Are state-space stock assessment model confidence intervals accurate? Case studies with SAM and Barents Sea stocks

Cadigan, Noel G.; Albertsen, Christoffer Moesgaard; Zheng, Nan; Nielsen, Anders

Published in:
Fisheries Research

Link to article, DOI:
10.1016/j.fishres.2024.106950

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Fisheries Research 272 (2024) 106950

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Are state-space stock assessment model confidence intervals accurate? Case studies with SAM and Barents Sea stocks

Noel G Cadigan [a,*], Christoffer Moesgaard Albertsen [c], Nan Zheng [b], Anders Nielsen [c]

[a] Centre for Fisheries Ecosystems Research, Fisheries and Marine Institute of Memorial University of Newfoundland, St. John's, NL A1C 5R3, Canada
[b] Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada
[c] National Institute of Aquatic Resources, Technical University of Denmark, Henrik Dams Allé, building 201, DK-2800 Kgs, Lyngby, Denmark

ARTICLE INFO

Handled by A.E. Punt

Keywords:
State-space model
Bias
Confidence intervals
Random effects
Conditional

ABSTRACT

Our main contribution is to examine the reliability of confidence intervals using the SAM state-space fish stock assessment model used for the assessment of many stocks by the International Council for the Exploration of the Seas. We focus on frequentist statistical inferences and more specifically on inference conditioned on specific values of the state-space model random effects drawn from their process distribution. This is somewhat consistent with simulation self-test procedures that are commonly used to examine the reliability of state-space assessment model results. However, recent research has indicated that some estimation bias may be expected in the conditional setting. Hence, we also investigate recently proposed bias corrected confidence intervals appropriate for the conditional inference setting. The SAM simulation coverage probabilities of 95% confidence intervals for SSB and Fbar were usually slightly larger than 95%, but in a small number of years these coverage probabilities could be much smaller than 95%. The bias corrected confidence intervals were more reliable. When averaged over years, the SAM and bias corrected confidence interval coverage probabilities were similar for the Northeast Artic cod and saithe case studies, but the bias corrected confidence intervals performed much better overall for the haddock case study.

1. Introduction

Fish stock assessment models are expressed naturally as state-space models, which are a class of probalistic models designed to describe how unobservable (latent) variables influence observed variables. In stock assessment models the observations commonly include time series of commercial catch estimates and stock size indices derived from scientific surveys. The latent variables are the time series of stock abundance, its age/size structure, and fishing mortality rates. The latent stock size is assumed stochastic because it is influenced by many factors in addition to fishery catches and assumed natural mortality rates. We cannot directly account for all of these other factors in our assessment models, and even if we could it would be impossible to predict exactly how many fish survive in a given time period. The observations available to estimate the stochastic stock size process are often indirect and always subject to observation/sampling errors. The main difference between a standard statistical (full parametric) assessment model and a state-space assessment model (SSAM) is that the latter allows for quantities which are unobserved to be random variables with a specified probability distribution. The aim of a SSAM is to estimate these latent time series, facilitate predictions, and reliably quantify the uncertainties of these estimates and predictions. SSAMs provide a consistent and natural framework for these purposes (e.g., Aeberhard et al., 2018). They provide the flexibility to formulate models where time-varying latent quantities follow a random walk or an autoregressive (AR) process, etc. A practical advantage of SSAMs compared to full parametric and deterministic models is that the method to do stochastic predictions is a natural part of the SSAM formulation.

Maximum marginal likelihood estimation (MMLE) is a common method to estimate SSAM parameters (e.g., Aeberhard et al., 2018). MMLE requires evaluations of high dimensional integrals, which until recently was often not feasible for estimation and simulation testing of full-scale assessment models. However, with recent advances in algorithms and software (i.e., Kristensen et al., 2016), the run-time to fully optimize such models is now often practically feasible. This is the main reason why SSAMs that integrate multiple sources of data related to stock productivity are increasingly being used to provide fisheries managers with advice on sustainable harvest rates and the consequences

* Corresponding author.
E-mail address: noel.cadigan@mi.mun.ca (N.G. Cadigan).

https://doi.org/10.1016/j.fishres.2024.106950
Received 15 November 2023; Received in revised form 8 January 2024; Accepted 8 January 2024
Available online 13 January 2024

Table 1
Definitions, notations, and parameters.

SSAM	state-space assessment model
AR	autoregressive
MMLE	Maximum marginal likelihood estimation
TMB	Template Model Builder
SE	standard error
MSE	Mean squared error
RMSE	Root mean squared error
EBP	Empirical bayes predictor
RE	random effect
pdf	probability density function
pmf	probability mass function
HDP	high dimensional parameter vector
GAM	Generalized additive model
CI	Confidence interval
CP	CI coverage probability
BC	Bias corrected
SSB	Spawning stock biomass
Fbar	Average fishing mortality
D	$n \times 1$ vector of data
θ	fixed-effects parameter vector
Ψ	vector of random-effects
$f(\Psi\|\theta)$	pdf of Ψ
$f(D\|\theta)$	marginal pdf/pmf of D
$\widehat{\theta}$	MMLE of θ
$\widehat{\Psi}$	EBP of Ψ
F	Fishing mortality
N	Stock abundance
c	user specified value for bias correction

of future fisheries (e.g., Schnute, 1994; Aanes et al., 2007; Nielsen and Berg, 2014; Cadigan, 2015; Aeberhard et al., 2018; Albertsen et al., 2018; Perreault et al., 2020; Stock and Miller, 2021; Liljestrand et al., 2023). SSAMs are considered to be an essential part of the next generation stock assessment package (Punt et al., 2020). Currently the most notable example is the SAM stock assessment package (Nielsen and Berg, 2014; Berg and Nielsen, 2016) used by many working groups of the International Council for the Exploration of the Seas (ICES; e.g., ICES, 2019a,b).

Recent versions of SAM are implemented in the Template Model Builder (TMB, Kristensen et al., 2016) package within R (R Core Team, 2018), which is a major driver for recent increases in the implementation of integrated SSAMs. TMB provides built-in and easy-to-use procedures for implementing nonlinear mixed-effects and state-space models. Generic descriptions of nonlinear mixed-effects models are available in Skaug and Fournier (2006) and Kristensen et al. (2016). However, the underlying statistical theory for standard errors (SEs) provided by TMB and SAM are not well described. For these models, Zheng and Cadigan (2021) and Zheng and Cadigan (2023) provided some statistical theory about the frequentist sampling properties of MMLEs of model parameters and empirical bayes predictors (EBPs) of random effects (REs), and also functions of both the model parameters and REs that are typically of interest from SSAMs. We summarize these papers briefly in Appendix A. A key assumption is whether random effects are truely random or actually high dimensional fixed parameters that are modelled as REs for nonparametric smoothing purposes, similar to Generalized Additive Models (e.g., Wood, 2020). To help describe these concepts, we first briefly review the nonlinear mixed-effects model framework Zheng and Cadigan (2021) and Zheng and Cadigan (2023) considered, which includes SSAMs.

Their framework involves random response data collected in a $n \times 1$ vector D that are assumed to have a multivariate probability density/mass function (pdf/pmf) $f(D|\Psi, \theta)$, given values of the $(p \times 1)$ vector of fixed-effects parameters θ and a $(q \times 1)$ vector of REs Ψ. For the SAM stock assessment model, D will typically involve age-based fishery catches and survey indices. Ψ will include the natural logarithms of stock numbers at age (i.e., logN's) and fishing mortality rates (i.e., logF's) for all assessment years. The means and covariances of D depend on θ and Ψ, possibly via nonlinear functions of θ, Ψ and covariates which we do not develop notation for. The pdf of Ψ is $f(\Psi|\theta)$. The marginal distribution of D is

$$f(D|\theta) = \int \cdots \int_q f(D|\Psi, \theta) f(\Psi|\theta) d\Psi_1, \ldots, d\Psi_q, \tag{1}$$

where $\Psi_1, \ldots, \Psi_q$ are the elements of Ψ. The MMLEs of θ are those values $\widehat{\theta}$ that maximize $f(D|\theta)$. Throughout this paper we use $\widehat{}$ to denote estimators. We can "estimate" Ψ as those values $\widehat{\Psi}$ that maximize $f(D, \Psi|\theta)$ when $\theta = \widehat{\theta}$. All mathematical notation and other notations/abbreviations are defined in Table 1.

Zheng and Cadigan (2021) developed frequentist variance approximations for $\widehat{\Psi}$ and smooth user-specified functions of $\widehat{\theta}$ and $\widehat{\Psi}$. The latter was based on the generalized delta method (e.g., Kristensen et al., 2016). Zheng and Cadigan (2021) also clarified the statistical basis for the SEs for $\widehat{\Psi}$ provided by TMB. The frequentist variance conceptually refers to the variability of $\widehat{\Psi}$ derived from repeated estimation with infinitely many data sets randomly drawn from $f(D, \Psi|\theta)$. This will include repeated sampling of Ψ from $f(\Psi|\theta)$ and D from $f(D|\Psi, \theta)$. Zheng and Cadigan (2021) showed that the TMB SEs were actually estimates of the marginal mean squared error (MSE) between $\widehat{\Psi}$ and Ψ, which is a commonly used approach to measure the variability of predictors of REs (e.g., Kackar and Harville, 1984; Datta and Lahiri, 2000; Das et al., 2004; Flores-Agreda and Cantoni, 2019). This is the squared differences between $\widehat{\Psi}$ and Ψ, when averaged over the joint distribution of the data and Ψ. In the SAM assessment model context, standard errors of model outputs are based on the MSE when averaged over the distributions of the data, the logN's and logF's, and other random effects that the SAM model includes. The TMB variance is not an approximation of the variance of $\widehat{\Psi}$ that occurs because of random sampling of Ψ and the data. Zheng and Cadigan (2021) derived an equation that is appropriate for that marginal variance. However, the variance will usually not be relevant for most objectives of fish stock assessment. While the variance of an unbiased estimator is commonly used to construct confidence intervals (CIs) for a fixed parameter, prediction intervals for REs like Ψ, a primary focus in fisheries applications, rely on the variances of prediction errors, $\mathrm{Var}(\widehat{\Psi} - \Psi)$ (see, e.g., Section 11.6 of McClave and Sincich, 2017. Zheng and Cadigan (2021) established that this latter variance corresponds to the TMB variance.

A different inferential setting was addressed in Zheng and Cadigan (2023), which involved the variability of $\widehat{\Psi}$ based on repeat sampling of D only from $f(D|\Psi, \theta)$. Zheng and Cadigan (2023) assumed Ψ was drawn once from the process model $f(\Psi|\theta)$ but then fixed at these values during repeated data generations from the observation model $f(D|\Psi, \theta)$. They referred to this as the Ψ-conditional (or just conditional) variation. Only sampling D from $f(D|\widehat{\Psi}, \widehat{\theta})$ is a common procedure used when simulation testing the efficacy of a fish stock assessment model (e.g., Nielsen and Berg, 2014; Cadigan, 2015; Perreault et al., 2020). Simulated data are generated only from $f(D|\widehat{\Psi}, \widehat{\theta})$ because we are really interested in quantifying how much our estimates will vary based on a fixed stock development (i.e., logN's and logF's) similar to the ones estimated. In stock assessment this is referred to as a simulation self-test (Deroba et al., 2015). However, the distribution of $\widehat{\Psi}$ can be considerably different than the distribution of Ψ (i.e., $f(\Psi|\theta)$) and an alternative procedure is to sample D from $f(D|\widetilde{\Psi}, \widehat{\theta})$ where $\widetilde{\Psi}$ is drawn once from $f(\Psi|D, \widehat{\theta})$. We consider this further in the Discussion.

Zheng and Cadigan (2023, Eqns. 8 and 9) provided conditional variance approximations (see Appendix A), which indicated that these variances of $\widehat{\theta}$ and $\widehat{\Psi}$ are smaller than the marginal variances provided by TMB. Zheng and Cadigan (2023) also derived equations for the biases of $\widehat{\theta}$ and $\widehat{\Psi}$, which indicated that both biases may be non-ignorable. In

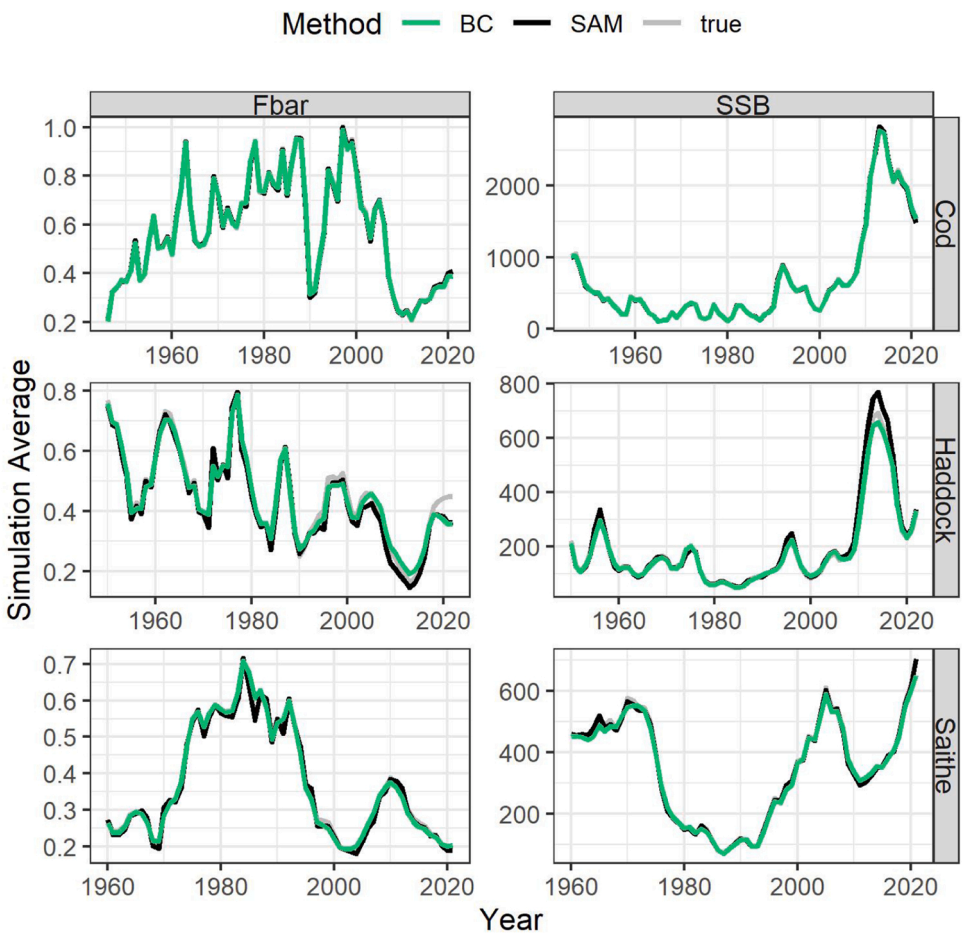

Fig. 1. Simulation results for three Barents Sea stocks (i.e. panels). Grey lines indicate true population values from the model used to generate simulated data, black lines are averages from the 2000 simulation SAM estimates, and green lines are simulation bias-corrected averages. The right-hand panels are for SSB and the left-hand panels are for average fishing mortality (Fbar).

this case MSE is a more comprehensive measure of variability than the variance. The biases are difficult to estimate reliably, especially when there is little or no data for some of the REs. Zheng and Cadigan (2023) suggested to approximate the bias-squared part of the MSE by averaging it over the distribution of REs, which results in the same MSE approximation as that used in TMB and SAM; that is, in the Ψ-conditional inference setting, the TMB variance can be used as the expected MSE, with substantial individual deviations anticipated for some data and specific values of the Ψ's (e.g., $\widehat{\Psi}$ or $\widetilde{\Psi}$).

In many cases Ψ is not actually a RE in reality, but Ψ is a high dimensional parameter vector (HDP) that is modeled as a RE for nonparametric smoothing purposes. In this case the conditional inference setting is also relevant. There are well-known connections between smoothing methods and RE models (e.g., Brown and De Jong, 2001; Wand, 2003; Wood et al., 2013). This is the case for some state-space fish stock assessment models in which subsets of Ψ will usually be time-dependent and could be considered to be complex but smooth functions of time. The generalized additive model (GAM) literature for non- or semi-parametric smoothing (i.e., Marra and Wood, 2012; Wood, 2020) suggests their MSE estimates and CIs are accurate only when they are averaged over the smoothing covariates, which for stock assessment models will usually be years; however, the MSE estimates and CIs may not be accurate for specific years, and smoothing bias can be expected. In this context, Zheng and Cadigan (2024) proposed a bias correction for EBP's of REs and improved CI methods. Their simulation studies for simple random walk models and GAMs, whose Ψ's are the basis coefficients, indicated that their new bias-corrected (BC) EBP of Ψ (denoted as $\widehat{\Psi}_{BC}$) and CIs led to substantial improvements in MSE for Ψ and improvements in the conditional coverage rate of Ψ CIs compared to the marginal inferences provided by TMB.

In this paper we use conditional self-test simulations based on three important Barents Sea SAM assessment models to investigate the accuracy of 1) SAM generalized delta-method SEs, 2) those based on the conditional variance approximations in Zheng and Cadigan (2023, Eqns. 8 and 9), and 3) conditional bias corrections in Zheng and Cadigan (2024), for spawning stock biomass (SSB) and average fishing mortality (Fbar). We also investigate the coverage accuracy of 95% CIs, year-by-year and averaged over years, based on marginal and conditional BC SEs.

2. Methods

2.1. SAM

A brief description of SAM is provided in Appendix B. The SAM logN and logF REs for the first model year do not have a distribution specified. In effect, the logN's and logF's in the first year are estimated like fixed-effects. The theory in Zheng and Cadigan (2023), but especially Zheng and Cadigan (2024), did not cover this situation. Hence, we used a slightly modified version of SAM in which these initial logN's and logF's were treated as parameters and not REs so that the theory in Zheng and Cadigan (2024) applied. We verified that this change in the SAM formulation, which we call SAM_init, had negligible impacts on model results.

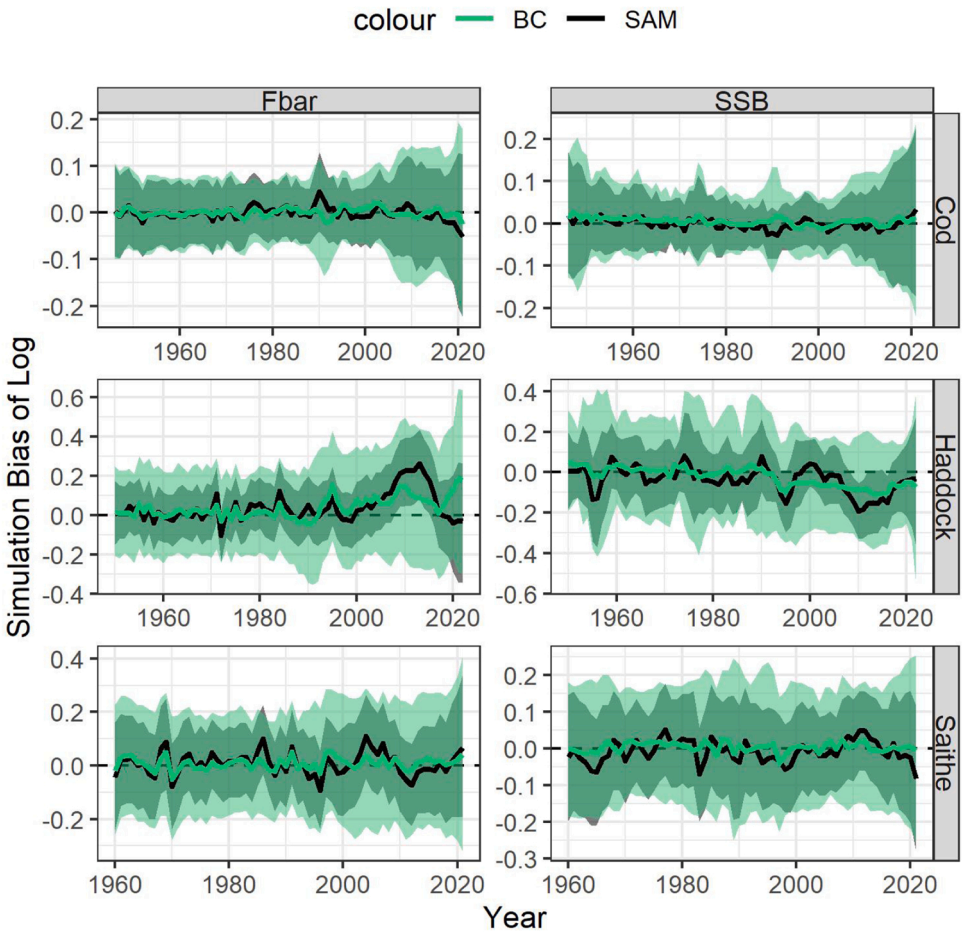

Fig. 2. Simulation bias results for three Barents Sea stocks (i.e. panels). Black lines are average SAM biases from the 2000 simulation log-estimates, and green lines are average biases using the BC estimates. The right-hand panels are for SSB and the left-hand panels are for average fishing mortality (Fbar). Shaded regions indicate lower 2.5% and upper 97.5% simulation percentiles.

There may be REs in SAM that have little or no information (i.e., data) to support their estimation. This could be caused by years with missing catch estimates, or REs involved in forecasts. In this case the EBP will usually be the marginal RE mean. However, as described in Zheng and Cadigan (2024), $\widehat{\Psi}_{BC}$ requires information for all Ψ's. For Ψ's with insufficient information, Zheng and Cadigan (2024) recommended marginal inferences be used. They provided a method to identify whether REs are supported by data sufficiently and give conditional or marginal CIs depending on the amount of information. This involved setting a small constant c to determine if there is sufficient information about Ψ's. The choice of c is somewhat subjective. It should be greater than zero but less than approximately Y^{-1}, where Y is the total number of model years. We used $c = 0.1$ but examined robustness using $c = 0.05$ and $c = 0.2$.

Another problem when applying the results in Zheng and Cadigan (2023) and Zheng and Cadigan (2024) involved the complicated joint distribution of all the SAM logN's. The conditional SEs and CIs in Zheng and Cadigan (2023) and Zheng and Cadigan (2024) require using the marginal means and covariances of the REs, and also a normal distribution assumption. However, these first two statistical moments for logN's are complicated and not easy to derive, and the normal distribution assumption is not correct since logN's depend on exp(logF) and values in previous years. In SAM's, logN's are a nonlinear Markov process with a complicated marginal distribution. Our solution for this problem was to modify SAM again (called SAM_dev), and treat the temporal deviations in logF's and logN's as REs, which have simple multivariate normal joint distributions so that the results of Zheng and Cadigan (2023) and Zheng and Cadigan (2024) apply directly. We also verified that this modification had negligible impacts on model results. A major disadvantage of the SAM_dev formulation is computational speed. Hence, we fit the model to simulated data using SAM_init, but derived conditional variances and confidence intervals using SAM_dev.

2.2. Case studies

The three Northeast Arctic (ICES subareas 1 and 2) case studies are based on SAM assessment inputs and model configuration files for cod (*Gadus morhua*), haddock (*Melanogrammus aeglefinus*), and Saithe (*Pollachius virens*). The assessment data and model configurations are described in Appendix B and more detailed descriptions are provided in ICES (2020). These stocks were selected because their SAM assessments are considered to be reliable and well-estimated.

2.3. Simulations

SAM provides easy to use options to conduct simulations. We used the sim.condRE=TRUE option and simstudy() to generate 2000 simulation data sets and fit the SAM's. The simstudy() procedure supports parallel processing which can greatly improve simulation speed. The sim.condRE=TRUE option causes simulated observations to be conditional on estimated time-series of fishing mortalities (F's) and stock abundances (N's).

The simulation operating models for each case study followed the SAM models specified in Appendix B with the exception that the process deviations were treated as latent variables instead of the processes

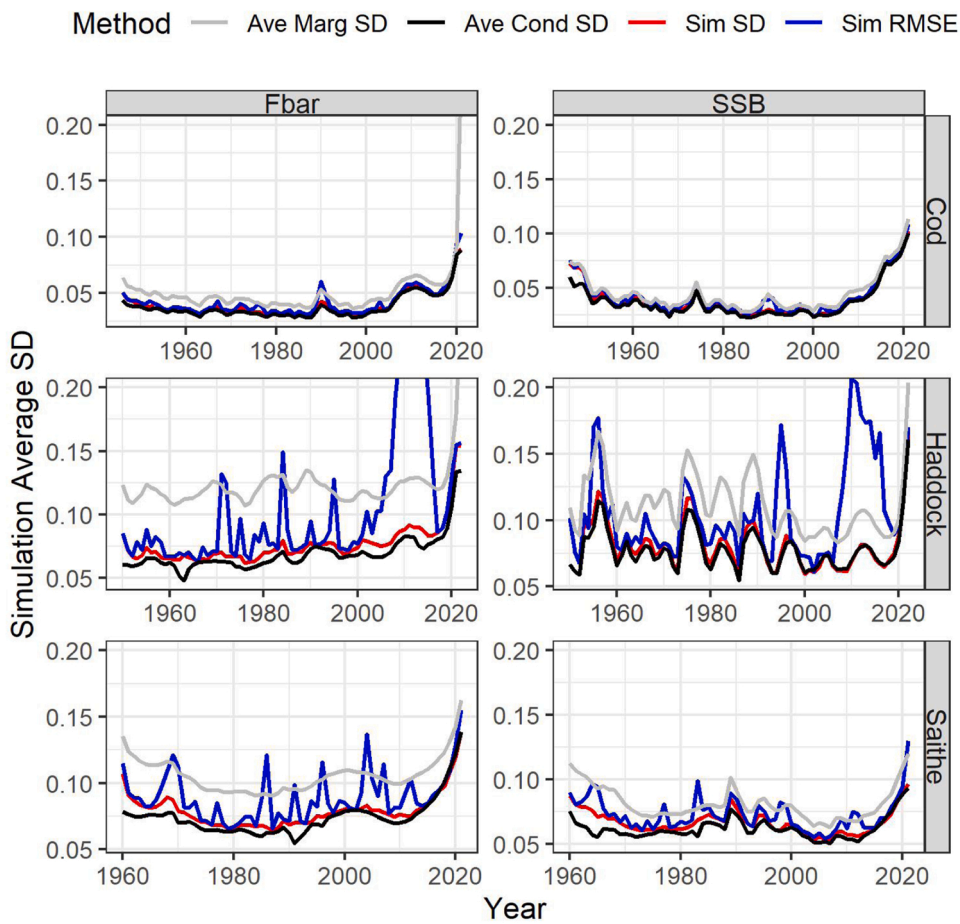

Fig. 3. Simulation average conditional standard errors (black) and marginal standard errors (grey) provided by SAM. The standard deviations (SDs; red) and root mean squared errors (RMSE; blue) of the 2000 simulation estimates of log Fbar and SSB are also indicated in each panel.

variables (i.e., SAM_dev). For example, $\epsilon_{a,y}^{(N)}$ was treated as the latent variable instead of $\log N_{a,y}$. The simulation estimation model was the standard version of SAM with the process variables as latent variables.

We focused our simulation analyses on estimates of SSB and Fbar, with the ages for Fbar the same as used in the assessments for each stock (see ICES, 2020). We computed CIs by exponentiating log CIs, $\exp\{\log(estimate) \pm Z_{0.975} \times SE\}$, where $Z_{0.975}$ is the standard normal quantile and SE is the marginal value for $\log(estimate)$ provided by SAM, or we used the CIs described in Zheng and Cadigan (2024). Deriving CIs from log-estimates is the default procedure provided by SAM and will tend to give more equal-tailed coverage probabilities. We summarize simulations results using the standardized bias (log-estimate bias/SE), SEs, root mean squared error (RMSE), and coverage probabilities of CIs (fraction of the simulated CIs that contained the population values used to generate simulated data). Large standardized biases (e.g., > 0.7 in absolute value) will produce CIs with coverage probabilities more different than the nominal value (e.g., < 0.9).

3. Results

Average estimates of SSB and Fbar from the 2000 simulations closely matched the population generating values (i.e., true values; Fig. 1) for most years, but occasionally the biases were large enough to potentially be a concern (e.g., Fbar in 2020 for Haddock) for fisheries management. Simulation biases of the logs of SSB and Fbar (Fig. 2) demonstrate that differences in estimates of log SSB and Fbar and their true values were usually close to zero. The averages of the conditional SEs based on Zheng and Cadigan (2023) were usually only slightly smaller than the standard deviations (SDs) of estimates of SSB and Fbar from the 2000 simulations (Fig. 3). The marginal SEs provided by SAM were usually larger but coarsely approximated the RMSE's, and the latter will depend on the specific values of Ψ's used to generate simulation data. Fig. 3 demonstrates that the asymptotic theory provided by Zheng and Cadigan (2023) was reasonably accurate for these case studies. Although the conditional SEs were more accurate for the simulation SDs, they produced poor CIs because they do not account for conditional biases (Fig. S1). Zheng and Cadigan (2024) also found that CIs based on conditional SEs were less reliable than CIs based on marginal RMSE or bias correction.

The SAM 95% CIs were usually somewhat conservative with simulated coverage probabilities that were usually greater than 0.95 (Fig. 4). The MSE approximation tends to overestimate the true MSE (Fig. 3), leading to conservative CIs. When averaged over assessment years, the SAM CIs were slightly conservative for our conditional simulations of cod and saithe, but not haddock (Table 2). However, in a few years the CIs contained the true population values in less than 95% of the simulations, and much less for Haddock. As expected, the years with poorer CI coverage were ones in which the standardized biases were large (Fig. 5). The BC CIs of Zheng and Cadigan (2024) were more reliable overall in that they never produced really unreliable intervals like SAM did in a few years, especially for the haddock case study. The BC CIs were slightly conservative when averaged over years, similar to the SAM CIs (Table 2). The BC standardized biases were never as large as the SAM values, but in many years these were slightly larger. The BC biases of the logs of SSB and Fbar (Fig. 2) are usually closer to zero in years when SAM biases were relatively large. To get an aggregate summary of the CI performance, we computed the averaged squared differences from the nominal 0.95 values, as well as the annual SD of the CI coverages. These

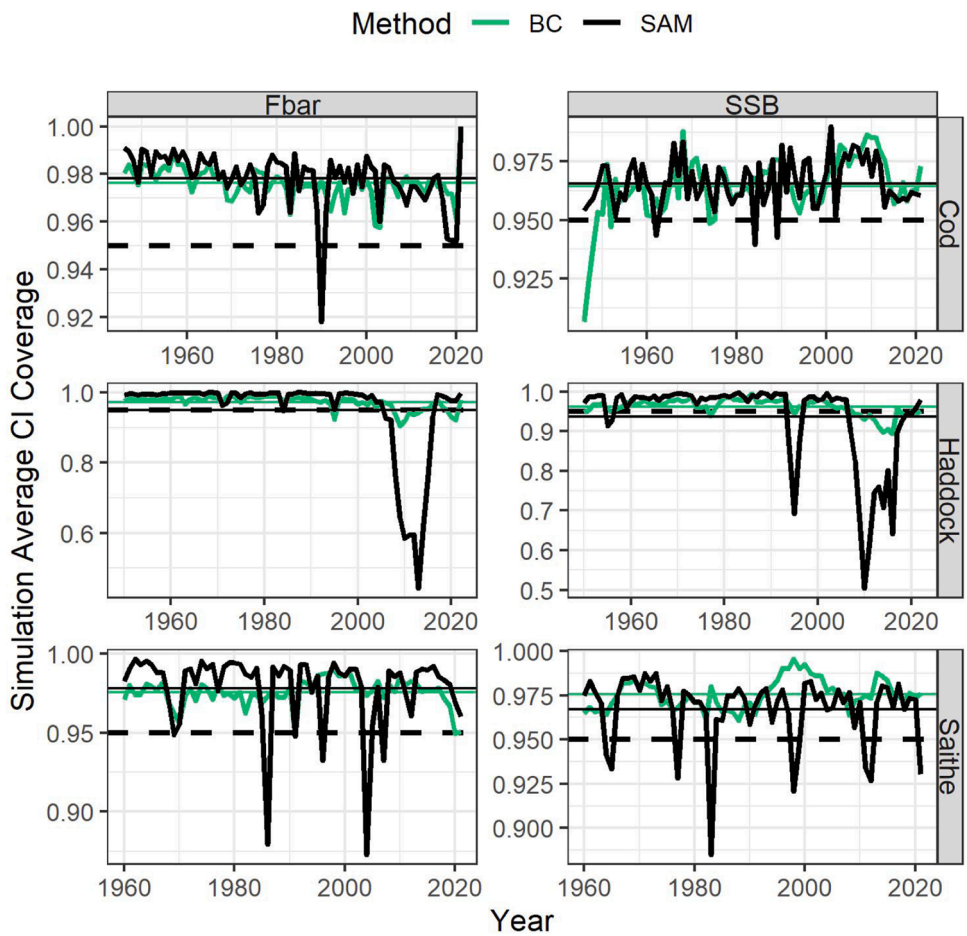

Fig. 4. Simulation coverages of 95% confidence intervals (CIs) versus year for Fbar and SSB (columns) and three stocks (rows) based on the bias corrected methodology (green lines) and marginal standard errors provided by SAM (black lines). Coverage is based on the fraction of the 2000 simulated CIs that contained the population values used to generate simulated data. Dashed lines indicate the nominal 0.95 level. Green and black lines indicate the average CI coverage across years.

Table 2
Simulated coverages of SAM and conditional bias-corrected (BC) 95% confidence intervals for the three case studies. Results are averaged over assessment model years.

	SSB		Fbar	
	SAM	BC	SAM	BC
Cod	0.966	0.965	0.978	0.976
Haddock	0.937	0.962	0.950	0.973
Saithe	0.967	0.976	0.979	0.975

results (Fig. 6) show that the BC CIs for Fbar were slightly more accurate than the SAM CIs on average (i.e., over years) for cod and saithe, but substantially more accurate for haddock. The SSB BC CIs were also substantially more accurate for haddock, but slightly less accurate for cod and saithe.

4. Discussion

We examined the reliability of standard errors (SEs) and confidence intervals (CIs) provided by SAM in the simulation self-test frequentist inferential setting in which random effects (REs) are fixed at their estimated values and variability is only considered in the distribution of the data conditional on the values of the REs. We clarified that SAM SEs are estimates of marginal (i.e., not conditioned on REs) root mean squared errors (RMSEs) and provide approximate estimates of conditional RMSEs. We also examined the reliability of conditional SEs using results from Zheng and Cadigan (2023). They provided good estimates of simulation self-test standard deviations (SDs) of SSB and average F (Fbar). SAM SSB and Fbar CIs had simulation coverage probabilities (CPs) that were usually slightly larger than 95%, but in a small number of years these CPs could be much smaller than 95%. The BC CIs proposed by Zheng and Cadigan (2024) were more reliable because their simulated CPs were never much different than 95%. However, when averaged over years, the SAM and BC CI CPs were similar for the cod and saithe case studies, but the BC CIs performed much better overall for haddock. CIs based on the conditional SEs in Zheng and Cadigan (2023) were not reliable because they do not account for the conditional bias.

SAM CIs are theoretically based on variability from resampling the model REs (including logN's and logF's) and the assessment data conditional on the REs. In a sense, the SAM CIs are designed to be accurate across different assessments. They are also more accurate when averaged across assessment years. However, for a specific stock and year, the SAM CIs can be considerably inaccurate. We caution fisheries managers and other users of stock assessment advice that in some years, SAM CIs may have simulation CPs substantially less than 95% and it is difficult to predict when this problem occurs. We suggest this is not a specific issue with SAM, but rather a general issue with state-space fish stock assessment models, mostly caused by the conditional bias.

Zheng and Cadigan (2023) demonstrated, for a simple random-walk state-space model, that the biases in estimates of the random-walk were like the smoothing bias that is a common feature of kernel and spline smoothers, where the smoothers have some bias attenuation towards the mean; that is, there is some under-estimation of the peaks and over-estimation of the valleys in the process being smoothed. Zheng and

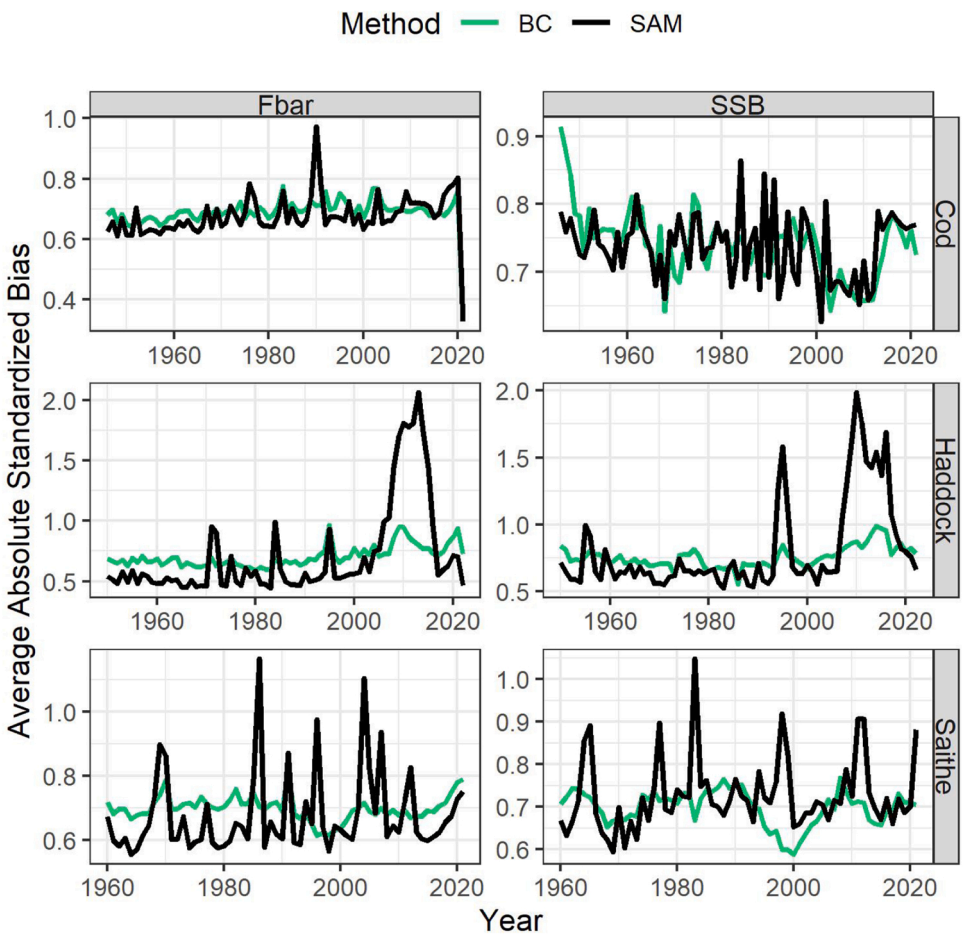

Fig. 5. Simulation average absolute standardized bias (log-estimate bias/SE) versus year for Fbar and SSB (columns) and three stocks (rows) based on the bias corrected methodology (green lines) and SAM (black lines).

Cadigan (2023) provided an analytical approximation of the bias (see their Equation 13) and showed that this gave a reasonably accurate approximation of the simulation bias for their random-walk example. However, SAM has two sets of REs, for stock size and for fishing mortalities. SAM is also fit to multiple data sets, all of which may conspire to produce small biases that are likely different and more complex than simple smoother biases. For the haddock case study, the SAM biases for Fbar and SSB were large in a few years around 2010, which led to substantially inaccurate CIs in those years. We are unsure why the bias was so large in these years. There were no unusual patterns in the assessment data that would indicate a bias issue. Diagnostics of when bias may be a problem is a useful area for future research. Zheng and Cadigan (2023) found that simple "plug-in" estimates of the bias did not lead to improved statistical inferences. Therefore they approximate the bias-squared term in the conditional MSE by averaging over REs, whereby the conditional MSE approximation is equal to the marginal MSE. Zheng and Cadigan (2024) addressed the bias issue by proposing a bias-corrected RE estimator and its corresponding conditional RMSE. This approach led to improved CIs compared to the ones currently provided by TMB and SAM, which are based on marginal RMSE.

The equations in Zheng and Cadigan (2023) assume that we are conditioning on RE values randomly drawn from their assumed distribution, and this will not be exactly correct for the simulation self-tests. Usually the estimated REs used in self-test data generation have less variability than the assumed RE distribution on which the marginal MSE is based on. An alternative is to sample the REs once from their "posterior" distribution conditional on the data (e.g., Thygesen et al., 2017). Although this distribution may still be substantially different than the assumed RE distribution, we expect that it can better capture the variability of the true REs compared to the RE estimators. For example, the REs that are not linked to data are estimated using values close to their marginal means, resulting in an underestimation of their variability. In contrast, REs generated from their posterior distribution reflect the variability of true REs and the information available for their inference. Improved simulations methods to test state-space model estimation requires further research.

CRediT authorship contribution statement

Cadigan Noel G: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Zheng Nan:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Albertsen Christoffer M:** Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Nielsen Anders:** Methodology, Software, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

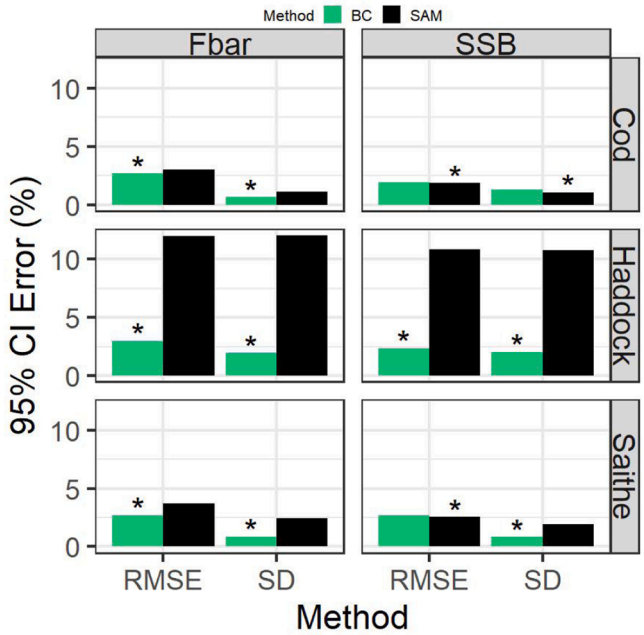

Fig. 6. Confidence interval (CI) errors, averaged over years. RMSE indicates the average squared difference (in %) between the simulated coverage and the nominal 0.95 level. SD is the standard deviation across years of the simulation coverage and indicates the consistency of the CI coverage. The method with the lowest RMSE or SD is denoted with a *.

Data Availability

Data will be made available on request.

Acknowledgements

Funding: Research funding to NC was provided by the Ocean Choice International Industry Research Chair program at the Marine Institute of Memorial University of Newfoundland. Research funding to NC and NZ was provided by the Ocean Frontier Institute, through an award from the Canada First Research Excellence Fund.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2024.106950.

References

Aanes, S., Engen, S., Sæther, B.E., Aanes, R., 2007. Estimation of the parameters of fish stock dynamics from catch-at-age data and indices of abundance: can natural and fishing mortality be separated? Can. J. Fish. Aquat. Sci. 64, 1130–1142.

Aeberhard, W.H., MillsFlemming, J., Nielsen, A., 2018. Review of state-space models for fisheries science. Annu. Rev. Stat. Its Appl. 5, 215–235.

Albertsen, C.M., Nielsen, A., Thygesen, U.H., 2018. Connecting single-stock assessment models through correlated survival. ICES J. Mar. Sci. 75, 235–244.

Berg, C.W., Nielsen, A., 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES J. Mar. Sci. 73, 1788–1797.

Brown, P.E., De Jong, P., 2001. Nonparametric smoothing using state space techniques. Can. J. Stat. 29, 37–50.

Cadigan, N.G., 2015. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Can. J. Fish. Aquat. Sci. 73, 296–308.

Das, K., Jiang, J., Rao, J., 2004. Mean squared error of empirical predictor. Ann. Stat. 32, 818–840.

Datta, G.S., Lahiri, P., 2000. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Stat. Sin. 613–627.

Deroba, J., Butterworth, D.S., Methot Jr, R., De Oliveira, J., Fernandez, C., Nielsen, A., Cadrin, S., Dickey-Collas, M., Legault, C., Ianelli, J., et al., 2015. Simulation testing the robustness of stock assessment models to error: some results from the ices strategic initiative on stock assessment methods. ICES J. Mar. Sci. 72, 19–30.

Flores-Agreda, D., Cantoni, E., 2019. Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Comput. Stat. Data Anal. 130, 1–17.

ICES, 2019a.Arctic Fisheries Working Group (AFWG). techreport 1:30. 934 pp. ICES Scientific Reports.10.17895/ices.pub.5292.

ICES, 2019b.North Western Working Group (NWWG). techreport 1:14. 830 pp. ICES Scientific Reports.10.17895/ices.pub.5298.

ICES, 2020.Report of the arctic fisheries working group (afwg) 2, 577.10.17895/ices.pub.6050.

Kackar, R.N., Harville, D.A., 1984. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. J. Am. Stat. Assoc. 79, 853–862.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M., 2016. TMB: automatic differentiation and Laplace approximation. J. Stat. Softw. 70, 1–21. https://doi.org/10.18637/jss.v070.i05.

Liljestrand, E.M., Bence, J.R., Deroba, J.J., 2023. Applying a novel state-space stock assessment framework using a fisheries-dependent index of fishing mortality. Fish. Res. 264, 106707 https://doi.org/10.1016/j.fishres.2023.106707. ⟨https://www.sciencedirect.com/science/article/pii/S0165783623001005⟩.

Marra, G., Wood, S.N., 2012. Coverage properties of confidence intervals for generalized additive model components. Scand. J. Stat. 39, 53–74.

McClave, J.T., Sincich, T.T., 2017. Statistics. Pearson Higher Ed.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101.

Perreault, A.M., Wheeland, L.J., Morgan, M.J., Cadigan, N.G., 2020. A state-space stock assessment model for american plaice on the grand bank of newfoundland. J. North. Atl. Fish. Sci. 51, 45–104.

Punt, A.E., Dunn, A., Elvarsson, B.Þ., Hampton, J., Hoyle, S.D., Maunder, M.N., Methot, R.D., Nielsen, A., 2020. Essential features of the next-generation integrated fisheries stock assessment package: a perspective. Fish. Res. 229, 105617.

R Core Team, 2018.R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.⟨https://www.R-project.org/⟩.

Schnute, J.T., 1994. A general framework for developing sequential fisheries models. Can. J. Fish. Aquat. Sci. 51, 1676–1688.

Skaug, H.J., Fournier, D.A., 2006. Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. Comput. Stat. Data Anal. 51, 699–709.

Stock, B.C., Miller, T.J., 2021. The woods hole assessment model (wham): a general state-space assessment framework that incorporates time-and age-varying processes via random effects and links to environmental covariates. Fish. Res. 240, 105967.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K., Nielsen, A., 2017. Validation of ecological state space models using the laplace approximation. Environ. Ecol. Stat. 24, 317–339.

Wand, M.P., 2003. Smoothing and mixed models. Comput. Stat. 18, 223–249.

Wood, S.N., 2020. Inference and computation with generalized additive models and their extensions. Test 29, 307–339.

Wood, S.N., Scheipl, F., Faraway, J.J., 2013. Straightforward intermediate rank tensor product smoothing in mixed models. Stat. Comput. 23, 341–360.

Zheng, N., Cadigan, N., 2021. Frequentist delta-variance approximations with mixed-effects models and tmb. Comput. Stat. Data Anal. 160, 107227.

Zheng, N., Cadigan, N., 2023. Frequentist conditional variance for nonlinear mixed-effects models. J. Stat. Theory Pract. 17, 3.

Zheng, N., Cadigan, N., 2024. Improved confidence intervals for nonlinear mixed-effects and nonparametric regression models. Submitted.