Testing and comparison of data-limited assessment models for estimating global and regional stock status

Olaf P. Jensen[1] and Christopher M. Free[2]

January 15, 2017

[1]146 Carter Rd.
Princeton, NJ 08540
Tel. +1 410-812-4842
olaf.p.jensen@gmail.com

[2] 235 Lincoln Ave.
Highland Park, NJ 08904
Tel. +1 610-999-4732
cfree14@gmail.com

Terms of Reference and Summary

1. "Description of methods (or reference to published work if available), description of changes made (if applicable), together with all data and codes used for at least 3 ***data-limited assessment methods***."

   Methods are described on pages 3-6 and all data and code is available in an online repository: [URL to come]

2. Detailed outputs should include estimates of stock status/categories (e.g., B/B_{MSY}), associated uncertainty and diagnosis of method accuracy/precision (e.g., how well a particular method assign a stock the right exploitation category)

   Outputs are summarized on pages 6-7 and in Tables 1-12 and Figures 1-6.

3. Draft decision tree to identify best method for FAO stock 'types'.

   The results suggest a simpler heuristic is more appropriate than a full decision tree. Essentially, rORCS should be preferred when sufficient ancillary information is available, though more testing of this method is needed. However, for many stocks this ancillary information is not currently available. In this case, the preferred method will depend on effort dynamics. In the absence of information on effort dynamics, cMSY should generally be preferred. This is discussed on pages 7-8.

4. Recommendations on future work, including: (i) further work needed on methods development; (ii) approaches to combine method outputs (e.g., super-ensembles); (iii) methods to improve priors on depletion; and (iv) priorities to improve the assessments in the future (e.g., what kind of information should be prioritized if resources for data collection become available, identify patterns and risks of misreporting, etc.).

   Recommendations for future work are given on pages 7-8.

1. Background

The State of World Fisheries and Aquaculture (SOFIA) published every two years by the FAO Fisheries and Aquaculture Department represents the most comprehensive global summary of marine fishery status in the world. However, our ability to accurately assess the status (i.e., biomass in relation to a reference point such as B_{MSY}) of the thousands of SOFIA stocks is limited by data availability. For many stocks, including many regionally important commercially harvested species, little data is available other than a time series of catch. Thus there is a pressing need for methods that can provide an accurate status estimate for these "catch-only stocks."

The past 15 years have seen a rapid expansion in the methods available for estimating status of catch-only stocks. We refer to these here as Catch-only Methods (COMs). This expansion has been driven in part by mandates to assess and set catch limits for a wider variety of fisheries, and also by efforts to determine global and regional stock status (Thorson et al. 2012, Costello et al. 2012, 2016). The proliferation of COMs and the lack of a consistent framework for assessing their relative performance across a wide range of fishery characteristics and performance metrics makes it difficult to decide which method(s) to use for a given application. The most recent general review of COMs (Carruthers et al. 2012) was published only five years ago, but several new methods have been developed since then. We provide a brief overview here and then evaluate the performance of six of the most promising methods through application to catch time series from assessed stocks and from simulated stocks.

There are many possible ways of categorizing COMs, but one useful distinction is between methods based on an underlying population dynamics model and those that are not. The former category is made up largely of methods based on a logistic or Schaefer population model. The major distinction among methods is the way in which parameters are estimated and treatment of process and observation errors. These methods include catch-MSY (cMSY, Martell and Froese 2013), the Catch-Only-Model with Sampling-Importance-Resampling (COM-SIR, Vasconcellos and Cochrane 2005), the State-Space Catch-Only Model (SSCOM, Thorson et al. 2013), and the Optimized Catch-Only Model (OCOM, Zhou et al. in review b). The first three of these methods have been tested in a consistent comparative framework (Rosenberg et al. 2014) and combined in a superensemble model that uses predictions from these methods and one other method as input into a boosted regression tree model (Anderson et al. in press). Results of these two analyses suggest that cMSY generally outperforms COM-SIR and SSCOM across a wide range of stock types and simulation scenarios. Therefore, we included cMSY, but not COM-SIR or SSCOM, in our analysis. OCOM is a new method not previously tested against other COMs in a consistent and rigorous comparison, and was therefore included here.

Catch-only methods which are not based on an underlying population model include a diverse array of approaches. Among the simplest are a set of decision rules to develop stock status plots based on a comparison of catch in the current year (C_{curr}) relative to the maximum observed catch in the time series (C_{max}). There are at least two variations on this approach. We included one older variation (Froese & Kesner-Reyes 2002) because it has been widely used in previous efforts to develop global and regional stock status plots and one newer variation (Kleisner et al. 2013) which was developed in response to critiques of the earlier method.

More complex COMs without population dynamics models include the Costello et al. (2012) panel regression model, the Zhou et al. (in review a) boosted regression tree model, and the refined Only Reliable Catch Stocks method (rORCS, Free et al. in review). The panel regression model was not included here because a modified version of this method performed quite poorly in head-to-head comparisons with other models (Anderson et al. in press). The Zhou et al. (in review a) boosted regression tree model, like OCOM, is a new method not previously tested against other COMs in a consistent and rigorous comparison, and was therefore included here. The rORCS method is somewhat unlike the other methods in that it is based partly on characteristics of the catch time series, but also on characteristics of the fishery which tend, in practice, to be associated with stock status. Such associations cannot be properly evaluated in a simulation context. Nevertheless, in testing against real stocks (Free et al. in review), the rORCS method performed substantially better than the other methods against which it was compared (including cMSY) and was therefore included here.

2. Methods

2.1 Overview

We evaluated and compared the performance of six COMs by testing them on data from both real and simulated fish stocks. Real fish stocks provide a potentially more realistic basis for testing as they reflect all of the uncertainties (e.g., observation and process errors) inherent in actual fisheries data. The simulated stocks provide increased sample size, the ability to compare model estimates against true stock status that is known without error, and the opportunity to compare performance among stocks of different characteristics using controlled factorial design. We evaluated the performance of all six methods categorically (i.e., how well does each method classify a stock as under, fully, or overexploited?) and three methods continuously (i.e., how well does each method estimate B/B_{MSY}?). These metrics are widely applicable and these datasets are publicly available, and we hope to provide a consistent framework for comparing the performance of future data-limited assessment methods.

2.2 Data sources

We applied six catch-only stock assessment methods (**Table 1**) to observed stocks (n=168) in the RAM Legacy Stock Assessment Database (RAMLDB v. 2.95; Ricard et al. 2012) and simulated stocks (n=5760) from Rosenberg et al. (2014).

The RAMLDB is a global database of catch data and stock assessment output, including reference points and time series of biomass and fishing mortality. We evaluated 172 of the 185 RAMLDB stocks used by Free et al. (in review) with catch time series ≥ 20 years; the six methods converged on status predictions for all but four stocks yielding a final sample size of 168 RAMLDB stocks (**Table 2**). These stocks include those that are currently underexploited (n=63; 37.5%), fully exploited (n=83; 49.4%), and overexploited (n=22; 13.1%) and represent a variety of taxa, geographic locations, and management agencies (**Figure 1**). The Rosenberg et al. (2014) simulated stocks represent a fully factorial dataset of simulated fisheries including three fish life histories, three levels of initial biomass depletion, four exploitation scenarios, two levels of recruitment variability, two levels of recruitment autocorrelation, and two levels of measurement error, with each combination of these parameters run through ten stochastic iterations (**Supp. Table 1**). The six methods converged on status predictions for 5759 of 5760 stocks (**Table 2**) including stocks that are underexploited (n=1488; 25.8%), fully exploited (n=3416; 59.3%), and overexploited (n=856; 14.9%) in the final year of their time series (**Figure 2**). Both the RAMLDB and Rosenberg et al. (2014) datasets are publicly available and can be used to compare future data-limited methods to those evaluated in the present analysis.

In addition to catch time series, some of the evaluated methods require estimates of resilience (used by cMSY) or natural mortality (used by Zhou-BRT and Zhou-OCOM; **Table 1**). For RAMLDB stocks, almost all resilience values were retrieved from FishBase (Froese & Pauly 2016) and most natural mortality estimates were taken from stock assessments, though a few values were also recorded from FishBase (**Supp. Figure. 1**). For simulated stocks, resilience and natural mortality were derived for each life history category (**Supp. Table 2**). Demersal and large pelagic fish were assigned low resilience while small pelagic fish were assigned medium resilience. Natural mortality rates for each life history were derived from assumed maximum ages (t_{max}) using the Hoenig (1983) estimator, which is thought to perform best amongst the life history invariant methods (Then et al. 2014). The refined ORCS (rORCS) approach requires answers to questions regarding both ecological and fishery characteristics of each stock (**Supp. Table 3**) and cannot be evaluated on simulated stocks. This method is only evaluated on RAMLDB stocks with the question answers provided by Free et al. (in review).

2.3 Catch-only assessment methods

We evaluated the following six catch-only stock assessment methods (**Table 1**):

1. **rORCS (refined ORCS approach)**: The refined ORCS approach, developed by Free et al. (in review) based on the work of Berkson et al. (2011), estimates stock status (i.e., under, fully, or overexploited) from twelve stock- and fishery-related predictors (**Supp. Table 3**; e.g., status of data-rich stocks in the fishery, ex-vessel price, life history, etc.) using a boosted classification tree model fit to stocks in the RAMLDB. We implemented this method using code from Free et al. (in review).

2. **cMSY (catch-MSY):** The catch-MSY approach, originally developed by Martell & Froese (2013) and updated by Froese et al. (in press), estimates stock status (B/B_{MSY}) and reference points using a stock reduction analysis based on a times series of catch and priors for *r, k*, and initial/final depletion derived from resilience. We implemented this method using code from Froese et al. (in press).

3. **SSP-2002 (original stock status plot method):** The original stock status plot method, developed by Froese & Kesner-Reyes (2002), estimates stock status (i.e., 'undeveloped', 'developing', 'fully exploited', 'overfished', or 'collapsed') based on a comparison of the current year's catch to the maximum year's catch. We implemented this method using the rules specified in **Supp. Table 4**.

4. **SSP-2013 (updated stock status plot method):** The original stock status plot method was updated by Kleisner et al. (2013) to include an additional 'rebuilding' category by considering the minimum catch occurring after the maximum catch as well as the current

year's catch relative to the maximum year's catch. We implemented this method using the rules specified in **Supp. Table 5**.

5. **Zhou-BRT (catch-only boosted regression trees):** The Zhou-BRT method estimates saturation (i.e., 1 – depletion) using a boosted regression tree (BRT) model fit to catch trends occurring within the catch times series of stocks in the RAMLDB (Zhou et al. in review a). We implemented this method using code from Zhou et al. (in review a).
6. **Zhou-OCOM (optimized catch-only model):** The Zhou-OCOM method estimates saturation (i.e., 1 – depletion) using a stock reduction analysis based on a time series of catch and priors for *r*, derived from natural mortality, and stock depletion, derived from the Zhou-BRT approach (Zhou et al. in review b). We implemented this method using code from Zhou et al. (in review b).

2.4 Method performance

The six catch-only assessment methods evaluated here provide a mixture of continuous (i.e., B/B_{MSY} or saturation) and categorical (i.e., exploitation or development status) status predictions. We compared predictive performance categorically for all six methods and continuously for the three methods providing continuous status predictions (cMSY, Zhou-BRT, Zhou-OCOM). To compare performance categorically, we standardized the benchmark statuses (RAMLDB=data-rich assessment status; simulated stocks=known status) and predicted statuses to the following three exploitation categories: underexploited, fully exploited, and overexploited. We mapped the SSP development categories to exploitation categories using the rules shown in **Supp. Tables 4 and 5** and B/B_{MSY} (cMSY and benchmark statuses) and saturation (Zhou-BRT and Zhou-OCOM) to exploitation categories using the rules shown in **Supp. Table 6.** To compare the Zhou-BRT and Zhou-OCOM methods' performance continuously, we converted estimates of saturation (S) to B/B_{MSY} using the simple formula: $B/B_{MSY} = 2*S$ (**Supp. Proof 1**).

We evaluated the classification accuracy (categorical performance) of each method using both percentage agreement and Cohen's kappa. Cohen's kappa measures inter-rate agreement between categorical items and is more robust than simple percentage agreement because it takes into account the probability of agreement occurring by chance alone (Cohen 1968). This is necessary for tests on both RAMLDB and simulated stocks given the uneven distribution of exploitation categories in both datasets (many fully exploited stocks, few overexploited stocks; **Figures 1E and 2**). For example, if a method misclassifies most overexploited stocks but correctly classifies most fully exploited stocks, it would still earn a high accuracy percentage, but it's kappa value would be appropriately penalized. Although there are no definitive rules for interpreting Cohen's kappa, general guidelines suggest that values >0.70 are 'excellent', 0.4-0.7 are 'good', 0.2-0.4 are 'fair', and <0.2 are 'poor' (Landis & Koch 1977; Fleiss 1981).

We evaluated the continuous performance of the cMSY, Zhou-BRT, and Zhou-OCOM methods by measuring each method's bias, accuracy, and ability to correctly rank or correlate across populations. We measured bias as the median proportional error (MPE) and accuracy as the median absolute proportional error (MAPE). Proportional error is calculated as $\left(\hat{\theta} - \theta\right) / |\theta|$, where $\hat{\theta}$ and θ represent predicted and "true" (or data-rich stock assessment) B/B_{MSY} values. We measured the ability to correctly rank populations as Spearman's rank-order correlation between predicted and "true" values. These same performance metrics were used in Anderson et al. (in press)'s assessment of four other data-limited assessment methods and could be used to measure and compare the performance of future data-limited methods.

2.4 Method performance across stock characteristics

We used the factorial design of the simulated stock dataset to evaluate and compare the performance of each method on stocks of varying: (1) life histories; (2) initial biomass depletion levels: (3) exploitation dynamics; and (4) catch time series lengths (see **Supp. Table 1** for the levels within each category). We assessed classification performance using the categorical framework described above. All analyses were performed in R v.3.3.2 (R Core Team 2016).

3. Results

3.1 Performance on RAMLDB stocks

The refined ORCS approach was the best classifier of stock status in terms of both Cohen's kappa and accuracy on RAMLDB stocks in both the full and test datasets **(Table 3)**. In the test dataset, the rORCS approach was 79.4% accurate (0.644 kappa), 38.2-61.8% more accurate (0.537-0.688 higher kappas) than the other methods. cMSY was second best classifier in terms of kappa for both datasets, but its performance was generally comparable to SSP-2002, Zhou-OCOM, and Zhou-OCOM in both kappa and accuracy (**Table 3**). SSP-2013 performed worse than a random classifier and was consistently the worst classifier (**Table 3**). The rORCS method was biased towards central or optimistic classification while the other methods were biased towards pessimistic classifications (**Tables 3-5**). cMSY was a more accurate predictor of B/B_{MSY} than the Zhou methods (**Figure 3**) and the methods were relatively uncorrelated (**Figure 4**).

3.2 Performance on simulated stocks

cMSY performed marginally better than other methods on the simulated stocks (though the rORCS could not be tested on these stocks) in terms of both kappa and accuracy (**Table 6**). Although SSP-2013 and SSP-2002 were the next best classifiers in terms of kappa, they failed to classify a single stock as either fully or underexploited, respectively (**Table 7**). Zhou-BRT performed better than Zhou-OCOM in terms of kappa and accuracy, though Zhou-OCOM was slightly less pessimistic than Zhou-BRT (**Tables 6 and 7**). cMSY was also a more accurate predictor of B/B_{MSY} than the Zhou et al. methods (**Figure 3**) and predictions from cMSY and the Zhou et al. methods were relatively uncorrelated (**Figure 4**) especially for scenarios of constant or increasing exploitation rate (**Figure 5**) where neither method performed particularly well. Zhou-OCOM rarely predicted B/B_{MSY} values between 1.0 and 1.3 for either the RAMLDB or simulated stocks (**Figure 6**). Although more accurate than other methods, cMSY was heavily pessimistic, correctly classifying only four (0.3%) stocks as underexploited (**Table 7**).

3.3 Performance among stock types

In terms of kappa, cMSY, SSP-2002, and SSP-2013 were consistently the best classifiers across the life history (**Table 8**), initial depletion (**Table 9**), exploitation dynamics (**Table 10**), and time series length (**Table 11**) groups represented in the simulated stocks. cMSY was the best classifier in terms of kappa for: small pelagic fish and stocks exhibiting biomass-coupled and roller coaster exploitation dynamics. SSP-2002 was the best classifier in terms of kappa for: large pelagic fish and stocks at 40% of carrying capacity, exhibiting increasing exploitation, and with 20 yr catch time series. SSP-2013 was the best classifier in terms of kappa for: demersal fish and stocks at 70% and 100% of carrying capacity, exhibiting constant exploitation, and with 60 yr catch time series. For many groups where cMSY was not the best classifier in terms of kappa, it was the best classifier in terms of accuracy (**Table 12**).

4. Discussion

Based on the results of this analysis and information from the scientific literature that resulted in our initial model selection, we offer two short-term recommendations for the application of data-limited models to FAO stocks and three longer-term recommendations. These recommendations apply to situations where the consequences of errors in status estimation are symmetric, e.g. where a 30% under-estimate of stock status is equally consequential as a 30% over-estimate of stock status. These recommendations do not necessarily apply to management of individual stocks where underestimation and overestimation of stock status may carry different consequences.

Short-term recommendations:

1. *When sufficient ancillary information is available to permit the use of rORCS, and the goal is a single estimate of current stock status, this method should be preferred.* However, caution is warranted because of the inherent inability to test rORCS with simulated data and the fact that rORCS was trained on data from assessed stocks which represent a non-random sample of global stocks. *In the absence of such ancillary information, or when estimates of status through time are needed, cMSY or a superensemble (Anderson et al. in press; not evaluated here) should generally be used.* The only exceptions are when stocks are known to have experienced severe initial depletion before the start of the catch time series or if exploitation rate is known to be constant or increasing.

2. *Avoid using SSP-2002 or SSP-2013 to determine stock status unless exploitation rate is known to be constant or increasing.* These two methods failed to classify a single stock as either fully (SSP-2013) or underexploited (SSP-2002). They also generally performed more poorly in terms of accuracy, though not for kappa, than the other four methods. The only exceptions were: (a) for stocks with initial depletion at the start of the catch time series of 60%, i.e., starting at 40% of carrying capacity, where SSP-2002 performed similarly to Zhou-BRT and better than cMSY in terms of kappa, but not accuracy; and (b) for stocks with a constant or increasing exploitation rate. For a constant exploitation rate, SSP-2013 performed better on both metrics than all other methods. For an increasing exploitation rate, SSP-2002 performed better on both metrics than all other methods.

Longer-term recommendations:

1. *Conduct additional testing of rORCS to evaluate its potential performance on unassessed stocks, and develop a data set of fishery characteristics to allow application of rORCS to the FAO stocks.* The rORCS method performs substantially better than all other methods in testing on real assessed stocks. This remains true in cross-validation against a random selection of real assessed stocks. However, assessed stocks represent a non-random subset of all stocks and are likely to differ systematically from the unassessed stocks to which rORCS would be applied. Further evaluation of rORCS is required before it can be recommended unequivocally. An additional limitation to widespread application of rORCS is the lack of a global database of scores for the rORCS table of attributes. Surveys or guided workshops with regional fishery experts may be used to develop such a database. It is also unclear how well rORCS would predict change in stock status through time as the method has never been evaluated for such an application.

2. *Develop and test superensemble approaches that incorporate the Zhou et al. (in review a, b) methods and, potentially, the two SSP methods*. When the superensembles were developed by Anderson et al. (in press), the Zhou et al. methods were not yet available. The relatively low correlation between status estimates from the Zhou et al. methods and cMSY (**Figures 4 and 5**) suggests that there is additional information to be gained from the combination of these two methods in a superensemble. Divergence of estimates from the two methods may also be a useful diagnostic of constant or increasing effort, the two effort dynamics scenarios in which both the Zhou et al. methods and cMSY performed relatively poorly.

3. *Develop and test methods for identifying effort dynamics and incorporate this information into data-limited assessment methods.* While cMSY generally performed better than the other methods, none of the methods can be recommended unequivocally. In particular, the performance of cMSY was quite poor (kappa = -0.029, accuracy = 0.245) for the constant exploitation rate scenarios (**Table 10**). Of all the variables examined in the simulations (life history, initial depletion, time series length, and effort dynamics), differences in performance of the different methods were most pronounced across different effort dynamics scenarios. This was not surprising. All of the methods attempt to interpret changes in catch, but catch can change because of a change in abundance or a change in fishing effort. Effort dynamics is also possibly the most difficult variable to identify from external sources. Time series length is clear and databases of life history characteristics exist (e.g., FishBase). Effort dynamics are not apparent from the catch time series alone. However, in most cases, experts familiar with a fishery can likely determine the general trend in effort (stable, increasing, or decreasing). Expert information or new analytical methods to identify effort dynamics combined with superensembles will likely prove to be the most powerful approach to stock status estimation based on only a catch time series.

References

Anderson, S.C, Cooper, A.B, Jensen, O.P., Minto, C., Thorson, J.T., Walsh, J.C., Afflerbach, J, Dickey-Collas, M. Kleisner, K.M., Longo, C., Osio, G.C., Ovando, D., Mosqueira, I., Rosenberg, A.A., Selig, E.R. (*in press*) Improving estimates of population status and trend with superensemble models. *Fish & Fisheries*.

Anderson, S.C., Afflerbach, J., Cooper, A.B., Dickey-Collas, M., Jensen, O.P., Kleisner, K.M., Longo, C., Osio, G.C., Ovando, D., Minte-Vera, C., Minto, C., Mosqueira, I., Rosenberg, A.A., Selig, E.R., Thorson, J.T., Walsh, J.C. (2016). datalimited: Stock assessment methods for data-limited fisheries. R package version 0.0.2. Available at: https://github.com/datalimited/datalimited

Berkson, J., Barbieri, L., Florida, F.W.C., Cass-Calay, S., Crone, P., Dorn, M., Friess, C., Kobayashi, D., Miller, T.J., Patrick, W.S., Pautzke, S., Ralston, S., Trianni, M. (2011) Calculating Acceptable Biological Catch for stocks that have reliable catch data only (Only Reliable Catch Stocks – ORCS). NOAA Technical Memorandum NMFS-SEFSC-616. NMFS-SEFSC: Miami, FL.

Carruthers, T.R., Walters, C.J., McAllister, M.K. (2012) Evaluating methods that classify fisheries stock status using only fisheries catch data. *Fisheries Research* 119-120: 66-79.

Cohen, J. (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4): 213-220.

Costello, C., Ovando, D., Hilborn, R., Gains, S.D., Deschenes, O., Lester, S.E. (2012) Status and solutions for the world's unassessed fisheries. *Science* 338: 517-520.

Costello, C., Ovando, D., Clavelle, T., Strauss, C.K., Hilborn, R., Melnychuk, M.C., Branch, T.A., Gaines, S.D., Szuwalski, C.S., Cabral, R.B., Rader, D.N. (2016) Global fishery prospects under contrasting management regimes. *Proceedings of the National Academy of Sciences* 113(18): 5125-5129.

Fleiss, J.L. (1973) Statistical methods for rates and proportions. John Wiley & Sons: New York, NY. 233 pp.

Free, C.M., Jensen, O.P., Wiedenmann, J., Deroba, J.J. (*in review*) The refined ORCS approach: a catch-based method for estimating stock status and catch limits for data-poor fish stocks. In review at *Fisheries Research*.

Froese, R., Demirel, N., Coro, G., Kleisner, K.M., Winker, H. (*in press*) Estimating fisheries reference points from catch and resilience. Fish &Fisheries: DOI:10.1111/faf.12190

Froese, R., Pauly, D. (2016) FishBase. Available at: www.fishbase.org

Hoenig, J.M. (1983) Empirical use of longevity data to estimate mortality rates. *Fishery Bulletin* 82: 898–903.

Kleisner, K., Zeller, D., Froese, R., Pauly, D. (2013) Using global catch data for inferences on the world's marine fisheries. *Fish and Fisheries* 14(3): 293-311.

Landis, J.R., Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159-174.

Martell, S., Froese, R. (2013) A simple method for estimating MSY from catch and resilience. *Fish and Fisheries* 14: 504-514.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. Available at: http://www.R-project.org/

Ricard, D., Minto, C., Jensen, O.P., Baum, J.K. (2012) Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. *Fish and Fisheries* 13(4): 380-398.

Rosenberg, A.A., Fogarty, M.J., Cooper, A. B., Dickey-Collas, M., Fulton, E.A., Gutiérrez, N.L., Hyde, K.J.W., Kleisner, K.M., Longo, C., Minte-Vera, C.V., Minto, C., Mosqueira, I., Osio, G.C., Ovando, D., Selig, E.R., Thorson, J.T., Ye, Y. (2014) Developing new approaches to global stock status assessment and fishery production potential of the seas. FAO Fisheries and Aquaculture Circular: Rome, Italy.

Then, A. Y., Hoenig, J. M., Hall, N. G., Hewitt, D. A. (2014) Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. *ICES Journal of Marine Science*: fsu136.

Thorson, J.T., Branch, T.A., Jensen, O.P. (2012) Using model-based inference to evaluate global fisheries status from landings, location, and life history data. *Canadian Journal of Fisheries and Aquatic Sciences* 69(4): 645-655.

Thorson, J.T., Minto, C., Minte-Vera, C.V., Kleisner, K.M., Longo, C. (2013) A new role for effort dynamics in the theory of harvested populations and data-poor stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences* 70(12): 1829-1844.

Vasconcellos, M., Cochrane, K. (2005) Overview of world status of data-limited fisheries: inferences from landings statistics. In: Kruse, G.H., Gallucci, V.F., Hay, D.E., Perry, R.I., Peterman, R.M., Shirley, T.C., Spencer, P.D., Wilson, B., Woodby, D. (eds.): Fisheries Assessment and Management in Data-Limited Situations. Alaska Sea Grant College Program, University of Alaska Fairbanks: Fairbanks, AK. 1–20 pp.

Zhou, S., Punt, A.E., Ye, Y., Ellis, N., Dichmont, C.M., Haddon, M., Smith, D.C., Smith, A.D.M. (in review a). Estimating stock depletion level from patterns of catch history. In review at *Fish & Fisheries*.

Zhou et al. (in review b) An optimised catch-only assessment method for data poor fisheries.

Tables & Figures

Supplemental Tables & Figures

Appendices

Table 1. Catch-only stock assessment methods evaluated in the present study.

Method			
Short name	**Full name *(reference)***	**Required inputs**	**Outputs**
SSP-2002	Stock status plot method *(Froese & Kesner-Reyes 2002)*	Catch	Development status
SSP-2013	Updated stock status plot method *(Kleisner et al. 2013)*	Catch	Development status
rORCS	Refined ORCS approach *(Free et al. in review)*	Catch, 12 questions	Exploitation status, OFL
cMSY	Catch-MSY method *(Froese et al. in press)*	Catch, resilience	B/B_{MSY}, MSY reference points
Zhou-BRT	Catch-only boosted regression trees *(Zhou et al. in review a)*	Catch, natural mortality	Saturation
Zhou-OCOM	Optimised catch-only assessment method *(Zhou et al. in review b)*	Catch, natural mortality	Saturation, MSY

Table 2. Convergence success for catch-only assessment methods fit to the RAMLDB and simulated stock datasets.

	Number (percentage) of stocks failing to converge	
Method	**RAMLDB stocks (n=172)**	**Simulated stocks (n=5760)**
SSP-2002	0	0
SSP-2013	0	0
rORCS	0	----
cMSY	3 (1.8%)	1 (0.02%)
Zhou-BRT	0	0
Zhou-OCOM	1 (0.6%)	0
Final sample size:	168	5759

Table 3. Status classification performance of catch-only assessment methods applied to the RAMLDB stocks used to both train and test (n=168; full dataset) and only test (n=34; test dataset) the refined ORCS approach (in order of decreasing kappa values).

Method	Kappa	Accuracy	Bias
Full dataset (162 stocks)			
rORCS	0.543	0.738	central
cMSY	0.137	0.405	pessimistic
SSP-2002	0.101	0.339	pessimistic
Zhou-OCOM	0.099	0.446	pessimistic
Zhou-BRT	0.089	0.452	pessimistic
SSP-2013	0.045	0.256	pessimistic
Test dataset (34 stocks)			
rORCS	0.644	0.794	slightly optimistic
cMSY	0.107	0.353	pessimistic
SSP-2002	0.095	0.324	pessimistic
Zhou-OCOM	0.049	0.412	slightly pessimistic
Zhou-BRT	-0.035	0.382	pessimistic
SSP-2013	-0.044	0.176	pessimistic

Table 4. Stock status predictions for the 168 RAMLDB stocks in the full dataset (grey shading indicates correct predictions; in order of decreasing success rates).

			n(predicted)			Success	Proportional
Method	**Status**	**n(observed)**	**Underexploited**	**Fully exploited**	**Overexploited**	**rate (%)**	**error (%)**
rORCS	underexploited	63	40	23	0	63.5%	-17.5%
	fully exploited	83	11	71	1	85.5%	22.9%
	overexploited	22	1	8	13	59.1%	-36.4%
	Total or mean:	168	52	102	14	(124) 73.8%	25.6% (absolute)
Zhou-BRT	underexploited	63	5	46	12	7.9%	-85.7%
	fully exploited	83	3	61	19	73.5%	42.2%
	overexploited	22	1	11	10	45.5%	86.4%
	Total or mean:	168	9	118	41	(76) 45.2%	71.4% (absolute)
Zhou-OCOM	underexploited	63	13	39	11	20.6%	-55.6%
	fully exploited	83	13	51	19	61.4%	19.3%
	overexploited	22	2	9	11	50.0%	86.4%
	Total or mean:	168	28	99	41	(75) 44.6%	53.7% (absolute)
cMSY	underexploited	63	0	33	30	0.0%	-100.0%
	fully exploited	83	0	48	35	57.8%	0.0%
	overexploited	22	0	2	20	90.9%	286.4%
	Total or mean:	168	0	83	85	(68) 40.5%	128.8% (absolute)
SSP-2002	underexploited	63	0	23	40	0.0%	-100.0%
	fully exploited	83	0	37	46	44.6%	-25.3%
	overexploited	22	0	2	20	90.9%	381.8%
	Total or mean:	168	0	62	106	(57) 33.9%	169.0% (absolute)
SSP-2013	underexploited	63	23	0	40	36.5%	-1.6%
	fully exploited	83	37	0	46	0.0%	-100.0%
	overexploited	22	2	0	20	90.9%	381.8%
	Total or mean:	168	62	0	106	(43) 25.6%	161.1% (absolute)

Table 5. Stock status predictions for the 34 RAMLDB stocks in the test dataset (grey shading indicates correct predictions; in order of decreasing success rates).

Method	Status	n(observed)	n(predicted) Underexploited	Fully exploited	Overexploited	Success rate (%)	Proportional error (%)
rORCS	underexploited	13	10	3	0	76.9%	0.0%
	fully exploited	17	3	14	0	82.4%	5.9%
	overexploited	4	0	1	3	75.0%	-25.0%
	Total or mean:	34	13	18	3	(27) 79.4%	10.3% (absolute)
Zhou-OCOM	underexploited	13	2	7	4	15.4%	-61.5%
	fully exploited	17	2	11	4	64.7%	17.6%
	overexploited	4	1	2	1	25.0%	125.0%
	Total or mean:	34	5	20	9	(14) 41.2%	68.1% (absolute)
Zhou-BRT	underexploited	13	1	8	4	7.7%	-84.6%
	fully exploited	17	1	12	4	70.6%	41.2%
	overexploited	4	0	4	0	0.0%	100.0%
	Total or mean:	34	2	24	8	(13) 38.2%	75.3% (absolute)
cMSY	underexploited	13	0	4	9	0.0%	-100.0%
	fully exploited	17	0	9	8	52.9%	-17.6%
	overexploited	4	0	1	3	75.0%	400.0%
	Total or mean:	34	0	14	20	(12) 35.3%	172.5% (absolute)
SSP-2002	underexploited	13	0	3	10	0.0%	-100.0%
	fully exploited	17	0	8	9	47.1%	-29.4%
	overexploited	4	0	1	3	75.0%	450.0%
	Total or mean:	34	0	12	22	(11) 32.4%	193.1% (absolute)
SSP-2013	underexploited	13	3	0	10	23.1%	-7.7%
	fully exploited	17	8	0	9	0.0%	-100.0%
	overexploited	4	1	0	3	75.0%	450.0%
	Total or mean:	34	12	0	22	(6) 17.6%	185.9% (absolute)

Table 6. Status classification performance of catch-only assessment methods applied to the Rosenberg et al. (2014) simulated stocks (n=5759; in order of decreasing kappa values).

Method	Kappa	Accuracy	Bias
cMSY	0.085	0.512	pessimistic
SSP-2013	0.082	0.274	to extremes
SSP-2002	0.080	0.442	pessimistic
Zhou-BRT	0.023	0.487	pessimistic
Zhou-OCOM	0.011	0.470	slightly pessimistic

Table 7. Stock status predictions for the 5759 Rosenberg et al. (2014) simulated stocks (grey shading indicates correct predictions; in order of decreasing success rates).

			n(predicted)			Success	Proportional
Method	**Status**	**n(observed)**	**Underexploited**	**Fully exploited**	**Overexploited**	**rate (%)**	**error (%)**
cMSY	underexploited	1488	4	1119	365	0.3%	-99.5%
	fully exploited	3415	2	2545	868	74.5%	20.6%
	overexploited	856	1	455	400	46.7%	90.8%
	Total or mean:	5759	7	4119	1633	(2949) 51.2%	70.3% (absolute)
Zhou-BRT	underexploited	1488	36	1158	294	2.4%	-88.8%
	fully exploited	3415	127	2471	817	72.4%	22.5%
	overexploited	856	4	555	297	34.7%	64.5%
	Total or mean:	5759	167	4184	1408	(2804) 48.7%	58.6% (absolute)
Zhou-OCOM	underexploited	1488	238	1061	189	16.0%	-40.4%
	fully exploited	3415	586	2238	591	65.5%	13.1%
	overexploited	856	63	565	228	26.6%	17.8%
	Total or mean:	5759	887	3864	1008	(2704) 47.0%	23.8% (absolute)
SSP-2002	underexploited	1488	0	976	512	0.0%	-100.0%
	fully exploited	3415	0	1942	1473	56.9%	-7.1%
	overexploited	856	0	253	603	70.4%	202.3%
	Total or mean:	5759	0	3171	2588	(2545) 44.2%	103.2% (absolute)
SSP-2013	underexploited	1488	976	0	512	65.6%	113.1%
	fully exploited	3415	1942	0	1473	0.0%	-100.0%
	overexploited	856	253	0	603	70.4%	202.3%
	Total or mean:	5759	3171	0	2588	(1579) 27.4%	138.5% (absolute)

Table 8. Status classification performance of catch-only assessment methods on life histories represented in the Rosenberg et al. (2014) simulated stocks.

Method	Kappa	Accuracy	Method	Kappa	Accuracy	Method	Kappa	Accuracy
Demersal fish (n=1920)			*Small pelagic fish (n=1919)*			*Large pelagic fish (n=1920)*		
SSP-2013	0.095	0.291	cMSY	0.127	0.552	SSP-2002	0.090	0.486
cMSY	0.054	0.466	SSP-2002	0.099	0.440	cMSY	0.077	0.519
SSP-2002	0.045	0.399	SSP-2013	0.079	0.263	SSP-2013	0.063	0.269
Zhou-BRT	0.042	0.489	Zhou-BRT	0.069	0.538	Zhou-OCOM	0.004	0.458
Zhou-OCOM	0.041	0.478	Zhou-OCOM	0.005	0.472	Zhou-BRT	-0.031	0.433

Table 9. Status classification performance of catch-only assessment methods on initial biomass depletion levels represented in the Rosenberg et al. (2014) simulated stocks.

Method	Kappa	Accuracy	Method	Kappa	Accuracy	Method	Kappa	Accuracy
100% of carrying capacity (n=1919)			*70% of carrying capacity (n=1920)*			*40% of carrying capacity (n=1920)*		
SSP-2013	0.130	0.308	SSP-2013	0.072	0.298	SSP-2002	0.173	0.562
cMSY	0.074	0.498	cMSY	0.064	0.428	cMSY	0.122	0.610
SSP-2002	0.028	0.403	SSP-2002	0.050	0.361	Zhou-BRT	0.100	0.567
Zhou-OCOM	-0.001	0.493	Zhou-BRT	-0.010	0.407	Zhou-OCOM	0.069	0.508
Zhou-BRT	-0.002	0.487	Zhou-OCOM	-0.018	0.408	SSP-2013	0.059	0.217

Table 10. Status classification performance of catch-only assessment methods for the four different effort dynamics scenarios represented in the Rosenberg et al. (2014) simulated stocks.

Method	Kappa	Accuracy	Method	Kappa	Accuracy	Method	Kappa	Accuracy	Method	Kappa	Accuracy
Constant F (n=1440)			*Biomass-coupled F (n=1440)*			*Increasing F (n=1440)*			*Roller coaster F (n=1439)*		
SSP-2013	0.088	0.610	cMSY	0.159	0.402	SSP-2002	0.454	0.735	cMSY	0.169	0.733
Zhou-OCOM	-0.017	0.288	SSP-2002	0.113	0.326	cMSY	0.277	0.668	Zhou-OCOM	0.076	0.603
Zhou-BRT	-0.022	0.211	SSP-2013	-0.009	0.141	Zhou-BRT	0.161	0.601	SSP-2002	0.066	0.518
cMSY	-0.029	0.245	Zhou-BRT	-0.026	0.476	Zhou-OCOM	0.148	0.576	Zhou-BRT	0.046	0.659
SSP-2002	-0.049	0.188	Zhou-OCOM	-0.031	0.410	SSP-2013	0.132	0.272	SSP-2013	0.029	0.074

Table 11. Status classification performance of catch-only assessment methods on time series lengths represented in the Rosenberg et al. (2014) simulated stocks.

Method	Kappa	Accuracy	Method	Kappa	Accuracy
20 yr catch time series (n=2880)			*60 yr catch time series (n=2879)*		
SSP-2002	0.158	0.534	SSP-2013	0.092	0.270
cMSY	0.133	0.568	Zhou-BRT	0.053	0.482
SSP-2013	0.072	0.278	cMSY	0.043	0.456
Zhou-OCOM	0.005	0.485	Zhou-OCOM	0.016	0.454
Zhou-BRT	-0.009	0.491	SSP-2002	0.015	0.350

Table 12. Best classifiers in terms of kappa and accuracy for stock types in the simulated stock dataset.

	Best method in terms of:	
Stock type	**Kappa**	**Accuracy**
Life history		
Demersal fish	SSP-2013	Zhou-BRT
Small pelagic fish	cMSY	cMSY
Large pelagic fish	SSP-2002	cMSY
Initial depletion		
100% carrying capacity	SSP-2013	cMSY
70% carrying capacity	SSP-2013	cMSY
40% carrying capacity	SSP-2002	cMSY
Exploitation dynamics		
Constant F	SSP-2013	SSP-2013
Biomass-coupled F	cMSY	Zhou-BRT
Increasing F	SSP-2002	SSP-2002
Roller coaster F	cMSY	cMSY
Time series length		
20 yr	SSP-2002	cMSY
60 yr	SSP-2013	Zhou-BRT

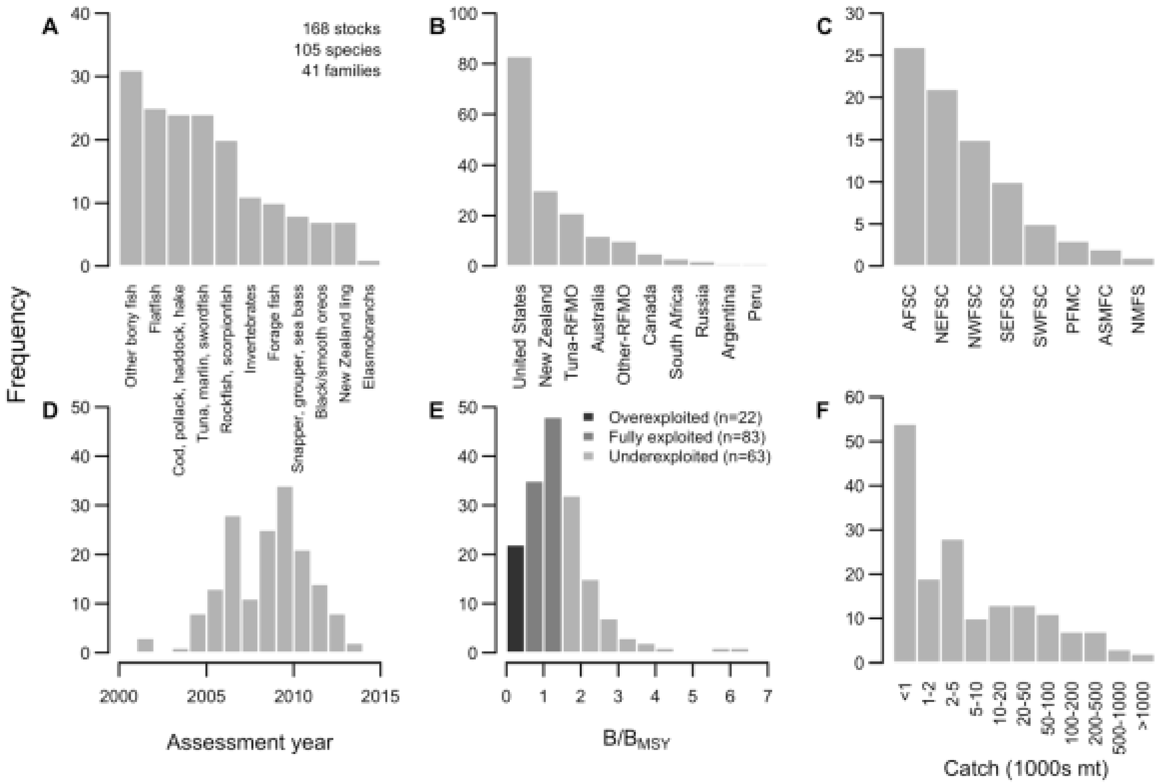


Figure 1. Demographics of the 168 data-rich stocks evaluated using catch-only assessment methods by: **(A)** taxonomic group; **(B)** managing country or multinational body; **(C)** U.S. assessment agency (U.S. stocks only; n=83, 49.4% of stocks); **(D)** assessment year; **(E)** stock status (B/B_{MSY}); and **(F)** fishery size (average annual catch over the most recent 5 years).

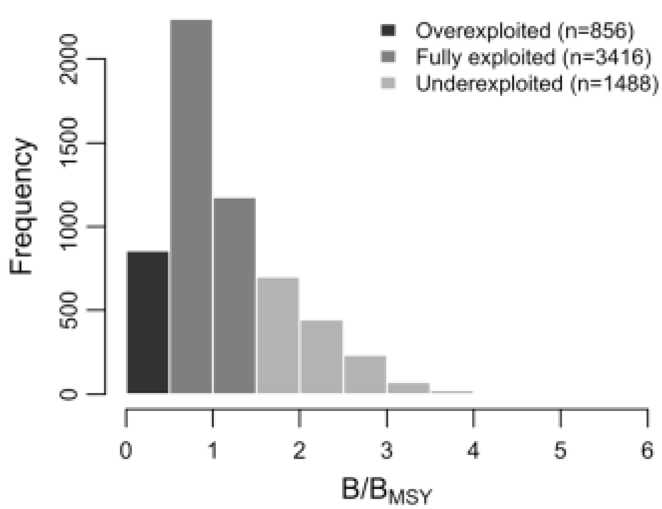


Figure 2. The status of the Rosenberg et al. (2014) simulated stocks in the final year of each time series (n=5760).

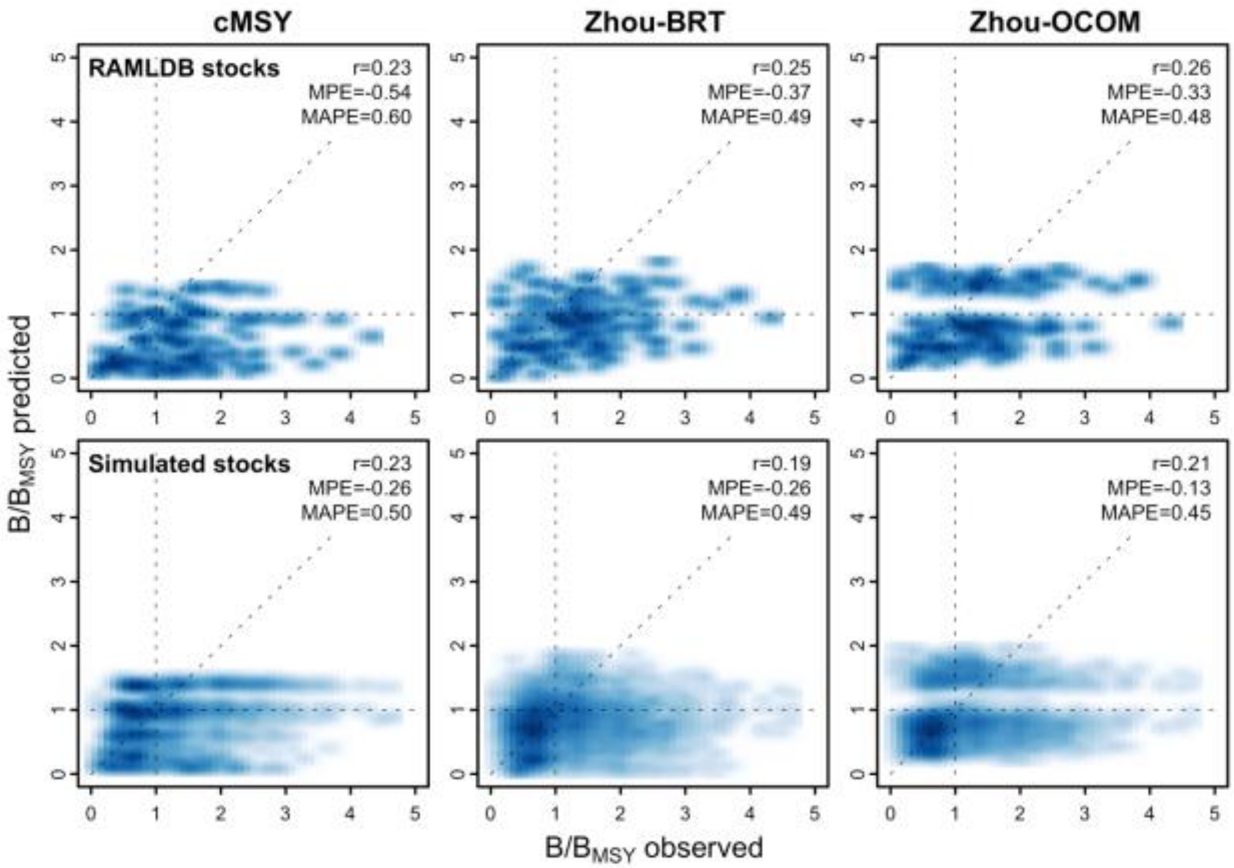


Figure 3. Observed population status vs. population status predicted by the cMSY, Zhou-BRT, and Zhou-OCOM methods for RAMLDB stocks (n=168) and simulated stocks (n=5759). Points are binned for visual presentation: darker areas indicate areas with greater density of points. Median proportional error (MPE) and median absolute proportional error (MAPE) measure bias and accuracy, respectively.

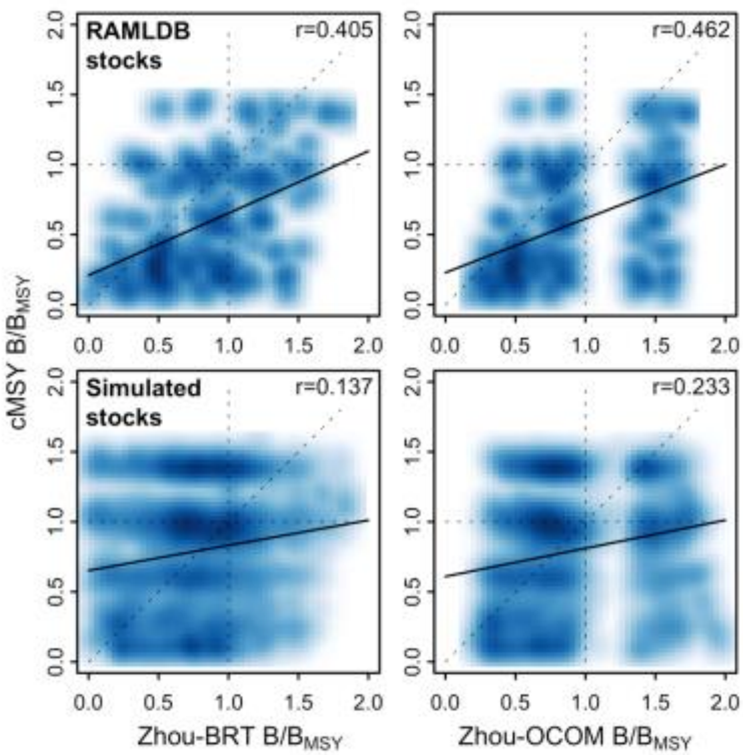


Figure 4. Correlation between B/B$_{MSY}$ predicted by cMSY and the Zhou-BRT and Zhou-OCOM methods for RAMLDB stocks (n=168) and simulated stocks (n=5759). Points are binned for visual presentation: darker areas indicate areas with greater density of points. Dark line indicates a linear regression fit to the data.

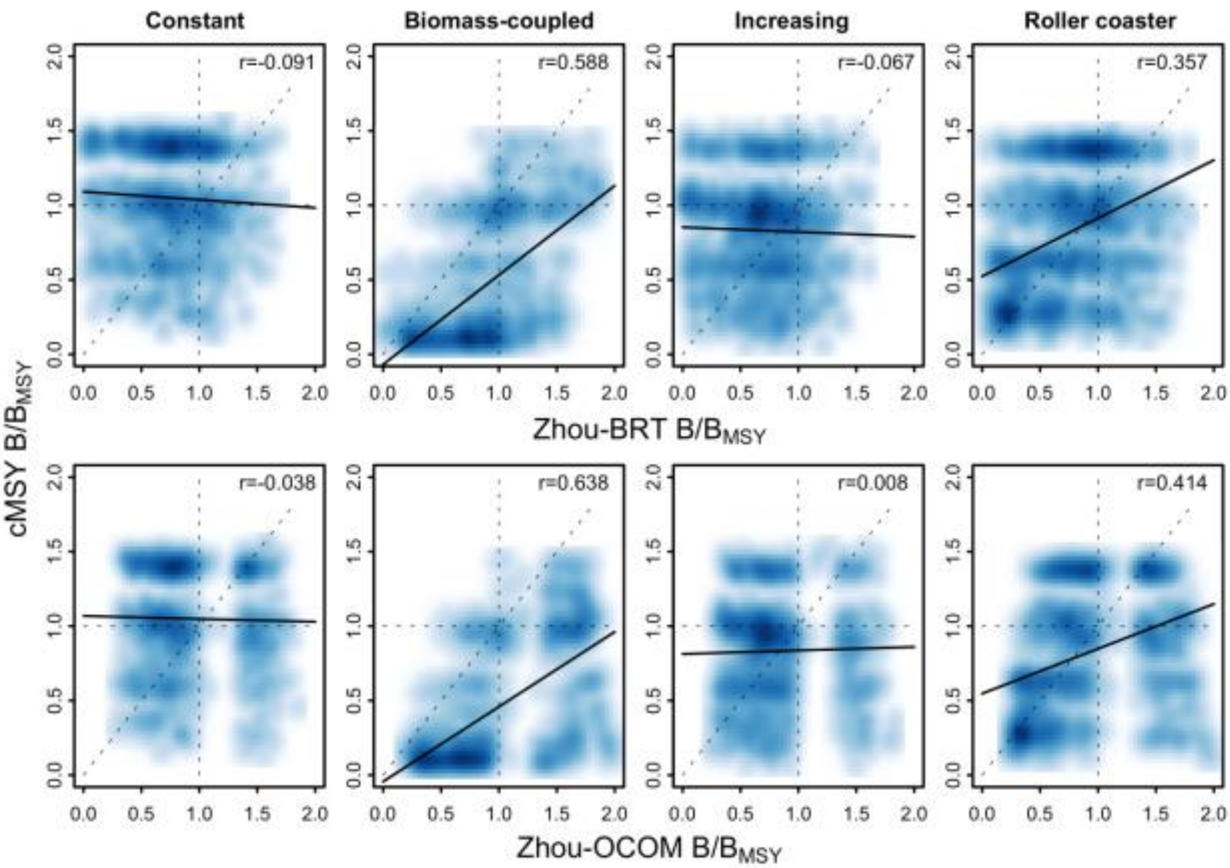


Figure 5. Correlation between B/B_{MSY} predicted by cMSY and the Zhou-BRT (top row) and Zhou-OCOM (bottom row) methods for Rosenberg et al. (2014) simulated stocks (n=5759) for each of the four different effort dynamics scenarios (columns). Points are binned for visual presentation: darker areas indicate areas with greater density of points. Dark line indicates a linear regression fit to the data.

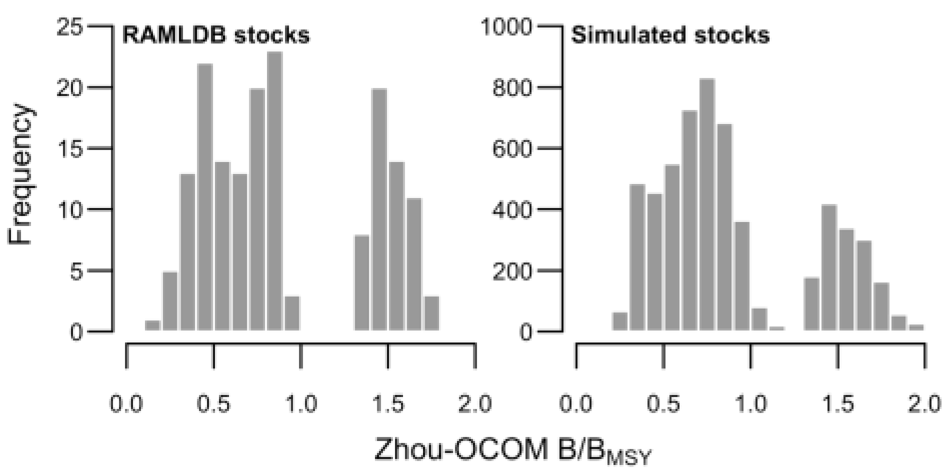


Figure 6. Zhou-OCOM B/B_{MSY} predictions for RAMLDB (n=168) and simulated (n=5759) stocks.

Supp. Table 1. Factorial design of the Rosenberg et al. (2014) simulated stocks.

Factor	# of levels	Levels
Life history	3	Demersal, small pelagic, or large pelagic
Initial biomass depletion	3	100%, 70%, or 40% of carrying capacity
Exploitation dynamics	4	Constant, biomass-coupled, increasing, or roller coaster rates
Recruitment variability	2	Low or high variability
Recruitment autocorrelation	2	With or without autocorrelation
Catch measurement error	2	With or without catch measurement error
Time series length	2	20 or 60 years
Iterations	10	Iterations for each combination of the above parameters
Total # of stocks:	5760	

Supp. Table 2. Resilience and natural mortality (M) values for the life histories represented in the Rosenberg et al. (2014) simulated stocks.

Life history category	Generic name	Resilience	L_{inf} (cm)	T_{max} (yr)	M (yr^{-1})*
Demersal	Gadoid	low	70	20	0.315
Small pelagic	Clupeoid	medium	30	8	0.729
Large pelagic	Scombrid	low	150	20	0.315

***** Estimated using the t_{max}-based Hoenig (1983) method: $M = 4.899*t_{max}^{-0.916}$

Supp. Table 3. Table of Attributes for the refined ORCS approach (Free et al. in review).

		Exploitation status		
#	**Attribute**	**Underexploited (1)**	**Fully exploited (2)**	**Overexploited (3)**
1	Status of assessed stocks in fishery	<10% overfished	10-25% overfished	>25% overfished
2	Refuge availability	***Not used in refined approach***		
3	Behavior affecting capture	-------------	No aggregation behavior	Exhibits aggregation behavior
4	Morphology affecting capture	***Not used in refined approach***		
5	Discard rate	<10% of catch discarded	10-25% of catch discarded	>25% of catch discarded
6	Targeting intensity	Not targeted	Occasionally targeted	Actively targeted
7	M compared to dominant species	Higher mortality rate	Equivalent mortality rates	Lower mortality rate
8	Occurrence in catch	Sporadic (in <10% of efforts)	Common (in 10-25% of efforts)	Frequent (in >25% of efforts)
9	Value (US$/lb, 5-year mean)	***Continuous value in refined approach***		
10	Recent trend in catch	Increasing last 5 years	Stable last 5 years	Decreasing last 5 years
11	Habitat loss	No time in threatened habitats	Part time in threatened habitats	Full time in threatened habitats
12	Recent trend in effort	Decreasing last 5 years	Stable last 5 years	Increasing last 5 years
13	Recent trend in abundance index	Increasing last 5 years	Stable last 5 years	Decreasing last 5 years
14	Proportion of population protected	Most of resource protected	Some of resource protected	None of resource protected

Supp. Table 4. Criteria used to classify stock status in SSP-2002 (Froese & Kesner-Reyes 2002).*

Stock status	SSP-2002 status	Criteria
Underexploited	Undeveloped	C_{curr} before C_{max} **AND** $C_{curr} < 0.1*C_{max}$
Underexploited	Developing	C_{curr} before C_{max} **AND** $0.1*C_{max} \leq C_{curr} \leq 0.5*C_{max}$
Fully exploited	Fully exploited	$C_{curr} > 0.5*C_{max}$
Overexploited	Overfished	C_{curr} after C_{max} **AND** $0.1*C_{max} \leq C_{curr} \leq 0.5*C_{max}$
Overexploited	Collapsed/closed	C_{curr} after C_{max} **AND** $C_{curr} < 0.1*C_{max}$

* C_{curr} = current catch; C_{max} = maximum catch

Supp. Table 5. Criteria used to classify stock status in SSP-2013 (Kleisner et al. 2013).*

Stock status	SSP-2013 status	Criteria
Underexploited	Developing	C_{curr} before C_{max} **AND** $C_{curr} \leq 0.5*C_{max}$ **OR** C_{max} in final year of time series
Fully exploited	Exploited	$C_{curr} > 0.5*C_{max}$
Overexploited	Overexploited	C_{curr} after C_{max} **AND** $0.1*C_{max} \leq C_{curr} \leq 0.5*C_{max}$
Overexploited	Collapsed	C_{curr} after C_{max} **AND** $C_{curr} < 0.1*C_{max}$
Overexploited	Rebuilding	C_{curr} after $C_{post\text{-}max\ min}$ **AND** $C_{post\text{-}max\ min}$ $0.1*C_{max}$ **AND** $0.1*C_{max} \leq C_{curr} \leq 0.5*C_{max}$

* C_{curr} = current catch; C_{max} = maximum catch; $C_{post\text{-}max\ min}$ = minimum catch after the maximum catch

Supp. Table 6. Mapping B/B_{MSY} and saturation values to exploitation status categories.

Category	B/B_{MSY}*	Saturation (S)**
underexploited	$B/B_{MSY} > 1.5$	$S > 0.75$
fully exploited	$0.5 < B/B_{MSY} < 1.5$	$0.25 < S < 0.75$
overexploited	$B/B_{MSY} < 0.5$	$S < 0.25$

* B/B_{MSY} estimated by cMSY and data-rich assessments
** Saturation estimated by Zhou et al. (in review a and b)

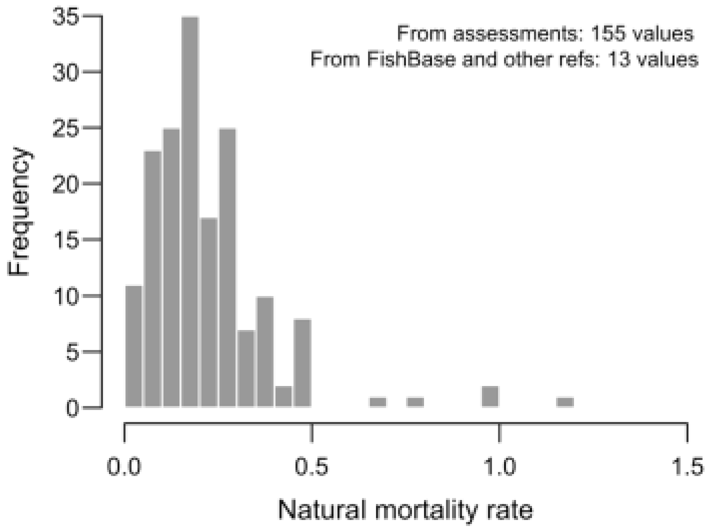
Supp. Proof 1. Method for estimating B/B_{MSY} from saturation (S).

In the logistic growth model, $dN/dt = rN(1-N/K)$, the following is true:

$B_{MSY} = K / 2$
$F_{MSY} = r / 2$
$MSY = rK / 4$

Saturation (S) is defined as $S = B/K$ so we can make the following substitution and rearrangement to estimate B/B_{MSY}:

$S = B / K$ & $B_{MSY} = K / 2$ therefore $K = 2B_{MSY}$
$S = B / 2B_{MSY}$
$2S = B/B_{MSY}$
$\mathbf{B/B_{MSY} = 2S}$



Supp. Figure 1. RAMLDB stock natural mortality rate values and sources.

Appendix A. Catch time series for the 168 RAMLDB stocks (TC=total catch, TL=total landings). Stocks are arranged and colored in order of ascending status (red=overexploited, orange=fully exploited, green=underexploited).

Appendix B. Catch time series for the 5,760 Rosenberg et al. (2014) simulated stocks (green=demersal, light blue=small pelagic, dark blue=large pelagic).