Environmental and Ecological Statistics manuscript No.
(will be inserted by the editor)

Validation of ecological state space models using the Laplace approximation

Uffe Høgsbro Thygesen ·
Christoffer Moesgaard Albertsen ·
Casper Willestofte Berg ·
Kasper Kristensen · **Anders Nielsen**

Received: date / Accepted: date

Abstract Many statistical models in ecology follow the state space paradigm. For such models, the important step of model validation rarely receives as much attention as estimation or hypothesis testing, perhaps due to lack of available algorithms and software. Model validation is often based on a naive adaptation of Pearson residuals, i.e. the difference between observations and posterior means, even if this approach is flawed. Here, we consider validation

Uffe Høgsbro Thygesen at National Institute of Aquatic Resources
Technical University of Denmark
2920 Charlottenlund
Denmark
E-mail: uht@aqua.dtu.dk
· Christoffer Moesgaard Albertsen at National Institute of Aquatic Resources
Technical University of Denmark
2920 Charlottenlund
Denmark
E-mail: cmoe@aqua.dtu.dk
· Casper Willestofte Berg at National Institute of Aquatic Resources
Technical University of Denmark
2920 Charlottenlund
Denmark
E-mail: cbe@aqua.dtu.dk
· Kasper Kristensen at National Institute of Aquatic Resources
Technical University of Denmark
2920 Charlottenlund
Denmark
E-mail: kaskr@aqua.dtu.dk
· Anders Nielsen at National Institute of Aquatic Resources
Technical University of Denmark
2920 Charlottenlund
Denmark
E-mail: an@aqua.dtu.dk

of state space models through one-step prediction errors, and discuss principles and practicalities arising when the model has been fitted with a tool for estimation in general mixed effects models. Implementing one-step predictions in the R package Template Model Builder (TMB), we demonstrate that it is possible to perform model validation with little effort, even if the ecological model is multivariate, has non-linear dynamics, and whether observations are continuous or discrete. With both simulated data, and a real data set related to geolocation of seals, we demonstrate both the potential and the limitations of the techniques. Our results fill a need for convenient methods for validating a state space model, or alternatively, rejecting it while indicating useful directions in which the model could be improved.

Keywords Statistical ecology · time series analysis · state space methods · maximum likelihood estimation · model validation · residual analysis

1 Introduction

The state space paradigm is of central importance in ecological modelling, as well as in other fields of science. In the context of time series analysis, state space methods are predominant in individual movement and behaviour (Patterson et al, 2008; Jonsen et al, 2013) and are gaining popularity in fisheries assessment models (Nielsen and Berg, 2014) as well as general population dynamics (de Valpine and Hastings, 2002; Clark, 2007). A main advantage of the state space approach is that understanding of ecological dynamics is mirrored in the state space structure of the model, while unknown parameters are estimated with rigorous statistical methods. This coupling of inductive and deductive modelling permits models with higher fidelity and predictive power than e.g. purely correlative descriptions, but also requires computational tools beyond the suite of standard statistical models. An additional appeal of the state space framework is that it is not limited to time series analysis, but also pivotal in qualitative and quantitative analysis of dynamic ecosystem models (Murray, 1989) as well as in dynamic optimisation models in behavioural ecology (Clark and Mangel, 2000; Thygesen et al, 2016).

The Kalman filter (Kalman, 1960; Harvey, 1989) is the classical technique for analysing time series using state space models. Originally applicable to linear systems, which have limited application in ecology, the Extended Kalman Filter and its many variants (Evensen, 2003; Wan and Van Der Merwe, 2000) deal with weakly nonlinear dynamics. More recently, Monte Carlo-based methods have gained popularity for inference in state space models; these include the particle filter (Liu and Chen, 1998) as well as Markov chain Monte Carlo methods such as Metropolis-Hastings or Gibbs sampling (Gilks et al, 1996; Jonsen et al, 2005).

While Monte Carlo algorithms apply to very general classes of models, they entail a significant computational burden, even for problems of seemingly moderate complexity (Pedersen et al, 2011; Bolker et al, 2013). Further, it is a non-trivial task to tune the algorithms and even determine if the runs have reached convergence. A viable alternative to Monte Carlo entails direct numerical optimisation of the likelihood function, using the Laplace approximation (Tierney and Kadane, 1986) to integrate out unobserved random variables (Skaug and Fournier, 2006). The appearance of computational tools such INLA (Rue et al, 2009), AD Model Builder (Fournier et al, 2012) and, recently, Template Model Builder (Kristensen et al, 2016) have been pivotal in this development. These tools have also been used for inference in ecological state space models, e.g. (Cadigan et al, 2014; Nielsen and Berg, 2014; Albertsen et al, 2015).

Despite the increasing popularity of state space models of ecological time series, few studies consider the issue of model validation. Validation is an important part of the procedure in all areas of statistical modeling: After a model structure has been conceived and parameters have been estimated, it is crucial to validate if this model describes the data adequately, assessing if it is plausible that the data were generated by a system identical to the model. We emphasize that the ambition is not to prove the model "right", which would be a futile exercise, but merely to check that the model cannot be falsified with the available data. Also, we emphasize that the question is not if the model is useful for a given purpose, but merely if it agrees with data.

Model validation thereby serves a purpose complementary to hypothesis testing and model selection: While estimation and model selection searches for the most suitable model within a specified family, and while hypothesis testing examines if the model structure can be reduced, model validation examines if the model family should be modified or extended.

Model validation typically relies on inspecting model residuals, which can be defined in a multitude of ways. In the simplest case of linear regression models, response residuals are defined as "observed values minus fitted values" (Anscombe and Tukey, 1963), while *Pearson residuals* are rescaled to have unit variance under the model assumptions. The model validation proceeds by inspecting and testing for patterns in the residuals, such as outliers, correlations with extra covariates, nonlinearities, correlations between the residuals, heteroscedasticity, and departure from Gaussianity (Cox and Snell, 1968). For linear dynamic models of time series (Box and Jenkins, 1970; Harvey, 1989), the fitted values are replaced with one-step predictions, leading to forecast residuals. For nonlinear and non-Gaussian models, the so-called quantile residuals are computed by evaluating the cumulated distribution function of the observation in the observed value (Dunn and Smyth, 1996). For time series, the cumulated distribution function is conditional on past observations (Smith, 1985), so that the computation of residuals agrees with recursive filtering techniques. These residuals are, under the model assumptions, independent and uniformly distributed on $[0, 1]$ (Rosenblatt, 1952) but are often

transformed to the Gaussian scale (Smith, 1985). Observations with discrete distributions require special attention to ensure that the residuals are continuously distributed; the typical approach is to add a random perturbation to the evaluated cumulated distribution function (Smith, 1985). In the previous, we have assumed that parameters in the model are known, in which case the quantile residuals are termed *theoretical* by Dunn and Smyth (1996) in contrast to *observed* quantile residuals when they are based on estimated model parameters. For observed quantile residuals, significance tests require corrections which take the parameter estimation error into account; asymptotic results are given by Kalliovirta (2012). The textbooks by Ljung (1999) and Madsen (2007) summarize techniques for validating time series models based on residuals.

The practical applicability of quantile residuals requires computational strategies which remain a hurdle except for linear systems and for Hidden Markov Models with finite state spaces (Zucchini and MacDonald, 2009). For this reason, attention has been given to systems of specific structures which make computations feasible. For example, Frühwirth-Schnatter (1996) considered linear system dynamics with a scalar linear predictor, conditional on which the observations may have aribtrary non-Gaussian distributions. Due to the computational complexity and shortage of available implementations, model validation based on residuals is rarely done outside such limited model classes.

The contribution of the present work is to present an algorithm for computing forecast quantile residuals in mixed-effect models using the Laplace approximation. This algorithm has been implemented in the Template Model Builder framework (TMB[1], (Kristensen et al, 2016)). In this framework, which applies to general non-linear mixed-effects models in which the unobserved random variables have continuous distributions, the modeller specifies the joint density of all observed and unobserved random variables in the model. Thereafter, built-in numerical algorithms integrate out unobserved random variables using the Laplace approximation, maximise the likelihood function, and thus provide facilities for statistical inference, i.e. parameter estimates, confidence regions, test statistics, etc. The contribution of the present work is to provide algorithms for computation of prediction residuals for general nonlinear mixed-effects model. The algorithms are implemented in TMB and thus available to the modeller with minimal coding effort, i.e. if a model has already been implemented to allow maximum likelihood estimation, prediction residuals can now be obtained with a single line of R code in the simplest situations. We demonstrate the technique with examples involving simulated time series, both linear and nonlinear, scalar and multivariate, and with continuous and discrete observations. Finally we demonstrate a case involving geolocation of seals, i.e., estimation of their position in space based on electronic tags. We

[1] TMB is an R package (R Core Team, 2015) available both at the Comprehensive R Archive Network (cran.r-project.org) and in a development version at GitHub (github.com/kaskr/adcomp)

discuss advantages and limitations to the approach. The source code for the simulation examples presented here is available as part of the TMB distribution; this allows the reader to reproduce the results presented here and modify the code to apply it to his or her own model.

2 State space models as mixed effect models

We consider discrete-time stochastic state space models with a sequence $(X_1,\dots,X_N)$ of states and associated measurements $(Y_1,\dots,Y_N)$. We focus on continuous state spaces, $X_i \in \mathbf{X} = \mathbf{R}^n$. The observations Y_i are real-valued m-vectors. The model is specified through:

1. The distribution of the initial state X_1, i.e. the p.d.f. $f_1(x;\theta)$.
2. Transition densities $f_X(x|x';\theta)$, i.e. the p.d.f. of X_{i+1} at x given $X_i = x'$.
3. Measurement densities $f_Y(y|x;\theta)$, i.e. the p.d.f. of Y_i at y given $X_i = x$.

Here, θ is a vector of model parameters. The observations may be continuously or discretely distributed, i.e. the p.d.f. f_Y is w.r.t. a reference measure μ which is either the Lebesgue measure or the counting measure on the integers, while the states are continuously distributed so that the p.d.f.'s f_1 and f_X are w.r.t. Lebesgue measure. The densities f_X and f_Y can be arbitrary, but we do require that $f_X(x|x';\theta)$ is twice differentiable in (x,x'), and that $f_Y(y|x;\theta)$ is twice differentiable in x.

We assume that the state process $\{X_i : i = 1,\dots,N\}$ has the Markov property w.r.t. its own filtration and that each measurement Y_i is conditionally independent of measurements Y_j and states X_j at any time point $j \neq i$. Thus, the joint density $f_{\bar{X}\bar{Y}}(\bar{x}|\bar{y};\theta)$ of all states and measurements is

$$f_{\bar{X}\bar{Y}}(\bar{x},\bar{y};\theta) = f_1(x_1)\cdot\left(\prod_{i=1}^{N-1} f_X(x_{i+1}|x_i;\theta)\right)\cdot\left(\prod_{i=1}^{N} f_Y(y_i|x_i;\theta)\right) \tag{1}$$

where $\bar{x} = (x_1,\dots,x_N)$ and $\bar{y} = (y_1,\dots,y_N)$. To estimate the parameters θ from observed measurements $\bar{y}$, the states X_i are unobserved random variables (a.k.a. latent variables or random effects), which must be integrated out in order to obtain the likelihood function:

$$L(\theta;\bar{y}) = \int_{\mathbf{X}^N} f_{\bar{X}\bar{Y}}(\bar{x},\bar{y};\theta)\; d\bar{x} \tag{2}$$

Our approach is based on the framework in Template Model Builder (TMB) (Kristensen et al, 2016). Here, the modeller supplies a code which evaluates the

joint density (1). Then, TMB computes the integral in (2) using the Laplace approximation (Tierney and Kadane, 1986), applying automatic differentiation (Rall, 1980; Griewank and Walther, 2008) to obtain the required derivatives. It is this step that requires that the densities are smooth functions of the state. The approach allows the likelihood function to be maximized numerically using standard quasi-Newton methods, where the needed derivatives are still obtained automatically. This approach seperates model specification from computational model analysis, which is a convenience to the modeller, and results in high computational efficiency (Kristensen et al, 2016).

3 A toy example: Random walk with an unmodelled drift

We consider first a simple example to demonstrate the use of prediction residuals in model validation, and in particular the flaw of a naive adaptation of Pearson residuals. Consider therefore the scalar ($n = m = 1$) random walk

$$X_{i+1} = X_i + \mu + E_i, \quad i = 1, \ldots N-1, \tag{3}$$

where $X_1 = 0$, μ is a constant drift term, and the E_i are Gaussian random variables distributed as $N(0, \sigma^2)$. This specifies the transition densities:

$$f_X(x|x';\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(x-x'-\mu)^2}{\sigma^2}\right) \quad .$$

Here, $\theta = (\mu, \sigma^2, s^2)$ is the vector of system parameters. We measure the states X_i with Gaussian measurement errors $W_i \sim N(0, s^2)$, obtaining

$$Y_i = X_i + W_i, \quad i = 1, \ldots, N, \tag{4}$$

which specifies the measurement densities

$$f_Y(y|x;\theta) = \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{1}{2}\frac{(y-x)^2}{s^2}\right) \quad .$$

All E_i and W_i are independent, so that this model is of the general structure described in section 2.

We simulate states and measurements with $N = 100$, $\mu = 0.75$, $\sigma = 1$, $s = 1$. Based on the measurements, we estimate states and parameters $\theta = (\mu, \sigma^2, s^2)$ using TMB; see figure 1. To recapitulate, TMB uses the Laplace approximation

to integrate out latent random variables, but in this case the model is linear and Gaussian so Laplace's formula involves no approximation error.

We envision a situation where we first conceive a model with no drift, and thus estimate parameters under the hypothesis $H_0 : \mu = 0$. We aim to validate or falsify the model M_0 fitted under H_0 without comparing to more complex models, using what is termed *pure significance tests* by Cox and Hinkley (1974). We stress that this model validation cannot compare model fit with more complex alternatives; any realistic modelling study includes a large number of simplifying assumptions, and it is not practical to analyse more complex alternatives where all these assumptions are relaxed. Thus, we expect that the model M_0 is falsified during the model validation step, and the alternative hypothesis $H_1 : \mu \neq 0$ is suggested. Next, we expect that the model M_1 fitted under the alternative hypothesis H_1 actually passes the validation step.

Figure 1 shows the true states $\{X_i\}$, the measurements $\{Y_i\}$, as well as posterior means

$$\hat{X}_i = \mathbf{E}^M\{X_i|\bar{Y}\} \ .$$

Here, on the right hand side, $\mathbf{E}^M$ means expectation assuming the model M, and we condition on all data $\bar{Y} = (Y_1, \ldots, Y_N)$. The figure includes both models M_0 and M_1, but at this stage we focus on M_0.

In order to assess model performance, a common practice (Nielsen and Berg, 2014; Cadigan et al, 2014) is to assess model fidelity by inspecting residuals. Too commonly, these residuals are computed naively as the response residuals $\tilde{Y}_i$, i.e. the difference between observed values and fitted values (posterior means):

$$\tilde{Y}_i = Y_i - \hat{X}_i \ . \tag{5}$$

These response residuals $\tilde{Y}_i$ are shown in figure 1 (top right panel) for both models. It is important to realize that they are unsuitable for model validation: They do not suggest that the model M_0 should be rejected while M_1 should be accepted. For both models, they appear equally unbiased; a correct drift term is not required to smoothen the observations. Moreover, the grossly incorrect model M_0 leads to residuals which are smaller in magnitude than those from the "correct" model M_1. The explanation for this seeming paradox is that the model M_0, not containing a bias term, has an exaggerated estimate of the process noise σ^2. It therefore adds too much credibility to the specific measurement Y_i, so that the residual $\tilde{Y}_i$ is too small; effectively overfitting $\hat{X}_i$. Thus, small response residuals $\tilde{Y}_i$ is not an indication of good model fit. Furthermore, the residuals $\tilde{Y}_i$ are not uncorrelated, not even when computed with the correct model. The fact that response residuals $\tilde{Y}_i$ are unsuitable for model validation

is well known within time series analysis, but evidently not widely appreciated by practitioning modellers everywhere.

To avoid this issue, classical methods for linear time series analysis (Box and Jenkins, 1970; Harvey, 1989; Madsen, 2007) focus on the prediction errors:

$$\tilde{Y}_{i|i-1} = Y_i - \hat{Y}_{i|i-1} \text{ where } \hat{Y}_{i|i-1} = \mathbf{E}^M\{Y_i|Y_1^{i-1}\} \tag{6}$$

Here, the subscript $i|i-1$ means that we predict the value Y_i based on information available at time $i-1$, i.e. on $Y_1^{i-1} = (Y_1, \ldots, Y_{i-1})$. The pivotal role of prediction errors is obvious when analysing time series with recursive filtering techniques, but is reduced when viewing this as a general mixed effect model. Technically, this central role stems from the likelihood function being written in terms of the prediction errors (Madsen, 2007). From a philosophical standpoint, one can reasonably argue that a mathematical model is best assesed by its ability to predict the outcome of future experiments (Ljung, 1999).

Here, we compute prediction residuals using TMB (computational details are given in the following) and transform them so they would be standard Gaussian if the true system equals the fitted model. Here, this amounts to computing Pearson residuals by rescaling with their standard deviation; for a general nonlinear model this is obtained by the prediction quantile residuals (Smith, 1985)

$$Z_i = \Phi^{-1}(U_i) \text{ where } U_i = \mathbf{P}^M(Y_i \leq y_i|Y_1^{i-1} = y_1^{i-1}) \tag{7}$$

and where Φ is the standard Gaussian cumulated distribution function. The terminology in the literature is somewhat inconsistent; we shall refer to these Z_i simply as prediction residuals. They are seen in figure 1. If M was the true system, these Z_i should be independent standard Gaussians (Smith, 1985). The validation proceeds by finding patterns in the residuals that are unlikely under this hypothesis. Visual inspection of residuals is not always sufficient, since a pattern may be obscured by noise; rather patterns are identified from simple summary statistics of the residuals, measuring bias, trends, correlation, heavy tails or outliers (Cox and Snell, 1968; Ljung, 1999). The residuals are best first viewed in isolation, but a complete validation would also include inputs, states and possible auxiliary time series: Under the true model, no time series which is independent of the state process $\{X_i\}$ should contain information about the prediction residuals.

In this example, the simple model M_0 generally under-predicts the next measurement, and the bias in the prediction errors is visually clear (figure 1, lower left panel, red dots and histogram). Testing if mean is 0 with a t-test, we find a critical p-value of $8 \cdot 10^{-7}$. This falsifies the model M_0 and suggests the remedy, namely to include a drift term in the process equation, i.e. the alternative

hypothesis H_1. For the model M_1, this bias is not present; in fact, the mean of the residuals is exactly 0 due to the bias being estimated. Thus, the model M_1 passes this step in the validation.

This example is available in the TMB distribution as `randomwalkvalidation.`

4 Computation of prediction residuals

When the time series analysis is done using recursive filtering, the prediction residuals may be obtained as part of the recursion; see for example (Frühwirth-Schnatter, 1996) for models with linear-Gaussian dynamics and scalar predictors. Conversely, when using computational methods for general nonlinear mixed effects models, the residuals require substantial extra computation. We now describe the computational approach, which we implemented in TMB. First, we rewrite the residuals U_i from (7) as

$$U_i = \frac{\mathbf{P}^M(Y_i \le y_i, Y_1^{i-1} = y_1^{i-1})}{\mathbf{P}^M(Y_1^{i-1} = y_1^{i-1})} = \frac{\int_{(-\infty,y_i]} f_i(y)\ d\mu(y)}{\int_{(-\infty,y_i]} f_i(y)\ d\mu(y) + \int_{(y_i,+\infty)} f_i(y)\ d\mu(y)} \tag{8}$$

where

$$f_i(y) = \int_{\mathbf{X}^N} f_{\bar{X}Y_1^i}(\bar{x}, [y_1^{i-1}, y]; \theta)\ d\bar{x} \tag{9}$$

Here, $f_{\bar{X}Y_1^i}$ is the joint density of all states $\bar{X} = (X_1, \ldots, X_N)$ and observations up to time i, $Y_1^i = (Y_1, \ldots, Y_i)$:

$$f_{\bar{X}Y_1^i}(\bar{x}, y_i^i; \theta) = f_1(x_1) \cdot \left(\prod_{i=1}^{N-1} f_X(x_{i+1}, x_i; \theta) \right) \cdot \left(\prod_{j=1}^{i} f_Y(y_j, x_j; \theta) \right) \tag{10}$$

while $[y_1^{i-1}, y]$ is the reduced data set where data points $y_1, \ldots, y_{i-1}$ are kept, the data point y_i is replaced with the integration variable y, and all future data points $y_{i+1,\ldots,N}$ are excluded. These joint densities may be obtained from the full joint density (1) simply by including the appropriate terms, so from the modeller's perspective, the only extra model specification required is the order in which the data points should be processed.

The next step is to approximate the integral over the states $\bar{x}$ in (9) using the Laplace approximation. This involves first maximizing the logarithm of the integrand (10) w.r.t. $\bar{x}$ and next determining the curvature at the maximum point; for both of these operations, automatic differentiation of the logarithm

of the integrand (10) is employed to yield the required derivatives. Finally, the line integals over y in (8) are computed by direct numerical integration.

We note that one could conceive several different numerical approaches to the computation of residuals based on Laplace approximations, and careful consideration to approximation errors must be given so that the patterns in the residuals can confidently be attributed to model misspecification and not to artifacts of the numerical methods. Writing the residuals as ratio of integrals, as opposed to an integral of conditional probabilities, in general has the advantage that leading error terms in the Laplace approximation cancel (Tierney and Kadane, 1986), yielding higher accuracy. To compare different candidate approaches, we performed simulation experiments where data were generated from simulation models, prediction residuals were computed based on the same models, and the residuals were tested for patterns. The method described in this section was superior to all other approaches we implemented. Still, our approach hinges on the assumed validity of the Laplace approximation. We will return to this issue in the discussion.

The previous describes the generic method in TMB for computing the residuals for general non-linear models. In the important special case of a fully Gaussian model, the computations may be greatly simplified. First, the joint precision matrix of states $\bar{X}$ and observations $\bar{Y}$ is found from the double derivatives of the logarithm of the joint density (1). Next, the covariance matrix of observations $\bar{Y}$ is determined from this joint precision by marginalization, using standard results for multivariate Gaussian distributions. Next, a Cholesky factorizaton of this covariance matrix yields the one-step predictors and the prediction variance, from which the standardized prediction residuals can be found.

We note that our computational approach to the prediction residuals is based solely on the joint density $f_{\bar{X}\bar{Y}}(\bar{x},\bar{y};\boldsymbol{\theta})$, and does not require that the underlying model follows the state-space paradigm, i.e. the particular form in (1). Although this paper focuses on validation of state space models based on time series data, TMB thus allows automated computation of one-step prediction residuals for any statistical mixed-effects model in which the joint density is specified and the data points are ordered, i.e. in the general framework of prequential statistics (Dawid, 1984).

5 Validation based on a single sample from the posterior

An alternative approach to validation utilises, instead of prediction errors, a single sample of the unobserved random variables in the model (Waagepetersen, 2006; Gelman et al, 2014). To this end, assume that the data $\bar{Y}$ has been generated by a model $\boldsymbol{M}$, and that we generate a single sample of the latent variables $\bar{X}$ from the posterior distribution of $\bar{X}$ given $\bar{Y}$ under $\boldsymbol{M}$. Then, the pair $(\bar{X},\bar{Y})$

would be jointly generated by the model M; in particular $\bar{X}$ would be a sample from the *prior* distribution. We can now search for patterns in $\bar{X}$, or in $(\bar{X},\bar{Y})$, that would be unlikely under the model. For a state space model, we can derive both state transitions and measurement errors from $(\bar{X},\bar{Y})$, and compare these with the assumptions in the model.

To illustrate this principle with an example, figure 1 shows the standardised process errors $\hat{\sigma}^{-1}E_i$. These are computed as

$$E_i = X_{i+1} - X_i - \hat{\mu}, \quad i = 1,\ldots N-1 \ , \tag{11}$$

where the state vector $\bar{X} = (X_1,\ldots,X_N)$ is a single random sample from the conditional distribution of states given observations.

When the sample is generated under M_0, i.e. with $\hat{\mu} = 0$, the process errors E_i feature a distinct bias which, according to M_0, should not be there. The hypothesis that the mean is 0 is rejected in a t-test with a critical p-value of 4e-07. This leads to rejecting the model M_0 and suggests the alternative M_1. The standardised process errors, generated under M_1, have a mean of exactly zero as was the case for the prediction errors, and the model M_1 passes this validation.

Since our computational tool of choice is direct likelihood optimisation and the Laplace approximation, we draw a sample of $\bar{X}$ as follows: First, we find the posterior mode of $\bar{X}$ given $\bar{Y}$ by maximizing the logarithm of the joint density (1) w.r.t. $\bar{x}$. Next, we find the Hessian using automatic differentiation. Approximating the posterior distribution of $\bar{X}$ given $\bar{Y}$ with a Gaussian, we have now identified the mean and the precision matrix, which allows us to sample from the distribution. We note that this is done directly from the precision matrix, i.e. without computing the covariance matrix. For large models, this avoids a costly matrix inversion and utilizes that the precision matrix is sparse while the covariance matrix is dense. While this is a simple and fast computational approach, and accurate for linear Gaussian models as in this example, approximation errors arise when the model is nonlinear and non-Gaussian. These errors are difficult to assess in general but can be investigated with a simulation study. Alternatively, the single sample of $\bar{X}$ may be drawn using Markov Chain Monte Carlo.

Although we in this paper focus on computations employing the Laplace approximation, we comment briefly on the use of Markov chain Monte Carlo. Here, generating a single sample $\bar{X}$ from the posterior is straightforward: We would let the chain run until convergence, stop at an arbitrary time, and let $\bar{X}$ be the value of the latent variables at the time of stopping. This simple approach seems to be underutilised, given the prominence of MCMC methods in ecological statistics.

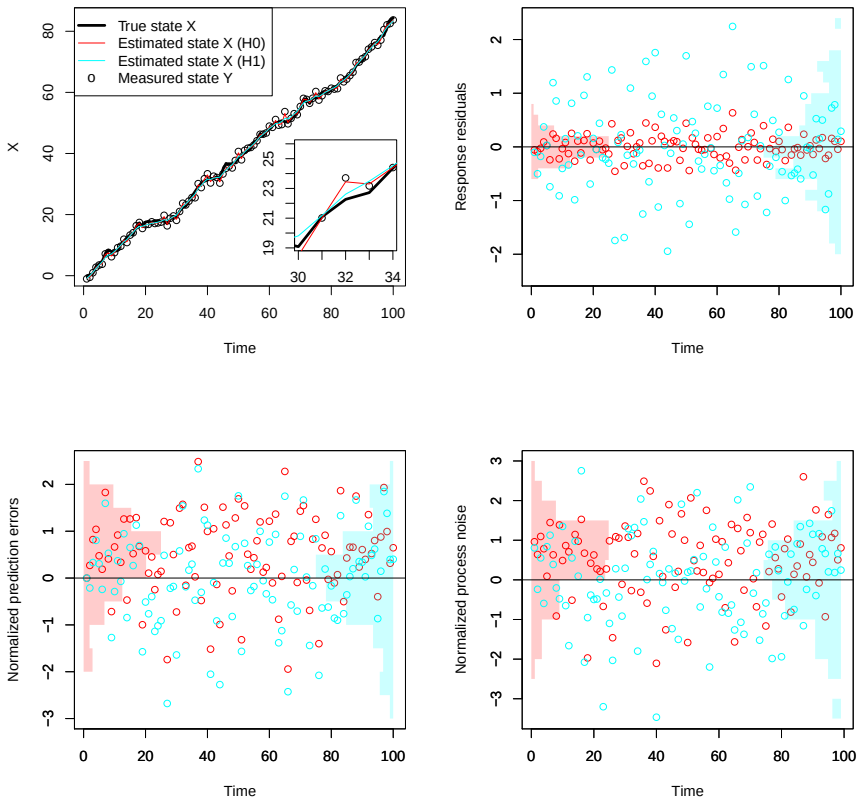

Fig. 1 Results from the toy example. In all panels, cyan colours are used for the "correct" model M_1 while red colours are used for the "incorrect" model M_0. Top left: Time series of true states, observations, and estimated states. The insert shows a zoom-in. Top right: Naive residuals as time series and histogram. Bottom left: Standardised prediction residuals from (7). Bottom right: Normalized sampled process errors from (11).

This technique is also implemented in the example `randomwalkvalidation` in the TMB distribution.

6 A multivariate random walk with correlations

To illustrate the multivariate case, consider an n-dimensional random walk

$$X_{i+1} = X_i + E_i, \quad i = 1, \ldots, N-1 \tag{12}$$

Model	ρ_X	ρ_Y
True	0.9	0.9
1	$\hat{\rho_X}$	$\hat{\rho_Y}$
2	0	0
3	$\hat{\rho_X}$	0
4	0	$\hat{\rho_Y}$

Table 1 Correlations used in the true data generating model and estimation models 1–4. Model 1 is the correct model, whereas models 2–4 are falsely assuming no correlations among processes, observations, or both.

where the increments $E_1, \ldots, E_{N-1}$ are iid and $N(0, \Sigma_X)$. We measure the state

$$Y_i = X_i + W_i \tag{13}$$

with measurement errors $W_1, \ldots, W_N$ that are iid and $N(0, \Sigma_Y)$, independent of the states. Here, Σ_X and Σ_Y are variance-covariance matrices of state increments and measurement errors within a time step. They are specified in terms of marginal standard deviations $\sigma_j = \sqrt{\Sigma_{j,j}}$ and a correlation coefficient ρ using an AR(1) correlation structure:

$$\Sigma_{j,\tilde{j}} = \rho^{|j-\tilde{j}|}\sigma_j\sigma_{\tilde{j}} \ . \tag{14}$$

These correlations imply that the prediction residuals (5) will be vectors with correlations between the different elements. Since validation tests rely on independent and standard Gaussian residuals, one approach is to standardise and decorrelate the vector residuals by multiplying with a square root of the inverse variance/covariance matrix. Conversely, the approach in TMB is to assume that measurements are made available to the estimation procedure one univariate variable at a time, also in the case of vector measurements. Thus, at each time step i, we perform n data updates, processing one element in the observation vector Y_i at a time. The natural default is to let the elements of a vector measurement be processed in the order they appear in the vector. This means that no theoretical extension and no coding effort is required to deal with multivariate measurements. However, this must be kept in mind when inspecting the residuals for patterns: For example, the residual associated with $Y_{i,j}$, i.e. the jth element in the vector Y_i, stems from the prediction of $Y_{i,j}$ based on $Y_{\tilde{i}}$ for $\tilde{i} < i$ as well as on $Y_{i,\tilde{j}}$ for $\tilde{j} < j$.

We simulate a data set from the system with $n = 4$ components and $N = 100$ time steps. We then estimate 4 models, differing in the correlations included (table 1): Model family 1 includes the data generating system, while the other model families lack correlation in state increments and/or measurement errors. We then compute prediction residuals and search for correlations in them. Figure 2 shows sample auto-correlation functions (ACF's) across time-steps and states for each fitted model. Table 2 shows true and estimated parameters

	True	Model 1	Model 2	Model 3	Model 4
ρ_X	0.900	0.877	0.000	0.995	0.000
ρ_Y	0.900	0.916	0.000	0.000	0.928
$\sigma_{X,1}$	0.500	0.480	0.496	2.692	0.329
$\sigma_{X,2}$	0.667	0.492	0.508	2.856	0.330
$\sigma_{X,3}$	0.833	0.379	0.408	2.894	0.219
$\sigma_{X,4}$	1.000	0.454	0.529	2.722	0.321
$\sigma_{Y,1}$	2.000	2.118	2.106	0.927	2.213
$\sigma_{Y,2}$	2.000	2.151	2.129	0.785	2.274
$\sigma_{Y,3}$	2.000	2.098	2.082	0.378	2.210
$\sigma_{Y,4}$	2.000	2.065	2.042	0.780	2.165
AIC		1320.385	1845.043	1408.499	1359.413
KS test		0.749	0.018	0.672	0.012
LB test (time)		0.064	0.000	0.000	0.002
LB test (state)		0.704	0.000	0.000	0.084

Table 2 True/estimated parameters, AIC, and p-values from the Kolomogorov-Smirnov and Ljung-Box tests of the residuals

and the Akaike Information Criterion (AIC; Akaike (1974)) for each model. The table also show the results of a two-sided Kolmogorov-Smirnov test for normality of the residuals, and of two Ljung-Box tests (Ljung and Box, 1978) for autocorrelation – one across time within a component (lags 1–10) and another across states within time-steps (lags 1–3).

The residuals tests from model 1 pass all tests, as expected since this model family includes the data generating system. Conversely, all tests from model 2 are rejected, as might be expected since both process- and measurement errors are wrongly assumed independent in this model. Models 3 and 4 have less obvious patterns in their residuals, illustrated by the fact that each passes one of the residual tests. Apparently the dynamics of the true model can be mimicked quite closely by a model which wrongly assumes independence in either the process or measurement equation.

In summary, the correct model 1 is validated, the incorrect models 2-4 are falsified, even if models 3 and 4 are only falsified weakly. Finally, when all four models are fitted as done here, the correct model would be selected with the AIC.

This example shows that the approach of one-step predictions is feasible also in the multivariate case, but also indicates the limitation of the approach: More than one test of the residuals may be needed to detect model misspecification; this in turn introduces the difficulties of multiple tests. Although model 4 has residuals with stastistically significant patterns, as indicated by the tests, the p-values are not as extreme as one might hope. Finally, although both models 3 and 4 are falsified, the patterns in the residuals do not indicate clearly in which direction the model should be extended.

This example is available in the TMB distribution as `MVRandomWalkValidation`.

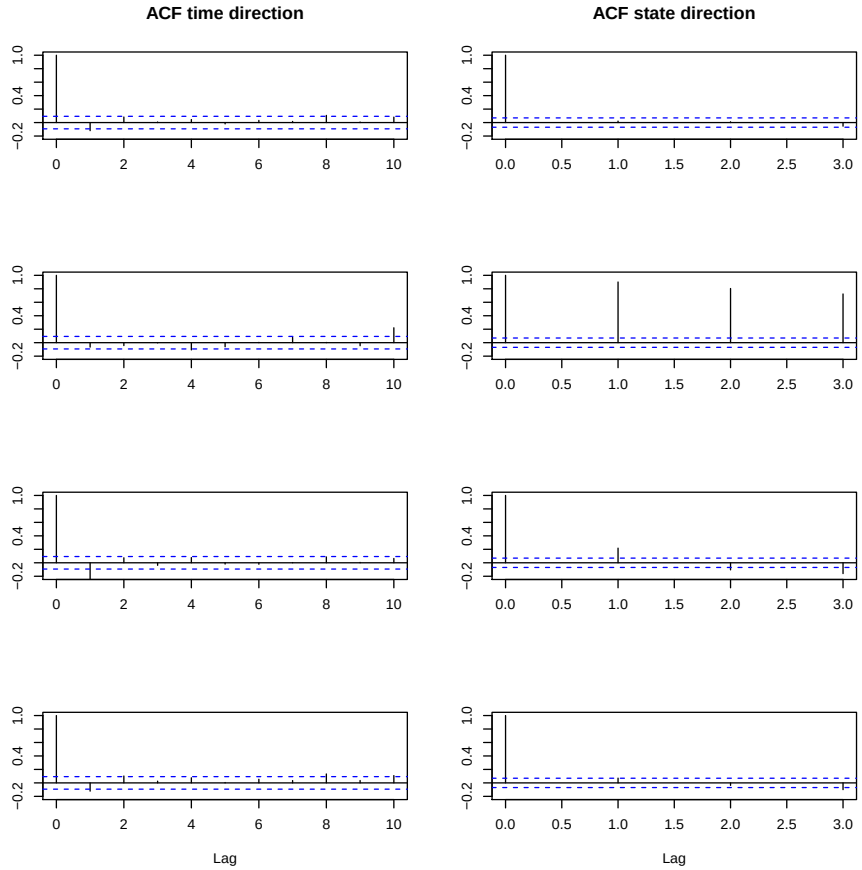

Fig. 2 Estimated auto-correlation of prediction residuals for models 1–4 (rows correspond to models) in the multivariate random walk example.

7 Nonlinear population dynamics with discrete measurements

The following example demonstrates that our framework and implementation applies also to strongly non-linear ecological models with discrete observations. We first simulate a state trajectory from the deterministic logistic map $N_{i+1} = RN_i(1-N_i)$. We take $R = 3.7$, i.e. the model is in the chaotic regime (May, 1974; Murray, 1989). We next generate a data set with Y_i being Poisson distributed with mean $S \cdot N_i$ with $S = 50$ while observations Y_i and Y_j are stochastically independent for $i \neq j$; see figure 3 (left panel). Finally we fit to this data a different model, namely the stochastic Ricker model with Poisson observations:

$$X_i|X_{i-1} \sim N(X_{i-1} + r(1 - e^{X_{i-1}}/K), Q), \quad Y_i|X_i \sim \text{Poisson}(Se^{X_i})$$

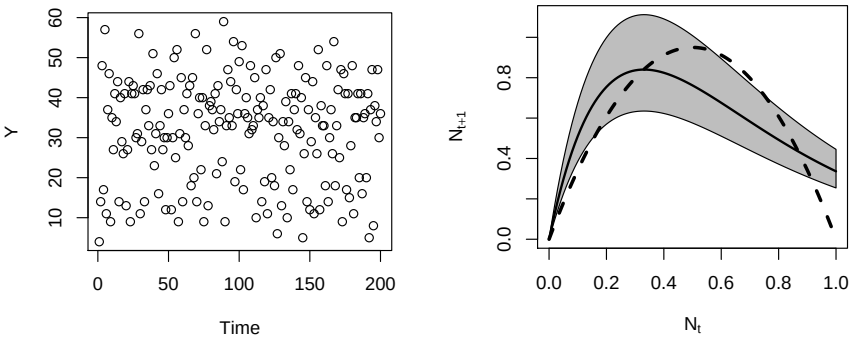

Fig. 3 The nonlinear model. Left panel: Simulated measurements. Right panel: The data generating deterministic map (dashed line) and the fitted model (thick line indicating median, grey zone indicating ± one standard deviation in log domain.

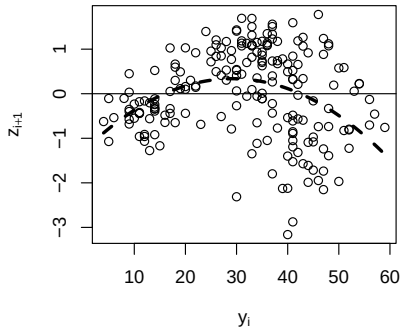

Fig. 4 Prediction residuals from the fitted stochastic Ricker map, plotted against the previous measurement. Included is a fitted parabola, based on assumed constant variance.

Here, the states X_i are log-abundances. We assume conditional independence as in section 2. We estimate these states X_i as random effects, in addition to the fixed effects r, K, and Q. The parameter S, describing the sample size of each measurement, and the carrying capacity K are statistically confounded, since they both express the scale of the data. We therefore fix S in the estimation. Figure 3 (right panel) shows the data generating logistic map as well as the fitted stochastic model in natural domain.

Based on the fitted model, we compute prediction residuals. Figure 4 shows these residuals plotted against the previous observation. Note that if the data

had been generated with the fitted model, these should be independent. Testing for a non-linear dependency between previous observations and residuals, we fit a linear model in which the response is the residual Z_{i+1} and covariates are the previous observation Y_i as well as the square of this previous observation, Y_i^2. The predictions from this model is included in the figure. In a likelihood ratio test for whether the quadratic term can be ommited, we find a p-value of 6e-07. We conclude that the state transition map in the fitted model does not adequately describe the data.

Knowing the data generating mechanism, we can explain this pattern in the residuals. From figure 3, right panel, it is clear that the fitted model over-predicts the future abundance when the current abundance is low or high, and correspondingly under-predicts the future abundance when the current abundance is intermediate. This difference in the models agrees with the bell-shaped pattern in the residuals in figure 4.

This example is available in the TMB distribution as `rickervalidation`.

8 Geolocation of seals

We consider the data set on the sub-adult ringed seal analysed by Albertsen et al (2015). Data available are position measurements obtained from the Argos satellite system; for brevity we focus on the latitudinal coordinate. Let Y_i be the observed latitude at time t_i; here, $i \in \{1,\ldots,N\}$ with $N = 3,583$. The sample times $\{t_i\}$ are irregularly spaced as they depend on contact between the tag on the seal and the satellite. Available is also a quality class C_i, which may take one of seven levels, attached to each latitude measurement Y_i and indicating the precision as assessed by the manufacturer. The observations $\{Y_i\}$ are noisy (figure 5) and the objective of the model is to estimate the positions by filtering the raw measurements. Following Albertsen et al (2015) we model the position Q_t and velocity V_t of the seal using the Ornstein-Uhlenbeck process (Øksendal, 2010), i.e. as continuous-time stochastic processes which satisfy the stochastic differential equation

$$dQ_t = V_t\ dt, \quad dV_t = -\beta V_t\ dt + \sigma\ dB_t$$

where the relaxation rate $\beta > 0$ and the noise intensity $\sigma > 0$ are parameters to be estimated. The transition probabilities in the state process $X_t = (Q_t, V_t)$ are Gaussian with known statistics (Albertsen et al, 2015):

$$\mathbf{E}\{X_{t+h}|X_t\} = \begin{bmatrix} 1 & (1-e^{-\beta h})/\beta \\ 0 & e^{-\beta h} \end{bmatrix} X_t$$

$$\mathbf{V}\{X_{t+h}|X_t\} = \sigma^2 \begin{bmatrix} \frac{1}{\beta^2}\left(h - 2\frac{1-e^{-\beta h}}{\beta} + \frac{1-e^{-2\beta h}}{2\beta}\right) & \frac{(1-e^{-\beta h})^2}{2\beta^2} \\ \frac{(1-e^{-\beta h})^2}{2\beta^2} & \frac{1-e^{-2\beta h}}{2\beta} \end{bmatrix}$$

We assume that the Argos observations Y_i can be written as

$$Y_i = Q_{t_i} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, s^2_{C_i})$ are normally distributed errors, independent of each other as well as of the states, and with a variance which depends on the quality class C_i. The states X_{t_i}, the movement parameters β, σ^2, and the seven measurement variance parameters s^2_c, are estimated using the R package `argosTrack` (Albertsen et al, 2015) which uses TMB for computations. Next, prediction residuals are computed with TMB. To re-iterate, the estimation of states and parameters, and the computation of prediction residuals, require only that the joint density of states $\{X_{t_i}\}$ and observations $\{Y_i\}$ in (1) can be evaluated, after which integration over the $2N = 7,166$ random effects (Q_{t_i}, V_{t_i}) is done automatically by TMB using the Laplace approximation, both when evaluating the likelihood function (2) and when computing the residuals, i.e. (8).

To investigate for bias in the prediction residuals, the residuals are averaged over each week (figure 6, left panel) to yield a mean and a confidence interval, which should include 0. To inspect for homoscedasticity, the figure includes prediction intervals, i.e. the sample mean of the prediction residuals plus and minus two times the standard deviation, computed for each week. To investigate auto-correlation in the residuals, since the times of measurements are not regularly spaced, we use the semivariogram (Pebesma, 2004) and convert this to an estimated auto-covariance function (figure 6, right panel). Since the residuals should be independent, the estimated auto-covariance function should be fluctuate around 0 with no clear pattern.

The residuals indicate several problems with the model: First, the mean of the latitude residuals is significantly different from zero in first weeks of the time series (figure 6, left panel). Although we have not tested for heteroscedasticity, a visual inspection of the width of the prediction intervals (figure 6, left panel) suggests that the model performs better at predicting in the middle of the time series, and particularly bad towards the end. Throughout the data set, but particularly towards the end, there are a substantial number of outliers. Finally, the autocovariance of the residuals is significantly different from 0 at all lags beyond 200 hours.

The patterns in the residuals of time-varying bias and variance, and the outliers, imply that the empirical distribution residuals are very far from normally

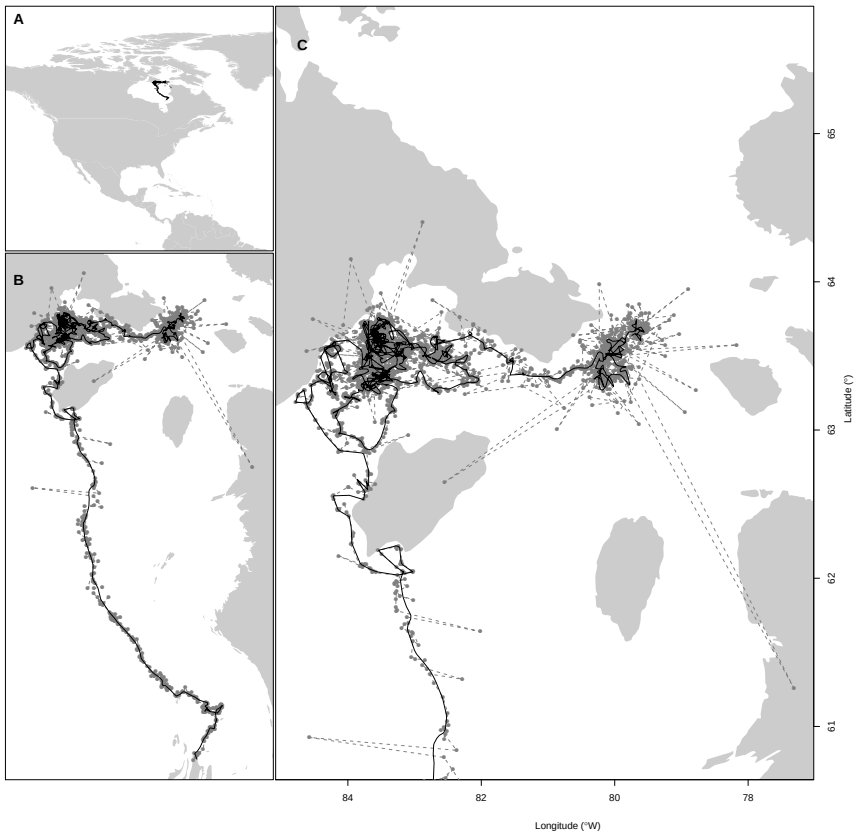

Fig. 5 Area overview (A), the full track (B), and close-up of the last part of the track (C). Grey dots connected with dashed grey line: Observed locations. Black line: Estimated locations.

distributed, and a Shapiro-Wilks test (Shapiro and Wilk, 1965) rejects normality with a critical p-value of less than 10^{-15}. While these issues are enough to falsify the model, they also suggest the problems with the model that could be improved in an iterative fashion. As also evident from the map in figure 5, the motion can be divided into periods of residence and periods of migration. The patterns in the residuals of time-varying bias and variance, and long-range correlation, suggest that the current model does not capture both types of behaviour. Introducing behavioural switching or slowly time-varying parameters into the model could perhaps remedy this. Further, the large number of outliers suggest that the measurement error distribution should be changed from a Gaussians to e.g. a t-distribution.

In summary, the prediction residuals indicate several shortcomings of the model and show us different paths to extend the model in an iterative mod-

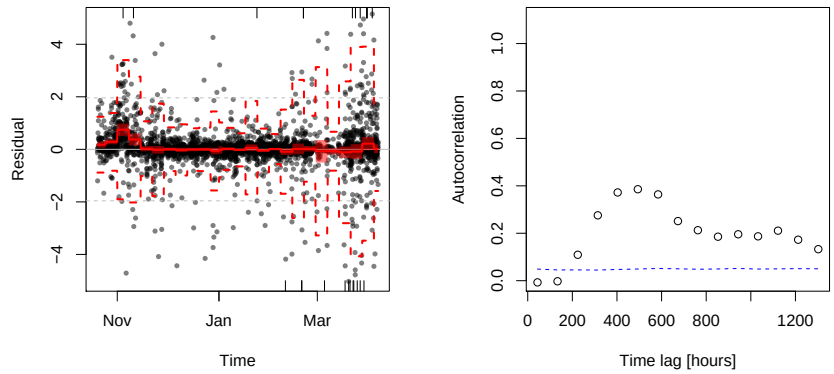

Fig. 6 Left panel: Prediction residuals against time. Outliers out of plotting range are indicated with tick marks on the bounding box. A bin smoother with weekly bins is added (red line) with 95% prediction intervals (dashed red line) and 95 % confidence intervals for the mean (red shaded region). Right panel: The sample auto-correlation of the residuals. Values above the dashed line are significantly different from 0 at the 95 % level.

elling process. Although this falsifies the model, we note that the model is still effective in removing noise from the Argos satelite data and providing confident position estimates by interpolating in the data set. The residuals suggest, however, that the mechanisms in the model do not capture actual movements of the seals, so that for example the parameter β should be viewed as an algorithmic tuning parameter rather than a property of the seal.

9 Discussion

Model validation is a crucial part of statistical modelling; it allows us to either reject the model for being too simple or wrongly specified, or proceed with confidence that the model describes the data adequately. The step of model validation is required to scrutinize the often numerous simplifying assumptions made at the onset of the modelling process. While model selection using e.g. AIC is common in state-space modeling of ecological time series, far fewer applied studies report results from model validation. We stress that model selection using AIC and model validation serve complementary purposes and cannot replace each other. Because model validation does not specify a precise alternative hypothesis, the tests performed during model validation are typically not as powerful against specific hypothesis. This was the case in the multivariate example in section 6, where the AIC with great confidence identified the correct model, while the residuals from models 3 and 4 were less able to clearly differentiate between models. Conversely, while AIC and other model selection techniques allow the modeller to identify the best model

within a specified family, these techniques do not address the question if this family of models was chosen sufficiently large, or if the modeler overlooked some important component that should have been in the model. Furthermore, the residuals provide information about not just overall model fit, but also when and how the model fails to predict the next data point. In contrast to the AIC, this information provides guidance for improving the model in an iterative process.

If the final model is falsified rather than validated, this casts doubt on the entire statistical analysis. Although a model may be useful even if it is wrong, as we argued was the case for geolocating seals in section 8, the theoretical basis for likelihood-based analysis relies on the data generating system belonging to the specified family. If this is not the case, then maximum likelihood estimates, their precision, likelihood ratio test statistics, and the AIC can be misleading. In particular, maximum likelihood estimation effectively tunes parameters in the model so as to produce the optimal one-step predictions. When the model family does not contain the data generating system, this tuning is often at the expense of poorer ability to generate long-term predictions. This is evident in figure 6, where the residuals are uncorrelated for short time lags but positively correlated at longer time lags; clearly the model does not capture this long-term persistence in the velocity of the seal. This tendency of the maximum likelihood estimator to give emphasis to short-term predictions should be kept in mind, in particular when using the fitted model to make long-term predictions.

For state space models of ecological time series, model validation seems to receive less attention than other steps in the modelling process, except in special cases such as linear models and Hidden Markov Models with finite state spaces. In the literature, validation is sometimes based on a naive use of response residuals such as the difference between observed values and posterior means, even if there is no justifying basis for this, as illustrated by our example in section 3. A likely partial cause for this unsatisfying state of the art is lack of general-purpose algorithms with accompanying software implementations, which motivated the research described in this paper. With the computational approach described in this paper, and the implementation in Template Model Builder, it now requires minimal effort to extend a statistical analysis to also include computation of prediction residuals. Examples of its use have already been published (Berg and Nielsen, 2016; Pedersen and Berg, 2016).

The methods and implementation apply to both linear and non-linear ecological models, Gaussian and non-Gaussian distributions, and both continuous and discrete observations, but only to continuous state spaces. As the techniques derive from the linear-Gaussian paradigm, it is an ongoing debate which techniques and algorithms are applicable to non-linear time series models. Here, we note that the example of section 7 is clearly non-linear in the sense of dynamic systems, even chaotic, but at the same time the Laplace approximation yields reasonable estimates of states and parameters, indicating that the model is

near linear from the point of view of statistical computations. The explanation for this seeming paradox is that the measurement uncertainty is small enough that the posterior distributions of the states are restricted to narrow regions where non-linearities have little effect.

The framework of Laplace approximations provide a unified approach to very general classes of nonlinear mixed effects models, including state space models, and is gaining popularity. The Achilles' heel of the framework remains the accuracy of the Laplace approximation. Although it is straightforward to implement models with strong nonlinearities, there is no gaurantee that the numerical algorithms for maximization converge, and even if they do, there is no guarantee that the Laplace approximation is sufficiently accurate: In general, the posterior distribution of the latent variables $\bar{X}$ conditional on data $\bar{Y}$ may be multimodal or have heavy tails. The practitioning ecological modeler may therefore have to choose between model structures which are ecologically plausible, and those for which computations are known to be feasible and accurate. More research is required to assess the accuracy of the Laplace approximation, and to improve it e.g. by coordinate transformations of state space. However, we believe that the potential of the framework is large enough that applications of the principle, such as the current contribution, should proceed in parallel with efforts to strengthen the fundament. With the current state of the art, it should be considered best practice to verify the computational methods by performing the analysis on an articifical data set from the model. This will make it possible to assess the accuracy of the Laplace approximation, in addition to illuminating parameter identifiability and other statistical properties of the model.

Our emphasis has been on validation through one-step predictions, although mentioning the alternative technique of sampling latent variables once from the posterior. With TMB and other approaches which are based on the Laplace approximation, the accuracy of this approximation is a major concern, which leads us to prefer the one-step predictions. If the posterior distribution of the states is Gaussian, the single sample from the posterior should be considered. Also, when the underlying computational engine is Markov Chain Monte Carlo, the single sample of latent variables may be obtained directly from the Markov chain. We believe that this technique is underutilised in ecological modelling.

The methods presented here apply not only to state-space models of time series, but to any statistical model. In a general setting, data points would be made available to the model one measurement at a time, and the concern is the ability of the model to predict the next data point. This is the general principle of prequential statistics (Dawid, 1984), even if it is most easy to interpret in the context of time series. Our frame of mixed-effects models is suitable for a classical, or frequentist, approach to state space models, where system parameters translate to fixed effects while system states are random effects. In principle, our computational methods could also be applied to a Bayesian description of state space models, where also system parameters are unobserved

random variables. However, the accuracy of the Laplace approximation could be more doubtful in this case.

We have ignored the issues that arise when we attempt to validate a model using the same data set that was used to estimate parameters in the model. This estimation introduces patterns in the residuals which should lead to corrections in the tests performed on the residuals. For example, when applying the Ljung-Box test (Ljung and Box, 1978) to residuals from a fitted ARMA model, the test statistic follows a chi-squared distribution, where the number of estimated parameters are deducted from the degrees of freedom. More generally, Kalliovirta (2012) gives asymptotic results for how parameter estimation affects test statistics computed from residuals. Alternatively, the error associated with parameter estimation can be assessed through simulation and controlled through cross validation.

When inspecting prediction residuals, or other residuals, one scans for a large number of possible patterns, e.g. bias, drift, skewness, heavy tails, correlation with states or driving inputs, and heteroscedasticity. One should avoid the dangers of hypothesis fishing and recall that if multiple true hypotheses are tested, it is likely that some of them are rejected. It is rarely possible to conceive a list of tests on the residuals before seeing them, which means that the hypotheses we are testing, implicitly or explicitly, are not proposed independently of data. This is not in agreement with the principles of statistics; this is a well-known problem of post-hoc analysis. For that reason (and several others, Wasserstein and Lazar (2016)) the ubiquitous significance level of 5 % should not be used uncritically. Similarly, it is recommendable to follow the general advice of reserving part of the data for validation, so that a pattern's significance is not tested on the same data set which suggested the pattern.

When a model is validated, it does not mean that the model is correct, but merely that the available data are insufficient to point out differences between the data generating system and the model. Such differences are ubiquitous: All models are wrong but some are useful (Box and Draper, 1987). Which model is selected for subsequent work should depend not only on available data, but also on the objectives of the analysis. For example, the model in section 8 is useful for removing noise from Argos data and improving the geolocation of seals, but not useful for predicting actual migrations done by the animals. Similarly, a stock assessment model may be useful for estimating the current abundance of a fish stock but inappropriate for long-term predictions of this stock under climate change. Thus, automated model selection that claims to be objective should be regarded with some scepticism, as should rigid procedures for model validation. While final modelling decisions remain the responsibility of the modeller, informed choice requires availability of diagnostic tools. This paper provides and demonstrates such tools.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automatic Control AC-19:716–723, system identification and time-series analysis

Albertsen CM, Whoriskey K, Yurkowski D, Nielsen A, Mills Flemming J (2015) Fast fitting of non-Gaussian state-space models to animal movement data via template model builder. Ecology 96(10):2598–2604

Anscombe FJ, Tukey JW (1963) The examination and analysis of residuals. Technometrics 5(2):141–160, DOI 10.1080/00401706.1963.10490071

Berg CW, Nielsen A (2016) Accounting for correlated observations in an age-based state-space stock assessment model. ICES Journal of Marine Science DOI 10.1093/icesjms/fsw046

Bolker BM, Gardner B, Maunder M, Berg CW, Brooks M, Comita L, Crone E, Cubaynes S, Davies T, Valpine P, et al (2013) Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. Methods in Ecology and Evolution 4(6):501–512

Box GE, Draper NR (1987) Empirical model-building and response surfaces. Wiley, New York

Box GEP, Jenkins GM (1970) Time series analysis: forecasting and control, 1976. ISBN: 0-8162-1104-3

Cadigan N, Morgan M, Brattey J (2014) Improved estimation and forecasts of stock maturities using generalised linear mixed models with auto-correlated random effects. Fisheries Management and Ecology 21(5):343–356

Clark C, Mangel M (2000) Dynamic State Variable Models in Ecology: Methods and Applications. Oxford University Press

Clark JS (2007) Models for ecological data: an introduction, vol 11. Princeton University Press

Cox D, Hinkley D (1974) Theoretical Statistics. Chapman & Hall

Cox DR, Snell EJ (1968) A general definition of residuals. Journal of the Royal Statistical Society Series B (Methodological) 30(2):248–275, URL http://www.jstor.org/stable/2984505

Dawid AP (1984) Present position and potential developments: Some personal views: Statistical theory: The prequential approach. Journal of the Royal Statistical Society Series A (General) pp 278–292

Dunn PK, Smyth GK (1996) Randomized quantile residuals. Journal of Computational and Graphical Statistics 5(3):236–244

Evensen G (2003) The ensemble kalman filter: Theoretical formulation and practical implementation. Ocean dynamics 53(4):343–367

Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, Nielsen A, Sibert J (2012) AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software 27(2):233–249

Frühwirth-Schnatter S (1996) Recursive residuals and model diagnostics for normal and non-normal state space models. Environmental and Ecological Statistics 3(4):291–309

Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Taylor & Francis

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov Chain Monte Carlo in Practice. Interdisciplinary Statistics, Chapman and Hall, London

Griewank A, Walther A (2008) Evaluating derivatives: principles and techniques of algorithmic differentiation. SIAM

Harvey AC (1989) Forecasting, structural time series models and the Kalman filter. Cambridge University Press

Jonsen I, Flemming J, Myers R (2005) Robust state-space modeling of animal movement data. Ecology 86(11):2874–2880

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

Kalliovirta L (2012) Misspecification tests based on quantile residuals. The Econometrics Journal 15(2):358–393, DOI 10.1111/j.1368-423X.2011.00364.x, URL http://dx.doi.org/10.1111/j.1368-423X.2011.00364.x

Kalman R (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering pp 35–45

Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM (2016) TMB: Automatic differentiation and Laplace approximation. Journal of Statistical Software 70(5):1–21, DOI 10.18637/jss.v070.i05

Liu JS, Chen R (1998) Sequential monte carlo methods for dynamic systems. Journal of the American Statistical Association 93:1032–1044

Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65(2):297, DOI 10.1093/biomet/65.2.297, URL + http://dx.doi.org/10.1093/biomet/65.2.297

Ljung L (1999) System Identification - Theory for the User, 2nd edn. Information and System Sciences Series, Prentice-Hall

Madsen H (2007) Time series analysis. Chapman & Hall/CRC

May RM (1974) Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. Science 186(4164):645–647

Murray J (1989) Mathematical Biology. Springer-Verlag

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

Øksendal B (2010) Stochastic Differential Equations - An Introduction with Applications, sixth edn. Springer-Verlag

Patterson T, Thomas L, Wilcox C, Ovaskainen O, Mathhiopoulos J (2008) State-space models of individual animal movement. Trends in Ecology & Evolution 23(2):87–94

Pebesma EJ (2004) Multivariable geostatistics in s: the gstat package. Computers & Geosciences 30:683–691

Pedersen MW, Berg CW (2016) A stochastic surplus production model in continuous time. Fish and Fisheries

Pedersen MW, Berg CW, Thygesen UH, Nielsen A, Madsen H (2011) Estimation methods for nonlinear state-space models in ecology. Ecological Modelling 222(8):1394–1400

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Rall LB (1980) Applications of software for automatic differentiation in numerical computation. In: Fundamentals of Numerical Computation (Computer-Oriented Numerical Analysis), Springer, pp 141–156

Rosenblatt M (1952) Remarks on a multivariate transformation. The annals of mathematical statistics 23(3):470–472

Rue H, Martino S, Chopin N (2009) Approximate bayesian inference for latent Gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71(2):319–392

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3-4):591, DOI 10.1093/biomet/52.3-4.591, URL + http://dx.doi.org/10.1093/biomet/52.3-4.591

Skaug HJ, Fournier DA (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Computational Statistics & Data Analysis 51(2):699–709

Smith J (1985) Diagnostic checks of non-standard time series models. Journal of Forecasting 4(3):283–291

Thygesen UH, Sommmer L, Evans K, Patterson TA (2016) Dynamic optimal foraging theory explains vertical migrations of bigeye tuna. Ecology To appear.

Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. Journal of the american statistical association 81(393):82–86

de Valpine P, Hastings A (2002) Fitting population models incorporating process noise and observation error. Ecological Monographs 72(1):57–76

Waagepetersen R (2006) A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. Scandinavian journal of statistics 33(4):721–731

Wan EA, Van Der Merwe R (2000) The unscented Kalman filter for nonlinear estimation. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000, IEEE, pp 153–158

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician (just-accepted):00–00

Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. CRC Press