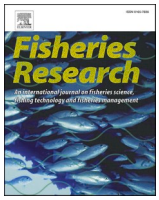
Fisheries Research 268 (2023) 106845

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



# Skill assessment of models relevant for the implementation of ecosystem-based fisheries management

Alexander Kempf [a,*], Michael A. Spence [b], Sigrid Lehuta [c], Vanessa Trijoulet [d], Valerio Bartolino [e], Maria Ching Villanueva [f], Sarah K. Gaichas [g]

[a] Thünen Institut für Seefischerei, Herwigstraße 31, 2750 Bremerhaven, Germany
[b] Centre for Environment, Fisheries and Aquaculture Science, Pakefield Road, Lowestoft, Suffolk NR33 0HT, UK
[c] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, INRAE, Institut Agro - Agrocampus Ouest, Nantes, France
[d] National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet, Building 201, 2800 Kgs. Lyngby, Denmark
[e] Department of Aquatic Resources, Swedish University of Agricultural Sciences, Turistgatan 5, 45330 Lysekil, Sweden
[f] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, INRAE, Institut Agro - Agrocampus Ouest, 29280, Nantes, France
[g] NOAA National Marine Fisheries Service, Northeast Fisheries Science Center, 166 Water Street, Woods Hole, MA, USA

ARTICLE INFO

Handled by A.E. Punt

*Keywords:*
Ecosystem based fisheries management
Skill assessment
Fisheries advice
Ecosystem models
Multispecies models

ABSTRACT

The advance of ecosystem-based fisheries management worldwide has made scientific advice on fisheries related questions more complex. However, despite the need to take interactions between fish stocks, and between stocks and their environment into account, multispecies and ecosystem models are still hardly used as a basis for fishery advice. Although reasons are numerous, the lack of high-level guidance for target-oriented skill assessments of such models contributes to the mistrust to use such models for advice. In this study, we propose a framework of guiding questions for a pragmatic and target-oriented skill assessment. The framework is relevant for all models irrespective of their complexity and approach. It starts with general questions on the advice purpose itself, the type of model(s) and data available for performance testing. After this, the credibility of the hindcasts are evaluated. A special emphasis is finally put on testing predictive skills. The skill assessment framework proposed provides a tool to evaluate a model's suitability for the purpose of providing specific advice and aims to avoid the bad practice of incomplete skill assessments. In the case of multiple models available, it can facilitate the evaluation or choosing of the best model(s) for a given advice product and intends to ensure a level playing field between models of different complexities. The suite of questions proposed is an important step to improve the quality of advice products for a successful implementation of ecosystem-based fisheries management.

## 1. Introduction

The questions asked to scientific bodies working in the field of fisheries management have become broader and more complex with the wish to implement ecosystem approaches to fisheries management (EAFM) and ecosystem-based fisheries management (EBFM) worldwide (Garcia et al., 2003; Link, 2010a; Pikitch et al., 2004). In this study we use the expression EBFM throughout. However, according to Dolan et al. (2016) EAFM focusing more on advice for single populations, EBFM focusing more on advice for communities as well as Ecosystem Based Management (EBM) including also other sectors than fishing form a hierarchical continuum and concepts are overlapping and often used interchangeably. Regardless of exact definitions, pure single species approaches are no longer sufficient as interactions among species and with their environment have to be considered.

As a result, the complexity of models needed to answer questions regarding the sustainable use of resources is increasing. Worldwide, a wide variety of such models, hereafter referred to as "ecosystem models" (EMs), exist (Plagányi, 2007) ranging from individual based models (e.g., Object-oriented Simulator of Marine ecOSystEms (OSMOSE; Shin and Cury, 2004)) over minimum realistic or intermediate complexity models (e.g., Stochastic Multi Species model (SMS; Lewy and Vinther, 2004), Globally applicable, Area Disaggregated, General Ecosystem Toolbox (Gadget; Begley, 2005; Plagányi et al., 2012; Collie et al., 2016; Trijoulet et al., 2020) and size spectrum models (e.g., Multi-species sIZE spectrum modelling in R (mizer; Blanchard et al., 2014), LEngth-based

* Corresponding author.
*E-mail address:* alexander.kempf@thuenen.de (A. Kempf).

https://doi.org/10.1016/j.fishres.2023.106845
Received 16 January 2023; Received in revised form 25 August 2023; Accepted 25 August 2023
Available online 30 August 2023

Multispecies Analysis by Numerical Simulation (LeMANS; Hall et al., 2006; Thorpe et al., 2015)) to end-to-end models (e.g., Ecopath with Ecosim (EwE; Christensen and Walters, 2004), Atlantis ecosystem model (Atlantis; Fulton et al., 2011)). These models serve different purposes. Models like EwE or Atlantis are mainly used to inform on strategic decisions based on a better understanding of processes within the ecosystem or are used as operating models in management strategy evaluations (MSEs; FAO, 2008), while models like SMS and Gadget can also be used to estimate historical stock status and make tactical decisions on e.g. next year's fishing opportunities (Plagány et al., 2012). The number of case studies where these models have been applied by scientists and managers is increasing steadily. However, when looking at a worldwide scale, environmental processes are hardly included in tactical fisheries management (Skern-Mauritzen et al., 2016; Karp et al., 2023) and single species approaches are in most cases still the preferred option for fish stock assessments but also e.g., as operating model in MSE simulations (e.g., ICES WKNSMSE, 2019; Thorpe and De Oliveira, 2019). Also, for more strategic decision support, the amount of examples where EMs have been implemented is still limited especially in Europe (Hyder et al., 2015; Lehuta et al., 2016).

The reasons for preferring single species approaches are numerous and span from a political avoidance of multispecies approaches because of the inherent trade-offs for decision making (Kempf et al., 2016), up to a lack of justification for increased complexity and appropriate supporting knowledge (Link et al., 2010b; Dickey-Collas et al., 2014) or inappropriate use (Rochet and Rice, 2009; Kraak et al., 2010; Planque, 2015). While issues related to policy integration, institutional adaptation and legitimacy of EBFM generally fall outside the areas of competence of natural science, we argue that a rigorous skill assessment and performance evaluation of models next to an efficient communication of uncertainties and results could represent important contributions to build trust among non-specialists (Pastoors et al., 2007; Lehuta et al., 2016).

A number of authors have already highlighted the importance of model skill assessments (MSA) and performance evaluation of models that could be used for advice on fisheries related questions (e.g., Link 2010b; Lehuta et al., 2013; Kaplan and Marshall, 2016; Olsen et al., 2016; Spence et al., 2021a). However, previous contributions often tend to focus on a specific model or certain aspects (e.g., sensitivity of parameters, hindcast or forecast, skill metrics or general model behavior). What is missing is a comprehensive framework (regardless of the complexity of the model) that would help model users and developers ask the right questions about the skill of their models(s) in a structured way. This may increase the trust in model-based studies for advice and may help to increase the uptake of results from more complex models (Karp et al., 2023). While a list of standard and generic diagnostics for single species assessments (e.g., residual patterns, retrospective patterns; Carvalho et al., 2021) or guidelines for single species MSE simulations (e.g., ICES WKGMSE, 2013; ICES WKGMSE2, 2019; ICES WKGMSE3, 2020) have emerged over time, the situation is much less simple for EMs. Indeed, their complexity and wide variety of approaches makes it difficult to automate tests and diagnostics. Objectives for EMs are also manifold with corresponding multiple outputs and scales difficult to be systematically scrutinized. A trade-off between rigor and flexibility is therefore needed. In addition, their currently still relatively limited impact on advice products creates no incentive for such laborious developments.

The International Council for the Exploration of the Sea (ICES) Working Group on Multispecies Assessment Methods (WGSAM) has developed the concept of so-called key-runs (e.g., ICES WGSAM, 2013; ICES WGSAM, 2015) for quality control of EMs and their parameterization. Output of key-runs contribute to specific aspects of the ICES advice. For example, key-runs with the multispecies model SMS are used to deliver natural mortality estimates for single species assessments of important North Sea and Baltic fish stocks (e.g., ICES HAWG, 2022; ICES WGBFAS, 2022; ICES WGNSSK, 2022). This is still one of the few examples worldwide where a multispecies model is used to contribute to tactical advice for fishing opportunities (Skern-Mauritzen et al., 2016). In recent years WGSAM developed a more structured way to reach conclusions on the suitability of a model to be used as key-run (ICES WGSAM, 2019).

Based on the experience gained from the WGSAM key-run approach and based on literature, a framework with guiding questions for targeted, pragmatic and flexible skill assessments has been developed to support a thorough review or benchmark before a model is used for advice. The aim was to develop a general framework applicable to models regardless of their complexity and approach as well as type of advice. The framework outlines the necessary steps to avoid the bad practice of incomplete skill assessments. For example, prediction skill is often hardly tested for EM's while the proposed framework includes this as an important step. We provide worked examples, but we do not propose best practice guidelines under each of the guiding questions given the large variety of potential advice questions and modeling approaches.

## 2. What is meant by the skill of a model?

In general, the skill of a model can be expressed as its ability to describe the true system state (Stow et al., 2009). However, because often the truth cannot be measured, the question is how well does a model reproduce the imperfect observations (Jolliff et al., 2009; Skogen et al., 2021). While this is relatively straightforward for less complex single species models, more parameters and time series of various types of observations have to be considered when evaluating the output from EMs.

We consider the true system state to be the modeled output plus a discrepancy (Kennedy and O'Hagan, 2001). A model is made up of "tuning parameters" and "input variables". "Tuning parameters", sometimes referred to as "free parameters" are adjusted to make the model look act like the real system, possibly fitting them to data, and "input variables" are variables that are taken as known inputs to the model (Brynjarsdóttir and O'Hagan, 2014; Spence et al., 2021b).

Strictly speaking "tuning parameters" do not have an interpretation outside of the model (Rougier and Beven, 2013), whereas "input variables" represent true variables outside of the model. What is a "tuning parameter" and an "input variable" is specific to the model. For example, in size-based models such as mizer, the predator–prey mass ratio is an "input variable," coming from other studies (e.g. Hatton et al., 2015), whereas in SMS it is treated as a "tuning parameter" and fitted to data (e.g. ICES WGSAM, 2021).

For stochastic models, we consider the random elements to be "tuning parameters" as in Spence and Blackwell (2016). Uncertainty in the "tuning parameters" is known as parameter uncertainty (or stochastic uncertainty for stochastic parameters). The discrepancy term corrects for errors in the model and uncertainty in this value is known as structural uncertainty. Therefore, the uncertainty of the prediction of the true system state depends on the uncertainties in both the parameters and the discrepancy. To learn about "tuning parameters" and discrepancies, one can use observations, however, it is often not possible to observe the exact value of interest, but a noisy, usually incomplete, version of it (e.g., Spence et al., 2018). This is known as observational uncertainty (Skogen et al., 2021).

The skill of the model is the inverse of the uncertainty of its prediction in Spence et al. (2021a). A model that has small uncertainty when predicting the true value of interest is more skilled than a model that has a higher uncertainty.

## 3. Guiding questions for a target-oriented skill assessment

### 3.1. General questions

A first set of questions deals with the general setting of the model

environment and the advice questions asked to the model (Fig. 1). Indeed, EMs produce a large variety of outputs at multiple scales and it might not be necessary for the model to display a good fit in all those dimensions to be used operationally. According to Sterman (1984) and NRC (2007), the validity of a model is somehow subjective and should rather be seen as its adequacy for its purpose, which can be more objectively assessed. Similarly, in this study, we consider a model as valid if it fits its purpose. Model validation is seen as the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended use of the model (AIAA, 1998). In a decision support context, we therefore advocate a target-oriented (i.e. focusing on the output of interest) skill assessment, which requires the following questions to be answered as a first step.

#### 3.1.1. What do we need?

The first question to be answered is what is needed for a successful advice delivery. This question is specifically relevant for EMs and delayed the development of MSA compared to other fields. For instance, in oceanographic or atmospheric modeling where the products are gridded maps of ocean or air variables, inter-comparison exercises and forecast requirements have led to the development of common standards for model evaluations (Hernandez et al., 2009). However, their expectations and outputs are often similar while the diversity of assumptions, structures, outputs and objectives in EMs makes standard protocols difficult to define. MSA for EMs includes justifying the need for more complex models than traditional single species models. To find the "sweet spot" between sufficient complexity (i.e. to avoid bias because important processes are not considered) and acceptable levels of model uncertainty is crucial (Collie et al., 2016). In this context often Models of Intermediate Complexity for Ecosystem assessments (MICE) are discussed as being preferable over more complex whole ecosystem models (Plagányi et al., 2012). However, the right level of complexity is also largely determined by the questions asked to the model and it has to be determined whether the model can produce the type of outputs and metrics needed to assess the objectives of the advice given the data at hand. One may need to give up complexity to robustness (i.e. stability in the outputs of interest) when providing advice. It is important to think about the relevant scales for each model output. It has to be clear early in the process whether the model can produce results at the right spatial (e.g., whole ecosystem, subareas, Marine Protected Areas) and temporal scale (e.g., inter-annual variability vs. long-term trends) including the question whether the focus is on past or current or future states. However, also other scales may be important. For example, there is a difference when advice is needed on a species or specific life stage level, for functional groups, trophic levels or biodiversity aspects in a whole ecosystem context. Depending on this, the MSA and the choice of the model need to be tailored towards the scales relevant for the advice.

The type of advice needed also determines which output from the model is highly relevant and which output is not as important. This allows focusing on specific outputs. For example, if the advice needed is on historical predation mortalities, a focus on diet related data and submodels is important as well as the robustness of the predation mortality estimates. In contrast, forecast skills are not relevant in this case. Also processes leading e.g., to a scaling of the total consumption of predators and because of this to a similar scaling of the prey abundance estimates may also be of less importance as long as the predation mortality estimate (influenced by predator consumption and prey abundance) is hardly impacted. The situation, however, is completely different if the same EMs are to be used to set total allowable catches (TACs) based on a target fishing mortality where the absolute level of abundance and catches are of greatest importance. Similarly, if the model is used as the operating model in an MSE, efforts should be put on assessing the realism of objective-related outputs and to whether they can be used in absolute or relative terms. It is recognized that these questions, although crucial, could be answered differently by individuals and the availability of an expert group to objectively judge on these aspects may be needed.

#### 3.1.2. What type of model is available?

Another question relates to the type of model itself, as the MSA process will fundamentally depend on this. Although the questions in this framework are universal, the type of analyses that can be carried out under each question depends on the type of model. For example, pure simulation models without an associated data fitting process do not work with "tuning parameters". Only a comparison with external time series of e.g., catch, survey indices or assessments can allow the possibility of testing whether the set of external input parameters results in a reasonable behavior of the model. As a complement, sensitivity tests can provide beneficial information on the impact of certain parameters on results and advice products (see below).

#### 3.1.3. Are there sufficient data available to test the model performance?

Before conducting a skill assessment, appropriate data for comparisons with model outputs have to be selected, and their availability and quality need to be evaluated. The availability and scales (e.g., temporal, spatial, age/length structure or whole population) of the observational data will ultimately determine at which level of detail a skill assessment can be carried out and how quantitative it is going to be.

Next to the availability, the uncertainty in the observations is also an important aspect to consider before any comparison is made. The more uncertain the observations, the more distance from modeling results may be allowed before a model is regarded as being unreliable. Observation quality and model expected accuracy constrain the type of methods to be used to assess model skill. For instance, a point to point evaluation might not be relevant to compare spatial model predictions to spatial observations if the localisation of observation is not precise or the model spatial resolution is too coarse. However, there could be situations where questions regarding the suitability of model outputs for advice products can no longer be answered in a quantitative way because of highly uncertain and/or misplaced observations. Ultimately, managers and other end-users of model advice should give their opinion on their comfort level with model performance in a data-poor situation by comparing with alternative methods for decision-making (e.g., can an uncertain model provide support for making decisions as opposed to deciding based on the poor data alone, or on no data or model at all?). Skill assessment provides measures to have these conversations in an informed and transparent manner, although in this study we focus on cases with sufficient data available (for examples see table S1) rather than data-poor situations.

### 3.2. Skill assessment of hindcasts

After successfully answering the general questions described above, the next step is to evaluate the calibration to historical data and dependent on the type of model its parameterisation with external information and/or internal parameter estimates (Fig. 1).

#### 3.2.1. What are the key parameters that influence model results most?

After, or even before, a first parameterization with "input variables" and/or estimation of "tuning parameters" has been finalized, sensitivity tests need to be carried out to get insights on the general model behavior and model formulation (i.e. structural uncertainty, uncertainty in the model discrepancy) and to learn what parameters require specific attention or refined estimation.

Although a local one-at-a-time (OAT) sensitivity analysis is a commonly used approach (Ferretti et al., 2016), it does not take into account interactions and associated non-linearities in the model response and correlations among the input parameters (Saltelli and Annoni, 2010). As an alternative, global sensitivity analysis (GSA) is increasingly used to develop and assess environmental models (Lehuta et al., 2010; Marzloff et al., 2013; Morris et al., 2014) and tool-boxes have been published (e.g., Pianosi et al., 2015). Although GSA comes at the cost of high computational time needed for complex models with many parameters, the access to high performance and parallel

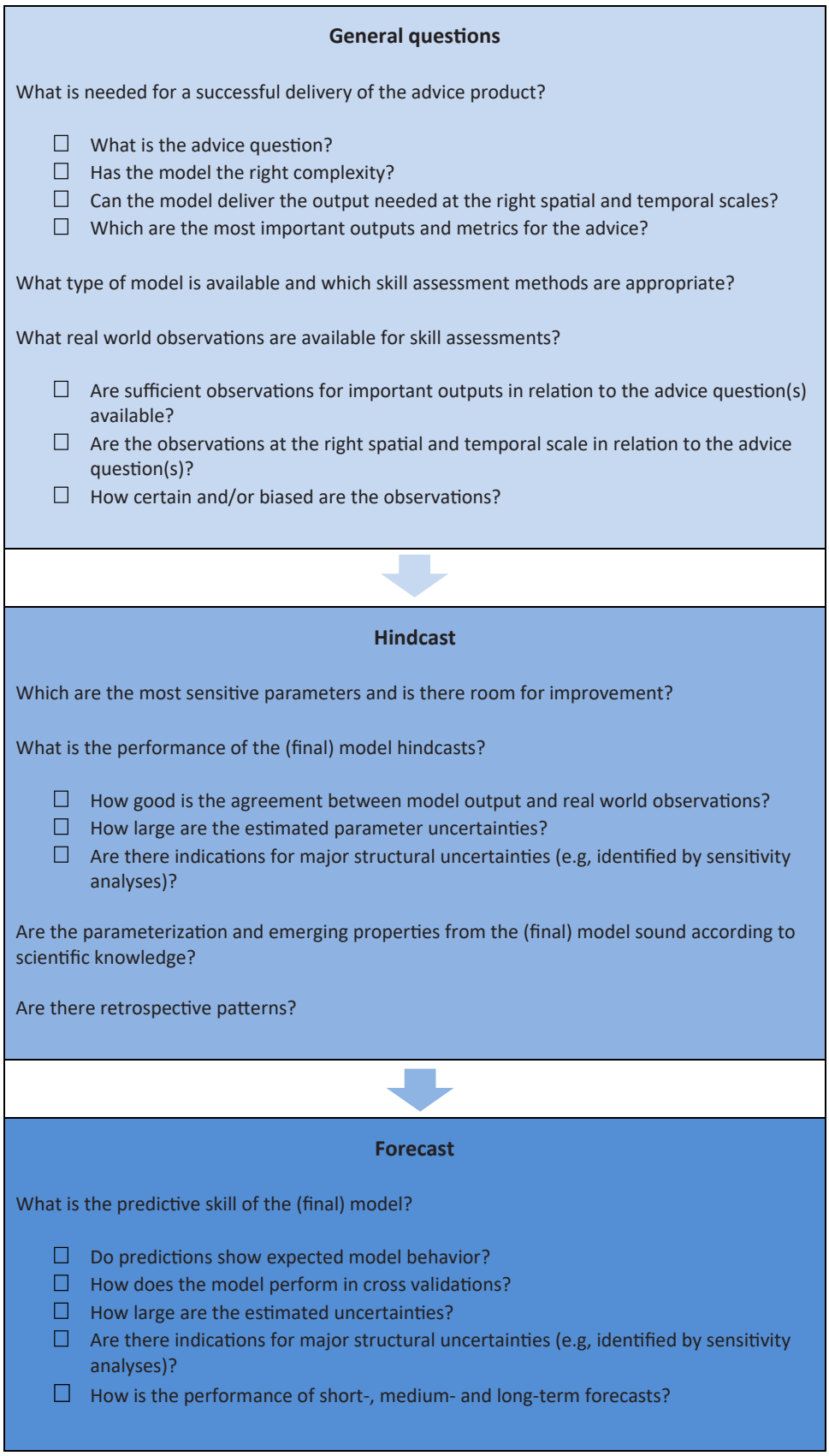



**Fig. 1.** Framework of questions proposed for a target-oriented skill assessment.

computing is opening the door for more GSA. In general, there is a need to decide on which approach is suitable on a case by case basis although one needs to be aware of correlations between parameters and a realistic uncertainty margin around parameters from outside the model in any case. An example, for an intermediate approach, Hansen et al. (2019) performed a sensitivity analysis on nine biomass dominating functional groups in an Atlantis ecosystem model for the Barents Sea to determine whether the species position in the food web influenced their sensitivity to certain parameter perturbations and whether responses (including non-additive responses) depend on the trophic level. Also, the so-called Morris method combines the idea of a global sensitivity analysis with a one step at a time approach in an efficient way and was successfully applied to an Atlantis model (Morris, 1991; Bracis et al., 2020).

In the particular case of estimation models (calibrated/fitted models, including tuned parameters), GSA is recommended both prior and after the calibration phase. Prior to calibration or tuning of the model, sensitivity analysis is advised both as informative of model behavior and as a means to prioritize factors to further investigate or include in calibration (Saltelli et al., 2006). In this latter case, the output of interest might directly be the objective function or likelihood measure of the estimation procedure and the SA will indicate which parameters are the most influential on its value. This might help restrict the estimation to the most influential parameters, which is particularly helpful when all the uncertain factors cannot be included in the estimation procedure, when over-parameterisation is suspected or alternative model structures or assumptions exist. SA can also point out the combinations that are mostly responsible for model realizations in the acceptable range and on the contrary identify the "worst" locations in the parameter space, therefore objectivising modelers choice. Depending on the method used, it may also unveil correlation structures among parameters, allowing it to be used in post-calibration SA or uncertainty analyses (see an example of a bayesian procedure in Da Veiga et al., 2021). After sensitivity analyses have been carried out, e.g., Lehuta et al. (2013) propose a classification of parameters into a hierarchy according to their sensitivity and the nature of their uncertainty (Fig. 2). Based on this hierarchy, it can be decided how parameters may be treated for further improvement of model skills and uncertainty estimation.

After the calibration or tuning phase, sensitivity analysis helps to assess the robustness (i.e. stability in the outputs of interest) of the parameterisation. A difficulty arises here, however, because of the correlation structure introduced between parameters by the calibration phase. If it has been revealed by the previous SA or the calibration procedure itself, it can be propagated in the SA using appropriate methods (Da Veiga et al., 2021). Otherwise, best practices imply that the model be refitted for each new combination of parameters tested (Turányi, 1990; Saltelli and Annoni, 2010; Lehuta et al., 2013). Recognising the potentially high computational cost of such a procedure, methods are emerging to assess the sensitivity of the internally estimated parameters to other parameters of the model (Jørgensen, 2023). Nonetheless, we recognise that these are novel, sophisticated methods possibly appearing cumbersome or even inapplicable to inexperienced SA practitioners. We recommend that an uncertainty analysis is at least performed, yet acknowledging its limits, because even imperfect, it is a step towards improving model transparency. Performed around parameter's estimates, uncertainty analysis will inform on the robustness of the model and the fit at least locally.

### 3.2.2. *How is the performance of hindcasts?*

Preferably, for all alternative parameter values and/or model formulations tested in sensitivity analyses, model outputs have to be compared to observations. Thereby, the focus has to be on the outputs, scales and skills relevant for the advice question(s). In general, different metrics might disagree with each other (e.g., correlation indices and Average Absolute Error (AAE)) and therefore the selected metrics and indicators need to reflect the skills that are critical for the purpose of the modeling exercise (e.g. trends in climate change issues, but point by point accuracy in tactical models). It is important to assess on which important aspects (e.g., predation mortalities or biomass trends or biodiversity), resolution (e.g., species level or functional groups or whole ecosystem), spatial (e.g., whole model area or for a sub-part of the model area only (e.g., areas around wind farms)) and temporal scales (e.g. accurate on yearly catches, but bias on seasonal catches) the model is reliable and on which it is less certain.

There is a large variety of methods available to compare model outputs/fits to historical data. These can be roughly divided into methods focusing on univariate statistics and methods focusing on a multivariate approach. A particular challenge is the skill assessment of spatially explicit models (see below under 3.2.2.3). Finally, for statistical models, testing the estimation skill of the model via simulation (data is simulated rather than observed) can be useful. These four aspects are further detailed below.

*3.2.2.1. Focus on univariate statistics.* A wide variety of metrics (e.g., AAE, Root Mean Squared Error (RMSE) or residual analysis) exist that can provide objective insights into the match between model output and observations (Stow, 2009). As example, Olsen et al. (2016) used a set of such metrics (Fig. 3) that allow for the identification of different types of discrepancies between model output and observations. It is best to decide beforehand the thresholds for each metric to indicate and discern between good or bad model performance (e.g., Modeling Efficiency (MEF) $> 0.3$ or correlations $> 0.6$, with MEF evaluating whether the model is an improvement over the mean of the observations (Stow et al.,

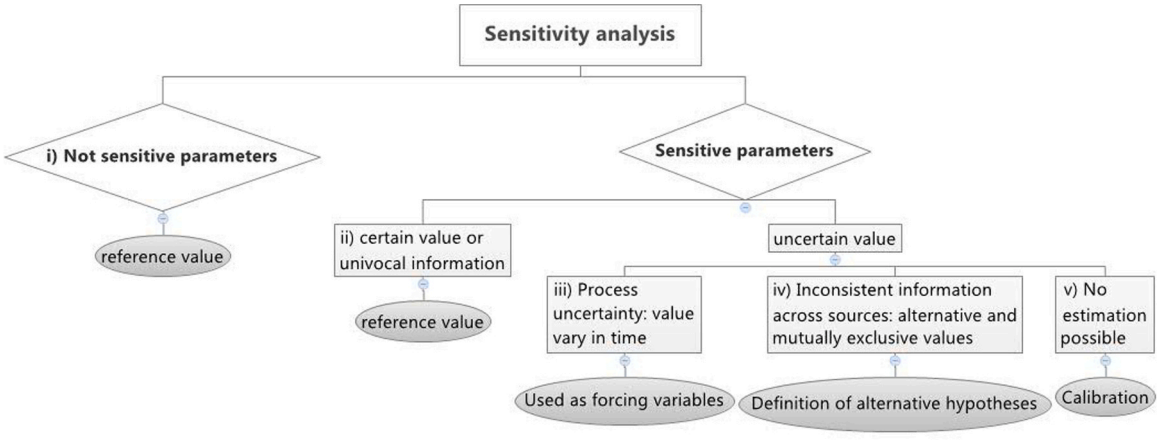



**Fig. 2.** Hierarchy of parameter sensitivity and uncertainty as well as suggested action.
Reprinted from Lehuta et al. (2013) with permission from Elsevier as copyright holder.

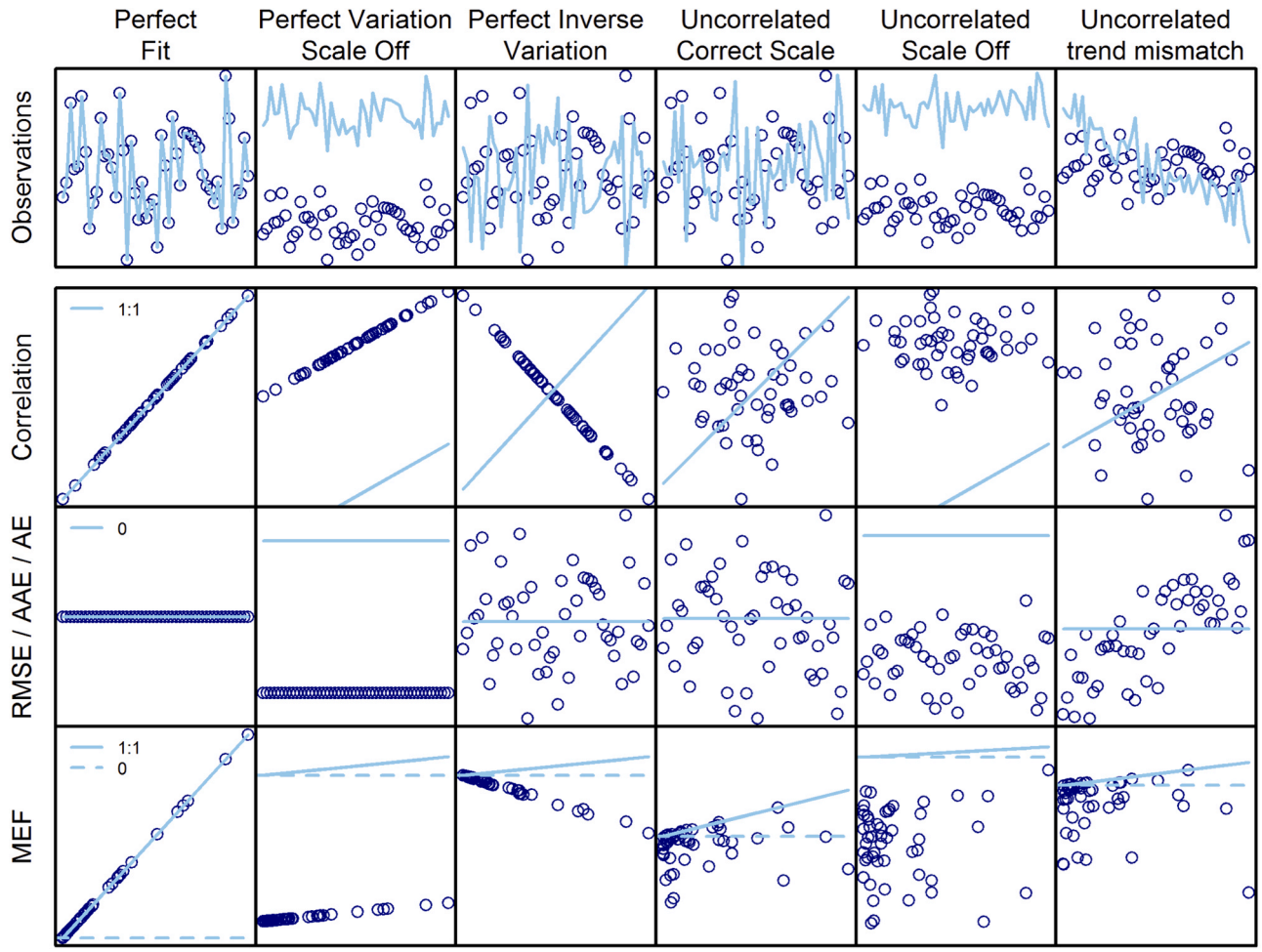



**Fig. 3.** Conceptual Figure comparing the skill metrics (Root Mean Squared Error (RMSE), Average Absolute Error (AAE), Average Error (AE), Modeling Efficiency (MEF) and correlation) performance using simulated observed (open circles) and modeled data (lines). Reprinted from Olsen et al. (2016).

2009). MEF is as good as the mean if MEF= 0. The larger the MEF, the better is the model performance). While some metrics come with theoretical thresholds (Dabrowski et al., 2014), scientific advisory bodies like ICES often still need to come up with general guidelines on appropriate thresholds (e.g., Annex 3 in ICES Benchmark Guidelines, 2023 provides thresholds for Mohn's rho (Mohn, 1999) as measure for retrospective bias in assessments or guidelines for RMSE runs tests (Carvalho et al., 2021)) that are accepted by their scientific community. Another aspect not often tackled by current MSA metrics is the account of uncertainties around observations in the computation of skill metrics. Ignoring uncertainties around observations can lead to an underestimation of model skill. Allen et al. (2010) enhance the halo of uncertainty surrounding both model and data.

*3.2.2.2. Focus on a multivariate approach (both variables and metrics).* More generally, for models with multiple response variables, independent, univariate comparisons of each response with its corresponding observations is informative, but it is also appropriate to compare responses and observations across all of the response variables simultaneously and check known relationships between variables to ensure structural realism (Friedrichs et al., 2006, 2007). A cost function or multivariate goodness of fit can be considered as single metrics of overall model performance (e.g., Kasibhatla et al., 2000; Stow et al., 2009).

Where real observations are scarce, not present, or too uncertain and/or misplaced in time and space, at least general patterns and overall model behavior can be tested based on consensus in the scientific community (e.g., Kaplan and Marshall, 2016). Perturbation analyses may also be used to check or compare model response with expected or observed past behavior (Maar et al., 2018).

*3.2.2.3. Spatially explicit models.* A particular challenge is the skill assessment of spatially explicit models. Atmospheric and oceanic models are well advanced in the validation of spatial outputs. Several approaches are for instance proposed by the Global Ocean Data Assimilation Experiment (GODAE; Hernandez et al., 2009; Ryan et al., 2015). They combine visual comparison of two- and three-dimensional fields (maps, possibly interpolated) of variables (e.g. mixed layer depth, temperature; class 1 metrics), confrontation of fine resolution model outputs with in-situ or remote-sensing observations over well observed sectors of interest (transects or moorings; class 2 metrics) and integrated quantities over or across special sections of model domain (e.g. heat transports; class3 metrics). Classical statistics of observation-model differences can be computed. Because visual comparison of maps is often subjective, area-integrated quantities may also be computed and compared (e.g., Schoener overlap index of similarity; Schoener, 1970) as for e.g., done by Püts et al. (2020) to judge on similarities between predicted and observed spatial distributions of various species. Also mapping of cell-by-cell misfits, Kappa tests (allowing or not for consideration of neighboring cells), or self-organizing maps (SOM) are further examples of methods to derive performance metrics (Allen et al., 2007; Stow et al., 2009; Mitchell et al., 2021). What EMs are concerned, observations are rarely available at a fine resolution and are often scarcer and more unevenly distributed than oceanographic data. The

metrics used to assess the ability of the model to mimic spatial patterns need to be adapted in consequence (Sandvik et al., 2016). What needs to be finally answered is whether the spatial resolution possible for a skill assessment is sufficient for the advice question in place. In addition, for any model feature but maybe even more for spatial patterns, one must be aware that spatial distributions in EMs are seldom fully emerging from mechanistic processes (but see for instance the Apex Predators ECOSystem Model (APECOSM; Maury, 2010) or OSMOSE (Shin and Cury, 2004)). An in-depth understanding of the level to which spatial patterns are forced or constrained by data is required and must be transparently communicated before a spatial MSA is performed. Also, circularity (especially when data-poor) needs to be avoided as much as possible to make sure tuning and test sets are different. Last but not least, any spatial MSA has to be conducted in conjunction with MSAs for the temporal and spatio-temporal dimensions. The best spatial representation of observations may not lead to the overall best ecosystem model performance when considering all dimensions and outputs (Püts et al., 2020).

*3.2.2.4. Bayesian model validation.* Bayesian model validation (Gelman et al., 2013) is another method to validate the model. This method, similar to hypothesis testing, tests if the model is able to recreate a dataset. The model uncertainties are quantified after it has been fitted to data, and then many pseudo-datasets, the posterior predictive distribution of the data, are generated. From these pseudo-datasets, summary statistics of the data are generated (e.g. mean, auto-correlation) creating a distribution of the summary statistics. The summary statistics are not limited to those that are sufficient statistics of the likelihood, in fact it is encouraged that they are not (Gelman et al., 2013). The summary statistics of the observations are compared with the distribution of the summary statistics of the pseudo data. If the observed values are in the tails of the distribution then one can conclude that these summary statistics are inconsistent with the model (see Spence et al., 2021c for an example). Unlike classical hypothesis testing, where the aim is to show that a null model is not correct, Bayesian model validation aims to answer if the summaries of the data could have arisen by chance under the model assumptions. This has the advantage that a "good model" is not rejected due to a lack of data, but only if there is evidence that it is not a "good model".

*3.2.2.5. Simulation testing for statistical models.* The estimation skill of statistical models can be tested via simulations. In this case, a simulation model, often called operating model (OM), is used to simulate a set of observations (including observation errors and potentially process errors) that the statistical estimation model (SEM) will fit to. Two cases can be differentiated: i) the SEM is the same as the OM; and ii) the SEM differs from the OM. In the first case, the simulations test the SEM consistency to estimate the truth when the model is correct (same structure and assumptions as the OM). This test is notably useful to understand the SEM behavior for different levels of errors around the observation and/or processes. In the second case, the SEM is not the full representation of the truth (OM), and often a simplification of it. The information provided by this test will depend on the difference between OM and SEM. This test is notably useful to understand, for instance, how the estimation skill of the SEM can be affected by different levels of data (Trijoulet et al., 2019), by alternative model structures or by different model complexity (e.g., trophic interactions and process errors in Trijoulet et al., 2020). When the OM is very complex (e.g., end-to-end models such as Atlantis), fitting a simpler SEM (e.g., single species or MICE models) could provide useful information on how the stock or fishery perceptions can be affected by model structure and assumptions and which model outputs are likely robust to structural uncertainty (ICES WGSAM, 2023). In the case where it is believed that the OM is a close representation of the real ecosystem, this simulation notably informs on the possible limitations that exist with worldwide use of single species assessment models.

*3.2.3. Are the parameterization and emerging properties from the model sound according to scientific knowledge?*

Based on knowledge on e.g., life history traits and ecosystem functioning, the "input variables", derived "tuning parameters" and outputs have to be checked for plausibility and consistency next to the comparison to observations described above. Cury et al. (2008) stress that many models can reproduce a historical time series but it takes structural realism to fit simultaneously several data components and advocates for "pattern-oriented modeling" (POM, Grimm et al., 2005). For example, prebalance (PREBAL) diagnostics (Link, 2010c) have been developed for ecological network models like Ecopath to increase the credibility of such models. Based on the PREBAL diagnostics, it can be evaluated whether the parameterization, estimates of parameters and emergent properties are in line with theoretical understanding of ecosystem structuring and functioning. While PREBAL has been developed for models like Ecopath, relevant similar diagnostics (e.g., production and consumption rates, biomass ratios, slope of removals or biomass across taxa and trophic levels) may be also used for any other type of model to check the general plausibility of its parameterization and emergent properties. In addition, intermediate outputs and estimated tuning parameters from the model that are not provided as standard outputs for decision making may be checked for realism (De Mora et al., 2016; Kaplan and Marshall, 2016). For example, it may be analyzed whether natural mortality rates decrease with age as expected or predicted age and length distributions are in line with observations.

*3.2.4. What is the influence of additional data points?*

Another important performance test that is common for e.g., single species assessment models is a retrospective analysis. It needs to be ensured that important model outputs do not change substantially just by adding or leaving one year of data out. So far this type of analysis is not always standard especially for more complex models, but it is an important addition to avoid especially retrospective bias (i.e. management decisions change if additional data points were available). One example for carrying out such a retrospective analysis is from WGSAM when testing the performance of the 2017 and 2020 SMS key-runs in estimating natural mortalities (ICES WGSAM, 2017; ICES WGSAM, 2021).

### 3.3. Skill assessment of forecasts

The predictive ability is often the most important model feature for e.g., advice on fisheries management strategies. Improvement of predictive capacities is one of the main reasons for increasing model complexity and moving towards more process-oriented models (Spence et al., 2021b). While hindcast performance gives an overview on the success of calibrating or fitting the model to observations, it does not allow conclusions on the forecast performance of the model. For example, if the model fits the historical data used to calibrate or fit the model well, it can still show poor performance in forecasting if the model e.g., over-fits the observations or processes like recruitment to the fish stock become more important in forecasts than in hindcasts and are not well understood. Trijoulet et al. (2020) found that single species models can fit observations well, but produce biased estimates of spawning stock biomass and recruitment relative to correctly specified multispecies models in a simulation analysis. In the absence of simulated "truth" or independent data available to perform cross validations, hindcast fit to data is often the primary model selection benchmark. In such situations often the well-fitting simpler model may be selected for advice, although an analysis of forecast skills could point in another direction. Performance of hindcast and forecast may also differ across model outputs and examining performance for the most important outputs to the question at hand is critical. For example, hindcast performance of an end-to-end model was variable across ecosystem metrics

that were not used in calibration, but forecast skill was better than hindcast skill for several of these metrics (Olsen et al., 2016; Fig. 4). In addition, there are several examples where relationships e.g., between environmental data and recruitment of fish stocks broke down although they were highly significant when fitted to historical data (e.g., Myers, 1998; Howell et al., 2013). Therefore, it is most important that the forecast skills are also investigated to ensure a common understanding of model capabilities for decision making (Fig. 1).

### 3.3.1. *Do predictions show expected model behavior?*

The performance test of forecasts is more difficult than for hindcasts because future observations are simply not available and historical observations are often needed for model calibration and it is not always easy to leave out enough data for meaningful cross validations. However, general questions on model behavior can be used to challenge model forecasts in any case. For Ecopath with Ecosim key-runs WGSAM tested whether the fishing mortalities and fishing effort leading to maximum sustainable yield are in a sensible order of magnitude and not considerably above fishing levels where it is known that the stocks seriously declined in the past (ICES WGSAM, 2015, 2016). This would indicate that the productivity of stocks could be overestimated and at least further investigations (e.g., is future productivity positively influenced by climate change) are needed. Another example of an expected model behavior is that in a simulation with no stochasticity, constant oceanographic forcing, and no fishing, the majority of species or functional groups should show no significant trend in biomass over the final 20 years of a long-term forecast (Kaplan and Marshall, 2016).

### 3.3.2. *How is the short-term, medium-term and long-term forecast performance?*

A common way of testing the forecast performance is to calibrate the model to a training data set and then compare model predictions to independent data points (e.g. years) not used for the calibration (cross validation). For example, one option is to leave out input data such as survey indices for some of the years and analyze if this induces bias in the survey and population predictions (Trijoulet et al., 2020). Univariate and multivariate skill metrics described for the hindcast performance testing can be used also for this purpose.

In addition, Mean Absolute Scaled Error (MASE; Hyndman and Köhler, 2006) is a form of cross validation that has become a popular metric for testing prediction skills. It is a measure for determining the effectiveness of forecasts by comparing the predictions with the output of a naïve forecasting approach. MASE has the desirable properties of scale invariance, so it can compare forecasts across data sets with different scales. Kell et al. (2021) and Carvalho et al. (2021) used it in an "hindcasting" approach (observations are peeled back from the terminal year) of assessment model outputs. An observation at time t was compared to a prediction of that observation made x time steps

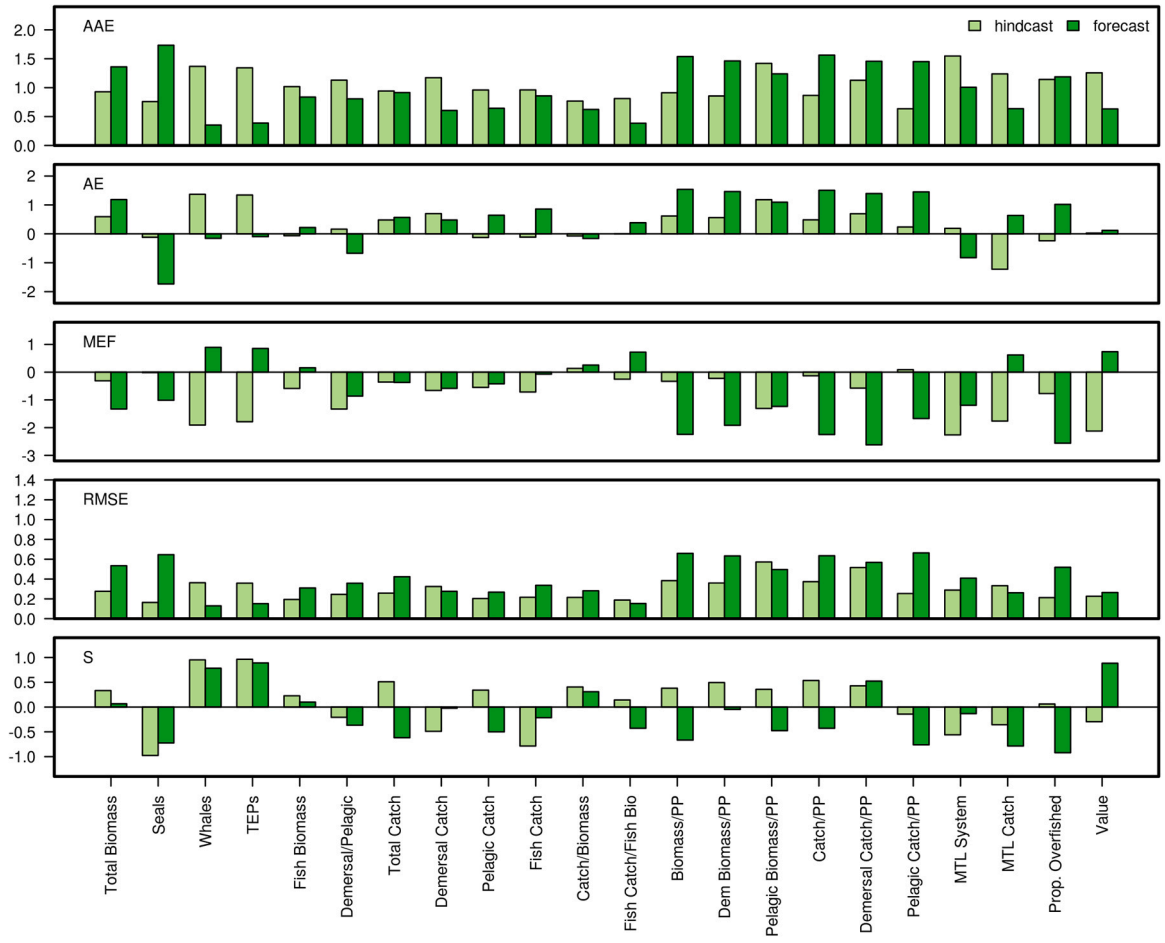



**Fig. 4.** Skill metrics for ecosystem indicators emulating emergent ecosystem properties using the Northeast US (NEUS) Atlantis model, demonstrating the importance of identifying and evaluating outputs of interest for skill assessment. Performance in hindcast and forecast differed by metric and was not necessarily consistent (e.g. opposite MEF indicating poor hindcast but good forecast performance for Whales, TEPs, MTL catch, and Value). Pairwise comparison of forecast and hindcast skill metric performance for 5 skill metrics: Average Absolute Error (AAE), Average Error (AE), Modeling Efficiency (MEF), Root Mean Squared Error (RMSE) and Spearman rank correlation (S) between observed and predicted values. Note that these indicators were not used in model calibration. TEPs: Threatened, endangered and protected species. PP: Primary production. MTL: Mean trophic level.
Reprinted from Olsen et al. (2016).

previously.

An issue with cross validation is that data are required to be omitted when fitting the model. One could recalibrate the model with all of the data after the skill assessment, however this will change the model and its `skill'. Therefore, the new model, with updated parameter estimates, has different `skill', which is unknown. An alternative is to leave the testing data out completely (e.g., several years of a time series). This has the disadvantage that the model is not utilizing all of the information when making predictions and therefore the estimates are more uncertain (and potentially biased if e.g., explaining variables become significant only if all data are used) than they could be. This requires careful thinking before performing cross validations.

It is of utmost importance to test the predictive skills at the scale required by the advice and where possible to extend the evaluation across the short-, medium- and long-term to better understand the limits of the model applicability in the context of advice.

### 3.4. Examples for applying the framework of guiding questions

In order to demonstrate that the described framework of guiding questions can be applied to a wide variety of advice questions and models with different complexity, we applied the framework to five examples (Table S1). The models tested ranged from a single species assessment model (State-space Assessment Model (SAM); Nielsen et al., 2021) for Western Baltic Spring Spawning Herring (WBSS) over MICE models (Georges Bank Hydra (Gaichas et al., 2017), Baltic Sea SMS (ICES WGSAM, 2019)) to ecosystem models (Georges Bank Rpath (R implementation of EwE; Lucey et al., 2020), Irish Sea EwE (Bentley et al., 2021)). The advice questions ranged from tactical advice on fishing opportunities over providing natural mortality rates to inform single species assessments to the use of ecosystem information as guidance on setting target fishing mortalities within single species $F_{MSY}$ ranges (Rindorf et al., 2016; Bentley et al., 2021). The examples are based on key-run evaluations from ICES WGSAM (ICES WGSAM, 2019; ICES WGSAM, 2023) and the ICES assessment for Western Baltic Spring Spawning Herring (ICES HAWG, 2023).

As envisaged, the proposed framework could be applied to any of the advice questions and models. The questions helped with scoping the needs for a successful advice delivery and conducting a structured target-oriented skill assessment. For Hydra and Rpath the framework was applied although the models were not yet fully developed. The questions helped here to plan future steps. This illustrates the iterative nature of the process as highlighted by Schmolke et al. (2010) in their recommendations on model documentation for advice.

While hindcast performance was tested in all examples, predictive skills were not fully tested because forecasts were either unnecessary to answer the advice question (Baltic Sea SMS, Irish Sea EwE), or the models were not fully developed (Georges Bank Hydra and Rpath) or mainly qualitative information on predictive skills were available (WBSS SAM). Our proposed framework will hopefully help to make the assessment of predictive skills a standard as already practiced in e.g., Olsen et al. (2016) or Carvalho et al. (2021).

### 3.5. Final judgment on model skills and limitations

After the questions proposed in this framework have been answered, the information has to be combined into a final recommendation regarding whether and how a certain modeling approach can be used as a basis for advice. A particular challenge is to combine information from different analyses and sources to get a complete picture on the strength and weakness of a modeling approach.

#### 3.5.1. Summary of information

Summary diagrams, such as Taylor diagrams (Taylor, 2001) can be useful in this respect. As example, Kell et al. (2016b) used Taylor diagrams to evaluate prediction skill for a variety of models conditioned on a wide range of scenarios based on different datasets and alternative model structures. Taylor diagrams are able to simultaneously present multiple skills over multiple variables and modeling options (find another example in Püts et al., 2020 comparing different parameterizations of an Ecospace model). They also make explicit the correlations between skill metrics, thus warning about redundancy in the assessment (Jolliff et al., 2009).

Radar plots are another way to summarize results. Taylor diagrams are mainly used for validation and selection across multiple models and scenarios, while Radar plots are often used to summarize how well multiple conflicting objects are met. However, Radar plots can be also used to summarize model skills on several variables as shown by Lehuta et al. (2013) or Vigier et al. (2018).

Both examples have in common that they can become quite complex and a good explanation/ training for managers/ end-users is needed to ensure that results are interpreted in the right way. More generally, a complete documentation on model development including the skill assessment steps is desirable for any model intended to support decision making (Schmolke et al., 2010).

In order to conclude in an objective way whether a model is fit for purpose, it needs to be ensured that the choices of outputs, metrics and thresholds are objective, transparent and relevant to the advice question (s). The uncertainty in relevant model output needs also to be communicated. The importance of using information from sensitivity analysis has to be highlighted again as often structural uncertainties are more important than parameter uncertainties, especially in the context of ecosystem modeling (Hill et al., 2007).

Finally, based on all information available a decision has to be made regarding which model output can be used to provide advice and which cannot. For example, WGSAM concluded for an Ecopath with Ecosim model for the Irish Sea that the model key-run was accepted as a basis for generating ecosystem indicator(s) but the direct use of modeled fishing mortality values was not recommended (ICES WGSAM, 2019).

All information leading to final conclusions and recommendations must be presented in a concise report. The set of questions provided can serve as a basis to present the evaluation results in a structured way (see e.g., ICES WGSAM, 2019). Because such reports can become quite extensive, e.g., model report cards could be added for stakeholders less familiar with technical details. Next to this, the results of the model and the results of the MSA have to be reproducible to allow especially other scientists to follow the decisions afterwards. This can be achieved by storing all information needed on repositories.

#### 3.5.2. Interaction with stakeholders

The interaction with stakeholders is a challenging task that, however, often decides whether a modeling approach is finally used for a certain advice product. We provide examples in table S1 on how outputs from ICES WGSAM key-runs have been perceived and used by stakeholders. The examples also show that work does not stop with a successful skill assessment and further steps (e.g., stakeholder workshops, benchmarks) are needed before model results are used as a basis for advice. Already during the skill assessment phase input from stakeholders is highly beneficial to build trust.

To make complex model results accessible to other scientists and stakeholders it is most important to think about the dissemination of products (Pastoors et al., 2007; Lehuta et al., 2016). Poor documentation and lack of accessibility to input data, model source codes and model outputs are bad practices in the communication of any quantitative advice and become particularly important when working at the forefront of new formats of advice. The best skill assessment will not help to improve the uptake if models and their results are not easily accessible. This may also include summaries at different levels of technical detail because not all stakeholders want to read a full report. Also, the communication between scientists with different backgrounds is important and e.g., people from WGSAM often participate in assessment working groups. This substantially improves the communication

between scientists (i.e. traditional stock assessment scientists and ecosystem modelers), as well as the dissemination and explanation of results.

#### 3.5.3. Communication of uncertainties

The communication of uncertainties is an integral part of the scientific advice and supporting the decision process for a correct use of information on uncertainties is especially challenging (Kell et al., 2016a; Levontin et al., 2017). The delivery of information in an easy and understandable way especially for stakeholders without a scientific background is difficult. Misinterpretation of uncertainties as "errors'' or turning towards models with lower parameter uncertainties, just because structural uncertainties are ignored (Bannister et al., 2021), needs to be avoided. It is often seen as problematic that, for example, fishing opportunities become lower the higher the number of sources of uncertainty incorporated - a phenomenon seen in MSE simulations (ICES WKNSMSE, 2019). To find a compromise between the number of sources of uncertainties tested (e.g., number of sensitivity runs) and delivering a useful advice product, especially from complex models, is challenging and needs to be found case by case. What is important is a level playing field when different models (of different complexity) are tested as candidates. Otherwise, there is a danger of uncertainties being underestimated and that the wrong model is chosen just because its weaknesses are not well understood or not presented. The framework of questions presented in this study can be an important step towards such a level playing field contributing to improve the quality of advice products for a successful implementation of EBFM.

## 4. Conclusions

In this work we propose a framework for target-oriented skill assessment of models to advance the implementation of EBFM. Several guidelines and best practices for model development and performance testing already exist (e.g., Link et al., 2012; Heymans et al., 2016; Kaplan and Marshall, 2016; ICES WKGMSE2, 2019, Carvalho et al., 2021). However, often such complex and detailed guidelines tend to be ignored (Schmolke et. al, 2010). A general framework for pragmatic and targeted skill assessments applicable to models irrelevant of their complexity and approach is so far missing. The framework of guiding questions presented here contributes to fill this gap and allows for systematically testing and benchmarking models such as EMs before they are used as a basis for advice. It also intends to ensure a level playing field between models of different complexity by asking the same questions to all candidates for a given advice product.

### 4.1. What the framework is not able to deliver

The suite of questions proposed provides a framework that ensures that all topics required in a target-oriented skill assessment are covered. However, they do not provide best practice guidelines on methods or analyses to be carried out under specific questions. The reason for this is that the variety of potential advice questions and modeling approaches is simply too large to propose one or several methods as best practice. The development of new methods and model types in times of big data and machine learning is also increasingly fast and any best practice guidelines on specific methods risk to be outdated soon after publication. Nevertheless, our guiding questions provide a framework for scientific communities to develop further technical guidelines for each question.

The framework also focuses on the “exploitation” phase of a model life cycle in the sense of its operational use (NRC, 2007). It provides guidance on how to test whether an existing model is fit for purpose. Although it is useful for other steps in a model life cycle, the framework of questions presented here is not intended to be complete for e.g., the model development phase. However, also for the development phase the framework of questions may be useful to plan future steps as demonstrated in the examples for Georges Bank Hydra and Rpath models.

### 4.2. What the framework is able to deliver

Answering the outlined questions (Fig. 1) provides the possibility to find the best model for a given advice product from alternatives tested and/or whether a certain model is fit for purpose. An overall picture of model performance and sources of uncertainties can be derived from answering the questions proposed in this framework. The results of the skill assessment provide a suitable basis to judge on the strength and weakness of the model to determine which model output and at which scales may be useful for advice. This question is much more difficult to answer for ecosystem and multispecies models than for e.g., a single species stock assessment model. For example, it depends on the advice question whether a good performance for several but not all species/stocks in the model is sufficient or not.

Following the proposed framework can improve the trust in complex models and in any case will improve transparency in the modeling work performed. For example, MSEs with single species models as operating model are used all over the world to test harvesting strategies. Missing important processes could lead to biased population dynamics in the operating model, finally leading to biased results and potentially wrong management decisions. It has to be highlighted that the same rigor must be applied to all models independent of their complexity. In order to be able to simulate more processes (e.g., food web interactions, climate change impacts), there is an increasing demand to use EMs in MSE simulations (Kaplan et al., 2021). While full ecosystem or end-to-end models (e.g., Atlantis) are mainly used as operating models to be able to test e.g., the robustness of management strategies in providing sustainable outcomes over a large variety of sources for bias and uncertainties, MICE models may also be used in the assessment part of the MSE loop to test their performance (see also above under Simulation testing for statistical models).

The guiding questions presented here are universal and can be applied to all models independent of their complexity and usage. The guiding questions can also easily be used for model ensembles. An increasing number of methods for multi-model inference is available for application in fisheries contexts (Gårdmark et al., 2013; Ianelli et al., 2016; Anderson et al., 2017; Spence et al., 2018). Multi-model ensembles are routinely used in weather and climate forecasting (e.g., Tracton and Kalnay, 1993; Tebaldi and Knutti, 2007; Semenov and Stratonovitch; 2010, Krishnamurti et al., 2016; Du et al., 2018). Model ensembles have been shown to outperform single models in terms of error compensation and improved consistency, and also lead to a more complete quantification of structural uncertainties (e.g., Hagedorn et al., 2005; Lotze et al., 2019). Skill assessment of a model ensemble can be done in a similar way as for individual models by comparison to historical observations and reserved current observations to evaluate forecast skill (e.g., Zhou and Du, 2010; Leonardo and Colle, 2017), or by using simulation analysis.

## CRediT authorship contribution statement

**Alexander Kempf:** Conceptualization, Investigation, Writing – original draft preparation. **Sarah K Gaichas:** Supervision, Conceptualization, Investigation, Writing – review & editing. **Michael A. Spence:** Investigation, Writing – review & editing. **Sigrid Lehuta:** Investigation, Writing – review & editing. **Vanessa Trijoulet:** Investigation, Writing – review & editing. **Valerio Bartolino:** Investigation, Writing – review & editing. **Ching Villanueva:** Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

## Acknowledgements

The authors are grateful to all other members of the ICES Working Group on Multispecies Stock Assessment Methods (WGSAM) for critical discussion on the framework during the last years. We specifically thank Jacob Bentley, Morten Vinther and Sean Lucey for their modeling work which was the basis for some of the examples we provide from WGSAM reports.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2023.106845.

## References

AIAA , 1998. Guide for the verification and validation of computational fluid dynamics simulations (AIAA G-077–1998(2002)). ⟨https://doi.org/10.2514/4.472855⟩.

Allen, J.I., Somerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. J. Mar. Syst. 64 (1–4), 3–14. https://doi.org/10.1016/j.jmarsys.2006.02.010.

Allen, J.I., Aiken, J., Anderson, T.R., Buitenhuis, E., Cornell, S., Geider, R.J., Haines, K., Hirata, T., Holt, J., Le Quéré, C., Hardman-Mountford, N., Ross, O.N., Sinha, B., While, J., 2010. Marine ecosystem models for earth systems applications: The MarQUEST experience. J. Mar. Syst. 81 (1–2), 19–33. https://doi.org/10.1016/j.jmarsys.2009.12.017.

Anderson, S.C., Cooper, A.B., Jensen, O.P., Minto, C., Thorson, J.T., Walsh, J.C., Afflerbach, J., Dickey-Collas, M., Kleisner, K.M., Longo, C., Osio, G.C., Ovando, D., Mosqueira, I., Rosenberg, A.A., Selig, E.R., 2017. Improving estimates of population status and trend with superensemble models. Fish Fish. 18 (4), 732–741. https://doi.org/10.1111/faf.12200.

Bannister, H.J., Blackwell, P.G., Hyder, K., Webb, T.J., 2021. Improving the visual communication of environmental model projections. Sci. Rep. 11, 19157. https://doi.org/10.1038/s41598-021-98290-4.

Begley, J., 2005. Gadget user guide. Mar. Res. Inst. Rep. Ser. 120, 90.

Bentley, J.W., Lundy, M.G., Howell, D., Beggs, S.E., Bundy, A., De Castro, F., Fox, C.J., Heymans, J.J., Lynam, C.P., Pedreschi, D., Schuchert, P., Serpetti, N., Woodlock, J., Reid, D.G., 2021. Refining fisheries advice with stock-specific ecosystem information. Front. Mar. Sci. 8. https://doi.org/10.3389/fmars.2021.602072.

Blanchard, J.L., Andersen, K.H., Scott, F., Hintzen, N.T., Piet, G., Jennings, S., 2014. Evaluating targets and trade-offs among fisheries and conservation objectives using a multispecies size spectrum model. J. Appl. Ecol. 51 (3), 612–622. https://doi.org/10.1111/1365-2664.12238.

Bracis, C., Lehuta, S., Savina-Rolland, M., Travers-Trolet, M., Girardin, R., 2020. Improving confidence in complex ecosystem models: the sensitivity analysis of an Atlantis ecosystem model. Ecol. Model. 431, 109133 https://doi.org/10.1016/j.ecolmodel.2020.109133.

Brynjarsdóttir, J., O'Hagan, A., 2014. Learning about physical parameters: the importance of model discrepancy. Inverse Probl. 30 (11), 114007 https://doi.org/10.1088/0266-5611/30/11/114007.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R., Maunder, M.N., Taylor, I., Wetzel, C.R., Doering, K., Johnson, K.F., Methot, R.D., 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. 240, 105959 https://doi.org/10.1016/j.fishres.2021.105959.

Christensen, V., Walters, C.J., 2004. Ecopath with Ecosim: methods, capabilities and limitations. Ecol. Model. 172 (2–4), 109–139. https://doi.org/10.1016/j.ecolmodel.2003.09.

Collie, J.S., Botsford, L.W., Hastings, A., Kaplan, I.C., Largier, J.L., Livingston, P.A., Plagányi, É., Rose, K.A., Wells, B.K., Werner, F.E., 2016. Ecosystem models for fisheries management: finding the sweet spot. Fish Fish. 17 (1), 101–125. https://doi.org/10.1111/faf.12093.

Cury, P., Shin, Y., Planque, B., Durant, J., Fromentin, J., Kramerschadt, S., Stenseth, N., Travers, M., Grimm, V., 2008. Ecosystem oceanography for global change in fisheries. Trends Ecol. Evol. 23 (6), 338–346. https://doi.org/10.1016/j.tree.2008.02.005.

Da Veiga, S., Gamboa, F., Iooss, B. & Prieur, C., 2021. Chapter 7: A case study in R: COVID-19 epidemic model. In: Basics and Trends in Sensitivity Analysis, 187–227. Computational Science & Engineering. Society for Industrial and Applied Mathematics, 2021. https://doi.org/10.1137/1.9781611976694.ch7.

Dabrowski, T., Lyons, K., Berry, A., Cusack, C., Nolan, G.D., 2014. An operational biogeochemical model of the North-East Atlantic: model description and skill assessment. J. Mar. Syst. 129, 350–367. https://doi.org/10.1016/j.jmarsys.2013.08.001.

De Mora, L., Butenschön, M., Allen, J.I., 2016. The assessment of a global marine ecosystem model on the basis of emergent properties and ecosystem function: a case study with ERSEM. Geosci. Model Dev. 9 (1), 59–76. https://doi.org/10.5194/gmd-9-59-2016.

Dickey-Collas, M., Payne, M.R., Trenkel, V.M., Nash, R.D.M., 2014. Hazard warning: model misuse ahead. ICES J. Mar. Sci. 71 (8), 2300–2306. https://doi.org/10.1093/icesjms/fst215.

Dolan, T.E., Patrick, W.S., Link, J.S., 2016. Delineating the continuum of marine ecosystem-based management: a US fisheries reference point perspective. Ices J. Mar. Sci. 73 (4), 1042–1050. https://doi.org/10.1093/icesjms/fsv242.

Du, J., Berner, J., Buizza, R., Charron, M., Houtekamer, P., Hou, D., Jankov, I., Mu, M., Wang, X., Wei, M., Yuan, H., 2018. Ensemble Methods for Meteorological Predictions. Handb. Hydrometeorol. Ensemble Forecast. 1–52. https://doi.org/10.1007/978-3-642-40457-3_13-1.

FAO ,2008. Best practices in ecosystem modelling for informing an ecosystem approach to fisheries. FAO Fisheries Technical Guidelines for Responsible Fisheries. No. 4, Suppl. 2, Add. 1, Rome, FAO, 1–78.

Ferretti, F., Saltelli, A., Tarantola, S., 2016. Trends in sensitivity analysis practice in the last decade. Sci. Total Environ. 568, 666–670. https://doi.org/10.1016/j.scitotenv.2016.02.133.

Friedrichs, M.A.M., Hood, R.R., Wiggert, J.D., 2006. Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data. Deep Sea Res. Part II: Top. Stud. Oceanogr. 53 (5–7), 576–600. https://doi.org/10.1016/j.dsr2.2006.01.026.

Friedrichs, M.A.M., Dusenberry, J.A., Anderson, L.A., Armstrong, R.A., Chai, F., Christian, J.R., Doney, S.C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D.J., Moore, J.K., Schartau, M., Spitz, Y.H., Wiggert, J.D., 2007. Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups. J. Geophys. Res. 112 (C8) https://doi.org/10.1029/2006jc003852.

Fulton, E.A., Link, J.S., Kaplan, I.C., Savina-Rolland, M., Johnson, P., Ainsworth, C., Horne, P., Gorton, R., Gamble, R.J., Smith, A.D.M., Smith, D.C., 2011. Lessons in modelling and management of marine ecosystems: the Atlantis experience. Fish Fish. 12 (2), 171–188. https://doi.org/10.1111/j.1467-2979.2011.00412.x.

Gaichas, S.K., Fogarty, M., Fay, G., Gamble, R., Lucey, S., Smith, L., 2017. Combining stock, multispecies, and ecosystem level fishery objectives within an operational management procedure: simulations to start the conversation. ICES J. Mar. Sci. 74 (2), 552–565. https://doi.org/10.1093/icesjms/fsw119.

Garcia, S.M.; Zerbi, A.; Aliaume, C.; Do Chi, T.; Lasserre, G. , 2003. The ecosystem approach to fisheries. Issues, terminology, principles, institutional foundations, implementation and outlook. FAO Fisheries Technical Paper. No. 443. Rome, FAO. 2003. 71 p.

Gårdmark, A., Lindegren, M., Neuenfeldt, S., Blenckner, T., Heikinheimo, O., Müller-Karulis, B., Niiranen, S., Tomczak, M.T., Aro, E., Wikström, A., Möllmann, C., 2013. Biological ensemble modeling to evaluate potential futures of living marine resources. Ecol. Appl. 23 (4), 742–754. https://doi.org/10.1890/12-0267.1.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data Anal. https://doi.org/10.1201/b16018.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. Science 310 (5750), 987–991. https://doi.org/10.1126/science.1116681.

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus A: Dyn. Meteorol. Oceanogr. 57 (3), 219. https://doi.org/10.3402/tellusa.v57i3.14657.

Hall, S.J., Collie, J.S., Duplisea, D.E., Jennings, S., Bravington, M., Link, J., 2006. A length-based multispecies model for evaluating community responses to fishing. Can. J. Fish. Aquat. Sci. 63 (6), 1344–1359. https://doi.org/10.1139/f06-039.

Hansen, C., Drinkwater, K.F., Jähkel, A., Fulton, E.A., Gorton, R., Skern-Mauritzen, M., 2019. Sensitivity of the Norwegian and Barents Sea Atlantis end-to-end ecosystem model to parameter perturbations of key species. PLOS ONE 14 (2), e0210419. https://doi.org/10.1371/journal.pone.0210419.

Hatton, I.K., McCann, K.S., Fryxell, J.M., Davies, T.J., Smerlak, M., Sinclair, A., Loreau, M., 2015. The predator-prey power law: biomass scaling across terrestrial and aquatic biomes. Science 349 (6252). https://doi.org/10.1126/science.aac6284.

Hernandez, F., Bertino, L., Brassington, G., Chassignet, E., Cummings, Davidson, F., Drévillon, M., Garric, G., Kamachi, M., Lellouche, J.-M., Mahdon, R., Martin, M., Ratsimandresey, A., Regnier, C., 2009. Validation and intercomparison studies within GODAE. Oceanography 22 (3), 128–143. https://doi.org/10.5670/oceanog.2009.71.

Heymans, J.J., Coll, M., Link, J.S., Mackinson, S., Steenbeek, J., Walters, C., Christensen, V., 2016. Best practice in ecopath with ecosim food-web models for ecosystem-based management. Ecol. Model. 331, 173–184. https://doi.org/10.1016/j.ecolmodel.2015.12.007.

Hill, S.L., Watters, G.M., Punt, A.E., McAllister, M.K., Quéré, C.L., Turner, J., 2007. Model uncertainty in the ecosystem approach to fisheries. Fish Fish. 8 (4), 315–336. https://doi.org/10.1111/j.1467-2979.2007.00257.x.

Howell, D., Filin, A.A., Bogstad, B., Stiansen, J.E., 2013. Unquantifiable uncertainty in projecting stock response to climate change: example from North East Arctic cod. Mar. Biol. Res. 9 (9), 920–931. https://doi.org/10.1080/17451000.2013.775452.

Hyder, K., Rossberg, A.G., Allen, J.I., Austen, M.C., Barciela, R.M., Bannister, H.J., Blackwell, P.G., Blanchard, J.L., Burrows, M.T., Defriez, E., Dorrington, T., Edwards, K.P., Garcia-Carreras, B., Heath, M.R., Hembury, D.J., Heymans, J.J., Holt, J., Houle, Jennifer, E., Jennings, S., Mackinson, S., 2015. Making modelling count - increasing the contribution of shelf-seas community and ecosystem models to policy development and management. Mar. Policy 61, 291–302. https://doi.org/10.1016/j.marpol.2015.07.015.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22 (4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Ianelli, J., Holsman, K.K., Punt, A.E., Aydin, K., 2016. Multi-model inference for incorporating trophic and climate uncertainty into stock assessments. Deep Sea Res.

Part II: Top. Stud. Oceanogr. 134, 379–389. https://doi.org/10.1016/j.dsr2.2015.04.002.
ICES Benchmark Guidelines, 2023. Guidelines for Benchmarks. Version 1. ICES Guidelines and Policies - Advice Technical Guidelines, 26 pp. https://doi.org/10.17895/ices.pub.22316743.
ICES HAWG, 2022. Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG). ICES Scientific Reports, 4:16, 745 pp. http://doi.org/10.17895/ices.pub.10072.
ICES HAWG, 2023. Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG). ICES Scientific Reports, 5:23 https://doi.org/10.17895/ices.pub.22182034.v1.
ICES WGBFAS, 2022. Baltic Fisheries Assessment Working Group(WGBFAS).ICES Scientific Reports, 4:44, 659pp. http://doi.org/10.17895/ices.pub.19793014.
ICES WGNSSK, 2022. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak (WGNSSK). ICES Scientific Reports, 4:43, 1376 pp. http://doi.org/10.17895/ices.pub.19786285.
ICES WGSAM, 2013. Interim Report of the Working Group on Multispecies Assessment Methods (WGSAM). ICES CM 2013/SSGSUE:10.
ICES WGSAM, 2016. Report of the Working Group on Multispecies Assessment Methods (WGSAM). ICES CM 2016/SSGEPI:20.
ICES WGSAM, 2017. Report of the Working Group on Multispecies Assessment Methods (WGSAM). ICES CM 2017/SSGEPI:20.
ICES WGSAM, 2019. Report of the Working Group on Multispecies Assessment Methods (WGSAM). ICES Scientific Reports, 1:91, 320 pp. http://doi.org/10.17895/ices.pub.5758.
ICES WGSAM, 2023. Working Group on Multispecies Assessment Methods (WGSAM; outputs from 2022 meeting). ICES Scientific Reports. 5:12. 233 pp. https://doi.org/10.17895/ices.pub.22087292.
ICES WGSAM, 2015. Report of the Working Group on Multispecies Assessment Methods (WGSAM). ICES CM 2015/SSGEPI:20.
ICES WGSAM, 2021. Working Group on Multispecies Assessment Methods (WGSAM; outputs from 2020 meeting). ICES Scientific Reports, 3:10, 231 pp. https://doi.org/10.17895/ices.pub.7695.
ICES WKGMSE, 2013. Report of the Workshop on Guidelines for Management Strategy Evaluations (WKGMSE). ICES CM 2013/ACOM: 39.
ICES WKGMSE2, 2019. Workshop on Guidelines for Management Strategy Evaluations (WKGMSE2). ICES Scientific Reports, 1:33, 162 pp. http://doi.org/10.17895/ices.pub.5331.
ICES WKGMSE3, 2020. The third Workshop on Guidelines for Management Strategy Evaluations (WKGMSE3). ICES Scientific Reports. 2:116. 112 pp. http://doi.org/10.17895/ices.pub.7627.
ICES WKNSMSE, 2019. WORKSHOP ON NORTH SEA STOCKS MANAGEMENT STRATEGY EVALUATION (WKNSMSE). ICES Scientific Reports, 1:12, 378 pp. http://doi.org/10.17895/ices.pub.5090.
Jolliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A.M., Helber, R., Arnone, R.A., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. J. Mar. Syst. 76 (1–2), 64–82. https://doi.org/10.1016/j.jmarsys.2008.05.014.
Jørgensen, T.H., 2023. Sensitivity to calibrated parameters. Rev. Econ. Stat. 105 (2), 474–481. https://doi.org/10.1162/rest_a_01054.
Kaplan, I.C., Marshall, K.N., 2016. A guinea pig's tale: learning to review end-to-end marine ecosystem models for management applications. ICES J. Mar. Sci. 73 (7), 1715–1724. https://doi.org/10.1093/icesjms/fsw047.
Kaplan, I.C., Gaichas, S.K., Stawitz, C.C., Lynch, P.D., Marshall, K.N., Deroba, J.J., Masi, M., Brodziak, J.K., Aydin, K.Y., Holsman, K., Townsend, H., Tommasi, D., Smith, J.A., Koenigstein, S., Weijerman, M., Link, J., 2021. Management strategy evaluation: Allowing the light on the hill to illuminate more than one species. Front. Mar. Sci. 8. https://doi.org/10.3389/fmars.2021.624355.
Karp, M.A., Link, J.S., Grezlik, M., Cadrin, S., Fay, G., Lynch, P., Townsend, H., Methot, R.D., Adams, G.D., Blackhart, K., Barceló, C., Buchheister, A., Cieri, M., Chagaris, D., Christensen, V., Craig, J.K., Cummings, J., Damiano, M.D., Dickey-Collas, M., Elvarsson, B.Þ, Gaichas, S., Haltuch, M.A., Haugen, J.B., Howell, D., Kaplan, I.C., Klajbor, W., Large, S.I., Masi, M., McNamee, J., Muffley, B., Murray, S., Plagányi, É., Reid, D., Rindorf, A., Sagarese, S.R., Schueller, A.M., Thorpe, R., Thorson, J.T., Tomczak, M.T., Trijoulet, V., Voss, R., 2023. Increasing the uptake of multispecies models in fisheries management. ICES J. Mar. Sci. Volume 80 (*Issue 2*), 243–257. https://doi.org/10.1093/icesjms/fsad001.
Inverse Methods in Global Biogeochemical Cycles. In: Kasibhatla, P., Heimann, M., Rayner, P., Mahowald, N., Prinn, R.G., Hartley, D.E. (Eds.), 2000. Geophysical Monograph Series. American Geophysical Union. https://doi.org/10.1029/gm114.
Kell, L., Levontin, P., Cambell, R., Pilling, G., Maunder, M., Sharma, R., 2016a. The Quantification and Presentation of Risk. In: Edwards, Dankel (Eds.), Management Science in Fisheries - An Introduction to Simulation-based Methods. Publisher: Routledge. ISBN: 9781317615170.
Kell, L.T., Kimoto, A., Kitakado, T., 2016b. Evaluation of the prediction skill of stock assessment using hindcasting. Fish. Res. 183, 119–127. https://doi.org/10.1016/j.fishres.2016.05.017.
Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., Fu, D., 2021. Validation of stock assessment methods: Is it me or my model talking. ICES J. Mar. Sci. 78 (6), 2244–2255. https://doi.org/10.1093/icesjms/fsab104.
Kempf, A., Mumford, J., Levontin, P., Leach, A., Hoff, A., Hamon, K.G., Bartelings, H., Vinther, M., Stäbler, M., Poos, J.J., Smout, S., Frost, H., van den Burg, S., Ulrich, C., Rindorf, A., 2016. The MSY concept in a multi-objective fisheries environment – Lessons from the North Sea. Mar. Policy 69, 146–158. https://doi.org/10.1016/j.marpol.2016.04.012.
Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. J. R. Stat. Soc.: Ser. B (Stat. Methodol. ) 63 (3), 425–464. https://doi.org/10.1111/1467-9868.00294.
Kraak, S.B.M., Kelly, C.J., Codling, E.A., Rogan, E., 2010. On scientists' discomfort in fisheries advisory science: the example of simulation-based fisheries management-strategy evaluations. Fish Fish. 11 (2), 119–132. https://doi.org/10.1111/j.1467-2979.2009.00352.x.
Krishnamurti, T.N., Kumar, V., Simon, A., Bhardwaj, A., Ghosh, T., Ross, R., 2016. A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. Rev. Geophys. 54 (2), 336–377. https://doi.org/10.1002/2015rg000513.
Lehuta, S., Mahévas, S., Petitgas, P., Pelletier, D., 2010. Combining sensitivity and uncertainty analysis to evaluate the impact of management measures with ISIS–Fish: marine protected areas for the Bay of Biscay anchovy (Engraulis encrasicolus) fishery. ICES J. Mar. Sci. 67 (5), 1063–1075. https://doi.org/10.1093/icesjms/fsq002.
Lehuta, S., Girardin, R., Mahévas, S., Travers-Trolet, M., Vermard, Y., 2016. Reconciling complex system models and fisheries advice: Practical examples and leads. Aquat. Living Resour. 29 (2), 208. https://doi.org/10.1051/alr/2016022.
Lehuta, S., Petitgas, P., Mahévas, S., Huret, M., Vermard, Y., Uriarte, A., Record, N.R., 2013. Selection and validation of a complex fishery model using an uncertainty hierarchy. Fish. Res. 143, 57–66. https://doi.org/10.1016/j.fishres.2013.01.008.
Leonardo, N.M., Colle, B.A., 2017. Verification of Multimodel Ensemble Forecasts of North Atlantic Tropical Cyclones. Weather Forecast. 32 (6), 2083–2101. https://doi.org/10.1175/waf-d-17-0058.1.
Levontin, P., Baranowski, P., Leach, A.W., Bailey, A., Mumford, J.D., Quetglas, A., Kell, L. T., 2017. On the role of visualisation in fisheries management. Mar. Policy 78, 114–121. https://doi.org/10.1016/j.marpol.2017.01.018.
Lewy, P., Vinther, M., 2004. A stochastic age-length structured multispecies model applied to North Sea stocks. Danish Institute for. Fish. Res., CM 2004/FF 20.
Link, J.S., 2010a. Why is an ecosystem approach now strongly heralded and merited? In: Ecosystem-Based Fisheries Management: Confronting Tradeoffs, 20–33, Cambridge: Cambridge University Press. doi:(10.1017/CBO9780511667091.004).
Link, J.S., 2010c. Adding rigor to ecological network models by evaluating a set of pre-balance diagnostics: a plea for PREBAL. Ecol. Model. 221 (12), 1580–1591. https://doi.org/10.1016/j.ecolmodel.2010.03.012.
Link, J.S., Ihde, T.F., Harvey, C.J., Gaichas, S.K., Field, J.C., Brodziak, J.K.T., Townsend, H.M., Peterman, R.M., 2012. Dealing with uncertainty in ecosystem models: the paradox of use for living marine resource management. Prog. Oceanogr. 102, 102–114. https://doi.org/10.1016/j.pocean.2012.03.008.
Link, J.S., Ihde, T.F., Townsend, H.M., Osgood, K.E., Schirripa, M.J., Kobayashi, D.R., Gaichas, S.K., Field, J.C., Levin, P.S., Aydin, K.Y., Harvey C.J., 2010b. Report of the 2nd National Ecosystem Modeling Workshop (NEMoW II): Bridging the Credibility Gap Dealing with Uncertainty in Ecosystem Models. US Department of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service. Silver Spring. Maryland USA.
Lotze, H.K., Tittensor, D.P., Bryndum-Buchholz, A., Eddy, T.D., Cheung, W.W.L., Galbraith, E.D., Barange, M., Barrier, N., Bianchi, D., Blanchard, J.L., Bopp, L., Büchner, M., Bulman, C.M., Carozza, D.A., Christensen, V., Coll, M., Dunne, J.P., Fulton, E.A., Jennings, S., Jones, M.C., Mackinson, S., Maury, O., Niiranen, S., Oliveros-Ramos, R., Roy, T., Fernandes, J.A., Schewe, J., Shin, Y.J., Silva, T.A.M., Steenbee, K.J., Stock, C.A., Verley, P., Volkholz, J., Walker, N.D., Worm, B., 2019. Global ensemble projections reveal trophic amplification of ocean biomass declines with climate change. Proc. Natl. Acad. Sci. USA 116 (26), 12907–12912.
Lucey, S.M., Gaichas, S.K., Aydin, K.Y., 2020. Conducting reproducible ecosystem modeling using the open source mass balance model Rpath. Ecol. Model. 427, 109057 https://doi.org/10.1016/j.ecolmodel.2020.109057.
Maar, M., Butenschön, M., Daewel, U., Eggert, A., Fan, W., Hjøllo, S.S., Hufnagl, M., Huret, M., Ji, R., Lacroix, G., Peck, M.A., Radtke, H., Sailley, S., Sinerchia, M., Skogen, M.D., Travers-Trolet, M., Troost, T.A., van de Wolfshaar, K., 2018. Responses of summer phytoplankton biomass to changes in top-down forcing: insights from comparative modelling. Ecol. Model. 376, 54–67. https://doi.org/10.1016/j.ecolmodel.2018.03.003.
Marzloff, M.P., Johnson, C.R., Little, L.R., Soulié, J.-C., Ling, S.D., Frusher, S.D., 2013. Sensitivity analysis and pattern-oriented validation of TRITON, a model with alternative community states: Insights on temperate rocky reefs dynamics. Ecol. Model. 258, 16–32. https://doi.org/10.1016/j.ecolmodel.2013.02.022.
Maury, O., 2010. An overview of APECOSM, a spatialized mass balanced "Apex Predators ECOSystem Model" to study physiologically structured tuna population dynamics in their ecosystem. Prog. Oceanogr. 84 (1–2), 113–117. https://doi.org/10.1016/j.pocean.2009.09.013.
Mitchell, P.J., Spence, M.A., Aldridge, J., Kotilainen, A.T., Diesing, M., 2021. Sedimentation rates in the Baltic Sea: A machine learning approach. Cont. Shelf Res. 214, 104325 https://doi.org/10.1016/j.csr.2020.104325.
Mohn, R., 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56 (4), 473–488. https://doi.org/10.1006/jmsc.1999.0481.
Morris, D.J., Speirs, D.C., Cameron, A.I., Heath, M.R., 2014. Global sensitivity analysis of an end-to-end marine ecosystem model of the North Sea: Factors affecting the biomass of fish and benthos. Ecol. Model. 273, 251–263. https://doi.org/10.1016/j.ecolmodel.2013.11.019.
Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33 (2), 161–174. https://doi.org/10.1080/00401706.1991.10484804.
Myers, R.A., 1998. Rev. Fish. Biol. Fish. 8 (3), 285–305. https://doi.org/10.1023/a:1008828730759.

Nielsen, A., Hintzen, N.T., Mosegaard, H., Trijoulet, V., Berg, C.W., 2021. Multi-fleet state-space assessment model strengthens confidence in single-fleet SAM and provides fleet-specific forecast options. ICES J. Mar. Sci. 78 (6), 2043–2052. https://doi.org/10.1093/icesjms/fsab078.

NRC, 2007. Models in Environmental Regulatory Decision Making. National Academies Press. https://doi.org/10.17226/11972.

Olsen, E., Fay, G., Gaichas, S., Gamble, R., Lucey, S., Link, J.S., 2016. Ecosystem model skill assessment. Yes We Can! PLOS ONE 11 (1), e0146467. https://doi.org/10.1371/journal.pone.0146467.

Pastoors, M.A., Poos, J.J., Kraak, S.B.M., Machiels, M.A.M., 2007. Validating management simulation models and implications for communicating results to stakeholders. ICES J. Mar. Sci. 64 (4), 818–824. https://doi.org/10.1093/icesjms/fsm051.

Pianosi, F., Sarrazin, F., Wagener, T., 2015. A matlab toolbox for global sensitivity analysis. Environ. Model. Softw. 70, 80–85. https://doi.org/10.1016/j.envsoft.2015.04.009.

Pikitch, E.K., Santora, C., Babcock, E.A., Bakun, A., Bonfil, R., Conover, D.O., Dayton, P., Doukakis, P., Fluharty, D., Heneman, B., Houde, E.D., Link, J., Livingston, P.A., Mangel, M., McAllister, M.K., Pope, J., Sainsbury, K.J., 2004. Ecosystem-based fishery management. Science 305 (5682), 346–347. https://doi.org/10.1126/science.1098222.

Plagányi, É.E., Punt, A.E., Hillary, R., Morello, E.B., Thébaud, O., Hutton, T., Pillans, R. D., Thorson, J.T., Fulton, E.A., Smith, A.D.M., Smith, F., Bayliss, P., Haywood, M., Lyne, V., Rothlisberg, P.C., 2012. Multispecies fisheries management and conservation: tactical applications using models of intermediate complexity. Fish Fish. 15 (1), 1–22. https://doi.org/10.1111/j.1467-2979.2012.00488.x.

Plagányi, E.E. , 2007. Models for an ecosystem approach to fisheries. FAO Fisheries Technical Paper. No. 477. Rome, FAO. 2007. 108p.

Planque, B., 2015. Projecting the future state of marine ecosystems, "la grande illusion"? ICES J. Mar. Sci.: J. Du Cons. 73 (2), 204–208. https://doi.org/10.1093/icesjms/fsv155.

Püts, M., Taylor, M., Núñez-Riboni, I., Steenbeek, J., Stäbler, M., Möllmann, C., Kempf, A., 2020. Insights on integrating habitat preferences in process-oriented ecological models – a case study of the southern North Sea. Ecol. Model. 431, 109189 https://doi.org/10.1016/j.ecolmodel.2020.109189.

Rindorf, A., Cardinale, M., Shephard, S., De Oliveira, J.A., Hjorleifsson, E., Kempf, A., Luzenczyk, A., Millar, C., Miller, D.C., Needle, C.L., Simmonds, J., Vinther, M., 2016. Fishing for MSY: Using "pretty good yield" ranges without impairing recruitment. ICES J. Mar. Sci. 74 (2), 525–534. https://doi.org/10.1093/icesjms/fsw111.

Rochet, M.-J., Rice, J.C., 2009. Simulation-based management strategy evaluation: ignorance disguised as mathematics. ICES J. Mar. Sci. 66 (4), 754–762. https://doi.org/10.1093/icesjms/fsp023.

Rougier, J., Beven, K., 2013. Model and data limitations: The sources and implications of epistemic uncertainty. In: Rougier, J., Sparks, S., Hill, L. (Eds.), Risk and Uncertainty Assessment for Natural Hazards. Cambridge University Press, Cambridge, pp. 40–63. https://doi.org/10.1017/CBO9781139047562.004.

Ryan, A.G., Regnier, C., Divakaran, P., Spindler, T., Mehra, A., Smith, G.C., Davidson, F., Hernandez, F., Maksymczuk, J., Liu, Y., 2015. GODAE OceanView Class 4 forecast verification framework: global ocean inter-comparison. J. Oper. Oceanogr. s98–s111. https://doi.org/10.1080/1755876x.2015.1022330. , 8(sup1).

Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environ. Model. Softw. 25 (12), 1508–1517. https://doi.org/10.1016/j.envsoft.2010.04.012.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2006. Sensitivity analysis practices: strategies for model-based inference. Reliab. Eng. Syst. Saf. 91 (10–11), 1109–1125. https://doi.org/10.1016/j.ress.2005.11.014.

Sandvik, A.D., Skagseth, Ø., Skogen, M.D., 2016. Model validation: Issues regarding comparisons of point measurements and high-resolution modeling results. Ocean Model. 106, 68–73. https://doi.org/10.1016/j.ocemod.2016.09.007.

Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting environmental decision making: a strategy for the future. Trends Ecol. Evol. 25 (8), 479–486. https://doi.org/10.1016/j.tree.2010.05.001.

Schoener, T.W., 1970. Nonsynchronous spatial overlap of lizards in patchy habitats. Ecology 51 (3), 408–418. https://doi.org/10.2307/1935376.

Semenov, M., Stratonovitch, P., 2010. Use of multi-model ensembles from global climate models for assessment of climate change impacts. Clim. Res. 41, 1–14. https://doi.org/10.3354/cr00836.

Shin, Y.-J., Cury, P., 2004. Using an individual-based model of fish assemblages to study the response of size spectra to changes in fishing. Can. J. Fish. Aquat. Sci. 61 (3), 414–431. https://doi.org/10.1139/f03-154.

Skern-Mauritzen, M., Ottersen, G., Handegard, N.O., Huse, G., Dingsør, G.E., Stenseth, N. C., Kjesbu, O.S., 2016. Ecosystem processes are rarely included in tactical fisheries management. Fish Fish. 17 (1), 165–175. https://doi.org/10.1111/faf.12111.

Skogen, M.D., Ji, R., Akimova, A., Daewel, U., Hansen, C.B., Hjøllo, S.S., Van Leeuwen, S. M., Maar, M., Macías, D., Mousing, E.A., Almroth-Rosell, E., Sailley, S.F., Spence, M., Troost, T.A., Van De Wolfshaar, K.E., 2021. Disclosing the truth: are models better than observations. Mar. Ecol. Prog. Ser. 680, 7–13. https://doi.org/10.3354/meps13574.

Spence, M., Griffiths, C., Waggitt, J., Bannister, H., Thorpe, R., Rossberg, A., Lynam, C., 2021a. Sustainable fishing can lead to improvements in marine ecosystem status: an ensemble-model forecast of the North Sea ecosystem. Mar. Ecol. Prog. Ser. https://doi.org/10.3354/meps13870.

Spence, M.A., Blackwell, P.G., 2016. Coupling random inputs for parameter estimation in complex models. Stat. Comput. 26 (6), 1137–1146. https://doi.org/10.1007/s11222-015-9593-2.

Spence, M.A., Muiruri, E.W., Maxwell, D.L., Davis, S., Sheahan, D., 2021c. The application of continuous-time Markov chain models in the analysis of choice flume experiments. J. R. Stat. Soc.: Ser. C. (Appl. Stat. ) 70 (4), 1103–1123. https://doi.org/10.1111/rssc.12510.

Spence, M.A., Thorpe, R.B., Blackwell, P.G., Scott, F., Southwell, R., Blanchard, J.L., 2021b. Quantifying uncertainty and dynamical changes in multi-species fishing mortality rates, catches and biomass by combining state-space and size-based multi-species models. Fish Fish. https://doi.org/10.1111/faf.12543.

Spence, M.A., Blanchard, J.L., Rossberg, A.G., Heath, M.R., Heymans, J.J., Mackinson, S., Serpetti, N., Speirs, D.C., Thorpe, R.B., Blackwell, P.G., 2018. A general framework for combining ecosystem models. Fish Fish. 19 (6), 1031–1042. https://doi.org/10.1111/faf.12310.

Sterman, J.D., 1984. Appropriate summary statistics for evaluating the historical fit of system dynamics models. Dynamica 10 (2), 51–66.

Stow, C.A., Jolliff, J., McGillicuddy, D.J., Doney, S.C., Allen, J.I., Friedrichs, M.A.M., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. J. Mar. Syst. 76 (1–2), 4–15. https://doi.org/10.1016/j.jmarsys.2008.03.011.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmosph. 106 (D7), 7183–7192. https://doi.org/10.1029/2000jd900719.

Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. R. Soc. A: Math., Phys. Eng. Sci. 365 (1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076.

Thorpe, R.B., De Oliveira, J.A.A., 2019. Comparing conceptual frameworks for a fish community MSY (FCMSY) using management strategy evaluation—an example from the North Sea. ICES J. Mar. Sci. 76 (4), 813–823. https://doi.org/10.1093/icesjms/fsz015.

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 Ecol. Evol. 6 (1), 49–58. https://doi.org/10.1111/2041-210x.12292.

Tracton, M.S., Kalnay, E., 1993. Operational ensemble prediction at the National Meteorological Center: practical aspects. Weather Forecast. 8 (3), 379–398. https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Trijoulet, V., Fay, G., Miller, T.J., 2020. Performance of a state-space multispecies model: What are the consequences of ignoring predation and process errors in stock assessments. J. Appl. Ecol. 57 (1), 121–135. https://doi.org/10.1111/1365-2664.13515.

Trijoulet, V., Fay, G., Curti, K.L., Smith, B., Miller, T.J., 2019. Performance of multispecies assessment models: insights on the influence of diet data. ICES J. Mar. Sci. 76 (6), 1464–1476. https://doi.org/10.1093/icesjms/fsz053.

Turányi, T., 1990. Sensitivity analysis of complex kinetic systems. Tools and applications. J. Math. Chem. 5 (3), 203–248. https://doi.org/10.1007/bf01166355.

Vigier, A., Mahévas, S., Bertignac, M., 2018. Towards a spatial integrated stock assessment model for European hake northern stock. Fish. Res. 199, 158–170. https://doi.org/10.1016/j.fishres.2017.12.001.

Zhou, B., Du, J., 2010. Fog prediction from a multimodel mesoscale ensemble prediction system. Weather Forecast. 25 (1), 303–322. https://doi.org/10.1175/2009waf2222289.1.