




## Food for Thought

# Operationalizing ensemble models for scientific advice to fisheries management

Ernesto Jardim <sup>1,2\*</sup>, Manuela Azevedo <sup>3</sup>, Jon Brodziak<sup>4</sup>, Elizabeth N. Brooks<sup>5</sup>, Kelli F. Johnson <sup>6</sup>, Nikolai Klibansky<sup>7</sup>, Colin P. Millar<sup>8</sup>, C oil n Minto<sup>9</sup>, Iago Mosqueira<sup>1,†</sup>, Richard D.M. Nash<sup>10,‡</sup>, Paraskevas Vasilakopoulos<sup>1</sup>, and Brian K. Wells<sup>11</sup>

<sup>1</sup>European Commission, Joint Research Centre (JRC), Directorate D, Sustainable Resources, Via E. Fermi, 2749, 21027 Ispra VA, Italy

<sup>2</sup>Marine Stewardship Council, Snow Hill 1, Marine House, London EC1A 2DH, UK

<sup>3</sup>Portuguese Institute for the Sea and Atmosphere (IPMA), Av. Doutor Alfredo Magalh es Ramalho, 6, Alg es 1495-165, Portugal

<sup>4</sup>Pacific Islands Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Honolulu, HI, USA

<sup>5</sup>Northeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Woods Hole, MA, USA

<sup>6</sup>Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, USA

<sup>7</sup>Southeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Beaufort, NC, USA

<sup>8</sup>International Council for the Exploration of the Sea (ICES), H. C. Andersens Boulevard 44-46, Copenhagen V 1553, Denmark

<sup>9</sup>Marine and Freshwater Research Centre (MFRC), Galway-Mayo Institute of Technology (GMIT), Dublin Road, Galway, Ireland

<sup>10</sup>Institute of Marine Research, P.O. Box 1870, Nordnes, Bergen 5817, Norway

<sup>11</sup>Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Santa Cruz, CA, USA

\*Corresponding author: tel: +393668918077; e-mail: [ernesto.jardim@msc.org](mailto:ernesto.jardim@msc.org).

†Present address: Wageningen Marine Research, Haringkade 1, 1976 CP, IJmuiden, the Netherlands.

‡Present address: Centre for Environment, Fisheries and Aquaculture Science (Cefas), Pakefield Road, Lowestoft, Suffolk NR33 0HT, UK.

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E. N., Johnson, K. F., Klibansky, N., Millar, C. P., Minto, C., Mosqueira, I., Nash, R. D.M., Vasilakopoulos, P., and Wells, B. K. Operationalizing ensemble models for scientific advice to fisheries management. – ICES Journal of Marine Science, 78: 1209–1216.

Received 13 August 2020; revised 7 January 2021; accepted 12 January 2021; advance access publication 1 March 2021.

This paper explores the possibility of using the ensemble modelling paradigm to fully capture assessment uncertainty and improve the robustness of advice provision. We identify and discuss advantages and challenges of ensemble modelling approaches in the context of scientific advice. There are uncertainties associated with every phase in the stock assessment process: data collection, assessment model choice, model assumptions, interpretation of risk, up to the implementation of management advice. Additionally, the dynamics of fish populations are complex, and our incomplete understanding of those dynamics and limited observations of important mechanisms, necessitate that models are simpler than nature. The aim is for the model to capture enough of the dynamics to accurately estimate trends and abundance, and provide the basis for robust advice about sustainable harvests. The status quo approach to assessment modelling has been to identify the “best” model and generate advice from that model, mostly ignoring advice from other model configurations regardless of how closely they performed relative to the chosen model. We discuss and make suggestions about the utility of ensemble models, including revisions to the formal process of providing advice to management bodies, and recommend further research to evaluate potential gains in modelling and advice performance.

**Keywords:** assessment, conservation, ensemble, exploitation, fisheries, management, , multi-model, natural resources, structural uncertainty

  International Council for the Exploration of the Sea 2021.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

## Introduction

Providing scientific advice to fisheries managers is a risky activity! It is not uncommon that a model that has been performing well suddenly fails to properly fit an additional year of data, or projections made in the past did not materialize when more recent observations became available. Fisheries scientists have to deal with a complex system, with many unknown or poorly understood processes and limited information. The emergence or increased importance of previously unmodelled processes, changes in processes that are assumed constant, conflicting information and data revisions, all have the insidious tendency to ruin what had been a perfectly acceptable assessment fit, invalidating advice and weakening confidence in future advice efforts.

The North Sea cod stock is a good example of assessment instability due to new data and changed model configurations. The 2015 benchmark meeting, after a thorough exploration of various model configurations and two different models, agreed on a single model (ICES, 2015a). The model fit showed moderate differences with the previous assessment and slight changes in PA and MSY reference points, but significant changes in limit reference points. The assessment subsequently carried out in 2015 (ICES, 2015b), with the new model configuration and updated data, doubled previous biomass estimates. For example, SSB estimates for 2014 were revised from 68.5 tonnes in the 2014 advice (ICES, 2014), to 124.7 tonnes in the 2015 advice (ICES, 2015b). Two years later, the 2017 assessment (ICES, 2017) revised the SSB estimates again, reducing recent values by about 20%, e.g. SSB estimates for 2016 were revised from 161.1 tonnes estimated in the 2016 assessment (ICES, 2016), to 133.4 tonnes estimated in the 2017 assessment (ICES, 2017). These revisions propagated through estimates of reference points, the perception of stock status, and catch advice, most likely impacting fishing opportunities for the industry as well. Worst of all, the EU management plan for cod was paused based on the new perception of stock status, only for a few years later having fishing mortality above the limit reference point and biomass approaching historical low levels (ICES, 2020). Needless to say, this instability and lack of robustness in the scientific advice for a major iconic stock, with large sums invested in studying the stock and fisheries dynamics, may have a major negative impact on the reputation of ICES and scientific advice in general.

Unfortunately, the tools currently used for advice are sensitive to alternative representations of the system, model assumptions and new data. To deal with the potential lack of robustness of fisheries advice, we suggest to expand the assessment modelling basis integrating across multiple sources of uncertainty with ensemble models. This paper presents the authors' ruminations about how ensemble models can be used to improve scientific advice, making it more robust to changes in data or system drivers, while still maintaining operational feasibility. No conclusive solution is provided here! We offer suggestions and speculations that will hopefully raise awareness about ensemble models and foster the creativity and interest of our fellow scientists.

Ensemble models are a class of methods that combine several individual models' predictions into quantities of interest (QoI) integrating across all models in the ensemble set. The same way an ecosystem is more resilient to changes if its diversity is high (e.g. Chapin *et al.*, 2000; Folke *et al.*, 2004), we are of the opinion that scientific advice could also be more robust if it incorporates results from more than one model (e.g. Anderson *et al.*, 2017).

Furthermore, in the case of substantial assessment or forecast model uncertainty, building multiple models to better explain and predict the target system seems a logical approach.

The ensemble model approach has been widely adopted in other scientific fields like weather and climate science (e.g. see Gneiting and Raftery, 2005; Tebaldi and Knutti, 2007; Semenov and Stratonovitch, 2010; Chandler, 2013; Bauer *et al.*, 2015), econometrics (e.g. see Bates and Granger, 1969; Clemen and Winkler, 1986; Wright, 2009; Cuaresma, 2010; Chakraborty and Joseph, 2017), medicine (e.g. see Muhlestein *et al.*, 2018; Caballero-Alfonso *et al.*, 2019), and geology (e.g. see Gulden *et al.*, 2008; Wellmann *et al.*, 2010).

In fisheries science, a fairly large portfolio of work using ensemble models has been published in the peer-reviewed literature. These papers use a variety of techniques, including simple arithmetic averages, Bayes factors, cross-validation, and machine-learning. Furthermore, the applications span models dealing with single-species, multi-species, and ecosystems.

Among single-species applications of ensemble modelling, Brodziak and Legault (2005) and Brodziak and Piner (2010) evaluated reference points, stock status, and rebuilding targets for commercially harvested finfish. Brandon and Wade (2006) explored model structure and the presence of density dependence for Bowhead whales, *Balaena mysticetus*. Bayes factors were used to construct model averaged results for the ensemble of models considered in these three studies. For Pacific halibut, Hippoglossus stenolepis, Stewart and Martell (2015) looked at the impact of three different weighting schemes (including equal weighting) on the statistical distribution of management quantities, while Stewart and Hicks (2018) explored the behaviour of model ensembles when additional data are added (equal weights were applied to the models in the ensemble). Scott *et al.* (2016) explored a range of uncertainties in model structure and biological processes for a single species using generalized cross-validation to weight individual models. Of these single-species studies, only Brandon and Wade (2006) and Stewart and Martell (2015) were used to inform managers, while the other studies focused on demonstrating a particular approach.

Ianelli *et al.* (2016) considered both single- and multi-species models, exploring temperature relationships and future climate scenarios. Due to differences in statistical weighting and the degree of data aggregation within the models, ensemble results were calculated as a simple arithmetic average of individual models. This study was illustrative rather than directly used to inform managers.

In the context of multi-species models, Thorpe *et al.* (2015) compared ensemble averages for reference points and response to management action for single species and multi-species communities. Spence *et al.* (2018) made projections from five different ecosystem models assuming no fishing, treating the component models as exchangeable units in a hierarchical analysis. This analysis decomposed QoIs into discrepancies between the ensemble estimate and the quantity being fit, and discrepancies between each component model and the ensemble estimate. Neither of these studies directly informed management advice.

Another type of ensemble models, "super-ensembles", have recently received attention in fisheries. Super-ensembles refer to a technique where the ensemble is built by modelling the predictions of the ensemble components, which may include co-variables that were not present in any of the models. Anderson *et al.* (2017) and Rosenberg *et al.* (2014) fitted data-limited models to data

from hundreds of global fisheries. Super-ensembles were first formed by fitting the data-limited models to simulated data, and estimating a statistical relationship between the model predictions and simulated values. The data-limited models were then fit to empirical data, and the previously fitted statistical model was used to create super-ensemble results from the data-limited model fits. These studies did not inform management, but rather they explored the super-ensemble approach and compared results with existing studies on the same datasets (Rosenberg *et al.*, 2014), or compared ensemble results with those from individual models in the ensemble (Anderson *et al.*, 2017).

The studies mentioned highlight both the interest and the ability to apply ensemble modelling approaches in fisheries science. However, it also highlights the limited current use of ensemble models to provide management advice. The standard process to provide scientific advice is still strongly grounded in selecting a single stock assessment framework, and a single configuration, from a set of competing candidate models and configurations.

The following sections will explore methodological issues (Ensemble models: methods and applications section) and discuss the utilization of ensembles (Discussion section) in support of stock assessment and provision of advice to fisheries managers and policy makers.

### Ensemble models: methods and applications

Ensemble models combine predictions of a set of models into unified QoIs, integrating across model structures and associated uncertainties. In order to develop ensemble models, two important subjects need to be explored (i) which models are included in the ensemble, the ensemble members, and (ii) which method is used to combine models' outcomes and estimate QoIs, potentially including a decision about weighting metrics. On the other hand, the objective of the analysis will dictate the data characteristics of the QoIs and their application for scientific advice. The following sub-sections will describe limitations and potential solutions related with the ensemble composition, review a variety of methods and metrics to combine models' results, and describe ensemble model data products and applications.

### Ensemble composition

A major crux of ensemble modelling relates to the ensemble composition and the decision of which models should be included in the ensemble, the ensemble members. Including models that are too similar may end up over-weighting a particular outcome. Whereas including very different models may generate results without any overlap in the solution space, leading to multimodal outcomes. Both cases would fail to provide a balanced representation of structural uncertainty.

Addressing this central issue involves identifying the core factors that affect the fisheries system. In particular, if ensembles are used to integrate across structural uncertainty, one should try to capture the several possible, although not necessarily equally likely, working hypotheses about alternative states of nature (Chamberlin, 1965). We refer to this theoretical set of models as the model space, a complete and continuous representation of the system dynamics by models with different structures.

Acknowledging that fisheries systems are too complex to be described by a single model (Chatfield, 1995; Draper, 1995; Tebaldi and Knutti, 2007; Millar *et al.*, 2015; Stewart and Martell, 2015), ensemble members may be chosen by their capacity to model

different parts of the system and thus capture structural uncertainty. Model structure may refer to assumed functional form of biological or fishery processes, model complexity or observation equations that attempt to deal with uncertainty about data. The ensemble members should be complementary and ensemble methods should integrate across distinct representations of the system to estimate QoIs, hopefully covering the most important processes.

In contrast to structural uncertainty, ensemble members may be chosen to deal with parametric uncertainty assuming different fixed values of an uncertain model parameter, such as natural mortality, to test the effect on QoIs. In such a case, the ensemble model integrates over the distribution of parameter values that were deemed plausible. This type of sensitivity analyses (Palmer *et al.*, 2005), which may be used to test the robustness of model results to parametric assumptions, is referred to as "perturbated-parameter ensemble" by Flato *et al.* (2013).

Finally, to integrate across uncertainty related with initial conditions, ensemble members may be chosen to reflect multiple starting points, e.g. different initial year (e.g. Stewart and Martell, 2015) or fishing history. A well-known case is weather forecasting where ensembles are built to deal with the chaotic tendency of weather dynamics and uncertainty in initial conditions (Palmer *et al.*, 2005; Tebaldi and Knutti, 2007).

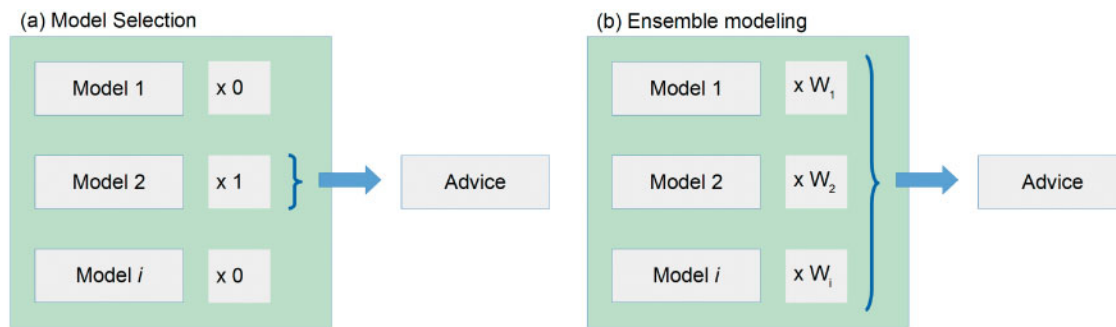
Understanding that structural uncertainty has a major impact in the ensemble outcomes forces the analysts to rethink their approach to model building. Instead of choosing the "best model" at the end of the model selection process, ensemble modelling requires a full range of models to be defined at the beginning of the modelling process. Figure 1 depicts simplified workflows of model selection and ensemble modelling. The differences between the two processes do not seem too extreme, although ensemble modelling will require much more emphasis on choosing models, metrics, methods, and QoIs than a conventional selection process, where models are discarded until the best one emerges.

Draper (1995) recognized the impossibility of identifying ensemble members, which fully cover the model space. The author suggested that instead of including every possible model only a set of plausible models needs to be identified. The author proposed a process of model expansion that extends an initial single model to include structural uncertainties expected to have non-zero probability of representing the true system. This model set would be sub-optimal, although if built in a standardized process could constitute the reference set to integrate structural uncertainty.

Operationally, the identification of plausible sets of ensemble models could be generalized to apply to many stocks or could be developed individually for each stock as part of specific Terms of Reference for the assessment work plan. Experience with either option will provide valuable feedback for improving the identification of ensemble model members in future applications.

### Methods and metrics

There are several methods that can be used to combine models' outcomes and estimate QoIs. The most common way to compute ensembles' estimates is to use some version of model weighting (Raftery *et al.*, 2005; Dormann *et al.*, 2018) and an analytical or resampling approach. For example, in the former case, a weighted average could be used to estimate a QoI, while for the latter the weights could be transformed into probabilities to draw



**Figure 1.** Simplified conceptual workflow comparison between conventional model selection (a) and ensemble modelling (b) in the context of stock assessment and advice provision. In the case of model selection (a), candidate models are analysed to find the “best” (weight set to one), which is then used for advice, while all the other models are discarded (weights set to zero). For ensemble modelling (b), all candidate models are kept and combined (curly bracket) using probabilities or weights ( $W_i$ ). The greenish square represents an Expert Working Group, which lays the ground for advice. The blue arrow represents the advisory process, which tends to differ across constituency.

resamples from each model and build the QoI empirical distribution. More sophisticated methods can be designed, though. In the machine-learning community, methods like boosting, bagging and stacking are commonly used (Breiman, 1996; Dietterich, 2000; Hastie *et al.*, 2001; Schapire and Freund, 2012; Yao *et al.*, 2018). These methods are mostly related with regression and classification analysis, which are of limited value for stock assessment and forecasting. Furthermore, super-ensembles provide a promising methodology where models’ weights are obtained through modelling the outcomes of each member using, e.g. linear models in a supervised learning framework (Anderson *et al.*, 2017).

In their comprehensive review of model averaging in ecology, Dormann *et al.* (2018) describe three approaches to set model weights: Bayesian, information theory based, and tactical. Each of these approaches differs in their assumptions, data requirements, treatment of individual candidate models, and numerical algorithms.

Bayesian approaches build model weights based on the posterior model probabilities of each model. A Bayesian ensemble prediction of a QoI can be calculated as the weighted average of individual model predictions by posterior model probabilities (Dormann *et al.*, 2018). An alternative, simplified Bayesian ensembles, can be built using the Bayesian information criterion approximation to Bayes factors (Kass and Raftery, 1995; Brodziak and Legault, 2005; Aho *et al.*, 2014).

Information theory metrics are based on statistics that reflect the information content of the model, like the Akaike information criterion (AIC; Burnham and Anderson, 2002) or some derivative of it. A disadvantage of information theory metrics is the potential to over-penalize models in the ensemble (for the AIC differences of more than four AIC points; Burnham and Anderson, 2002), resulting in all the weight being given to one or very few models. A restriction to using information theory metrics is that the data must be the same (Burnham and Anderson, 2002). In assessment models, this restriction would also extend to the data weighting that is sometimes specified, i.e. scores between models would not be comparable if different data weights are assumed in each model.

Tactical weights are based on the models’ capability of forecasting or predicting QoIs. Historical performance of each model, hindcasts, cross-validation, experts’ opinions, or a mix of several of the aforementioned methods can be used to compute these

metrics. The idea is to capture a model feature that is relevant for the analysis’ objective. For example, if the ensemble is used to forecast, then using each members’ forecast skills seems intuitive. An advantage of this approach is that one could relax the restrictions for information theory metrics and potentially extend tactical metrics to encompass several modelling approaches.

Otherwise, assigning equal weights avoids the decision about the weighting type, although it may simply shift the focus to decisions about ensemble’s composition. Assuming all models are equally likely representations of the natural system is probably unrealistic and equal weighting of an unlikely model could degrade the ensemble performance.

To address the possibility that models portraying the same or similar states of nature are over-represented in the ensemble, a model clustering two-step combination procedure could be used to build model weights and mitigate the impact of correlated models in the ensemble’s composition. This is similar to what Burnham and Anderson (2002) did to deal with model redundancy. Distinct model groups were given equal prior weights which were then shared equally among redundant models within a group. A difficulty with these authors approach is that it requires the analyst to identify the redundant models *a priori*, which is usually not possible in fisheries science. Our suggestion is to use a *post hoc* clustering procedure. In both cases, there will be difficulties associated with the fact that the several QoIs these models produce may cluster in different ways.

An open issue related to model weights is how to take into account the historical performance of metrics. A metric could be designed to vary along the period included in the analysis, e.g. it may have time blocks with different values. Such approach is not referred to in the literature, although it may be interesting to explore, considering how regime shifts or changes in fleet behaviour affect the historical performance of individual models.

## Applications

Ensemble modelling can generate several QoIs, which provide diverse insights into the dynamics of stocks and fisheries. Consequently several applications can be foreseen in the context of scientific advice to fisheries managers and policy makers. Nevertheless, it is important to bear in mind that QoIs have certain numerical characteristics, which will determine both the



complexity of their estimation and utility for applications. A single variable and its statistical distribution are a lot simpler to compute than a full matrix of population abundance and the complex multi-variate distribution associated with it. On the other hand, the utility of both cases is also very different, with the former limiting much more the analysis than the latter.

In our opinion, the most promising applications for scientific advice are estimating stock status, setting future fishing opportunities, and building operating models. Estimating stock status, which requires estimating fishing mortality, biomass, and reference points, combines multiple stock assessment models' estimates to derive QoIs. Setting future fishing opportunities, which in the European Union policy most of the times refers to setting Total Allowable Catches, uses projections of future catches or fishing effort limits estimated by several models to build an ensemble estimation of such QoIs. In this case, the distinct models take into account their own estimates of stock dynamics and pre-defined management options and objectives. Finally, to build operating models, complementary representations of stocks and fleets' dynamics by multiple models and approaches can be used in simulation testing and Management Strategies Evaluation (MSE) analysis.

In relation to the characteristics of QoIs derived from ensemble models, we suggest the following classification regarding their numerical characteristics, in the ascending order of complexity:

- Univariate: the outcome of the ensemble is a single QoI, e.g. MSY and its distribution. These can often be derived with analytical methods.
- Multivariate: the outcome is a set of QoIs, which may be related with each other, e.g. stock status in the final year of the model ( $B/B_{MSY}$  and  $F/F_{MSY}$ ). It is not usually possible to derive such a distribution analytically; resampling methods will typically be needed.
- Time series: the ensemble outcome is a time series, e.g. spawning stock biomass. An analytical solution may be difficult to derive and using resampling methods may be the best option, in which case it is important to take into account auto-correlation.
- Matrix or array: the outcome is a matrix, e.g. population numbers at age. An analytical solution may be difficult to derive and using resampling methods may be the best option, in which case it is important to take into account within model correlations across ages and years.
- Full stock and fisheries dynamics: the ensemble is used to build operating models that require several matrices. In such cases, metrics that need to have some degree of coherence across them have to be combined, e.g. abundance in numbers at age and fishing mortality at age. Analytical solutions are not available and using resampling methods seem to be the only alternative, in which case correlation structures need to be accounted for, both internal to the variable and across variables.

The complexity level of the different applications - stock status, forecast, and operating models - will determine how many of these QoIs will be necessary. To estimate the status of a stock, a single or bivariate variable may be sufficient. When it comes to forecasts, a full understanding of the stock exploitation history and productivity will be necessary, and QoIs will be time series of

projections under certain conditions. In data-rich situations, forecasts will also use matrices, like population abundance and selectivity by age or length. Obviously, information about the status of the stock(s), mentioned above, will be needed to set proper conditions for future fishing opportunities analysis. With regard to building operating models, all of the previous will be needed plus several age or length structures of the population, fleet selectivity, population productivity, and, although less common, socio-economic information. In this case, several correlated matrices will need to be included in the ensemble results.

## Discussion

In our opinion, ensemble modelling can be useful in the context of providing scientific advice to fisheries managers and policy makers in the following non-mutually exclusive situations: (i) to include structural uncertainty across different models of the same system, (ii) to better report scientific uncertainty, and (iii) to integrate across alternative, and potentially complementary, processes or parameterisation. Furthermore, there are three main applications that can be improved by using ensemble models: (i) estimate stock status, (ii) forecast future fishing opportunities, and (iii) build operating models.

Nevertheless, ensemble models are not a panacea. [Dormann et al. \(2018\)](#) showed situations where use of ensembles improves the individual models' predictions and others where it has no effect or even degrades individual estimates. [Stewart and Hicks \(2018\)](#) showed that correlation across ensemble members can jeopardize the ensemble utility in integrating structural uncertainty.

There are a number of challenges to overcome in order to fully integrate ensemble models outputs in advice. Some are not different to those faced by other methods, like how to frame probabilistic outcomes in the advisory context or engage stakeholders. Communication is a key step if results are to be successfully used and accepted ([Miller et al., 2019](#)), no matter which model is behind those results. Ensemble models should make use of generic approaches applied elsewhere. However, the added complexity of multi-model integration may exacerbate those difficulties. A facilitating factor would be to fully disclose the analysis algorithm and provide full replicable results. Although a non-technical audience of policy makers and other stakeholders may not fully understand the technical details of the analysis, there will clearly be more confidence in the results if both data and analysis algorithm are fully disclosed for public scrutiny.

Other challenges are specific to ensemble modelling, like choosing ensemble members and model weights. On the one hand, including similar models may overweight a specific model configuration, not due to their representativeness but to biases introduced in the ensemble's composition. On the other hand, if model predictions are correlated, despite all being legitimate representations of relevant states of nature, one may end up penalizing realistic models and possibly biasing results to extreme or unlikely fits. A potential solution in the context of scientific advice would be to decide the ensembles' composition and methods during a benchmark exercise, and keep that setting for a number of years (see model expansion by [Draper, 1995](#)). In addition, a two-level weighting process, where hypotheses are on the first level and model skill nested within, would not be too complex to implement and could create the necessary interest to further develop and refine the methodology, making it more operational for stock assessment working groups. A number of technicalities

could surface, e.g. how to compute reference points from an ensemble model, or how to provide fishing opportunities advice. Nevertheless, these issues should not be any different from approaches taken for other probabilistic models, risk analysis, or analysis of scenarios.

Notably, the same careful decisions about data inclusion and justifiable model structure that are taken to arrive at a single best model should be maintained when deciding on an ensemble's members. The ensemble composition should not be treated as a dumpster for group indecision, nor should non-credible model structures be included with the hope that the analysis will reject or severely penalize them. While these decisions can be difficult or even contentious, they should be confronted at the start of the ensemble building, and justifications clearly documented. Such an approach, using benchmark workshops to explore the utility of ensemble models, could foster collaboration among scientists, promote transparency, and maintain the objectiveness of the scientific process.

Moving from the current single best model approach to an ensemble approach is not as big a step as it may seem. Current practices already require fitting and setting up several models for the same stock. This practice could be compared to an ensemble modelling exercise, where one model will have all the weight and all others have none (Figure 1). For example, the work done choosing the best model for a stock during a benchmark, or sensitivity analysis carried out to evaluate if the assessment results are robust to misspecifications of model assumptions, could both be the starting point of an ensemble modelling exercises. It is not common to build ensembles from these model trials, taking instead a decision about the "best" model, discarding all the other candidates and not reporting the uncertainty of the selection process itself. It should not be a surprise that often the chosen models fail to fit properly when new information is added. After all, one model is just one simplified representation of a very complex system among the several possible. Ensemble models would make use of many models and integrate across the uncertainty of the selection process itself (Chatfield, 1995; Brodziak and Legault, 2005; Raftery et al., 2005; Grueber et al., 2011; Claeskens, 2016) avoiding overconfidence in results. This would be helpful in situations where major changes in estimates of stock status, stock magnitude, and management advice have resulted from data revisions, changes in model assumptions (e.g. natural mortality), or changes in model structure from one assessment to the next. We expect an ensemble model framework to be more stable than any single model and therefore to provide a more robust advice.

The current spectrum of stock assessment methods is very diverse. Analytical methods, which require age- or length-based data, range from virtual population analysis to state-space models including statistical catch-at-age methods. Data-limited methods include dozens of alternatives. Such diversity is important to maintain. Limiting the scientific community to a small set of models would definitely have a high impact on the resilience and creativity of scientific advice. Ensembles could be used to integrate across these models provided QoIs are in comparable units. In theory, there is no limitation to the types of models that can be used in an ensemble. One should be able to combine their results as long as their outcomes can be transformed into common variables. In practice though, if models have very different structures it may be difficult to find a common metric (Kaplan et al., 2018) imposing limits to the diversity of models that can be included in an ensemble.

Further development of general, modular, extensible, well-tested, and well-documented software systems is required. The lack of consistency in the output from the plethora of available stock assessment frameworks is probably one of the main factors limiting an immediate trial of ensemble models. Although difficulties are inevitable when dealing with real cases, having a common framework should allow solutions to be discussed and shared within a large group of people dealing with similar problems. We therefore emphasize the importance of standardizing formats of assessment outputs to facilitate collaboration and model comparisons and make the process of ensemble modelling more efficient.

Processes to build ensemble models and develop performance metrics, algorithms, etc., require additional work before they become fully functional for scientific advice. In our opinion, future studies should explicitly test the process of building the ensemble, comparing the feasibility of combining outcomes from models of varying complexity, and exploring the frequency of updating model weights. Simulation studies like those supporting MSEs could be useful to test these methods. Operating models based on theoretical ecology, not a particular stock assessment model fit, could provide the data generation mechanism to test different estimators. The estimator is the MSE component that mimics the stock assessment working group, where pseudo-observations are transformed into QoIs for the advisory process, for example stock status estimates to feed a harvest control rule (HCR). It can encompass anything, from a single data-limited methodology up to a complex ensemble model, providing the simulation testing framework required. Best practices on developing MSEs would need to be followed to avoid the expected optimistic outcomes that models generate, e.g. using more than one operating model, testing several sampling mechanisms, adjusting the HCR to the estimator outcome, etc. (e.g. see Punt et al., 2016).

In our opinion, pursuing these paths of research will provide tools to improve the robustness and stability of scientific advice and will promote transparency regarding scientific uncertainty.

## Acknowledgements

This work was funded by the European Commission Joint Research Centre Exploratory Research Programme. The scientific results and conclusions, as well as any views and opinions expressed herein, are those of the author(s) and do not necessarily reflect those of NOAA or the Department of Commerce. There are no new data associated with this article.

## References

- Aho, K., Derryberry, D., and Peterson, T. 2014. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95: 631–636.
- Anderson, S. C., Cooper, A. B., Jensen, O. P., Minto, C., Thorson, J. T., Walsh, J. C., Aflerbach, J., et al. 2017. Improving estimates of population status and trend with superensemble models. *Fish and Fisheries*, 18: 732–741.
- Bates, J. M., and Granger, C. W. J. 1969. The combination of forecasts. *Journal of the Operational Research Society*, 20: 451–468.
- Bauer, P., Thorpe, A., and Brunet, G. 2015. The quiet revolution of numerical weather prediction. *Nature*, 525: 47–55. doi: 10.1038/nature14956
- Brandon, J., and Wade, P. 2006. Assessment of the Bering-Chukchi-Beaufort Seas stock of bowhead whales using Bayesian model averaging. *Journal of Cetacean Research Management*, 8: 225–239.

- Breiman, L. 1996. Bagging predictors. *Machine Learning*, 24: 123–140.
- Brodziak, J., and Legault, C. M. 2005. Model averaging to estimate rebuilding targets for overfished stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, 62: 544–562.
- 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.
- Burnham, K., and Anderson, D. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer Verlag, New York.
- Caballero-Alfonso, A. Y., Cruz-Montegudo, M., Tejera, E., Benfenati, E., Borges, F., Cordeiro, M. N. D. S., Armijos-Jaramillo, V., *et al.* 2019. Ensemble-based modeling of chemical compounds with antimalarial activity. *Current Topics in Medicinal Chemistry*, 19: 957–969.
- Chakraborty, C., and Joseph, A. 2017. Machine learning at central banks. Staff Working Paper (674).
- Chamberlin, T. C. 1965. The method of multiple working hypotheses. *Science*, 148: 754–759.
- Chandler, R. E. 2013. Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371: 20120388–20120388.
- Chapin III, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., *et al.* 2000. Consequences of changing biodiversity. *Nature*, 405: 234–242.
- Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158: 419–466.
- Claeskens, G. 2016. Statistical model choice. *Annual Review of Statistics and Its Application*, 3: 233–256.
- Clemen, R. T., and Winkler, R. L. 1986. Combining economic forecasts. *Journal of Business & Economic Statistics*, 4: 39–46.
- Cuaresma, J. C. 2010. Can emerging asset price bubbles be detected? OECD Economics Department Working Papers (772).
- Dietterich, T. G. 2000. Ensemble methods in machine learning. *In Multiple Classifier Systems*, pp. 1–15. Ed. by J. Kittler and F. Roli. Springer, Berlin, Heidelberg.
- Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., *et al.* 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88: 485–504.
- Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57: 45–97.
- Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., *et al.* 2013. Evaluation of climate models. In *Climate change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. Stocker, *et al.* Cambridge University Press, Cambridge.
- Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., and Holling, C. 2004. Regime shifts, resilience, and biodiversity in ecosystem management. *Annual Review of Ecology, Evolution, and Systematics*, 35: 557–581.
- Gneiting, T., and Raftery, A. E. 2005. Weather forecasting with ensemble methods. *Science*, 310: 248–249.
- Grueber, C. E., Nakagawa, S., Laws, R. J., and Jamieson, I. G. 2011. Multimodel inference in ecology and evolution: challenges and solutions: multimodel inference. *Journal of Evolutionary Biology*, 24: 699–711.
- Gulden, L. E., Rosero, E., Yang, Z.-L., Wagener, T., and Niu, G.-Y. 2008. Model performance, model robustness, and model fitness scores: a new method for identifying good land-surface models. *Geophysical Research Letters*, 35:
- Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The Elements of Statistical Learning*, 1. Springer Series in Statistics. Springer, Berlin.
- Ianelli, J., Holsman, K., Punt, A., and Aydin, K. 2016. Multi-model inference for incorporating trophic and climate uncertainty into stock assessments. *Deep-Sea Research II*, 134: 379–389.
- ICES. 2014. *Report of the ICES Advisory Committee. Book 6. North Sea*.
- ICES. 2015b. Cod (*Gadus morhua*) in Subarea IV and Divisions VIIId and IIIa West (North Sea, Eastern English Channel, Skagerrak). ICES advice on fishing opportunities, catch, and effort. Greater North Sea and Celtic Seas Ecoregions, 1–16.
- ICES. 2015a. Report of the Benchmark Workshop on North Sea Stocks (WKNSEA), 2–6 February, Copenhagen, Denmark.
- ICES. 2016. Cod (*Gadus morhua*) in Subarea 4 and Divisions 7.d and 3.a.20 (North Sea, Eastern English Channel, Skagerrak). ICES Advice on fishing opportunities, catch, and effort. Greater North Sea and Celtic Seas ecoregions (update), 1–20.
- ICES. 2017. Cod (*Gadus morhua*) in Subarea 4 and Divisions 7.d and Subdivision 20 (North Sea, Eastern English Channel, Skagerrak). ICES Advice on fishing opportunities, catch, and effort. Greater North Sea and Celtic Seas ecoregions, 1–23.
- ICES. 2020. Cod (*Gadus morhua*) in Subarea 4 and Divisions 7.d and Subdivision 20 (North Sea, Eastern English Channel, Skagerrak). ICES Advice on fishing opportunities, catch, and effort. Greater North Sea and Celtic Seas ecoregions, 1–23.
- Kaplan, I., Francis, T., Punt, A., Koehn, L., Curchitser, E., Hurtado-Ferro, F., Johnson, K., *et al.* 2018. A multi-model approach to understanding the role of Pacific sardine in the California Current food web. *Marine Ecology Progress Series*, 1–15.
- Kass, R., and Raftery, A. 1995. Bayes factors. *Journal of the American Statistical Association*, 90: 773–795.
- Millar, C. P., Jardim, E., Scott, F., Osio, G. C., Mosqueira, I., and Alzorric, N. 2015. Model averaging to streamline the stock assessment process. *ICES Journal of Marine Science*, 72: 93–98.
- Miller, S. K., Anganuzzi, A., Butterworth, D. S., Davies, C. R., Donovan, G. P., Nickson, A., Rademeyer, R. A., *et al.* 2019. Improving communication: the key to more effective mse processes. *Canadian Journal of Fisheries and Aquatic Sciences*, 76: 643–656.
- Muhlestein, W. E., Akagi, D. S., Kallos, J. A., Morone, P. J., Weaver, K. D., Thompson, R. C., and Chambless, L. B., 2018. Using a guided machine learning ensemble model to predict discharge disposition following meningioma resection. *Journal of Neurological Surgery Part B: Skull Base*, 79: 123–130.
- Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., and Weisheimer, A. 2005. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360: 1991–1998.
- Punt, A. E., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A. A., and Haddon, M. 2016. Management strategy evaluation: best practices. *Fish and Fisheries*, 17: 303–334.
- Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. 2005. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133: 1155–1174.
- Rosenberg, A. A., Fogarty, M. J., Cooper, A., Dickey-Collas, M., Fulton, E., Gutiérrez, N., Hyde, K. J. W., *et al.* 2014. *Developing New Approaches to Global Stock Status Assessment and Fishery Production Potential of the Seas (No. 1086)*. Food & Agriculture Organization.
- Schapire, R. E., and Freund, Y. 2012. *Boosting—Foundations and Algorithms*. The MIT Press.
- Scott, F., Jardim, E., Millar, C. P., and Cerviño, S. 2016. An applied framework for incorporating multiple sources of uncertainty in fisheries stock assessments. *PLoS One*, 11: e0154922.

- Semenov, M. A., and Stratonovitch, P. 2010. Use of multi-model ensembles from global climate models for assessment of climate change impacts. *Climate Research*, 41: 1–14.
- Spence, M. A., Blanchard, J. L., Rossberg, A. G., Heath, M. R., Heymans, J. J., Mackinson, S., Serpetti, N., *et al.* 2018. A general framework for combining ecosystem models. *Fish and Fisheries*, 19: 1031–1042.
- Stewart, I. J., and Hicks, A. C. 2018. Interannual stability from ensemble modelling. *Canadian Journal of Fisheries and Aquatic Sciences*, 75: 2109–2113.
- Stewart, I. J., and Martell, S. J. D. 2015. Reconciling stock assessment paradigms to better inform fisheries management. *ICES Journal of Marine Science*, 72: 2187–2196.
- Tebaldi, C., and Knutti, R. 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365: 2053–2075.
- Thorpe, R. B., Le Quesne, W. J. F., Luxford, F., Collie, J. S., and Jennings, S. 2015. Evaluation and management implications of uncertainty in a multispecies size-structured model of population and community responses to fishing. *Methods in Ecology and Evolution*, 6: 49–58.
- Wellmann, J. F., Horowitz, F. G., Schill, E., and Regenauer-Lieb, K. 2010. Towards incorporating uncertainty of structural data in 3d geological inversion. *Tectonophysics*, 490: 141–151.
- Wright, J. H. 2009. Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, 28: 131–144.
- Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. 2018. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, 13: 917–1003.

*Handling editor: Shijie Zhou*