Canadian Journal of Fisheries and Aquatic Sciences

OPEN ACCESS | Article

# Detecting significant retrospective patterns in state space fish stock assessment

**Olav Nikolai Breivik**[a], **Magne Aldrin**[a], **Edvin Fuglebakk**[b], **and Anders Nielsen**[c]

[a]Norwegian Computing Center, Gaustadaleen 23A, 0373 Oslo, Norway; [b]Institute of Marine Research, Bergen, Norway; [c]National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet, 2800 Kgs. Lyngby, Denmark

Corresponding author: **Olav Nikolai Breivik** (email: Olavbr@nr.no)



## Abstract

Retrospective patterns are commonly investigated to validate fish stock assessment models. A widely applied measure for retrospective bias is Mohn's $\rho$ and corresponding retrospective plots. However, retrospective patterns can be interpreted differently by experts. To make decisions regarding significant retrospective patterns less subjective, we proposed a post-sample Mohn's $\rho$ significance test. As case studies, we applied the state space assessment model SAM with data on Northeast Arctic cod and Norwegian coastal cod north of 67°N. We showed that the acceptance regions of Mohn's $\rho$ depends on both the data available and the assessment model complexity. We also assessed the test power under a range of assumption violations and conclude that Mohn's $\rho$ is useful for detecting violations associated with bias, but not for violations associated with variances and correlations.

**Key words:** Mohn's $\rho$, validation, SAM, post-sample evaluation

## 1. Introduction

A retrospective pattern in fish stock assessment is a systematic inconsistency in an estimate, e.g., spawning stock biomass (SSB) or fishing mortality, in recent years of an assessment. Such inconsistencies indicate model misspecification. In particular, they may indicate that parameters have been systematically over- or underestimated in some consecutive years. Retrospective patterns often occur when a model parameter is assumed constant, while it is in fact time-varying in the true system. Detecting this kind of error is crucial for producing sustainable quota advice, e.g., a systematic overestimation of current SSB may cause quota advice that will make the fishery collapse. Mohn's $\rho$ (Mohn 1999) is a widely applied measure for retrospective patterns (ICES 2019, 2020*a*, 2020*b*, 2020*c*, 2020*d*, 2020*e*, 2021*a*, 2021*b*; Legault 2020; Carvalho et al. 2021; Kell et al. 2021) Historic estimates and therefore Mohn's $\rho$, are however, subjected to randomness in data, and relatively little is known about the statistical distribution of Mohn's $\rho$.

Hurtado-Ferro et al. (2015) used a simulation study and suggested an acceptable SSB Mohn's $\rho$ interval $(-0.15, 0.2)$ for medium and long-lived species. This interval is widely used in practice as a rule of thumb when validating assessments (ICES 2020*c*, 2020*d*, 2020*e*, 2021*b*; Carvalho et al. 2021; Kell et al. 2021). However, the distribution of Mohn's $\rho$ depends on the unknown true data-generating mechanism and the assessment model applied to the generated data. The data-generating mechanism here is the combination of the process generating the latent population and the mechanism generating the data given the population. Therefore, it is essential to evaluate the significance of the Mohn's $\rho$ in the context of the specific assessment.

Post-sample evaluation is a leave-out procedure for evaluating whether recent data are inconsistent with historic data and an assumed model (Harvey 1990, section 5.6). In post-sample evaluation, data are divided into two subsets, (1) data prior to the post-sample period are used to estimate the model and (2) data within the post-sample period are used to define a test statistic with known distribution under the model. The distribution of the test statistic is obtained analytically or approximated using a parametric bootstrap (Efron and Tibshirani 1994; Davison and Hinkley 1997). In this research, we introduced a Mohn's $\rho$ post-sample significance test by defining the years within the retrospective analysis as the post-sample period and using Mohn's $\rho$ as the test statistic. Mohn's $\rho$ is used as the test statistic because it is widely applied in practice. The null distribution of Mohn's $\rho$ is calculated using the parametric bootstrap. The resulting acceptance regions are tailored to the data and the complexity of the assessment model used, and can be wider or smaller than the standard interval suggested by Hurtado-Ferro et al. (2015). To illustrate the proposed Mohn's $\rho$ significance test, we applied the state space assessment model SAM (Nielsen and Berg 2014; Berg and Nielsen 2016; Breivik et al. 2021) on two case studies, Northeast Arctic (NEA) cod and Norwegian coastal cod north of 67°N (NCC cod).

We further seek to determine the power of a Mohn's $\rho$ test for detecting model misspecifications and data issues, i.e., the

probability for detecting an assumption violation given that it is present. The null distribution of Mohn's $\rho$ in the proposed significance test is constructed using the parametric bootstrap, and the distribution under a given assumption violation can be obtained by modifying the bootstrap accordingly. By comparing the distributions of Mohn's $\rho$ with and without misspecification, we can quantify the corresponding power of the proposed significance test in the hypothetical case of a single violation of the null hypothesis.

Miller and Legault (2017) used a parametric bootstrap procedure to estimate the distribution of Mohn's $\rho$ by sampling from distributional assumptions about the observations. Our approach differs in that we bootstrap both the latent processes, i.e., fishing mortality, recruitment and survival, and observations conditioned on the simulated processes. By sampling both the latent processes and observations, we bootstrap all random variables in the model and provide the distribution of Mohn's $\rho$ in the post-sample significance test. State space assessment models have been increasingly popular in fisheries science (Nielsen and Berg 2014; Cadigan 2015; Miller et al. 2016; Aanes 2016; Berg and Nielsen 2016; Perreault et al. 2020; Breivik et al. 2021; Newman et al. 2022), and the proposed procedure provides the distributions of Mohn's $\rho$ tailored to the data and complexity of the applied model. Retrospective patterns can be interpreted differently by experts, and the motivation of our research is to reduce the subjectivity of decisions regarding significant retrospective patterns when applying state space assessment models.

## 2. Materials and methods

### 2.1. Model

In our research, we used the state space assessment model SAM (Nielsen and Berg 2014; Berg and Nielsen 2016; Breivik et al. 2021). SAM incorporates standard stock equations and includes the abundance (**N**) and fishing mortality (**F**) as latent effects (states). The applied state space framework automatically weights data sources on how well the data fit into the population dynamic structure, and by applying the estimated processes, we can propagate the states into the future to produce fish quota advice.

We will now define SAM mathematically. Let

$$\text{(1a)}\quad \log N_{a_0,y} = \log R\left(\mathbf{N}_{y-1}\right) + \eta_{a_0,y}$$

$$\text{(1b)}\quad \log N_{a,y} = \log N_{a-1,y-1} - F_{a-1,y-1} - M_{a-1,y-1} + \eta_{a,y} \quad \text{when } a_0 < a < A$$

$$\text{(1c)}\quad \log N_{A,y} = \log\left(N_{A-1,y-1}e^{-F_{A-1,y-1}-M_{A-1,y-1}} + N_{A,y-1}e^{-F_{A,y-1}-M_{A,y-1}}\right) + \eta_{A,y}$$

Here, $R$ is a recruitment function and $M_{a,y}$ is the assumed known natural mortality rate at age $a$ in year $y$. The age span is defined from $a_0$ to $A$, where all fish of age $A$ or older are included in the oldest age group (plus group). Furthermore, $\eta_{a,y}$ is assumed independent mean zero Gaussian distributed, and typically, there are separate variance parameters for the recruitment ($\eta_{a_0,y}$) and the survival process ($\eta_{a,y}$ when $a > a_0$). The fishing mortality vector $\boldsymbol{F}_y = \{F_{1,y},..., F_{A,y}\}$ is assumed to follow a random walk as in Nielsen and Berg (2014); Berg and Nielsen (2016); Breivik et al. (2021).

$$\text{(2)}\quad \log \boldsymbol{F}_y = \log \boldsymbol{F}_{y-1} + \boldsymbol{\xi}_y^F$$

Here, $\boldsymbol{\xi}_y^F$ is mean zero multivariate Gaussian distributed, and it is often assumed a first-order autoregressive structure in the age dimension (Berg and Nielsen 2016).

The model is fit to data time series of estimated commercial catch at age ($C_{a,y}$) and survey indices at age ($I_{a,y}$). These time series inform the system through standard observation equations,

$$\text{(3a)}\quad \log C_{a,y} = \log\left(\frac{F_{a,y}}{F_{a,y}+M_{a,y}}\left(1-\mathrm{e}^{-F_{a,y}-M_{a,y}}\right)N_{a,y}\right) + \varepsilon_{a,y}^{(C)}$$

$$\text{(3b)}\quad \log I_{a,y}^{(S)} = \log\left(Q_a^{(S)}\mathrm{e}^{-\left(F_{a,y}+M_{a,y}\right)\mathrm{day}^{(S)}/365}N_{a,y}\right) + \varepsilon_{a,y}^{(S)}$$

Here, $Q_a^{(S)}$ is the unknown survey catchability at age $a$ for survey $S$, and $\mathrm{day}^{(S)}$ is the number of days into the year when the survey is typically half done. Furthermore, $\boldsymbol{\varepsilon}^{(C)}$ and $\boldsymbol{\varepsilon}^{(S)}$ are mean zero multivariate Gaussian distributed. Note that fleet specific age plus groups may be included in the observation equations. In our case studies we included a total stock biomass index ($I_{\mathrm{TSB},y}$). $I_{\mathrm{TSB},y}$ is assumed to be proportional to the true total stock biomass with noise, and we included it in the model by

$$\text{(4)}\quad \log I_{\mathrm{TSB},y} = \log\left(Q_{\mathrm{TSB}}\sum_{a=a_0}^{A} w_{a,y}\mathrm{e}^{-\left(F_{a,y}+M_{a,y}\right)\mathrm{day}/365}N_{a,y}\right) + \varepsilon_{\mathrm{TSB},y}$$

Here, $Q_{\mathrm{TSB}}$ is the unknown survey catchability, $w_{a,y}$ is the assumed known mean stock weight at age $a$ in year $y$, and $\epsilon_{\mathrm{TSB},y}$ is mean zero Gaussian distributed.

### 2.2. Mohn's $\rho$

Denote $Y$ as the terminal year of an assessment, i.e., the most recent year with data. Mohn's $\rho$ of an estimated quantity, e.g., SSB, was introduced in Mohn (1999) and is defined as

$$\text{(5)}\quad \rho = \frac{1}{n}\sum_{y=y_0}^{Y-1}\frac{\widehat{X}_{y|y}-\widehat{X}_{y|Y}}{\widehat{X}_{y|Y}}$$

Here, $n$ is the number of years sequentially removed in the retrospective analysis, $\widehat{X}_{y|Y}$ is the estimate at year $y$ using all available data in year $Y$, $\widehat{X}_{y|y}$ is the estimate using data available in year $y$ (i.e., when peeling off the $Y-y$ last years of data), and $y_0 = Y-n$. Typically, $n$ is set to 5 in working groups arranged by the International Council for the Exploration of the Sea (ICES). An often applied rule of thumb is that retrospective patterns should be addressed if Mohn's $\rho$ (eq. 5) for SSB is outside of the interval $(-0.15, 0.2)$ for medium and long-lived species (Hurtado-Ferro et al. 2015; Kell et al. 2021).

Data on catch in the terminal year are sometimes not available, resulting in an estimated fishing mortality in the terminal year that is very uncertain. In such scenarios, we followed common practice in ICES to relabel $\widehat{X}_{y|y}$ and $\widehat{X}_{y|Y}$ in eq. 5 to be estimates in year $y-1$ and $Y-1$ given data available in year $y$ and $Y$, respectively, when referring to fishing mortality. This is the practice for NEA cod and haddock (ICES 2020*a*).

## 2.3. Inference

The model is estimated by maximum likelihood using the R-package template model builder (TMB) (Kristensen et al. 2016) combined with the optimization routine *nlminb* (Core-Team and contributors worldwide 2022). TMB utilizes automatic differentiation and efficiently integrates over latent effects with the Laplace approximation by utilizing Markov structures. *nlminb* further utilizes the gradient provided by TMB for quasi-Newton optimization of the marginal likelihood. The latent effects are the log abundance ($\log \mathbf{N}$) and log fishing mortality ($\log \mathbf{F}$). Our inference procedure is identical to Nielsen and Berg (2014); Berg and Nielsen (2016); and Breivik et al. (2021).

## 2.4. Post-sample Mohn's $\rho$ significance test

The goal in diagnostics for model fitting is to determine whether the observed data are representative of the data expected given an assumed model (Cressie and Wikle 2011, page 40). In particular, Mohn's $\rho$ provides insight about whether the data in the $n$ most recent years systematically shifted important estimates in a way that is inconsistent with the model conditioned on data prior to the post-sample period. The proposed test investigates whether the sequential inclusion of the post-sample data results in a significant retrospective pattern under the null hypothesis (H0): *The assessment model estimated with data prior to the post-sample period generated the post-sample data.* We let Mohn's $\rho$ be the test statistic and provide the distribution of Mohn's $\rho$ under H0. In general, leaving out recent data and applying them to construct a test statistic with known distribution, given the remaining data are known as post-sample evaluation (Harvey 1990, section 5.6). We therefore refer to the proposed Mohn's $\rho$ test as the post-sample Mohn's $\rho$ significance test. The distribution of Mohn's $\rho$ under H0 and the corresponding significance level are constructed using the parametric bootstrap (Efron and Tibshirani 1994; Davison and Hinkley 1997). By comparing the assessment value of Mohn's $\rho$ with it's distribution under H0, we can objectively determine whether the post-sample data resulted in a Mohn's $\rho$ that significantly signals an inconsistency with respect to the model conditioned on data prior to the post-sample period, or if it was likely caused by randomness given the model state in year $y_0$. The proposed post-sample significance test is a special case of a parametric bootstrap test (Davison and Hinkley 1997, page 148).

We are interested in the distribution of Mohn's $\rho$ conditioned on the data prior to the post-sample period under H0, i.e., the distribution of $\rho\left(\mathbf{D}_{\text{new}}|\mathbf{D}_{y_0}, \text{H0}\right)$ where $\mathbf{D}_{\text{new}}$ is (random) post-sample data, $\mathbf{D}_{y_0}$ is the set of observations available in year $y_0$ and $\rho(\cdot)$ maps the data to Mohn's $\rho$ value given the applied assessment model. The distribution of $\rho\left(\mathbf{D}_{\text{new}}|\mathbf{D}_{y_0}, \text{H0}\right)$ is unavailable analytically, so we obtained it using the parametric bootstrap. The proposed parametric bootstrap procedure generates realizations of $\mathbf{D}_{\text{new}}$ by first sampling $\boldsymbol{N}_{y_0} = \left\{N_{a_0,y_0}, \cdots, N_{A,y_0}\right\}$ and $\boldsymbol{F}_{y_0} = \left\{F_{a_0,y_0}, \cdots, F_{A,y_0}\right\}$ given $\mathbf{D}_{y_0}$, then utilizing the Markov properties of the latent effects to project the population dynamics into the future with eqs. 1 and 2, and finally sampling $\mathbf{D}_{\text{new}}$ with observation eqs. 3 and 4. We will now define the bootstrap procedure in detail. Let $\log \widehat{\boldsymbol{N}}_{y_0|y_0}$ and $\log \widehat{\boldsymbol{F}}_{y_0|y_0}$ be estimated log abundance and fishing mortality in year $y_0$ given $\mathbf{D}_{y_0}$. Denote $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y_0}|\boldsymbol{y_0}}$ as the corresponding estimated joint covariance matrix. The bootstrap procedure is defined as follows:

(i) Simulate $\left(\log \boldsymbol{N}_{y_0}^{(b)}, \log \boldsymbol{F}_{y_0}^{(b)}\right)$ from $N\left(\left(\log \widehat{\boldsymbol{N}}_{y_0|y_0}, \log \widehat{\boldsymbol{F}}_{y_0|y_0}\right), \widehat{\boldsymbol{\Sigma}}_{y_0|y_0}\right)$ for bootstrap sample $b$.
(ii) Given $\left(\boldsymbol{N}_{y_0}^{(b)}, \boldsymbol{F}_{y_0}^{(b)}\right)$, simulate the processes up to year $Y$ by applying eqs. 1 and 2. One exception is included for recruitment and is elaborated below.
(iii) Simulate observations in the post-sample period by applying eqs. 3 and 4 and the simulated processes.
(iv) Fit SAM with observations available prior the post-sample period ($\mathbf{D}_{y_0}$) and the simulated observations ($\mathbf{D}_{\text{new}}^{(b)}$).
(v) Calculate Mohn's $\rho$ (eq. 5).
(vi) Repeat the steps (i)–(v) a large number ($B$) of times.

Denote $q_\alpha$ as the $\alpha$-quantile of the parametric bootstrap. If Mohn's $\rho$ falls outside of the bootstrap prediction interval $(q_{\alpha/2}, q_{1-\alpha/2})$, the post-sample data updated the estimate significantly inconsistent with respect to the model conditioned on data prior to the post-sample period. We set $B = 1000$ in our research, and we further denote the $(1-\alpha)\,100\%$ prediction interval of Mohn's $\rho$ as a $(1-\alpha)\,100\%$ acceptance interval.

The recruitment function in our case studies follows a random walk and is typically associated with a large variance. A random walk with high dispersion will predict recruitment poorly a few years forward in time. To produce realistic recruitment, we sample it with replacement from the set of all previously estimated recruitment. Note that similar recruitment sampling procedures are commonly applied when providing quota advice (e.g., see ICES 2021*a*).

## 2.5. Mohn's $\rho$ distribution with model misspecifications

We can obtain the distribution of Mohn's $\rho$ for a specific misspecification in the post-sample period by modifying the parametric bootstrap described in Section 2.4. For example, the uncertainty of commercial catch observations in the post-sample period can be modified in step iii. In our research, we calculated the test power in eight different scenarios, each with *one* violation of H0. Table 1 provides information about these scenarios. They are constructed by modifying steps ii and iii in Section 2.4, meaning that the reality within the simulation study is modified and *not* the assessment model assumptions. All scenarios, except for one, are constructed such that the misspecification gets more severe linearly in time, e.g., in scenario M3, the natural mortality is increased with a factor $\frac{3}{5}$ in year $y_0 + 1$ and so on until it is increased with a factor of 3 in year $Y$. For each scenario, we calculated the average number of Mohn's $\rho$ samples outside the post-sample Mohn's $\rho$ acceptance region. This average is an approximation of the test power.

**Table 1.** Description of misspecification scenarios.

| Scenario | Description | Linear change |
|---|---|---|
| M3 | Natural mortality scaled with a factor of 3 | Yes |
| M/3 | Natural mortality scaled with a factor of 1/3 | Yes |
| Q3 | Survey catchabilities scaled with a factor of 3 | Yes |
| C/3 | Catches are reported with a scaling factor of 1/3 | Yes |
| sdF | Variance of log **F** increments scaled with a factor of 3 | Yes |
| SdI | Variance of index observations scaled with a factor of 3 | Yes |
| sdC | Variance of catch observations scaled with a factor of 3 | Yes |
| corObs | Correlation structure in observations are removed | No |

**Note:** Misspecifications are included only in the five most recent years including terminal year.



## 2.6. Retrospective plot

Retrospective plots are commonly investigated to identify retrospective patterns. A retrospective plot shows the estimated quantity using all data and corresponding estimates obtained when sequentially removing the last year's data. In our research, we included prediction intervals on how estimates change in the retrospective analysis under H0, which can be used to determine whether an estimate is significantly modified. To create these intervals, we defined $d_y^{(b)}$ as the difference between the estimated quantity using all data and the corresponding estimate obtained when removing the last year's data for bootstrap sample $b$, i.e., $d_y^{(b)} = \widehat{X}_{y|y}^{(b)} - \widehat{X}_{y|Y}^{(b)}$. Let $d_{y,\alpha}$ be the corresponding $\alpha$-quantile and note that the interval $(d_{y,\alpha/2}, d_{y,1-\alpha/2})$ is a (1-$\alpha$)100% prediction interval for how much the estimate in year $y$ changes in the retrospective analyses under H0. To objectively investigate whether an estimate changes significantly, we included the prediction interval

$$PI_y = \left(\widehat{X}_{y|Y} - d_{y,\alpha/2}, \widehat{X}_{y|Y} + d_{y,1-\alpha/2}\right) \quad (6)$$

in our retrospective plots. If $\widehat{X}_{y|y}$ falls outside of this interval, we conclude that the estimate in year $y$ changed significantly in the retrospective analysis. Note that the prediction intervals are for estimates within individual years, and not for systematic patterns. ICES (2020*e*) recommends investigating estimates from each step in the retrospective analysis closely even if there are no patterns. With the proposed procedure, we can determine whether each estimate in the retrospective analysis is significantly different from the terminal estimate. However, one should be cautious when selecting the significance level because we performed several tests.

# 3. Results

We applied our procedure to NEA cod and NCC cod case studies. Both stocks are officially assessed with SAM (ICES 2021*a*, 2022). During the most recent benchmark for NEA cod (ICES 2021*a*), flexibility in the latent fishing mortality process was reduced because it removed retrospective patterns as determined by inspection of retrospective plots. This makes NEA cod ideal as a case study. The NCC cod assessment has much less data compared to the NEA cod assessment and is included because we want to highlight that Mohn's $\rho$ acceptance intervals should be different for different data sources. In our research, we sequentially removed 5 years of data in our retrospective analysis, i.e., let $n = 5$ in eq. 5 as applied at the recent NEA cod benchmark (ICES 2021*a*).

Table 2 provides a short description of the data used. A detailed description of these data and model configurations for NEA cod and NCC cod are provided in the corresponding benchmark reports (ICES 2021*a*, 2022). In the official NEA cod assessment, stomach data are utilized to include cannibalism in the natural mortality by running SAM sequentially. For simplicity, we do not apply this sequential procedure, but assume the same natural mortality as applied in the final SAM run (ICES 2021*a*).

In our case study, we calculated Mohn's $\rho$ for SSB, recruitment, and average fishing mortality ($\bar{F}$) for ages 5–10 for NEA cod and 4–8 for NCC cod. These age ranges are the same as those used in the official assessments. Quota advice is based on targeting $\bar{F}$ at certain levels depending on SSB, and these quantities are therefore especially important to be estimated with consistency.

Figure 1 illustrates 95% acceptance intervals of Mohn's $\rho$ for NEA and NCC cod assessments. The green lines in Fig. 1 illustrate the rule of thumb acceptable SSB Mohn's $\rho$ intervals (Hurtado-Ferro et al. 2015; Kell et al. 2021). For NEA cod, the 95% acceptance interval based on the proposed post-sample Mohn's $\rho$ test is narrower than what is suggested in Hurtado-Ferro et al. (2015), while for NCC cod the interval is wider. The red points in Fig. 1 illustrate Mohn's $\rho$ with corresponding $P$ values, which are obtained by doubling the one sided $P$ value. Figure 2 illustrates SSB retrospective plots that include 95% prediction intervals for how much the estimates are modified in the retrospective analysis under H0 (see Section 2.6). Note that the median Mohn's $\rho$ for recruitment of NCC cod deviates clearly from zero in Fig. 1*b*. This difference is because the recruitment of NCC cod is estimated to be relatively high in year $y_0$ and the random walk structure in the recruitment is neglected in the bootstrap.

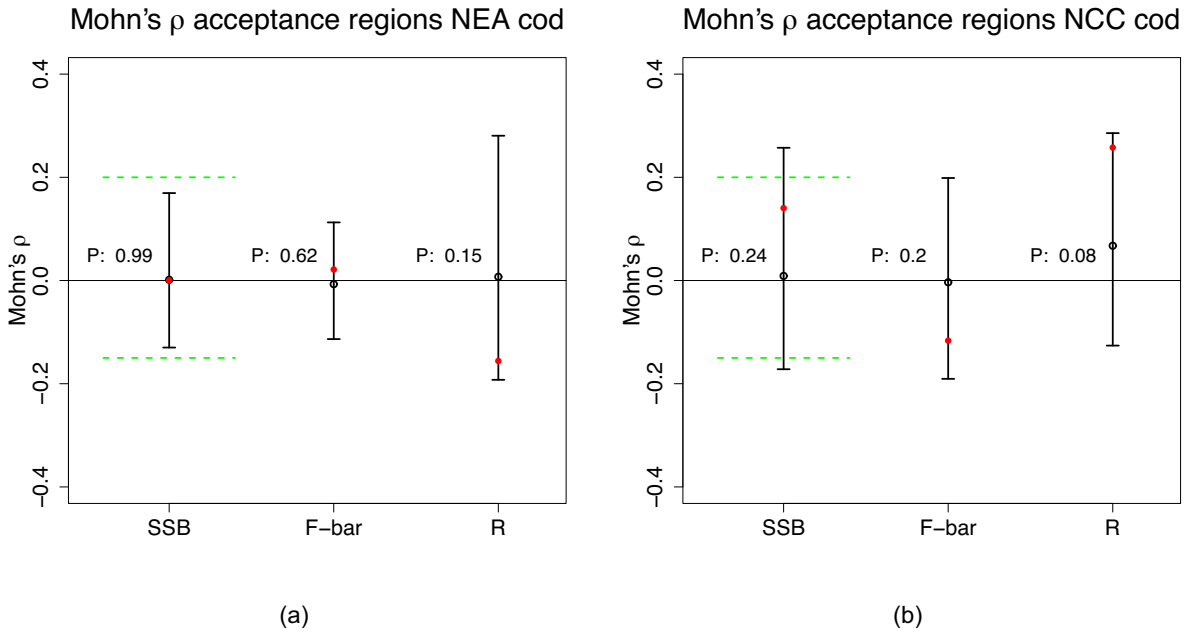
Our research is motivated by model adjustments at the most recent benchmark for NEA cod (ICES 2021*a*). At this benchmark, a correlation structure in the fishing mortality increments (see eq. 2) was removed from the assessment model based on fishing mortality retrospective patterns (ICES 2021*a*). Figure 3*a* illustrates Mohn's $\rho$ along with acceptance intervals when including the first-order autoregressive correlation structure between ages that was discarded at the benchmark. Mohn's $\rho$ for fishing mortality has a $P$ value of 0.1 and is therefore not highly significant. The correlation structure was, however, not discarded based on Mohn's $\rho$ but based on visualization of retrospective patterns (ICES 2021*a*). Figure 3*b* shows the retrospective plot for $\bar{F}$. We con-

**Table 2.** Short summary of NEA (top) and NCC (bottom) cod data.

| Data source | Years | Short description |
|---|---|---|
| Catch at age | 1946–2019 | Catch at age estimates from coastal states harvesting the stock |
| Survey 1 | 1981–2020 | Joint Norwegian–Russian Winter survey, swept-area index (Mehl et al. 2016) |
| Survey 2 | 1985–2020 | Joint Norwegian–Russian Winter survey, acoustic index (Mehl et al. 2016) |
| Survey 3 | 1982–2017 | Russian survey, ended permanently in 2017 |
| Survey 4 | 2004–2019 | Joint Norwegian–Russian Ecosystem survey, swept-area index (Johannesen et al. 2019) |
| Catch at age | 1994–2020 | Catch at age estimates from Norwegian fisheries, provided with ECA model (Hirst et al. 2012) |
| Survey 1 | 2003–2020 | Norwegian coastal survey, swept-area abundance-at-age index (Aglen et al. 2021) |
| Survey 2 | 1995–2020 | Norwegian coastal survey, acoustic total stock biomass index (Aglen et al. 2021) |

**Note:** All NEA cod indices are provided by age.

**Fig. 1.** (*a*) Estimated 95% acceptance regions for Mohn's $\rho$ for NEA and (*b*) NCC cod. Green lines illustrate the rule of thumb acceptance region (−0.15, 0.2) for SSB. Red points show obtained Mohn's $\rho$ in assessments and numbers provide *P* values. Black solid points show median Mohn's $\rho$.




clude that all estimates of $\bar{F}$ within the post-sample period are reduced when including the post-sample data, but it is difficult to determine whether the pattern is significant. Figure S1 illustrates 24 retrospective plots for $\bar{F}$ obtained in the parametric bootstrap (in step v) and provides an indication on how we can expect retrospective patterns to look like if the model has generated the post-sample data. As indicated by the close to significant Mohn's $\rho$, the observed retrospective pattern in Fig. 3*b* is at the borderline of what to expect, and we consider it to be only a weak indication of model misspecification.

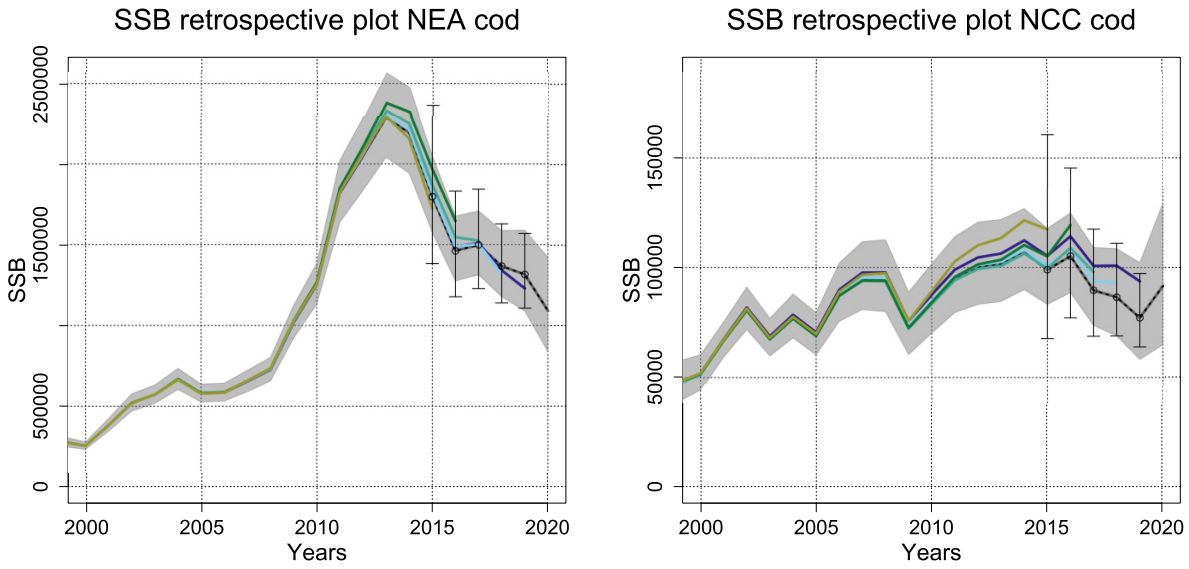
Figure 4 shows 95% prediction intervals for Mohn's $\rho$ and test powers for the assumption violation scenarios elaborated in Table 1. The case specific power shown in Fig. 4 is to be interpreted as the probability of detecting a significant Mohn's $\rho$ (significance level 0.05) when the only violation of the null hypothesis is given by the scenarios elaborated in Table 1. For example, Figs. 4*a* and 4*d* illustrate that there is approximately 88% and 97% probability for detecting significant SSB Mohn's $\rho$ in scenario M3 for NEA cod and NCC cod, respectively.

Convergence problems may occur when modifying input data. For NEA cod, we did not have convergence problems in our research. However, for NCC cod, approximately 4% of the simulated data sets resulted in convergence problems when calculating Mohn's $\rho$. All bootstrapped data sets that resulted in convergence problems were discarded and replaced with a new sample.

## 4. Discussion

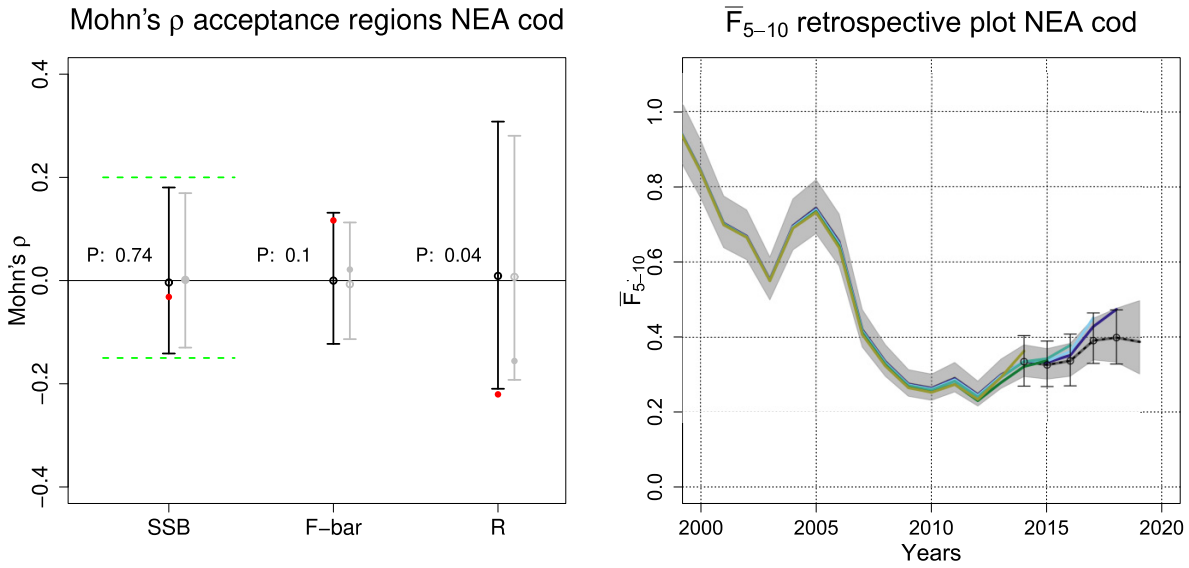
In this research, we introduced a post-sample significance test for Mohn's $\rho$ and constructed corresponding acceptance intervals using the parametric bootstrap. A Mohn's $\rho$ outside of the corresponding acceptance region implies that the post-sample data updated the estimate of interest significantly inconsistent with respect to the model conditioned on data prior to the post-sample period, and the Mohn's $\rho$ is therefore indicative of model or data misspecifications. We illustrated that the post-sample acceptance regions depend on the data and complexity of the assessment model, and the accep-

**Fig. 2.** (*a*) Retrospective SSB plots for NEA and (*b*) NCC cod. Solid black lines illustrate SSB estimated with all data and shaded areas are corresponding 95% confidence intervals. Colored lines show estimates in the retrospective analysis. Vertical intervals are 95% prediction intervals for how much the estimates are modified in the retrospective analysis under H0 (see eq. 6).




(a) (b)

**Fig. 3.** (*a*) Black intervals show 95% acceptance intervals for NEA cod Mohn's $\rho$ using the model with correlation structure in fishing mortality increments. Red points illustrate Mohn's $\rho$, corresponding $P$ values are given to the left of each interval, and the green line provides the ICES rule of thumb acceptance region for SSB. Grey intervals illustrate 95% acceptance intervals in the official assessment (same as in Fig. 1*a*). (*b*) Retrospective $\bar{F}$ plot for NEA cod using the model with correlation structure in fishing mortality increments, see Fig. 2 for general figure description.

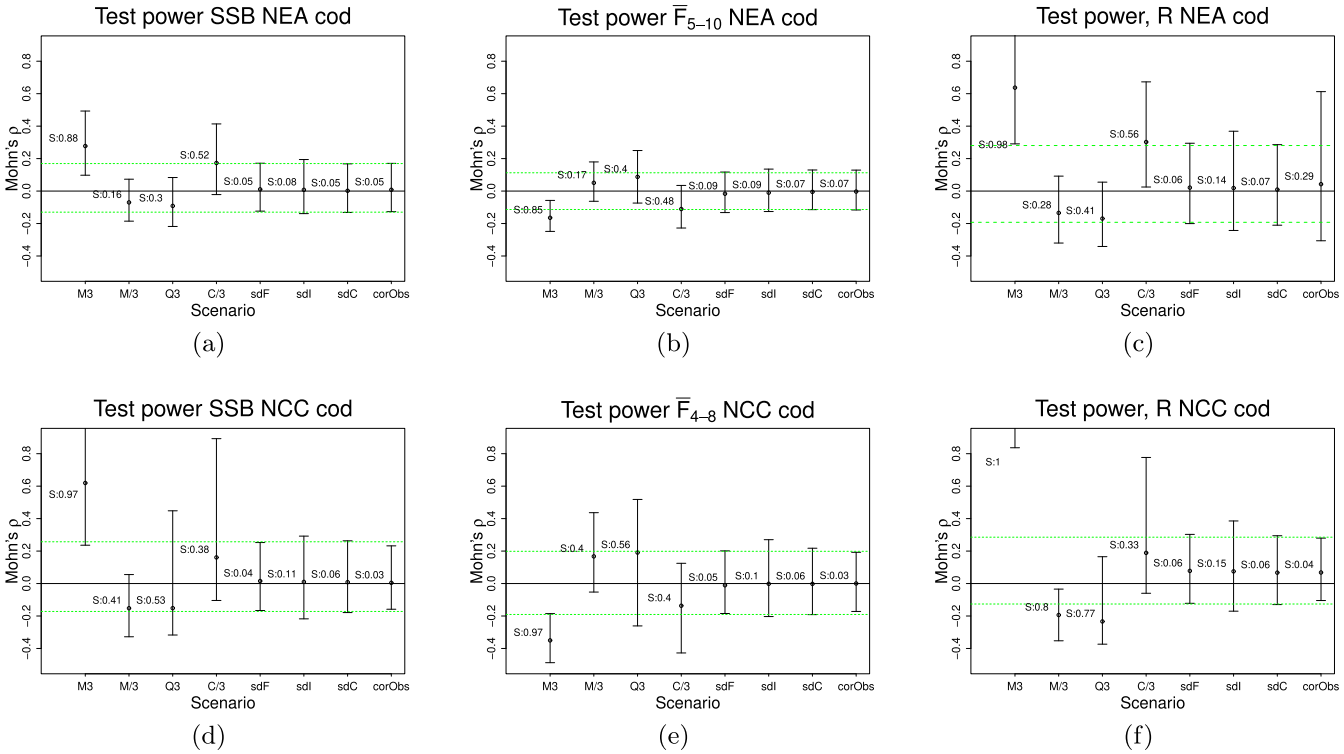



(a) (b)

tance regions can be both wider and narrower than the rule of thumb proposed by Hurtado-Ferro et al. (2015). If Mohn's $\rho$ falls within the corresponding acceptance interval, it provides no significant evidence of inconsistency in the post-sample data with respect to the estimated model prior to the post-sample period.

We demonstrated that the post-sample Mohn's $\rho$ significant test is able to detect model misspecification. Figure 4 illustrates test power when the only violation of the null hypothesis is specified in Table 1. We observed that the power varies between data sets. The only scenarios that introduced clear shifts in median Mohn's $\rho$ are those that introduce bias

**Fig. 4.** Estimated power of the post-sample Mohn's $\rho$ test with significance level 0.05. (*a*), (*b*), and (*c*) are for NEA cod and (*d*), (*e*), and (*f*) are for NCC cod. Green horizontal lines illustrate 95% acceptance intervals. Vertical intervals and points show 95% prediction intervals of Mohn's $\rho$ and median Mohn's $\rho$ under given misspecification.




in the quantity of interest, i.e., changes in natural mortality, systematic misreporting of catch, and change in index survey catchability. The scenarios that did not produce changes in median Mohn's $\rho$ are those scenarios related to changes in variance and correlation parameters, i.e., change in observation variances, process variance, and observation correlation. This shows that Mohn's $\rho$ is useful for detecting violations that introduce bias in the parameter of interest and indicate that it is not useful for detecting violations related to variance and correlation parameters. Mohn's $\rho$ is subjected to randomness in data and a Mohn's $\rho$ test will sometimes fail to detect a misspecification. Indeed, the examples we have investigated represent rather drastic changes in parameters like survey catchability and catch reporting, and in many cases they are likely to not be detected (Fig. 4). Carvalho et al. (2017) also found Mohn's $\rho$ to have relatively little power compared to other model diagnostics, when inspecting 10-year retrospective patterns using the rule of thumb proposed by Hurtado-Ferro et al. (2015). Conversely, a reduction in retrospective patterns does not necessarily translate into less biased estimates. Szuwalski et al. (2018) illustrated through simulations that it was possible to make model modifications that reduce Mohn's $\rho$ while still increasing the error of reference point estimates. These and our results highlight the importance of justifying the model formulations independently of validation criteria and applying several validation procedures to choose between justified model candidates, e.g., inspection of residuals (Thygesen et al. 2017). We highlighted that the power of the Mohn's $\rho$ tests, provided in Table 1, are obtained by assuming that the estimated assessment model is the true data-generating process. The applied assessment model is never the correct model and the power will differ in reality from Table 1. However, the power analysis provides an indication about what types of model misspecifications a Mohn's $\rho$ test is able to detect.

The distribution of Mohn's $\rho$ depends on the complexity of the assessment model. Figure 3*a* illustrates changes in the acceptance regions for NEA cod Mohn's $\rho$ when an autoregressive structure is included in the fishing mortality increments. We found that the regions are slightly larger when the additional flexibility is included, and we find this intuitively reasonable since the latent process is then more flexible. This illustrates that model modifications may reduce Mohn's $\rho$ by affecting its distributions without improving the model specification.

The bootstrap procedure we propose can be applied to improve our intuition on how retrospective plots vary due to randomness under H0. If we are unsure whether a retrospective plot indicates model misspecification, we can compare it with simulated retrospective plots and see whether the real one is anomalous, e.g., if a retrospective plot looks one-sided and suggests that a parameter may have been over- or underestimated in recent years, we can investigate how often such a structure occurs due simply to random variation given the estimated model prior to the post-sample period.

In our research, we included prediction intervals in retrospective plots to illustrate how much estimates typically adjust when all data are included for estimation (see Fig. 2). Such intervals can be applied to objectively determine whether an estimate has changed significantly in a retrospective analysis. We highlighted that these intervals can be wider than the corresponding confidence intervals that utilize all data. This is reasonable because we are typically surer about an estimate a few years back in time compared to in the terminal year.

The post-sample Mohn's $\rho$ significance test conditions Mohn's $\rho$ inference on data prior to the post-sample period. This differs from Miller and Legault (2017), who bootstrapped the entire data set based on distributional assumptions about the observations. Following the idea of bootstrapping the entire data set, we could sample realizations from the posterior distribution of the latent effects and further sample new observations conditioned on these effects. However, this procedure could be problematic because model misspecifications may be included in the posterior distribution of the latent effects. For example, in scenario M3 in Table 1, it is intuitive that the survival process posterior has a mean significantly smaller than zero in the post-sample period to accommodate the increased natural mortality. Miller and Legault (2017) did not investigate state–space models, so they did not have to address this problem.

A rule of thumb to reduce subjectivity is to interpret that a retrospective pattern is significant if the confidence interval of the estimate (e.g., SSB) does not include the $\rho$-adjusted value (Brooks and Legault 2016). That is, interpret the Mohn's $\rho$ as significant if $\widehat{X}_{Y|Y}\frac{1}{1+\rho_X} \notin \mathrm{CI}_{X_{Y|Y}}$, where $\widehat{X}_{Y|Y}$ is the estimate of interest, $\rho_X$ is the corresponding obtained Mohn's $\rho$, and $\mathrm{CI}_{X_{Y|Y}}$ is the confidence interval of $X_{Y|Y}$. The intuition behind the rule of thumb is that if the $\rho$-adjusted estimate is within $\mathrm{CI}_{X_{Y|Y}}$, then the adjustment was not significant with respect to the uncertainty of $\widehat{X}_{Y|Y}$. Our research differs because we directly provide confidence intervals of Mohn's $\rho$ under H0.

Mohn (1993) bootstrapped survey observations based on a VPA model and found that the average bootstrap estimates did not show retrospective patterns. A similar result is found in this research because the median Mohn's $\rho$ is centered around zero, see Fig. 1. This is because under H0, the bias of an estimate should be close to zero. If it is not centered at zero that may imply parameter identifiability issues.

In a state space model, where recruitment is a random walk (as in our case studies), rather than a function of a stock recruitment curve, there is little to inform the pattern in later years, and the estimates tend to become constant. Additional years of catch data can retroactively inform the recruitment leading to large retrospective patterns that are not indicative of model misspecifications. If there is a stock–recruitment relation in the state space model, we recommend utilizing that structure in the significance test.

The post-sample Mohn's $\rho$ significance test investigates whether the post-sample data are consistent with the model estimated with data prior to the post-sample period. Table 1 provides the test power in scenarios where a misspecification was introduced in the post-sample period. If a misspecification enters before the post-sample period, the data-generating process under H0 partially includes the misspecification. As a result, the test power will depend on how the model has incorporated the misspecification at the start of the retrospective analysis.

When a significant retrospective pattern is found, we recommend reviewing both the data and assessment model in detail. Such a review demands team effort by scientists with expert knowledge about all aspects of the assessment. We will now highlight two examples from benchmarks where we have first-hand experience on how the data and assessment model were reviewed and modified to remove retrospective patterns. (1) During the benchmark for NEA cod ICES (2021*a*), a clear retrospective pattern was observed for SSB. For this particular stock, a corresponding survey area was extended some years prior to the benchmark, and it was therefore suggested that a separate catchability parameter be included before and after the extension. The inclusion of a break point in the catchability parameter removed the retrospective pattern in SSB. (2) During the benchmark for NEA haddock ICES (2020*b*), retrospective patterns were removed by including survey plus groups in eq. 3*b*. Previously, survey observations of the oldest haddock were discarded. The abundance of old haddock had increased in the last decade, and it was therefore intuitive that the inclusion of separate survey plus groups could solve retrospective issues.

Retrospective analysis is useful for identifying consistent over- or underestimation in consecutive years, pinpointing exactly the kind of errors that are of particular concern for the management of fish stocks. Evaluating retrospective patterns is very important in operational stock assessment. To avoid misinterpretation of retrospective patterns, it is important to evaluate their statistical significance. In this respect, it is important to recognize that the distributions of these patterns are determined by the specific model applied and the data available. We have presented tools that address these concerns, and that can assist in deciding on whether a retrospective pattern is significant and indicative of model misspecification. In particular, our proposed procedures are well suited to guard against over-interpretation of retrospective patterns.

## Acknowledgments

The authors want to thank Sondre Aanes, Sondre Hølleland, Knut Korsbrekke, and Geir Storvik for discussion about the paper, and Johanna Fall for helping with NCC cod data and assessment set-up. The authors also want to thank Chris Legault and three anonymous reviewers for constructive comments that improved the article.

## Article information

### History dates

Received: 28 October 2022
Accepted: 17 May 2023
Accepted manuscript online: 1 June 2023
Version of record online: 17 July 2023

## Data availability
Data and scripts to reproduce results are available on GitHub (https://github.com/OlavNikolaiBreivik/cjfasPaperMonhsRho2022).

# Author information

## Author ORCIDs
Olav Nikolai Breivik https://orcid.org/0000-0002-9336-4297
Magne Aldrin https://orcid.org/0000-0003-2718-8528
Edvin Fuglebakk https://orcid.org/0000-0002-3192-6768
Anders Nielsen https://orcid.org/0000-0001-9683-9262

## Author contributions
Conceptualization: ONB, MA, EF, AN
Formal analysis: ONB
Methodology: ONB, MA, AN
Project administration: EF
Software: ONB, AN
Writing – original draft: ONB
Writing – review & editing: MA, EF, AN

## Competing interests
The authors declare there are no competing interests.

## Funding information
The work of ONB, MA, and EF was funded by the Norwegian Institute of Marine Research through the project “ASAM”. The work of AN was funded by EU and the Danish Agri-Fish Agency via the EMFF project "Emergency action for calculating stock size and reference points in case of massive data deficiency, (33113-B-20-168)".

# Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/cjfas-2022-0250.

# References

Aanes, S. 2016. A statistical model for estimating fish stock parameters accounting for errors in data: applications to data for Norwegian Spring-spawning herring. *In* WD4 for ICES WKPELA 2016.

Aglen, A., Fall, J., Gjøsæter, H., and Staby, A. 2021. Abundance indices for Norwegian coastal cod north of 62°N. Rapport fra havforskningen. (6): 93. *In* Available from https://www.hi.no/hi/nettrapporter/rapport-fra-havforskningen-en-2021-6 [accessed 18 June 2023].

Berg, C.W., and Nielsen, A. 2016. Accounting for correlated observations in an age-based state–space stock assessment model. ICES J. Mar. Sci. **73**(7): 1788–1797. doi:10.1093/icesjms/fsw046.

Breivik, O.N., Nielsen, A., and Berg, C.W. 2021. Prediction-variance relation in a state–space fish stock assessment model. ICES J. Mar. Sci. **78**(10): 3650–3657. doi:10.1093/icesjms/fsab205.

Brooks, E.N., and Legault, C.M. 2016. Retrospective forecasting—evaluating performance of stock projections for new england groundfish stocks. Can. J. Fish. Aquat. Sci. **73**(6): 935–950. doi:10.1139/cjfas-2015-0163.

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**(2): 296–308. doi:10.1139/cjfas-2015-0047.

Carvalho, F., Punt, A.E., Chang, Y.J., Maunder, M.N., and Piner, K.R. 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. **192**: 28–40. doi:10.1016/j.fishres.2016.09.018.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., et al. 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. **240**: 105959. doi:10.1016/j.fishres.2021.105959.

Core-Team and contributors worldwide 2022. stats. R package version 4.2.0.

Cressie, N., and Wikle, C.K. 2011. Statistics for spatio-temporal data. John Wiley & Sons.

Davison, A.C., and Hinkley, D.V. 1997. Bootstrap methods and their application. Number 1. Cambridge University Press. doi:10.1017/CBO9780511802843.

Efron, B., and Tibshirani, R.J. 1994. An introduction to the bootstrap. CRC press. doi:10.1201/9780429246593.

Harvey, A.C. 1990. Forecasting, structural time series models and the Kalman filter. Cambridge University Press. doi:10.1017/CBO9781107049994.

Hirst, D., Storvik, G., Rognebakke, H., Aldrin, M., Aanes, S., and Vølstad, J.H. 2012. A bayesian modelling framework for the estimation of catch-at-age of commercially harvested fish species. Can. J. Fish. Aquat. Sci. **69**(12): 2064–2076. doi:10.1139/cjfas-2012-0075.

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., et al. 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. **72**(1): 99–110. doi:10.1093/icesjms/fsu198.

ICES 2019. Benchmark workshop on baltic cod stocks (WKBALTCOD2). ICES, **9**(1). doi:10.17895/ices.pub.4984.

ICES 2020a. Artic Fisheries Working Group (AFWG), *In* ICES Scientific Reports. 2:52. p. 577. doi:10.17895/ices.pub.6050.

ICES 2020b. Benchmark Workshop for Demersal Species (WKDEM). *In* ICES Scientific Reports. 2:31. p. 136. doi:10.17895/ices.pub.5548.

ICES 2020c. Inter-benchmark process on Sandeel (*Ammodytes* spp.) in Area 2r (central and southern North Sea, Dogger Bank), and Area 3r (Skagerrak, northern and central North Sea)(IBPSandeel). *In* ICES Scientific Reports. 2:11. p. 23. doi:10.17895/ices.pub.5553.

ICES 2020d. Joint NAFO/ICES Pandalus Assessment Working Group (NIPAG). *In* ICES Scientific Reports. 2:19. p. 22. doi:10.17895/ices.pub.5554.

ICES 2020e. Workshop on Catch Forecast from Biased Assessments (WKFORBIAS; outputs from 2019 meeting). doi:10.17895/ices.pub.5997.

ICES 2021a. Benchmark Workshop for Barents Sea and Faroese Stocks (WKBARFAR 2021). *In* ICES Scientific Reports. 3:21. p. 205. doi:10.17895/ices.pub.7920.

ICES 2021b. Benchmark Workshop on herring (Clupea harengus) in the Gulf of Bothnia (WKCLuB). *In* ICES Scientific Reports. 3:9. p. 113. doi:10.17895/ices.pub.5989.

ICES 2022. Workshop on the evaluation of northern Norwegian coastal cod harvest control rules (WKNCCHCR). **4**(49). doi:10.17895/ices.pub.20012459.v1.

Johannesen, E., Johnsen, E., Johansen, G.O., and Korsbrekke, K. 2019. StoX applied to cod and haddock data from the Barents Sea NOR–RUS ecosystem cruise in autumn. Fisken og Havet. (6): 40. *In* Available from https://www.hi.no/templates/reporteditor/report-pdf?id=29879&59805751 [accessed 18 June 2023].

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., and Fu, D. 2021. Validation of stock assessment methods: is it me or my model talking? ICES J. Mar. Sci. **78**(6): 2244–2255. doi:10.1093/icesjms/fsab104.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. 2016. TMB: automatic differentiation and Laplace approximation. J. Stat. Softw. **70**(5): 1–21. doi:10.18637/jss.v070.i05.

Legault, C.M. 2020. Rose vs. Rho: a comparison of two approaches to address retrospective patterns in stock assessments. ICES J. Mar. Sci. **77**(7–8): 3016–3030. doi:10.1093/icesjms/fsaa184.

Mehl, S., Aglen, A., and Johnsen, E. 2016. Re-estimation of swept area indices with CVs for main demersal fish species in the Barents Sea winter survey 1994–2016 applying the Sea2Data StoX software. *In* Fisken og Havet(10): 44.

Miller, T.J., and Legault, C.M. 2017. Statistical behavior of retrospective patterns and their effects on estimation of stock and harvest status. Fish. Res. **186**: 109–120. doi:10.1016/j.fishres.2016.08.002.

Miller, T.J., Hare, J.A., and Alade, L.A. 2016. A state–space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to southern New England yellowtail flounder. Can. J. Fish. Aquat. Sci. **73**(8): 1261–1270. doi:10.1139/cjfas-2015-0339.

Mohn, R. 1993. Bootstrap estimates of adapt parameters, their projection in risk analysis and their retrospective patterns. Can. Special Pub. Fish. Aquat. Sci. 173–184.

Mohn, R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. **56**(4): 473–488. doi:10.1006/jmsc.1999.0481.

Newman, K., King, R., Elvira, V., de Valpine, P., McCrea, R.S., and Morgan, B.J. 2022. State–space models for ecological time-series data: practical model-fitting. Methods Ecol. Evol. doi10.1111/2041-210X.13833.

Nielsen, A., and Berg, C.W. 2014. Estimation of time-varying selectivity in stock assessments using state–space models. Fish. Res. **158**: 96–101. doi:10.1016/j.fishres.2014.01.014.

Perreault, A.M., Wheeland, L.J., Morgan, M.J., and Cadigan, N.G. 2020. A state–space stock assessment model for American plaice on the Grand Bank of Newfoundland. J. Northwest Atl. Fish. Sci. **51**: 45. doi:10.2960/J.v51.m727.

Szuwalski, C.S., Ianelli, J.N., and Punt, A.E. 2018. Reducing retrospective patterns in stock assessment and impacts on management performance. ICES J. Mar. Sci. **75**(2): 596–609. doi:10.1093/icesjms/fsx159.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K., and Nielsen, A. 2017. Validation of ecological state space models using the Laplace approximation. Environ. Ecol. Stat. **24**(2): 317–339. doi:10.1007/s10651-017-0372-4.