Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

Operationalizing ensemble models for scientific advice to fisheries management

Ernesto Jardim[1,2*], Manuela Azevedo[3], Jon Brodziak[4], Elizabeth N. Brooks[5], Kelli F. Johnson[6], Nikolai Klibansky[7], Colin P. Millar[8], Coilin Minto[9], Iago Mosqueira[1**], Richard D.M. Nash[10***], Paris Vasilakopoulos[1], and Brian K. Wells[11]

[1]European Commission, Joint Research Centre, Sustainable Resources Directorate, Water and Marine Resources Unit, Via Enrico Fermi, 21027 Ispra (VA), Italy
[2]Marine Stewardship Council, Marine House, Snow Hill 1, London EC1A 2DH, UK
[3]Portuguese Institute for the Sea and Atmosphere (IPMA). Av. Doutor Alfredo Magalhães Ramalho, 6, 1495-165 Algés, Portugal
[4]Pacific Islands Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Honolulu, HI, United States
[5]Northeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Woods Hole, MA, United States
[6]Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, United States
[7]Southeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Beaufort, NC, United States
[8]International Council for the Exploration of the Sea (ICES), H. C. Andersens Boulevard 44-46, 1553 Copenhagen V, Denmark
[9]Marine and Freshwater Research Centre (MFRC), Galway-Mayo Institute of Technology (GMIT), Dublin Road, Galway, Ireland.
[10]Institute of Marine Research, P.O. Box 1870 Nordnes, 5817 Bergen, Norway
[11]Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Santa Cruz, CA, United States
[*]Corresponding author: ernesto.jardim@msc.org
[**]Present address: Wageningen Marine Research. Haringkade 1. 1976CP, IJmuiden, The Netherlands
[***]Present address: Centre for Environment, Fisheries and Aquaculture Science (Cefas), Pakefield Road, Lowestoft, Suffolk, NR33 0HT, UK

August 2, 2020

 doi:10.20944/preprints202008.0078.v1

Abstract

There are uncertainties associated with every phase of the stock assessment process, ranging from the collection of data, assessment model choice, model assumptions and interpretation of risk to the implementation of management advice. 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 to provide advice to managers about sustainable harvests. The *status quo* approach to assessment modelling has been to identify the 'best' model, based on diagnostics and model selection criteria, and to generate advice from that model, mostly ignoring advice from other model configurations regardless of how closely they performed relative to the chosen model. We review the suitability of the ensemble modelling paradigm to more fully capture uncertainty in stock assessment model building and the provision of advice. We recommend further research to evaluate potential gains in modelling performance and advice from the use of ensemble modelling, while also suggesting revisions to the formal process for reviewing models and providing advice to management bodies.

keywords: assessment, conservation, exploitation, management, multimodel, natural resources, ensemble, model, fisheries

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

1 Introduction

Providing scientific advice to fisheries managers is a risky activity! It is not uncommon that a model which was performing well suddenly fails to properly fit an additional year of data or projections made in the past did not materialise when more recent information 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 have the insidious tendency to ruin what had been a perfectly acceptable assessment fit, invalidating one's advice and weakening confidence in future advice efforts.

Unfortunately, tools currently used to provide fisheries 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 expanding the assessment modelling basis by integrating across multiple sources of uncertainty using 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 the data or system drivers, while still maintaining operational feasibility. No conclusive solution is provided here! We offer ideas and speculations which hopefully will raise awareness about ensemble models and foster the creativity and interest of 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 III 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 adopted in other scientific fields like weather and climate science (*e.g.*, see Bauer et al., 2015; Gneiting & Raftery, 2005; Semenov & Stratonovitch, 2010; Tebaldi & Knutti, 2007; Chandler, 2013), econometrics (*e.g.*, see Wright, 2009; Bates & Granger, 1969; Clemen & Winkler, 1986; Cuaresma, 2010; Chakraborty & Joseph, 2017), medicine (*e.g.*, see Muhlestein et al., 2018) and geology (*e.g.*, see Gulden et al., 2008; Wellmann et al., 2010).

In fisheries science, a fairly large portfolio of work using ensemble modelling 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; the applications span models dealing with single species, multiple species and ecosystems.

Among single species applications of ensemble modelling, Brodziak & Legault (2005) and

Brodziak & Piner (2010) evaluate reference points, stock status and rebuilding targets for commercially harvested finfish; Brandon & 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 & Martell (2015) looked at the impact of three different weighting schemes (including equal weighting) on the statistical distribution of management quantities, while Stewart & Hicks (2018) explored the behavior 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 the individual models of the overall ensemble. Of these single-species studies, only Brandon & Wade (2006) and Stewart & Martell (2015) were used to inform managers, while the other studies focused more 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 intended to directly inform managers.

In the context of multi-species models, Thorpe et al. (2015) compared ensemble averages for reference points and response to management actions for single- 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's components, which may include co-variates that were not present in any of the models. Anderson et al. (2017) and Rosenberg et al. (2014) fit data-limited models to data from hundreds of global fisheries. Super-ensembles were then 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 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 current interest and the ability to apply ensemble modelling approaches in fisheries science. Although, they also point to the limited use of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

ensemble models to provide management advice. The current process of 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 explore methodological issues (section 2) and discuss the utilization of ensembles (section 3) in support of stock assessments and provision of advice to fisheries managers and policy makers.

2 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, also called ensemble members and (ii) which weighting method is used to combine models' outcomes and estimate QoIs. 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 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.

2.1 Ensemble composition

A major crux of ensemble modelling relates to the ensemble's composition and the decision of which models should be included in the ensemble. 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 and multi-modal outcomes without a clear message that are sensitive to choices about both ensemble metrics and weighting methods.

Addressing this central issue involves identifying the core causal 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 theorectical 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 (Tebaldi & Knutti, 2007; Chatfield, 1995; Draper, 1995; Stewart & Martell, 2015; Millar et al., 2015), ensemble members may be chosen by their capacity to model different parts of the system and capture structural uncertainty. The ensemble members should be complementary and ensemble methods should integrate across distinct representations of the system, hopefully covering the most important processes, to estimate QoIs.

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, and test their 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), commonly 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 to initial conditions, ensemble members may be chosen to reflect multiple starting points. A well known case is weather forecasting where ensembles are built to deal with the chaotic tendencies of weather dynamics (Palmer et al., 2005; Tebaldi & Knutti, 2007).

Understanding that structural uncertainty has a major impact in ensemble modelling, as it forces the analyst to rethink one's modelling approach, is key to the approach. Instead of choosing the 'best model' at the end of the model selection process, ensemble modelling requires defining a full range of models at the very beginning. 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, it could provide a reference set used 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 the discussion for specifying Terms of Reference for the assessment work plan. Future experience with both options will provide feedback for improving the identification of ensemble model members in future applications.

2.2 Methods and metrics

There are several methods that can be used to combine model outcomes and estimate QoIs. The most common way to compute ensemble 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 case, the weights could be transformed into probabilities used to resample from each model and build the QoI empirical distribution. Though, more sophisticated methods can be designed. In the machine learning community, methods like boosting, bagging and stacking are commonly used (Schapire & Freund, 2012; Breiman, 1996; Yao et al., 2018; Dietterich, 2000; Hastie et al., 2001). These methods are mostly related to regression and classification analysis, which are of limited value for stock assessment and forecasting. Furthermore, super-ensembles also 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) describes three approaches to set model weights, Bayesian, information-theory based and tactical. Each of these approaches differ 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 the individual model predictions times the posterior model probabilities (Dormann et al., 2018). An alternative, simplified Bayesian ensembles, can be built using the Bayesian information criterion (BIC) approximation to Bayes factors (Aho et al., 2014; Kass & Raftery, 1995; Brodziak & Legault, 2005).

Information theory metrics are based on statistics that reflect the information content of the model, like Akaike information criterion (AIC; Burnham & Anderson 2002) or some derivative of it. A disadvantage of information theory metrics is the potential to over-penalize models in the ensemble (for AIC, any model different by greater than 4 AIC points; Burnham & Anderson 2002), resulting in all the weight being given to one or very few models. A further restriction to using information theory metrics is that the data must be the same (Burnham & 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 feature of the model that is relevant for the analysis' objective. For example, if the ensemble is used to forecast, then using each members' forecast capability,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

also called model 'skill', 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 weighting type, although it may simply shift the focus to decisions about ensemble's composition, since assuming all models are equally likely representations of the natural system is probably unrealistic.

To address the possibility that models portraying the same, or a similar, state of nature are over-represented in the ensemble, Garthwaite & Mubwandarikwa (2010) suggest using a process similar to Principal Component Analysis (PCA) projections to build independent/orthogonal weights. In our opinion, for very complex systems, like exploited marine ecosystems, this approach seems of limited value as a large number of potentially overlapping models will need to be projected in a high dimensional space. Another potential solution could be to use model clustering and a two step combination procedure to build model weights and mitigate the impact of correlated models in the model space. Nevertheless, clustering represents an extra challenge because these models provide several QoIs which could cluster in different ways.

An open issue related to model weights is how to take into account the metric's historical performance. It is possible to conceptualize metrics that are not constant along the period included in the analysis and require revision at certain periods, or having 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 behavior affect the historical performance of individual models.

2.3 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 estimating and their utility for applications. A single variable and its statistical distribution is simpler to compute than a full matrix of population abundance and the complex multi-variate distribution associated with it. On the other hand, each of them provide very different material for analysis, the former limiting it much more than the latter.

In our opinion, the most promising applications of model ensembles for fisheries advice are estimating stock status, setting future fishing opportunities and building operating models. Estimating stock status can be accomplished by combining multiple stock assessment models' estimates to derive QoIs. Setting future fishing opportunities can use projections of future catches, or fishing effort limits, from 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 predefined management options and objectives. Finally, when building 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. Not very distinct from a multi operating model MSE.

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

- Univariate variable: The outcome of the ensemble is a single QoI, *e.g.*, a reference point, like B_{MSY}, and its distribution, which can often be derived using analytical methods.
- Multivariate variable: The outcome is a set of QoIs which may be related to each other, *e.g.*, biological reference points B_{MSY} and F_{MSY}. It is usually possible to derive a multivariate distribution analytically.
- Time series: The ensemble outcome is a time series, *e.g.*, spawning stock biomass by year. 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 auto-correlation into account.
- Matrix or array: The outcome is a matrix, *e.g.*, population numbers by year and 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 years and ages.
- Full stock and fisheries dynamics: The ensemble is used to build operating models that require several matrices. In such cases metrics which need to have some degree of coherence across them have to be combined, *e.g.*, abundance in numbers by year and age and fishing mortality by year and age. Analytical solutions are not available and using resampling methods seems to be the only alternative, in which case correlation structures need to be accounted for, both internal to the variable as well as across variables.

To clarify the relationship between QoIs and applications, Table 1 shows the linkage between the two. With the increasing complexity of the applications – stock status, forecast and operating models – the complexity of the data product also increases. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

or length. Obviously, information about the status of the stock(s), mentioned above, will be needed to set proper conditions for the analysis of future fishing opportunities. With regards 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 commonly used, socio-economic information. In this case, several correlated matrices will need to be included in the ensemble results.

3 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 parametrisations. Furthermore, there are three steps that could benefit from the use of ensemble models: (a) estimation of stock status; (b) forecasting of future fishing opportunities; and (c) building of operating models.

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

Choosing ensemble members seems to be one of the major challenges of ensemble modelling. Including similar models may overweight a specific model configuration, not due to the accuracy of its representation of the states of nature, due to biases introduced in the ensemble space. On the other hand, if model predictions are correlated, not due to model similarities, due to 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 for scientific advice would be to decide the ensembles' composition and methods during a benchmark exercise and maintain that agreement for a number of years (see model expansion by Draper, 1995). Such a procedure could foster collaboration among scientists, promote transparency and maintain objectiveness of the scientific process, and is already being followed to that effect in the development of operating models for MSE analyses (Sharma et al., n.d.). Unsurprisingly, 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 the ensemble 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.

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 the sensitivity analysis carried out to evaluate if the assessment results are robust to mis-specifications of model assumptions, could both be the starting point for ensemble modelling exercises. Despite all this work having been done, it is not common to build ensembles of these model trials, opting instead for making a decision about which model to choose, discarding all the other candidates and not reporting the uncertainty of the selection process itself. It should not be a surprise that often such models fail to fit properly when new information is added. After all, one model is just one simplified representation, amongst the several possible, of a very complex system. Ensemble models would make use of many models and integrate across the uncertainty of the selection process itself (Chatfield, 1995; Claeskens, 2016; Grueber et al., 2011; Brodziak & Legault, 2005; Raftery et al., 2005) avoiding overconfidence in results.

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 modelling frameworks 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, develop performance metrics, algorithms, etc. require additional work before becoming 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 ideal frequency of updating model weights. Simulation studies will be useful to develop and test diagnostics about individual model convergence and fit, as well as the weighted ensemble results. Furthermore, new data products may be generated which will require modifications in the way we communicate scientific information to managers, namely uncertainty and risk. 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.

4 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.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636.

Anderson, S. C., Cooper, A. B., Jensen, O. P., Minto, C., Thorson, J. T., Walsh, J. C., ... Selig, E. R. (2017). Improving estimates of population status and trend with superensemble models. *Fish and Fisheries*, *18*(4), 732–741.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451-468.

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*, 47. doi: 10.1038/nature14956

Brandon, J., & Wade, P. (2006). Assessment of the Bering-Chukchi-Beaufort Seas stock of bowhead whales using Bayesian model averaging. *Journal of Cetacean Research Management*, *8*(3), 225–239.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brodziak, J., & Legault, C. M. (2005). Model averaging to estimate rebuilding targets for overfished stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, *62*(3), 544–562.

Brodziak, J., & Piner, K. (2010). Model averaging and probable status of North Pacific striped marlin, *Tetrapturus audax*. *Canadian Journal of Fisheries and Aquatic Sciences*, *67*(5), 793–805.

Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* (2nd ed.). Springer Verlag, New York.

Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks. *Staff Working Paper*(674).

Chamberlin, T. C. (1965). The method of multiple working hypotheses. *Science*, *148*(3671), 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*(1991), 20120388–20120388.

Chapin III, F., Zavaleta, E., Eviner, V., Naylor, R., Vitousek, P., Reynolds, H., ... Díaz, S. (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*(3), 419–466.

Claeskens, G. (2016). Statistical model choice. *Annual Review of Statistics and Its Application*, *3*(1), 233–256.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*(1), 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). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., ... Hartig, F. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504. doi: 10.1002/ecm.1309

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 45–97.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., ... Rummukain, M. (2013). Climate change 2013: The physical science basis. contribution of working group i to the fifth assess-ment report of the intergovernmental panel on climate chan. In T. Stocker et al. (Eds.), (chap. Evaluation of Climate Models). Cambridge University Pres.

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., & Holling, C. (2004). Regime shifts, resilience, and biodiversity in ecosystem management. *Annual Review of Ecology, Evolution, and Systematics*, *35*, 557–581.

Garthwaite, P. H., & Mubwandarikwa, E. (2010). Selection of weights for weighted model averaging: Prior weights for weighted model averaging. *Australian & New Zealand Journal of Statistics*, *52*(4), 363–382.

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248-249.

Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions: Multimodel inference. *Journal of Evolutionary Biology*, *24*(4), 699–711.

Gulden, L. E., Rosero, E., Yang, Z.-L., Wagener, T., & 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*(11).

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Ianelli, J., Holsman, K., Punt, A., & Aydin, K. (2016). Multi-model inference for incorporating trophic and climate uncertainty into stock assessments. *Deep-Sea Research II*, *134*, 379–389.

Kaplan, I., Francis, T., Punt, A., Koehn, L., Curchitser, E., Hurtado-Ferro, F., … Levin, P. (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., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Millar, C. P., Jardim, E., Scott, F., Osio, G. C., Mosqueira, I., & Alzorriz, N. (2015). Model averaging to streamline the stock assessment process. *ICES Journal of Marine Science*, *72*(1), 93–98.

Muhlestein, W. E., Akagi, D. S., Kallos, J. A., Morone, P. J., Weaver, K. D., Thompson, R. C., … Chambless, L. B. (2018). Using a guided machine learning ensemble model to predict discharge disposition following meningioma resection. *J Neurol Surg B Skull Base*, *79*(2), 123–130.

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., & 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*(1463), 1991–1998.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174.

Rosenberg, A. A., Fogarty, M. J., Cooper, A., Dickey-Collas, M., Fulton, E., Gutiérrez, N., … Ye, Y. (2014). *Developing New Approaches to Global Stock Status Assessment and Fishery Production Potential of the Seas.* (No. 1086). Food & Agriculture Org.

Schapire, R. E., & Freund, Y. (2012). *Boosting. foundations and algorithms.* The MIT Press.

Scott, F., Jardim, E., Millar, C. P., & Cerviõ, S. (2016). An applied framework for incorporating multiple sources of uncertainty in fisheries stock assessments. *PLOS ONE*, *11*(5).

Semenov, M. A., & Stratonovitch, P. (2010). Use of multi-model ensembles from global climate models for assessment of climate change impacts [Journal Article]. *Climate Research*, *41*(1), 1-14.

Sharma, R., Levontin, P., Kitakado, T., Kell, L., Mosqueira, I., Kimoto, A., … Magnusson, A. (n.d.). Operating model design in tuna regional fishery management organizations: Current practice, issues and implications. *Fish and Fisheries*, *n/a*(n/a). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/faf.12480 doi: 10.1111/faf.12480

Spence, M. A., Blanchard, J. L., Rossberg, A. G., Heath, M. R., Heymans, J. J., Mackinson, S., … Blackwell, P. G. (2018). A general framework for combining ecosystem models. *Fish and Fisheries*, *19*(6).

Stewart, I. J., & Hicks, A. C. (2018). Interannual stability from ensemble modelling. *Canadian Journal of Fisheries and Aquatic Sciences*(12), 2109–2113.

Stewart, I. J., & Martell, S. J. D. (2015). Reconciling stock assessment paradigms to better inform fisheries management. *ICES Journal of Marine Science*, *72*(8), 2187—2196.

Tebaldi, C., & 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*(1857), 2053–2075.

Thorpe, R. B., Le Quesne, W. J. F., Luxford, F., Collie, J. S., & 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*(1), 49–58.

 doi:10.20944/preprints202008.0078.v1

Wellmann, J. F., Horowitz, F. G., Schill, E., & Regenauer-Lieb, K. (2010). Towards incorporating uncertainty of structural data in 3d geological inversion. *Tectonophysics*, *490*(3), 141–151.

Wright, J. H. (2009). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, *28*(2), 131–144.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3), 917–1003.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

QoIs	Stock status	Future fishing opportunities	Operating models
Univariate	x	x	x
Multivariate	x	x	x
Time series		x	x
Matrix		x	x
Full dynamics			x

Table 1: Relationship between QoIs and applications.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2020 doi:10.20944/preprints202008.0078.v1

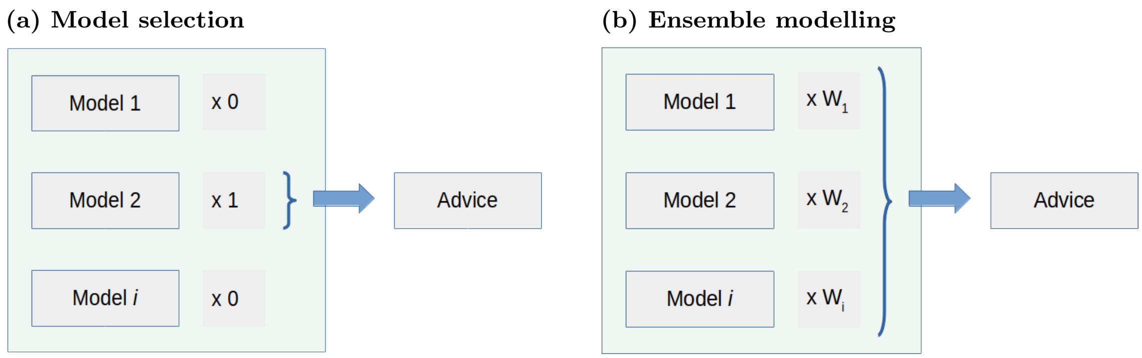

Figure 1: Simplified conceptual workflow comparison between conventional model selection (a) and ensemble modelling (b) in the context of stock assessment and advice provision. In the case of model selection (a), candidate models are analysed to find the 'best' (weight set to one) which is then used 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.