i) = \alpha_{ij}$ and $\sum_{j=1}^{2} \alpha_{ij} = 1$ for $i, j = 1, 2$). When γ_1 is close to 0 and and γ_2 is close to 1, then the movement path switches between tortuous and directed movement. While simple HMMs are simpler to fit to data, SSMs that have both continuous and discrete states are not, and thus such SSMs require the use of some of the more complex fitting methodology described in Section 3.

2.8 Is an SSM necessary?

The examples above have demonstrated the advantage of accounting for process variation and observation error, and why the flexibility of SSMs makes them attractive analytical tools. However, their complexity may deter ecologists, or result in models that are difficult to fit to data. So one may ask: when is it necessary to use a SSM? When could we use a simpler alternative? To our knowledge there has not been an in-depth investigation into when SSMs are necessary, but their use has been primarily promoted to analyze time-series data that are affected by both relatively large process variation and observation error. In fisheries science, early papers on SSMs showed that they were particularly superior to simpler models when both the process variance and the observation error are large (e.g., de Valpine and Hastings, 2002). When one of the sources of stochasticity is small, and the model dynamics are not too complex, simpler models that account for either just the process variance or observation error give adequate results (de Valpine and Hastings, 2002). A key point, however, is that the simpler model performs adequately only if the model is well specified with regards to which source of stochasticity is most important. Thus, using a process variance-only model, is only suitable if we are certain that the observation error is negligible. Similarly, using an observation error-only model, is only suitable if we are certain that the process variation is small. Chapter 11 of Bolker (2008) provides further discussion of the merits of modeling process variance only, observation error only, and both simultaneously.

For all of the examples above, the observation error and process variation are so large that we would not hesitate to recommend the use of a SSM. This is not always the case. For example, while a model similar to the DCRW presented in Section 2.4 could be fitted to GPS data rather than Argos data, the level of accuracy of GPS data may not require the use of a SSM to infer better location estimates. In fact, GPS data can be used to validate the fit of the DCRW to Argos data (see Fig. 2, Appendix S1, and Auger-Méthé et al., 2017). As this topic has been relatively neglected in the literature, we believe there is a need for guidelines on when the levels of process variation and observation error are such that SSMs are recommended.

3 Fitting SSMs

The goals of fitting a SSM to data include estimating the parameters, $\boldsymbol{\theta}$, the states, $\mathbf{z}$, or both. In ecology, we regularly need to estimate both, as we rarely know the value of $\boldsymbol{\theta}$ *a priori* and estimating the states is often a primary goal of the analysis. In movement ecology, researchers often fit SSMs similar to that described in Section 2.4 because the states provide better estimates of the true locations of the animal than the data. In the SSM literature, a distinction is often drawn between three different types of state estimation processes (Shumway and Stoffer, 2016). Using all of the observations, $y_{1:T}$, to estimate the states is referred to as 'smoothing'. Smoothing is common with ecological SSMs, as we often have the complete dataset in hand when we start the analysis. We denote the smoothed state estimate as $\hat{z}_{t|1:T}$, with the subscript $t|1{:}T$ identifying that the state at time t is estimated using the observations from time 1 to T. In the original engineering application and in other fields, states are often estimated while data continues to be collected, so only observations up-

to and including time t, $y_{1:t}$, are used to estimate the state $\hat{z}_{t|1:t}$. This ubiquitous estimation procedure is referred to as 'filtering'. Finally, we can use a subset of the observations that ends s time steps before time t, $y_{1:t-s}$, to predict the state at time t, $\hat{z}_{t|1:t-s}$, a procedure we refer to as 'forecasting'. A common forecast is the one-step-ahead prediction, $\hat{z}_{t|1:t-1}$, which is also used within fitting algorithms (see Section 3.1.1) and to validate models (see Section 6).

The states are random variables, and thus have probability distributions. The states are often referred as random effects or latent variables, and all state estimation procedures are sometimes referred as prediction, rather than estimation, even when smoothing and filtering are used. The inferences about the states can include a variety of summary measures of their probability distributions. Above, the state estimates (e.g., $\hat{z}_{t|1:T}$) referred to point estimates such as the expected value. However, one can also calculate interval estimates (e.g., 95% prediction intervals) and single measures of uncertainty (e.g., standard deviations or variances).

Methods for fitting SSMs can be divided into the two main inferential approaches: frequentist and Bayesian. Frequentists maximize the likelihood, while Bayesians focus on the posterior density. A key difference between the two statistical frameworks is the extent to which they differentiate between the estimation of parameters, $\boldsymbol{\theta}$, and states, $\mathbf{z}$. The frequentist approach considers each parameter as a fixed unknown constant and the states as random variables, while the Bayesian approach considers both parameters and states as random variables. The frequentist approach is generally a two-stage procedure, estimating $\boldsymbol{\theta}$ first and then the states given the estimated $\boldsymbol{\theta}$. In contrast, the Bayesian approach focuses on calculating the joint posterior distribution of states and parameters. As we show below, despite these differences, both approaches involve high-dimensional integration, which is at the crux of the difficulties associated with fitting SSMs to data. The many tools developed for fitting SSMs are essentially different solutions to this high-dimensional integration problem.

3.1 Frequentist approach

When we fit a SSM with a frequentist approach, we search for the parameter and state values that maximize the likelihood, a method called maximum likelihood estimation with the resulting estimates called maximum likelihood estimates (MLEs). For our toy SSM (Eqs. 1-2), the joint likelihood for $\boldsymbol{\theta}$ and $\mathbf{z}_{1:T}$ would be defined as:

$$L_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} g(y_t|z_t, \boldsymbol{\theta}_o) f(z_t|z_{t-1}, \boldsymbol{\theta}_p), \tag{39}$$

where T is the length of our time series and $\boldsymbol{\theta}$ is a vector of (unknown) model parameters that contains the parameters for the process equation, $\boldsymbol{\theta}_p$, and the observation equation, $\boldsymbol{\theta}_o$, and in this example the initial state, z_0. Maximizing the joint likelihood with respect to both parameters and the states is challenging. Instead, one can use a two-step process.

First, to estimate the parameters, we maximize the marginal likelihood:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, L_{\text{M}}(\boldsymbol{\theta}|\mathbf{y}_{1:T}), \tag{40}$$

where

$$L_{\text{M}}(\boldsymbol{\theta}|\mathbf{y}_{1:T}) = \int L_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})\text{d}\mathbf{z}_{1:T}. \tag{41}$$

The key here is that we integrate out the hidden states and thus have a function that only depends on the observations.

Second, we can estimate the hidden states using the conditional distribution of the states given the observations and the estimated parameter values, $p(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \hat{\boldsymbol{\theta}})$; for example, see Sections 3.1.1 and 3.1.3. As an approximation, one can also maximize the joint likelihood (Eq. 39), where we use the MLE values, $\hat{\boldsymbol{\theta}}$, for the fixed parameters:

$$\hat{\mathbf{z}}_{1:T} = \underset{\mathbf{z}_{1:T} \in \mathbf{Z}^T}{\arg\max}\, L_{\text{J}}(\hat{\boldsymbol{\theta}}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T}), \tag{42}$$

where $\mathbf{Z}^T$ is the set of all possible values for the states. This maximization is akin to hierarchical- and pseudo-likelihood methods, where the states are treated as if they were equivalent to fixed parameters in an ordinary likelihood (Aeberhard et al., 2018) and is often used when the parameters are estimated with the Laplace approximation method (see Section 3.1.2). While Eq. 42 treats the parameters as known when estimating the states, one can propagate the estimation variability when reporting the state estimate variance (e.g., see `TMB` function `sdreport`, Kristensen et al., 2016).

The maximization used to estimate the parameters, and thus the states, requires the evaluation of the marginal likelihood, and thus, the computation of the high-dimensional integral of Eq. 41. This computation is difficult to achieve for most SSMs and the frequentist inference methods discussed below are different ways to either evaluate the marginal likelihood (e.g., Kalman filter) or to approximate it (e.g., Laplace and simulation-based approximations).

3.1.1 Kalman filter

For simple linear SSMs with Gaussian errors (i.e., NDLMs), the state estimates and marginal likelihood can be directly calculated using the Kalman filter (Kalman, 1960). The Kalman filter provides an algorithm that, using only elementary linear algebra operations, sequentially updates the filtering mean and variance of the states (Harvey, 1990; Durbin and Koopman, 2012). Using the toy model described in Section 2.1, we can describe the Kalman filter as follows. To initialize the Kalman filter recursions, one needs to either specify the initial state distribution as $z_0 \sim \text{N}(m_0, C_0)$ or consider z_0 to be an additional fixed unknown parameter (equivalent to setting $m_0 = z_0$ and $C_0 = 0$). For all other time steps, $t = 1, \ldots, T$, the Kalman filter sequentially calculates the following distributions. First, the prior distribution of z_t given observations up to time $t-1$, also known as the one-step-ahead predictive distribution of the state, is calculated as:

$$z_{t|1:t-1} \sim \text{N}(a_t, R_t), \tag{43}$$

where $a_t = \beta m_{t-1}$, $R_t = \beta^2 C_{t-1} + \sigma_p^2$, and β is the autocorrelation parameter and σ_p^2 is the variance in the process equation (Eq. 1) and m_{t-1} and C_{t-1} are defined below. Second,

the predictive distribution of y_t given observations up to time $t-1$, also known as the one-step-ahead predictive distribution of the observation, is calculated as:

$$y_{t|1:t-1} \sim \mathrm{N}(f_t, Q_t), \tag{44}$$

where $f_t = \alpha a_t$, $Q_t = \alpha^2 R_t + \sigma_o^2$, and α is the bias parameter and σ_o^2 is the variance in the observation equation (Eq. 2). Finally, the posterior distribution of z_t given observations up to time t, also known as the filtered distribution of the state, can be calculated as:

$$z_{t|1:t} \sim \mathrm{N}(m_t, C_t), \tag{45}$$

where $m_t = a_t + \alpha R_t(y_t - f_t)/Q_t$ and $C_t = R_t - \alpha^2 R_t^2/Q_t$. The mean of each filtered distribution, m_t, provides a good point estimate of the state value at that time and the variance, C_t, quantifies the uncertainty around this estimate. We can view m_t (the mean of the filtered distribution, i.e., $E[z_t|y_{1:t}]$), as a correction of a_t (the mean of the one-step-ahead predictive distribution of the state, i.e., $E[z_t|y_{1:t-1}]$). The correction is based on the one-step-ahead prediction residual $(y_t - f_t)$, namely the difference between the observed value, y_t, and the mean of the one-step-ahead predictive distribution of the observation, f_t (i.e., $E[y_t|y_{1:t-1}]$). The mean and variance of each smoothing distribution (i.e., distribution of $z_{t|1:T}$) can be obtained with a backward recursion in time, using an algorithm analogous to the Kalman filter called the Kalman smoother (Harvey, 1990; Durbin and Koopman, 2012).

The Kalman filter output can also be used to evaluate the marginal likelihood and thus to find the MLE. We can define the marginal likelihood as:

$$L_M(\boldsymbol{\theta}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} p(y_t|\mathbf{y}_{1:t-1}). \tag{46}$$

Since the one-step-ahead predictive distribution of the observation, $p(y_t|\mathbf{y}_{1:t-1})$ is defined as $\mathrm{N}(f_t, Q_t)$ in the Kalman filter (Eq. 44), we can easily calculate the marginal likelihood as:

$$L_M(\boldsymbol{\theta}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} (2\pi Q_t)^{-1/2} \exp\left\{-\frac{1}{2Q_t}(y_t - f_t)^2\right\}. \tag{47}$$

In Appendix S1, we demonstrate how to use the `R` package `dlm` (Petris, 2010) to perform Kalman filtering and smoothing, as well as forecasting. It can also be used to find MLEs of unknown fixed parameters. The package is flexible enough to allow univariate and multivariate NDLMs, accounting for constant or time-varying distributions of states and observations. More details about Kalman filter and smoother and `dlm` can be found in Petris et al. (2009) and Petris (2010). See also Chapter 6 of Shumway and Stoffer (2016) for description of filtering, smoothing, forecasting and for maximum likelihood estimation.

While the Kalman filter is an important algorithm for fitting SSMs to data, it does not work with nonlinear and non-Gaussian SSMs. Approximate techniques based on the Kalman filter are available for linear models whose observations follow an exponential family distribution (e.g., Poisson, see Durbin and Koopman, 2012, Ch. 9). Other approximate filtering and smoothing methods based on the Kalman filter, such as the extended Kalman filter and the unscented Kalman filter (e.g., Durbin and Koopman, 2012, Ch. 10) are useful for some nonlinear and/or non-Gaussian SSMs. However, for more complex, nonlinear, and non-Gaussian models, one must use one of the methods described below.

3.1.2 Laplace approximation

The Laplace approximation is a useful tool for obtaining an approximation of the marginal likelihood that is commonly used to fit SSMs to data (Fournier et al., 2012; Kristensen et al., 2016). The general idea is that if the marginal likelihood (Eq. 41) is a well-behaved unimodal function, it can be approximated with a Normal density function. We can use this approximation to find the MLE.

The Laplace approximation of the marginal likelihood is derived by first approximating the natural logarithm of the joint likelihood ($\log L_{\text{J}}$, Eq. 39), denoted $\ell_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})$. The approximation uses a second order Taylor polynomial with respect to the states evaluated at $\hat{\mathbf{z}}_{1:T}$ (the value of $\mathbf{z}_{1:T}$ that maximizes ℓ_{J}):

$$\begin{aligned}\ell_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T}) \approx \ell_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T}) + \ell'_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T}) \\ + \frac{1}{2}(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T})^{\text{tr}}\ell''_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T}),\end{aligned} \tag{48}$$

where $\ell'_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})$ and $\ell''_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})$ are the first and second derivatives of the joint log likelihood evaluated at points $\hat{\mathbf{z}}_{1:T}$, and $(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T})^{\text{tr}}$ is the transpose of $(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T})$. Because $\hat{\mathbf{z}}_{1:T}$ maximizes ℓ_{J}, the first derivative evaluated at $\hat{\mathbf{z}}_{1:T}$ equals zero and the approximation simplifies to:

$$\ell_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T}) \approx \ell_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T}) + \frac{1}{2}(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T})^{\text{tr}}\ell''_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T}). \tag{49}$$

The marginal likelihood can be approximated by exponentiating the approximation of the log joint likelihood and then integrating over the states. Noting that the joint likelihood evaluated at $\hat{\mathbf{z}}_{1:T}$, $L_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})$, is a constant term with respect to $\mathbf{z}_{1:T}$, we have the following:

$$L_{\text{M}} \approx L_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T}) \int \exp\left(\frac{1}{2}(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T})^{\text{tr}}\ell''_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})(\mathbf{z}_{1:T} - \hat{\mathbf{z}}_{1:T})\right) \text{d}\mathbf{z}_{1:T} \tag{50}$$

The integrand is as a quantity proportional to the density function of a multivariate Gaussian known as the kernel. Thus, the integral will be the normalizing constant (i.e., the constant with which we divide the unnormalized function) of this multivariate Gaussian distribution:

$$L_{\text{M}} \approx L_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T})\sqrt{\frac{(2\pi)^T}{\det(-\ell''_{\text{J}}(\boldsymbol{\theta}, \hat{\mathbf{z}}_{1:T}|\mathbf{y}_{1:T}))}}. \tag{51}$$

This method is quite flexible, and a variety of SSMs can be fitted using the Laplace approximation. However, the method does assume that the states can be locally-approximated with a Gaussian distribution, which means that the states are assumed to have an unimodal distribution. Because the method uses the second derivative of the log likelihood, we cannot use the Laplace approximation with categorical states or other state distributions that are not twice differentiable. One of the main advantages of the Laplace approximation approach over simulation-based approaches described below is the speed at which SSMs are fitted to data (see Auger-Méthé et al., 2017). Many software use the Laplace approximation approach (e.g., Fournier et al., 2012), but one that is particularly useful for SSMs, and is described in detail in Appendix S1, is the `R` package `TMB` (Kristensen et al., 2016).

3.1.3 Iterative filtering

Monte Carlo methods can be used to estimate the states and evaluate the integral needed to obtain the marginal likelihood. Monte Carlo methods are computer intensive sampling procedures that generate random samples from specific probability distributions, which can then be used to evaluate integrals. For example, if X is a continuous random variable with a probability density function $f(x)$, the expected value of X is $E[X] = \int xf(x)dx$. A Monte Carlo solution is to generate a random sample $x_{1:N}$ from $f(x)$ and the estimate $E[X]$ with $\frac{1}{N}\sum_{i=1}^{N} x_i$. In the context of SSMs fitted with a frequentist approach, these Monte Carlo methods sample the state space by generating samples using the process equation and weighting the samples with the observation equation. While in this section we discuss Monte Carlo methods in the context of a frequentist inference approach called iterative filtering, we will see in Section 3.2 that Monte Carlo methods are commonly used for Bayesian inference.

Sequential importance sampling (SIS; Doucet et al., 2001) can be used to numerically approximate one of the important integrations associated with fitting SSMs to data. To estimate the smoothed state at time t, $\hat{z}_{t|1:T}$, one needs to perform the following high-dimensional integration:

$$E[z_t|y_{1:T}, \boldsymbol{\theta}] = \int z_t p(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) \mathrm{d}\mathbf{z}_{1:T}. \tag{52}$$

SIS is a procedure that uses an importance function to sequentially generate N samples of the state for each time step t, z_t^i, i=1,..., N. In the context of SSMs, the importance function is often the probability density function of the state at time t: $f(z_t|z_{t-1}, \boldsymbol{\theta}_p)$ (Eq. 3). Each sequence or sample trajectory $z_{1:t}^i$ is referred to as a particle. SIS also keeps track of the importance weights, w_t^i, of each sample i at time t in order to approximate an overall importance value for each sample trajectory, $w_T^{*,i}$. In the context of SSMs, the importance weights at time t are functions of the probability density function of the observations, $g(y_t|z_t^i, \boldsymbol{\theta}_o)$ (Eq. 4), and are updated sequentially as follows:

$$w_t^i \propto w_{t-1}^i g(y_t|z_t^i, \boldsymbol{\theta}_o). \tag{53}$$

For example, for our toy model (Eqs. 1-2), the i^{th} sample of the state at time t, z_t^i, would be sampled from N($\beta z_{t-1}^i, \sigma_p^2$) and the weights are calculated as:

$$w_t^i \propto w_{t-1}^i \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left(-\frac{1}{2\sigma_o^2}(y_t - \alpha z_t^i)^2\right). \tag{54}$$

A smoothed estimate of the state at time t is the weighted average of the N simulated states for that time point:

$$\hat{z}_{t|1:T} = \sum_{i=1}^{N} w_T^{*,i} z_t^i, \tag{55}$$

where $w_T^{*,i} = \frac{w_T^i}{\sum_{j=1}^{N} w_T^j}$. The final normalized importance weight, $w_T^{*,i}$, represents how likely the i^{th} particle is compared to the other particles. However, while relatively simple to implement

(see Appendix S1 for example), SIS is impractical for even moderately long time series (e.g., T=20). The problem is that the weights of particles become increasingly uneven, with only a small proportion of the particles typically having the majority of the weight. This leads to state estimates with unacceptably large variances. The gradual erosion of support for many of the particles, known as particle depletion, is a serious problem with SIS and other particle filters.

The bootstrap filter (Gordon et al., 1993) is a procedure designed to remedy particle depletion. The bootstrap filter assesses the weight of a particle through time and iteratively removes particles with low weights and replaces them with duplicates of particles with higher weights. There are various algorithms for the bootstrap filter. The one described below, and implemented in Appendix S1, is the simplest and involves resampling the particles after each iteration. At iteration t:

1. Generate a state value at time t for each particle i, $\tilde{z}_t^i$, with the probability density function of the state (i.e., with the process equation), $f(z_t|z_{t-1}^i, \boldsymbol{\theta}_p)$.
2. Set $\tilde{z}_{1:t}^i = (z_{1:t-1}^i, \tilde{z}_t^i)$.
3. Calculate the normalized importance weights of each proposed state $\tilde{z}_t^i$, $w_t^{*,i}$, using the observation equation, $g(y_t|z_t, \boldsymbol{\theta}_o)$.
4. Randomly sample N times (with replacement) from the N particle vectors with probabilities, $w_t^{*,i}$, and denote the selected vectors $z_{1:t}^i$, i=1,...,N. Set the unnormalized weights w_t^i equal to 1.

Because the simulated particles are resampled after each iteration, the simple average of the simulated states based on the selected particles at time t is the filtered state estimate, $\hat{z}_{t|1:t}$, while the simple average based on the final set of particles at time T is the smoothed state estimate, $\hat{z}_{t|1:T}$. While the bootstrap filter can reduce particle depletion, it does not completely solve the problem particularly for long time series. There are various remedies aimed at reducing particle depletion including more sophisticated importance sampling distributions that include information from the observations (Pitt and Shephard, 1999) or changing the resampling methods (Liu and Chen, 1998). SIS and bootstrap filtering is explained in further detail in Appendix S2 and implemented in `R` in Appendix S1.

While one can use SIS or the bootstrap filter (or any of the broader class of related algorithms known as 'particle filters') to estimate the likelihood, in order to perform frequentist inference, one must be able to maximize the likelihood. This presents a number of difficulties. First, the likelihood estimates given by the particle filter algorithms are noisy due to their Monte Carlo nature. As the variance in these estimates grows, the utility of the estimates is diminished. Second, these algorithms can be computationally expensive, since each likelihood calculation requires simulation of N realizations of the latent state process, where N must be large enough to yield precise likelihood estimates. Finally, to maximize the likelihood, one must explore what is in many cases a complicated likelihood surface, marked by features such as multiple local maxima, narrow, curved ridges, and precipitous cliffs. Exploring this surface in search of its highest peaks will necessarily involve numerous likelihood evaluations. Thus, while it is in principle possible to use a general-purpose optimization algorithm such as Nelder-Mead to maximize the likelihood computed by a particle

filter, in practice this is typically prohibitively expensive. Iterated filtering is an attractive alternative.

As its name implies, iterated filtering (IF) involves repeatedly applying a particle filter as a means of maximizing the likelihood. Repeated application of a particle filter at a single parameter value merely results in multiple estimates of the likelihood. To explore the likelihood surface, IF effectively modifies the model by turning parameters into state variables. In particular, the fixed parameters of the model of interest are perturbed, at each observation time, by a random amount. This random perturbations forestall particle depletion by continually re-injecting random variability into the filter. This reduces the Monte Carlo noise of the likelihood estimates. Thus, paradoxically, it is frequently possible when using IF to maximize the likelihood using a much smaller N than is needed to precisely estimate the likelihood itself. Finally, the random perturbations of the parameters act to harvest information from a local neighborhood; IF exploits this information to increase the likelihood at each iteration, on average. However, despite all of these performance enhancements, because it applies artificial perturbations to parameters, IF is learning not about the model of interest (i.e., model with fixed parameters), but only about a modified model (i.e., model where fixed parameters have been transformed into state variables). Therefore, as filtering iterations proceed, one gradually cools (i.e., reduces the magnitude of) the artificial perturbations, so that the modified model approaches the model of interest as the iterations proceed. In theory, if the cooling is sufficiently slow, and the number of iterations sufficiently large, the algorithm will converge to some local likelihood maximum. In practice, as with related algorithms such as simulated annealing or Markov Chain Monte Carlo, the number of iterations required by the theory can be prohibitively large. Moreover, because statistical inference hinges on identification of the global likelihood maximum and exploration of its neighborhood, a single IF computation is never sufficient. For these reasons, it is usually advisable to perform many independent, relatively short, IF computations, starting from widely dispersed starting points. In such a global search strategy, one relies on the ability of each independent IF computation to rapidly ascend the surface in the vicinity of its starting point. Because this strategy is easily parallelizable, one can efficiently exploit modern high-performance computing technology to search for the global maximum. IF can be easily implemented using the `R` package `pomp` (see Appendix S1, King et al., 2009).

3.1.4 Other methods

The methods described above represent, in our view, the most commonly used methods to fit SSMs to ecological data in a frequentist framework. These methods are associated with comprehensive `R` packages that facilitate their implementation. However, many other methods exist. Of note, Kitagawa (1987) provided a general iterative algorithm for non-Gaussian SSMs similar to the Kalman filter, but that uses piecewise linear functions (first-order splines) to approximate the non-normal one-step-ahead predictive distributions and the filtered distributions of the states. In essence, this method approximates the needed distributions by discretizing them. It can be viewed as discretizing the continuous state space and reformulating the model as a HMM (Pedersen et al., 2011). de Valpine and Hastings (2002) demonstrated how flexible this approach was to fit nonlinear non-Gaussian population dynamics models. This approach appears particularly promising for population

modeling, where the states are counts, and thus the state space is already discretized (Besbeas and Morgan, 2019). Pedersen et al. (2011) demonstrated that while this method is general and can provide results similar to the Laplace approximation and Bayesian methods, it is computationally limited to problems with only a few state dimensions.

3.2 Bayesian framework

When we fit a SSM with a Bayesian approach, the function of interest (also known as the target distribution) is the posterior distribution for the states and parameters given the observations:

$$p(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\lambda}) = \frac{L_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int\int L_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\mathrm{d}\mathbf{z}_{1:T}\mathrm{d}\boldsymbol{\theta}}, \tag{56}$$

where $L_{\text{J}}(\boldsymbol{\theta}, \mathbf{z}_{1:T}|\mathbf{y}_{1:T})$ is the joint likelihood (i.e., $p(\mathbf{y}_{1:T}|\boldsymbol{\theta}, \mathbf{z}_{1:T})$, see for example Eq. 39), and $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is the prior distribution(s) for the parameters with fixed hyperparameters, $\boldsymbol{\lambda}$. Eq. 56 is an application of Bayes' theorem ($P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta)}{P(\mathbf{y})}$) and the denominator of Eq. 56 represents the probability of the data (i.e., the marginal likelihood, which is the probability of the data for all possible values of the states and parameters). In Bayesian analyses, both the states, $\mathbf{z}_{1:T}$, and what we have been referring to as fixed parameters, $\boldsymbol{\theta}$, are considered random variables. The posterior distribution is a complete characterization of these random variables given the data and prior information. As such, the first inferential goal of a Bayesian analysis is often to evaluate the posterior distribution. While point estimates for the parameters and the states are not necessarily the primary goal of a Bayesian analysis, they can be obtained by summarizing the center of the posterior distribution (e.g. mean or mode of the posterior distribution). Similarly we can use the posterior distribution to obtain interval estimates and single measures of variation.

As for the frequentist framework, the fitting procedures are complicated by high-dimensional integrals and it is common to avoid calculating the integral and the posterior distribution explicitly. Instead, quantities of interest are generally approximated using Monte Carlo methods (see also Section 3.1.3), where large samples of states and parameters are randomly drawn from the posterior distribution. For example, one can approximate the point estimate of a parameter with the sample mean of the draws from the posterior distribution (often referred as the posterior mean). Simulating independent draws from Eq. 56 is typically impossible. However, there are various algorithms that can approximate the posterior distribution with large samples of dependent draws. In particular, Markov Chain Monte Carlo (MCMC) methods are a broad class of algorithms that obtain samples from the target distribution (here the posterior distribution Eq. 56), by sampling from a Markov chain rather than sampling from the target itself. By definition a Markov chain is a sequence of random variables, where at each time step of the chain the value of the random variable only depends on the values from previous time step. To be able to use the samples from a Markov chain as an approximation of the posterior distribution, the Markov chain needs the additional characteristic of having an invariant distribution (also known as a stationary or equilibrium distribution) equal to the target distribution (Geyer, 2011). An invariant distribution refers to when the probability distribution remains unchanged as samples are

drawn, a quality which is dependent on the initial condition of the chain and the transition probabilities, and related to the importance of chain convergence as a diagnostic in MCMC sampling. MCMC algorithms fall into two broad families: Metropolis-Hastings samplers (which include Gibbs samplers) and Hamiltonian Monte Carlo.

3.2.1 Metropolis-Hastings samplers

Metropolis-Hastings samplers are at the base of most MCMC algorithms used to sample the posterior distribution in a Bayesian analysis. Metropolis-Hastings samplers are iterative algorithms that create an appropriate Markov chain to sample the target distribution. From demonstrative purposes, let say we have a very simple model with one parameter, θ, no latent states, and where the observations $y_i \overset{\text{i.i.d.}}{\sim} f(y|\theta)$. We denoted the likelihood function with $L(\theta|\mathbf{y}_{1:N})$ and the prior with $\pi(\theta)$. An Metropolis-Hastings algorithm samples the values of the parameter at step j of the chain, θ^j, with the following iterative rules. For time step j of the chain, first generate a candidate value, $\tilde{\theta}^j$, with a proposal distribution, $q(\tilde{\theta}^j|\theta^{j-1})$. For example, the proposal distribution for the candidate value could be a Gaussian distribution centered on the previous value of the chain (i.e., $\tilde{\theta}^j \sim \mathrm{N}(\theta^{j-1}, \sigma^2_{\mathrm{MH}})$). Second, accept the candidate value with probability $\alpha(\theta^{j-1}, \tilde{\theta}^j) = \min(1, p_\theta)$, where

$$p_\theta = \frac{L(\tilde{\theta}^j|\mathbf{y}_{1:N})\pi(\tilde{\theta}^j)}{L(\theta^{j-1}|\mathbf{y}_{1:N})\pi(\theta^{j-1})} \frac{q(\theta^{j-1}|\tilde{\theta}^j)}{q(\tilde{\theta}^j|\theta^{j-1})}. \tag{57}$$

If the proposed value is accepted $\theta^j = \tilde{\theta}^j$; if not, the previous value of the chain is used: $\theta^j = \theta^{j-1}$. The first value of the chain, θ^0, will affect the chain and the convergence towards the target distribution (see Section 3.2.3).

In the context of SSMs, we have a multivariate posterior distribution for the states and the parameters. Using Metropolis-Hastings algorithms to sample for more than one random variables is more complex, but there are various implementation tools to do so. For example, for each iteration j of the chain, one can first sample sequentially all parameter values, and then sequentially sample the state values (Newman et al., 2014). Let us denote the parameters of our toy example as $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_o, z_0) = (\beta, \sigma_p, \alpha, \sigma_o, z_0)$ and the elements of $\boldsymbol{\theta}$ with $r = 1, ..., 5$. For iteration j and $r > 1$, we will have already updated some of the parameters. For example, when updating $r = 2$ we will have $\boldsymbol{\theta}^{j-1}_{[2]} = (\beta^j, \sigma_p^{j-1}, \alpha^{j-1}, \sigma_o^{j-1}, z_0^{j-1})$ as our current parameters and $\tilde{\boldsymbol{\theta}}^j_{[2]} = (\beta^j, \tilde{\sigma}_p^j, \alpha^{j-1}, \sigma_o^{j-1}, z_0^{j-1})$ as our candidate value. For each parameter r, we would accept the candidate value with probability $\alpha_r = \min(1, p_{\theta_r})$, where

$$p_{\theta_r} = \frac{\prod_{t=1}^T g(y_t|z_t, \tilde{\boldsymbol{\theta}}^j_{[r]}) f(z_t|z_{t-1}, \tilde{\boldsymbol{\theta}}^j_{[r]}) \pi(\tilde{\boldsymbol{\theta}}^j_{[r]}|\boldsymbol{\lambda})}{\prod_{t=1}^T g(y_t|z_t, \boldsymbol{\theta}^{j-1}_{[r]}) f(z_t|z_{t-1}, \boldsymbol{\theta}^{j-1}_{[r]}) \pi(\boldsymbol{\theta}^{j-1}_{[r]}|\boldsymbol{\lambda})} \frac{q(\boldsymbol{\theta}^{j-1}_{[r]}|\tilde{\boldsymbol{\theta}}^j_{[r]})}{q(\tilde{\boldsymbol{\theta}}^j_{[r]}|\boldsymbol{\theta}^{j-1}_{[r]})}, \tag{58}$$

Once all r parameters have been sampled, we would sequentially sample the states for iteration j. We would sequentially accept a candidate state value with probability $\alpha_t = \min(1, p_{z_t})$, where

$$p_{z_t} = \frac{g(y_t|\tilde{z}^j_t, \boldsymbol{\theta}^j) f(\tilde{z}^j_t|z^j_{t-1}, \boldsymbol{\theta}^j) f(z^{j-1}_{t+1}|\tilde{z}^j_t, \boldsymbol{\theta}^j)}{g(y_t|z^{j-1}_t, \boldsymbol{\theta}^j) f(z^{j-1}_t|z^j_{t-1}, \boldsymbol{\theta}^j) f(z^{j-1}_t|z^{j-1}_{t-1}, \boldsymbol{\theta}^j)} \frac{q(z^{j-1}_t|\tilde{z}^j_t)}{q(\tilde{z}^j_t|z^{j-1}_t)}. \tag{59}$$

Note that all other terms of the posterior cancel out, since they use the same previously sampled values (e.g., $\pi(\boldsymbol{\theta}^j)$ and $f(y_{t-1}|z_{t-1}^j)$). Note that if groups of variables are related they can be sampled simultaneously from a multivariate distribution rather than sequentially. In practice, states and parameters are often correlated, and thus it may be difficult to implement an efficient MCMC sampler that doe not require very long simulations before convergence (Newman et al., 2014).

Gibbs samplers are commonly-used Metropolis-Hastings samplers for multivariate distributions, where the proposal distributions are conditional distributions of the target distribution and thus the candidate values are always accepted (Geyer, 2011). Deriving these conditional distributions is facilitated by the fact that they are proportional to the target distribution and that functions that do not include the variable of interest will be cancelled out. For example, the full conditional distribution of the state z_t for a general SSM is:

$$q(z_t|z_{t-1}, z_{t+1}, \boldsymbol{\theta}^j, y_t) \propto g(y_t|z_t, \boldsymbol{\theta}^j) f(z_t|z_{t-1}, \boldsymbol{\theta}^j) f(z_{t+1}|z_t, \boldsymbol{\theta}^j), \tag{60}$$

which is a much simpler expression than Eq. 56 and have some of the same components as Eq. 59. When the conditional distribution is available in closed form, it is easy to sample from it exactly. For the toy model, the full conditional distribution of z_t is Normal, with mean

$$\left(\frac{\alpha^2}{\sigma_o^2} + \frac{1+\beta^2}{\sigma_p^2}\right)^{-1} \left(\frac{\alpha y_t}{\sigma_o^2} + \frac{\beta(z_{t-1}+z_{t+1})}{\sigma_p^2}\right), \tag{61}$$

and variance

$$\left(\frac{\alpha^2}{\sigma_o^2} + \frac{1+\beta^2}{\sigma_p^2}\right)^{-1}. \tag{62}$$

In this case, simulating from this normal distribution is straightforward and it allows us to forego the acceptance step of other Metropolis-Hastings algorithms. For NDLMs, the entire sequence $\mathbf{z}_{0:T}$ can be simulated at once from its conditional distribution, given the data $\mathbf{y}_{1:T}$ and the time-invariant parameter $\boldsymbol{\theta}$, using the Forward Filtering Backward Sampling (FFBS) algorithm described in Carter and Kohn (1994). Note that the FFBS algorithm can also be used to conduct inference for the SSMs that are conditionally linear and Gaussian. However, Gibbs samplers for nonlinear and non Gaussian models often require sampling from each conditional distribution sequentially (see chapter 4.5 of Prado and West, 2010, for an overview). A drawback of this particular Gibbs sampler design is that consecutive draws of $z_{0:T}^j$ and $z_{0:T}^{j-1}$ tend to be highly correlated, slowing the convergence of sample averages to their theoretical expected values, thus deteriorating the quality of the Monte Carlo approximations.

3.2.2 Hamiltonian Monte Carlo

An efficient alternative to Metropolis-Hastings sampling is provided by Hamiltonian Monte Carlo (HMC) methods, which have gained popularity in recent years thanks in part to their implementation in the Stan software (Stan Development Team, 2012). These methods are

inspired by analogies drawn from physics and rely heavily on deep differential geometric concepts, which are beyond the scope of this review. HMC can be a more efficient sampler than Metropolis-Hastings as fewer iterations are typically required and fewer rejections occur. This is achieved by the addition of a momentum variable with the target distribution that helps the Markov chain to remain within the typical set of the target distribution, rather than conducting random walk to explore the target distribution as is frequently done in Metropolis-Hastings. The interested reader can explore the details in Neal (2011) or Betancourt (2017). Conducting inference for general SSMs via HMC is possible when all parameters and states are continuous or when the posterior distribution can be marginalized over any discrete parameters or states. Continuous distributions are required because density gradients of the target distribution are required to direct the sampling through the typical set of the target distribution (Betancourt, 2017). Unlike Metropolis-Hastings samplers, HMC methods draw samples from the joint posterior distribution directly and can scale well to high dimensional spaces. General SSMs can be fitted either by defining Eq. 56 or by marginalization over the state process to derive the posterior distribution of the time-invariant parameters only, $p(\boldsymbol{\theta}|\mathbf{y}_{1:T}, \boldsymbol{\lambda})$.

3.2.3 Convergence diagnostics

Regardless of the sampling method, it is important to assess whether it has reached the target distribution. Convergence between multiple chains usually indicates that they have reached the invariant distribution. As such, multiple approaches have been developed to assess whether chain convergence has been achieved. In general, samples from the first iterations are discarded, as these likely occurred before the chain has reached the target distribution (Gelman and Shirley, 2011, but see Geyer 2011). In the Metropolis-Hastings setting, this period is referred to as 'burn-in'. A somewhat similar initial period, referred as the 'warm-up', is discarded with HMC. Then, as a first step, convergence within and between chains can be assessed visually via traceplots (see Appendix S1). More formal metrics exist. The Gelman-Rubin metric, $\hat{R}$ (Gelman and Rubin, 1992, see Brooks and Gelman 1998 for the multivariate analogue), is one of the most popular multi-chain diagnostics. Although $\hat{R} < 1.1$ generally indicates convergence (Gelman et al., 2013), recent research has indicated that a threshold closer to one may be more suitable in some scenarios (Vats and Knudson, 2018). The $\hat{R}$ diagnostic has been further developed by Gelman et al. (2013) to better diagnose split chains. Note that pseudo-convergence can occur in many different scenarios. For example, if the target distribution has multiple modes that are not well connected by the Markov chain dynamics, the sampler can get caught in one mode (Geyer, 2011). Running the chain for a long period can help limit these pseudo-convergence problems (Geyer, 2011). A detailed summary of convergence methods is available in Cowles and Carlin (1996) and further research on convergence diagnostics includes Boone et al. (2014), VanDerwerken and Schmidler (2017), and Vats and Knudson (2018).

3.2.4 Priors

Selection of priors is a significant part of a Bayesian analysis because priors affect the resulting posterior distribution (Robert, 2007). Several approaches can be taken depending

on the information available about the model parameters and the philosophy of the modeler. Ecologists often use 'noninformative' priors or vaguely informative priors. These priors (e.g., a uniform distribution over the parameter space) are often thought to be objective and are generally chosen with the goal of maximizing the influence of the data on the posterior. However, noninformative priors may still have important effects on the posterior, and they should not be used naively (Gelman et al., 2017). For example, Lele (2015) showed that noninformative priors could significantly influence the parameter and state estimates of ecological SSMs. Alternatively, ecologists can use informative priors, which are created using knowledge of the parameter or previously collected data (e.g., Meyer and Millar, 1999; Dunham and Grand, 2016). As there are many advantages to using informative priors, they are increasingly used in ecological models (Hooten and Hobbs, 2015). For example, informative priors can be used to supplement SSMs with limited time-series data (Chaloupka and Balazs, 2007) and can improve state estimates (Dunham and Grand, 2016). In most cases, noninformative and informative priors are used in the same model on different parameters. For technical reasons, it can be sometime advantageous to use conjugate priors (i.e., priors with the same distribution as the conditional posterior distribution or the posterior distribution). Kass and Wasserman (1996) and Millar (2002) have summarized priors typically used in fisheries models, including many SSMs. More generally, Robert (2007) and Gelman et al. (2013) provide a thorough review of available priors, selection and examples for a variety of models.

3.2.5 Software for Bayesian inference

There are many packages available to fit Bayesian models in `R`. Some packages, such as `dlm`, allow to fit specific types of SSMs, but those listed below are general Bayesian software and `R` packages that can be easily used for ecological SSMs. Generating draws from the posterior distributions using a Gibbs sampler can done using software from the BUGS (Bayesian inference Using Gibbs Sampling) project and their associated `R` packages: `WinBUGS` can be called in `R` via `R2WinBUGS` (Lunn et al., 2000), `OpenBUGS` via `BRugs` (Lunn et al., 2009), while `MultiBUGS` `R` interface is in development (Goudie et al., 2017). Gimenez et al. (2009a) provide a tutorial on how to fit ecological models (including some of the SSMs of Section 2.2) with `WinBUGS` in `R`. `JAGS` (Just Another Gibbs Sampler; Plummer, 2003) is an alternative to BUGS project software that is written for UNIX, thus preferred by Mac and Linux users. `JAGS` is available through the `R` package `rjags` (Plummer, 2018). The `R` package `NIMBLE` (de Valpine et al., 2017) is a recent alternative to `JAGS` and BUGS software that is more transparent in how the sampling is performed. `NIMBLE` allows users to write custom Gibbs samplers that perform block updating or implement a variety of other techniques (e.g., particle filtering). Note that both `JAGS` and BUGS project software, as well as the `R` package `coda` (Plummer et al., 2006), provide several methods to assess convergence. The only software program that uses Hamiltonian Monte Carlo is `Stan`, available in `R` through the package `rstan` (Stan Development Team, 2018). All these software, allow one to write general models in a language based on BUGS (Bayesian analysis Using Gibbs Sampling, see Lunn et al., 2013). The user can set up the sampler in `R`, and once compiled, can use it to simulate draws to make inference about states and unknown parameters. See Appendix S1 for detailed examples in `JAGS`, `NIMBLE`, and `Stan`.

3.3 Information reduction approaches

While uncommonly used with SSMs, information reduction approaches, such as synthetic likelihood or Approximate Bayesian Computation (ABC), appear promising to fit complex, highly nonlinear, SSMs (Fasiolo et al., 2016). These methods bypass the calculation of the exact likelihood (Csilléry et al., 2010; Fasiolo et al., 2016). Instead, these methods generate samples from the model and transform them into a vector of summary statistics that describe the data in the simplest manner possible (Csilléry et al., 2010; Fasiolo et al., 2016). The simulated summary statistics are then compared to observed summary statistics using a predefined distance measure (Fasiolo et al., 2016). Information reduction approaches require little tuning and smooth the likelihood, reducing some of the common implementation problems encountered with other fitting methods. However, the results from information reduction approaches are often imprecise and, thus, may be most useful in the model development phase (Fasiolo et al., 2016; Fasiolo and Wood, 2018). Interested readers are refered to Csilléry et al. (2010), Fasiolo et al. (2016), and Fasiolo and Wood (2018).

3.4 When to use each method?

Choosing from this multitude of fitting methods can appear daunting, but can be guided by a choice of inference framework and the limitations of each methodology. In Table 1, we list the methods discussed above, with some pros and cons. We simply state the associated inferential framework (frequentist vs. Bayesian), and we let the readers decide their favorite inferential framework.

Note that in some cases, it may also be easier to use one of the more specific ecological SSM model `R` packages. For example, the package `MARSS` (which stands for Multivariate Auto-Regressive State-Space, Holmes et al., 2012, 2018) can be useful to model multiple populations, if these can be reasonably formulated with a linear and normal SSM. Those interested in fisheries stock assessment SSMs should look at the package `stockassessment` (available on GitHub at https://github.com/fishfollower/SAM, Nielsen and Berg, 2014). Those interested in SSMs for animal movement should explore `bsam` (Jonsen et al., 2005; Jonsen, 2016) and `crawl` (Johnson et al., 2008; Johnson and London, 2018).

4 Formulating an appropriate SSM for your data

SSMs are powerful tools, but their inherent flexibility can tempt ecologists to formulate models that are far too complex for the available data. The model structure or the characteristics of the specific dataset may make it impossible to estimate every parameter reliably. In such cases, parameter estimates will no longer provide key information on the underlying biological process and state estimates may become unreliable (e.g., Auger-Méthé et al., 2016). Formulation of SSMs needs to be guided by the inference objectives and the available data. In this section, we discuss how to assess whether a model is adequate for your data and how one can alleviate potential estimation problems.

Table 1: Comparison of the fitting methods discussed in Section 3.

Method	Framework	Pros	Cons	R package
Kalman filter & MLE	Frequentist	Efficient & exact	Only applicable to linear Gaussian SSMs	`dlm, MARSS`
Laplace approximation	Frequentist	Efficient & flexible	States need to be approximatable with a continuous unimodal distribution (e.g., no discrete states)	`TMB`
Particle filter & iterative filtering	Frequentist	Flexible	Can be slow and sensitive to starting values	`pomp`
MCMC-MH	Bayesian	Flexible	Can be slow and sensitive to convergence problems	`rjags`, `NIMBLE`, `R2WinBUGS`, `BRugs`
MCMC-HMC	Bayesian	Efficient & flexible	Require continuous parameters and states or marginalization	`rstan`
Information reduction	Bayesian	Flexible & fewer estimation problems	Can be slow and imprecise	`EasyABC`

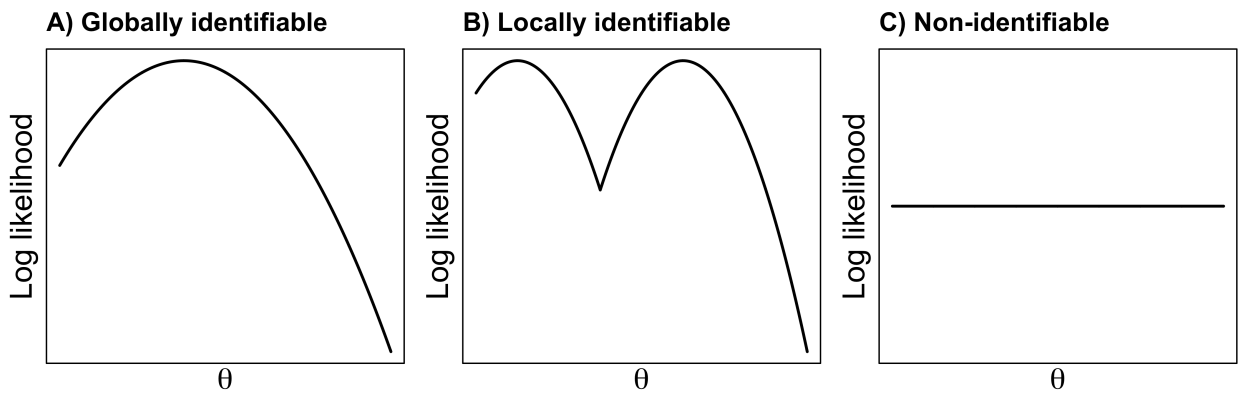

Figure 3: Examples of log-likelihood profiles for a parameter θ under various identifiability scenarios: A) globally identifiable, B) locally identifiable, and C) non-identifiable model.

4.1 Identifiability, parameter redundancy and estimability

When we estimate the parameters of a model, denoted here as $M(\boldsymbol{\theta})$, we want to find the set of parameter values, $\boldsymbol{\theta}$, that results in the best fit to the data. For this to be possible, the model needs to be identifiable. Identifiability refers to whether or not there is a unique representation of the model. A model is globally identifiable if $M(\boldsymbol{\theta}_1) = M(\boldsymbol{\theta}_2)$ implies that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. For example, in a frequentist framework, an identifiable model would have only a single $\boldsymbol{\theta}$ value that would maximise the likelihood (Fig. 3A). A model is locally identifiable if there exists a neighbourhood of $\boldsymbol{\theta}$ where this is true (Fig. 3B). Otherwise a model is non-identifiable (Fig. 3C; Rothenberg, 1971; Cole et al., 2010).

An obvious case of non-identifiability is if a model is overparameterised and can be reparameterised with a smaller set of parameters. For example, if two parameters only appear as a product in a model (e.g., $y = \alpha\beta x$); that model could be reparameterised with a single parameter replacing that product (e.g., $y = \gamma x$, where $\gamma = \alpha\beta$). The parameter redundancy of the original model will result in non-identifiability (Catchpole and Morgan, 1997) and we refer to non-identifiability caused by the inherent structure of a model as intrinsic parameter redundancy (Gimenez et al., 2004) or structural non-identifiability (Cobelli and DiStefano III, 1980). Regardless of the amount or quality of data, it is impossible to estimate all the parameters in such a model.

Having a structurally identifiable model does not guarantee that one can estimate its parameters with the data at hand. Non-identifiability can be caused by a specific dataset with, for example, missing or sparse data (Gimenez et al., 2004). This problem is known as extrinsic parameter redundancy (Gimenez et al., 2004) or practical non-identifiability (Raue et al., 2009). A parameter is defined as practically non-identifiable if it has a confidence interval that is infinite (Raue et al., 2009). It is also possible for a dataset to create estimation problems with an otherwise structurally and practically identifiable model, a phenomenon we refer to as statistical inestimability (Campbell and Lele, 2014). If a model is statistical inestimable, a confidence interval for a parameter will be extremely large but not infinite. This often occurs because the model is very similar to a submodel that is parameter redun-

dant for a particular dataset, which is known as near redundancy (Catchpole et al., 2001; Cole et al., 2010).

Having a non-identifiable model (either structurally or practically) leads to several problems. First, there will be a flat ridge in the likelihood of a parameter redundant model (Catchpole and Morgan, 1997), resulting in more than one set of MLEs. However, despite the parameter redundancy, numerical methods for parameter estimation usually converge to a single set of MLEs. Therefore, without further diagnostics, one may not realise that the MLEs are not unique. Second, the Fisher information matrix will be singular (Rothenberg, 1971) and therefore the standard errors will be undefined in a non-identifiable model. However, the exact Fisher information matrix is rarely known and standard errors are typically approximated using a Hessian matrix obtained numerically. The numeric error can lead to explicit (but incorrect) estimates of standard errors. Third, most model selection methods are based on the assumption that a model is identifiable. For example, when calculating AIC, the number of estimated parameters in a non-identifiable model should be the number of unique parameters that can be estimated, rather than the number of parameters in the model (Gimenez et al., 2004). If a model is statistically inestimable or near redundant, similar issues will occur, as the model is close to being non-identifiable. For example the log-likelihood profile will be almost flat.

Checking for identifiability and estimability should become part of the model fitting process and several methods are available to do so. A clear sign of problems is a flat log-likelihood profile (Fig. 3C), and plotting the log-likelihood profile for each parameter can serve as a diagnostic for this (Fig. 3; Raue et al., 2009; Auger-Méthé et al., 2016). Correlation between parameters can also be indicative of estimation problems, and it may be useful to inspect the log-likelihood or posterior surface of pairs of parameters (Campbell and Lele, 2014; Auger-Méthé et al., 2016). Depending on model complexity and computation time, simulations can be one of the easiest ways to investigate the estimability of SSMs (Auger-Méthé et al., 2016). For a specified SSM and a known set of parameters, one simulates the state process and observation time series, and then estimates the parameters and states. One then compares estimated parameter and state values with the known true values. Parameter estimates from non-identifiable models will usually be biased with large variances.

In addition to these simple checks, three advanced methods to assess estimability and identifiability problems exist. First, data cloning has been shown to be useful with ecological models (Peacock et al., 2016). Data cloning involves using Bayesian methodology with a likelihood based on K copies of the data (clones). The posterior variance of a parameter will tend towards K times the asymptotic variance of the parameter, so that if a parameter is identifiable the posterior variance will tend to zero as K tends to infinity. If a parameter is not identifiable, the posterior variance will tend to a fixed (non-zero) value (Lele et al., 2010). Campbell and Lele (2014) show how this method can be extended to find estimable parameter combinations in non-identifiable models.

Second, one can use the fact that the Hessian matrix in a non-identifiable model will be singular at the MLE. As a singular matrix has at least one zero eigenvalue, the Hessian method involves finding the eigenvalues of the Hessian matrix. If the Hessian matrix is found numerically the eigenvalues for a singular matrix may be close to zero rather than exactly zero. Therefore, if any of the eigenvalues are zero or close to zero, the model is deemed non-identifiable or parameter redundant, at least with this particular dataset (Viallefont et al.,

1998). The Hessian matrix will also have eigenvalues close to zero if the model is statistically inestimable or near redundant (Catchpole et al., 2001).

Third, one can use the symbolic method. A model can be represented by an exhaustive summary, which is a vector of parameter combinations that uniquely define the model. To investigate non-identifiability a derivative matrix is formed by differentiating each term of the vector with respect to each parameter. Then, the rank of the derivative matrix is found. If the rank is less than the number of parameters, then the model is non-identifiable or parameter redundant (Catchpole and Morgan, 1997; Cole et al., 2010). The rank is also the number of estimable parameters (or parameter combinations). For SSMs, suitable exhaustive summaries are given in Cole and McCrea (2016). This method can be used to investigate practical identifiability as well as structural identifiability by choosing an exhaustive summary that includes the specific dataset (Cole et al., 2012). In some more complex models, the computer can run out of memory calculating the rank of the derivative matrix. Cole et al. (2010) and Cole and McCrea (2016) provide symbolic algebra methods for overcoming this issue. The alternative is a hybrid symbolic-numerical method, which involves finding the derivative matrix using symbolic algebra, but then finding the rank at five random points in the parameter space (Choquet and Cole, 2012).

Each of the numerical methods (log-likelihood profile, simulation, data cloning, Hessian method) can be inaccurate. Also, they are also not able to distinguish between estimability, practical identifiability and structural identifiability when applied to a specific dataset, although in some cases a large simulated dataset could be used to test structural identifiability. Being able to distinguish between these problems is useful as it can help us assess whether gathering more data will help. The symbolic method is accurate, but is more complicated to use as it involves using a symbolic algebra package. Code for assessing estimability using simulations and the Hessian method is given in Appendix S1. Code for the symbolic algebra method is given in Appendix S3.

In Bayesian analysis, identifiability and estimability issues have a different focus because priors can affect our capacity to differentiate between parameters (Cressie et al., 2009). In general, parameters are said to be weakly indentifiable when the posterior distribution significantly overlaps with the prior (Garrett and Zeger, 2000; Gimenez et al., 2009b). If priors are well informed by previous data or expert knowledge, their strong influence on the posterior distribution is no longer an identifiability/estimability issue but one of the benefits of Bayesian analysis. However, misusing informed priors (e.g., when the information is not reliable) may hide identifiability issue or cause the estimability problems (Yin et al., 2019). Thus, one should choose priors with great care. To help ensure that the data inform the model and that the posterior is well behaved, Gelfand and Sahu (1999) suggested to use informative priors that are not too precise. Weak identifiability can result in multiple implementation issues, including slow convergence (Gimenez et al., 2009b). Diagnostics for parameter identifiability in the Bayesian framework include some of the tools described above and the visual or numerical assessment of the overlap between priors and posterior distributions (Garrett and Zeger, 2000; Gimenez et al., 2009b).

4.2 Remedies for identifiability issues

When we fit a SSM to our data, we hope that it will provide accurate and precise estimates of our parameters and states. But how can we achieve these goals? First, we need to have a structurally identifiable model. Second, one needs a dataset appropriate for the model and vice-versa, otherwise one can face estimation problems even with structurally identifiable models. Generally, we assume that having more data will allow us to better estimates parameters and states. However, as discussed below, increasing the length of the time series may not be the best way to increase the dataset.

4.2.1 Reformulating the SSM

To create a structurally identifiable model, one should start by avoiding overparametrization. As mentioned above, models where some parameters only appear as products of each other should be simplified. The same holds for models where parameters only appear as sums (e.g., $y = (\alpha + \beta)x$), or differences, or fractions. Models where the magnitude of two sources of error are simply additive are also problematic (e.g., $Y \sim N(X, \sigma^2)$ and $X \sim N(\mu, \tau^2)$, will result in $Y \sim N(\mu, \sigma^2 + \tau^2)$ where σ and τ cannot be uniquely identified). As such, one needs to check that none of the parameters are confounded and carefully inspect the combination of the sources of variability in all hierarchical models, including SSMs (see below). Some of the tools discussed above can help construct structurally identifiable models. In particular, the symbolic method can be used to identify the parameters that are confounded in a non-identifiable model, and thus can be used to select estimable parameter combinations. This involves solving a set of partial differential equations formed from the same derivative matrix used to check identifiability (Catchpole et al., 1998; Cole et al., 2010).

4.2.2 External estimates of measurement errors

SSMs can be associated with significant estimability problems, particularly when trying to estimate the two main sources of variability (Knape, 2008; Auger-Méthé et al., 2016). As a result, researchers often fix some of the parameters to known values, or use informed priors if they are working in a Bayesian framework. In particular, many use fixed values for the measurement errors (e.g., Jonsen, 2016). In some instances, one can use independent estimates of measurement errors (Jonsen et al., 2005) or can take advantages of time series with multiple data sources to estimate the errors of each data source (e.g., double tagged individuals in movement SSMs, Winship et al., 2012). While such method can alleviate estimation problems (Knape, 2008), one must be careful not to use biased or misspecified values.

4.2.3 Integrating additional data

Covariates that provide additional information about a state or a process (e.g., survival) may be a means of overcoming identifiability problems. Polansky et al. (2019) showed that non-identifiability in the estimation of a fecundity parameter and an observation bias parameter could be overcome by including a covariate in the model for fecundity.

Similarly, identifiability issues can be overcome by combining a SSM with a model for another data set that has parameters in common with the SSM. For example, in integrated population model, SSMs for time series of census data are combined with capture-recapture or mark-recovery data (Besbeas et al., 2002; Abadi et al., 2010). Adding additional data sources can be extremely useful but may not remove all identifiability issues. Methods for checking identifiability in integrated models are discussed in Cole and McCrea (2016).

4.2.4 Replicate observations

Having replicated observations through time (e.g., two independent population surveys) can help differentiate process variation from observation error, improve the parameter estimates accuracy, and improve the capacity of model selection methods to identify the correct observation distribution (Dennis et al., 2010; Knape et al., 2011). In many instances such replicated observations have already been collected, but are aggregated. For example, in population monitoring studies, subsamples (e.g., transect portions) are often aggregated into one overall estimate of abundance. Dennis et al. (2010) demonstrated that using these as replicates, rather than aggregating them, can improve the estimates. For animal movement models, individuals can been seen as replicates of the same process, but often the SSMs are fitted separately to each individual track. To improve inference, one can create a population model, where each individual track is linked to a distinct state time series but all share the same parameters (Jonsen, 2016). While the gains that can be made with replications are significant, one must understand the assumptions of models for replicated data. Simple population models for replicated datasets may assume that the replicates are independent (Dennis et al., 2010). However, many temporally varying factors (e.g., weather) may affect the sampling conditions and/or the behaviour of animals and result in correlations between replicates. Knape et al. (2011) demonstrated how to account for such dependence in population dynamics models. For animal movement, one may want to consider whether it is appropriate to assume that the behavioral mechanism driving movement is identical across individuals and, if not, may want to modify the model accordingly. However, as the gains that can be made with replications far surpasses those that could be made with longer time-series (Dennis et al., 2010), one should consider using replication in their models and when designing their studies. For example, Knape et al. (2011) suggested that in some cases managers may want to sample a population twice every second year rather than once a year. As SSMs are becoming the prime method to fit ecological time series, such study design issues should be explored further in the literature.

4.2.5 Matching temporal resolution for states and observations

The temporal resolution of the data can affect the parameter and state estimates and it is important to define a model at a resolution that is appropriate for the data. In many cases, adequate temporal span or resolution is more important than increased data quantity. For example, if a model describes a long-term cycle, then collecting data from more individuals is unlikely to make parameters estimable if the dataset is not long enough time to span the cycle being described (Peacock et al., 2016). If developing a model to classify a movement path into distinct behavioral modes, one must sample the movement track at a high enough

frequency so that multiple locations are recorded in each movement bout (Postlethwaite and Dennis, 2013). If one has a dataset with locations every 8 hrs, it would be extremely challenging to estimate behavioral states lasting less than 16-24 hrs. One can use pilot data, simulations, and data cloning to identify the temporal (and spatial) scale of sampling appropriate for the model, in something akin to a power analysis (Peacock et al., 2016). Overall, finding an appropriate model for your data, or collecting the appropriate data for your questions, can be an iterative process where one assess the estimability of different models under different data conditions.

4.2.6 Overcoming label-switching in Bayesian models

One important implementation problem with Bayesian SSMs with discrete states (e.g., HMMs) is label-switching (Jonsen et al., 2013). The labels given to the N discrete states are arbitrary, and thus there are $N!$ potential label assignments (Zucchini et al., 2016). The different label permutations result in the same model, a clear example of an identifiability problem. Thus when the MCMC chains have reached convergence, all possible labels will have been assigned to each state and inference on the states will be difficult. For example, we will no longer be able to take the mean of the posterior distribution to estimate the states because all $\hat{z}_{t,1:T} \approx N/2$. One solution is to impose constraints on the parameters that would be violated when labels are permuted (Zucchini et al., 2016). For example, in the two-behavior movement model described in Section 2.7 we would constrain $\gamma_1 \leq \gamma_2$. Label-switching is irrelevant to maximum likelihood and only relevant to Bayesian models with discrete states (Zucchini et al., 2016).

5 Model selection and comparison

Model comparison (or selection) can be used to compare the relative fit of models representing multiple working hypotheses, and to identify the model amongst these that best describe the data (see Section 6 for methods to evaluate the absolute fit of a model). Because different model structures can affect the estimated states and parameters (Knape et al., 2011), model comparison can be extremely useful in helping to refine state estimates (Auger-Méthé et al., 2017). Model comparison is common in ecology and has been used to compare SSMs (e.g., Siple and Francis, 2016). However, it is not uncommon for users to fit only a single SSM, likely due to the computational burden of fitting complex SSMs and some of the known limitations of applying model selection methods to SSMs (Jonsen et al., 2013). With the improved efficiency of fitting algorithms and advancements in model selection measures, model comparison of SSMs is becoming more attainable.

One common view is that ecological systems are so complex that it is impossible to develop a model that truly describes them, and that the goal of model selection is to find the best approximation of the truth (Burnham and Anderson, 2002). Under this paradigm, a useful way to compare models is by assessing how well they can be used to predict new data. Comparing the out-of-sample predictive accuracy of models can be done with cross-validation (see Section 6), and many advocate this as the best method to do so (Gelman et al., 2014). However, cross-validation is rarely done with ecological SSMs because it requires

fitting the same model multiple times, and thus can add significant computational burden to the analysis. While the gains in efficiency of fitting algorithms should make cross-validation increasingly feasible (but see Section 6 for potential problems), in this section we focus on what can be considered approximations of predictive accuracy. In particular, we discuss information criteria measures used with the frequentist and Bayesian approaches.

5.1 Frequentist approach

The most common model comparison measure in ecology is Akaike's Information Criterion (AIC; Aho et al., 2014). AIC was derived to estimate the expected and relative distance between the fitted model and the unknown true data-generating mechanism (Burnham and Anderson, 2002):

$$\text{AIC} = -2 \log L(\hat{\boldsymbol{\theta}}_{\text{MLE}}|\mathbf{y}) + 2k, \tag{63}$$

where $L(\hat{\boldsymbol{\theta}}_{\text{MLE}}|\mathbf{y})$ is the likelihood of the model at the MLE (i.e., the probability of the observed data given the model) and k is the number of parameters estimated. The model with the lowest AIC, thus the shortest distance from the truth, is considered the best model. Any model with an AIC within 4 units of the smallest AIC value should be considered as having some support. AIC is a measured aimed at selecting a model with the highest likelihood while penalizing for the number of parameters in the model. Models with more parameters will be more flexible and will tend to fit the data better by chance alone. As a result, they will often have poorer prediction performance than a simpler model. Gelman et al. (2014) provide an alternative description of AIC as equal to -2 times an approximation of the predictive accuracy of the model: $\log L(\hat{\boldsymbol{\theta}}_{\text{MLE}}|\mathbf{y}) - k$. This approximation accounts for the fact that we are using the same data to estimate the parameters and assess the model's predictive ability.

There are many issues related to using AIC with SSMs, and some have cautioned against this practice (e.g., Jonsen et al., 2013). We identified five different concerns. The first three concerns are related to the fact that the states of a SSM can be considered as random effects. First, using AIC to understand whether including random effects improves the model is difficult because some of the models may have parameters at the boundary of parameter space (Bolker et al., 2009). For example, testing whether or not there is process variance in SSMs (e.g., comparing our toy model to a model with no process variance, where $\sigma_p = 0$) could result in boundary problems, and is not recommended. Second, when you have random effects it is difficult to quantify the effective number of parameters (Bolker et al., 2009). For SSMs, it is difficult to know to what extent the states should be counted as estimated parameters and contribute to k. These two issues are most problematic when one uses AIC to select for models that differ in their random effects. If all SSMs compared have the same number of states and no additional random effects, we would expect the bias in the penalty k to be the same across models and thus have little effect on the difference in AIC across models. Third, one must decide whether the marginal likelihood or the conditional likelihood should be used when calculating AIC of a model with random effects (Müller et al., 2013). In contrast to the marginal likelihood, where we integrate out the states (Eq. 41), the conditional likelihood considers the states as known: $L_{\text{C}}(\boldsymbol{\theta}_o|\mathbf{z}_{1:T}, \mathbf{y}_{1:T}) = \prod_{t=1}^{T} g(y_t|z_t, \boldsymbol{\theta}_o)$. When the conditional likelihood is used in the AIC framework, both the parameter and state

estimates are plugged in and the penalty must account for the number of states (Vaida and Blanchard, 2005; Müller et al., 2013). The conditional AIC is a measure of the model's ability to predict new observations that share the same latent states, while the marginal AIC does not assume that the latent states are shared with the new observations and measures the model's ability to predict new observations from the same process (Vaida and Blanchard, 2005). For example, for a SSM describing the population dynamics of a fish species, we would interpret the conditional AIC as assessing the ability to predict another survey of the same population during the same time period. The marginal AIC would be assessing the ability of the model to predict a survey from a similar population of the same species. To our knowledge, the marginal likelihood has always been used with SSMs fitted in a frequentist framework, and no-one has investigated the potential advantages of using the conditional likelihood in the frequentist SSM literature (but see Section 5.2). The fourth source of concern is related to the problems associated with using AIC to choose the number of components in mixture models, which are particularly relevant for choosing the number of states in HMMs (Jonsen et al., 2013). Pohle et al. (2017) outline solutions to this HMM-specific problem.

The final concern, which is specific to cases with small sample size, is one that has been studied in the SSM literature. When the sample size, n, is small and the number of parameters, k, is relatively large (e.g., when $k \approx n/2$), the $2k$ penalty is inadequate and AIC has a tendency to favor more complex models (Cavanaugh and Shumway, 1997). Instead, many use the corrected AIC for small sample size:

$$\text{AICc} = -2 \log L(\hat{\boldsymbol{\theta}}_{\text{MLE}}|\mathbf{y}) + 2k \left(\frac{n}{n-k-1} \right). \tag{64}$$

However, Cavanaugh and Shumway (1997) noted that AICc may be inadequate for many SSMs, and suggested an alternative: the bootstrap-corrected measure, AICb. AICb has been used for ecological SSMs (Ward et al., 2010; Siple and Francis, 2016), especially by users of the `R` package `MARSS` (Holmes et al., 2012). This package for estimating the parameters of linear multivariate auto-regressive SSMs with Gaussian errors (i.e., multivariate dynamic linear models) has a function that calculates various versions of AICb. AICb was developed in the context of linear Gaussian SSMs, but is thought to be relatively robust to violations to normality (Cavanaugh and Shumway, 1997). We can describe AICb as:

$$\text{AICb} = -2 \log L_{\text{M}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}|\mathbf{y}_{1:T}) + 2 \left(\frac{1}{N} \sum_{i=1}^{N} -2 \ \log \frac{L_{\text{M}}(\hat{\boldsymbol{\theta}}^i|\mathbf{y}_{1:T})}{L_{\text{M}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}|\mathbf{y}_{1:T})} \right), \tag{65}$$

where $\hat{\boldsymbol{\theta}}^i$ is the i^{th} bootstrap replicate of $\hat{\boldsymbol{\theta}}$, N is the number of replicates, and $L_{\text{M}}(\hat{\boldsymbol{\theta}}^i|\mathbf{y}_{1:T})$ is the marginal likelihood of the model with the bootstrapped parameter sets given the original data. This bootstrap replicate can be achieved by simulating a time series from our model with $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ and estimating the parameters using this new time series. AICb was shown to outperform AIC and AICc when used with SSMs that had relatively small sample size for the number of parameter estimated (Cavanaugh and Shumway, 1997). The disadvantage of AICb is that it requires fitting the model N times. In the case of models that are computationally demanding to fit, one may need to continue to rely on AICc when sample

sizes are small. While AICc tends to erroneously choose more complex models compared to AICb, it is better than AIC and many other metrics for SSMs with small sample size (Cavanaugh and Shumway, 1997). Another similar computationally-intensive AIC variant for SSMs fitted to small samples has been developed by Bengtsson and Cavanaugh (2006), but its use in ecology has been limited by some of its constraints (e.g., Ward et al., 2010). For large datasets, some ecologists prefer to use BIC over AIC because AIC tends to choose more complex models as sample size increases. However, these two measures are used to achieve different inferential goals, and choosing between them is largely a philosophical question (see Aho et al., 2014; Hooten and Hobbs, 2015).

Overall, AIC and its small-sample alternatives can be used with SSMs in many instances, especially when the number of states and random effects are the same. AIC has been used for decades with SSMs (Harvey, 1990), and simple simulation studies have shown that AIC can be used to reliably select between SSMs (Auger-Méthé et al., 2017). Further research on the capacity of AIC to compare the predictive abilities of SSMs under various conditions would be useful. In the meantime, one should be aware of the limitations outlined above, and interpret the results accordingly.

Other frequentist methods may be used to select between SSMs. For example, likelihood ratio tests can be used to select between nested models, especially when conducting planned hypothesis testing (e.g., Karban and de Valpine, 2010). However, likelihood ratio tests will suffer from some of the same issues as the those highlighted for AIC. Newman et al. (2014) also highlighted the potential use of score tests, transdimensional simulated annealing, and other methods. To our knowledge, these alternative methods have not been used in the SSM literature, but may be the focus of future research.

5.2 Bayesian approaches

Two Bayesian information criteria, the Deviance Information Criteria (DIC) and the Watanabe-Akaike information criterion (WAIC), are popular with hierarchical models, and have been used with SSMs. These two measures are similar to AIC, but unlike AIC they both estimate the effective number of parameters using data-based bias correction rather than a fixed rule. These data-based methods attempt to account for the effects of priors and the hierarchical structure (e.g., the characteristics of the random effects) on the flexibility of the model.

DIC uses a formulation similar to AIC (Gelman et al., 2014; Hooten and Hobbs, 2015):

$$\text{DIC} = -2 \log L(\hat{\boldsymbol{\theta}}_{\text{POST}}|\mathbf{y}) + 2p_{\text{DIC}}. \tag{66}$$

Here, the measure of fit, $L(\hat{\boldsymbol{\theta}}_{\text{POST}}|\mathbf{y})$, is the likelihood of the model at the mean of the posterior distribution of the parameters rather than at the MLE. There are alternatives for how to calculate the effective number of parameters (Gelman et al., 2014), p_{DIC}, but a common one is -2 times the difference between the posterior mean log likelihood and the log likelihood evaluated at the posterior mean of the parameter:

$$p_{\text{DIC}} = -2 \left(\int \log L(\boldsymbol{\theta}|\mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} - \log L(\hat{\boldsymbol{\theta}}_{\text{POST}}|\mathbf{y}) \right). \tag{67}$$

In practice, this would be calculated as: $p_{\text{DIC}} = -2(\frac{1}{S}\sum_{s=1}^{S} \log p(\mathbf{y}|\boldsymbol{\theta}^s) - \log L(\hat{\boldsymbol{\theta}}_{\text{POST}}|\mathbf{y}))$, where $\boldsymbol{\theta}^s$ are samples from the posterior distribution.

DIC has been shown to accurately select ecological SSMs in certain contexts (Wilberg and Bence, 2008, but see Chang et al. 2015), and has been used to do so (e.g., Michielsens et al., 2006). DIC is easy to use because MCMC sampler software (e.g., JAGS, Plummer, 2003) and `R` packages like `rjags` (Plummer, 2018) have functions that compute it easily. However, DIC has multiple drawbacks. DIC performs better when the number of effective parameters is much smaller than the sample size (Hooten and Hobbs, 2015), a condition likely uncommon with SSMs because the number of latent states scales with the sample size. In addition, DIC is known to be problematic for mixture models, can poorly estimate the effective number of parameters (e.g., can return negative numbers), relies on approximate posterior normality, and is not fully Bayesian because its measure of fit relies on the posterior mean of $\boldsymbol{\theta}$ (i.e., a point estimate) instead of the entire posterior distribution (Gelman et al., 2014; Hooten and Hobbs, 2015; Kai and Yokoi, 2019). These limitations may explain why Chang et al. (2015), in contrast to Wilberg and Bence (2008), showed that DIC had difficulties selecting amongst SSMs.

Many now favor WAIC, a recently developed Bayesian information criterion (Gelman et al., 2014; Hooten and Hobbs, 2015):

$$\text{WAIC} = -2\sum_{i=1}^{T}\log\int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} + 2p_{\text{WAIC}}. \tag{68}$$

The first component of WAIC is also a measure of fit, but unlike DIC it uses the entire posterior distribution for $\boldsymbol{\theta}$ rather than a point estimate. As such, we can consider this measure of fit as truly Bayesian. This measures also implicitly assumes that the observations are independent of each other (i.e., that $-2\log p(y_{1:T}|\mathbf{y}) = -2\sum_{t=1}^{T}\log(p(y_t|\mathbf{y})) = -2\sum_{t=1}^{T}\log\left[\int p(y_t|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\right]$). In practice, the log pointwise predictive density, $\sum_{i=1}^{T}\log\int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}$, is calculated using S samples from the posteriors: $\sum_{i=1}^{T}\log(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\boldsymbol{\theta}^s))$. As for DIC, there are different ways to estimate the effective number of parameters, p_{WAIC}. Gelman et al. (2014) recommend using $\sum_{i=1}^{T}\text{Var}_{\text{POST}}(\log p(y_i|\boldsymbol{\theta}))$ as it gives results closer to the leave-one-out (loo) cross validation. In practice, it can be computed as $\sum_{i=1}^{T}(\frac{1}{S-1}\sum_{s=1}^{S}(\log p(y_i|\boldsymbol{\theta}^s) - \overline{\log p(y_i|\boldsymbol{\theta}^*)})^2)$, where $\overline{\log p(y_i|\boldsymbol{\theta}^*)}$ is the mean log probability of data point y_i across all S parameter samples.

WAIC has been used to compare ecological SSMs (e.g., Baldwin et al., 2018; Ferretti et al., 2018) and can be computed using the R package `loo` (Vehtari et al., 2017). Recent reviews of Bayesian model comparison methods tend to favor WAIC over DIC (Gelman et al., 2014; Hooten and Hobbs, 2015) because it is a fully Bayesian metric, it is not affected by parametrization, and will not return negative values for the effective number of parameters. However, WAIC has a few shortcomings. Most importantly, both parts of WAIC are computed by using the sum over each data point i, and thus rely on partitioning the data into disjoint, ideally conditionally independent, pieces (Gelman et al., 2014). Such partitioning can be problematic with SSMs since the time-series nature of the data generally results in dependence structure in the observations. While AIC and DIC do not require this explicit partitioning, they also assume conditional independence. A solution to further explore for WAIC could be to write $\log p(y_i|\boldsymbol{\theta}^s)$ in terms of conditional distributions: $\log p(y_1|\boldsymbol{\theta}^s) + \sum_{t=2}^{i}\log p(y_t|y_{1:t-1}, \boldsymbol{\theta}^s)$.

Just as for AIC, we can use either the conditional or marginal likelihood with DIC and WAIC (Kai and Yokoi, 2019; Merkle et al., 2019). With the Bayesian approach, the likelihood is generally defined as fully conditional on both parameters and latent states and both are generally sampled when sampling the posterior. Thus, the conditional likelihood is usually used with Bayesian metrics even though this is rarely specified (Millar, 2018; Merkle et al., 2019). While computing the marginal likelihood version of these Bayesian metrics is more computationally expensive, DIC and WAIC appear to more reliably select the true underlying SSM when the marginal likelihood is used (Kai and Yokoi, 2019).

As Gelman et al. (2014) noted, we are asking close to the impossible from these information criteria measures: an unbiased estimate of out-of-sample prediction error based on data used to fit the model that works for all model classes and requires minimum computation. As such, metrics such as WAIC can be unreliable estimates of the predictive ability of ecological models (Link et al., 2017). While further research is needed to assess when WAIC is appropriate for SSMs and to identify data partitioning schemes that resolve some of the potential biases, WAIC is likely the best information criterion for Bayesian SSMs at this point. If one is unwilling to accept the shortcomings of WAIC, the best solution is to compare models using the more computer intensive cross-validation methods (Gelman et al., 2014; Link et al., 2017; Vehtari et al., 2017). Cross-validation will also require one to partition data intelligently, but this may be more easily implemented with blocking (Gelman et al., 2014; Roberts et al., 2017).

Other methods could be used to compare models in a Bayesian framework (e.g. Newman et al., 2014). For example, reversible-jump MCMC has been used to compare SSMs (McClintock et al., 2012), but is known to be difficult to implement (Hooten and Hobbs, 2015). The importance of multiple covariates in a model (e.g., the effect of temperature and precipitation on bird survival) can be assessed by multiplying coefficients in a model by indicator variables which when equal to one include the covariate and when equal to zero exclude the covariate (O'Hara et al., 2009). Such techniques have been used to compare ecological SSMs (Sanderlin et al., 2019), but such an approach is designed for nested models only. Posterior predictive loss approaches appear to be suitable for time-series data (Hooten and Hobbs, 2015) and have been used to compare ecological SSMs (Mills Flemming et al., 2010). While these alternative approaches may not be as commonly used to compare ecological SSMs, and will have drawbacks, many of them warrant further exploration.

5.3 Model averaging

Model averaging can combine the strength of several models and account for model uncertainty, something model selection cannot offer (Buckland et al., 1997; Hooten and Hobbs, 2015). Wintle et al. (2003) argued against using a single model to make predictions because uncertainty about model structure is often high in ecology, and alternative models can have prediction differences with important repercussions for management decisions. When one selects a single model, and presents the parameter and state estimates based on this best model, one implicitly assumes that the model is true and that the uncertainty is only in the estimation process (Buckland et al., 1997; Wintle et al., 2003). One can instead use model averaging, where, for example, each model is weighted and the predictions are a weighted sum across the plausible models (Wintle et al., 2003). Both the parameters and

the predictions could be averaged, but this must be done with care and we would generally caution against averaging parameters. In many cases, differences in model structure result in changing the meaning of parameters, thus making their average nonsensical (Dormann et al., 2018). Model averaging has been used in a few studies applying SSMs to ecological data (e.g., Maunder and Deriso, 2011; Moore and Barlow, 2011) and was shown to provide unbiased estimates (Wilberg and Bence, 2008). However, simulations studies have shown that model averaging may not always provide more accurate point estimates than the best SSMs (Wilberg and Bence, 2008; Chang et al., 2015). In addition, while model averaging generally reduces prediction errors compared to each of the contributing models, these gains can be counteracted by factors such as uncertainty in the model weights and covariance between models (Dormann et al., 2018). In addition, calculating weights using parametric methods such as AIC can perform poorly (Dormann et al., 2018). We refer interested readers to a recent review by Dormann et al. (2018), which provides an in-depth discussion of model averaging in ecology.

6 Diagnostics and model validation for SSMs

While model selection can help us identify which of the fitted models best describes the data, it does not provide an assessment of the absolute fit of that model. As such, the selected model could be a poor representation of the data generating process (i.e., could poorly describe the ecological process and/or measurement process) and relative measures of fit, such as AIC, do not quantify how closely the model matches the data. Thus, before interpreting model results, it is crucial to carry out some of the following model diagnostics. First, it is essential to examine whether estimated parameters seem biologically reasonable. For example, our understanding of the system may stipulate that a response variable should increase with a covariate. A model with parameter estimates inconsistent with such *a priori* understanding will be suspect. Second, it is important to assess the influence of individual observations on estimated parameters. For example, outliers can have a strong influence on parameter estimates. Third, one should examine whether the model assumptions are reasonable. For example, with SSMs, assumptions are made about the probability distributions for states and observations (e.g., Eqs. 1-2 assume both are normal). Fourth, it is useful to examine the goodness of fit, which defines how well the model describes the data used to fit the model. At the individual observation level, it measures how far an observation is from its predicted value (e.g., $|y_t - \hat{y}_t|$, t=1,..., T). At the model level, it summarizes the overall fit of a model to all observations (e.g., the average squared errors). Fifth, one ought to assess the model's predictive accuracy, or how well the model predicts an outcome for an observation that was not used to fit the model (e.g., via cross-validation). With time-series models, including SSMs, one can use the first t observations to fit the model, and then use the model to predict the $t + 1$ observation, or fit the model to all T observations and see how well future observations are predicted.

6.1 Challenges with SSMs

For simple statistical models, such as linear regression, diagnostics for most of the above features are well established. Diagnostics for SSMs, however, can be challenging for two reasons. First, observations are temporally dependent. Many diagnoses rely on response or conventional residuals, which we define as follows for our toy model:

$$e_{t|1:T} = y_t - \hat{y}_{t|1:T}, \tag{69}$$

where $\hat{y}_{t|1:T}$ is the predicted observation at time t given all observations. This predicted value depends on the smoothed state estimate at time t, $\hat{z}_{t|1:T}$, and the observation equation. For example, for our toy model (Eqs. 1-2), $\hat{y}_{t|1:T} = E[y_t|y_{1:T}] = \alpha\hat{z}_{t|1:T}$. Harvey (1990) notes that these response residuals are not serially independent. Their use can impair one's capacity to identify model misspecification (Harvey, 1990), and can have negative consequences for model inference and further model diagnosis (e.g., inflated goodness of fit, Thygesen et al., 2017). Second, as for most hierarchical models, we generally do not have direct observations of the hidden states, z_t, thus one cannot directly compare predicted states with their "true" values.

Because of these challenges, researchers often fail to check the absolute fit of SSMs, and thus risk making conclusions based on a misspecified model or risk having biased parameter and state estimates. Here, we provide a list of tools to help researchers perform this essential model-checking step. We start with the tools commonly used to assess Bayesian hierarchical models. These tools can be easily used with frequentist and Bayesian SSMs alike, but have important limitations. We then discuss the tools that have been the focus of model validation developments for SSMs, which specifically address the issue of temporal dependence in the residuals. We end with methods relying on out-of-sample validation (e.g., cross-validation), which we believe is the gold standard for assessing the predictive ability of a model, and we hope will become the focus of future developments for SSMs. This order also reflects an increased division between the data used to estimate the model parameters and hidden states and the data used to perform the diagnostics.

6.2 Posterior predictive measures

Posterior predictive checking is a common Bayesian method to quantify the discrepancies between the data and the model, and to assess model assumptions (Gelman et al., 2013; Conn et al., 2018). It has been used to verify the fit of SSMs to ecological data (e.g., Hobbs et al., 2015). The idea behind posterior predictive checking is that if the model fits the data well, then data generated from the model should have characteristics similar to those of the observed data (Gelman et al., 2013). These posterior predictive checks often involve calculating a posterior predictive p-value, p_B, which is similar to the p-value used in frequentist inference, and is defined as follows:

$$p_B = Pr(T(\mathbf{y}^i, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}), \tag{70}$$

where each $\mathbf{y}^i$ is a time series that has been simulated from the fitted model (i.e., representing a replicate time series that could have been observed from the model), $\mathbf{y}$ is the observed

data, $\boldsymbol{\theta}$ contains the model parameters, and $T(\mathbf{y}, \boldsymbol{\theta})$ is a test quantity summarizing the data (e.g., the mean) or a discrepancy function (e.g., χ^2 measure). This p-value measures the probability, under the model of interest, of finding a test quantity as extreme as that associated with the data. Posterior predictive checks are estimated via simulations, where we sample a set of posterior $\boldsymbol{\theta}$ values and from each simulate one $\mathbf{y}^i$. For each $\mathbf{y}^i$, we calculate the test quantity and compare it to the value for the observed data. We estimate the p-value with the proportion of these replicates that have a test quantity value greater or equal to that of the real data. When the posterior predictive p-value is near 0 or 1, it indicates that the pattern observed with the data would be unlikely if the model were true. The relevance of the p-value largely depends on the choice of test quantity, which generally measures a feature of the data that is not already incorporated in the model. For example, Hobbs et al. (2015) used the mean and standard deviation of the observed data, as well as: $T(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^{T}(y_t - \hat{y}_{t|1:T})^2 = \sum_{t=1}^{T} e_{t|1:T}^2$. The last of these quantities is a discrepancy function, which by definition measures the disagreement between the model and the data. Newman et al. (2014) and Conn et al. (2018) provide lists of alternative functions. While we described posterior predictive checks in a Bayesian framework and have defined the test quantity as a function of $\mathbf{y}$, King et al. (2015) have applied similar concepts in a frequentist framework and used test quantities that describe characteristics of the estimated hidden states, $\mathbf{z}$ (e.g., autocorrelation function at lag 1 of the states).

Although posterior predictive p-values are a common model validation method, they have important limitations (Conn et al., 2018). Because they use the data twice, once to fit the model and once to test the model fit, they tend to be conservative (i.e., tend to return value closer to 0.5 than to 0 or 1), and often have insufficient power to detect lack of fit. One can alter the method described above and generate all the observation replicates using only a single sample from the posterior parameter distribution. This sampled posterior p-values method was shown to have better theoretical properties, with better Type I error rate control, and shown to detect lack of fit more reliably for some ecological hierarchical models (Conn et al., 2018). Following King et al. (2015), we recommend assessing discrepancies between the model and the data by looking at where $T(\mathbf{y}, \boldsymbol{\theta})$ falls in the frequency distribution of $T(\mathbf{y}^i, \boldsymbol{\theta})$ (Fig. 4 A, D). This graphical method is also more useful in assessing the ecological importance of the discrepancies than looking at the p-value, and can provide a better sense of why the model may be inappropriate for the data (Conn et al., 2018).

While posterior predictive checks (e.g., Bayesian p-values) are often used to assess how well the model describes the data, they can also be used to assess the validity of model assumptions (Gelman et al., 2013). We can use a single sample from the posterior distribution of the hidden states to assess the assumptions associated with the process equation (Thygesen et al., 2017). For example, we can sample a time series of state, $\mathbf{z}^i$, from the posterior state distribution of our toy model to calculate the process variation as $\epsilon_t^i = z_t^i - \beta z_{t-1}^i$, and verify whether the ϵ_t^i are normally distributed with a mean of 0 as assumed by Eq. 1. Departures from the assumed distribution (e.g., if the mean of the process variation is far from 0), indicate that the model is not adequate for the data. This method is generally recommended for assessing the assumptions of Bayesian hierarchical models (Gelman et al., 2013), but Thygesen et al. (2017) used the Laplace approximation implemented in `TMB` to create a posterior distribution of the states for non-Bayesian models.

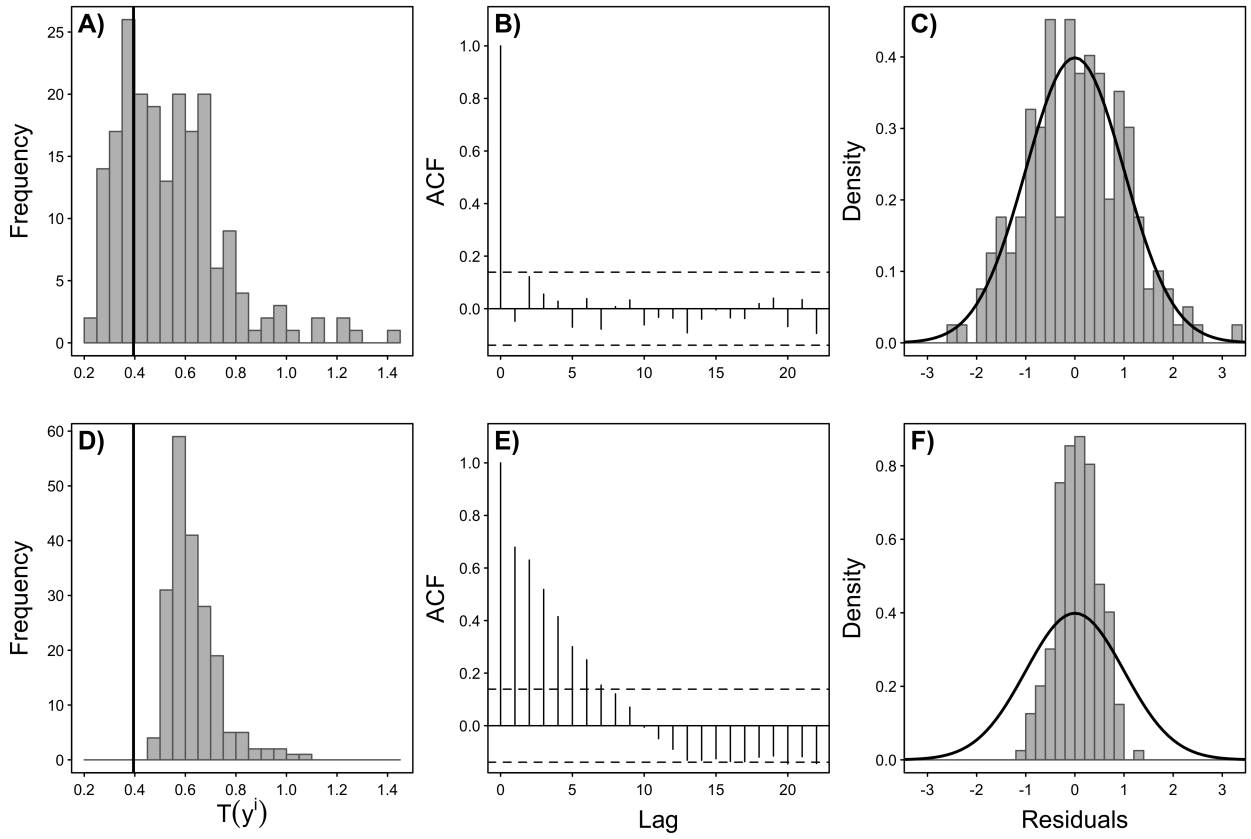

Figure 4: Examples of diagnostic plots for a well specified (panels A, B, C) and a misspecified model (panels D, E, F). The data for all plots were simulated from the toy model (Eqs. 1-2) with $\alpha = \beta = 1$ and the $\sigma_p = \sigma_o = 0.1$. The correctly specified model used the correct known values of α and β, and estimated σ_p and σ_a. While the misspecified model also used the correct known values of α and β, it wrongly assumed a value of $\sigma_o = 0.5$ and then estimated σ_p. Panels A) and D) represent a frequentist version of the posterior predictive check, where the test quantity is the standard deviation of the observations, $T(\mathbf{y}) = \sqrt{\sum_{t=1}^{T} (y_t - \bar{y})^2/(T-1)}$. The histograms represent the frequency of test quantity for 200 datasets simulated using the estimated parameters with the original dataset. The vertical bar is the test quantity for the original dataset. Panels B) and E) represent the autocorrelation function of the one-step-ahead residuals. Panels C) and F) compare the distribution of the observed one-step-ahead residuals (histograms) to a standard normal probability density function (curves).

6.3 One-step-ahead residuals and their extensions

The model diagnostic that has received the most attention in the SSM literature is the one-step-ahead residuals (Harvey, 1990; Thygesen et al., 2017), also known as recursive residuals (Frühwirth-Schnatter, 1996). Unlike the response residuals (Eq. 69), the one-step-ahead residuals should not have temporal dependence when the model is adequate because the residual for the t^{th} observation uses the expected observation at time t given observations only up to time $t-1$:

$$e_{t|1:t-1} = y_t - \hat{y}_{t|1:t-1}. \tag{71}$$

Effectively, for response residuals we use the smoothed estimates of states, $\hat{z}_{t|1:T}$, to predict the observation at time t, while for one-step-ahead residuals, we use the prediction of the states, $\hat{z}_{t|1:t-1}$. In the context of a Kalman filter, we can calculate $\hat{y}_{t|1:t-1}$ using the one-step-ahead forecast prediction that is already calculated as part of the recursive algorithm. As more information is available for fitting the model as t increases, the variance of prediction residuals will tend to decrease with t. To account for this change in variance, it is useful to scale the prediction residuals by their standard deviations (a procedure equivalent to calculating standardized Pearson residuals):

$$\tilde{e}_{t|1:t-1} = \frac{e_{t|1:t-1}}{sd(e_{t|1:t-1})}. \tag{72}$$

For the special case of SSMs with normally distributed states and observations, such standardized residuals are independent and identically distributed with a standard normal distribution and can be used to test a variety of assumptions. Diagnostic procedures include qq-normal plots to check for normality, auto-correlation function plots to see if the residuals are independent, and plots of the residuals against observed values (y_t) to check for non-constant variance.

For non-normal SSMs, the probability distribution of these standardized residuals are not standard normal, making the exploration of residuals harder. Probability scores (P-scores), and their transformed version, prediction quantile residuals, are useful alternatives (Frühwirth-Schnatter, 1996; Thygesen et al., 2017). A P-score, u_t, is the cumulative distribution function for the predicted observations evaluated at the t^{th} observed value:

$$u_t = F_{Y_t|y_{1:t-1}}(y_t) = Pr(Y_t \leq y_t | Y_{1:t-1} = y_{1:t-1}). \tag{73}$$

If $F_{Y_t|y_{1:t-1}}$ describes the cumulative distribution function of the true model, then the resulting u_t are uniformly distributed (Conn et al., 2018). Deviations from uniformity suggest model misspecification. As this is simply an application of the probability integral transformation (i.e., if Y has the cumulative density function F_Y, then $F_Y(Y)$ is distributed with Uniform(0,1), Smith, 1985), these are a specific case of probability integral transform (PIT) residuals (Warton et al., 2017). To get normally distributed residuals, we can transform the P-scores to prediction quantile residuals, v_t as follows:

$$v_t = \Phi^{-1}(u_t), \tag{74}$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function (also known as the standard normal quantile function). When the model is true, v_t should be an independent sample from a standard normal. Thus, we can assess whether the data fits the model assumptions using the same diagnostic procedures available for standardized one-step-ahead prediction residuals in the case of normally distributed SSMs (see Fig. 4 B-C, E-F; and Newman et al., 2014; Thygesen et al., 2017).

P-scores and prediction quantile residuals can be difficult to estimate for non-normal SSMs because their calculation requires knowledge of the cumulative distribution function (cdf) for $Y_t|y_{1:t-1}$, which in many cases will not be known nor have an analytical form. However, Thygesen et al. (2017) developed methods for approximating the cdf based on the Laplace approximation that can be implemented easily in `TMB`. Because this method depends on the Laplace approximation, it is important to assess the accuracy of this approximation (see Appendix S1). The quantile residuals of Thygesen et al. (2017) are applicable to a broad range of frequentist SSMs, although there are some limitations in using them with multivariate time series. We are not aware of equivalent methods for as broad a range of Bayesian SSMs, although some exist for a limited class (Frühwirth-Schnatter, 1996; Gamerman et al., 2013).

6.4 Cross-validation

While one-step-ahead residuals and their extensions remove data when calculating the expected value of the observation at time t, they use the complete dataset to estimate the global model parameters. Thus, these residuals cannot be used to fully assess the predictive ability of the model. The gold standard for assessing the predictive ability of a model is thought to be achieved with out-of-sample data, where two independent datasets are used: one to fit (or train) the model and one to validate (or test or evaluate) it (Hooten and Hobbs, 2015). In practice, this is often achieved by dividing a dataset in two. Because of this division, a large initial dataset is required to allow for a large enough sample for fitting the model and large enough sample to assess the predictive performance of the model. While this validation procedure is rarely done with SSMs, we have seen a few examples where one of several data streams is used as validation data (Hobbs et al., 2015), part of a single time series was set as a validation period (Holdo et al., 2009), or independent information on the true values of the hidden states was collected (see Auger-Méthé et al., 2017, and Appendix S1).

When leaving out a single subset of the data, we can only assess the predictive ability for these specific observations. Thus, it is common to use a cross-validation method that looks at the predictive ability of all data points by sequentially leaving out small subsets of the data (Hooten and Hobbs, 2015). k-fold cross validation is a ubiquitous statistical method that divides the data into k groups of similar size. The k groups sequentially serve as the validation dataset while the remaining $k-1$ groups are collectively used as the training set. Leave-one-out (loo) is a common version of k-fold cross validation that leaves each of the data points out sequentially. Score or discrepancy functions can then be used to assess the predictive ability of the model. A common score function is the mean squared prediction error (MSPE) for group k, which if we assume that T/k is an integer n, can be written as

follows:

$$\text{MSPE}_k = \sum_{i=1}^{n} (\mathbf{y}_{i,k,\text{oos}} - \hat{\mathbf{y}}_{i,k,\text{oos}})^2 / n, \tag{75}$$

where *oos* means out of sample, $\mathbf{y}_{i,k,\text{oos}}$ is the i^{th} observation in subsample k, and $\hat{\mathbf{y}}_{i,k\text{oos}}$ is the expected observation based on the model fitted to the dataset without sample k. As an overall value, we can then average the k MSPE_k. Such functions directly assess the predictive ability of the model and thus are intuitive measures of how good a model is. As mentioned in Section 5, cross-validation is also often deemed the preferred method for model comparison (Gelman et al., 2014; Hooten and Hobbs, 2015). While cross-validation can be implemented relatively easily, it can be computationally demanding (Link and Sauer, 2016; Vehtari et al., 2017). Cross-validation generally requires refitting the models k times, which can be a daunting task with Bayesian models (but see Hooten and Hobbs, 2015, for suggested solutions). In addition, cross-validation assumes that the training and evaluation datasets are independent (Roberts et al., 2017). The main challenge with using cross-validation with SSMs is that, due to the temporal dependency in the data, removing only a few data points will underestimate the prediction error and removing many will lead to propagation of error (Newman et al., 2014). Despite these drawbacks cross-validation is a powerful tool. Unfortunately, there are only a few examples of cross-validation methods for SSMs (Ansley and Kohn, 1987; de Jong, 1988). However, the time-series literature (e.g., Bergmeir and Benítez, 2012) and the suggestions of Roberts et al. (2017) on block cross-validation methods to account for dependence structure in ecological data are useful starting points for the development of such methods for SSMs.

The topic of model validation for SSM is one that has been relatively poorly studied, with a few notable exceptions (e.g., Frühwirth-Schnatter, 1996; King et al., 2015; Thygesen et al., 2017). Because of the additional parameter identifiability and estimability problems discussed in Section 4, we believe this topic deserves more attention. Beyond the tools we have outlined above, SSM developers and users should gain inspiration from the tools developed for hierarchical models (e.g., PIT-trap residuals, Warton et al., 2017). For researchers using Bayesian SSMs, we point readers towards the review of Conn et al. (2018) on model checking methods for Bayesian hierarchical models. Finally, we would like to remind readers that, while it is crucial to perform a model validation step, passing this step does not mean that the model is representing the truth. It simply means that one could not find difference between the data generating system and the model. This could be due to a low sample size or the conservative nature of some of the methods described above.

7 Conclusion

Through a diverse set of examples, we have demonstrated that SSMs are flexible models for time series that can be used to answer a broad range of ecological questions. They can be used to model univariate or multivariate time series. SSMs can be linear or nonlinear, and have discrete or continuous time steps. They can have normal or non-normal sources of stochasticity, and thus can model continuous, count, binary, or categorical data. They are particularly useful when one has significant process variation and observation error.

Accounting for these sources of uncertainty can substantially affect management decisions, making SSMs the perfect modeling tool in many contexts (e.g., Jamieson and Brooks, 2004; Hobbs et al., 2015).

As we have outlined, a variety of tools to fit SSMs to data exist. Historically, many researchers wrote SSMs so they could be fitted with the Kalman filter and its extensions. However, the diversity of fitting procedures available now allows researchers to create models that are more representative of the structure of their data and the ecological processes they are interested in. In addition, flexible fitting tools now exist in both the frequentist and Bayesian frameworks, allowing researchers to choose their preferred inferential framework rather than have their model dictate the framework they can use. Within each inferential framework, the choice of a fitting procedure will be a compromise between flexibility and efficiency. In particular, highly efficient fitting methods (e.g., Laplace approximation and Hamiltonian Monte Carlo) have more restrictions than their slower alternatives (e.g., particle filter and Gibbs).

While these tools provide the means to fit complex SSMs, it is crucial to appropriately formulate the model. As discussed, SSMs can suffer from parameter estimability problems, but various tools exist to assess whether this is the case and to identify the type of study design or model simplification that will resolve these problems. In general, making use of replication or including covariates can help reduce some of the common estimation problems.

Researchers often forgo doing model selection and validation with SSMs, but we advocate that these should become part of every SSM user's workflow. Model mispecification can affect ecological inferences and the accuracy of state estimates. While no model selection measure is perfect for SSMs, AIC and WAIC, can be useful. Similarly, while model validation is difficult with SSMs, posterior predictive measures, and one-step-ahead residuals and their extensions are relatively easy ways to assess whether the model describes the data well and whether some of the model assumptions are met.

While there are many tools already available to fit, compare, and validate SSMs, five topics warrant further research. First, it would be important to pinpoint the conditions under which SSMs, rather than simpler alternatives, are needed (e.g., when do SSMs provide more reliable parameter and state estimates than models without observation error). Such research should account for the additional identifiability and estimability issues that comes with fitting SSMs and the types of datasets that allow SSMs to return reliable estimates. Second, as SSMs are often the primary tools used to analyze time series, it is important to explore the study designs that optimize the estimation of SSMs, so that the best data possible are collected. Third, there is a need for further research on model selection procedures for SSMs. For example, understanding whether one should use the marginal versus the conditional likelihood with AIC and WAIC should be further explored (e.g., do the marginal and conditional likelihoods affect the predictive accuracy of the states and parameters differently?). Fourth, while there have been a few important advances in model validation methods for SSMs, this remains a relatively untouched research area. Given the increasing use of SSMs in management, it is crucial that a broader range of validation methods be developed for these complex models. Fifth, with the increasing efficiency of fitting procedures, cross-validation is becoming a feasible procedure to assess predictive accuracy and compare models. As such, the time is ripe to start developing proper cross-validation procedures that will account for dependencies in the data.

Overall, we provided a review of the topics needed to develop and fit SSMs to ecological data, and Appendix S1 provide an extensive set of examples of methods to facilitate this process. We hope this guide will help researchers develop and apply SSMs to their own data, and foster the development of SSMs in multiple fields of ecology.

8 Acknowledgements

This paper was instigated during a Banff International Research Station (BIRS) workshop hosted at the Casa Matemática Oaxaca (CMO) entitled *New perspectives on state-space models*. We thank BIRS and CMO for their support and the lead organizer of the workshop, David Campbell, and all participants for their insights. This effort was also supported by the Canadian Statistical Sciences Institute through a Collaborative Research Team Project led by JMF. MAM thanks the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program. We thank Devin Lyons for his comments on previous versions of the manuscript. We also thank Andrew Derocher for the polar bear track used in Fig. 2 and Appendix S1.

References

Abadi, F., Gimenez, O., Ullrich, B., Arlettaz, R., and Schaub, M. (2010). Estimation of immigration rate using integrated population models. Journal of Applied Ecology, 47:393–400.

Aeberhard, W. H., Mills Flemming, J., and Nielsen, A. (2018). Review of state-space models for fisheries Science. Annual Review of Statistics and Its Application, 5:215–235.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. Ecology, 95:631–636.

Albertsen, C. M., Whoriskey, K., Yurkowski, D., Nielsen, A., and Mills Flemming, J. (2015). Fast fitting of non-Gaussian state-space models to animal movement data via Template Model Builder. Ecology, 96:2598–2604.

Anderson-Sprecher, R. and Ledolter, J. (1991). State-space analysis of wildlife telemetry data. Journal of the American Statistical Association, 86:596–602.

Ansley, G. F. and Kohn, R. (1987). Efficient Generalized Cross-Validation for State Space Models. Biometrika, 74:139–148.

Auger-Méthé, M., Albertsen, C. M., Jonsen, I. D., Derocher, A. E., Lidgard, D. C., Studholme, K. R., Bowen, W. D., Crossin, G. T., and Mills Flemming, J. (2017). Spatiotemporal modelling of marine movement data using Template Model Builder (TMB). Marine Ecology Progress Series, 565:237–249.

Auger-Méthé, M., Field, C., Albertsen, C. M., Derocher, A. E., Lewis, M. a., Jonsen, I. D., and Flemming, J. M. (2016). State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. Scientific Reports, 6:26677.

Baldwin, J. W., Leap, K., Finn, J. T., and Smetzer, J. R. (2018). Bayesian state-space models reveal unobserved off-shore nocturnal migration from Motus data. Ecological Modelling, 386:38–46.

Bengtsson, T. and Cavanaugh, J. E. (2006). An improved Akaike information criterion for state-space model selection. Computational Statistics and Data Analysis, 50:2635–2654.

Berg, C. W. and Nielsen, A. (2016). Accounting for correlated observations in an age-based state-space stock assessment model. ICES Journal of Marine Science: Journal du Conseil, 73:1788–1797.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191:192–213.

Besbeas, P., Freeman, S. N., Morgan, B. J. T., and Catchpole, E. A. (2002). Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. Biometrics, 58:540–547.

Besbeas, P. and Morgan, B. J. T. (2019). Exact inference for integrated population modelling. Biometrics, 75:475–484.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv e-prints, page arXiv:1701.02434.

Bolker, B. M. (2008). Ecological models and data in R. Princeton University Press, Princeton, NJ.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution, 24:127–135.

Boone, E. L., Merrick, J. R., and Krachey, M. J. (2014). A Hellinger distance approach to MCMC diagnostics. Journal of Statistical Computation and Simulation, 84:833–849.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7:434–455.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. Biometrics, 53:603–618.

Burnham, K. P. and Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach. Springer-Verlag, New York, NY, 2nd edition.

Campbell, D. and Lele, S. (2014). An anova test for parameter estimability using data cloning with application to statistical inference for dynamic systems. Computational Statistics and Data Analysis, 70:257–267.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. Biometrika, 81:541–553.

Catchpole, E. A., Kgosi, P. M., and Morgan, B. J. T. (2001). On the near-singularity of models for animal recovery data. Biometrics, 57:720–726.

Catchpole, E. A. and Morgan, B. J. T. (1997). Detecting parameter redundancy. Biometrika, 84:187–196.

Catchpole, E. A., Morgan, B. J. T., and Freeman, S. N. (1998). Estimation in parameter-redundant models. Biometrika, 85:462–468.

Cavanaugh, J. and Shumway, R. (1997). A bootstrap variant of the AIC for state-space model selection. Statistica Sinica, 7:473–496.

Chaloupka, M. and Balazs, G. (2007). Using Bayesian state-space modelling to assess the recovery and harvest potential of the Hawaiian green sea turtle stock. Ecological Modelling, 205:93–109.

Chang, Y.-J., Brodziak, J., O'Malley, J., Lee, H. H., DiNardo, G., and Sun, C.-L. (2015). Model selection and multi-model inference for Bayesian surplus production models: a case study for Pacific blue and striped marlin. Fisheries Research, 166:129–139.

Choquet, R. and Cole, D. J. (2012). A hybrid symbolic-numerical method for determining model structure. Mathematical Biosciences, 236:117–125.

Cobelli, C. and DiStefano III, J. (1980). Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 239:7–24.

Cole, D. J. and McCrea, R. S. (2016). Parameter redundancy in discrete state-space and integrated models. Biometrical Journal, 58:1071–1090.

Cole, D. J., Morgan, B. J. T., Catchpole, E. A., and Hubbard, B. A. (2012). Parameter redundancy in mark-reovery models. Biometrical Journal, 54:507–523.

Cole, D. J., Morgan, B. J. T., and Titterington, D. M. (2010). The parametric structure of models. Mathematical Biosciences, 228:16–30.

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., and Hooten, M. B. (2018). A guide to Bayesian model checking for ecologists. Ecological Monographs, 88:526–542.

Costa, D. P., Robinson, P. W., Arnould, J. P. Y., Harrison, A. L., Simmons, S. E., Hassrick, J. L., Hoskins, A. J., Kirkman, S. P., Oosthuizen, H., Villegas-Amtmann, S., and Crocker, D. E. (2010). Accuracy of ARGOS locations of pinnipeds at-sea estimated using fastloc GPS. PLoS ONE, 5:e8677.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91:883–904.

Cressie, N., Calder, C. A., Clark, J. S., Ver Hoef, J. M., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications, 19:553–570.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. Trends in Ecology and Evolution, 25:410–418.

de Jong, P. (1988). A cross-validation filter for time series models. Biometrika, 75:594–600.

de Valpine, P. (2004). Monte carlo state-space likelihoods by weighted posterior kernel density estimation. Journal of the American Statistical Association, 99:523–536.

de Valpine, P. and Hastings, A. (2002). Fitting population models incorporating process noise and observation error. Ecological Monographs, 72:57–76.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. Journal of Computational and Graphical Statistics, 26:403–413.

Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L., and Staples, D. F. (2006). Estimating density dependence, process noise, and observation error. Ecological Monographs, 76:323–341.

Dennis, B., Ponciano, J. M., and Taper, M. L. (2010). Replicated sampling increases efficiency in monitoring biological populations. Ecology, 91:610–620.

Dennis, B. and Taper, M. L. (1994). Density dependence in time series observations of natural populations: Estimation and testing. Ecological Monographs, 64:205–224.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Wood, S. N., Wüest, R. O., and Hartig, F. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs, 88:485–504.

Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential monte carlo methods. In Sequential Monte Carlo methods in practice, pages 3–14. Springer.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google flu trends data and a state-space SEIR model. Journal of the American Statistical Association, 107:1410–1426.

Dunham, K. and Grand, J. B. (2016). Effects of model complexity and priors on estimation using sequential importance sampling/resampling for species conservation. Ecological Modelling, 340:28–36.

Durbin, J. and Koopman, S. J. (2012). Time series analysis by state space methods. Oxford University Press, Oxford, UK, 2nd edition.

Fasiolo, M., Pya, N., and Wood, S. N. (2016). A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. Statistical science, 31:96–118.

Fasiolo, M. and Wood, S. N. (2018). ABC in Ecological Modelling. In Sisson, S. A., Fan, Y., and Beaumont, M., editors, Handbook of approximate Bayesian computation, pages 597–622. Chapman and Hall/CRC, New York, NY.

Ferretti, F., Curnick, D., Liu, K., Romanov, E. V., and Block, B. A. (2018). Shark baselines and the conservation role of remote coral reef ecosystems. Science Advances, 4:eaaq033.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software, 27:233–249.

Frühwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. Environmental and Ecological Statistics, 3:291–309.

Gamerman, D., Rezende dos Santos, T., and Franco, G. C. (2013). A non-Gaussian family of state-space models with exact marginal likelihood. Journal of Time Series Analysis, 34:625–645.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. Biometrics, 56:1055–1067.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. Journal of the American Statistical Association, 94:247–253.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis. CRC Press, Boca Raton, FL, 3rd edition.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24:997–1016.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7:457–511.

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, Handbook of Markov chain Monte Carlo, chapter 6, pages 163–174. Chapman and Hall/CRC, New York, NY.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. Entropy, 19:555.

Geyer, C. (2011). Introduction to Markov Chain Monte Carlo. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, Handbook of Markov Chain Monte Carlo, chapter 1, pages 3–48. Chapman and Hall/CRC, New York, NY.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). Markov chain Monte Carlo in practice. Chapman and Hall/CRC, New York, NY.

Gimenez, O., Bonner, S., King, R., A. Parker, R., Brooks, S., Jamieson, L., Grosbois, V., Morgan, B., and Thomas, L. (2009a). Winbugs for population ecologists: Bayesian modeling using markov chain monte carlo methods. Environmental and Ecological Statistics, 3:883–915.

Gimenez, O., Morgan, B. J. T., and Brooks, S. P. (2009b). Weak identifiability in models for mark-recapture-recovery data. In Thomson, D. L., Cooch, E. G., and Conroy, M. J., editors, Modeling demographic processes in marked populations, pages 1055–1067. Environmental and Ecological Statistics Series.

Gimenez, O., Viallefont, A., Catchpole, E. A., Choquet, R., and Morgan, B. J. T. (2004). Methods for investigating parameter redundancy. Animal Biodiversity and Conservation, 27:1–12.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. IEE Proceedings F - Radar and Signal Processing, 140:107–113.

Goudie, R. J. B., Turner, R. M., De Angelis, D., and Thomas, A. (2017). MultiBUGS: A parallel implementation of the BUGS modelling framework for faster Bayesian inference. arXiv e-prints, page arXiv:1704.03216.

Grewal, M. S. and Andrews, A. P. (2010). Applications of Kalman filtering in aerospace 1960 to the present. IEEE Control Systems Magazine, 30:69–78.

Harvey, A. C. (1990). Forecasting, structural time series models and the Kalman filter. Cambridge university press, Cambridge, UK.

Hobbs, N. T., Geremia, C., Treanor, J., Wallen, R., White, P. J., Hooten, M. B., and Rhyan, J. C. (2015). State-space modeling to support management of brucellosis in the Yellowstone bison population. Ecological Monographs, 85:525–556.

Holdo, R. M., Sinclair, A. R. E., Dobson, A. P., Metzger, K. L., Bolker, B. M., Ritchie, M. E., and Holt, R. D. (2009). A disease-mediated trophic cascade in the Serengeti and its implications for ecosystem C. PLoS Biology, 7:e1000210.

Holmes, E., Ward, E., Scheuerell, M., and Wills, K. (2018). MARSS: Multivariate Autoregressive State-Space Modeling. R package version 3.10.10.

Holmes, E. E., Ward, E. J., and Wills, K. (2012). MARSS: Multivariate autoregressive state-space models for analyzing time-series data. The R Journal, 4:11–19.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to bayesian model selection for ecologists. Ecological Monographs, 85:3–28.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. Proceedings of the National Academy of Sciences, 112:719–724.

Jamieson, L. E. and Brooks, S. P. (2004). Density dependence in North American ducks. Animal Biodiversity and Conservation, 27.1:113–128.

Johnson, D. S. and London, J. M. (2018). crawl: an r package for fitting continuous-time correlated random walk models to animal movement data.

Johnson, D. S., London, J. M., Lea, M.-A., and Durban, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. Ecology, 89:1208–1215.

Jonsen, I. (2016). Joint estimation over multiple individuals improves behavioural state inference from animal movement data. Scientific Reports, 6:20625.

Jonsen, I., Basson, M., Bestley, S., Bravington, M., Patterson, T., Pedersen, M., Thomson, R., Thygesen, U., and Wotherspoon, S. (2013). State-space models for bio-loggers: a methodological road map. Deep Sea Research Part II: Topical Studies in Oceanography, 88-89:34–46.

Jonsen, I. D., Mills Flemming, J., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. Ecology, 86:2874–2880.

Kai, M. and Yokoi, H. (2019). Performance evaluation of information criteria for estimating a shape parameter in a Bayesian state-space biomass dynamics model. Fisheries Research, 219:105326.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of basic Engineering, 82:35–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. Journal of basic engineering, 83:95–108.

Karban, R. and de Valpine, P. (2010). Population dynamics of an Arctiid caterpillar-tachinid parasitoid system using state-space models. Journal of Animal Ecology, 79:650–661.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules stable. Journal of the American Statistical Association, 91:1343–1370.

Kindsvater, H. K., Dulvy, N. K., Horswill, C., Juan-Jordà, M.-J., Mangel, M., and Matthiopoulos, J. (2018). Overcoming the data crisis in biodiversity conservation. Trends in Ecology & Evolution, 33:676–688.

King, A. A., Domenech de Cellès, M., Magpantay, F. M. G., and Rohani, P. (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proceedings of the Royal Society of London, Series B, 282:20150347.

King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S., Kendall, B., Wearing, H., Ferrari, M. J., Lavine, M., and Reuman, D. C. (2009). pomp: Statistical inference for partially observed markov processes. URL http://pomp. r-forge. r-rproject. org.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. Journal of the American Statistical Association, 82:1032–1041.

Knape, J. (2008). Estimability of density dependence in models of time series data. Ecology, 89:2994–3000.

Knape, J., Jonzén, N., and Sköld, M. (2011). On observation distributions for state space models of population survey data. The Journal of animal ecology, 80:1269–1277.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. (2016). TMB: automatic differentiation and Laplace approximation. Journal of Statistical Software, 70:1–21.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. Ecology, 93:2336–2342.

Lele, S. R. (2015). Is non-informative Bayesian analysis appropriate for wildlife management: survival of San Joaquin Kit Fox and declines in amphibian populations. arXiv e-prints, page arXiv:1502.00483.

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. Journal of the American Statistical Association, 10:1617–1625.

Link, W. A. and Sauer, J. R. (2016). Bayesian cross-validation for model evaluation and selection, with application to the North American Breeding Bird Survey. Ecology, 97:1746–1758.

Link, W. A., Sauer, J. R., and Niven, D. K. (2017). Model selection for the North American Breeding Bird Survey: a comparison of methods. The Condor, 119:546–556.

Liu, J. S. and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. Journal of the American statistical association, 93:1032–1044.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). The BUGS book. Chapman & Hall/CRC, Boca Raton, FL.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: evolution, critique and future directions. Statistics in medicine, 28:3049–3067.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325–337.

Maunder, M. N. and Deriso, R. B. (2011). A statespace multistage life cycle model to evaluate population impacts in the presence of density dependence: illustrated with application to delta smelt (*Hyposmesus transpacificus*). Canadian Journal of Fisheries and Aquatic Sciences, 68:1285–1306.

McClintock, B. T., Johnson, D. S., Hooten, M. B., Ver Hoef, J. M., and Morales, J. M. (2014). When to be discrete: the importance of time formulation in understanding animal movement. Movement Ecology, 2:21.

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., and Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. Ecological Monographs, 82:335–349.

Mendelssohn, R. (1988). Some problems in estimating population sizes from catch-at-age data. Fishery Bulletin, 86:617–630.

Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian Comparison of latent variable models: conditional versus marginal likelihoods. Psychometrika, 84:802–829.

Meyer, R. and Millar, R. B. (1999). Bugs in bayesian stock assessments. Canadian Journal of Fisheries and Aquatic Sciences, 56:1078–1087.

Michielsens, C. G. J., McAllister, M. K., Kuikka, S., Pakarinen, T., Karlsson, L., Romakkaniemi, A., Perä, I., and Mäntyniemi, S. (2006). A Bayesian state-space mark-recapture model to estimate exploitation rates in mixed-stock fisheries. Canadian Journal of Fisheries and Aquatic Sciences, 63:321–334.

Millar, R. B. (2002). Reference priors for Bayesian fisheries models. Canadian Journal of Fisheries and Aquatic Sciences, 59:1492–1502.

Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. Statistics and Computing, 28:375–385.

Mills Flemming, J., Jonsen, I. D., Myers, R. A., and Field, C. A. (2010). Hierarchical state-space estimation of leatherback turtle navigation ability. PLoS ONE, 5:e14245.

Monnahan, C. C., Thorson, J. T., and Branch, T. A. (2017). Faster estimation of bayesian models in ecology using hamiltonian monte carlo. Methods in Ecology and Evolution, 8:339–348.

Moore, J. E. and Barlow, J. (2011). Bayesian state-space model of fin whale abundance trends from a 1991-2008 time series of line-transect surveys in the California Current. Journal of Applied Ecology, 48:1195–1205.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. Statistical Science, 28:135–167.

Neal, R. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X., editors, Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, Boca Raton, FL.

Newman, K. B., Buckland, S. T., Morgan, B. J. T., King, R., Borchers, D. L., Cole, D. J., Besbeas, P., Gimenez, O., and Thomas, L. (2014). Modelling population dynamics: model formulation, fitting and assessment using state-space methods. Springer, New York, NY.

Nielsen, A. and Berg, C. W. (2014). Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research, 158:96–101.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. Bayesian analysis, 4:85–117.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State-space models of individual animal movement. Trends in Ecology and Evolution, 23:87–94.

Peacock, S. J., Krkošek, M., Lewis, M. A., and Lele, S. (2016). Study design and parameter estimability for spatial and temporal ecological models. Ecology and Evolution, 7:1–9.

Pedersen, M. W., Berg, C. W., Thygesen, U. H., Nielsen, A., and Madsen, H. (2011). Estimation methods for nonlinear state-space models in ecology. Ecological Modelling, 222:1394–1400.

Petris, G. (2010). An R package for Dynamic Linear Models. Journal of Statistical Software, 36:1–16.

Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic Linear Models with R. Springer.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filters. Journal of the American Statistical Association, 94:590–599.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing, 124:1–10.

Plummer, M. (2018). rjags: Bayesian Graphical Models using MCMC. R package version 4-8.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for MCMC. R News, 6:7–11.

Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. Journal of Agricultural, Biological, and Environmental Statistics, 22:270–293.

Polansky, L., Newman, K. B., and Mitchell, L. (2019). Improving inference for nonlinear state-space models of animal population dynamics given biased sequential life stage data. arXiv e-prints, page arXiv:1909.09111.

Postlethwaite, C. M. and Dennis, T. E. (2013). Effects of temporal resolution on an inferential model of animal movement. PLoS ONE, 8:e57640.

Prado, R. and West, M. (2010). Time series: modeling, computation, and inference. Chapman & Hall/CRC, Boca Raton, FL.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25:1923–1929.

Robert, C. P. (2007). The Bayesian choice: from decision theoretic foundations to computation implementation. Springer Sciences + Business Media, LLC, New York, NY, 2nd edition.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-arroita, G., Hauenstein, S., Lahoz-monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical ,or phylogenetic structure. Ecography, 49:913–929.

Rothenberg, T. J. (1971). Identification in parametric models. Econometrica, 39:577–591.

Sanderlin, J. S., Block, W. M., Strohmeyer, B. E., Saab, V. A., and Ganey, J. L. (2019). Precision gain versus effort with joint models using detection/non-detection and banding data. Ecology and Evolution, 9:804–817.

Schick, R. S., Kraus, S. D., Rolland, R. M., Knowlton, A. R., Hamilton, P. K., Pettis, H. M., Kenney, R. D., and Clark, J. S. (2013). Using hierarchical Bayes to understand movement, health, and survival in the endangered North Atlantic right whale. PLoS ONE, 8:e64166.

Shumway, R. H. and Stoffer, D. S. (2016). Time series analysis and its applications: with R examples. Springer International Publishing, New York, NY, 4th edition.

Siple, M. C. and Francis, T. B. (2016). Population diversity in Pacific herring of the Puget Sound, USA. Oecologia, 180:111–125.

Smith, J. (1985). Diagnostic checks of non-standard time series models. Journal of Forecasting, 4:283–291.

Stan Development Team (2012). Stan: A C++ library for probability and sampling. http://mc-stan.org.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.

Sullivan, P. J. (1992). A Kalman filter approach to catch-at-length analysis. Biometrics, 48:237–257.

Thygesen, U. H., Albertsen, C. M., Berg, C. W., Kasper, K., and Nielsen, A. (2017). Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Statistics, 24:317–339.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. Biometrika, 92:351–370.

VanDerwerken, D. and Schmidler, S. C. (2017). Monitoring joint convergence of MCMC samplers. Journal of Computational and Graphical Statistics, 26:558–568.

Vats, D. and Knudson, C. (2018). Revisiting the Gelman-Rubin diagnostic. arXiv e-prints, page arXiv:1812.09384.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27:1413–1432.

Viallefont, A., Lebreton, J.-D., and Reboulet, A.-M. (1998). Parameter identifiability and model selection in capture-recapture models: A numerical approach. Biometrical Journal, 40:313–325.

Viljugrein, H., Stenseth, N. C., Smith, G. W., and Steinbakk, G. H. (2005). Density dependence in North American ducks. Ecology, 86:245–254.

Ward, E. J., Chirakkal, H., González-Suárez, M., Aurioles-Gamboa, D., Holmes, E. E., and 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.

Warton, D. I., Thibault, L., and Wang, Y. A. (2017). The PIT-trap - a "model-free" bootstrap procedure for inference about regression models with discrete, multivariate responses. PLoS ONE, 12:e0181790.

Wilberg, M. J. and Bence, J. R. (2008). Performance of deviance information criterion model selection in statistical catch-at-age analysis. Fisheries Research, 93:212–221.

Winship, A. J., Jorgensen, S. J., Shaffer, S. A., Jonsen, I. D., Robinson, P. W., Costa, D. P., and Block, B. A. (2012). State-space framework for estimating measurement error from double-tagging telemetry experiments. Methods in Ecology and Evolution, 3:291–302.

Wintle, B. A., McCarthy, M. A., Volinsky, C. T., and Kavanagh, R. P. (2003). The use of Bayesian model averaging to better represent uncertainty in ecological models. Conservation Biology, 17:1579–1590.

Yin, Y., Aeberhard, W. H., Smith, S. J., and Mills Flemming, J. (2019). Identifiable state-space models: a case study of the Bay of Fundy sea scallop fishery. Canadian Journal of Statistics, 47:27–45.

Zucchini, W., Macdonald, I. L., and Langrock, R. (2016). Hidden Markov models for time series: an introduction using R. Taylor & Francis, Boca Raton, FL, 2nd edition.