In the estimation process of the stock assessment, one of key contribution to the overall likelihood is fitness in abundance indices. Considering the statistical nature of the indices, a Sum of Squared Errors (SSE) between observed and predicted indices in log-space is normally used as a measure of fitness. However, for comparison of models, the SSE is not usable because complex models tend to be flexible and advantageous for fitting. And that is why information criteria for model selection such as AIC have been developed to account for the difference in complexity in models. Nevertheless, for example AIC is only a relative measure of appropriateness of models and is not an absolute measure, and therefore a diagnostic test is also sometimes employed for a statistical model validation in addition to the statistical model comparison.

In stock assessment, a different type of model validation than the conventional model validation has been proposed by Mohn (1999) to quantify a retrospective bias. As described in earlier sections, the retrospective analysis can be conducted by sequentially refitting the model to reduced data sets by removing some recent years' data to see if there are any systematic pattern or inconsistency within a model. The retrospective bias is then evaluated as the so-called Mohn's rho as

$$\rho = \sum_{t=T-n}^{T-1} \frac{\hat{y}_{(1:t),t} - \hat{y}_{(1:T),t}}{\hat{y}_{(1:T),t}},$$

where $\hat{y}$ denotes in general a value like estimated biomass, 1+population size, or predicted abundance index, and the value with suffix $\hat{y}_{(1:t'),t}$ means such a value estimated at time t of a full series from 1 to T using a retrospective data window from 1 to $t'(\leq T)$. In this paper, we will use a variant of the original ρ as the mean (average) like

$$\rho_r = \frac{1}{n} \sum_{t=T-n}^{T-1} \frac{\hat{y}_{(1:t),t} - \hat{y}_{(1:T),t}}{\hat{y}_{(1:T),t}} \quad \text{[rho for retro-bias]}, \tag{1}$$

This metric is an average of relative differences at the final time of each window, and therefore it is a measure of relative retrospective 'bias' (scale-free) in a statistical sense. Also note that the metric tends to be applied not the log-scale value but the original scale because both the directions of positive and negative biases are regarded equivalent unless any precautional aspects like prioritization of avoidance of falsely overestimation of populations size is considered.

The approach of hindcasting, which is a primally focus in this paper and a sort of retrospective cross-validation, is an extension of retrospective analysis to project forward beyond the retrospective data window to several step ahead to quantify the prediction skill of models. Theoretically the projection period is ultimately to the end of full time period, but practically the step size is one or several years ahead considering non-small process stochasticity in fishery population dynamics and non-ignorable extents of observation uncertainty. For evaluating the prediction skill, we shall propose use of several metrics as model-dependent and model-free validations. .

We now retrospectively use terminologies 'retro-period' and 'hc-period' as 'the period of shrunken data set for retrospective model fitting' and 'future time period with a certain projection step (say $S \geq 1$) for hindcasting after retro-period". And let $\hat{y}_{(1:t),t+s}$ be an projected value at time $t+s$ ($s \leq S$) in an hc-period based on the conditioned model with data in a retro-period $(1, t)$.

Modified Mohn's rho for prediction bias and absolute error:

$$\rho_p = \frac{1}{(n-S+1)S} \sum_{t=T-n}^{T-S} \sum_{s=1}^{S} \frac{\hat{y}_{(1:t),t+s} - \hat{y}_{(1:T),t+s}}{\hat{y}_{(1:T),t+s}} \text{ [rho for projection-bias]} \tag{2}$$

This is a simple extension of Mohn's rho to evaluate the prediction skill within the assumed model because all the values are produced by the model assumption. In this sense, this is used for a

model-dependent consistency check in prediction. In addition, if we need to evaluate the absolute error for prediction, we may be able to use the following absolute one.

$$|\rho_p| = \frac{1}{(n-S+1)S} \sum_{t=T-n}^{T-S} \sum_{s=1}^{S} \frac{\left|\hat{y}_{(1:t),t+s} - \hat{y}_{(1:T),t+s}\right|}{\hat{y}_{(1:T),t+s}}. \text{ [rho for projection-absolute-error]} \quad (3)$$

In case that only the prediction result for the S-th ahead for the S-step ahead by ignoring prediction outcomes between $t+1$ an $S-1$, the definition becomes simpler as

$$|\rho_p| = \frac{1}{(n-S+1)} \sum_{t=T-n}^{T-S} \frac{\left|\hat{y}_{(1:t),t+S} - \hat{y}_{(1:T),t+S}\right|}{\hat{y}_{(1:T),t+S}}. \text{ [rho for projection-absolute-error]} \quad (4)$$

I might be wrong: do we calculate errors for 3 predicted values when using 3-step ahead prediction for each hindcasting period? Or just the 3rd year in each 3-step ahead prediction? If the latter is the case, I need to remove the second summation and change the denominator of the fraction into just $n-S+1$ for the above and below...

Next three metrics are kinds of model-free validation. For this purpose, the error is basically defined as the difference between predicted value ($\hat{y}_{(1:t),t+s}$ as above) and actually observed value (say y_{t+s}) such as the model-based predicted CPUE using a retro-period data and observed CPUE used for model fitting.

Mean Absolute Percentage Error (MAPE) for projection:

$$MAPE = \frac{1}{(n-S+1)S} \sum_{t=T-n}^{T-S} \sum_{s=1}^{S} \frac{\left|\hat{y}_{(1:t),t+s} - y_{t+s}\right|}{y_{t+s}} \times 100 \quad (5)$$

This is another simple extension of the modified Mohn's rho for quantifying a relative difference between prediction and observation. This metric is also known as a scaled version of Mean Absolute Error (MAE). A problem with the MAE is that the relative size of the error is not always obvious. Sometimes it is hard to tell a big error from a small error. The MAPE can be calculated instead to deal with this problem, and this treatment allows forecasts of different series in different scales to be compared.

Root Mean Squared Error (RMSE) for projection error:
As an alternative measure of distance, the Mean Squared Error (MSE) is also commonly used in statistical literatures. To make comparison easier, the following squared root variant of MSE can be used:

$$RMSE = \sqrt{\frac{1}{(n-S+1)S} \sum_{t=T-n}^{T-S} \sum_{s=1}^{S} \left(\hat{y}_{(1:t),t+s} - y_{t+s}\right)^2} \quad (6)$$

Different from ρ_p and MAPE, RMSE is not scale-invariant and can be influenced by large discrepancies only in a single or a few data points. However, there is a useful feature that the squared RMSE can be in general expressed as a simple arithmetic. For a notational simplicity, let us set S at 1, and then

$$\begin{aligned} RMSE^2 &= \frac{1}{n} \sum_{t=T-n}^{T-1} \left(\hat{y}_{(1:t),t+1} - y_{t+1}\right)^2 \\ &= \frac{1}{n} \left(\hat{y}_{(1:t),t+1} - y_{t+1} - \bar{E}\right)^2 + \bar{E}^2 \\ &= E'^2 + \bar{E}^2 \end{aligned} \quad (7)$$

where

$$\begin{aligned} \bar{E} &= \frac{1}{n} \sum_{t=T-n}^{T-1} \left(\hat{y}_{(1:t),t+1} - y_{t+1}\right), \\ E'^2 &= \frac{1}{n} \left(\hat{y}_{(1:t),t+1} - y_{t+1} - \bar{E}\right)^2. \end{aligned} \quad (8)$$

The a centored mean squared error, E'^2 can be also expressed as

$$E'^2 = \sigma_o^2 + \sigma_f^2 - 2\sigma_o\sigma_f Cor, \tag{9}$$

where σ_o and σ_f are respectively the standard deviation of observation y_t and prediction, and Cor is their correlation. This means that E', ρ and σ_f can be summarised simultaneously (Taylor, 2001). Taylor diagrams therefore provide a concise statistical summary of how well patterns match each other and are therefore especially useful for evaluating multiple aspects or in gauging the relative skill of different models (Griggs and Noguer, 2002). It should be remarked that RMSE can be extended for a percentage measure as MAPE, but with a reason stated below, we use RMSE as it is.

Mean absolute scaled error (MASE) for projection:
A more robust and easier to interpret statistic for evaluating prediction skill is the MASE (Hyndman and Koehler, 2006). MASE evaluates a model' s prediction skill relative to a naíve baseline prediction, which is the observation. A prediction is said to have skill if it improves the model forecast compared to the baseline. A widely used baseline forecast for time series is the persistence algorithm that simply takes the value at the previous time step to predict the expected outcome at the next time step as a naíve in-sample prediction, i.e. tomorrow will the same as today. The original definition of MASE for 1-step ahead perdition is

$$MASE = \frac{\frac{1}{n}\sum_{t=T-n}^{T-1} \left|\hat{y}_{(1:t),t+1} - y_{t+1}\right|}{\frac{1}{n-1}\sum_{t=T-n+1}^{T-1} |y_{t+1} - y_t|}, \tag{10}$$

and this can be extended as actuallt not straighforward

$$MASE = \frac{\frac{1}{(n-S+1)S}\sum_{t=T-n}^{T-S}\sum_{s=1}^{S} \left|\hat{y}_{(1:t),t+s} - y_{t+s}\right|}{\frac{1}{(n-1)S}\sum_{t=T-n}^{T-1} |y_{t+1} - y_t|}, \tag{11}$$

The MASE has the desirable properties of scale invariance, predictable behavior, symmetry, interpretability and asymptotic normality. Also, compared to MAPE, which relies on division of observation for scaling, MASE does not necessarily skew its distribution even when the observed values are close to zero. In addition, a MASE score of 0.5 indicates that the model forecasts twice as accurate as a naíve baseline prediction; the model thus has prediction skill.

The best statistical measure to use depends on the objectives of the analysis and using more than one measure can be helpful in providing insight into the nature of observation and process error structures. Here for the evaluation of models, we will use the following metrics:

- Original Mohn's rho (ρ) for checking the retrospective bias
- Modified Mohn's rho for prediction bias and absolute error, which? both might be meaningful though but it becomes noisy... as checking model-based self-consistency check
- MASE and RMSE for model-free validation with different angles.

References

Griggs, D. J. and Noguer, M. (2002). Climate change 2001: the scientific basis. contribution of working group i to the third assessment report of the intergovernmental panel on climate change. *Weather*, 57(8):267–269.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 106(D7):7183–7192.