**Can autocorrelated recruitment be estimated using integrated assessment models, and how does it affect population forecasts?**

Kelli Johnson1,\*, Elizabeth Councill1,2, James T. Thorson1, Elizabeth Brooks3, Richard D. Methot4, André E. Punt2

1School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA

2Fishery Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

3Northeast Fisheries Science Center, 166 Water Street, Woods Hole, MA 02543, USA

4NOAA Senior Scientist for Stock Assessments, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

\*Corresponding author:

telephone: +1 206 543 4270; fax: +1 206 616 8689; email: kfjohns@uw.edu

**Highlights**

* bullet 1
* bullet 2
* bullet 3
* bullet 4
* bullet 5

**Abstract**

The addition of juveniles to marine populations (termed “recruitment”) is highly variable due to variability in survival for larvae and early juvenile stages. Recruitment estimates are often positive or negative for several years in a row (termed “autocorrelated” recruitment). Recruitment may be autocorrelated due to numerous factors including regime shifts and periodicity in environmental drivers affecting juvenile survival rates. The ability of stock assessments to accurately estimate the magnitude of recruitment autocorrelation, and its effect on the quality of forecasts of spawning stock biomass, has not generally been analyzed. We used a simulation experiment to evaluate how well Stock Synthesis (an ‘integrated’ age-structured stock assessment method used extensively in the assessment of fish stocks) estimates autocorrelation in the presence of a range of levels of autocorrelation in recruitment deviations. The precision and accuracy of estimated autocorrelation, and the ability of the stock assessment framework to forecast the true dynamics of the system, were compared for scenarios where autocorrelation parameter within the assessment was fixed at zero, fixed at its true value, internally estimated, or input as a fixed value determined using an external estimation procedure. Estimates of autocorrelation produced by Stock Synthesis were biased toward extreme values (i.e., towards 1.0 when true autocorrelation positive and -1 when true autocorrelation was negative). Less biased estimates of autocorrelation were obtained by externally estimating it from the recruitment deviations estimated within Stock Synthesis. We also explore how neglecting or estimating recruitment affects the statistical performance of predictions of future biomass following the period with available data (termed “population forecasts”). We show that ignoring autocorrelation when true recruitment is autocorrelated results in poor forecast interval coverage (i.e., a large proportion of simulation replicates where true biomass is outside the predictive interval for the forecast). However, the “external estimate” of autocorrelation generally improves forecast interval coverage. Collectively, our results suggest that autocorrelation estimates have good statistical performance when calculated from the estimated recruitment deviations. However, estimates are likely to be imprecise whenever there are relatively few years of data to estimate recruitment (i.e., less than 40 years of recruitment estimates).

**Keywords:** five keywords go here

**1. Introduction**

Under the United States Magnuson-Stevens Fishery Conservation and Management Act (MSA; United States Public Law 104-297), all stocks included in United States Fishery Management Plans must have target and limit reference points and all overfished stocks must have a rebuilding plan. Rebuilding plans involve specifying management measures to rebuild the stock to a biomass associated with maximum sustainable yield () within 10 years (or, if rebuilding within 10 years is impossible, then one generation time plus the median time for rebuilding in the absence of fishing). Legally, rebuilding plans must be more likely than not to succeed, i.e., be based upon a probabilistic forecast of future population dynamics given the agreed level of fishing that implies recovery with ≥ 50% probability.

Stock assessment models represent the link between collected data and scientific advice in fisheries management. Assessments are expected to use fits to historical data and prescribed harvest policies to forecast future stock abundance and catch levels. These predicted “Acceptable Biological Catches” must account for scientific uncertainty and provide ≤ 50% probability that overfishing will occur (Methot et al., 2013). Variability in recent recruitment to the stock is a major contribution to this scientific uncertainty. As National Marine Fisheries Service (NMFS) works to reduce the number of overfished stocks, projection success is being examined more critically, and the accuracy of probabilistic forecasts in rebuilding plans is receiving increased research attention (Neubauer et al., 2013; NRC, 2013).

Reference points and rebuilding forecasts are often estimated using a stock assessment model that treats fluctuations in recruitment as a random process around a prediction derived from a presumed relationship between spawning output and recruits (Clark, 1993; Methot and Wetzel, 2013). Stock assessments are increasingly conducted using “integrated” population dynamics models that typically incorporate many data types, including samples of compositional data from fisheries and surveys, indices of abundance, and information regarding total fishery harvest (Maunder and Punt, 2013). These data are combined to estimate values for population productivity (parameters in the stock-recruit relationship) and status (spawning biomass in each year relative to reference points). Probabilistic forecasts of future population dynamics can then be made given assumed fishing mortality rates.

Recent studies illustrate that recruitment for many fishes is non-random over time and includes high and low periods (Hollowed et al., 2001; Szuwalski et al., 2014). This could be driven by environmental factors acting on recruit survival (citation) or adult reproduction (citation), or both (citation), or changes in the abundance of predators (citation). Ideally, researchers can identify measureable environmental factors that are correlated with recruitment deviations or regime shifts, and which can be forecast into the future (Haltuch and Punt, 2011). If an environmental factor that helps predict future recruitment can be identified, it can then be used to inform rebuilding forecasts (Holt and Punt 2009, Punt 2011) and reference point calculations (Schirripa et al., 2009). If an environmental factor cannot be identified, population forecasts are sometimes calculated for different “states-of-nature”, where each state-of-nature depends upon a hypothetical scenario for expected future recruitment (e.g., high, average, and low productivity scenarios; citation).

When correlated measurable environmental factors remain unidentified, the influence of regime shifts can still be accounted for by invoking autocorrelation in future recruitment deviations (i.e., where future recruitment deviations are greater or less than zero for many years in a sequence). Including “autocorrelated recruitment” in the population dynamics model may result in wider forecasting intervals compared with assuming recruitment follows a white-noise process. This wider forecast interval may, in some cases, have better statistical coverage (e.g., a 75% forecast interval that contains the true value 75% of the time). Well-calibrated statistical coverage is a pre-requisite of probabilistic methods used for forecasting and reference point determination (Shertzer et al. 2008).

In this study, we explore and evaluate the performance of population forecasts obtained from an integrated, age-structured assessment model when recruitment is autocorrelated. We conduct a simulation experiment using a factorial design involving six plausible levels of autocorrelation, ρ, in recruitment deviations and four alternative configurations for estimating ρ in the assessment model. We explore model performance by answering two questions:

1. How well can the magnitude of autocorrelation be estimated? and
2. Does accounting for autocorrelation improve the accuracy and predictive coverage of forecasts compared with ignoring autocorrelation in recruitment deviations?

We conclude by outlining a practical strategy to test and account for autocorrelated recruitment when generating forecasts in real-world assessment models.

**2. Methods**

We conduct a simulation experiment using the Stock Synthesis (SS) assessment software (Methot and Wetzel 2013), which is widely used in the Unites States and provides a generic implementation of an integrated assessment model. SS estimates recruitment at the same time as other parameters that govern stock productivity and status, and uses the delta-method to propagate uncertainty about past and future recruitment when calculating standard errors for population forecasts. Simulations and analyses were accomplished using the *ss3sim* software package (Anderson et al. 2014a, 2014b), and a public online repository houses the simulation code (github.com/kellijohnson/AR-perf-testing) to ensure the results are reproducible.

The simulation framework consists of three components: (1) an operating model that generates the true population dynamics; (2) a sampling model that generates data from the operating model; and (3) an estimation method that is applied to the simulated data, where the parameter estimates and derived quantities (i.e., forecasted future population abundances) from the estimation method can be compared with their true values from the operating model. We use a factorial design involving six levels of recruitment ρ and four alternative configurations for the estimation method. One hundred simulation replicates were generated for each scenario, where each replicate has a different realization of process (recruitment deviations) and observation errors. Each replicate involves simulating population dynamics over 100 years, which we divide into three periods:

1. “Burn-in period” – Years 1-25 are simulated without any fishing;
2. “Fishing period” – Years 26-80 include a simulated fishery, which fishes at *MSY,* and the potential for data from the fishery and/or survey, which is used to fit an assessment model conducted in year 80; and
3. “Forecast period” – Years 81-100 are simulated without any fishing, which can be compared to forecasts based on parameter estimates derived from the estimation method.

**2.1 Operating model**

The operating model represents a cod-like (i.e., slow-growing and long-lived) life history based on biological parameters estimated from the stock assessment for North Sea cod (*Gadus morhua*; Deroba et al., 2015) with some simplifications facilitating interpretation of the results (Table 1). Simplifications include: one fishery and one survey, combined sexes, and selectivity parameters based on the maturity ogive.

We used the steepness-parameterization of the Beverton-Holt stock-recruit function:

(1)

where *rt* and *bt* are the estimate of recruitment output and spawning biomass, respectively, in year *t*, *h* and *r0* are estimated parameters representing steepness (the strength of recruitment compensation) and average recruitment at unfished spawning biomass *b0*, and recruitment deviation *εt* is calculated as:

(2)

where *δt* is a normally distributed coefficient representing recruitment variability:

(3)

where is the marginal variance of recruitment deviations and *ρ* is the magnitude of autocorrelation in recruitment. Equation (1) includes the term , which has an average value of 1.0. This term is included to ensure that *r0* is equal to the mean (not the median) recruitment given unfished spawning biomass.

Each replicate of the operating model involved simulating true dynamics over 100 years, where recruitment is variable each year but the same across scenarios given an iteration (i.e., the values of for the first replicate of the *ρ* = 0.0 scenario were the same as for the first replicate of the *ρ* = 0.9 scenario). Years 1 through 25 had no fishing and are included to ensure that the population age-structure in year 25 had plausible deviations away from its expectation in an unfished state. In years 26-80, fully-selected fishing mortality, *F*, was fixed at the value that produced maximum sustainable yield. Fishery selectivity was logistic, based on fish length, and was identical to the maturity ogive. Survey selectivity was similar, except that the length at which 50% of individuals were selected was specified as 80% of the length at which 50% of individuals were mature to ensure that the survey sampled younger fish than were caught in the fishery.

We simulated data for six scenarios that differed in the value of autocorrelation used to generate recruitment: -0.25, 0, 0.25, 0.5, 0.75 and 0.9. Included levels of ρ are centered approximately around estimates from recent meta-analyses (Mueter et al. 2007, Thorson et al. 2014).

**2.2 Sampling model**

Annual catch was reported without error from the start of the fishery (year 26) to the year of the assessment (year 80; see Fig. 2). Fishery length- and age-composition data were simulated every other year for years 26-80, and were drawn from a Dirichlet distribution with twice the standard deviation of a multinomial distribution given an annual sample size of 100. The Dirichlet distribution was used to more accurately reflect the reduced information contained in data collected from fisheries compared to surveys (Aanes and Pennington, 2003; Hulson et al., 2011). Survey length- and age-composition data were simulated every other year for years 41-79, and were drawn from a multinomial distribution with an annual sample size of 100. The survey was simulated every other year providing an index of relative abundance for years 41-79, and the abundance index was drawn from a lognormal distribution with log-standard deviation of 0.1 and log-mean equal to logarithm of stock biomass available to the survey in that year.

**2.3 Estimation model**

An age-structured stock assessment model was fit to each simulated data set, using data generated during the “fishing period” (years 26-80), and we refer to this as the “estimation model” (see Table 1 for a list of estimated parameters). Each estimation method generates forecasts of population abundance during years 81 to 100, and estimates recruitment deviations for years 1-100. For clarity of communication, we refer to recruitment deviations during the three periods of the model:

1. Recruitment deviations during years 1-25: These recruitment deviations occur prior to the collection of any data, and estimated so that estimated age-structure in the first year of data (year 26) has plausible deviations away from the unfished age-distribution;
2. Recruitment deviations for years 26-80: These recruitment deviations occur during available data, and are generally estimated with some precision;
3. Recruitment deviations for years 81-100: These recruitment deviations occur during the forecast interval, and ensure that dynamics during this period includes a plausible magnitude of recruitment variation.

All estimation models have no additional data during the forecast period (years 81-100), so recruitment deviations for years 81-100 are estimated at their expected value (i.e., zero when autocorrelation is absent, or decaying towards zero from the value of the estimated recruitment deviation in year 80 when autocorrelation is nonzero).

The estimation model uses identical equations to the operating model, except it also includes annually varying bias-correction for estimated recruitment:

(1b)

where Eq. 1b replaces Eq. 1 from the operating model, and is the fraction of bias-correction included for each year. The bias-correction term is included to ensure that *r0* is equal to the mean (not the median) recruitment given unfished spawning biomass, and we use the Methot and Taylor (2011) approach to estimating . This approach involves the following steps for each simulation replicate:

1. Run the model once to identify maximum likelihood estimates and standard errors for all parameters including ;
2. Extract standard error estimates, , and estimate the bias-correction for each year,
3. Fit a five-parameter bias-correction “ramp” to annual bias correction calculations, ;
4. Use predictions of bias-correction, , for each year in Eq. 1, and re-run the estimation model to identify maximum likelihood estimates and standard errors for all parameters.

This bias-correction algorithm can be derived under the assumption that recruitment deviations are a random effect (Thorson and Kristensen, 2016). For estimation models with autocorrelated recruitment, the bias correction is sometimes greater than 0.0 during the forecast period. Bias-correction is included during the forecast period because recruitment deviations at the end of the fishing period (e.g., year 80) will inform recruitment deviations during the forecast period (e.g., year 81) whenever recruitment is autocorrelated. SS uses the delta-method when calculating uncertainty in population abundance during the forecast period. Therefore, forecast period abundance has a standard error that includes uncertainty about future recruitment deviations, and this uncertainty is a function of the level of recruitment autocorrelation.

**2.3.1 Estimation model configurations**

The following four estimation methods were investigated for each level of ρ:

1. “True” – an estimation model where the autocorrelation parameter was fixed at the level used to generate the recruitment deviations in the operating model. This estimation model is not plausible for any real-world assessment (given that the true ρ will never be unknown), but is included as a reference case, to demonstrate model performance if the extent of autocorrelation were known exactly.
2. “Zero” –an estimation model where the autocorrelation parameter is fixed at zero. This estimation model represents the most common assumption in stock assessment models to date.
3. “Internal” – an estimation model where the autocorrelation parameter is estimated as a fixed effect. This scenario will likely result in biased estimates of autocorrelation, given that SS implements “penalized likelihood” estimation rather than true “mixed-effect” estimation (Thorson and Minto, 2015). Previous research demonstrates that penalized likelihood estimation results in negative bias when estimating the variation in the recruitment deviations (*σr*, Thorson *et al*., 2014). However, its performance when estimating the magnitude of ρ has not been previously explored.
4. “External” – an estimation model where the autocorrelation parameter is estimated externally to SS. This involves extracting estimates of recruitment deviations from the “Zero” estimation model, and then estimating the first-order autocorrelation of these estimates using the *acf* function in R (R Core Development Team 2015). This level of autocorrelation is then set as a fixed value in SS, and SS is run a second time to estimate other parameters for a given data set. This estimation model will likely have different estimation performance than the “internal” estimation model, given that sample- and population-level estimates are often different in maximum likelihood estimates of mixed-effects models (citation).

In each scenario, the marginal log-standard deviation of recruitment was fixed at the true value (Table 1).

For each estimation model, we specified that fishing mortality was zero during the forecast period, and this matches the operating model which has no fishing during the forecast period. Given that fishing rate is correctly specified during the forecast period, any bias or imprecision in population abundance during the forecast period arises either from (1) bias and imprecision of estimated parameters during the fishing period, or (2) the impact of mis-specifying ρ during the forecast period. The correct input sample size for both the multinomial and Dirichlet composition samples (*Ninput* = 100 and 100/22, respectively) were specified in each estimation method (i.e., the estimation model had correct weighting for age- and length-composition sampling data). Convergence of the estimation model was determined using the maximum gradient of the objective function, where models with a maximum gradient of less than 0.01 and a positive definite Hessian matrix were assumed to have converged. Models that failed to converge were removed from the analysis, and exploratory analysis confirms that results are qualitatively similar when changing the gradient threshold used to identify model convergence.

**2.3.2 Evaluating model performance**

Estimation performance was evaluated using two performance statistics:

1. relative error, , where and are estimated and true parameter values, respectively, where a well-performing estimation model will have a relative error close to zero for all simulation replicates; and
2. forecast interval coverage, defined as the proportion of simulation replicates where the forecast interval contains the true value from the operating model. A well-calibrated model will have approximately nominal forecast interval coverage, i.e., where a 50% forecast interval will contain the true value in 50% of simulation replicates.

**3. Results**

**3.1 Estimating autocorrelation**

We first seek to determine whether an integrated assessment model can provide an accurate and precise estimate of recruitment autocorrelation. We therefore evaluate estimates produced either when treating autocorrelation as a fixed effect (“internal”) or when calculating the sample autocorrelation of estimated recruitment deviations (“external”). “Internal” estimation is biased towards extreme values in all scenarios (i.e., towards 1.0 when true autocorrelation is positive and towards -1.0 when true autocorrelation is negative; Fig. 3 top row). The “internal estimation” model also has a high proportion of simulation replicates that do not converge when the true autocorrelation is 0.9. In these cases, the estimated autocorrelation approaches the bound at 1.0 and the hessian is generally not positive definite. By contrast, external estimates of autocorrelation are approximately unbiased for all levels of autocorrelation (Fig. 3 bottom row). The “external estimation” model also has a larger proportion of replicates that are converged.

**3.2 Impact of autocorrelation on population forecasts**

We next seek to determine the impact of autocorrelated recruitment on population forecasts, and whether including estimates of autocorrelation improves model performance. To do so, we first illustrate the effect of autocorrelated recruitment on estimated spawning output for all years (years 1-100) for the first replicate of the simulation experiment (Fig. 5). As expected, fixing autocorrelation at its true value results in a forecast interval that expands rapidly during the forecast period (years 81-100) whenever autocorrelation is substantially different from zero. Most noteable, the lower confidence bound for forecasts of spawning output declines over time when recruitment autocorrelation is 0.9, despite the forecast model correctly assuming that fishing is absent during this period (top-right panel of Fig. 5).

These patterns also hold when summarizing relative error in estimates of spawning output across all simulation replicates (Fig. 6). During the “fishing” period (years 26-80), average absolute relative error (AARE) in estimates of spawning output is generally less than 0.07 for all estimation models and all levels of true autocorrelation. We therefore conclude that increased recruitment autocorrelation, or mis-specifying recruitment autocorrelation, has relatively little impact on the precision and accuracy of estimates of spawning output during the period with information to estimate recruitment deviations. However, increased autocorrelation leads to a large increase in AARE during the forecast period (years 81-100), such that AARE is 0.2-0.26 when autocorrelation is 0.9. All estimation models have a AARE of 0.11 during the forecast period when recruitment is not autocorrelated, but when autocorrelation is high () the “true” and “external” models have lower AARE (0.17-0.18 and 0.0.20-0.21) than the “zero” estimation model (0.19 and 0.26). All models have a small positive bias in spawning biomass during the forecast period when autocorrelation is 0.75 and even more so when autocorrelation is 0.9. Exploratory analysis indicates that this bias arises due to the nonlinear stock-recruit function, i.e., because calculating forecasts based on the mean of the stock-recruit function is not identical to the expectation of the forecast due to this nonlinearity.

Finally, we illustrate 50% forecast interval coverage for each estimation model, defined as the proportion of simulation replicates where true spawning output falls within a 50% forecast interval (Fig. 7). A well-performing model will have nominal coverage probability, i.e., 50% of simulation replicates will fall within the 50% interval. When autocorrelation is absent (column “0.00” in Fig. 7), all estimation models have approximately nominal coverage, although all models exhibit a less-than-50% coverage (indicating too narrow of forecast intervals) in years 84-87. When autocorrelation is fixed at its true value (Fig. 7, upper rows), coverage remains close to 50% for all levels of true autocorrelation. However, increasing autocorrelation leads a large decline in coverage for the estimation model that neglects autocorrelation (Fig. 7, row two). For this model, coverage is close to 20% in year 90 (only 10 years into the forecast period) when autocorrelation is 0.75, and is approximately 10% in this year when autocorrelation is 0.9. By contrast, coverage is slightly smaller than 50% for the external estimation model when autocorrelation is 0.75 or 0.9. We therefore conclude that external estimation has substantially improved forecast interval performance relative to the model that neglects autocorrelated recruitment.

**4. Discussion**

Fisheries management in the United States and worldwide increasingly uses integrated stock assessment models to evaluate the likely impact of alternative management measures on fish population abundance. The United States and Europe both seek to end overfishing and rebuild overfished stocks (citation). Rebuilding plans in the United States are based upon forecasts of population abundance, and each United States Regional Fisheries Management Council is required to developed an approved Rebuilding Plan that will result in rebuilding within a pre-determined time frame. Rebuilding Plans are also required to be more likely than not to succeed in their stated timeframe, i.e., rebuilding plans are premised on a probabilistic interpretation of the forecasts generated from integrated stock assessment models. A probabilistic interpretation of catch advice arising from stock assessment models is also used in many United States regions to incorporate scientific uncertainty when defining catch limits (Shertzer et al., 2008) or when interpreting stock status relative to biological reference points (Stewart et al., 2013).

In this study, we have demonstrated that autocorrelated recruitment has a substantial impact upon both the accuracy of forecasts (i.e., how close they are to the true value) as well as the width of forecast intervals (i.e., how large is the estimated standard error during forecasts). In particular, high levels of autocorrelation (i.e., ρ>0.5) result in significant increases in the relative error of population forecasts, regardless of whether the stock assessment model includes autocorrelation or not. Also, a model where autocorrelation is fixed at its true value showed that forecast interval width is substantially increased when autocorrelation is high. These results confirm that the certainty of population forecasts is highly dependent upon the presence or absence of recruitment autocorrelation. Presumably, high recruitment autocorrelation could contribute to the lack of rebuilding for some fishes under rebuilding plans worldwide, particularly if forecasted biomass is overestimated as in our results (Hutchings, 2001; Neubauer *et al.*, 2013). Previous analysis of model output from stock assessment models suggests that recruitment may have intermediate, positive autocorrelation for marine fishes (Thorson et al., 2014). However, these previous results are based on model-output, which is fraught with statistical issues (Brooks and Deroba, 2015; Thorson et al., 2015a).

We have also shown improvements in forecast interval performance when fixing autocorrelation at the sample autocorrelation of estimated recruitment deviations. It is not necessary to accurately estimate forecast interval width as if management only currently utilize the median of the forecast, but the accuracy of the median will be very important. However, if fisheries managers use other quantities from the forecast (i.e., seek a management procedure that achieves a target biomass with 75% probability), or have Harvest Control Rules where the percentile for catch advice depends on the degree of depletion, then it is necessary to have accurate estimates of forecast interval width. Our simulation results show that the “external” estimate of autocorrelation results in less biased estimates of autocorrelation than estimating autocorrelation as a fixed effect.

The poor forecast interval performance when estimating autocorrelation as a fixed effect likely arises from our use of penalized-likelihood estimation methods. Penalized likelihood has previously been shown to result in a negative bias when estimating the variance of recruitment deviations (Thorson et al., 2015b), and a sample-based statistic has therefore been developed for estimating this variance (Methot and Taylor, 2011). However, we note that fixing autocorrelation externally does not propagate uncertainty about the magnitude of autocorrelation when estimating standard errors for other parameters or derived quantities for management (e.g., the CV of average unfished spawning biomass may be different when ρ is estimated compared to when ρ is fixed).

Based on our results here, we also identify several useful avenues for future research:

1. Most obviously, research could explore whether a mixed-effects estimate of autocorrelation could improve performance when estimating autocorrelation as a model parameter. Mixed-effects estimation is increasingly feasible using either the Lapace approximation (Skaug and Fournier, 2006; Thorson et al., 2015b) or Markov-chain Monte Carlo sampling (Stewart et al., 2013).
2. Future research could also explore the impact of autocorrelated recruitment on harvest strategy performance when either estimating or ignoring autocorrelation. Autocorrelated errors during forecast intervals are likely to impact the performance of different harvest strategies (Wiedenmann et al., 2015), but it remains unclear whether the magnitude of improvements from estimating autocorrelated recruitment ouweight the additional complexity when developing and explaining the model.
3. Finally, many parameters are likely to vary over time in stock assessment models, including growth, maturity, selectivity, and productivity (Martell and Stewart, 2014; Thorson et al., In preparation). These processes (e.g., time-varying growth) could affect the interpretation of length composition samples, so neglecting time-varying growth could in some cases appear as autocorrelated recruitment. We have not explroed the impact of multiple time-varying parameters on estimates of recruitment autocorrelation, and its potential impact remains difficult to predict. We therefore recommend ongoing research to develop tools to identify and account for time-varying parameters in stock assessment models.

**5. Conclusions**

We conclude that “external” estimation will likely result in better performance when estimating the magnitude of autocorrelated recruitment when estimation is based on penalized likelihood.

* Following my comment on section 3.1 (line 236), this might be a good place to summarize the detectability vs false-positive question. I’ve used SCAAs with and without ρ terms, and I suspect readers fall into that same category. Given that results in fishing years didn’t seem to vary across ρ cases or estimation configurations, and the fact that often times a separate forecast tool is used, it would be good to provide advice as to whether you can leave ρ=0 during the assessment model, then get a decent estimate of ρ externally and just use it in forecasts.
* The reliability of forecasts degrades for high autocorrelation, especially beyond the first 5-10 years. Rebuilding success is supposed to occur within 10 years, if feasible given life history. This suggests that progress towards rebuilding should be evaluated, and perhaps the assessment and forecast updated, within 5 years to avoid a “balloon payment” if overestimates in biomass are allowed to accumulate for too many years.

**6. Acknowledgements**

We gratefully acknowledge help from Ian Taylor, Chantel Wetzel, and Allan Hicks on implementation issues regarding ss3sim and forecasting within Stock Synthesis. This publication is partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement No. ?, Contribution No. ?. Partial funding for KFJ was provided by a grant from Washington Sea Grant, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA140AR4170078. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

**References**

Aanes, S., Pennington, M., 2003. On estimating the age composition of the commercial catch of Northeast Arctic cod from a sample of clusters. ICES J. Mar. Sci. 60, 297-303.

Anderson, S. C., Monnahan, C.C., Johnson, K.F., Ono, K., Valero, J.L., 2014a. ss3sim: An R package for stock assessment simulation with Stock Synthesis. Plos One 9, e92725.

Anderson, S.C., Monnahan, C.C., Johnson, K.F., Ono, K., Valero, J.L., Cunningham, C.J., Hurtado-Ferro, F., Licandeo, R., McGilliard, C.R., Szuwalski, C.S., Vert-pre, K.A., Whitten, A.R., 2014b. ss3sim: Fisheries stock assessment simulation testing with Stock Synthesis. R package version 0.8.9.9.

Clark, W.G., 1991. Groundfish exploitation rates based on life history parameters. Can. J. Fish. Aquat. Sci. 48, 734-750.

Deroba, J.J., Butterworth, D.S., Methot, R.D., Jr., De Oliveira, J.A.A., Fernandez, C., Nielsen, A., Cadrin, S.X., Dickey-Collas, M., Legault, C.M., Ianelli, J., Valero, J.L., Needle, C.L., O’Malley, J.M., Chang, Y-J., Thompson, G.G., Canales, C., Swain, D.P., Miller, D.C.M., Hintzen, N.T., Bertignac, M., Ibaibarriaga, L., Silva, A., Murta, A., Kell, L.T., de Moor, C.L., Parma, A.M., Dichmont, C.M., Restrepo, V.R., Ye, Y., Jardim, E., Spencer, P.D., Hanselman, D.H., Blaylock, J., Mood, M., Hulson, P.-J.F., 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.

Hollowed, A.B., Hare, S.R., Wooster, W.S., 2001. Pacific Basin climate variability and patterns of Northeast Pacific marine fish production. Prog. Oceanogr. 1-4, 257-282.

Holt, C.A., Punt, A.E., 2009. Incorporating climate information into rebuilding plans for overfished groundfish species of the U.S. west coast. Fish. Res. 100, 57-67.

Hulson, P-J.F., Hanselman, D.H., Quinn, T.J., II., 2011. Effects of process and observation errors on effective sample size of fishery and survey age and length composition using variance ration and likelihood methods. ICES J. Mar. Sci. 68, 1548-1557.

Methot, R.D., Taylor, I.G., 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Can. J. Fish. Aquat. Sci. 68, 1744-1760.

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.

Methot, R.D., Jr., Tromble, G.R., Lambert, D.M., Greene, K.E., 2013. Implementing a science-based system for preventing overfishing and guiding sustainable fisheries in the U.S. ICES J. Mar. Sci. 71, 183-194. 10.1093/icesjms/fst119

Maunder, M.N., 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. Fish. Res. 109, 92-99.

Mueter, F.J., Boldt, J.L., Megrey, B.A., Peterman, R.M., 2007. Recruitment and survival of Northeast Pacific Ocean fish stocks: temporal trends, covariation, and regime shifts. Can. J. Fish. Aquat. Sci. 64, 911-927.

Neubauer, P., Jensen, O.P., Hutchings, J.A., Baum, J. K., 2013. Resilience and recovery of overexploited marine populations. Sci. 340, 347-349.

NRC, 2013. Evaluating the Effectiveness of Fish Stock Rebuilding Plans in the United States. The National Academies Press, Washington, D.C. http://www.nap.edu/catalog.php?record\_id=18488

Punt, A.E., 2011. The impact of climate change on the performance of rebuilding strategies for overfished groundfish species of the U.S. west coast. Fish. Res. 109, 320-329.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Schirripa, M.J., Goodyear, C.P., Methot, R.M., 2009. Testing different methods of incorporating climate data into the assessment of US West Coast sablefish. ICES J. Mar. Sci. 66, 1605-1613.

Shertzer, K.W., Prager, M.H., Williams, E.H. 2008. A probability-based approach to setting annual catch levels. Fish. Bull. 106, 225-232.

Thorson, J.T., Jensen, O.P., Zipkin, E.F., 2014. How variable is recruitment for exploited marine fishes? A hierarchical model for testing life history theory. Can. J. Fish. Aquat. Sci. 71, 973–983. doi: 10.1139/cjfas-2013-0645

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth Edition. Springer-Verlag.

Vert-pre, K.A, Amoroso, R.O., Jensen, O.P., Hilborn, R., 2013. Frequency and intensity of productivity regime shifts in marine fish stocks. Proc. Nat. Acad. Sci. 110, 1779-1784.

Table . Parameter specifications used in the operating models (OMs) and estimation methods (EMs). Parameter specifications that vary among scenarios (combinations of OMs and EMs) are denoted in the table.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter |  | OM | EM |
| Name | Symbol | True value | Fixed (F) orEstimated (Est) |
| Natural mortality rate | *M* | 0.2 yr-1 | F |
| Length at age 1 | *La=1* | 20 cm | Est |
| Asymptotic maximum length | *L∞* | 132 cm | Est |
| Von Bertalanffy growth coefficient | *k* | 0.2 yr-1 | Est |
| Coefficient of variation for length at age 1 | *CVa=1* | 0.1 | Est |
| Coefficient of variation for asymptotic maximum length | *CV∞* | 0.1 | Est |
| Length at 50% maturity |  | 38.2 cm | F |
| Length at 95% maturity |  | 48.9 cm | F |
| Average recruits for the unfished population (natural log) | *ln(r0)* | 18.7 | Est |
| Steepness of the Beverton-Holt stock recruit function | *h* | 0.65 | F |
| Marginal log-standard deviation of recruitment | *σR* | 0.4 | F |
| Magnitude of autocorrelated recruitment | *ρ* | varies | varies |
| Random coefficients for recruitment variability (years 1-100) | *δt* | varies | Est |
| Catchability coefficient for survey index of abundance (natural log) | *ln(q)* | 0 | Est |
| Length at 50% selection in the fishery |  | 38.2 cm | Est |
| Length at 95% selection in the fishery |  | 48.9 cm | Est |
| Length at 50% selection in the survey |  | 30.6 cm | Est |
| Length at 95% selection in the survey |  | 39.1 cm | Est |

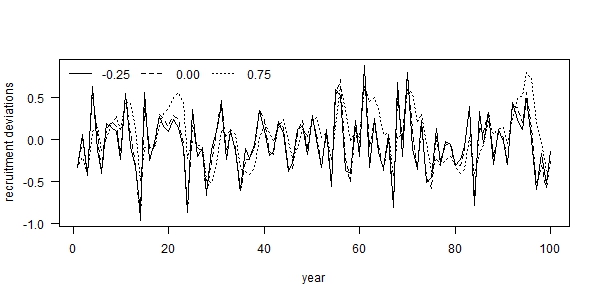


Fig. 1. Examples of autocorrelated recruitment deviations for three levels of autocorrelation: (i) -0.25 (solid line), (ii) 0.00 (dashed line), and 0.75 (dotted line), where each example used the same set of process error deviations ().

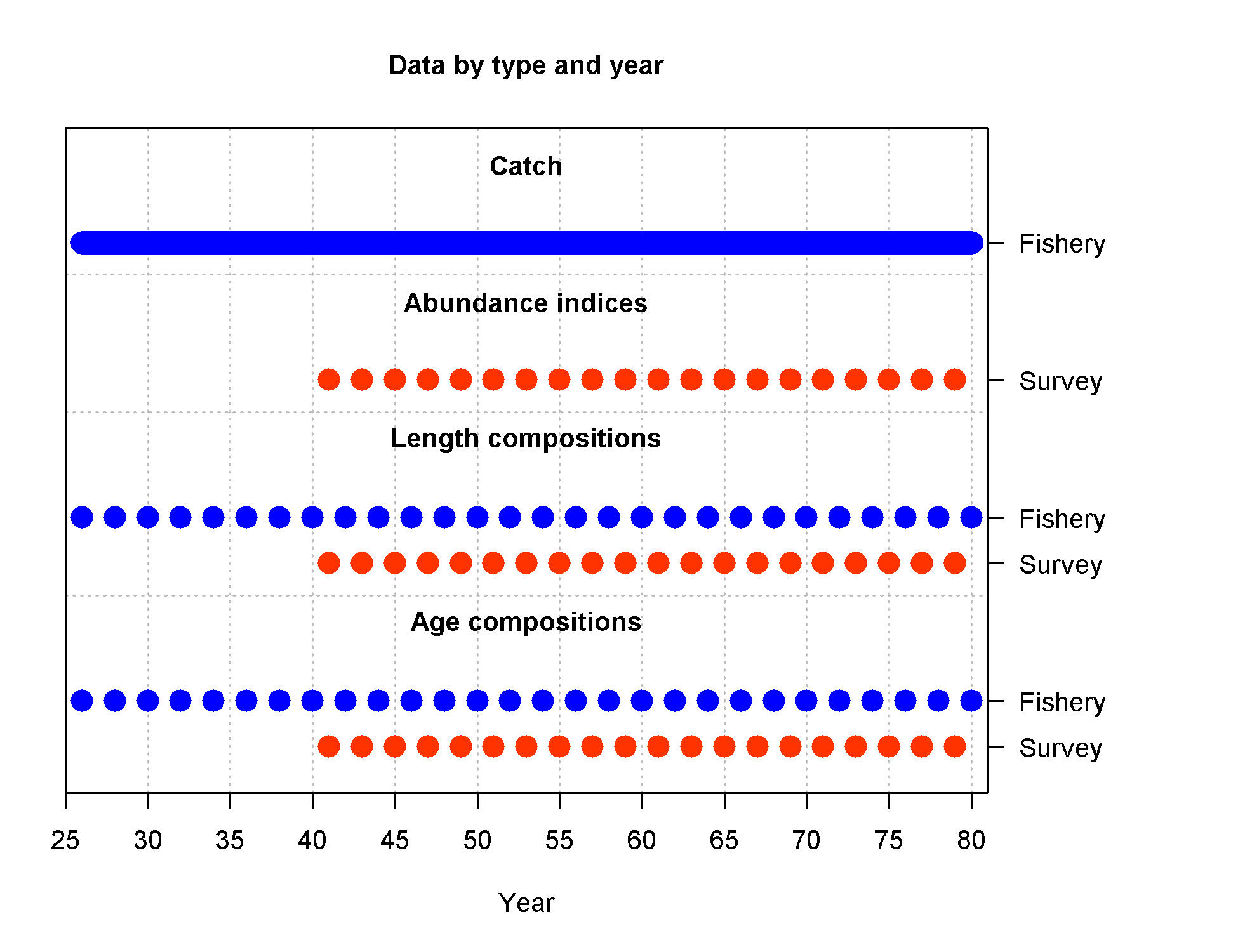
****

Fig. . Summary of simulated data available to the estimation model during the fishing period (years 26-80).

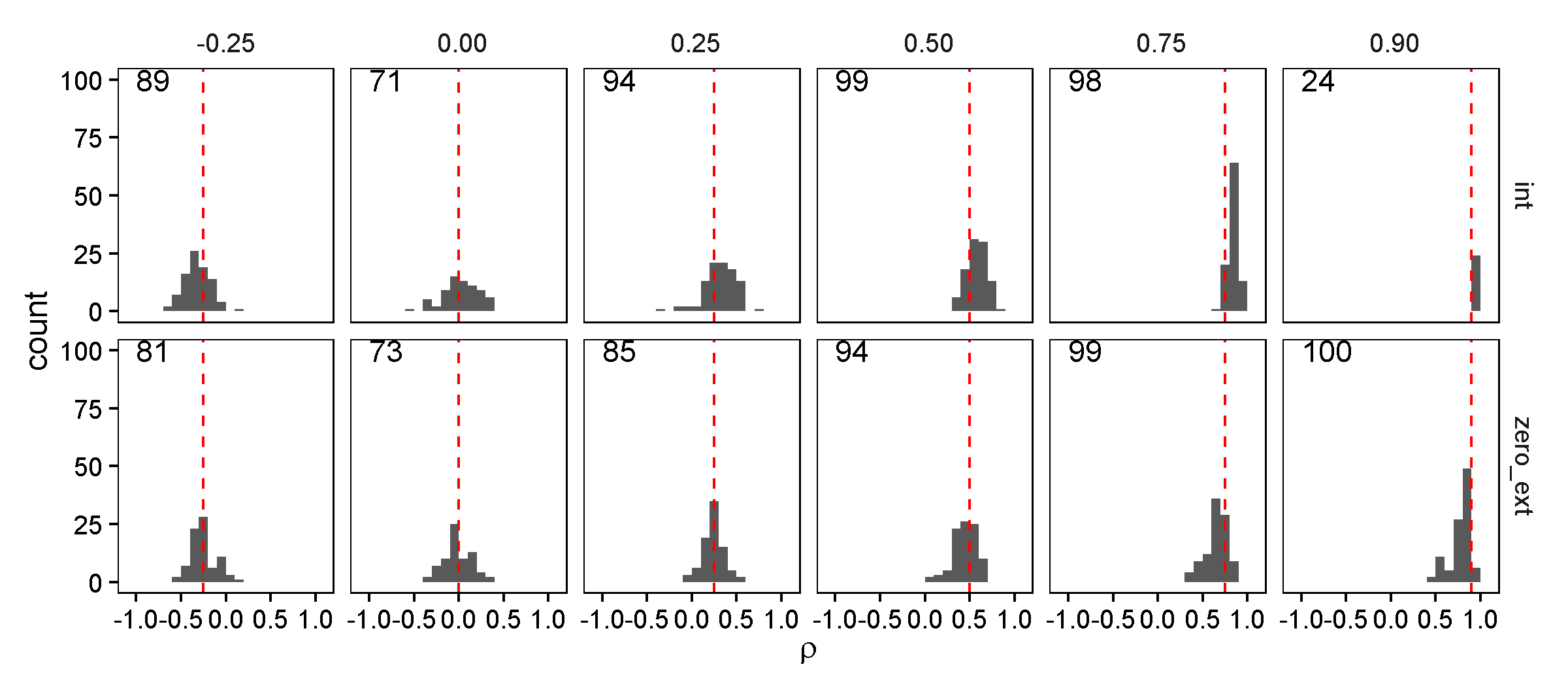


Fig. . Estimates of recruitment autocorrelation (*ρ*) from two estimation models: (i) calculated as the sample autocorrelation of recruitment deviations estimated in Stock Synthesis (“external”; top row) and (ii) estimated as a fixed effect within Stock Synthesis simultaneously with other parameter estimation (“internal”; bottom row), for six levels of recruitment autocorrelation (columns). The dashed red line illustrates the true level of autocorrelation, while the black shaded area is a histogram representing the simulation distribution for each scenario and estimation model.

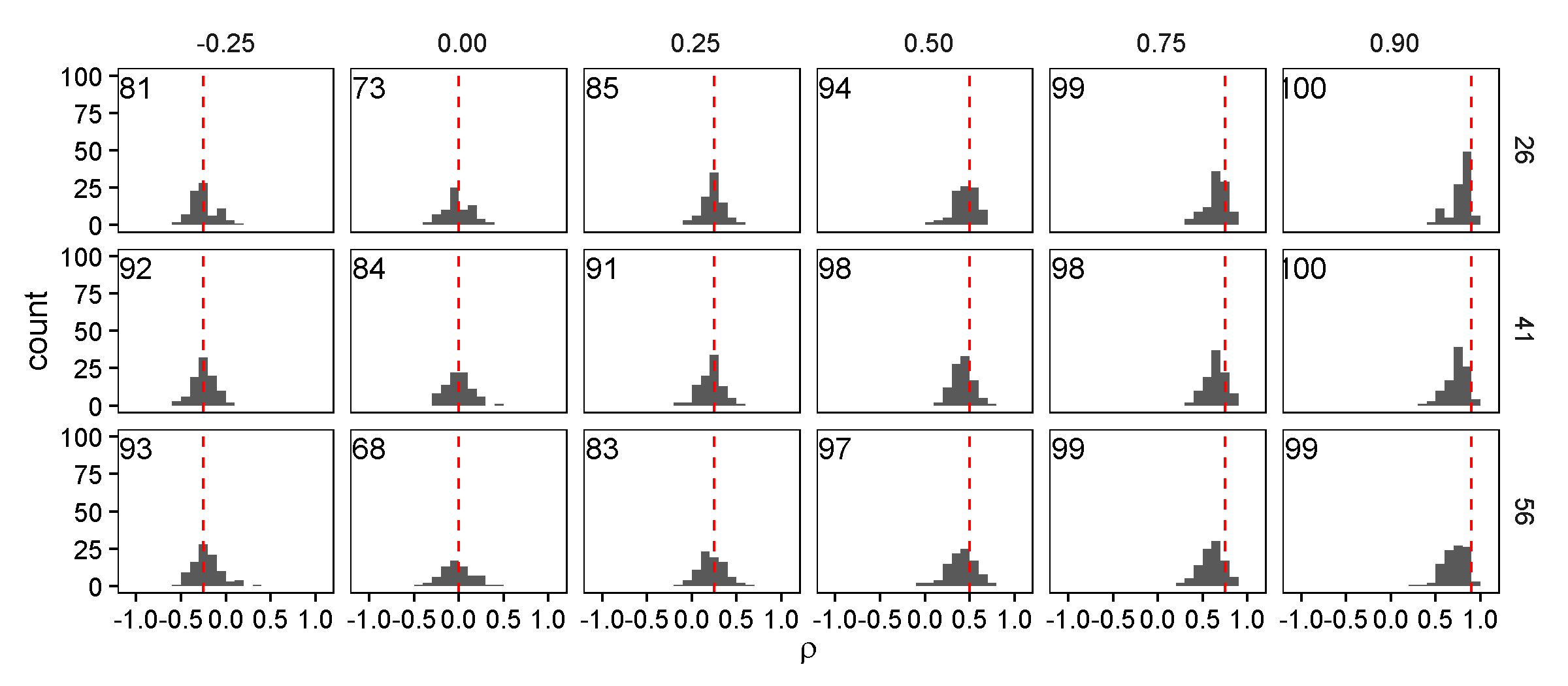


Fig. 4 -- Estimates of recruitment autocorrelation (*ρ*) from the “internal” estimation scenario, where it is calculated as the sample autocorrelation of recruitment deviations estimated in Stock Synthesis, for six levels of recruitment autocorrelation (columns) and three different starting years for fishery length and age-composition samples. The dashed red line illustrates the true level of autocorrelation, while the black shaded area is a histogram representing the simulation distribution for each scenario and estimation model.

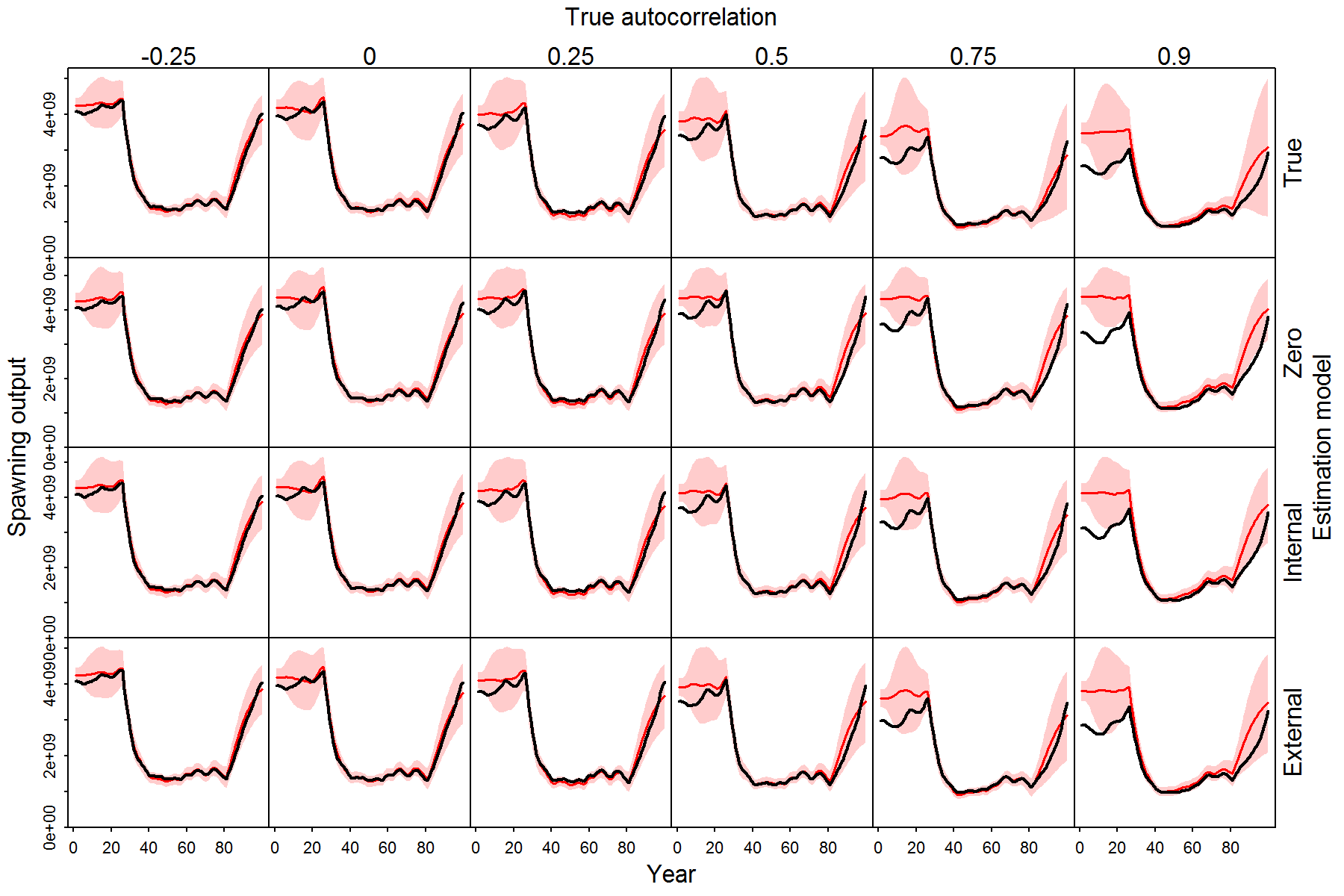


Fig. 5. Illustration of estimated spawning stock biomass (SSB) during 100 simulated years for different scenarios (columns, where recruitment autocorrelation is ρ={-0.25, 0.0, 0.25, 0.5, 0.75, 0.9}), and four estimation models (rows: external estimation, internal estimation, fixed at the true value, or fixed at zero), where each panel shows the true spawning biomass (black line) and the red shaded area shows the 95% confidence and forecasting interval for the estimating spawning biomass.

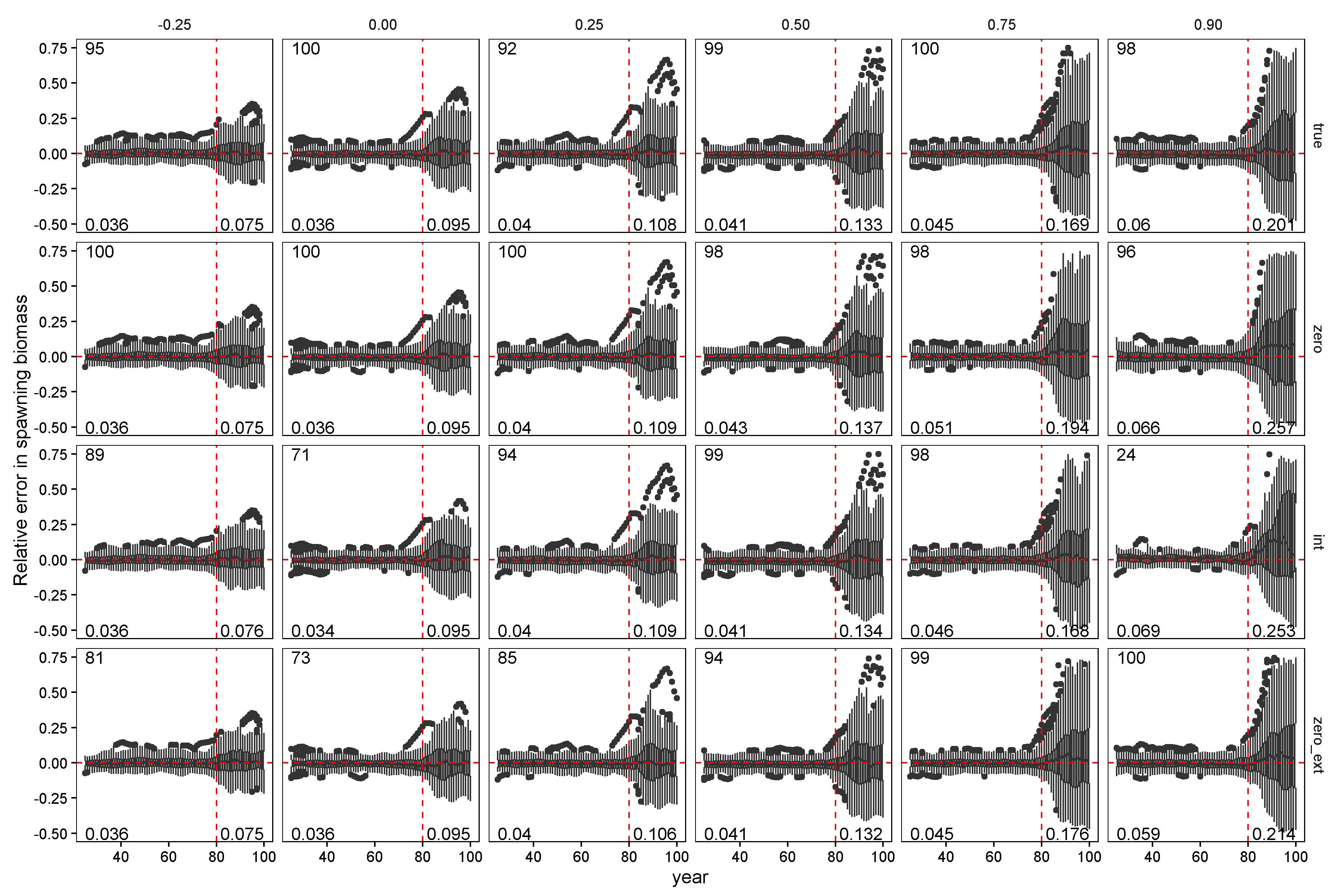


Fig. 6. Relative error in spawning stock biomass during years for which the assessment method was provided data (years 26 through 80) and the forecast period (years 81 through 100, to the right of vertical red dashed lines) for six levels of autocorrelation in the simulated data (columns) and four estimation methods (rows). Horizontal dashed red lines indicate a relative error of zero. Upper and lower edges of the boxes correspond to the first and third quartiles (the 25th and 75th percentiles) and the whiskers correspond to 1.5 times the distance between the first and third quartiles.

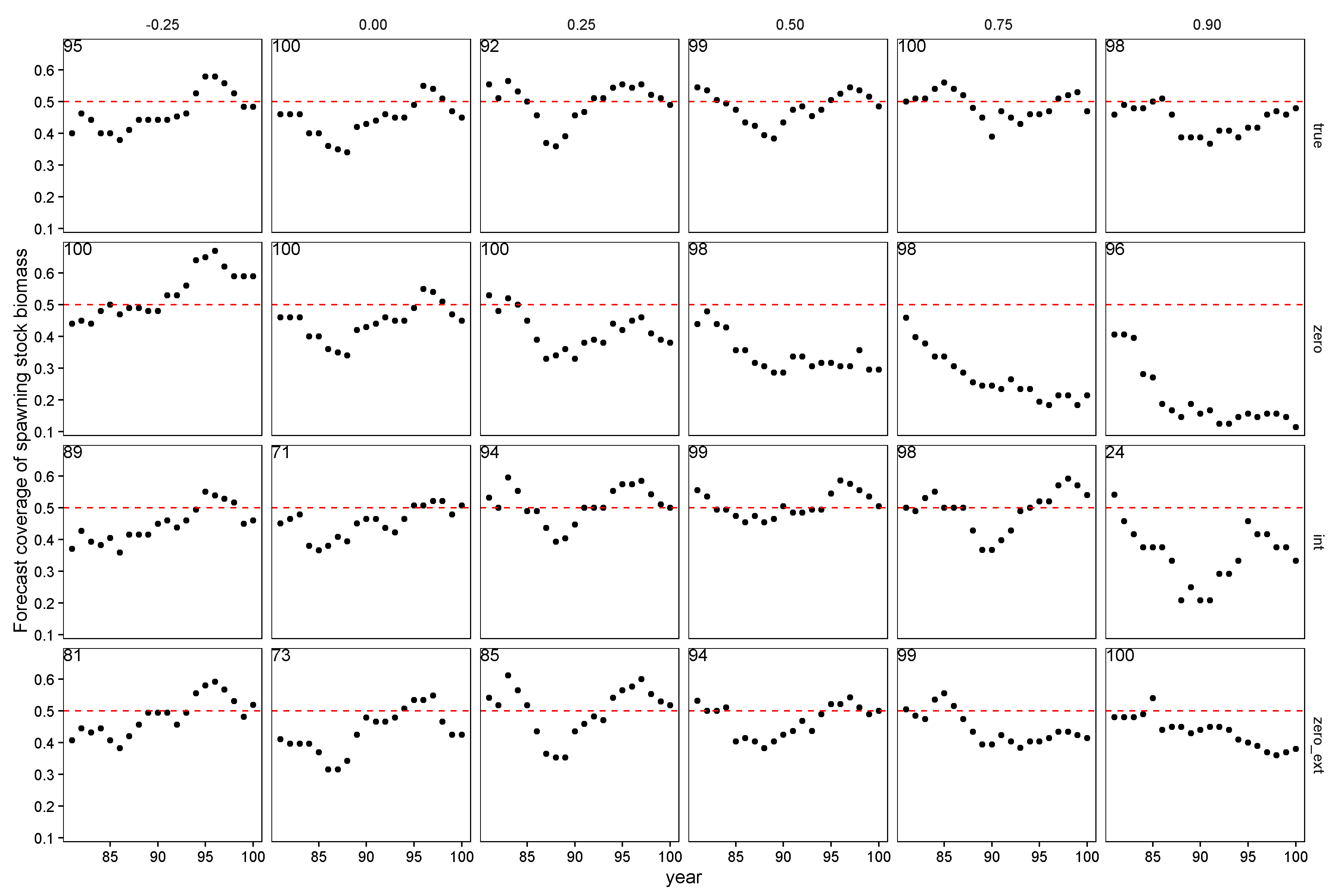


Fig. 7. Performance of forecast interval estimates for different estimation models (rows) and levels of autocorrelation (columns), where each panel shows the proportion of 50% forecast intervals for spawning stock biomass that contain the true value. A well calibrated 50% forecast interval will contain the true value 50% of the time, and this value is indicated by a red dashed line in each panel. Points above or below the line indicate forecast intervals were too conservative (wide) or permissive (not wide enough), respectively.