Received Date : 21-May-2016
Revised Date : 16-Nov-2016
Accepted Date : 01-Dec-2016
Article type : Original Article

Improving estimates of population status and trend with superensemble models

Alternative 1: Combining stock-assessment output with ensemble modelling

Alternative 2: Ensembles of data-limited stock assessments improve accuracy and reduce bias of B/B_{MSY} estimates

Sean C. Anderson[1*], Andrew B. Cooper[1], Olaf P. Jensen[2], Cóilín Minto[3], James T. Thorson[4], Jessica C. Walsh[1], Jamie Afflerbach[5], Mark Dickey-Collas[6,7], Kristin M. Kleisner[8], Catherine Longo[9], Giacomo Chato Osio[10], Daniel Ovando[11], Iago Mosqueira[10], Andrew A. Rosenberg[12], Elizabeth R. Selig[13]

[1]School of Resource and Environmental Management, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

[2]Institute of Marine & Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, NJ, 08901-8525, USA

[3]Marine and Freshwater Research Centre, Galway-Mayo Institute of Technology, Galway, 00000, Ireland

[4]Fisheries Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 2725 Montlake Boulevard E., Seattle, WA, 98112, USA

[5]National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, 735 State Street, Suite 300, Santa Barbara, CA, 93103, USA

[6]International Council for the Exploration of the Sea, H.C. Andersens Boulevard 44-46, DK 1553, Copenhagen, Denmark

[7]DTU Aqua National Institute of Aquatic Resources, Technical University of Denmark (DTU), Jægersborg Alle 1, 2920 Charlottenlund, Denmark

[8]Ecosystem Assessment Program, Northeast Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 166 Water St., Woods Hole, MA, 02543, USA

[9]Marine Stewarship Council, Marine House, 1 Snow Hill, London, EC1A 2DH, United Kingdom

[10]European Commission, DG Joint Research Centre, Directorate D – Sustainable Resources, Unit D.02 Water and Marine Resources, Via E. Fermi 2749, 21027 Ispra VA, Italy

[11]Bren School of Environmental Science and Management, University of California Santa Barbara, Santa Barbara, CA, 93106-5131, USA

[12]Union of Concerned Scientists, Cambridge, MA, USA

[13]Conservation International, Arlington, VA, USA

[*]Corresponding author, present address: School of Aquatic and Fishery Sciences, Box 355020, University of Washington, Seattle, WA 98195, USA E-mail: sandrsn@uw.edu

Running title: Superensembles of population status

Abstract: Fishery managers must often reconcile conflicting estimates of population status and trend. Superensemble models, commonly used in climate and weather forecasting, may provide an effective solution. This approach uses predictions from multiple models as covariates in an additional "superensemble" model fitted to known data. We evaluated the potential for ensemble averages and superensemble models ("ensemble methods") to improve estimates of population status and trend for fisheries. We fit four widely applicable data-limited models that estimate stock biomass relative to the equilibrium biomass at maximum sustainable yield (B/B_{MSY}). We combined these estimates of recent fishery status and trends in B/B_{MSY} with four ensemble methods: an ensemble average and three superensembles (a linear model, random forest, and boosted regression tree). We trained our superensembles on 5760 simulated stocks and tested them with cross-validation and against a global database of 249 stock assessments. Ensemble methods substantially improved estimates of population status and trend. Random forest and boosted regression trees performed the best at estimating population status: inaccuracy (median absolute proportional error) decreased from 0.42–0.56 to 0.32–0.32, rank-order correlation between predicted and true status improved from 0.02–0.32 to 0.44–0.48, and bias (median proportional error) declined from -0.22–0.31 to -0.12–0.03. We found similar improvements when predicting trend and when applying the simulation-trained superensembles to catch data for global fish stocks. Superensembles can optimally leverage multiple model predictions; however, they

must be tested, formed from a diverse set of accurate models, and built on a dataset representative of the populations to which they are applied.

Keywords: CMSY, data-limited fisheries, ensemble methods, multi-model averaging, population dynamics, sustainable resource management

Introduction

Methods

Individual models of population status

Simulated dataset to build the superensemble

Building the superensemble models

Additional covariates

Applying the superensemble models and testing performance

Results

Discussion

Acknowledgements

References

Supporting Material

Introduction

Status and trend are two of the most fundamental values to quantify in the management of ecological populations (e.g., Hutchings *et al.* 2010; IUCN 2015). However, managers are often faced with reconciling multiple uncertain and potentially conflicting estimates of status and trend (e.g., Brodziak and Piner 2010; Branch *et al.* 2011; Deroba *et al.* 2015). For example, one model may suggest a population is at risk and declining in abundance while others may suggest it is not at risk and stable.

One solution is to take the average or weighted average of several model predictions, i.e., an ensemble. Such ensembles are typically more accurate and less biased than individual model estimates and can incorporate various types of uncertainty, such as uncertainty in model structure, initial conditions, and parameter estimation (Dietterich 2000; Araújo and New 2007). Ensembles are superior to individual models in at least three ways: (1) statistically by averaging across models and therefore being less likely to pick the "wrong" model, (2) computationally by reducing the risk of getting stuck in a local optima, and (3) representationally by expanding the range of hypotheses explored (Dietterich 2000). This approach forms the basis of many machine learning methods (e.g., Dietterich 2000), has helped reconcile climate forecasts from dozens of models (e.g., Murphy *et al.* 2004; Tebaldi and Knutti 2007; IPCC 2013), and even improved early warning signs of malaria outbreaks (Thomson *et al.* 2006). In ecology, ensemble methods are sometimes used to improve species distribution modelling (e.g., Araújo and New 2007; Breiner *et al.* 2015) and indeed have been used to combine estimates of population status and trend (e.g., Brodziak and Piner 2010).

Whereas averages or weighted averages of model estimates may improve predictions compared to a single model, they may not optimally leverage available data. The best prediction does not necessarily lie in the middle of multiple model predictions, some models may perform better than others in certain conditions, and the covariance between models may contain information that can improve predictive accuracy. For example, one model might perform well at estimating high levels of abundance but be biased at low levels of abundance, while another model might have the opposite properties. An optimal combination of these models is not simply an average of the two.

We can exploit these characteristics by using the predictions from a group of models as inputs into a separate statistical model. This technique, sometimes called superensemble modelling (Krishnamurti *et al.* 1999), is common in climate and weather forecasting (e.g., Yun *et al.* 2005; Mote *et al.* 2015). The superensemble is fit to a training dataset where outcomes are well known and then used to predict on a dataset of interest. For example, Krishnamurti *et al.* (1999) combined predictions of wind and precipitation in Asian monsoons via a superensemble regression fit to observed data. Their superensemble was considerably more accurate than any individual prediction or an average of the predictions.

In fisheries science, the commonly used operational models for determining status and trend of exploited fish populations are stock assessments, i.e., population models coupled to an observation model that incorporate all appropriate data (e.g., catches, size and age distributions, surveys, and tagging information) to quantify values such as the biomass of a stock that can produce maximum sustainable yield (B_{MSY}) (Hilborn and Walters 1992). However, the broad range of data required to conduct these stock assessments are not available for the majority of fish populations, including those of conservation concern and of economic interest to fisheries (FAO 2014). Therefore, a number of models have been proposed to assess B/B_{MSY} based on the limited data available for the majority of fish stocks: (1) a time series of the total weight of catch and (2) a basic understanding of population productivity (e.g., Vasconcellos and Cochrane 2005; Martell and Froese 2013). Recently, Rosenberg *et al.* (2014) investigated the performance of four data-limited models through a large-scale simulation experiment. Three of these models were based on Schaefer (logistic) biomass dynamics and one was an empirical model fitted to more data-rich stock-assessment output. The four models frequently disagreed about population status (e.g., Fig. 1), no one model had strong performance across all fish stocks, and some models performed better than others depending on circumstances.

Here, we estimate population status and trend of exploited fish populations using ensembles and superensembles (collectively "ensemble methods") of these four data-limited models. We apply four ensemble-method approaches of varying complexity to both simulated and real-world fish stocks and compare their predictive performance against each other and the individual models.

Methods

To test the ability of superensembles to improve estimates of status and trend in data-limited fish stocks, we first fit four individual assessment models to a large simulated dataset of fish stocks. We then built and tested the performance of superensembles using cross-validation of the simulated dataset. Finally, we tested superensembles built with the entire simulated dataset against a database of global fish stocks. We describe these steps in detail below and illustrate the general approach both with illustrations and pseudocode in Fig. 2.

Individual models of population status

We fit four individual data-limited models that use catch data and basic life-history parameters to estimate B/B_{MSY}. We chose these models because they can be fit to the vast majority of fisheries around the world, are established in the literature, and have been extensively simulation tested (Rosenberg *et al.* 2014).

Three of the models are mechanistic and based generally on Schaefer biomass dynamics (Schaefer 1954) of the form

$$\hat{B}_{t+1} = B_t + rB_t(1 - B_t/B_0) - C_t,$$

where $\hat{B}_{t+1}$ represents predicted biomass at time t plus one year, B_t represents biomass at time t, r represents intrinsic population growth rate, B_0 represents unfished biomass or carrying capacity K, and C represents catch. The fourth model is an empirically derived model based on the RAM Legacy Stock Assessment Database (Ricard *et al.* 2012). Rosenberg *et al.* (2014) provide a full background on these four methods (`http://www.fao.org/docrep/019/i3491e/i3491e00.htm`, last accessed 2016-11-08) and code to fit all the models is available in an accompanying package datalimited for the statistical software R (R Core Team 2015) `https://github.com/datalimited/datalimited` (last accessed 2016-11-08). In summary:

- CMSY (catch-MSY) implements a stock-reduction analysis with Schaefer biomass dynamics (Martell and Froese 2013). It requires a prior distributions on r and K as well as priors on the relative proportion of biomass at the beginning and end of the time series compared to unfished biomass (depletion). The version of the model used in Rosenberg *et al.* (2014) was modified from Martell and Froese (2013) to generate biomass trends from all viable r-K pairs and produce an estimate of B/B_{MSY} from the median trend.
- COM-SIR (catch-only-model with sampling-importance-resampling) is a coupled harvest-dynamics model (Vasconcellos and Cochrane 2005). Biomass is assumed to follow a Schaefer model and harvest dynamics are assumed to follow a logistic model. The model is fit with a sampling-importance-resampling algorithm (Rosenberg *et al.* 2014).

- SSCOM (state-space catch-only model) is a hierarchical model that, similar to COM-SIR, is based on a coupled harvest-dynamics model (Thorson *et al.* 2013). SSCOM estimates unobserved dynamics in both fishing effort and the fished population based on a catch time series and priors on *r*, the maximum rate of increase of fishing effort, and the magnitude of various forms of stochasticity. The model is fit in a Bayesian state-space framework to integrate across three forms of stochasticity: variation in effort, population dynamics, and fishing efficiency (Thorson *et al.* 2013).
- mPRM (modified panel regression model) is a modified version of the panel-regression model from Costello *et al.* (2012). Unlike the other models, mPRM is empirical and not mechanistic — it uses the RAM Legacy Stock Assessment database to fit a regression model to a series of characteristics of the catch time series and stock with stock-assessed B/B_{MSY} as the response. The model used in this paper is modified from the original — it condenses the life-history categories into three categories to match the simulated dataset, removes the maximum catch predictor since the absolute catch in the simulated dataset is arbitrary, and does not implement the bias correction needed in Costello *et al.* (2012) since we do not derive estimates of median status across multiple stocks.

Simulated dataset to build the superensemble

We first developed and tested ensemble methods on a fully factorial simulated dataset of fisheries with known status (Rosenberg *et al.* 2014). Briefly, these simulations were implemented with the FLR packages (Kell *et al.* 2007) for the statistical software R, and, in particular, the FLBRP package. The framework takes a series of life-history parameters and fishery characteristics to generate population projections and resulting catch timeseries. Life-history values (e.g. mean asymptotic length) for three fish life histories (small pelagic, demersal, and large pelagic) were translated into a complete set of parameters for a von Bertalanffy growth model, a maturity ogive, natural mortality, a selectivity function, and a Beverton-Holt stock-recruitment function using the life-history relationships derived in Gislason *et al.* (2008).

Fishing scenarios included three levels of initial biomass depletion compared to carrying capacity: biomass at 100%, 70%, and 40% of carrying capacity; and four exploitation patterns: (1) a constant exploitation rate, (2) an exploitation rate coupled with biomass to mimic an open-access single-species fishery, (3) a scenario where exploitation rate increased continuously, and

(4) a "roller-coaster" scenario where the exploitation rate increased and then decreased. Process noise (recruitment variability; i.e., unexplained variability in population dynamics) was introduced to the models at two magnitudes in log space, $N(0,0.2^2)$ and $N(0,0.6^2)$, and was either uncorrelated through time or had first-order autoregressive correlation of 0.6. The simulation also included a scenario with $N(0,0.2^2)$ measurement error around log catch and one scenario without measurement error. Rosenberg *et al.* (2014) ran ten iterations for each combination of factors adding stochastic draws of recruitment and catch-recording variability each time to generate a total of 5760 stocks. Code to generate the simulations is available at `https://github.com/datalimited/stocksims` (last accessed 2016-11-08).

Building the superensemble models

The individual models we seek to combine with superensembles provide time series of stock status (B/B_{MSY}). Therefore, we can use superensembles to estimate any property of these time series. Here, we focus on two properties: the mean and slope of B/B_{MSY} in the last five years. Together, these quantities address the recent state and trend of stock status, which are both of management and conservation interest (e.g., Hutchings *et al.* 2010; IUCN 2015). To avoid undue influence of the time series end points on the calculated slope, we measured the slope as the Theil-Sen estimator of median slope (Theil 1950).

We used the mean or slope of B/B_{MSY} as the response variable and the predictions from the individual models as predictors in our superensemble models (Fig. 2a). When modelling mean B/B_{MSY} — a ratio bounded at zero — we fit the superensemble models in log space and exponentiated the predictions. For the estimates of B/B_{MSY} slope, which are not bounded at zero, we fit superensemble models on the natural untransformed scale.

We compared an ensemble average and three superensembles of varying complexity: a linear model with two-way interactions, a random forest, and a boosted regression tree. We describe these models as estimating $\hat{\theta}$, which represents either the ensemble estimated log B/B_{MSY} or slope of B/B_{MSY}. The individual model estimates of log B/B_{MSY} or slope of B/B_{MSY} are represented as $\hat{b}$ for models *i* 1 through 4 (CMSY, COM-SIR, SSCOM, mPRM). The ensemble average for each fishery *i* was calculated as:

$$\hat{\theta}_1 = \left(\hat{b}_{i,1} + \hat{b}_{i,2} + \hat{b}_{i,3} + \hat{b}_{i,4}\right)/4, \quad \text{for } i = 1, \ldots, n.$$

We fit the linear model superensemble with all second-order interactions:

$$\hat{\theta}_i = \beta_0 + \beta_1 \hat{b}_{i,1} + \ldots + \beta_{1,2} \hat{b}_{i,1} \hat{b}_{i,2} + \ldots + \epsilon_i, \quad \epsilon \sim \text{Normal}(0, \sigma^2), \quad \text{for } i = 1, \ldots, n.$$

For this illustrative example we chose this level of model complexity *a priori* but a modeller could apply model selection via information-theoretic or cross-validation approaches.

Our two machine learning superensemble models, a random forest and a generalized boosted model (GBM), were based on regression trees. Regression trees sequentially determine what value of a predictor best splits the response data into two branches based on a loss function (Breiman *et al.* 1984). In random forests, a series of regression trees are built on a random subset of the data and random subset of the covariates of the model (Breiman 2001). In GBMs, each subsequent model is fit to the residuals from the previous model; data points that are fit poorly in a given model are given more weight in the next model (Elith *et al.* 2008). Random forests and GBMs can provide strong predictive performance and fit highly non-linear relationships (Elith *et al.* 2008; Hastie *et al.* 2009). We fit random forest models with the randomForest package (Liaw and Wiener 2002) for R with the default argument values. We fit boosted regression tree models with the gbm package (Ridgeway 2015) for R. We fit GBMs with 2000 trees, an interaction depth of 6, a learning rate (shrinkage parameter) of 0.01, and all other arguments at their default values.

Additional covariates

Superensemble models allow us to incorporate additional covariates and potentially leverage interactions between these covariates and individual model predictions. Additional covariates could be, for example, life-history characteristics, information on exploitation patterns, or statistical properties of the data. We tested the performance benefits of including one set of additional covariates: spectral properties of the catch time series. Spectral analysis decomposes a time series into the frequency domain and provides a means of describing the cyclical shape of the catch series that is independent of time series length (except in affecting precision) and independent of absolute magnitude of catch. We fit spectral models to the scaled catch time series (catch divided by maximum catch) with the `spec.ar` function in R and recorded representative short- and long-term spectral densities at frequencies of 0.20 and 0.05, which correspond to 5- and 20-year cycles. For the linear model superensemble, we incorporated the two spectral covariates (*S*1, *S*2) along with all second-order interactions as:

$$\hat{\theta}_i = \beta_0 + \beta_1 \hat{b}_{i,1} + \ldots + \beta_{1,2} \hat{b}_{i,1} \hat{b}_{i,2} + \ldots + \beta_{S1} S1_i + \beta_{S2} S2_i + \beta_{S1,S2} S1_i S2_i + \beta_{1,S2} \hat{b}_{i,1} S2_i + \ldots + \epsilon_i,$$

with $\varepsilon \sim N(0, \sigma^2)$ and for simulated fisheries *i* 1 through *n*. We include the results of adding these additional covariates in the supplementary materials.

Applying the superensemble models and testing performance

Once the superensemble models are built and trained using the simulated stocks (or any dataset with “known” status), we can use the superensembles to estimate the status of new stocks (Fig. 2b). To do this we applied the individual models to our stocks of interest (i.e., CMSY, COM-SIR, SSCOM, mPRM) and then used these individual model estimates of status or trend as data in our already built superensemble models. In this paper we applied the superensemble models to subsets of the simulated data as a cross-validation test to test predictive performance and to the RAM Legacy Stock Assessment database to test predictive performance on real stocks.

We used repeated three-fold cross validation: we randomly divided the dataset into three sets, built superensemble models on two-thirds of the data, and evaluated predictive performance on the remaining third. We repeated this across each of the three splits and then repeated the whole procedure 50 times to account for bias that may result from any one set of validation splits. In the simulated dataset, there were 10 replicates of each unique combination of simulation parameters that differed only in stochastic variability. Since the dynamics of these populations were often similar, we grouped these stocks in the cross-validation process into either the training or testing split.

We also tested our ensemble methods on the RAM Legacy Stock Assessment Database — a compilation of stock-assessment output from hundreds of exploited marine populations around the world. Our analysis of the stock-assessment database was based on version 2.5. After removing stocks for which at least one of the individual models did not converge (121), this database included 249 stocks. We removed these stocks for all methods — both for the individual and superensemble models. An alternative would be to fit separate superensemble models to subsets of the individual models that did converge, but for simplicity we only used superensemble models fitted to all four individual models.

In the case of the RAM Legacy Stock Assessment Database, we used superensembles trained on the entire simulation dataset. However, since mPRM is built on the same stock-assessment database, we applied three-fold cross-validation to the data underlying the mPRM model so that the dataset with which mPRM was trained (for the individual model and superensemble) was separate from the dataset with which it was tested. This meant that, for each iteration of cross validation, we split the RAM database into three, fit the mPRM model to two-thirds of the RAM database, fit a superensemble with this version of mPRM, and then tested the performance of the superensemble on the third of the RAM database we had withheld.

Predictive performance can be evaluated with metrics that represent a variety of modelling goals. For continuous response variables such as the mean and slope of population status, performance metrics often measure some form of bias, precision, accuracy (a combination of bias and precision), or the ability to correctly rank or correlate across populations (e.g., Walther and Moore 2005). Here, we measure proportional error, defined as $(\hat{\theta} - \theta)/|\theta|$, where $\hat{\theta}$ and θ represent estimated and "true" (or stock-assessed) mean or slope of B/B_{MSY}. We calculated median proportional error to measure bias, median absolute proportional error to measure accuracy, and Spearman's rank-order correlation between predicted and "true" values to measure the ability to correctly rank populations. When testing with the RAM Legacy Stock Assessment database, we treated the estimates from these data-rich stock assessments as known without error. Thus, any error in the stock-assessment estimates of the mean or slope of B/B_{MSY} also contributes to our estimates of prediction error for each of the four data-limited models and the ensembles. Code to reproduce our analysis is available at `https://github.com/datalimited/ensembles` (last accessed 2016-11-08).

Results

Applied to the simulated dataset of known stock status, the individual models had variable success at estimating the mean (status) and slope (trend) of B/B_{MSY} in the last five years. All models exhibited a high degree of scatter around the one-to-one line of perfect status prediction (Fig. 3). In contrast to the known unimodal distribution of status, CMSY exhibited bimodal predictions (Fig. 3a), but had the best rank-order correlation and accuracy scores (Fig. 4a). COM-SIR and SSCOM both correctly identified a number of stocks with low status, but frequently predicted a high status when status was in fact low (Fig. 3b, c). mPRM had relatively poor ability to predict status for the simulated dataset (Fig. 3d). There was generally little correlation between true and predicted recent trend in status for any of the individual models (rank-order correlation = 0.02–0.25) with the exception of SSCOM (correlation = 0.54; Figs S1a–d).

Ensemble methods, and in particular the machine learning superensemble models (random forest and GBM), generally improved estimates of status and trend over any individual model (Fig. 3e–h, Fig. S1e–h). Compared to the individual models, machine learning superensembles decreased inaccuracy (median absolute proportional error) from 0.42–0.56 to 0.32–0.32, increased rank-order correlation from 0.02–0.32 to 0.44–0.48, and reduced bias (median proportional error) from -0.22–0.31 to -0.12–0.03 (Fig. 4a). These superensembles also generally had better ability to distinguish if simulated stocks were above or below B/B_{MSY}=0.5 (Fig. S2). Results were similar when predicting trend: compared to individual models, machine learning superensembles decreased inaccuracy from 0.04–0.06 to 0.03–0.03, increased rank-order correlation from 0.02–0.54 to 0.61–0.65, and reduced bias from -0.009–0.014 to -0.002—0.002 (Fig. S3). The ensemble models that simply took a mean of the individual models ranked slightly behind the best individual model for estimating fish stock status (CMSY; Fig. 4a) and had slightly lower correlation but higher accuracy than the best individual model at predicting the trends of status (SSCOM; Fig. S3).

The superensemble models were able to improve the predictive performance by harnessing the best properties of individual models, the covariance between individual models, and interactions with other covariates. For example, SSCOM had strong predictive ability when it predicted low B/B_{MSY} (Fig. 3c, Fig. S4c) and CMSY predictions were approximately linearly related to B/B_{MSY} within the low and high clusters of predictions (Fig. S4). SSCOM contributed most strongly on its own to determining trend (Fig. S5). Superensembles also exploited the covariance between individual model predictions. For instance, both the linear model and GBM ensemble suggest that if mPRM and SSCOM predict high status, the true status also tends to be high (Figs S6, S7f). The addition of spectral density covariates helped the superensemble models correctly predict higher status values (Fig. S8g, h). The performance of the ensembles was only marginally improved by including these covariates (Fig. S9 vs. Fig. 4).

When applied to the stock-assessment database, the superensemble models — trained exclusively on the simulated dataset — generally performed as well or better than the best individual models. The mean, random forest, and GBM ensembles outperformed the mPRM method which is trained directly on the RAM Legacy Stock Assessment database itself (Fig. 4b, Fig. S10). Compared to the individual models, the machine learning superensembles increased accuracy by 0–30%, improved correlation from 0.19–0.36 to 0.35–0.38, and reduced bias from -0.25–0.45 to -0.05–0.02.

Discussion

Ensemble methods provide a useful approach to situations where environmental resource management decisions must be made on the basis of multiple, potentially contrasting estimates of status. Compared to individual models of fish population status, ensemble methods were consistently the best or among the best across three performance dimensions (accuracy, bias, and rank-order correlation), two response variables (status and trend), two datasets (simulated and global fisheries), and multiple ensemble methods (from a simple average to machine learning superensembles). Our results suggest choosing a superensemble model that allows for non-linear relationships, such as machine learning methods; these models provided added insight into individual model behaviour and generally performed the best.

Certain conditions will make some ensemble models more effective than others. First, ensembles will be most effective when they are comprised of diverse individual models that choose different structural model forms, explore contrasting but plausible ranges of parameter values, and make uncorrelated errors (Ali and Pazzani 1996; Dietterich 2000; Tebaldi and Knutti 2007). We would expect such models to perform well in different conditions and an ensemble model can exploit the best predictive performance of each. Second, ensemble models will be most effective when they are not overfit to the training dataset. Cross-validation testing (Caruana *et al.* 2004; Hastie *et al.* 2009) and methods that are robust to overfitting such as random forests (Breiman 2001), may help avoid overfitting ensemble models. We note that our simplest ensemble model, an average of individual model predictions, performed approximately as well as complex machine learning models when we trained our superensembles on the simulation dataset and tested them on a separate "real" dataset (i.e., the RAM Legacy Stock Assessment database, Fig. 4b). Third, ensemble models will be most effective when they are trained on data that are representative of the dataset of interest (Knutti *et al.* 2009; Weigel *et al.* 2010). Cross-validation within a training dataset will provide an optimistically biased impression of predictive performance if the training dataset fundamentally differs from the dataset of interest (Hastie *et al.* 2009).

We illustrated that superensembles can improve point estimates of population status and trends in status; however, there is no reason why superensembles cannot also be used to provide measures of uncertainty around those point estimates. The same approaches to deriving measures of uncertainty from any regression model are available to a superensemble. For example, likelihood profile confidence intervals or Bayesian credible intervals are available for superensembles fit via maximum likelihood or Bayesian procedures, respectively. Measures of predictive uncertainty can be generated for machine learning methods such as random forests or GBMs using bootstrap procedures (e.g. Hastie *et al.* 2009; Finnegan *et al.* 2015). Furthermore, uncertainty from the component models could be included in superensembles. These superensembles could be fit using any errors-in-variables or measurement-error modelling approach (e.g. Carroll *et al.* 2006).

Multi-model inference in the form of coefficient averaging weighted by information theoretics such as the Akaike Information Criterion (AIC) is a common analytical approach in fisheries and ecology (e.g., Burnham and Anderson 2002; Johnson and Omland 2004; Grueber *et al.* 2011). The ensemble methods described in this paper share similarities with coefficient averaging but differ in other important ways. Ensemble methods and coefficient averaging share the long-held notion that multiple working hypotheses can contribute useful information for inference (Chamberlin 1890). A fundamental difference is that coefficient averaging focuses on averaging *coefficients* whereas ensembles instead average *predictions*. Thus, ensembles provide a general purpose tool: they do not require information theoretics and they can combine different types of models (e.g., parametric and non-parametric models or frequentist and Bayesian predictions). Furthermore, superensembles extend these benefits by allowing model predictions to be combined via non-linear functions that are tuned to known data.

A strength of superensembles is that they can be tailored to predict specific response variables. For example, we built separate superensemble models of mean B/B_{MSY} and the slope of B/B_{MSY}. The same set of model weights or non-linear relationships need not hold across different response variables. For instance, SSCOM contributed little to the GBM superensemble estimate of status at higher levels of predicted B/B_{MSY} (Fig. S4), but contributed strongly to estimates of trend (Fig. S5). Formally, fitting superensemble models to specific quantities of interest (such as the slope of B/B_{MSY}) provides an additional calibration step to a quantity of interest (Rykiel 1996). This ensemble calibration could include a loss function tailored to the goals of the model, say placing greater weight on accuracy at lower rather than higher status levels. Conversely, because superensembles are tailored to a specific response and loss function, superensembles force a modeller to choose an operational purpose for their model upfront (*sensu* Dickey-Collas *et al.* 2014). For instance, one could have an ensemble estimate of B and an ensemble estimate of B_0, but their ratio may not be the same as an ensemble estimate of B/B_0. A modeller might therefore choose to focus on B/B_0, which provides a unitless ratio, is easier to compare across stocks, and the ratio is often a more stable estimate across models (Deroba *et al.* 2015).

As Box and Draper (1987) noted, all models are wrong, but some may still be useful. The ensemble methods we investigated attempt to piece together the useful parts of candidate models to build a model with improved performance. Instead of viewing the superensemble as a black box, we think considerable mechanistic understanding can be gained by studying its structure. For example, when SSCOM estimates low status this is likely the case, conversely when COMSIR estimates low status, the true status is more likely to be high (Fig. S4). These models have two main differences: (1) the form of effort dynamics and (2) the allowance for both measurement and process error in SSCOM, whereas the implemented COMSIR admits measurement error only. Were the methods to differ only in effort dynamics, the results point towards a more suitable representation of effort dynamics at low biomasses in SSCOM. We think that such investigation of the structure of a superensemble may lead to improvement in the mechanisms assumed in individual models.

Combining predictions from multiple models via superensemble methods is broadly useful in other subfields of fisheries science and ecology in general. In fisheries science, superensembles provide an additional tool to assist with some longstanding issues. For example, superensembles are helpful since modelers need not decide on one model — instead of deciding on dome versus asymptotic fisheries selectivity (e.g., Sampson and Scott 2012), or on whether to fix or estimate natural mortality (e.g., Johnson *et al.* 2015), superensembles can use multiple models to draw inference. Furthermore, the relative contributions of individual models can help tease apart the conditions under which various model assumptions result in the most accurate predictions. Finally, superensembles can be used to directly estimate other quantities of interest in fisheries science. For instance, superensembles could help assess overfishing by estimating fishing mortality compared to fishing mortality at MSY (F/F_{MSY}) or be trained to estimate natural mortality.

More broadly, in ecology, predictions about extinction risk are widely used at national (e.g., the US Endangered Species Act and the Canadian Species at Risk Act) and international (e.g., the IUCN Red List, IUCN 2015) levels. These risk assessments generally involve fitting regression models to outcomes for individual species along with predictors of extinction risk (e.g., Anderson *et al.* 2011; Pinsky *et al.* 2011), or fitting population-dynamic models to data for individual species (e.g., DFO 2010). Both types of models are prone to error caused by model-misspecification and therefore results are sensitive to decisions about model structure (Brooks and Deroba 2015). Although there are options to account for potential model-misspecification in determination of species risk (e.g., coeffcient averaging, Burnham and Anderson 2002; generalized modeling, Yeakel *et al.* 2011; or semi-parametric methods, Thorson *et al.* 2014), ensemble methods are a relatively simple way to combine predictions in a transparent manner. Beyond estimates of status and trend, ensemble methods could be used, for example, to increase the robustness of spatial predictions when designing networks of protected areas (Rassweiler *et al.* 2014) or to forecast potential spatial shifts in species distribution given climate impacts (Harsch *et al.* 2014). In any case, superensembles are not a panacea and are ultimately limited by the quality, breadth, and representativeness of simulated or trusted data to which they are calibrated.

Acknowledgements

We thank members of Phase I of the working group “Developing new approaches to global stock status assessment and fishery production potential of the seas” who contributed to developing the data-limited methods and simulations used in our analysis. We thank E. Jardim, F. Scott, and J.A. Hutchings for helpful comments during the development of this project, and R.D. Methot for comments on an earlier version of the manuscript. We also thank two anonymous reviewers for their thoughtful reviews, comments, and criticisms. We thank the Gordon and Betty Moore Foundation for funding the working group "Applying data-limited stock status models and developing management guidance for unassessed fish stocks".

References

Ali, K.M. and Pazzani, M.J. (1996) Error reduction through learning multiple descriptions. *Mach. Learn.* **24**, 173–202.

Anderson, S.C., Farmer, R.G., Ferretti, F., Houde, A.L.S. and Hutchings, J.A. (2011) Correlates of vertebrate extinction risk in Canada. *BioScience* **61**, 538–549.

Araújo, M.B. and New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* **22**, 42–47.

Box, G.E. and Draper, N.R. (1987) *Empirical model-building and response surfaces*. Wiley, New York, NY, USA.

Branch, T.A., Jensen, O.P., Ricard, D., Ye, Y. and Hilborn, R. (2011) Contrasting global trends in marine fishery status obtained from catches and from stock assessments. *Conservation Biology* **25**, 777–786.

Breiman, L. (2001) Random forests. *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, USA.

Breiner, F.T., Guisan, A., Bergamini, A. and Nobis, M.P. (2015) Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution* **6**, 1210–1218.

Brodziak, J. and Piner, K. (2010) Model averaging and probable status of North Pacific striped marlin, *Tetrapturus audax*. *Canadian Journal of Fisheries and Aquatic Sciences* **67**, 793–805.

Brooks, E.N. and Deroba, J.J. (2015) When "data" are not data: the pitfalls of post hoc analyses that use stock assessment model output. *Canadian Journal of Fisheries and Aquatic Sciences* **72**, 634–641.

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition*. Springer, New York, NY, USA.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, FL.

Caruana, R., Niculescu-Mizil, A., Crew, G. and Ksikes, A. (2004) Ensemble selection from libraries of models. In: *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*. New York, NY, USA.

Chamberlin, T.C. (1890) The method of multiple working hypotheses. *Science* **15**, 92–96.

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

Deroba, J.J., Butterworth, D.S., Methot, R.D., Oliveira, J.A.A.D. *et al.* (2015) Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. *ICES Journal of Marine Science* **72**, 19–30.

DFO (2010) Stock assessment update for British Columbia canary rockfish. Technical report, Canadian Science Advisory Secretariat, Ottawa, Canada.

Dickey-Collas, M., Payne, M.R., Trenkel, V.M. and Nash, R.D.M. (2014) Hazard warning: model misuse ahead. *ICES Journal of Marine Science* **71**, 2300–2306.

Dietterich, T.G. (2000) Ensemble methods in machine learning. In: *Multiple classifier systems*. Springer, Berlin Heidelberg, pp. 1–15.

Elith, J., Leathwick, J.R. and Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–13.

FAO (2014) The state of the world fisheries and aquaculture. Technical report, Food and Agriculture Organization of the United Nations, Rome.

Finnegan, S., Anderson, S.C., Harnik, P.G., Simpson, C. *et al.* (2015) Paleontological baselines for evaluating extinction risk in the modern oceans. *Science* **348**, 567–570.

Gislason, H., Pope, J.G., Rice, J.C. and Daan, N. (2008) Coexistence in North Sea fish communities: implications for growth and natural mortality. *ICES Journal of Marine Science* **65**, 514–530.

Grueber, C.E., Nakagawa, S., Laws, R.J. and Jamieson, I.G. (2011) Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology* **24**, 699–711.

Harsch, M.A., Zhou, Y., HilleRisLambers, J. and Kot, M. (2014) Keeping pace with climate change: stage-structured moving-habitat models. *American Naturalist* **184**, 25–37.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. Springer.

Hilborn, R.W. and Walters, C. (1992) *Quantitative Fisheries Stock Assessment: Choice, Dynamics, and Uncertainty*. Chapman and Hall, London.

Hutchings, J.A., Minto, C., Ricard, D., Baum, J.K. and Jensen, O.P. (2010) Trends in the abundance of marine fishes. *Canadian Journal of Fisheries and Aquatic Sciences* **67**, 1205–1210.

IPCC (2013) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IUCN (2015) The IUCN Red List of Threatened Species. Version 2015.1. Technical report.

Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution* **19**, 101–108.

Johnson, K.F., Monnahan, C.C., McGilliard, C.R., Vert-pre, K.A. *et al.* (2015) Time-varying natural mortality in fisheries stock assessment models: identifying a default approach. *ICES Journal of Marine Science* **72**, 137–150.

Kell, L.T., Mosqueira, I., Grosjean, P., Fromentin, J.M. *et al.* (2007) FLR: an open-source framework for the evaluation and development of management strategies. *ICES Journal of Marine Science* **64**, 640–646.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G.A. (2009) Challenges in combining projections from multiple climate models. *Journal of Climate* **23**, 2739–2758.

Krishnamurti, T.N., Kishtawal, C.M., LaRow, T.E., Bachiochi, D.R. *et al.* (1999) Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**, 1548–1550.

Liaw, A. and Wiener, M. (2002) Classification and regression by random forest. *R News* **2**, 18–22.

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

Mote, P.W., Allen, M.R., Jones, R.G., Li, S. *et al.* (2015) Superensemble regional climate modeling for the western US. *Bulletin of the American Meteorological Society* .

Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M. and Stainforth, D.A. (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772.

Pinsky, M.L., Jensen, O.P., Ricard, D. and Palumbi, S.R. (2011) Unexpected patterns of fisheries collapse in the world's oceans. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 8317–8322.

R Core Team (2015) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rassweiler, A., Costello, C., Hilborn, R. and Siegel, D.A. (2014) Integrating scientific guidance into marine spatial planning. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **281**, 20132252.

Ricard, D., Minto, C., Jensen, O.P. and 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**, 380–398.

Ridgeway, G. (2015) *gbm: generalized boosted regression models*. R package version 2.1.1.

Rosenberg, A.A., Fogarty, M.J., Cooper, A.B., Dickey-Collas, M. *et al.* (2014) Developing new approaches to global stock status assessment and fishery production potential of the seas. FAO Fisheries and Aquaculture Circular 1086, Rome, Italy.

Rykiel, Jr., J. (1996) Testing ecological models: the meaning of validation. *Ecological Modelling* **90**, 229–244.

Sampson, D.B. and Scott, R.D. (2012) An exploration of the shapes and stability of population–selection curves. *Fish and Fisheries* **13**, 89–104.

Schaefer, M.B. (1954) Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Inter-American Tropical Tuna Commission Bulletin* **1**, 23–56.

Tebaldi, C. and Knutti, R. (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London. Series A: Physical, Mathematical and Engineering* **365**, 2053–2075.

Theil, H. (1950) A rank-invariant method of linear and polynomial regression analysis, I, II, and III. *Nederlandsche Akad. van Wetenschappen Proc.* **53**, 386–92, 521–525, and 1397–1412.

Thomson, M.C., Doblas-Reyes, F.J., Mason, S.J., Hagedorn, R. *et al.* (2006) Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature* **439**, 576–579.

Thorson, J.T., Minto, C., Minte-Vera, C.V., Kleisner, K.M. and 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**, 1829–1844.

Thorson, J.T., Ono, K. and Munch, S.B. (2014) A Bayesian approach to identifying and compensating for model misspecification in population models. *Ecology* **95**, 329–341.

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

Walther, B.A. and Moore, J.L. (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **28**, 815–829.

Weigel, A.P., Knutti, R., Liniger, M.A. and Appenzeller, C. (2010) Risks of model weighting in multimodel climate projections. *Journal of Climate* **23**, 4175–4191.

Yeakel, J.D., Stiefs, D., Novak, M. and Gross, T. (2011) Generalized modeling of ecological population dynamics. *Journal of Theoretical Biology* **4**, 179–194.

Yun, W.T., Stefanova, L., Mitra, A.K., Kumar, T.S.V.V., Dewar, W. and Krishnamurti, T.N. (2005) A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. *Tellus A* **57**, 280–289.

Figure captions

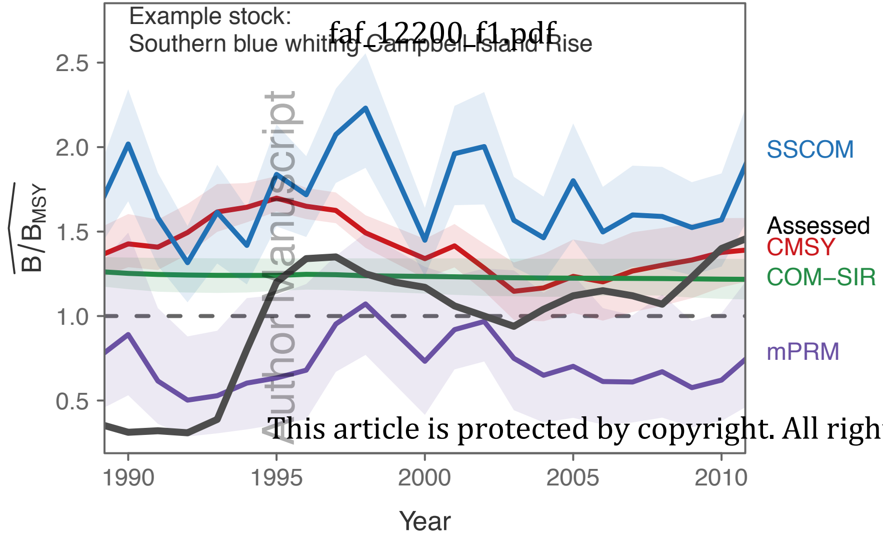
Figure 1: Different models can suggest conflicting population statuses and trends. Shown are trajectories of estimated B/B_{MSY} from four data-limited assessment methods (colours) and a data-rich stock assessment (black) for Southern blue whiting (*Micromesistius australis*) on Campbell Island Rise, New Zealand. Lines indicate median fits and shaded regions indicate interquartile ranges. Dashed horizontal line indicates $B/B_{MSY}=1$.

Figure 2: Using a superensemble model to predict population status from two individual models. The process is illustrated graphically on the left and with R pseudocode on the right. (a) Individual models (red and blue lines) are fit to training data (dots) from populations of known or assumed status (known status shown by black line). The shaded gray boxes indicate the recent time period that we are interested in for this paper. Estimates of status from these individual models ($\hat{b}_{i,1}$ and $\hat{b}_{i,2}$), potentially combined with additional covariates, are then used as covariates in a statistical model fitted to the known or assumed population status as the response (here represented as a linear model). The symbols β and ε represent parameters and error in the linear model, respectively. The i subscripts represent individual fish stocks from 1 to n, and θ represents the known status. (b) The superensemble can then be used to make predictions for new stocks of interest. The same individual models are fit to populations of interest and then combined using the previously fitted superensemble model. Here, the j subscripts represent

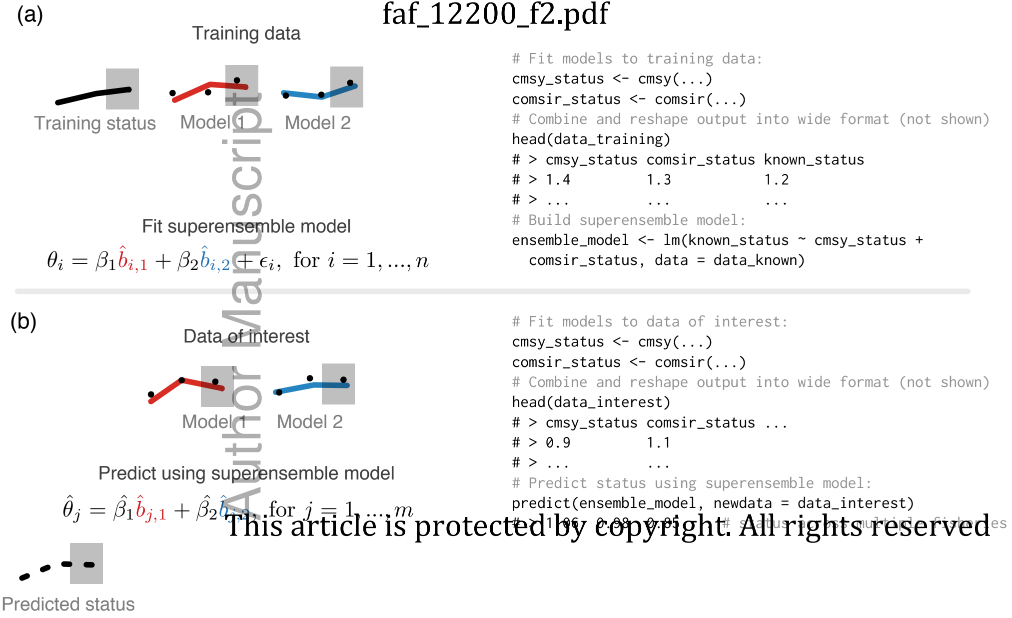
individual fish stocks from 1 to *m*, and $\hat{\theta}$ represents the predicted status. The $\hat{\beta}$ represent the parameters estimated when the superensemble was fit in panel a.

Figure 3: True (or assessed) population status (x axis) vs. predicted population status from individual models and ensemble methods with cross-validation (y axis). These scatterplots represent the aggregate results of repeated three-fold cross-validation tests where the ensemble models are built on two-thirds of the data and tested on the remaining third. (a–d) Individual data-limited model estimates of mean \ (biomass divided by biomass at maximum sustainable yield) in the last five years for a simulated dataset of known population status. (e–h) Ensemble estimates for the same populations. Shown are a mean ensemble, a linear superensemble model with two-way interactions (LM), a random forest superensemble (RF), and a generalised boosted regression model superensemble (GBM). (i–l) The same ensemble models, which were trained on the simulated dataset, applied to the RAM Legacy stock assessment database and compared to data-rich stock assessed status. In the case of the RAM Legacy stock assessment data, we refit the modified panel regression model (mPRM) on each cross-validation split. We binned the data into hexagons for visual presentation. Darker areas indicate areas with greater density of data. Yellow-red shading and yellow-blue shading distinguishes individual models from ensemble methods.

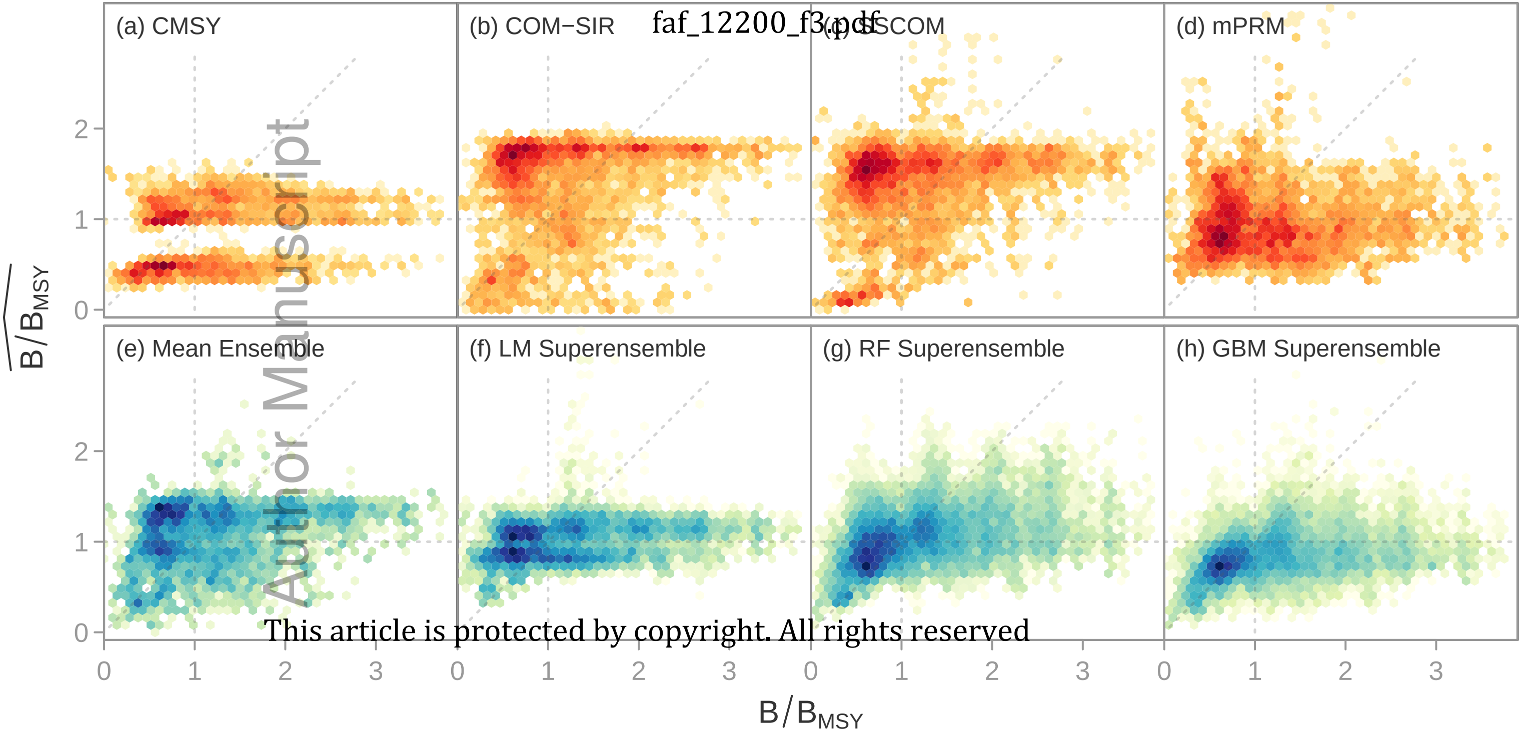
Figure 4: Performance metrics of individual and ensemble models predicting B/B_{MSY} (mean biomass divided by biomass at maximum sustainable yield) in the last five years fitted to a dataset with (a) known population status and (b) the RAM Legacy stock assessment database. The x-axis represents within-population inaccuracy: median absolute proportional error (MAPE). The y-axis represents across-population Spearman rank-order correlation. The top-left corner contains methods with the best performance across the two metrics. The colour shading represents bias (median proportional error; MPE): white points are unbiased, blue points represent methods that predict B/B_{MSY} values that are too high, red points represent methods that predict B/B_{MSY} values that are too low. These performance metrics are derived from the data in Fig. 3 and based on repeated three-fold cross-validation testing.


Example stock:
Southern blue whiting Campbell Island Rise
2.5
2.0
1.5
1.0
0.5
B/B_MSY
1990
1995
2000
2005
2010
Year
SSCOM
Assessed
CMSY
COM−SIR
mPRM

faf_12200_f2.pdf



faf_12200_f3.pdf



faf_12200_f4.pdf