



Quo Vadimus

A guinea pig's tale: learning to review end-to-end marine ecosystem models for management applications[†]

Isaac C. Kaplan^{1*} and Kristin N. Marshall²

¹Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2725 Montlake Boulevard E., Seattle, WA 98112, USA

²University of Washington School of Aquatic and Fisheries Sciences, PO Box 355020, Seattle, WA 98195, USA

*Corresponding author: tel: +1-206-302-2446; fax: +1-206-860-3394; e-mail: isaac.kaplan@noaa.gov

Kaplan, I. C., and Marshall, K. N. A guinea pig's tale: learning to review end-to-end marine ecosystem models for management applications. – ICES Journal of Marine Science, doi: 10.1093/icesjms/fsw047.

Received 10 June 2015; revised 3 March 2016; accepted 7 March 2016.

A shift towards ecosystem-based management in recent decades has led to new analytical tools such as end-to-end marine ecosystem models. End-to-end models are complex and typically simulate full ecosystems from oceanography to foodwebs and fisheries, operate on a spatial framework, and link to physical oceanographic models. Most end-to-end approaches allow multiple ways to implement human behaviours involving fishery catch, fleet movement, or other impacts such as nutrient loading or climate change effects. Though end-to-end ecosystem models were designed specifically for marine management, their novelty makes them unfamiliar to most decision makers. Before such models can be applied within the context of marine management decisions, additional levels of vetting will be required, and a dialogue with decision makers must be initiated. Here we summarize a review of an Atlantis end-to-end model, which involved a multi-day, expert review panel with local and international experts, convened to challenge models and data used in the management context. We propose nine credibility and quality control standards for end-to-end models intended to inform management, and suggest two best practice guidelines for any end-to-end modelling application. We offer our perspectives (as recent test subjects or “guinea pigs”) on how a review could be motivated and structured and on the evaluation criteria that should be used, in the most specific terms possible.

Keywords: Atlantis, best practices, end-to-end models, evaluation criteria, marine ecosystem models, model performance, peer review.

What are end-to-end models, and why is it complex to review them?

In recent years, the shift toward ecosystem-based management of marine resources (Pikitch *et al.*, 2004; McLeod and Leslie, 2009) has led to the development of new analytical tools that simultaneously consider multiple human impacts and multiple species. Unlike traditional single species fishery stock assessments (Maunder and Punt, 2013; Methot and Wetzel, 2013) used for tactical management such as setting annual quotas, ecosystem models are typically viewed as strategic tools for exploring qualitative patterns, conducting risk assessment, and ranking policy alternatives (Plagányi, 2007; Fulton *et al.*, 2014). Critically, these tools are intended to capture trade-offs between species, fisheries, and human uses that may occur given

future marine management actions or environmental conditions. Evaluation of these trade-offs across species and sectors is a central part of ecosystem-based management such as the Marine Strategy Framework Directive (European Commission, 2008) and US national ocean policy (Interagency Ocean Policy Task Force of the White House Council on Environmental Quality, 2010; Obama, 2010).

End-to-end marine ecosystem models (Travers *et al.*, 2007; Fulton, 2010; Rose *et al.*, 2010) are one type of modelling tool that simulates full ecosystems from oceanography to foodwebs and fisheries. These are spatially explicit simulation models that are forced by physical oceanographic models, and that include human actions such as fishery catch, fleet movement, nutrient loading, or climate change. Examples include the Atlantis model (Fulton *et al.*

[†]Intended for ICES Journal of Marine Science *Quo Vadimus*, “describing the future landscape/potential of a topic, issue, discipline, or technology”.

2011), OSMOSE (Shin and Cury, 2004), SEAPODYM (Lehodey *et al.*, 2008), and Ecopath/Ecosim/Ecospace models (Christensen and Walters, 2004), though each of these models involves different levels of complexity and focuses on particular aspects of the ecology and fishing.

End-to-end models differ from single species models in some key aspects that ultimately lead to different criteria for model review and application. Most end-to-end models do not estimate parameters internally, but instead typically require the analyst to obtain parameter estimates outside the model, build the model, then calibrate key parameters to obtain satisfactory model behaviour. Model run times are often long, approximately hours to days, and therefore a final parameter set that may be adequate is not necessarily globally optimal—searching the full parameter space is simply not computationally possible. Thus, the simulations provide one realization of ecosystem dynamics, but cannot claim to have the unique answer. Finally, ecology has no agreed-upon set of mathematics that dictates system dynamics, akin to the Navier-Stokes equations of physical oceanographers. The addition of ecological interactions into population dynamics leads to a series of choices regarding functional relationships for predation, movement, and growth, all of which have important implications for model dynamics (FAO, 2008; Rose *et al.*, 2010; Hunsicker *et al.*, 2011). Additionally, spatial and taxonomic resolution (lumping vs. splitting areas or species) is a necessary consideration in building end-to-end models, but can influence model dynamics (Fulton, 2004; Pinnegar *et al.*, 2005) via structural uncertainty. The net result of these properties of end-to-end models is that they have different purposes, behaviours, and uncertainties than those which have typically been used for marine fisheries management. Below, we draw on recent experience to address how, despite these challenging properties, end-to-end models can engage in a rigorous external peer review process to move the models into the management arena.

Motivation for reviewing ecosystem models

Though end-to-end ecosystem models have clear utility for marine management, these new purposes, behaviours, and aspects of uncertainty inherent in end-to-end models are unfamiliar to most decision makers. To provide advice within the context of US marine management, Link *et al.* (2010b) note that ecosystem models' "credibility will need to be established and the rigor of quality control/assurance and peer review will need to be at a comparable level as what is done for single species and protected species stock assessments." By "peer review," the authors imply multi-day, expert review panels with both local and international experts, convened to challenge models and data used in the management context. The same authors noted that peer review panels are also necessary if where expert judgment is required to weight alternative models that may be statistically incomparable (e.g. based on different data) (Link *et al.*, 2012). Collie *et al.* (2016) noted that such expert panels are necessary for evaluating alternate model forms and identifying the level of model complexity appropriate for specific management questions. Within ICES, the Working Group on Multispecies Assessment Methods has begun a procedure to review "key runs" (ICES, 2013) of multispecies models that can provide management advice.

The goal of holding expert review panels for marine ecosystem models is a relatively high bar. Link *et al.* (2010b) and Townsend *et al.* (2008) found that many US fishery ecosystem models were at most peer reviewed in refereed literature. Prominent exceptions include a suite of ecosystem models in the Northeast US and

Alaska, reviewed by the Center of Independent Experts in 2005 and 2011 (http://www.nefsc.noaa.gov/ecosys/modeling_review.html, <https://www.st.nmfs.noaa.gov/science-quality-assurance/cie-peer-reviews/cie-review-2005>), and periodic review of Alaskan ecosystem modelling by stock assessment teams and fishery management council subcommittees (Link *et al.*, 2010b). However, the established process for reviewing single species stock assessment models within fisheries management (e.g. NOAA Northeast Fisheries Science Center, 2014; Pacific Fishery Management Council 2012) can serve as a rough template for review of end-to-end models (Figure 1), and the decades of communication between managers and scientists regarding single species models serves as a reminder that much work is to be done by end-to-end modellers in this arena. Our perspective benefits from these previous and ongoing efforts.

Below, we summarize a review of an Atlantis end-to-end ecosystem model. The review was held in Seattle, Washington, USA, during June and July 2014. We describe this process to consider how such reviews should be structured and organized (Figure 1), and we propose a very specific set of model evaluation criteria for marine end-to-end models. These differ from previous sets of best practices provided by other marine and terrestrial ecosystem modelers (FAO, 2008; Townsend *et al.*, 2008; Link *et al.*, 2010b; Schmolke *et al.*, 2010; Bennett *et al.*, 2013), which are extremely valuable but not necessarily tailored to end-to-end models. We offer our perspective (as recent test subjects or "guinea pigs") to other ecosystem modellers, first in terms of how a review could be motivated and structured, and second regarding the evaluation criteria that should be used, in the most specific terms possible. Full materials from the 2014 Atlantis review, including agendas, terms of reference, and reviewer reports, are available at <http://www.nwfsc.noaa.gov/research/divisions/cb/ecosystem/marineecology/aem.cfm>, and below we synthesize lessons learned and potential evaluation criteria.

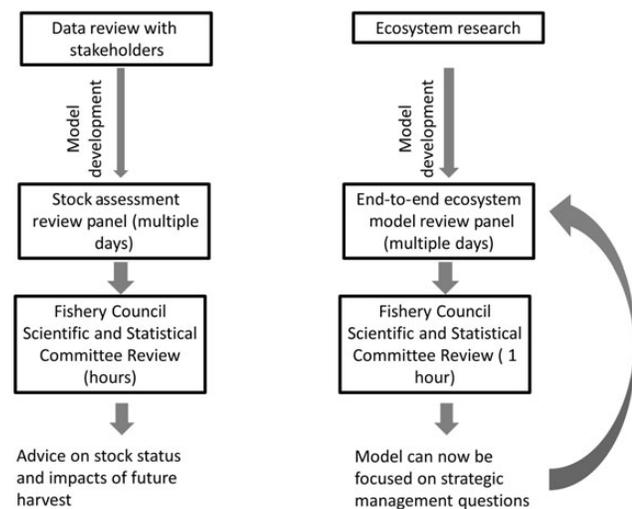


Figure 1. Schematic of the stock assessment review process typical on the US West Coast (left) vs. our recent Atlantis end-to-end model review (right). The ultimate aim of the stock assessment is to provide advice that leads to implementation of annual or biannual catch limits, whereas an end-to-end model review evaluates whether it can be used for strategic ecosystem management questions. If so, end-to-end modellers can expect additional reviews (looping arrow) on focused topics or species.

How to initiate the process?

It is unlikely that fishery managers will review a model without specific management needs; the time and expense are too high to justify the review on purely scientific grounds. In our case, for the California Current Atlantis model on the US West Coast, momentum built for the review after ~5 years of research and peer reviewed publication (Kaplan *et al.*, 2010, 2012a,b, 2013, 2014; Kaplan and Leonard, 2012), largely within the context of an Integrated Ecosystem Assessment (Levin *et al.*, 2009). The California Current Integrated Ecosystem Assessment applied Atlantis and several other models to evaluate and rank fisheries management options, and results from these analyses were presented to fishery managers (Levin *et al.*, 2013). Additionally, the review was initiated as this Atlantis model was being considered for inclusion in a strategic environmental impact assessment (Pacific Fishery Management Council and National Marine Fisheries Service, 2014). This marked a transition between use of the model for research (by federal scientists) to direct inclusion in the fishery management process. In that process, the Pacific Fishery Management Council (PFMC) advises the National Marine Fisheries Service (US Department of Commerce), which reviews and implements regulations. Readers unfamiliar with US federal fisheries management and the role of Fishery Management Councils are referred to deReynier (2014).

Our review was greatly facilitated by the availability of formal terms of reference (see below) for a Methodology Review Process within the PFMC. This fishery management council is the primary management body responsible for US West Coast federal fisheries, and the primary target audience for management advice from the model. Critically, these terms of reference are not designed for evaluation of single species stock assessments; assessments have their own distinct evaluation criteria. Instead, the Methodology Review Process was designed to address broader model types, with criteria structured around general model behaviour and technical performance. These terms of reference were supplemented with additional criteria specific to the purpose of the review, aimed at evaluating whether the model could be applied to a series of ecosystem-based management questions in the California Current.

Review panel composition, terms of reference, and agenda

Following standard practice for our regional fishery management council's Methodology Review Process, our panel consisted of technical advisors to the council—the Scientific and Statistical Committee (SSC). These seven experts primarily had backgrounds in stock assessment or economics. They were augmented with one senior ecosystem modeller from a different laboratory within our agency, and three external reviewers provided by the Center of Independent Experts (CIE). These external reviewers greatly increased the depth of ecological and oceanographic knowledge available to the panel, a step forward in terms of breaking down the disciplinary silos that can stymie end-to-end modelling (Rose *et al.*, 2010). Their ecological modelling expertise offered novel perspectives, in particular because the reviewed Atlantis modelling framework is one of many end-to-end models (Plagányi, 2007), each with its strengths and weaknesses, and much was gained by comparison with approaches in other regions and countries.

The terms of reference were a critical tool for focusing discussion and organizing the review. Highlights of these terms of reference are listed in Table 1.

Table 1. Terms of reference for California Current Atlantis end-to-end model review.

<p>Technical merits and/or deficiencies of the methodology and recommendations for remedies (criteria from Pacific Fishery Management Council (2012)).</p> <ul style="list-style-type: none"> • What are the data requirements of the methodology? • What are the situations, management uses, and spatial scales for which the methodology is applicable? • What are the assumptions of the methodology? • Is the methodology correct from a technical perspective? • How robust are results to departures from the assumptions of the methodology? • Does the methodology provide estimates of uncertainty? How comprehensive are those estimates? • What is the process of model fitting and calibration? <p>Strengths, weaknesses, appropriate uses, and potential areas of improvement for the Atlantis models with respect to these management needs, in the context of ecosystem-based management.</p> <ul style="list-style-type: none"> • Foodweb impacts of groundfish fisheries, pelagic fisheries, and other anthropogenic impacts. • Ranking of potential fishery management strategies, including spatial management, harvest rates, and quota systems. This expands beyond trophic impacts to include habitat, bycatch, and economic indicators. • Evaluation of risks of climate change and ocean acidification. • Informing parameters within single species assessments, e.g. natural mortality. • Formal Management Strategy Evaluation to 'simulation test' new methods of stock assessment, data collection, and decision making.
--

The two-and-a-half day agenda was structured around these terms of reference, and began with the history, goals, and evolution of model development, and an introduction to the potential role of the model for management. Following that, we presented an overview of the Atlantis modelling framework, including model mechanics, assumptions, and functional relationships. We then focused on our local implementation of Atlantis, including geography and functional groups, data, calibration and fits to historical data, and treatment of uncertainty and sensitivity. The review panel then considered recent publications and current and potential model applications for research and fishery management.

Though these topics parallel the components of model evaluation and documentation proposed by others (Bart, 1995; Schmolke *et al.*, 2010), the details of our review were tailored to end-to-end models. For instance, the review explicitly included the mathematics, functional forms, assumptions, and dynamics of the model, but excluded the implementation within the C++ code. In end-to-end models such as Atlantis, there is simply too much code to review in a few days, and the review panelists would need expertise in computer science rather than ecology, fisheries, and economics. Rather than the detailed code appendices or descriptions suggested by Schmolke *et al.* (2010) or Bart (1995), we referred reviewers to technical documents that summarized the mathematics of the model. A formal code review by computer scientists is a daunting challenge likely requiring at least weeks of time from several skilled programmers. For a recent code review by Mozilla engineers of “500 line snippets” of other scientific code (<1% the length of Atlantis code), see Petre and Wilson (2013). In the meantime, our approach

is to exclude the code from science review, while the Atlantis code developers (Elizabeth Fulton and Rebecca Gorton, CSIRO Australia) make it visible to all users via SVN code share, for testing, application, and feedback via a wiki. This approach is supported by the Mozilla study’s authors, who noted that “hour for hour, nothing is better at improving code quality than having someone other than the author read it carefully”, and that “Today, code review is routine in every large open source project” via sharing of code and feedback between programmers.

Descriptions of model calibration and treatment of uncertainty and sensitivity were also tailored to end-to-end models, primarily to accommodate practical limitations on the total number of simulations that we could explore. Atlantis models take hours to days to run, and therefore the formal Monte-Carlo approaches or parameter estimation that might be used to calibrate and test other models are not yet possible. Instead, we explored bounded scenarios (Fulton *et al.*, 2011) that compare the base ecosystem model to parameterizations that are at the upper and lower bound of abundance or productivity of relevant species. One example bounded the uncertainty in biomass estimates of an abundant fishery target species, Pacific hake (*Merluccius productus*), by applying high and low biomass estimates based on the coefficients of variation between stock assessments (Ralston *et al.*, 2011); this has been presented elsewhere (Collie *et al.* 2016). We also presented bounded scenarios for rate parameters, creating high-productivity vs. low-productivity Atlantis parameterizations characterized by different groundfish stock–recruit steepness, natural mortality, and unfished recruitment (Figures 2 and 3). Finally, we ran a limited number of sensitivity tests, varying initial conditions by $\pm 50\%$ and varying rate parameters by $\pm 20\%$. These rate parameters included assimilation rates, linear mortality, diet interaction strengths, and growth rates of primary producers or consumers (Figure 4). These sensitivity tests involved 15 separate simulations, with run times of ~ 3 days per simulation. With additional computing power, more formal sensitivity tests could be

conducted to identify which parameter sets lead to the variation in several types of model outputs (Lehuta *et al.*, 2013), though in our case computing needs would be more extreme than required in published examples.

When presenting the results of model calibration, we qualitatively compared model predictions of biomass to observed estimates of biomass (or estimated historical biomass from stock assessment). The approach was to tally the number of cases where the Atlantis model predicted the correct magnitude and/or trend in all or part of historical time series. We also described initial phases of model calibration, checking for persistence of species, and testing a range of fixed fishing mortality rates to check the productivity of the simulated stocks. The review panel requested that we develop a set of minimum performance standards for this calibration process, which we propose in the next section based on other published examples of ecosystem model review, our personal experience, and what is practical with end-to-end simulation models such as Atlantis.

These proposed standards are focused on determining whether a model should be used for strategic fisheries management purposes. We recommend that this be preceded by documented model development and calibration using a broader set of tools, for instance the pattern-oriented modelling (POM) approach of Grimm *et al.* (2005). The POM approach leads to refinement of model structure as model outputs are compared with patterns and observations on multiple scales of time, space, and complexity. POM should also further improve our estimation or understanding of interactions between parameters and processes (Grimm *et al.*, 2005). A key concept in POM is that multiple patterns of observations can be used as filters to test both parameters and processes (Kramer-Schadt *et al.*, 2007). For Atlantis specifically, such patterns could include fleet catches, biomass trends, and seasonal and spatial patterns of vertebrates, invertebrates, and phytoplankton (see, for example, Link *et al.*, 2010a). However, we expect this POM approach to be most useful during model development and with ecologically oriented journal peer reviewers. In our experience, after peer reviewed publications, review panels considering potential fisheries management applications will focus on the following topics.

		Ecosystem model productivity:		
		Low productivity	Base productivity	High productivity
Catch streams from assessments:	Low catch stream			
	Recent average catch stream			
	Moderately high catch stream			
	High catch stream			

Figure 2. Example of simulation design using the bounded scenarios approach. The Atlantis model with base productivity (centre column) was modified to create low- and high-productivity parameterizations (left and right columns). These parameterizations were informed directly by low- and high-productivity estimates from groundfish stock assessments, specifically via each species’ stock-recruit and natural mortality parameters. Each of the three Atlantis parameterizations was then subject to a range of fishing catches (rows) to evaluate foodweb impacts of the catch policies (Pacific Fishery Management Council and National Marine Fisheries Service, 2014).

Proposed standards for end-to-end ecosystem model performance

Mandatory standards

Mandatory standards listed below are nine minimum performance standards for use of end-to-end ecosystem models in strategic management applications. As performance standards, these focus on model output, assuming that input data and model structure, and assumptions have been vetted (‘technical merits and/or deficiencies’ within Table 1).

Population dynamics

- *Persistence of functional groups:* All biological functional groups should persist. Groups that go extinct (i.e. $< 1\%$ of initial biomass) in unfished prototype simulations should be investigated. If groups cannot persist in any simulation, they should be removed, and implications of this for model behaviour should be investigated.

Meeting this criterion is not a trivial task. Previous foodweb modelling by Gaichas *et al.* (2012) testing parameterization of predation functional responses suggests that somewhere between 1 in 1000 and 1 in 10 000 randomly drawn parameter sets lead to stable, persistent species dynamics. Thorpe *et al.*

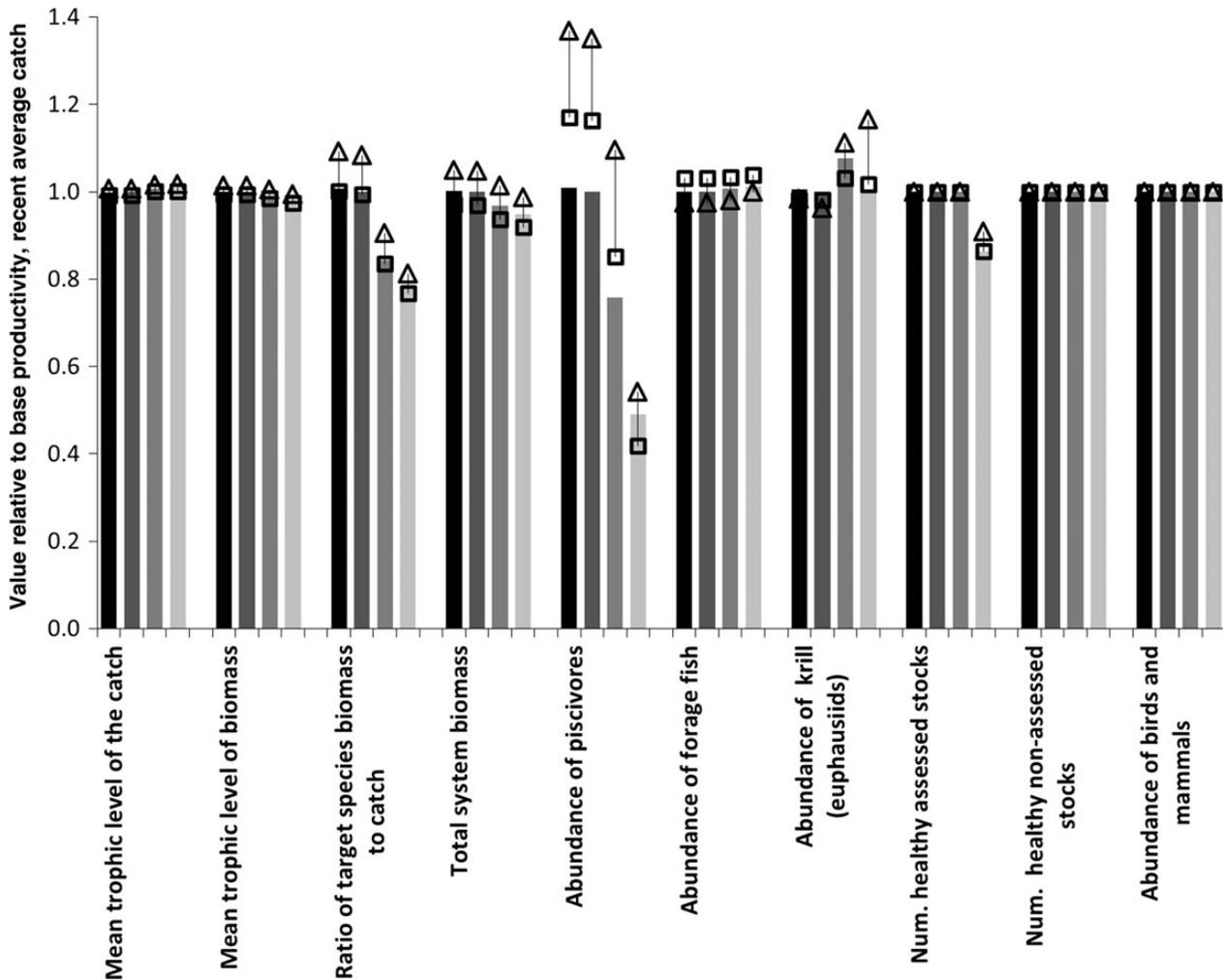


Figure 3. Results from the bounded scenarios approach introduced in Figure 2. Value of ecosystem metrics predicted by the California Current Atlantis model when subjected to four catch streams ranging from low (black) to high (light grey). The Atlantis model with base productivity (bars) was modified to create low- and high-productivity parameterizations (squares and triangles, respectively) (Pacific Fishery Management Council and National Marine Fisheries Service, 2014).

(2015) used the 1% persistence criterion as well, and similarly found low acceptance of randomly drawn parameter sets. Although in reality marine extinctions may occur in the absence of fishing, they are likely rare over simulation time-scales (~100 years) in any given ecosystem, and at the taxonomic resolution of end-to-end models. Multispecies fisheries modelling convention is that “in the absence of detailed information to the contrary. . . general evidence of stability [not proceeding to extinction or to an infinitely great size] forms a valuable guide to an adequate formulation of theoretical population models” (Beverton and Holt, 1957).

- **Model achieves equilibrium:** During calibration, an unfished model should be run to quasi-equilibrium (Ainsworth and Walters, 2015). In a simulation with no stochasticity, constant oceanographic forcing, and no fishing, the majority of vertebrate species or groups should show no significant trend in biomass over the final 20 years. Age structures should appear stable. Similar to the requirement of persistence of functional groups, this is a general rule and for particular species there may be detailed information to the contrary—for instance, species that

show cycles due to cannibalism (Botsford and Hobbs, 1995) or chaotic behaviour due to very high-population growth rates. Starting at this unfished quasi-equilibrium, the model can be perturbed with historical or future fishery or other human actions.

The California Current ecosystem includes many long-lived species, not only marine mammals but also fish such as rockfish (*Sebastes* spp.) that can live as long as 80 years. Thus, we expect 80- to 100-year simulations to address this criterion, though end-to-end models with shorter-lived species may require fewer years.

- **Hindcast:** Model biomass trajectories for a historical period (i.e. hindcast) should be compared against both survey time series and stock assessment output, where available. These should include confidence intervals around biomass (see next point).
Note that confidence intervals around survey or assessment biomass estimates may include estimation uncertainty, but also the impacts of structural uncertainty and variation in abundance estimates across models and modellers (Ralston et al. 2011). The call for more hindcast comparisons has been made by other authors (Bart, 1995; Link et al. 2010b; Schmolke et al. 2010; ICES, 2013; Collie et al. 2014) and on the US West Coast by a

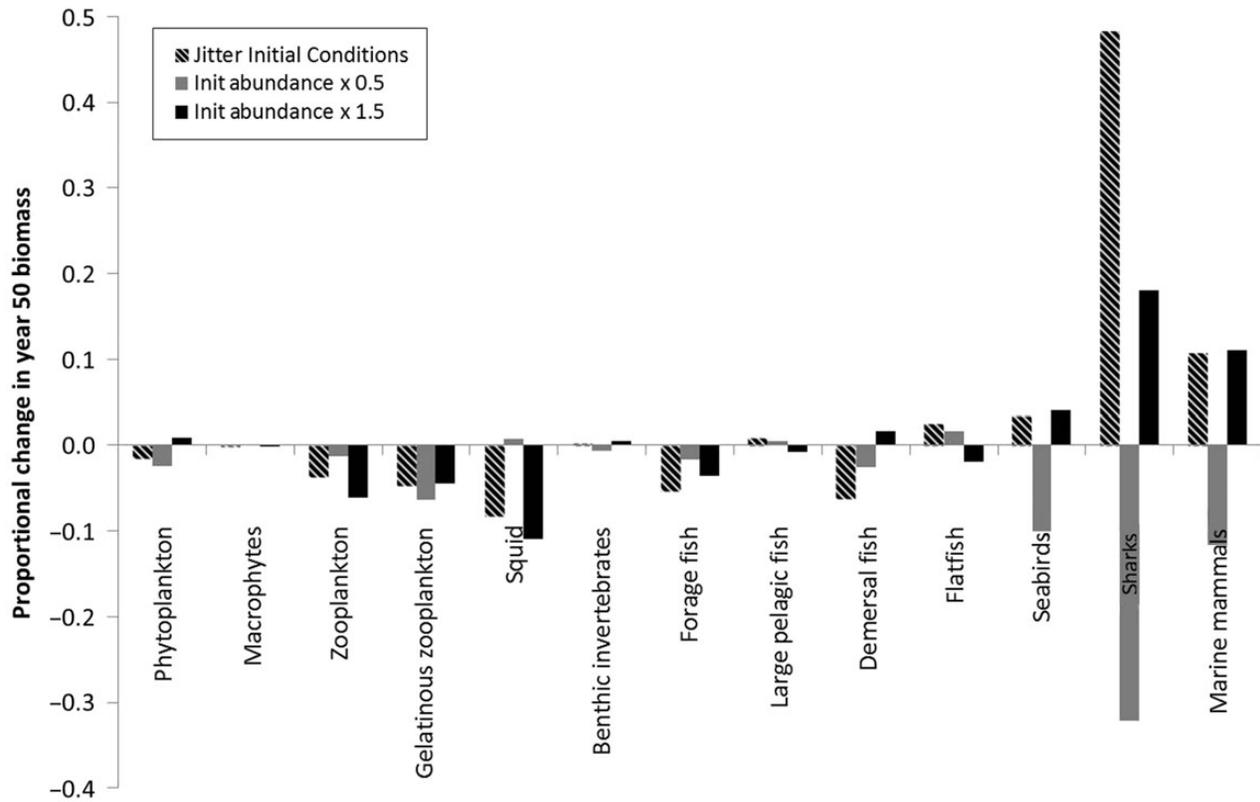


Figure 4. Sensitivity tests of the California Current Atlantis model to initial conditions. Y-axis illustrates proportional change in biomass at year 50, relative to a base model at year 50. Initial abundances of all vertebrates were multiplied by $1.5 \times$ initial abundance, or by $0.5 \times$, or jittered randomly (per vertebrate group) within the uniform range from 0.5 to $1.5 \times$ initial abundance. For simplicity, functional groups are aggregated into coarse taxa. Most species show < 0.05 change in predicted biomass at year 50 when initial conditions are changed. Seabirds, sharks, and marine mammals exhibit slow dynamics (longer lifespans and generation times) and therefore effects of initial conditions are still evident at year 50.

fishery management advisory committee (Pacific Fishery Management Council, 2013). Hindcasts should include not only fish but also other biological groups (Mackinson, 2013).

- **Hindcast agreement with data:** Qualitative model comparisons with survey data and assessment outputs are adequate. Model authors should demonstrate to reviewers that the majority of functional groups match the biomass trends (but not magnitude) observed in survey data or estimated from stock assessments. More specifically, species comprising 80% of the total biomass (for species with stock assessment output or survey data) should yield trends similar to those in survey data or assessment output.

Effectively, this requires good agreement between model and available data for abundant stocks. In the California Current, this means capturing dynamics of abundant pelagic and mid-water stocks, such as sardine (*Sardinops sagax*) and Pacific hake. The goal should be to capture overall ecosystem dynamics, hence the focus on 80% of biomass, rather than focusing solely on the number of species (some quite rare) for which model predictions match survey or assessment estimates. We focus on fitting trends rather than total magnitude because the total magnitude of biomass will differ between models simply due to model differences regarding the age of recruitment (first age class present in the model) and the model representation (or omission) of predation on fish younger than the harvested ages. Similarly, total magnitude of biomass estimated from surveys is heavily influenced by net characteristics (catchability and selectivity), which makes

direct model comparison with survey trends more appropriate than to survey magnitude. Our criteria are less stringent than those proposed as best practices by FAO (2008), primarily because end-to-end models such as Atlantis are too computationally intensive to allow statistical parameter estimation. Nonetheless, we feel their guidelines on sensitivity analyses are useful, “For dynamic models: 1) fit to as much data as possible using appropriate likelihood structures; . . . in cases of fixing parameter values, additional sensitivity analyses should be used to assess model sensitivity to the assumptions; and use results of sensitivity analyses to guide future data collections and the continuation of key time series.”

- **Reproducing patterns of temporal variability at many time-scales:** Dynamics of abundant, periodic species (e.g. hake and sardine in the California Current) should be at least qualitatively captured by the model, even if they must be forced with historical recruitment time series. These periodic dynamics should arise, or alternatively be forced, in future projections.

The challenge of capturing dynamics of periodic small pelagic species has been noted by other authors (Rose *et al.* 2010) but called for by fishery managers (Pacific Fishery Management Council, 2013). In the California Current even single species sardine stock assessments only represent the years from 1993 to present, and earlier stock behaviour and dynamics were distinctly different. Therefore, we anticipate needing to force the Atlantis model to capture multi-decadal trends in recruitment and

abundance, rather than expecting these cyclic patterns to emerge. Currently, we are forcing Atlantis recruitment patterns using recruitment deviations from single species models. An alternative would be to test scenarios for primary productivity time series, which may also explain cyclical patterns of small pelagic species (e.g. Mackinson *et al.*, 2009).

Productivity, life history parameters, and ecology

Based on our experience, review panels considering potential fisheries management applications will focus on the standards above related to population dynamics, but end-to-end ecosystem models should also meet the following mandatory standards related to productivity, life history, and ecological interactions. These could be handled during journal peer review before potential fishery management consideration. These tests and visualizations mentioned below are similar to what Bart (1995) labelled secondary predictions—“intermediate outputs of the model that are not provided as standard output or would usually not be used in making management decisions but that can be used to assess the reliability of the model.”

- *Productivity*: The majority of functional groups and 80% of biomass, summed over all vertebrate species, should qualitatively match expected productivity (FMSY) from stock assessments or life history theory (Ainsworth and Walters, 2015).
- *Natural mortality*: Natural mortality (M), including predation, should be plotted as a function of age over time. The variable M should decrease with age for the majority of species or groups, and should be consistent with expectations from life history theory or literature parameters.
- *Age and length structure*: Predicted age and length structure from the model should qualitatively match expected age and length structure, for the majority of species or groups.
- *Diets*: Diet predictions from the model should qualitatively match diet data from empirical studies, for the majority of species or groups. Predicted diets should fall within the range of empirical diet compositions, acknowledging high empirical diet variability (and uncertainty) across space, season, and year. Fulton *et al.* (2007) provide examples of this for a Southeast Australia Atlantis model.

Best practices for any application

These best practices are specific to end-to-end models, and are not captured in other valuable best practices marine ecosystem modelling guides (FAO, 2008; Link *et al.*, 2010b).

- *Multiple parameterizations*: Multiple model parameterizations should be retained, where all alternatives have approximately similar agreement with data.
This approach has been used in Australia (Fulton *et al.*, 2007, 2014) and is a straightforward way to capture some aspects of parameter uncertainty. Fulton and colleagues developed an Atlantis model with three parameterizations, all of which have similar agreement to scientific surveys and fishery logbook data.
- *Bounded scenarios*: should be developed to handle parameter uncertainty, *sensu* Fulton *et al.* (2011).
An example of the bounded parameterization approach was provided in Figure 2. In that example, uncertainty in productivity

of groundfish stocks (related to uncertainty in recruitment and natural mortality) was translated into high- and low-productivity parameterizations of an Atlantis model. Effects of fishing groundfish were then predicted using a base Atlantis model parameterization, and compared with predictions from this high- and low-productivity scenario.

Results of one end-to-end ecosystem model review

Our perspective is strongly influenced by the 2014 California Current Atlantis model review. Readers interested in specific results and recommendations from the review panel and external CIE reviewers are referred to the website provided above. Most importantly, the panel found that the Atlantis end-to-end model for the California Current could be used to address several ecosystem-based fishery management initiatives, including those related to protecting unfished forage fish, social and economic effects of fisheries management, effects of climate shifts, and prediction of ecological responses (tracked via ecosystem indicators) to fishery management actions. This was tempered by the review panel's calls for continued model improvement, and for additional technical reviews of the newest model version, focused on particular management questions, before additional use in the policy arena. The reviewers called for ongoing engagement between end-to-end modellers and fishery managers; as noted above, this is already common practice in at least one other US region (Alaska; Link *et al.*, 2010b). The review panel was cognizant of the strategic nature of end-to-end models, and emphasized the importance of conveying this when presenting model results in the management arena. For instance, they generally supported the “radar plot” or “spider diagram” approaches used to rank relative performance of management policies, as measured against multiple ecosystem goals (e.g. Kaplan *et al.*, 2012b and Figures 5–8 therein; Fulton *et al.* 2014 and Figure 5 therein). They cautioned against presenting results to management audiences in a manner that could be misinterpreted as overly precise, such as decision tables (e.g. Kaplan *et al.*, 2010; Tables 2–3 therein).

Encouragingly, the review panel findings were endorsed by decision makers (PFMC), and this allows future policy applications following model improvements and additional technical review. In the US federal fishery management system, journal peer review generally does not confer enough confidence for policy applications, and thus vetting via a Methodology Review Process like the one described here is necessary for models to move forward in the management arena.

Conclusions

Here, we developed guidelines to implement rigorous, multi-day peer reviews of end-to-end ecosystem models for use in management decisions. Model reviews for other regions and nations could be initiated, organized, and structured along these lines, with adaptations for specific management needs. Our proposed set of evaluation criteria for end-to-end models complements more over-arching best practices for marine ecosystem modelling (FAO, 2008; Townsend *et al.*, 2008; Link *et al.*, 2010b). Most importantly, we have met the challenge of organizing a form of peer review which has been called for by end-to-end modellers (Link *et al.*, 2010b) and that fisheries managers demand before policy applications. Other US ecosystem model reviews, notably those mentioned above in Alaska and the Northeast, have paved the way for this. Though our effort has many limitations it can benefit future

developments in this realm, and we encourage authors from outside the USA to add to this conversation.

Our proposed standards are applicable to many end-to-end models that are currently used to inform management, or are moving into management arenas (e.g. Atlantis, Ecosim/Ecospace, OSMOSE, and NEMURO-SAN; Fiechter *et al.*, 2015). However, we acknowledge that particular applications and different regulatory environments may require different criteria to judge model performance. Additionally, each model type has a unique set of state variables, and may not include all processes discussed in the Atlantis review. One particular avenue for expansion is criteria to evaluate fleet dynamics. End-to-end models increasingly include simulated fleet dynamics, and we expect a parallel set of criteria will be required to evaluate model behaviour with respect to fleet effort, catch, bycatch, and spatial distribution. For instance, Lehuta *et al.* (2013) assessed skill of an ISIS-Fish model for Bay of Biscay anchovy against observed total annual catch, monthly catch per fleet, and age composition of the catch.

Our evaluation criteria and format of the model review were informed *a priori* by our familiarity with best practices guides for marine ecosystem models (FAO, 2008; Link *et al.*, 2010b, 2012), but they also echo best practices established for other model types. In terrestrial contexts (primarily related to Northern Spotted Owl), Bart (1995) suggested that peer review was necessary before applying individual-based models to management questions. Bart (1995) identified key steps that echo many we proposed for marine end-to-end models, such as using bounded scenarios and quantifying uncertainty stemming from different sources. Schmolke *et al.* (2010) proposed the TRACE (transparent and comprehensive ecological modelling) documentation structure, which echoes many themes we have discussed here: problem formulation, design and formulation, model description, calibration, verification, sensitivity analyses, validation, results, uncertainty analysis, and recommendations. Our experience suggests that fishery managers will focus reviews on the dynamics of species' biomass and abundance rather than for instance spatial patterns or broader ecological characteristics, but we also urge modellers to consider broader criteria (and more generally Pattern-Oriented Modelling approaches).

New developments in end-to-end model review are likely to include improvements in evaluating model skill in three areas: (i) spatially explicit output from models, (ii) hindcasts, and (iii) forecasts. First, model evaluation and calibration initially focuses on spatially aggregated fits of models to survey data or stock assessment abundance, since these are standard metrics in the fisheries management realm. However, the spatial match between model predictions and data is also important; to date this is usually handled via qualitative, visual comparisons, but in the future formal spatial comparisons using metrics such as the kappa statistic (a cell-to-cell comparison of model to observation, Cohen (1960)), or fuzzy kappa statistic (which makes spatial comparisons over a neighbourhood, Hagen (2003)) are a likely next step. These metrics have been applied to ecosystem models by Rose *et al.* (2009) and software exists to facilitate these spatial comparisons (Map Comparison Kit; www.risks.nl/mck; Hagen-Zanker *et al.* 2006; Visser and de Nijs 2006). Second, the formal metrics to compare model output to observational data have not been well defined. Metrics familiar to fisheries modellers might be used (coefficient of determination or R^2 , Akaike's information criterion or AIC), but metrics of model skill more familiar to biogeochemical marine models may be more appropriate. These include root mean-squared error, reliability index, average error (bias), average absolute error, modelling

efficiency, and skill score. Readers are referred to Stow *et al.* (2009) for a summary of candidate metrics for skill assessment, and Jolliff *et al.* (2009) for useful 'target diagrams' and Taylor diagrams that visualize some of these metrics. Third, as we transition from evaluating ecosystem models in hindcast mode into testing their ability as forecasting tools, additional metrics will be required, for instance receiver-operating curves and area-under-curve or AUC (Fielding and Bell, 1997) criteria that test the ability of a model to discriminate between presence and absence of a species (e.g. in a model region), or a phenomenon (e.g. pH falling below some threshold). We expect that these spatial and non-spatial metrics of model hindcast and forecast skill will become increasingly important in future reviews of end-to-end models.

Multi-model approaches will be required to fully understand the impacts of structural assumptions in end-to-end models and resulting structural uncertainty, and this will require even more input from reviewers and fishery management bodies. Multi-model inference formally integrates results from multiple models, weighted by the skill metrics discussed above. This extends beyond model comparison (e.g. Kaplan *et al.* (2013) considered in the Atlantis review panel to either ensemble modelling (Gårdmark *et al.*, 2012) or expert weighting of model results by review panels. In both cases, management bodies must expect additional calls for reviews of new ecosystem models spanning a range of complexity.

An informal poll of US West Coast stock assessors suggests that rather gruelling, multi-day stock assessment review panels are considerably more in depth and challenging than traditional journal peer review. For end-to-end ecosystem models, traditional peer review has been the primary means for challenging and improving the science to date. We suggest that in-person reviews of ecosystem models are necessary to prove the utility of these models in the management context and build familiarity of decision makers with their use for strategic advice. Moving ecosystem models from the frying pan of refereed journals to the fire of review panels may be somewhat uncomfortable for the analysts (including these authors), but is the most efficient way to improve the science and strategic advice for management decisions.

Acknowledgements

Many thanks are due to members of the Atlantis Methodology Review Panel for comments in June and July, 2014. These reviewers included Martin Dorn, Kerim Aydin, Pete Lawson, Cindy Thomson, Galen Johnson, André Punt, Will Satterthwaite, Tien-Shui Tsou, Reg Watson, Kenneth Frank, and Daniel Howell. The Pacific Fishery Management Council and NMFS Office of Science and Technology facilitated the review, and Kit Dahl was the PFMC staff representative. This work was supported in part by the NOAA Ocean Acidification Program and NOAA Centers for Coastal Ocean Science. Kirstin Holsman and Chris Harvey (NOAA), Sigrid Lehuta (IFREMER), and two anonymous reviewers provided helpful comments on an early draft.

References

- Ainsworth, C. H., and Walters, C. J. 2015. Ten common mistakes made in Ecopath with Ecosim modelling. *Ecological Modelling*, 308: 14–17.
- Bart, J. 1995. Acceptance criteria for using individual-based models to make management decisions. *Ecological Applications*, 5: 411–420.
- Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., *et al.* 2013. Characterising

- performance of environmental models. *Environmental Modelling Software*, 40: 1–20.
- Beverton, R. J. H., and Holt, S. J. 1957. On the dynamics of exploited fish populations. *Fisheries Investigation Series 2*, volume 19, UK Ministry of Agriculture, Fisheries and Food. London, U.K.
- Botsford, L. W., and Hobbs, R. C. 1995. Recent advances in the understanding of cyclic behavior of Dungeness crab (*Cancer magister*) populations. In *ICES Marine Science Symposia*, pp. 157–166. International Council for the Exploration of the Sea, 1991, Copenhagen, Denmark.
- Christensen, V., and Walters, C. J. 2004. Ecopath with Ecosim: methods, capabilities and limitations. *Ecological Modelling* 172: 109–139.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.
- Collie, J. S., Botsford, L. W., Hastings, A., Kaplan, I. C., Largier, J. L., Livingston, P. A., Plagányi, É., *et al.* 2016. Ecosystem models for fisheries management: finding the sweet spot. *Fish Fish*, 17: 101–125.
- deReynier, Y. L. 2014. U.S. Fishery Management Councils as ecosystem-based management policy takers and policymakers. *Coastal Management*, 42: 512–530.
- European Commission. 2008. Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). *Official Journal of the Eur Union Lex*, 164: 19–40.
- FAO. 2008. Fisheries management. 2. The ecosystem approach to fisheries. 2.1 Best practices in ecosystem modelling for informing an ecosystem approach to fisheries. Technical Guidelines, FAO, Rome. <ftp://ftp.fao.org/docrep/fao/011/i0151e/i0151e00.pdf> (last accessed 16 September 2015).
- Fiechter, J., Rose, K. A., Curchitser, E. N., and Hedstrom, K. S. 2015. The role of environmental controls in determining sardine and anchovy population cycles in the California Current: Analysis of an end-to-end model. *Progress in Oceanography*, 138: 381–398.
- Fielding, A. H., and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24: 38–49.
- Fulton, E. 2004. Effects of spatial resolution on the performance and interpretation of marine ecosystem models. *Ecological Modelling*, 176: 27–42.
- Fulton, E. A. 2010. Approaches to end-to-end ecosystem models. *Journal of Mar Systems*, 81: 171–183.
- Fulton, E. A., Link, J. S., Kaplan, I. C., Savina-Rolland, M., Johnson, P., Ainsworth, C., Horne, P., *et al.* 2011. Lessons in modelling and management of marine ecosystems: the Atlantis experience. *Fish Fish*, 12: 171–188.
- Fulton, E. A., Smith, A. D. M., and Smith, D. C. 2007. Alternative management strategies for southeast Australian Commonwealth Fisheries: stage 2: quantitative management strategy evaluation. Australian Fisheries Management Authority Report, Canberra Australia. http://atlantis.cmar.csiro.au/www/en/atlantis/mainColumnParagraphs/02/text_files/file/AMS_Final_Report_v6.pdf (last accessed 14 January 2009).
- Fulton, E. A., Smith, A. D. M., Smith, D. C., and Johnson, P. 2014. An Integrated Approach Is Needed for Ecosystem Based Fisheries Management: Insights from Ecosystem-Level Management Strategy Evaluation. *PLoS ONE*, 9: e84242.
- Gaichas, S. K., Odell, G., Aydin, K. Y., Francis, R. C., and Rochet, M.-J. 2012. Beyond the defaults: functional response parameter space and ecosystem-level fishing thresholds in dynamic food web model simulations. *Canadian Journal of Fisheries and Aquatic Sciences*, 69: 2077–2094.
- Gårdmark, A., Lindegren, M., Neuenfeldt, S., Blenckner, T., Heikinheimo, O., Müller-Karulis, B., Niiranen, S., *et al.* 2012. Biological ensemble modeling to evaluate potential futures of living marine resources. *Ecological Applications*, 23: 742–754.
- Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., *et al.* 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science*, 310: 987–991.
- Hagen, A. 2003. Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, 17: 235–249.
- Hagen-Zanker, A., Engelen, G., Hurkens, J., Vanhout, R., and Uljee, I. 2006. Map comparison kit 3: user manual. Research Institute for Knowledge Systems, Maastricht, The Netherlands.
- Hunsicker, M. E., Ciannelli, L., Bailey, K. M., Buckel, J. A., Wilson White, J., Link, J. S., Essington, T. E., *et al.* 2011. Functional responses and scaling in predator-prey interactions of marine fishes: contemporary issues and emerging concepts. *Ecology Letters*, 14: 1288–1299.
- ICES. 2013. Interim Report of the Working Group on Multispecies Assessment Methods (WGSAM), 21–25 October 2013. ICES, Stockholm, Sweden. <http://ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/SSGSUE/2013/WGSAM2013.pdf> (last accessed 15 November 2015).
- Interagency Ocean Policy Task Force of the White House Council on Environmental Quality. 2010. Final recommendations of the interagency ocean policy task force. Washington, DC, White House Council on Environmental Quality, Washington D.C. <http://www.whitehouse.gov/files/documents/OPTFFinalRecs.pdf> (last accessed 3 June 2013).
- Jolliff, J. K., Kindle, J. C., Shulman, I., Penta, B., Friedrichs, M. A., Helber, R., and Arnone, R. A. 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems*, 76: 64–82.
- Kaplan, I. C., Brown, C. J., Fulton, E. A., Gray, I. A., Field, J. C., and Smith, A. D. M. 2013. Impacts of depleting forage species in the California Current. *Environmental Conservation*, 40: 380–393.
- Kaplan, I. C., Gray, I. A., and Levin, P. S. 2012a. Cumulative impacts of fisheries in the California Current. *Fish Fish*, 14: 515–527.
- Kaplan, I. C., Holland, D. S., and Fulton, E. A. 2014. Finding the accelerator and brake in an individual quota fishery: linking ecology, economics, and fleet dynamics of US West Coast trawl fisheries. *ICES Journal of Marine Science*, 71: 308–319.
- Kaplan, I. C., Horne, P. J., and Levin, P. S. 2012b. Screening California current fishery management scenarios using the Atlantis end-to-end ecosystem model. *Progress in Oceanography*, 102: 5–18.
- Kaplan, I. C., and Leonard, J. 2012. From krill to convenience stores: Forecasting the economic and ecological effects of fisheries management on the US West Coast. *Marine Policy*, 36: 947–954.
- Kaplan, I. C., Levin, P. S., Burden, M., and Fulton, E. A. 2010. Fishing catch shares in the face of global change: a framework for integrating cumulative impacts and single species management. *Canadian Journal of Fisheries and Aquatic Sciences*, 67: 1968–1982.
- Kramer-Schadt, S., Revilla, E., Wiegand, T., and Grimm, V. 2007. Patterns for parameters in simulation models. *Ecological Modelling*, 204: 553–556.
- Lehodey, P., Senina, I., and Murtugudde, R. 2008. A spatial ecosystem and populations dynamics model (SEAPODYM)—modeling of tuna and tuna-like populations. *Progress in Oceanography*, 78: 304–318.
- Lehuta, S., Petitgas, P., Mahévas, S., Huret, M., Vermard, Y., Uriarte, A., and Record, N. R. 2013. Selection and validation of a complex fishery model using an uncertainty hierarchy. *Fisheries Research*, 143: 57–66.
- Levin, P. S., Fogarty, M. J., Murawski, S. A., and Fluharty, D. 2009. Integrated ecosystem assessments: developing the scientific basis for ecosystem-based management of the ocean. *PLOS Biology*, 7: e1000014.
- Levin, P. S., Wells, B. K., and Sheer, M. B. 2013. California Current Integrated Ecosystem Assessment: Phase II. NOAA Northwest

- Fisheries Science Center, Seattle, USA. Available from <http://www.noaa.gov/iea/CCIEA-Report/index> (last accessed 23 January 2014).
- Link, J. S., Fulton, E. A., and Gamble, R. J. 2010a. The northeast US application of ATLANTIS: A full system model exploring marine ecosystem dynamics in a living marine resource management context. *Progress in Oceanography*, 87: 214–234.
- Link, J. S., Ihde, T. F., Harvey, C. J., Gaichas, S. K., Field, J. C., Brodziak, J. K. T., Townsend, H. M., *et al.* 2012. Dealing with uncertainty in ecosystem models: the paradox of use for living marine resource management. *Progress in Oceanography*, 102: 102–114.
- Link, J. S., Ihde, T. F., Townsend, H. M., Osgood, K. E., Schirripa, M. J., Kobayashi, D. R., Gaichas, S. K., *et al.* 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.
- Mackinson, S. 2013. Combined analyses reveal environmentally driven changes in the North Sea ecosystem and raise questions regarding what makes an ecosystem model's performance credible? *Canadian Journal of Fisheries and Aquatic Sciences*, 71: 31–46.
- Mackinson, S., Deas, B., Beveridge, D., and Casey, J. 2009. Mixed-fishery or ecosystem conundrum? Multispecies considerations inform thinking on long-term management of North Sea demersal stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, 66: 1107–1129.
- Maunder, M. N., and Punt, A. E. 2013. A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142: 61–74.
- McLeod, K., and Leslie, H. 2009. *Ecosystem-Based Management for the Oceans*. Island Press, Washington, DC.
- Methot, R. D., Jr, and Wetzel, C. R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, 142: 86–99.
- NOAA Northeast Fisheries Science Center. 2014. Term of Reference 3: Peer Review Process. http://www.nefsc.noaa.gov/program_review/pdfs/TOR3.pdf (last accessed 15 November 2014).
- Obama, B. 2010. Executive Order: Stewardship of the ocean, our coasts, and the great lakes. *In* Executive Order. <http://www.whitehouse.gov/the-press-office/executive-order-stewardship-ocean-our-coasts-and-great-lakes> (last accessed 20 November 2014).
- Pacific Fishery Management Council. 2012. Terms of Reference for the Methodology Review Process for Groundfish and Coastal Pelagic Species. Briefing Book, Portland, OR. <http://www.pcouncil.org/resources/archives/briefing-books/june-2012-briefing-book/> (last accessed 20 November 2014).
- Pacific Fishery Management Council. 2013. Report of the Pacific Sardine Harvest Parameters Workshop. Pacific Fishery Management Council, Portland, Oregon, USA. <http://www.pcouncil.org/resources/archives/briefing-books/april-2013-briefing-book/> (last accessed 28 January 2016).
- Pacific Fishery Management Council, and National Marine Fisheries Service. 2014. Draft Environmental Impact Statement (DEIS) for proposed Harvest Specifications and Management Measures for the Pacific Coast Groundfish Fishery and Amendment 24 to The Pacific Coast Groundfish Fishery Management Plan. PFMC and NMFS, Portland, Seattle, OR, WA. <http://www.westcoast.fisheries.noaa.gov/publications/nepa/groundfish/1516spexdeis.pdf> (last accessed 8 January 2015).
- Petre, M., and Wilson, G. 2013. PLOS/Mozilla Scientific Code Review Pilot: Summary of Findings. Cornell University, Ithaca New York USA. ArXiv13112412 Cs. <http://arxiv.org/abs/1311.2412> (last accessed 16 November 2014).
- Pikitch, E. K., Santora, C., Babcock, E. A., Bakun, A., Bonfil, R., Conover, D. O., Dayton, P., *et al.* 2004. Ecosystem-based fishery management. *Science*, 305: 346.
- Pinnegar, J. K., Blanchard, J. L., Mackinson, S., Scott, R. D., and Duplisea, D. E. 2005. Aggregation and removal of weak-links in food-web models: system stability and recovery from disturbance. *Ecological Modelling*, 184: 229–248.
- Plagányi, É. E. 2007. Models for an ecosystem approach to fisheries. Food and Agriculture Organization Rome. 108 pp.
- Ralston, S., Punt, A. E., Hamel, O. S., DeVore, J. D., and Conser, R. J. 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. *Fishery Bulletin*, 109: 217.
- Rose, K., Allen, J. I., Artioli, Y., Barange, M., Blackford, J., Carlotti, F., Cropp, R., *et al.* 2010. End-to-end models for the analysis of marine ecosystems: challenges, issues, and next steps. *Marine and Coastal Fisheries Dynamics, Management, and Ecosystem Science*, 2: 115–130.
- Rose, K. A., Roth, B. M., and Smith, E. P. 2009. Skill assessment of spatial maps for oceanographic modeling. *Journal of Marine Systems*, 76: 34–48.
- Schmolke, A., Thorbek, P., DeAngelis, D. L., and Grimm, V. 2010. Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology & Evolution*, 25: 479–486.
- Shin, Y.-J., and Cury, P. 2004. Using an individual-based model of fish assemblages to study the response of size spectra to changes in fishing. *Canadian Journal of Fisheries and Aquatic Sciences*, 61: 414–431.
- Stow, C. A., Jolliff, J., McGillicuddy, D. J., Jr, Doney, S. C., Allen, J., Friedrichs, M. A., Rose, K. A., *et al.* 2009. Skill assessment for coupled biological/physical models of marine systems. *Journal of Marine Systems*, 76: 4–15.
- Thorpe, R. B., Le Quesne, W. J. F., Luxford, F., Collie, J. S., and Jennings, S. 2015. Evaluation and management implications of uncertainty in a multispecies size-structured model of population and community responses to fishing. *Methods in Ecology Evolution*, 6: 49–58.
- Townsend, H. M., Link, J. S., Osgood, K. E., Gedamke, T., Watters, G. M., Polovina, J. J., Levin, P. S., *et al.* 2008. Report of the NEMoW (National Ecosystem Modeling Workshop). NOAA Tech. Memo., Silver Spring Maryland USA. NMFS-FSPO-87: 93.
- Travers, M., Shin, Y. J., Jennings, S., and Cury, P. 2007. Towards end-to-end models for investigating the effects of climate and fishing in marine ecosystems. *Progress in Oceanography*, 75: 751–770.
- Visser, H., and De Nijs, T. 2006. The map comparison kit. *Environmental Modelling & Software*, 21: 346–358.

Handling editor: Morgane Travers-Trolet