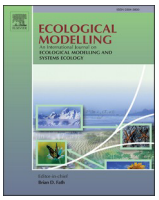
Ecological Modelling 471 (2022) 110059

Contents lists available at ScienceDirect

Ecological Modelling

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



A standard protocol for describing the evaluation of ecological models

Benjamin Planque [a,*], Johanna M. Aarflot [b], Lucie Buttay [a], JoLynn Carroll [c,d], Filippa Fransner [e], Cecilie Hansen [b], Bérengère Husson [b], Øystein Langangen [f], Ulf Lindstrøm [a,d], Torstein Pedersen [d], Raul Primicerio [a,d], Elliot Sivel [a,d], Morten D. Skogen [b], Evelyn Strombom [g], Leif Christian Stige [f], Øystein Varpe [h,i], Nigel G. Yoccoz [d]

[a] Institute of Marine Research, Tromsø, Norway
[b] Institute of Marine Research, Bergen, Norway
[c] Akvaplan-niva, Tromsø, Norway
[d] UiT the Arctic University of Norway, Tromsø, Norway
[e] Geophysical Institute, University of Bergen, and Bjerknes Centre for Climate Research, Norway
[f] University of Oslo, Oslo, Norway
[g] University of Minnesota, Minneapolis, United States of America
[h] Norwegian Institute for Nature Research, Bergen, Norway
[i] University of Bergen, Bergen, Norway

ARTICLE INFO

Keywords:
Standardisation
Best practice
Ecological patterns
Skill assessment
Transparency

ABSTRACT

Numerical models of ecological systems are increasingly used to address complex environmental and resource management questions. One challenge for scientists, managers, and stakeholders is to appraise how well suited these models are to answer questions of scientific or societal relevance, that is, to perform, communicate, or access transparent evaluations of ecological models. While there have been substantial developments to support standardised descriptions of ecological models, less has been done to standardise and to report model evaluation practices. We present here a general protocol designed to guide the reporting of model evaluation. The protocol is organised in three major parts: the *objective(s)* of the modelling application, the ecological *patterns* of relevance and the *evaluation* methodology proper, and is termed the OPE (objectives, patterns, evaluation) protocol. We present the 25 questions of the OPE protocol which address the many aspects of the evaluation process and then apply them to six case studies based on a diversity of ecological models. In addition to standardising and increasing the transparency of the model evaluation process, we find that going through the OPE protocol helps modellers to think more deeply about the evaluation of their models. From this last point, we suggest that it would be highly beneficial for modellers to consider the OPE early in the modelling process, in addition to using it as a reporting tool and as a reviewing tool.

1. Introduction

Scientists, managers, and stakeholders increasingly rely on numerical models of ecological systems. One challenge is to appraise the efficiency of these models to tackle complex environmental questions. Providing clear evaluations of model performance is one way to address this challenge. Models can be constructed, analysed, and used by different actors, from scientists to policymakers, and these actors have different understandings and expectations from models. Assessing how good a model is at addressing specific problems is difficult when ecological modellers use a variety of model types, have different modelling cultures and practices, and use different vocabularies. This can hinder communication, transparency, reproducibility, and the general development of good practices within the modelling community. It is therefore essential to provide tools to support a collective understanding of what can be expected from a model and how a model is to be evaluated (Cartwright et al., 2016; Eker et al., 2018; Heymans et al., 2020).

Transparency and reproducibility are at the core of the scientific method. However, the complexity of the tools used to observe and model ecological systems challenges reproducibility and transparency (Powers and Hampton 2019). The ongoing so-called reproducibility or

* Corresponding author at: Institute of Marine Research, P.O. Box 6606, 9296 Tromsø, Norway.
E-mail address: benjamin.planque@hi.no (B. Planque).

https://doi.org/10.1016/j.ecolmodel.2022.110059
Received 25 January 2022; Received in revised form 16 June 2022; Accepted 19 June 2022
Available online 21 July 2022

replicability crisis reflects this difficulty. The crisis has primarily been identified in the fields of psychology (Pashler and Wagenmakers 2012), clinical studies (Begley and Ioannidis John 2015) and economics (Camerer et al., 2016) and is much less discussed in ecological research (but see Ives 2018; Nichols et al., 2019, 2021). It may not be possible to strictly replicate ecological observations, but transparency in workflow and data analyses can facilitate reproducibility. It should be possible to reproduce ecological model simulations, given that the relevant information is provided for that purpose. In addition to replicating a model and the associated simulations, it is equally important to be able to understand, assess and replicate how model performance was evaluated. This step is critical, given that almost every new method published claims to outperform existing ones, but are seldom re-evaluated (Boulesteix et al., 2020). Providing relevant and comprehensive information is a first step towards replicability, which needs to be complemented by appropriate communication and quality standards. How is the information communicated? Is it accessible? Is it unambiguous? Is it sufficient? A standardised protocol for reporting model evaluation procedures would address these questions and contribute to increased transparency and reproducibility of ecological models.

There have been considerable collective efforts in recent decades to develop standardized modelling practices, from model building to evaluation of model performances. A major advancement has been the development of standardised protocols such as the ODD (Overview, Design concepts, and Details, Grimm et al., 2006). The ODD protocol was originally developed to respond to the lack of a standard protocol for describing individual based models (IBMs). The protocol was reviewed and updated twice since its original publication (Grimm et al., 2010, 2020b) and it is now commonly used by ecological modellers, beyond the original IBM community, to describe their models in reports and publications. The ODD protocol has been inspirational to groups of modellers with diverse focus, such as on model optimisation (ODDO, Mahévas 2019), data-mapping (ODD+2D, Laatabi et al., 2018), and inclusion of human decisions (ODD+*D*, Müller et al., 2013). In each case these groups have borrowed from the original ODD protocol idea and extended it for their specific purpose, thereby contributing to the harmonisation and communication of modelling practices.

A major step in the development and application of ecological models is the evaluation phase. There exists a large body of literature on how to perform model evaluation for various classes of models (e.g. Stow et al., 2009; Allen and Somerfield 2009; Bennett et al., 2013; Conn et al., 2018; Hipsey et al., 2020), but much less work has been done to standardise the reporting of model evaluations. The TRACE (TRAnsparent and Comprehensive Ecological modelling) documentation (Grimm et al., 2014) is a notable exception which provides a framework for documenting the modelling process, including several aspects of model evaluation. Standardised protocols for reporting model evaluation can constitute useful tools for modellers and end-users to easily understand and compare evaluation procedures and appreciate the performance of models in relation to specific objectives. Making such tools available is therefore anticipated to benefit the scientific community and model end-users.

The issue of model validation and evaluation in environmental science has been the subject of extensive research and debate. Oreskes (1998) argued that quantitative models cannot be validated but only evaluated. In Oreskes' view, evaluation is described as "an assessment in which both positive and negative results are possible, and where the grounds on which a model is declared good enough are clearly articulated". This assessment implies an examination of model outputs against pre-specified performance criteria. In the literature, the term *model validation* has remained pervasive (Eker et al., 2019) although often overlapping with the concept of evaluation as originally presented by Oreskes. In their 10-step procedure for developing and evaluating environmental models, Jakeman et al. (2006) introduced a stepwise approach in which every stage is open to critical review and revision, in consort with end-users. The evaluation step is left to the end and is concerned with the model being *fit for purpose*, although the criteria for achieving this goal are not fully developed by these authors. More recently, Parker (2020) explores the meaning of a model being *adequate for purpose* for different classes of models, whether pedagogical, explanatory or predictive. In the works of Jakeman and Parker, model evaluation is primarily achieved by measuring the performance of a model against pre-specified objectives, thereby following the original argument of Oreskes. This excludes the idea of a *general validity* of a model and favours the principle of an evaluation of a model for a specific objective (or a set of objectives). This mirrors George Box's notorious statement that "*all models are wrong; the practical question is how wrong do they have to be to not be useful*" (Box and Draper 1987), where *useful* implies use and therefore purpose. This is also in line with Augusiak's review of the literature on model evaluation and validation which concludes that despite little agreement on terms and underlying notions in the literature, it has repeatedly been pointed out that the evaluation of a model should depend on its *purpose* (Augusiak et al., 2014).

Evaluating that an ecological model is fit for purpose implies that the same model can (and should) be evaluated each time it is used for a new purpose. This is a rather trivial implication of the *fit for purpose* evaluation, however examples of re-evaluation of the performance of complex ecological models are scarce. Complex ecological models require extensive development efforts, and these materialise in the first publication of the model, together with a global evaluation (or validation) of the model (see e.g., Radach and Moll 2006; Link et al., 2010; Travers-Trolet et al., 2014; Pedersen et al., 2021). A fit-for-purpose approach would require that this first model evaluation be revised and reported for each new application of the model. One challenge in doing so is that the task of reporting model evaluation, which is already substantial when the model is first published, may seem daunting if it is to be repeated for every new model application. This can possibly be eased by reporting primarily on aspects of the model evaluation that are specific to each new application. An additional help can be provided by following a template in which a set of questions can guide the modeller through the reporting process.

By taking inspiration from the success and utility of the ODD protocol and the following extensions, we here present a complementary protocol for the reporting of ecological model evaluation procedures: the OPE (Objectives, Patterns, Evaluation) protocol. We discuss the rationale for the different elements of this protocol and provide a list of questions that can guide modellers to report in OPE format. We summarise the protocol (Table 1) and provide an easy-to-use Word template to support documenting model evaluations. (Supplementary material S1). Finally, we test the protocol on six case studies taken from a collection of marine ecosystem models with which the authors are familiar. These case studies are presented in detail in the supplementary material (S2). These modelling applications pre-existed the OPE protocol. The OPE has therefore not been used to guide the model evaluations presented here, but only to report how these evaluations were performed.

2. Elements of the OPE protocol

The elements of the OPE protocol are divided into three sections: Objectives, Patterns and Evaluation. Each section is then divided in subsections which contain one to six questions.

2.1. Objectives

2.1.1. Context and motivations

In our experience, many ecological models are not developed with the sole purpose of answering a single, well circumscribed question. Rather, complex ecological models take time to develop, are often built to address multiple, and sometimes diffuse, purposes and are gradually applied to a range of questions (Fulton et al., 2011; Planque and Mullon 2020). For example, dynamic global vegetation models (DGVM) which

were originally conceived to assess ecosystem-level responses to atmospheric-CO_2 concentration (Prentice et al., 2000) have later been applied to deliver projections based on future climate scenarios (Sitch et al., 2008) and are being continuously developed to address new questions, such as the impacts of different management practices on terrestrial ecosystems (Prentice et al., 2007). In a similar fashion, the ecosystem model Atlantis developed for the northwest Atlantic shelf (Link et al., 2010) was first used to explore the combined effects of climate and fishing (Nye et al., 2013), then to address the impact of ocean acidification (Fay et al., 2017) and more recently to quantify combined effects of acidification, fishing and marine protection (Olsen et al., 2018). Models of natural systems are inevitably embedded with multiple sources of uncertainty, and modellers make decisions during model construction (e.g., on which processes to include or simplify) which will affect the final outcome (Babel et al., 2019). There is a risk that assumptions which are reasonable for one particular model application are inadequate for another (Parker 2020; Saltelli et al., 2020). It is therefore essential that model suitability and performance are assessed and described for each application, and a crucial first step is describing the purpose of the specific application. In other words, to evaluate that a model is fit for purpose one must first specify the purpose. In this contribution, we refer to *model* as the generic description of the modelling tool (e.g. Ecopath with Ecosim, Polovina 1984a, 1984b; Christensen and Walters 2004), we use the terms *goal, purpose* and *objective* in an interchangeable manner to express the motivation driving the study and we refer to a model *application* when the model is applied towards a pre-defined objective or set of objectives. Central in the OPE framework is our conception that it is sensical to evaluate the same model against different patterns or data when applied for different purposes.

Describing the main objectives of the study and how modelling will contribute to reach these objectives is perhaps the most crucial step in evaluating model performance and suitability, and it should be a key reference point throughout the evaluation process. Without a clear understanding of the purpose, it becomes difficult to communicate credibility and generate trust in the modelling work. Furthermore, it may be sensible to evaluate the same model against quite different patterns or data when applied for different purposes. Defining the aims and objectives of the model application early in the research process can save time, for instance with the realisation that objectives may depend on key processes for which the model of choice lacks functionality.

The aims and objectives of a model application should be stated in simple, clear language. We suggest using active sentences (e.g., construct, produce, test, document) and avoid vague wordings (e.g., explore, study, investigate). Beware that ambiguity in the description of the purpose of a model often leads to multiple (subjective) interpretations of whether an outcome was successful or not (Parker 2020). This hinders a reliable evaluation process. The following questions guide the reporting of objectives:

1 What are the objectives of the model application?
2 Why is the model suitable to address the objectives?
3 What would count as successful in achieving these objectives?

2.1.2. Specific model setup

Ideally, the ecological model has already been fully described following a standardized protocol such as the ODD. It is possible that the original description is adequate for a new application of the model, but specific applications may also require adjustments of the model structure, parameters, or assumptions. Assumptions are particularly important to report when the model is used to perform predictions at other points in time and space, which requires that the model has some degree of transferability (Wenger and Olden 2012; Yates et al., 2018). This is the case when the objective of the model is to produce forecasts or to predict ecosystem properties in one region based on a model developed in another. It is wise to explicitly state what lies behind the often-implicit assumption of *ceteris paribus* (everything else being equal). For example, are trophic interactions assumed to follow the same rules in different regions? Are spatial distributions or environmental conditions assumed to be unchanged in the future? When models are used for conditional forecasting, one should also report assumptions about expected changes that can affect the system studied. For example, how are possible future changes in water temperature, fishing effort, accidental oil spill or increase in noise due to shipping represented in the model? A model can be revised to better reproduce the ecological components or processes that are relevant to a new application. It is also possible that revised model structure, estimates of input parameters or new data on the forcing conditions of the model become available. All these updates should be reported in this section which describes any changes or additions which have been made since the original model description.

4. Are there any deviations from the original model description?
   a In the model assumptions,
   b in the model structure (e.g., addition of submodels, variables, components, modifications of spatial or temporal scales),
   c in the model details (e.g., changes in parameter values, functional relationships),
   d in the model forcing (e.g., initial conditions, boundary conditions, forcing time series and maps).

2.2. Patterns

2.2.1. Selected patterns

A *pattern* may be defined as a characteristic and clearly identifiable structure in nature, or in data extracted from nature (e.g., population cycles, animal space use, species diversity etc.), that can be attributed to a generative process (Levin 1992; Grimm et al., 1996). Thus defined, a pattern is key to ecological understanding and prediction. Ecological patterns emerge from multiple ecological processes, which operate at multiple spatial and temporal scales and levels of organization (individual, population, community, and ecosystem). Understanding the causal mechanisms responsible for pattern formation is a primary goal of ecology (Levin 1992).

Modelling complex adaptive systems (see Levin 1992), such as marine ecosystems, is challenging, but pattern-orientated modelling (POM) may facilitate the task (Grimm et al., 1996, 2005; Grimm and Railsback 2012). POM "starts with identifying a set of patterns observed at multiple scales and levels that characterize a system with respect to the particular problem being modelled" (Grimm and Railsback 2012). In other words, the selection of patterns to be used in model evaluation, depends on the objective(s) or hypothesis of the study.

Relevant ecological patterns may be related to numbers, biomass, production, or consumption of relevant ecological entities, to dynamic behaviour at equilibrium, or to character of state transitions in perturbation studies or in systems undergoing change (e.g. Beisner et al., 2003). Other examples are spatial patterns such as spatial synchrony or travelling waves (e.g. Sherratt and Smith 2008). More complex emerging patterns (e.g., spatial structure described by a variogram, degree of spatial overlap between species) may also be candidate targets for model evaluation. The selection of specific patterns is motivated by the objectives of the modelling application and is generally driven by the hypotheses that can explain the emergence of these patterns. As pointed by Cury et al. (2008), it might be relatively easy to reproduce a single ecological pattern with all kinds of alternative models, but simultaneously reproducing an entire set of patterns is much more demanding and requires that the model is *structurally realistic*. Rather than tying a model to a specific pattern, via heavy calibration, it can be more useful consider several *weak patterns* at the same time - because then the risk that we force the model to look right, but for the wrong reasons, is reduced. This is particularly true in the case of complex ecosystem models which include many processes and parameters that can be adjusted to tune the model to few selected outputs. While some patterns

may be used to inform the model construction (e.g., some empirical relationships between ecological variables), other are emergent properties of the model. Model evaluation based on these emergent patterns may be of greater interest since models that succeed in getting emergent patterns right may also have greater potential for transferability to other time, place or systems (Radchuk et al., 2019). It is therefore critical to report on the selection of patterns and on the justification for this selection.

5. Which ecological patterns are used for the model evaluation?
   a temporal patterns such as cycles, regime shift or trends, measures of temporal variability, and autocorrelation.
   b spatial patterns such as spatial synchrony, travelling waves, patchiness, and autocorrelation.
   c structural and functional patterns, such as taxonomic diversity, biomass ratios, integrated production, diet fractions, and trait distributions.
   d Other relevant patterns
6. Why are these patterns important/essential to address the objectives?

In the following part of the OPE one must describe the data used for evaluation purposes, which can include both data from the model output and data which are independent of the model. Information on data used for model building should be provided in the model description (typically, an ODD protocol) and data used for optimization should be reported in the optimization description (e.g. in an ODDO protocol, Mahévas 2019).

2.2.2. Independent data

Independent data – that is data that exists independently of the model being built – are often derived from field observations, and procedures for collecting and processing these observations should briefly be summarized in this part of the OPE. Relevant information includes i) whether the data originate from a dedicated field survey, an open database, or another model, ii) the spatial/ temporal/ taxonomic/ etc. extent and resolution of the data, iii) data representativeness, and iv) accuracy, precision, bias, or uncertainty. Data representativeness is the degree to which data can be used to represent the ecological patterns that are relevant for the objective of the study. For example, daily, weekly, or monthly time-series will have different representativeness if the ecological pattern of interest is related to phenology. Similarly, the representativeness of data collected at a single sampling station is also expected to vary with the spatial scale of the ecological question of concern, being more representative for small scale modelling studies centred around the sampling station than for larger scale investigations. Deriving ecological patterns (Section 2.2.1) from observations can involve extensive data processing, and this should be reported here. When the same type of data can be used for model optimisation and evaluation (as in cross-validation) this should be reported in this section. In some cases, although the data is collected independently of the model being built, the model and data may not be completely independent from each other (for example, knowledge from historical data used to build the model, or input data in an Ecopath model is also expressed as an output of the model) and this should be reported. The following questions guide the collection of information about the independent data used to evaluate the model, given selected pattern(s).

7. Where do the independent data originate from? (e.g. field survey, open database, another model, …)
8. What are the extent and resolution of the independent data? (spatially, temporally, taxonomically, …)
9. How representative of the ecological process are the independent data?
10. Are there estimates of independent data accuracy, precision, bias, or uncertainty?
11. How are the independent data processed to represent the selected patterns? Are assumptions made to derive these patterns from the data?

2.2.3. Model outputs

Often, only parts of the model outputs are used in a specific application and the aim of this section is to describe which outputs have been used and evaluated. In some cases, the data may be post-processed (e.g., aggregation of results by guild, geographical region, or integration in time). The purpose of post-processing can be to generate indicators of the relevant patterns (ex. species spatial distribution, biomass ratios, index of seasonality, see Section 2.2.1) or to generate model outputs that are comparable with independent data (Section 2.2.2). The post processing step can require new assumptions (e.g., assume that conversion rates such as C:Chla are constant in time/space/taxa). The aim of this section is to describe the selection of model outputs, the post-processing operations, and to report on quality, quantity, representativeness, uncertainties, or potential bias in the model outputs.

12. Which model outputs are used for the evaluation?
13. Have the outputs been post-processed, and how?
14. Are there estimates of model outputs accuracy, precision, bias, or uncertainty?
15. Are additional assumptions made when deriving patterns from model outputs?

2.3. Evaluation

2.3.1. Evaluation methodology

We refer here to the evaluation method applied in the context of a specific application of a model to address stated objectives (Section 2.1.1). Model verification (sensu Gräbner 2018) - the act of testing whether the model does what it is supposed to do, i.e., that it is technically functional - should precede any application of the model and is not considered here. A first model evaluation step is often to conduct *sanity checks*. These are rapid explorations of the model outputs which ensure that, even though the model is technically functional, it is not behaving poorly. Sanity checks are often non-quantitative and based on domain knowledge rather than on quantitative comparisons of observations vs. model outputs. Though these are not often reported in model evaluation procedures, they inform about key conditions that the model must satisfy to be considered useful. Examples of sanity checks can include an inspection that population sizes or biomasses are within plausible ranges, that seasonal patterns are plausible or that emerging spatial patterns are visually credible. These can be done via Fermi estimations, often referred to as *back of the envelope* calculations of plausible ranges. Sanity checks are often performed in an unformal way and the intention of this section is to clarify and document this step. In cases when no sanity checks are performed, this should be justified.

16. Are sanity checks conducted? If so, what is the method used? If not, explain why.
    a Which data and patterns are used for this?
    b Does this apply to patterns that are not otherwise evaluated for this model application?

The core of the evaluation process is the comparison of patterns emerging from model outputs against those obtained from independent observations. This first raises the issue of the comparability between independent observations and model outputs, i.e., whether model outputs and independent data are directly comparable and whether modelled patterns are directly comparable to observed patterns. For example, are modelled biomass integrated over a large continuous geographical domain comparable with biomass field observations from a limited number of sampling sites? The second issue is the methodology used to compare ecological patterns derived from observations to those

derived from the model. There can be many methodological approaches, ranging from qualitative visual comparisons to fully quantitative estimates of the model performance at reproducing observed patterns (Allen and Somerfield 2009; Bennett et al., 2013). The latter can include univariate or multivariate approaches, and can be based on error-based measures, information theory measures, parametric tests, non-parametric tests, distance-based measures, and combined measures (Hora and Campos 2015). This stage of the evaluation is sometimes referred to as *skill assessment*.

The choice of methods and metrics used in model skills evaluation will depend on the relevant patterns. For example, when dealing with cycles, the degree of congruence between modelled and observed cycles amplitude and frequency should be reported. When modelling state transitions, agreement in the rate of change of a trend should be reported. With ecosystem models addressing ecological stability or temporal variability, the stability measure should be reported at multiple levels of organisation (e.g., species, functional group, community etc.). The quantitative criteria to evaluate the match between observed and simulated patterns must be reported. For example, if the mean of the simulations is within a certain range (e.g. 1 standard deviation) of the observed pattern, the model satisfactorily addresses the pattern (e.g. Kramer-Schadt et al., 2007). The selection (or lack of selection) of particular skill assessment methods can also be partially dictated by existing skills, available software or discipline culture and habits. Some evaluation methods may have been tried without success. In those cases, one should report on the attempted evaluation steps with some discussion on how and why these were deemed unsuccessful.

Each methodology usually comes with associated assumptions that need fulfilling for the method to be valid, and these should also be reported here.

The core issue at the end of the evaluation process is whether the model outputs can be considered satisfying for the purpose of answering the modelling objective, i.e., that the grounds on which a model is declared good enough are clearly articulated (Oreskes 1998).

17. What is the methodology used to compare ecological patterns derived from independent data with patterns derived from the model?
    a What is the rationale for choosing this method?
    b How are observational and/or model output uncertainties handled?
    c Does the methodology rely on specific assumptions?
    d Were other methods experimented? If they didn't succeed, explain why.
18. Is there a threshold level (in the match between observed and modelled patterns) that can separate acceptable from unacceptable models?
19. How comparable are the patterns derived from the model and those derived from the independent data?

By answering the above questions, researchers should also discuss if there are patterns that cannot be well evaluated with the chosen method.

2.3.2. Sensitivities

We distinguish between two types of sensitivities to be reported. First, *model sensitivity* which is the result of a sensitivity analysis (SA), usually performed on model structure and parameters. Second, *evaluation method sensitivity*, which refers to the sensitivity of the model evaluation to the choice of evaluation methodology and available observational data.

Sensitivity analysis scrutinizes how variations in model inputs influence variations in model outputs, a fundamental step in model evaluation and corroboration (EPA 2009). A sensitivity analysis informs about which input parameters the model is most sensitive to (and therefore which parameters should be obtained with greater precision and accuracy), and about the relative importance of processes in the model. A diverse array of SA approaches has been developed to help cope with the various needs dictated by differing model assumptions, computational complexity, and availability of relevant information (Saltelli et al., 2004; EPA 2009). Reviews and guidelines for best SA practice in the context of ecological and environmental modelling are an important aid to SA planning, implementation, and reporting (Saltelli et al., 2004; A. 2021; EPA 2009; Thiele et al., 2014; Pianosi et al., 2016).

Attributes of SA methods worth considering in reporting include: independence of model linearity and additivity assumptions, ability to address interaction effects amongst input factors, capacity to handle differences in scale and shape of input probability distribution functions, ability to deal with differences in input spatial and temporal dimensions, and capacity to evaluate the effect of an input while all other inputs are allowed to vary as well (Frey 2002; Saltelli et al., 2004).

In this section, one should consider the sensitivity of the model outputs that are relevant to the objective of the study i.e., the modelled *patterns* (Section 2.2.3). Priority should be given to reporting sensitivity analyses that were conducted specifically for the model application. Sensitivity analyses performed in earlier stages of model development can be reported if also relevant for the objective(s) of the study.

20. Has a model sensitivity analysis been performed? If so, how? If not, explain why.
    a on the model structure?
    b on the model parametrization?
    c on other aspects of the model?
21. Which elements are the modelled patterns most sensitive to?
    a input parameters
    b priors and assumptions
    c structural elements
    d processes
22. How sensitive are the modelled patterns to the choice of initial conditions, boundary conditions, spatial and temporal resolution?

While there is no perfect model to address a specific ecological question, there is no perfect method either to evaluate the performance of a model (Makridakis et al., 2020). Typically, the choice of the sensitivity analyses depends on the availability of observational data with which the model can be compared, on the computational requirements to perform certain types of model evaluation, on the availability of evaluation methodologies to the modellers, and on modellers skill sets. This section reports on the rationale and criteria for choosing a particular approach to evaluate the model performance. It stresses when the choices are dictated by the objectives of the study as opposed to computational constraints, lack of relevant information or other considerations. For example, models with complex architecture and high computational costs - two common features for ecosystem models (Steenbeek et al., 2021) - impose restrictions on the exploration of the parameter space. This in turns limits the scope for global SA and simultaneous exploration of known sources of uncertainty, which are two desirable features of SA. This section also reports on how sensitive the evaluation method is to the data used for evaluation (section 2.4). Could the model evaluation give significantly different results if supported by other/new/more precise data or if other skill assessment methods had been used? It is also the place where one can report failed attempts to evaluate the model or discuss possible future development in evaluation methodology. Alternative or complementary approaches to standard sensitivity analyses (e.g., robustness analysis, Thiele and Grimm 2015; Grimm and Berger 2016) can also be reported here. In summary, this section highlights the relevant attributes of the model evaluation, caveats, possible limitations, and possible developments, clarifying the performance of the model evaluation in relation to the objectives.

Table 1
The 25 questions of the OPE protocol, grouped into three headings: Objectives, Patterns and Evaluation. A brief comment accompanies each question to guide the reporting. A template form is provided in appendix S1, in which reporting can be directly entered.

		#	Question	*Comments*
OBJECTIVES	CONTEXT AND MOTIVATIONS	1	**What are the objectives of the model application?**	*Describe here the motivation and context for using the model. What is the purpose of the study? Do not describe the model, or its general objectives but focus on study-specific objectives. Use active sentences (e.g., produce, test, quantify, reconstruct dynamics) and avoid vague wordings (e.g., explore, study, investigate, understand).*
		2	**Why is the model suitable to address the objectives?**	*Provide the main rationale for why this specific model approach is suited to address the objective(s) raised in question 1. For example, is the model representing a process that is central to addressing the objectives?*
		3	**What would count as successful in achieving these objectives?**	*Explain here which criteria are used to determine if the model can address the objective or not. For example, if the objective of the model is to quantify a variable/process, is success defined based on the uncertainty around these estimated quantities?*
	SPECIFIC MODEL SETUP	4	**Are there any deviations from the original model description?** a **In the model assumptions?** b **In the model structure – submodels, variables, components, scales?** c **In the model details – parameter values, functional relationships** d **In the model forcing – initial conditions, boundary conditions, observation forcing, maps?**	*If this is the first time the model is presented, a full ODD description should be provided (Grimm et al., 2006, 2010, 2020b). If the model has already been presented elsewhere, only deviations from the original description should be provided here. Models are often adjusted to address a specific ecological question/objective. It is these adjustments that should be reported here.*
PATTERNS	SELECTED PATTERNS	5	**Which ecological patterns are used for the model evaluation?** a **Temporal patterns – cycles, shifts, trends, variability, autocorrelation** b **Spatial patterns – synchrony, travelling waves, patchiness, autocorrelation** c **Structural, functional patterns – diversity, biomass ratio, integrated production, diet, traits** d **Other relevant patterns**	*The term "ecological pattern" refers to Pattern-Oriented Modelling (POM, Grimm et al., 1996, 2005; Grimm and Railsback 2012). Relevant ecological patterns can be observed at various scales and characterize the ecological system with respect to the particular problem being modelled. The patterns listed in a, b, and c are by no mean required or exhaustive, but are provided as examples of possibly relevant patterns.*
		6	**Why are these patterns important/essential to address the objectives?**	*Explain here how the selection of ecological patterns is justified in relation to the objectives of the modelling application. Are there hypotheses that can explain the emergence of these patterns? Do not discuss how these patterns can be derived from observations or model outputs, this is addressed in questions 11–15.*
	INDEPENDENT DATA	7	**Where do the independent data originate from?**	*Independent data refers to data that exists independently from the current model being developed. These can be observational data or outputs from other models. Do not discuss outputs from the modelling study, these are addressed in questions 12–15.*
		8	**What are the extent and resolution of the independent data?**	*Report here the spatial, temporal, taxonomic extent and resolution of the independent data identified in question 7. For example, if a data series is presented, what are the starting and ending time and the time-frequency of data acquisition; if biodiversity data is provided, what is the taxonomic resolution and the method used to determine taxonomic units.*
		9	**How representative of the ecological process are the independent data?**	*This is a follow-up from question 8 to link data with key processes and patterns. For example, if a central process in the study is interannual variations in population numbers, and observational data of population numbers are available: do these data appropriately represent the annual abundance, or do they represent a snapshot in time or space? Do not report on uncertainty estimates here, this is addressed in question 10.*
		10	**Are there estimates of independent data accuracy, precision, bias, or uncertainty?**	*Uncertainty estimates for the independent data should be reported here (uncertainty estimates for the model outputs are reported in question 14).*
		11	**How are the independent data processed to represent the selected pattern? Are assumptions made to derive these patterns from the data?**	*Independent data – whether observational or modelled – may provide a representation of the patterns of interest (question 5) only after further processing. For example, survey data may be spatially interpolated to derive spatial distribution patterns. Another example: biomasses from several taxonomic units may be grouped to derive patterns of interannual changes in biomass for particular functional groups. Report these post-processing steps here.*
	MODEL OUPUTS	12	**Which model outputs are used for the evaluation?**	*This is a list of model outputs that have been selected based on the modelling objectives and related ecological patterns. The full set of raw outputs, which is often large, unprocessed, and not targeted towards the specific objectives of the modelling study, should not be reported here.*
		13	**Have the outputs been post-processed, and how?**	*As for independent data, model outputs may provide a representation of the patterns of interest only after further processing (see question 11). Report here the post-processing steps that are used to go from raw model outputs to ecologically relevant patterns.*
		14	**Are there estimates of model output accuracy, precision, bias, or uncertainty?**	*Uncertainty estimates for the model outputs should be reported here. Focus should be on model outputs that are used for the model evaluation.*
		15	**Are additional assumptions made when deriving patterns from model outputs?**	*Report here when some assumptions may be required to derive outputs at the appropriate scale or in the appropriate units. For example, a dry:wet-weight ratio may be assumed across species/seasons/areas to derive weight wet estimates (the relevant pattern) from dry weight (the model output).*

(*continued on next page*)

Table 1 (*continued*)

		#	Question	*Comments*
EVALUATION	EVALUATION METHODOLOGY	16	**Are sanity checks conducted? If so, what is the method used? If not, explain why.** a **Which data and patterns are used for this?** b **Does this apply to patterns that are not otherwise evaluated for this model application?**	*Sanity checks are informal steps that are taken throughout model development to ensure that the model is not behaving badly. They inform on key conditions that the model must satisfy to be considered useful. For example, checking that a population neither becomes extinct nor grows to unrealistic size.*
		17	**What is the methodology used to compare ecological patterns derived from independent data with patterns from the model?** a **What is the rationale for choosing this method?** b **How are observational and/or model output uncertainties handled?** c **Does the methodology rely on specific assumptions?** d **Were other methods experimented? If they didn't succeed, explain why.**	*This section describes how model outputs are evaluated against independent data. This is sometimes referred to as model "skill assessment". This section should describe the methodology used as well as the rationale for the choice of methods,* i.e., *how the methods relate to data, model outputs, objectives of the study, and relevant ecological patterns.*
		18	**Is there a threshold level (match between observed and modelled patterns) that can separate acceptable from unacceptable models?**	*When are the model outputs reliable enough to be used to answer the main question of the study? Answering this question is critical to evaluate when the model can address the main objective of the study. One should not discuss here the conclusions of the study, but only the skill level required to consider the model useful.*
		19	**How comparable are the patterns derived from the model and those derived from the independent data?**	*This section describes the result of the model skill assessment, plus any other qualitative features (patterns) that can be compared between model outputs and independent data.*
	SENSITIVITIES	20	**Has a model sensitivity analysis been performed? If so, how? If not, explain why.** a **on the model structure?** b **on the model parametrization?** c **on other aspects of the model?**	*This section describes the approach used to conduct model sensitivity analyses (SA), in a broad sense, from individual parameter SA to global SA. Various aspect of the methods used for SA can be reported here, including sensitivity to parameters, model structure, boundary/initial conditions, simulation design, and so on (see* e.g., *Pianosi et al., 2016).*
		21	**Which elements are the modelled patterns most sensitive to?** a **input parameters** b **priors and assumptions** c **structural elements** d **processes**	*If applicable, report here the results of the SA on parameters, model structure, processes, and assumptions.*
		22	**How sensitive are the modelled patterns to the choice of initial conditions, boundary conditions, spatial and temporal resolution?**	*If applicable, report here the results of the SA on the choice of initial conditions, spin-up time, boundary conditions, spatial and temporal resolution.*
		23	**How sensitive is the model evaluation to the independent data availability and uncertainty?**	*Could the model evaluation give significantly different results if other, new, or more precise data were used than those described in question 7?*
		24	**How much is the model evaluation constrained by computational or theoretical limits?**	*Models that are structurally simple and computationally fast can generally be explored through in-depth SA. It is more demanding to run appropriate SA on models that are structurally complex or that use substantial CPU resources to run. For some models, complexity & run time make SA non-achievable in practice. These issues should be reported here.*
		25	**How does the perceived performance of the model depend on the chosen evaluation methodology?**	*Could the model evaluation give significantly different results if another evaluation approach had been used (other than reported in question 17)?*

23. How sensitive is the model evaluation to availability and uncertainty of the independent data?
24. How much is the model evaluation constrained by computational or theoretical limits?
25. How does the perceived performance of the model depend on the chosen evaluation methodology?

3. OPE template

As a practical tool, we provide in Table 1 a summary of the OPE protocol which highlights the main sections of the protocol, the 25 questions as well as guidelines on how to answer them. We also provide in supplementary material (S1), a Word template that can be used to provide information relevant to a modelling study.

4. Applications

We provide in the supplementary material (S2) examples of applications of the OPE protocol in the context of six modelling applications:

1 an Individual Based Model (IBM) used to quantify uncertainties in the estimates of mean biomass of the copepod *Calanus finmarchicus* as a function of sampling design (Hjøllo et al., 2021),
2 a statistical food-web model used to quantify the association between capelin (*Mallotus villosus*) and its main two prey (krill and *Calanus* species) (Stige et al., 2018),
3 simulations from the Non-Deterministic Network Dynamics (NDND) model to assess the persistence of trophic controls in the Barents Sea (Sivel et al., 2021),
4 an Ecopath model to estimate trophic positions for ecological groups in the Barents Sea (Pedersen 2022),
5 the Nordic and Barents Seas Atlantis Model (NoBa) simulations to assess cumulative impact of fisheries and climate in the Norwegian and Barents Seas (Hansen et al., 2019), and
6 the reconstructions and predictions of selected physical and biogeochemical properties using the NorCPM1 model in the Barents Sea (Bethke et al., 2021).

These case studies cover a range of modelling practices, modelling tools and study objectives. Knowledge about context within which a model is developed and of the history of the model development is essential to understand the evaluation approach. We realise that the OPE case studies presented in this manuscript can be difficult to read without prior knowledge of each model context and history. In standalone modelling studies, model descriptions would normally be provided in full, but this is not the case here. To correct for this, we included introductory paragraphs that describe the models that were used in each case study and provide a brief history of the models, i.e., where they originate from and how they evolved to finally be used in the current case studies.

5. Discussion

The OPE protocol as we present it here is complementing other reporting protocols, in particular the ODD protocol and the extensions (e.g., ODDO, ODD+*D*), by focusing on the model evaluation. We argue that such a protocol can significantly contribute to improving model evaluation and can in general increase transparency and reproducibility of published models. Following Oreskes (1998); Augusiak et al., (2014); Edmonds et al., (2019); Grimm et al., (2020a); Parker (2020), and others, we contend that model evaluation is purpose-dependant and that a clear description of the purpose of a modelling application must be an integral part of the evaluation process, whether the model goal is pedagogical, explanatory or predictive.

Model evaluation is essential and should accompany all model studies. We have therefore developed the OPE protocol for model evaluation, which is generic enough to apply to a wide range of ecological modelling studies, from coupled physical-chemical-biological systems (NORWECOM.E2E, NorCPM1, Atlantis), to simpler models focussed on food-webs interactions (NDND, Ecopath, Gompertz). In our experience, most modellers consider their model as somewhat special (i.e., not like other models) and therefore presume that it would be difficult to evaluate models using a standardised protocol like the OPE. Indeed, we found that it was often work-demanding for modellers to answer the 25 questions of the OPE protocol. Through the six case studies, we identified several challenges in documenting the OPE. Documenting model evaluation is not a standard step in most modelling studies. Lack of experience and training in doing so made it a time-consuming and demanding task that required several iterations, and substantial amount of thinking and discussion. At times, the OPE exercise was perceived as too time-consuming, little rewarding in the short term and easy to postpone. It was often difficult to find the balance between providing informative answers and remaining concise. In several cases, it was not always obvious what was the right amount of contextual information required to inform readers about the model. The amount of evidence to be presented in support of OPE statements was also debated. When sensitivity analysis had been performed in earlier studies, it could be unclear how much this should be reported. At first sight, some questions appeared unclear or redundant, though these issues were usually resolved after some iterations. Some questions were also of little relevance for some of the model applications explored here. Nevertheless, it was possible to successfully apply the OPE protocol to each specific case study, despite the diverse collection of model types. We therefore anticipate that the protocol will be applicable to many ecological modelling studies.

The protocol can be used from the start of a modelling study, to guide model evaluation throughout the study. Though the primary motivation for this protocol was to construct a tool to help modellers reporting how they evaluated their models given specific objectives, we found that answering the protocol questions for the individual case studies led to additional discussions and reflections on model evaluation. In some instances, it was identified that additional evaluation steps could be taken or that some steps in the evaluation process could have been better specified. In the case of the Gompertz case study, documenting the OPE revealed that posterior predictive checks could have been considered to improve the evaluation. In the NDND case study, it was only after the OPE was documented that the issue of determining a threshold between acceptable and unacceptable models became clear. In the NoBa case study, it became apparent that many aspects of model evaluation for a complex end-to-end model like Atlantis, were still under-developed, and that the OPE could guide future work towards improved model evaluation methodology. In all case studies the OPE helped to clarify existing evaluation procedures and identify possible improvements. Had the OPE been available at the start of these studies, the model evaluation would likely have been conducted more thoroughly. A lesson learned from the exercise is that documenting the OPE is more easily done if modellers take relevant notes about model evaluation while developing their model, rather than leaving the OPE questionnaire to the end. This highlights the potential utility of the OPE to stimulate higher standard of model evaluation, in addition to its original goal of merely reporting how evaluation was conducted.

It is important to note that the OPE protocol goes far beyond model skill assessment. Assessing the prediction skill of ecological models has been the focus of recent literature (see e.g., Stow et al., 2009; Olsen et al., 2016; Steenbeek et al., 2021 and references therein). Skill assessment is an integral part of model evaluation and is clearly identified in the first part of the *Evaluation* section of the OPE protocol (questions 17–19). The OPE protocol expands beyond skill assessment by addressing issues related to objective, patterns, data, and sensitivity analyses and puts balanced focus on these different elements.

Documenting model evaluation is not yet standard practice. The 25 questions outlined in the OPE protocol are a guide to present an

extensive – but not exhaustive – description of a model evaluation. A full description of the evaluation is often too long to be included in the core part of a published manuscript. We advocate that the OPE documentation be presented as a technical supplement. By documenting the details of the model evaluation procedure, the OPE provides essential information for the peer-review of a modelling study and directly contributes to higher transparency. Even when not all OPE questions are answered, it makes sense to present an OPE. We encourage modellers to try the OPE protocol by using the word template (S1) and get help and inspiration from the answers provided in the six case studies (S2). We also encourage reviewers to use the OPE questions as a guide when evaluating modelling studies.

The current OPE template is qualitative, thus providing high flexibility in reporting, but makes the evaluation report hard to appraise or to enter in automated systems that prefer numbers over free text. Possible future developments of the OPE may focus on adding standardised evaluation metrics or standardised evaluation vocabularies that could be automatically populated while performing evaluation exercises This in turn would facilitate analyses and comparisons within and between models. Further development of the OPE might also include other aspects of model evaluation that were not explicitly addressed here, such as robustness analysis (Grimm and Berger 2016). The questionnaire structure could possibly be hierarchised to highlight questions that have the highest priority (e.g., questions 1, 2, 3 and 19), or it could eventually be formally linked to other existing tools like TRACE (Grimm et al., 2014; Ayllón et al., 2021).

As noted by Grimm et al., (2014), building a 'culture' of model reporting is about *doing all these things as well as you can because you know that peers and model clients are expecting you to; there is no point any more in complaining about "additional effort" for these things*. We recognise that we are not there yet. Promoting the OPE and similar documentation during the peer review process would help in getting this culture in place.

The current version of the OPE protocol is a work-in-progress. Model evaluation is complex and the development of tools for reporting how evaluation is conducted is not a simple problem. The case studies presented here all originate from high-latitude marine ecosystem modelling research, which reflects the expertise of the authors. Further applications of the OPE will show how much the experience gained from developing and applying the OPE protocol on these few examples can benefit other modelling approaches on other ecological system types. During the discussions that formed the basis for the current protocol, a central point was that modellers have various cultures, experiences, and practices when it comes to model evaluation. These points of view are not always easy to reconcile with each other. Further discussions based on the use of the protocol on a wider range of models are expected to lead to revisions of the OPE protocol in the future.

6. Conclusion

The OPE protocol is proposed as a tool to report the evaluation of ecological models. The reporting template is organised along 25 questions which make it easier and faster for modellers to report model evaluation. The OPE structure further promotes comprehensive reporting of the evaluation process, ranging from objectives, to data, skill assessment, and sensitivity analyses. Our experience is that structured reporting of model evaluation helps modellers to think more deeply about the evaluation of their models. From this last point, we suggest that it would be highly beneficial for modellers to consider the OPE early in the modelling process, in addition to using it as a reporting tool (as we have done here) and as a reviewing tool.

Funding

This work was funded by the Research Council of Norway through the project Nansen Legacy (RCN No. 276730).

CRediT authorship contribution statement

Benjamin Planque: conceptualization, synthesis and interpretation, manuscript preparation, editing and revision; **Johanna M. Aarflot**: conceptualization, synthesis and interpretation, manuscript preparation, editing and revision; **Lucie Buttay**: conceptualization, case study construction and interpretation, manuscript preparation and editing; **Filippa Fransner**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation, editing and revision; **Cecilie Hansen**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation, editing and revision; **Bérengère Husson**: conceptualization, case studies interpretation, manuscript preparation and editing; **Øystein Langangen**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation and editing; **Ulf Lindstrøm**: conceptualization, synthesis and interpretation, manuscript preparation, editing and revision; **Torstein Pedersen**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation and editing; **Raul Primicerio**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation, editing and revision; **Elliot Sivel**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation, editing and revision; **Morten D. Skogen**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation and editing; **Evelyn Strombom**: conceptualization, synthesis and interpretation, manuscript preparation and editing; **Leif Christian Stige**: conceptualization, case study construction, synthesis and interpretation, manuscript preparation and editing; **Øystein Varpe**: conceptualization, synthesis and interpretation, manuscript preparation and editing; **Nigel G. Yoccoz**: conceptualization, synthesis and interpretation, manuscript preparation, editing and revision

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.

Data Availability

No data was used for the research described in the article.

Acknowledgments

The authors thank Noel Keenlyside at the University of Bergen and Ina Nilsen at the Institute of Marine Research for their contribution during the first and second workshops that led to the development of the OPE protocol.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2022.110059.

References

Allen, J.I., Somerfield, P.J., 2009. A multivariate approach to model skill assessment. J. Mar. Syst. 76 (1), 83–94. https://doi.org/10.1016/j.jmarsys.2008.05.009.

Augusiak, J., Van den Brink, P.J., Grimm, V., 2014. Merging validation and evaluation of ecological models to "evaludation": a review of terminology and a practical approach. Ecol. Modell. 280, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009.

Ayllón, D., Railsback, S.F., Gallagher, C., Augusiak, J., Baveco, H., Berger, U., Charles, S., Martin, R., Focks, A., Galic, N., Liu, C., van Loon, E.E., Nabe-Nielsen, J., Piou, C., Polhill, J.G., Preuss, T.G., Radchuk, V., Schmolke, A., Stadnicka-Michalak, J., Thorbek, P., Grimm, V., 2021. Keeping modelling notebooks with TRACE: good for

you and good for environmental research and management support. Environ. Model. Softw. 136, 104932 https://doi.org/10.1016/j.envsoft.2020.104932.
Babel, L., Vinck, D., Karssenberg, D., 2019. Decision-making in model construction: unveiling habits. Environ. Model. Softw. 120, 104490 https://doi.org/10.1016/j.envsoft.2019.07.015.
Begley, C.G., Ioannidis John, P.A., 2015. Reproducibility in Science. Circ. Res. 116 (1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819.
Beisner, B.E., Haydon, D.T., Cuddington, K., 2003. Alternative stable states in ecology. Front. Ecol. Environ. 1 (7), 376–382. https://doi.org/10.1890/1540-9295(2003)001[0376:ASSIE]2.0.CO;2.
Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V, 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.
Bethke, I., Wang, Y., Counillon, F., Keenlyside, N., Kimmritz, M., Fransner, F., Samuelsen, A., Langehaug, H., Svendsen, L., Chiu, P.-G., Passos, L., Bentsen, M., Guo, C., Gupta, A., Tjiputra, J., Kirkevåg, A., Olivié, D., Seland, Ø., Solsvik Vågane, J., Fan, Y., Eldevik, T., 2021. NorCPM1 and its contribution to CMIP6 DCPP. Geosci. Model Dev. 14 (11), 7073–7116. https://doi.org/10.5194/gmd-14-7073-2021.
Boulesteix, A., Hoffmann, S., Charlton, A., Seibold, H., 2020. A replication crisis in methodological research? Significance 17 (5), 18–21. https://doi.org/10.1111/1740-9713.01444.
Box, G.E.P., Draper, N.R., 1987. Empirical Model Building and Response Surfaces. John Wiley & Sons, New York, NY.
Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351 (6280), 1433. https://doi.org/10.1126/science.aaf0918.
Cartwright, S.J., Bowgen, K.M., Collop, C., Hyder, K., Nabe-Nielsen, J., Stafford, R., Stillman, R.A., Thorpe, R.B., Sibly, R.M., 2016. Communicating complex ecological models to non-scientist end users. Ecol. Modell. 338, 51–59. https://doi.org/10.1016/j.ecolmodel.2016.07.012.
Christensen, V., Walters, C.J., 2004. Ecopath with Ecosim: methods, capabilities, and limitations. Ecol. Modell. 172, 109–139. https://doi.org/10.1016/j.ecolmodel.2003.09.003.
Conn, P.B., Johnson, D.S., Williams, P.J., Melin, S.R., Hooten, M.B., 2018. A guide to Bayesian model checking for ecologists. Ecol. Monogr. 88 (4), 526–542. https://doi.org/10.1002/ecm.1314.
Cury, P.M., Shin, Y.-J., Planque, B., Durant, J.M., Fromentin, J.-M., Kramer-Schadt, S., Stenseth, N.C., Travers, M., Grimm, V., 2008. Ecosystem oceanography for global change in fisheries. Trends Ecol. Evol. (Amst.) 23 (6), 338–346.
Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H., Squazzoni, F., 2019. Different modelling purposes. JASSS 22 (3), 6. https://doi.org/10.18564/jasss.3993.
Eker, S., Rovenskaya, E., Langan, S., Obersteiner, M., 2019. Model validation: a bibliometric analysis of the literature. Environ. Model. Softw. 117, 43–54. https://doi.org/10.1016/j.envsoft.2019.03.009.
Eker, S., Rovenskaya, E., Obersteiner, M., Langan, S., 2018. Practice and perspectives in the validation of resource management models. Nat. Commun. 9 https://doi.org/10.1038/s41467-018-07811-9.
EPA, 2009. Guidance on the development, evaluation, and application of environmental models. Technical Report. Environ. Protection Agency. Available from. https://www.epa.gov/sites/production/files/2015-04/documents/cred_guidance_0309.pdf.
Fay, G., Link, J.S., Hare, J.A., 2017. Assessing the effects of ocean acidification in the Northeast US using an end-to-end marine ecosystem model. Ecol. Modell. 347, 1–10. https://doi.org/10.1016/j.ecolmodel.2016.12.016.
Frey, H.C., 2002. Introduction to Special Section on Sensitivity Analysis and Summary of NCSU/USDA Workshop on Sensitivity Analysis. Risk Anal. 22 (3), 539–545. https://doi.org/10.1111/0272-4332.00037.
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 Fisheries 12 (2), 171–188. https://doi.org/10.1111/j.1467-2979.2011.00412.x.
Gräbner, C., 2018. How to relate models to reality? An epistemological framework for the validation and verification of computational models. Jasss 21 (3), 26. https://doi.org/10.18564/jasss.3772.
Grimm, V., Augusiak, J., Focks, A., Frank, B.M., Gabsi, F., Johnston, A.S.A., Liu, C., Martin, B.T., Meli, M., Radchuk, V., Thorbek, P., Railsback, S.F., 2014. Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. Ecol. Modell. 280, 129–139. https://doi.org/10.1016/j.ecolmodel.2014.01.018.
Grimm, V., Berger, U., 2016. Robustness analysis: deconstructing computational models for ecological theory and applications. Ecol. Modell. 326, 162–167. https://doi.org/10.1016/j.ecolmodel.2015.07.018.
Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., 2006. A standard protocol for describing individual-based and agent-based models. Ecol. Modell. 198 (1–2), 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023.
Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., Railsback, S.F., 2010. The ODD protocol: a review and first update. Ecol. Modell. 221 (23), 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019.
Grimm, V., Frank, K., Jeltsch, F., Brandl, R., Uchmański, J., Wissel, C., 1996. Pattern-oriented modelling in population ecology. Sci. Total Environ. 183 (1), 151–166. https://doi.org/10.1016/0048-9697(95)04966-5.
Grimm, V., Johnston, A.S.A., Thulke, H.-H., Forbes, V.E., Thorbek, P., 2020a. Three questions to ask before using model outputs for decision support. Nat. Commun. 11 (1), 4959. https://doi.org/10.1038/s41467-020-17785-2.
Grimm, V., Railsback, S.F., 2012. Pattern-oriented modelling: a "multi-scope" for predictive systems ecology. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367 (1586), 298–310. https://doi.org/10.1098/rstb.2011.0180.
Grimm, V., Railsback, S.F., Vincenot, C.E., Berger, U., Gallagher, C., DeAngelis, D.L., Edmonds, B., Ge, J.Q., Giske, J., Groeneveld, J., Johnston, A.S.A., Milles, A., Nabe-Nielsen, J., Polhill, J.G., Radchuk, V., Rohwader, M.S., Stillman, R.A., Thiele, J.C., Ayllon, D, 2020b. The ODD protocol for describing agent-based and other simulation models: a second update to improve clarity, replication, and structural realism. Jasss 23 (2). https://doi.org/10.18564/jasss.4259.
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.
Hansen, C., Nash, R.D.M., Drinkwater, K.F., Hjøllo, S.S., 2019. Management scenarios under climate change – a study of the nordic and barents seas. Front. Mar. Sci. 6, 668. https://doi.org/10.3389/fmars.2019.00668.
Heymans, J.J., Bundy, A., Christensen, V., Coll, M., de Mutsert, K., Fulton, E.A., Piroddi, C., Shin, Y.-J., Steenbeek, J., Travers-Trolet, M, 2020. The ocean decade: a true ecosystem modeling challenge. Front. Mar. Sci. 7, 554573 https://doi.org/10.3389/fmars.2020.554573.
Hipsey, M.R., Gal, G., Arhonditsis, G.B., Carey, C.C., Elliott, J.A., Frassl, M.A., Janse, J. H., de Mora, L., Robson, B.J., 2020. A system of metrics for the assessment and improvement of aquatic ecosystem models. Environ. Model. Softw. 128, 104697 https://doi.org/10.1016/j.envsoft.2020.104697.
Hjøllo, S., Hansen, C., Skogen, M., 2021. Assessing the importance of zooplankton sampling patterns with an ecosystem model. Mar. Ecol. Prog. Ser. 680, 163–176. https://doi.org/10.3354/meps13774.
Hora, J., Campos, P., 2015. A review of performance criteria to validate simulation models. Expert Syst. 32 (5), 578–595. https://doi.org/10.1111/exsy.12111.
Ives, A.R., 2018. Informative irreproducibility and the use of experiments in ecology. Bioscience 68 (10), 746–747. https://doi.org/10.1093/biosci/biy090.
Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21 (5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004.
Kramer-Schadt, S., Revilla, E., Wiegand, T., Grimm, V., 2007. Patterns for parameters in simulation models. Ecol. Modell. 204 (3–4), 553–556. https://doi.org/10.1016/j.ecolmodel.2007.01.018.
Laatabi, A., Marilleau, N., Nguyen-Huu, T., Hbid, H., Babram, M.A., 2018. ODD+2D: an ODD based protocol for mapping data to empirical ABMs. Jasss 21 (2). https://doi.org/10.18564/jasss.3646.
Levin, S.A., 1992. The problem of pattern and scale in ecology. EcologyEcology 73, 1943–1967. https://doi.org/10.2307/1941447.
Link, J.S., Fulton, E.A., Gamble, R.J., 2010. The northeast US application of ATLANTIS: a full system model exploring marine ecosystem dynamics in a living marine resource management context. Prog. Oceanogr. 87 (1–4), 214–234. https://doi.org/10.1016/j.pocean.2010.09.020.
Mahévas, S.P. 2019. Practical guide for conducting calibration and decision-making optimisation with complex ecological models. preprints **2019120249**. doi: 10.20944/preprints201912.0249.v1.
Makridakis, S., Spiliotis, E., Assimakopoulos, V, 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. Int. J. Forecast. 36 (1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014.
Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., Schlüter, M., Schulze, J., Weise, H., Schwarz, N., 2013. Describing human decisions in agent-based models – ODD + D, an extension of the ODD protocol. Environ. Model. Softw. 48, 37–48. https://doi.org/10.1016/j.envsoft.2013.06.003.
Nichols, J.D., Kendall, W.L., Boomer, G.S., 2019. Accumulating evidence in ecology: once is not enough. Ecol. Evol. 9 (24), 13991–14004. https://doi.org/10.1002/ece3.5836.
Nichols, J.D., Oli, M.K., Kendall, W.L., Boomer, G.S., 2021. Opinion: a better approach for dealing with reproducibility and replicability in science. Proc. Natl. Acad. Sci. 118 (7), e2100769118 https://doi.org/10.1073/pnas.2100769118.
Nye, J.A., Gamble, R.J., Link, J.S., 2013. The relative impact of warming and removing top predators on the Northeast US large marine biotic community. Ecol. Modell. 264, 157–168. https://doi.org/10.1016/j.ecolmodel.2012.08.019.
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.
Olsen, E., Kaplan, I.C., Ainsworth, C., Fay, G., Gaichas, S., Gamble, R., Girardin, R., Eide, C.H., Ihde, T.F., Morzaria-Luna, H.N., Johnson, K.F., Savina-Rolland, M., Townsend, H., Weijerman, M., Fulton, E.A., Link, J.S., 2018. Ocean futures under ocean acidification, marine protection, and changing fishing pressures explored using a worldwide suite of ecosystem models. Front. Mar. Sci. 5 (64) https://doi.org/10.3389/fmars.2018.00064.
Oreskes, N., 1998. Evaluation (not validation) of quantitative models. Environ. Health Perspect. 106 (Suppl 6), 1453–1460. https://doi.org/10.1289/ehp.98106s61453.
Parker, W.S., 2020. Model evaluation: an adequacy for purpose view. Philos. Sci. 87 (3), 457–477. https://doi.org/10.1086/708691.
Pashler, H., Wagenmakers, E.-J., 2012. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect. Psychol. Sci. 7 (6), 528–530. https://doi.org/10.1177/1745691612465253.

Pedersen, T., 2022. Comparison between trophic positions in the barents sea estimated from stable isotope data and a mass balance model. Front. Mar. Sci. 9, 813977 https://doi.org/10.3389/fmars.2022.813977.

Pedersen, T., Mikkelsen, N., Lindstrøm, U., Renaud, P.E., Nascimento, M.C., Blanchet, M.-A., Ellingsen, I.H., Jørgensen, L.L., Blanchet, H., 2021. Overexploitation, recovery, and warming of the barents sea ecosystem during 1950–2013. Front. Mar. Sci. 8, 732637 https://doi.org/10.3389/fmars.2021.732637.

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ. Model. Softw. 79, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008.

Planque, B., Mullon, C., 2020. Modelling chance and necessity in natural systems. ICES J. Mar. Sci. 77 (4), 1573–1588. https://doi.org/10.1093/icesjms/fsz173.

Polovina, J.J., 1984a. An overview of the Ecopath model. Fishbyte 2, 5–7.

Polovina, J.J., 1984b. Model of a coral reef ecosystem: I. The ECOPATH model and its application to French Frigate Shoals. Coral Reefs 3 (1), 1–11. https://doi.org/10.1007/BF00306135.

Powers, S.M., Hampton, S.E., 2019. Open science, reproducibility, and transparency in ecology. Ecol. Appl. 29 (1), e01822. https://doi.org/10.1002/eap.1822.

Prentice, I.C., Bondeau, A., Cramer, W., Harrison, S.P., Hickler, T., Lucht, W., Sitch, S., Smith, B., Sykes, M.T., 2007. Dynamic global vegetation modeling: quantifying terrestrial ecosystem responses to large-scale environmental change. In: Canadell, J. G., Pataki, D.E., Pitelka, L.F. (Eds.), Terrestrial Ecosystems in a Changing World. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 175–192. https://doi.org/10.1007/978-3-540-32730-1_15.

Prentice, I.C., Heimann, M., Sitch, S., 2000. The carbon balance of the terrestrial biosphere: ecosystem models and atmospheric observations. Ecol. Appl. 10 (6), 1553–1573. https://doi.org/10.1890/1051-0761(2000)010[1553:TCBOTT]2.0.CO;2.

Radach, G., and Moll, A. 2006. Review of three-dimensional ecological modelling related to the North Sea shelf system. Part II: model validation and data needs. *In* Oceanography and marine biology. An annual review., 1st edition. pp. 1–60. Available from https://www.taylorfrancis.com/chapters/mono/10.1201/9781420006391-4/review-three-dimensional-ecological-modelling-related-north-sea-shelf-system-part-ii-model-validation-data-needs-gibson-atkinson-gordon?context=ubx&refId=41590f7c-dd00-46e8-84b1-4d56031bd254.

Radchuk, V., Kramer-Schadt, S., Grimm, V., 2019. Transferability of mechanistic ecological models is about emergence. Trends Ecol. Evol. (Amst.) 34 (6), 487–488. https://doi.org/10.1016/j.tree.2019.01.010.

Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Nelson Espeland, W., Kay, J., Lo Piano, S., Mayo, D., Pielke Jr, R., Portaluri, T., Porter, T.M., Puy, A., Rafols, I., Ravetz, J.R., Reinert, E., Sarewitz, D., Stark, P.B., Stirling, A., van der Sluijs, J., Vineis, P., 2020. Five ways to ensure that models serve society: a manifesto. Nature 582 (7813), 482–484. https://doi.org/10.1038/d41586-020-01812-9.

Saltelli, A., Jakeman, A., Razavi, S., Wu, Q., 2021. Sensitivity analysis: a discipline coming of age. Environ. Model. Softw. 146, 105226 https://doi.org/10.1016/j.envsoft.2021.105226.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in practice. A Guide to Assessing Scientific Models. John Wiley and Sons, Chichester, New York.

Sherratt, J.A., Smith, M.J., 2008. Periodic travelling waves in cyclic populations: field studies and reaction–diffusion models. J. R. Soc., Interface 5 (22), 483–505. https://doi.org/10.1098/rsif.2007.1327.

Sitch, S., Huntingford, C., Gedney, N., Levy, P.E., Lomas, M., Piao, S.L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C.D., Prentice, I.C., Woodward, F.I., 2008. Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs): UNCERTAINTY IN LAND CARBON CYCLE FEEDBACKS. Glob. Chang. Biol. 14 (9), 2015–2039. https://doi.org/10.1111/j.1365-2486.2008.01626.x.

Sivel, E., Planque, B., Lindstrøm, U., Yoccoz, N.G., 2021. Multiple configurations and fluctuating trophic controls in the Barents Sea food-web. PlosOne 16 (7), e0254015. https://doi.org/10.1371/journal.pone.0254015.

Steenbeek, J., Buszowski, J., Chagaris, D., Christensen, V., Coll, M., Fulton, E.A., Katsanevakis, S., Lewis, K.A., Mazaris, A.D., Macias, D., de Mutsert, K., Oldford, G., Pennino, M.G., Piroddi, C., Romagnoni, G., Serpetti, N., Shin, Y.-J., Spence, M.A., Stelzenmüller, V., 2021. Making spatial-temporal marine ecosystem modelling better – a perspective. Environ. Model. Softw., 105209 https://doi.org/10.1016/j.envsoft.2021.105209.

Stige, L.C., Kvile, K.Ø., Bogstad, B., Langangen, Ø., 2018. Predator-prey interactions cause apparent competition between marine zooplankton groups. Ecology 99 (3), 632–641. https://doi.org/10.1002/ecy.2126.

Stow, C.A., Jolliff, J., McGillicuddy Jr, 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.

Thiele, J.C., Grimm, V., 2015. Replicating and breaking models: good for you and good for ecology. Oikos 124 (6), 691–696. https://doi.org/10.1111/oik.02170.

Thiele, J.C., Kurth, W., Grimm, V., 2014. Facilitating parameter estimation and sensitivity analysis of agent-based models: a cookbook using NetLogo and "R. JASSS 17 (3), 11. https://doi.org/10.18564/jasss.2503.

Travers-Trolet, M., Shin, Y.-J., Field, J., 2014. An end-to-end coupled model ROMS-N $_2$ P $_2$ Z $_2$ D $_2$ -OSMOSE of the southern Benguela foodweb: parameterisation, calibration and pattern-oriented validation. Afr. J. Mar. Sci. 36 (1), 11–29. https://doi.org/10.2989/1814232X.2014.883326.

Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol. Evol. 3 (2), 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., Fielding, A.H., Bamford, A.J., Ban, S., Barbosa, A.M., Dormann, C.F., Elith, J., Embling, C.B., Ervin, G.N., Fisher, R., Gould, S., Graf, R.F., Gregr, E.J., Halpin, P.N., Heikkinen, R.K., Heinänen, S., Jones, A.R., Krishnakumar, P.K., Lauria, V., Lozano-Montes, H., Mannocci, L., Mellin, C., Mesgaran, M.B., Moreno-Amat, E., Mormede, S., Novaczek, E., Oppel, S., Ortuño Crespo, G., Peterson, A.T., Rapacciuolo, G., Roberts, J.J., Ross, R.E., Scales, K.L., Schoeman, D., Snelgrove, P., Sundblad, G., Thuiller, W., Torres, L.G., Verbruggen, H., Wang, L., Wenger, S., Whittingham, M.J., Zharikov, Y., Zurell, D., Sequeira, A.M.M., 2018. Outstanding Challenges in the Transferability of Ecological Models. Trends Ecol. Evol. (Amst.) 33 (10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001.