Environmental Modelling and Software 128 (2020) 104697

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

A system of metrics for the assessment and improvement of aquatic ecosystem models

Matthew R. Hipsey [a,b,*], Gideon Gal [c], George B. Arhonditsis [d], Cayelan C. Carey [e], J. Alex Elliott [f], Marieke A. Frassl [g], Jan H. Janse [h], Lee de Mora [i], Barbara J. Robson [j]

[a] Aquatic Ecodynamics, UWA School of Agriculture and Environment, The University of Western Australia, Perth, Australia
[b] UWA Oceans Institute, The University of Western Australia, Perth, Australia
[c] Yigal Allon Kinneret Limnological Laboratory, Israel Oceanographic and Limnological Research, Migdal, Israel
[d] Physical and Environmental Sciences, University of Toronto, Scarborough, Canada
[e] Department of Biological Sciences, Virginia Tech, Blacksburg, USA
[f] Centre for Ecology and Hydrology, Lancaster, United Kingdom
[g] Australian Rivers Institute, Griffith University, Brisbane, Australia
[h] Netherlands Institute of Ecology, Wageningen, the Netherlands
[i] Plymouth Marine Laboratory, Plymouth, United Kingdom
[j] Australian Institute of Marine Science and AIMS@JCU, Townsville, Australia

ARTICLE INFO

Keywords:
AEM
CSPS
Ecological modelling
Uncertainty
Water quality
Freshwater
Marine

ABSTRACT

In this paper, we introduce the CSPS framework for the hierarchical assessment of aquatic ecosystem models built on a range of metrics and characteristic signatures relevant to aquatic ecosystem condition. The framework is comprised of four levels: 0) conceptual validation; 1) comparison of simulated state variables with observations ('state validation'); 2) comparison of fluxes with measured process rates ('process validation'); and 3) assessment of system-level emergent properties, patterns and relationships ('system validation'). Of these, only levels 0 and 1 are routinely undertaken at present. To highlight a diverse range of contexts relevant to the aquatic ecosystem modelling community, we present several case studies of improved validation approaches using the level 0–3 assessment hierarchy. We envision that the community–driven adoption of these metrics will lead to more rigorously assessed models, ultimately accelerating advances in model structure and function, and improved confidence in model predictions.

1. Introduction

Models of catchments, lakes, wetlands, rivers, estuaries and marine systems are now in widespread use to simulate water quality responses to anthropogenic change and to unravel nutrient and pollutant pathways (Hipsey et al., 2015; Janssen et al., 2015). Perhaps more than any other field of environmental modelling, aquatic ecosystem modelling spans a large diversity of environments, scales and disciplines; ranging from small wetlands and lakes to the global ocean. Numerous authors have conducted reviews of the diversity of modelling approaches used to simulate lakes (Mooij et al., 2010; Janssen et al., 2015), wetlands (Coletti et al., 2017), rivers (Rode et al., 2010), and marine systems (e.g., Gentleman, 2002; Glibert et al., 2010; Rose et al., 2010; Anderson et al., 2010; Steele et al., 2013; Robson, 2014b). Looking across the geographic diversity of (process-based) aquatic ecosystem models (AEMs), it is notable that whether the focus is freshwater or marine applications, two main thematic areas of commonly-used model approaches have dominated the literature: (i) coupled physical-biogeochemical models with high spatial resolution and a focus on the biophysical environment and lower trophic levels (typically nutrients, phytoplankton and zooplankton), and (ii) models that are lumped in space and focus on resolving high trophic complexity (see Mooij et al., 2010 for fresh water systems, Fulton, 2010, for marine systems; Fig. 1). This distinction reflects the different backgrounds and research questions being asked by aquatic ecosystem modellers, and also the trade-offs between conceptual complexity, spatial resolution, data requirements and computational demands.

But how good are these models? Over the past decade or so there has been a significant expansion in the scope and capability, however, several authors have argued that AEMs have failed to keep up with

* Corresponding author. Aquatic Ecodynamics, UWA School of Agriculture and Environment, The University of Western Australia, Perth, Australia.
E-mail address: matt.hipsey@uwa.edu.au (M.R. Hipsey).

https://doi.org/10.1016/j.envsoft.2020.104697
Received 8 October 2019; Received in revised form 2 March 2020; Accepted 10 March 2020
Available online 13 March 2020

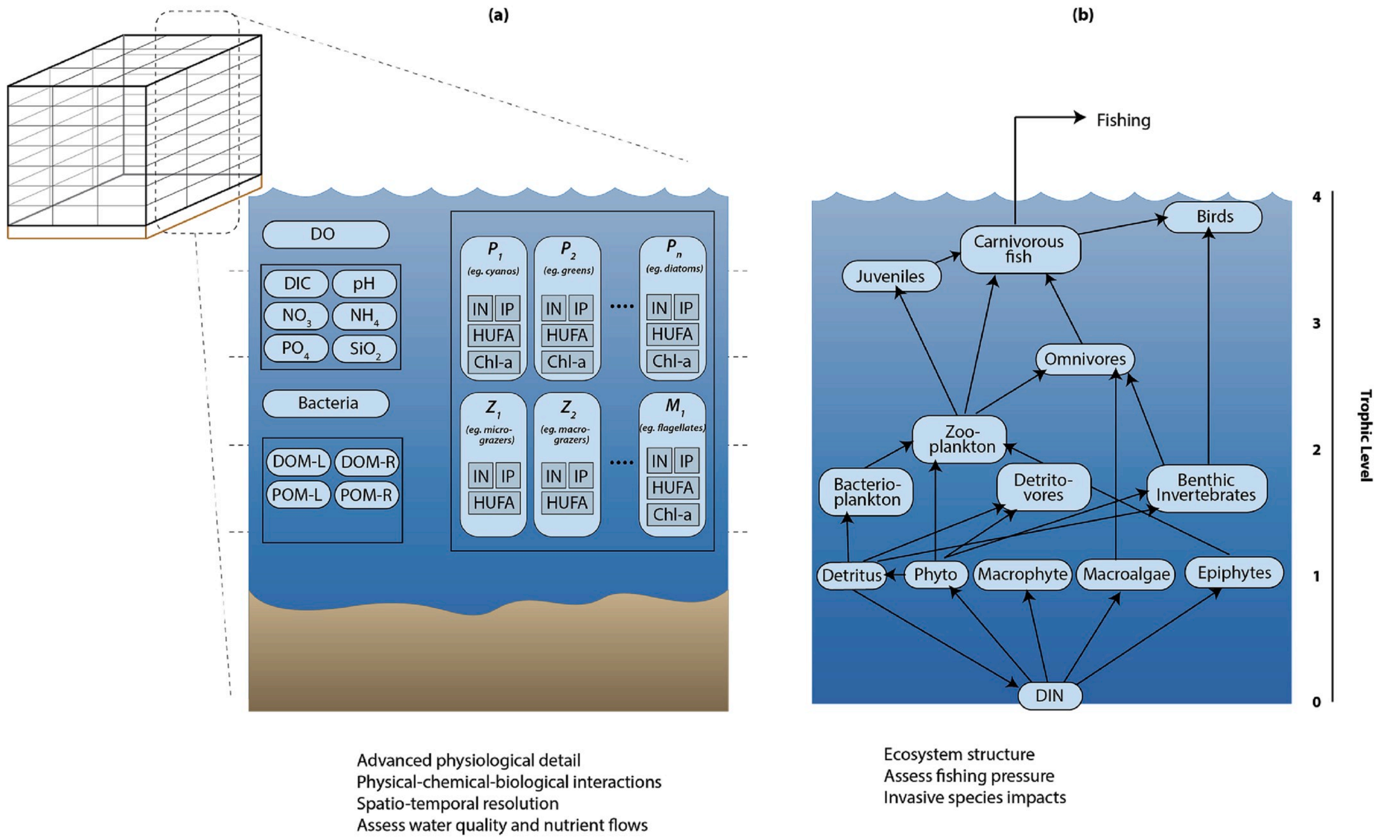

Fig. 1. Examples of two different aquatic ecosystem model conceptualisations, focused on (a) resolution of physical and biogeochemical processes and interactions, and (b) resolving trophic complexity and interactions. Both examples have been used in both fresh and marine studies.

advances in scientific understanding (Flynn, 2005; Anderson and Mitra, 2010; Oliver et al., 2012; Hellweger, 2017), are not interdisciplinary (Mooij et al., 2010; Ward et al., 2019), fail to assess or reign in uncertainty (Arhonditsis et al., 2006, 2008a; Dietzal and Reichert, 2012), and "fail to fail" when they should fail if they were true tests of conceptual understanding (Franks, 2009). Despite continuing advances in process understanding of aquatic biogeochemistry and ecology, and the emergence of a plethora of model approaches and platforms in the literature, it could be argued that the level of predictability in many practical applications of AEMs has not significantly improved over the past two decades (Arhonditsis et al., 2014).

These challenges motivate us to find new ways to assess our models so that we can transparently compare different models and model approaches, whether they are simple or complex, and understand the level of predictability they provide. Oreskes et al. (1994) argued the use of models is heuristic, which is consistent with a common view that decisions about defining when a model is "validated" are largely *ad hoc.* The chosen level of validation often depends on available pre-existing data sets, the background and experience of the individual modeller and a general desire to report favourably on the performance of the model. A common framework and established standards for model assessment and documentation would create opportunities for synthesis between diverse model studies and transferability of knowledge between applications. This would facilitate us to benchmark AEMs, that is, to compare which models are better under which circumstances, and what level of complexity tends to achieve a given level of accuracy and performance for a given application context.

Here, we review current approaches that can be used to assess the performance of AEMs, and formalize a general strategy to improve confidence in model predictions. The shortcomings of current model assessment and the need for a systematic approach are detailed further in Section 2. A framework is outlined in Section 3 for the hierarchical assessment of models to encourage modellers to assess not only state variable predictions, but also process behaviour and system-scale dynamics. A range of metrics and characteristic signatures relevant to aquatic ecosystem condition are exemplified for a range of physical, chemical and ecological contexts in Section 4, spanning the diversity of aquatic system models - from ponds and lakes to the global ocean. Our goal is to demonstrate the utility of hierarchical assessment to more comprehensively assess when a model is "fit for purpose". In doing so, our aim is to create standards, a common vocabulary, and encourage a more rigorous, multi-level approach to model assessment that will facilitate comparisons among different AEMs, their applications, and thereby increase their predictive value and usefulness.

2. Approaches to model assessment and need for a standard framework

In a review of model assessment frameworks, Pohjola et al. (2013) highlighted four kinds of model assessment: (i) quality assurance (best practise procedures); (ii) uncertainty analysis; (iii) technical assessment; and (iv) assessment of effectiveness in achieving social, environmental, or policy outcomes. In this paper, we focus on the technical assessment of model performance. We note here the extensive history of literature describing appropriate measures of fit for objectively evaluating model performance (Mayer and Butler, 1993; Power, 1993; Alewell and Manderscheid, 1998; Stow et al., 2003, 2009; Bennett et al., 2013; de Mora et al., 2013; Kubicek et al., 2015), providing guidance on specific mathematical measures of model accuracy. In general, these include varied methods for error calculation, correlation and model efficiency measures (Table 1). We do not focus this analysis on the specific suitability of these measures, but rather on providing a framework within which they can be used. For a review of the strengths and weaknesses of different measures of model-data comparison see the summary by

Table 1
Summary of quantitative techniques used for assessment of aquatic ecosystem models. Refer to Bennett et al. (2013) for more detailed overview and categorisations of available assessment approaches.

Abbreviation	Technique	Description
V	*Visual inspection*	*Visual inspection of time-series,* ***TS****, is often undertaken ("chi-by-eye"), but weak relative to the quantitative metrics listed below.* *Visual inspection of more complex model outputs may be warranted, and is frequently undertaken, for example:* ***TZ****: Time vs Depth for vertical profile assessment* ***XZ****: Distance vs Depth for cross section ("curtain") assessment* ***XY****: Plan view spatial ("sheet") comparison* ***TX****: Time vs Distance contour comparison*
BIAS, MAE, NMAE	*Bias Mean Absolute Error Normalised Mean Absolute Error*	$BIAS = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i);$ $MAE = \frac{1}{n}\sum_{i=1}^{n}(\lvert P_i - O_i\rvert);$ $NMAE = \frac{MAE}{\overline{O}};$ *Can be applied to demonstrate localisation of error in spatial models, denoted as MAE(ΔXZ), for example.*
RMSE	*Root Mean Square Error*	$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}}$
MEF, NSE, B	*Model Efficiency, Nash-Sutcliffe Efficiency, Bardsley coefficient*	$MEF = NSE = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2}$ Nash and Sutcliffe (1970) *&* Murphy and Epstein (1989); Bardsley (2013) *presents a variant that better accounts for bias:* $B = \frac{R^2}{2 - MEF}$
d_2	*Index of Agreement Model Skill Score Willmott index*	$d_2 = MSM = \frac{\sum_{i=1}^{N}\lvert P_i - O_i\rvert^2}{\sum_{i=1}^{N}(\lvert P_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert)^2}$ Willmott (1981) *introduces the above skill score to consider both correlation and variance, allowing a choice regarding how to weight error at extremes versus around the mean.*
R, SR	*Correlation coefficient Spearman Rank Correlation*	$R = \frac{\sum_{i=1}^{N}(P_i - \overline{P})(O_i - \overline{O})}{\left[\sum_{i=1}^{N}(P_i - \overline{P})^2\sum_{i=1}^{N}(O_i - \overline{O})^2\right]^{1/2}}$ $SR = 1 - 6\frac{\sum_n \tau^2}{n(n^2 - 1)}$ *where τ is the difference in rank between the predicted and observed*
TD	*Taylor/Target Diagram*	*Visualisations of patterns of statistics to assess and quantify model performance; see* Taylor (2001) *and* Jolliff et al. (2009).
DF	*Distribution Functions*	*Plots showing variance of data set, including box-plots, violin plots and cumulative distributions*
FFT, WT, WC	*Fast Fourier Transform Wavelet Transform Wavelet Coherence*	*Approaches to demonstrate spectral power localisation at distinct frequencies; Wavelet transform demonstrates this localisation as a function of time (or space); Wavelet coherency computes the correlation in wavelet power between variables.*
CCF, ACF	*Cross-correlation Function Auto-correlation Function*	*Measure of the similarity of a series to itself (autocorrelation) or another series (cross-correlation) as a function of displacement in time.*

Bennett et al. (2013) and previous discussions by others (Elliott et al., 2000; Allen et al., 2007).

To-date, no well-established guidelines have been developed on how to approach aquatic ecosystem simulation and how to decide when models are fit for purpose. The reason can be found in the large diversity of model approaches, which is compounded by the fact that AEMs have been applied over wide environmental (lake, river, ocean) and application contexts (e.g., forecasting, system understanding, scenario comparison etc.) (Janssen et al., 2015). For any given application context, typical approaches in the literature adopt a wide range of variables, spatial dimensionalities, spatial scales and simulation time-frames. Whilst the diversity of these applications is a good thing overall, the consequence has been that it is difficult to compare model performance between studies and approaches in a way that allows a clear definition of the limits of their predictions.

In the past decades there has been limited true improvement in the level of model predictability. Given the rapid uptake of AEMs (e.g. Trolle et al., 2012; Janssen et al., 2015), it would be expected that over time, higher quality datasets, more refined model process descriptions and increasing computer power would lead to improved predictions. This is of course true in some cases, however, there is evidence that in general terms model approaches and their ability to accurately capture trends in observed data have not considerably improved compared to some of the pioneering studies (e.g., Thomann and Fitzpatrick, 1982). This trend was first noted by Arhonditsis and Brett (2004) and subsequent analyses have pointed to a similar conclusion (e.g., Arhonditsis et al., 2006; Robson, 2014b; Paraska et al., 2014; Arhonditsis et al., 2014). During routine applications of models, we anecdotally hear that use of R^2 is a "waste of time" for coupled hydrodynamic-biogeochemical model variables, and that modellers often resort to "chi-by-eye". That is, the modeller analyses the model output in the context of the quality and noise in the data and uses a subjective visual comparison as the ultimate determinant of suitability. Kubicek et al. (2015) found that 92% of studies reported in the journal, *Ecological Modelling*, reported visual inspections as the only or main method of model evaluation. This approach is not always due to the lack of willingness by the modeller to validate more deeply, but rather due to confusion as to what features need to be tested within the specific application. The question therefore remains: are there ways in which we can further refine our efforts so that they may lead to more robust model predictions?

A large challenge remains to improve parameterisation of AEMs. For most applications, information exists about the values and variability of the state variables, but little is known about the values of the model parameters. Modellers then adjust parameters to find the best agreement between modelled and observed data, either by adjusting the model parameter vector by trial and error (manual calibration) or through optimisation algorithms. The advantage of optimisation methods is that they are objective and repeatable methodologies that are more likely to result in an optimal parameter set. Any significant lack of fit is then due to the inadequacy of the model structure and not due to poor parameter choice (Chapra and Canale, 2010). The conventional practice of seeking a single "optimal" parameter set reflects two major assumptions: (i) that there exists a single calibration vector for faithfully reproducing a wide range of ecosystem dynamics; and (ii) that our empirical knowledge (monitoring data, experimental work) adequately depicts the patterns of the "real world", and thus offers an objective standard for testing our models. The credibility of both statements has been extensively debated in the literature and there are sound arguments to cast doubt on the legitimacy of such a deterministic approach to mathematical modelling. A recent attempt to address knowledge gaps and improve practice in setting parameter values has been made by Robson et al. (2018), who provide a tool to facilitate modellers' exploration of evidence for process rates and traits relevant to biogeochemical model parameters.

Nonetheless, model practitioners often encounter the problem that several distinct choices of model inputs result in equivalent model outputs, i.e. many sets of parameters fit the data equally well. The non-uniqueness of the model solutions, known as equifinality (Arhonditsis et al., 2008a), is a consequence of insufficient data or in the case when internal process pathways are of substantially higher order than what can be externally observed (Beck, 1987). As a result, our ability to set quantitative (or even qualitative) constraints on model ecological structure is significantly reduced, and thus we are often faced with a situation whereby our models give "good results for the wrong reasons" (Arhonditsis et al., 2007).

Aside from parameter uncertainty, models contain errors that arise from its structure or its inputs (Omlin and Reichert, 1999). Model structural error is associated with (i) errors in the selection of appropriate state variables or processes to reproduce ecosystem behaviour, (ii) errors and necessary simplifications in selection of mathematical formulations for describing the processes, and (iii) the fact that our models are based on equations derived from controlled laboratory environments that may not yield an accurate picture of the real world variability in biological systems and complicated interactions between forcing factors (Hellweger, 2017). Essentially, models are simplifications of reality, and all parameters are effectively applied as spatially and temporally averaged values that in reality are unlikely to be represented by fixed constants. In addition, it should be recognised that observational data are also uncertain approximations, and are in fact also models of reality.

An important and increasing area of model application is to capture shifts in system function. Shifts occur in response to a varied range of external drivers, such as climate change, the cumulative loading of nutrients and pollutants and/or management measures (e.g., Trolle et al., 2008; Skerratt et al., 2013). In this case, modelled ecosystems are non-stationary; they are being pushed out of their typical state-space range upon which they were trained and predictability in the past is no guarantee a model can capture future trajectories. From the broader ecological literature, we know that ecosystems are vulnerable to deterioration when key system functions are pushed over thresholds, resulting in the loss of resilience and the emergence of a regime shift (Scheffer et al., 2009). Often these dynamics are at the core of what we need models to help us understand, yet we have limited confidence that the models are capturing these shifts and non-linear ecosystem dynamics (Hipsey et al., 2015).

The lack of universally accepted performance criteria impedes our capacity to impartially determine what an acceptable model is. Thus, an emerging imperative in the field of aquatic systems modelling is the development of a predetermined standard, considering model complexity, the spatiotemporal domain or even the question being asked. Given the complexities highlighted above and the diversity of simulation contexts, this is unlikely to take the form of a simple cut-off value for an acceptable goodness-of-fit metric. In some cases, it is sufficient to be able to confidently predict that one management option will have a better outcome than another (e.g. shorter algal bloom duration), while in other cases, the decision will hinge on being able to predict *how much better* a certain scenario may be. Hence, we need to be able to evaluate not only how uncertain our prediction may be, but also what our models can confidently predict (e.g. that a bloom will occur) despite prediction uncertainty.

In light of these conceptual and technical challenges, there are several areas where the AEM community would benefit from improved tests and reporting of model performance. These include:

- greater emphasis in model publications to highlight the assessment approach and the variables validated (or tested in sensitivity analysis);
- adoption of assessment standards to facilitate inter-comparison of diverse model approaches;
- improved assessment of process pathways in models as a means to help resolve concerns around equifinality, including assessment of spatio-temporal variability in process rates;
- exploration of the degree to which different scales of variability are captured by models, considering not just state variables, but also flux pathways;
- assessment of model performance in reproducing theoretically relevant, system-scale responses – that is, even if models capture trends at a sampling point, they must also demonstrate ability to capture emergent behaviours; and
- employing a wider range of validation approaches able to accommodate new monitoring technologies and high-frequency data streams to support model-data fusion efforts.

Addressing the criteria above will improve credibility and transparency in aquatic ecosystem modelling. For example, capturing emergent properties is particularly relevant when models are used to explore non-stationarity (systems undergoing change), where a model may be out of its calibrated range, or if the goal is to explore uncertain future conditions like climate change. We used the list to formulate the core principles of a comprehensive model assessment framework that can be used to understand, discuss and evaluate underlying principles in AEMs, and to broaden their applicability and transferability.

3. Overview of the multi-level assessment framework

The CSPS framework introduces a hierarchical assessment of a range of metrics and theoretically-relevant signatures relevant to aquatic ecosystem structure and function, depicted schematically in Fig. 2. The approach includes four levels of assessment, with an *a- priori* or pre-application assessment of the model (indicated as Level 0), while the other three levels are post-simulation assessments of the model. The four levels are summarised as:

0. ***Concept***: Conceptual validation to ensure that sub-models are consistent with ecological theory and valid over the range of conditions for which the model will be applied;
1. ***State***: Comparison of simulated state variables with observed properties;
2. ***Process***: Comparison of simulated energy and mass fluxes with measured process rates; and
3. ***System***: Comparison of system-scale emergent properties, patterns and relationships with observed and theorised phenomena.

These levels are further described generically below given our desire for consistency across both inland and marine waters. Specific examples for different application contexts and specialisations relevant to the modelling community are expanded upon in Section 4.

3.1. Level zero: Conceptual validation

A process for model assessment that includes conceptual validation of model structure and sub-model algorithms is a precursor to empirical validation of the model as a whole (Bert et al., 2014). At this level, we ask, "does the conceptual basis of the model accord with current scientific understanding of how the system functions?" This level of evaluation is usually undertaken explicitly when developing a new model,

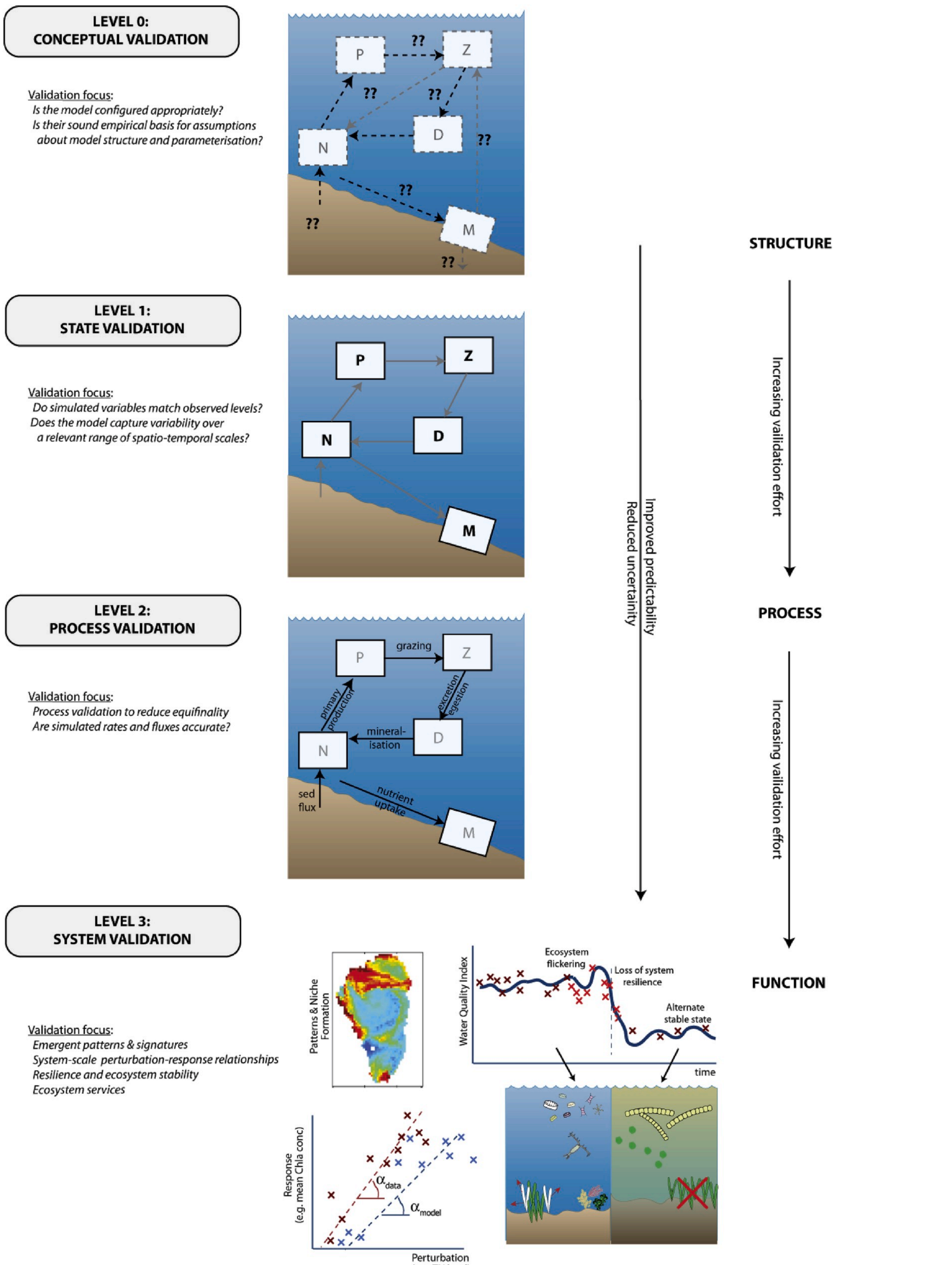

Fig. 2. Conceptual diagram illustrating the CSPS model and assessment workflow highlighting the shifting focus from assessment of state, to process, and ultimately to ecosystem function.

but is often overlooked when applying an existing model in a new context or when it is adapted to include additional functionality. Questions to consider include:

- Does the model structure adequately reflect our conceptual understanding of the system and its key drivers, having regard for processes that may change within the range of scenarios being considered? An example of a situation in which a previously successful model may fail this conceptual validation is when a model developed for a deep-sea application is to be applied to a coastal ecosystem, where a more detailed representation of benthic processes is needed. Another example is where a model previously applied within a narrow range of temperature conditions is to be applied to climate change scenarios and the conceptual basis of temperature response functions may need to be reconsidered.
- Does the mathematical representation of ecosystem processes produce realistic system dynamics? For example, the widely-used Michaelis-Menten kinetics are best suited to steady-state conditions and may be considered dysfunctional in dynamic simulations where nutrient concentrations vary considerably (Flynn, 2005; Frassl et al., 2014; Hellweger, 2017).
- Does the model structure reflect recent advances in ecosystem understanding that may be important in this application? For example, many ecosystem models in common use have not been updated to include anammox or dissimilatory nitrate reduction to ammonium (Robson, 2014a,b).
- Does the model reflect the system understanding of relevant local disciplinary experts and stakeholders? If not, this may be a hurdle to acceptance of decisions based on model results, as well as not making good use of local knowledge.
- Is the model mathematically valid and dimensionally consistent?
- Is their evidence that the implementation of the model as software correctly reflects its conceptual and mathematical basis? For example, does it maintain conservation of mass?
- Where two or more possible model structures have been identified, what process has been followed to compare the options? In some cases, it may be appropriate to develop an ensemble of models to test the range of possible results and the trade-offs in speed and accuracy.

This conceptual validation phase should be revisited after state, process and system validation to consider whether the results suggest the need for re-evaluation of the model structure or implementation.

3.2. Level one: State validation

The comparison of time-series of physical, chemical and biological state variables is the main form of model validation. However, it is not the only means by which the accuracy of model state can be assessed, and when used in isolation may not give a complete picture of model performance. In most cases, the frequency of observations is significantly less than the time-step of aquatic models, particularly in the case of coupled hydrodynamic-biogeochemical models. For variables that exhibit rapid changes in time, such as algal biomass during a bloom, standard metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) can be fundamentally weak (Elliott et al., 2000). Similarly, in terms of a spatial distribution, an *in situ* observation is usually the mean of a small sample volume, being compared to model output that may represent the mean value of a much larger quantity of water. Various alternatives for assessing state can be considered, including assessment of scales of variability, and derived metrics, whereby observations and simulated data are subject to some form of transformation.

3.2.1. 1a – Direct comparison

Classically, model simulation results are compared with measured data at specific points in time and space where the measurement was taken. This type of assessment is undisputedly where most effort in model validation has been concentrated and will no doubt remain the focus of most model assessments. The variables selected for assessment are inherently linked to the nature of the investigation and associated choice of model approach and structural complexity. Nonetheless, there has been a tendency for modellers to validate against an arguably small subset of simulated water chemistry variables, a product of both a desire to emphasise aspects of model output relevant to the application, but also commonly due to lack of observations for remaining state variables.

Measures of model fit (as in Table 1) can be computed for one or more sampling stations and are generally based on calculation of residuals (i.e., the difference between model outputs and the corresponding observations). A potentially useful exercise that adds to the common error calculation, is to calculate skewness or kurtosis in predictions. Spatial assessments of multi-dimensional models against data from remote sensing platforms are increasingly being introduced. Other methods do not necessarily involve error calculation but rather assess patterns in data series. For example, Spearman Rank correlation (SR) may be more useful to identify the degree to which the order of predictions and observations from small to large magnitude are captured, for example, where assessment of the ability of a model to reproduce seasonal and inter-annual variability is required without focusing on exact values. The use of a cross correlation function (CCF), can allow a modeller to look for correlations in how the simulated and observed data vary with a delay across time and may be used for looking at lags in time-series, or alternatively may be useful in spatially resolved models to determine measures such as patch length.

3.2.2. 1b – Derived metrics describing model state

This category refers to metrics that do not involve a direct assessment of state variable time-series or spatial data, but are derived from the simulated variables. The focus of this class of metrics is to test the model against theoretically relevant indicators of ecosystem state, such as relationships between variables. They can provide additional evidence for demonstrating that a model is fit for purpose even if direct value comparisons (e.g., R^2) are not possible due to data limitations or if direct validation indicated a weak level of predictability. Examples of derived metrics could include simple stoichiometric indicators (e.g., TN:TP), or other ratios of simulated variables (e.g., DOC:TOC; Chl-a:TSS). Other derived quantities include assessment of relevant dimensionless numbers, for example, the Richardson number as a measure of stratification intensity, or in the case of algal bloom dynamics, metrics derived from analysis of the raw state variable time-series, such as the average duration of a bloom or time of bloom onset.

3.2.3. 1c – Metrics describing multi-scale variability in model state

Metrics at this level describe how well various scales of temporal or spatial variability are reproduced in our models. Depending on the model structure, its spatial dimension and time-step, any given simulation will have limits on the scales that it can predict. The inherent difficulty in defining what model approaches are most appropriate for a given scale (e.g., at what temporal scale would a 1D and 3D model converge?) remains a large challenge in the AEM community.

Simulations may be assessed from a probabilistic point of view, for example, through comparison of exceedance probabilities (CDFs) in order to ascertain whether a model is able to capture the proportion of time (or spatial domain) over which a certain concentration is experienced. However, assessment of probability distributions alone may not be suited under non-stationary conditions, and they do not inform us if expected modes of variability are adequately represented. For example, a coupled hydrodynamic-biogeochemical model might be designed to reproduce phytoplankton biomass in response to diurnal changes in productivity, but also over intermediate scales due to dynamic hydrological and meteorological conditions, over seasonal scales due to changes in temperature and nutrients, and potentially up to decadal scales if the fundamental drivers of phytoplankton biomass are shifting.

Approaches to assess the performance of models over these scales are described in Bennett et al. (2013) as “data transformation methods” and relevant techniques include Fourier transformation (FFT) or wavelet transforms (WT). To date there has been limited application of these methods for assessing AEMs, partly due to the fact that there are few cases of modelled applications that have observational data that span such a large range of time-scales (reviewed by Kara et al., 2012). The routine application of real-time sensors in aquatic systems over the past decade, however, is providing diverse datasets that span from time-scales of minutes to decades, and offer new opportunities to assess models through this approach (e.g., Hamilton et al., 2015). Where spatially rich observational data are available, these approaches can also be used to examine model performance across multiple spatial scales. Spatial maps can be quantitatively compared with observational maps, particularly from remote sensing observations, using a variety of methods (Stow et al. (2009).

3.3. Level two: Process validation

Process validation refers to assessment of model performance against the underlying rates of transformation that drive changes in model state variables (i.e., the arrows connecting “stocks” or “pools”); process validation is therefore specific to process-based models. Indeed, the most commonly cited advantage of process models is their ability to resolve the interaction of the different mechanisms that shape ecosystem state, yet rarely do we rigorously assess whether they are correctly captured. This is particularly relevant if we consider that most model applications adopt process parameterisations reported broadly in the literature, potentially from sites that may be inherently different, or from laboratory or mesocosm studies conducted under controlled conditions. Similarly, the associated parameter estimates for these algorithms may also be chosen from within large ranges reported from diverse model applications, or from laboratory assessments such as phytoplankton or sediment incubations. Therefore, it is not obvious that models with complex interactions accurately represent spatial and temporal variability in process pathways correctly, even if several model state variables are seemingly reproduced well at Level 1. Consequently, validating models with regard to the individual flux pathways that connect individual state variables is a way of reducing equifinality, helping modellers to get “good results for the right reasons”. Comparing modelled with measured flux rates also provides a way to pinpoint sources of structural and conceptual error in the model.

In practice, this approach remains rare in aquatic ecosystem modelling, as measuring variability in process rates through time and/or space is resource-intensive and difficult, and in some cases may not yet be directly possible *in situ*. However, given that it has the potential to greatly improve confidence in the underlying function of models, we believe it should be actively promoted, and several examples are outlined in the following sections that may be more routinely adopted. These are classified next as either being from direct measurements or indirect rate estimates.

3.3.1. 2a – Comparison with raw process measurements

When developing a model, it is necessary for us to distinguish between process measurements that are required to assist model parameterisation and parameter assignment, and process measurements that can be used for validation. A range of *in situ* process measurements are relevant to assess physical, chemical and biological model attributes and their temporal and spatial variation. These include rates of mixing, fluxes across the air-water or sediment-water interfaces, and kinetic transformations (e.g. nutrient uptake, rates of primary production, or grazing). Specific examples highlighted in Section 4 relevant to a range of different model applications are reviewed.

3.3.2. 2b – Process metrics interpreted from raw data

Process information can also be extracted from raw data series, either derived from changes in the observed data record via inverse modelling, or potentially through more sophisticated data-driven models designed to estimate bulk process rates. As a simple example, the trend of oxygen depletion in the hypolimnion of a lake may be used to estimate the net sediment oxygen demand if other sinks are minor. The use of environmental tracers and isotopic data also has potential to support validation of the flow of oxygen, carbon or nitrogen, for example, if assumptions are made about the relative fractionation and transformation rates that occur for individual process pathways.

These approaches are not reported widely in the literature, and usually depend on the validity of several simplifying assumptions that are made when interpreting the data. However, it is highlighted here as an area of increasing interest for the growing area of model-data fusion, as a means to compare modelled process rates against those estimated from empirical means (Robson, 2014a; Hipsey et al., 2015).

3.4. Level three: System validation

The complex non-linear interactions and feedbacks that govern the response of aquatic systems to changes in internal or external conditions can lead to ecosystem-scale emergent patterns, relationships and dynamics. These emergent properties are not necessarily predictable directly from the underlying model formulation, and are outcomes from the model that are “*not a direct extrapolation of the choices made in model design*” (Allen, 2010). A classic example of an emergent property is the behaviour of a flock of birds in flight, which emerges in a way that is not obvious from the behaviours of individual birds.

Although not widespread, there are several examples where aquatic models have been assessed in terms of their ability to produce emergent dynamics, though generally not with the direct purpose to refine model accuracy or to justify whether it is fit for purpose. These provide valuable insights to ecosystem behaviour, and we advocate for more effort to be placed in assessing if our models are able to capture higher order behaviour or patterns. As highlighted by Anderson et al. (2010), a different choice of model structure may lead to different emergent dynamics. Where two models perform equally well at Level 1 but predict different emergent system dynamics, these can be treated as competing hypotheses regarding system dynamics, and measurement programmes can be devised to invalidate one or both hypotheses.

Examples of system properties that might be captured by models can include simple metrics such as scaling relationships (e.g. nutrient loading vs. chlorophyll-a response), or more complex spatial or temporal patterns in nutrient cycles and community dynamics. In many applications, these patterns may be expected based on empirical experience, such as the succession of different plankton functional groups, or spatial niches in a habitat. Multivariate comparison methods are available to explore performance of models in capturing inter-relationships between variables (e.g., a Taylor Diagram, TD), and these may be particularly useful for identifying cases where models resolve emergent dynamics that are not known *a priori*. Self Organising Mapping (SOM) is one example, among others, of a machine learning method that may be suited to identification of emergent patterns in both model output and observational data that are rich in two or more dimensions (e.g., Williams et al., 2014). Where observational data are both spatially and temporally rich, Empirical Orthogonal Function (EOF) decomposition can be used to analyse major modes of variance and patterns of variation, which can be compared with the results of the same analysis applied to model output (e.g. Rocha et al., 2019).

Another area being explored is the response of ecosystem state-space to perturbations, considering threshold effects, hysteresis and alternative stable states. This level of validation is especially important if the model’s purpose is to define the stability or resilience of key ecosystem attributes to climate change, fishing and/or eutrophication. At this stage only a qualitative or semi-quantitative comparison may be possible.

4. What metrics should I use when … ? Examples for different application contexts

In this section, we consider the literature through the lens of the framework outlined in Section 3, by combining the framework proposed with specific assessment techniques summarised in Table 1. Given the predominance of Level 1 validation in the literature, we did not target a comprehensive identification of all published model studies that consider validation. Rather, we aimed to collate a range of examples that have been applied across a broad range of application contexts. Through this review, we also aimed to identify gaps and potential areas for further development of new Level 2 and 3 validation approaches.

4.1. Hydrodynamic applications for ecosystem assessment

Relative to water quality and ecosystem modelling applications, approaches for characterizing the performance of physical models of aquatic environments are well established, most notably from the engineering and geophysical sciences literature. They share some common metrics across lacustrine, riverine and oceanographic model applications, depending on the underlying hydrology, model dimension and time-frame of the simulation. A range of general metrics have been categorised related to prediction of a) the water balance, waves and water circulation, b) the heat and salt balance, c) stratification, and d) bottom morphometry and sediment transport (Table 2).

In general, the validation of hydrodynamic models is achieved by conducting multiple levels of assessment. First, researchers can conduct time-series assessment of water level, temperature and salinity at fixed points, and/or horizontal or vertical variation derived from profile cross sections (Level 1). Simulation of surface ice dynamics can also be undertaken by time-series assessments of ice thickness, complemented with derived metrics such as ice-on and ice-off dates (Level 1, e.g., Yao et al., 2014). For models that simulate surface or internal waves, spectral plots are commonly reported to demonstrate power across waves of different frequency. Derived indices such as the Richardson number as a measure of stratification intensity, or Schmidt stability (e.g., Bruce et al., 2018), are also useful as Level 1 metrics. Considering mixed layer depth (Acreman and Jeffery, 2007; Bayer et al., 2013), or thermocline/-pycnocline thickness and changes to surface layer thickness as a function of various driving factors, can prove useful in diagnosing problems with model parameterisation. For cases where periodicity varies over time, wavelet plots can highlight how power is localised in frequency space over different seasons or time-frames. Isotherm/isopycnal displacement power spectra are a useful metric to demonstrate the time-scale over which models are able to reproduce the internal wave field (e.g. Hodges et al., 2000).

Level 2 validation of hydrodynamic models can include assessment of evaporative mass fluxes, estimation of albedo, measurement of velocities (at a fixed location or from drifter tracks, e.g., Dissanayake et al., 2019), and shear stresses, for example, at the sediment-water interface or within macrophyte beds. The use of direct measurements or observation-based estimates of turbulent mixing could also be considered, and a range of novel tracers have been adopted for dilution experiments to characterise contaminant dispersion (e.g., caffeine and pharmaceuticals in waters impacted by wastewater effluent, Cantwell et al., 2016).

Level 3 assessments in hydrodynamic models consider the formation of residual currents and eddy structures. For example, different, but otherwise similar, models reveal the emergence of different eddy structures in the North Atlantic Ocean (Holt et al., 2014). Hetland and DiMarco (2012) undertook an assessment of a 3D hydrodynamic model of the Texas–Louisiana continental shelf using data from moorings by presenting maps of model skill; in this example, surface and bottom variance ellipses are used to demonstrate the model captures the point-scale and residual field. In lakes and coastal environments, currents created by differential heating and cooling may be assessed by indirectly comparing against profile cross sections (Woodward et al., 2017). Spatial variability in water currents creates patterns of water age distributions that emerge as a system-scale property, but which are not easily able to be validated. The potential for using methods such as assessing against empirical estimates from conservative tracers, for example from radium isotope measurements (Tomasky-Holmes et al., 2013), may be able to be applied in the future. The increasing applications of models for complex aquatic domains (e.g. estuaries, inter-tidal wetlands, river floodplains, reef structures, or island archipelagos) requires efforts to validate the relative pattern of connectivity across modelled sub-domains.

At smaller spatial scales, Level 3 assessment of other patterns that may emerge in physical models can be conducted. Examples include travelling waves in spatial patterns of plant-wrack in intertidal zones (Sun et al., 2010), wave attenuation in seagrass beds (Chen et al., 2007), and the canopy structure such as the dynamics of canopy deflection properties (Dijkstra and Uittenbogaard, 2010). Models with morphodynamic ability can also be assessed by examining spatial patterns in temporal changes in bottom morphometry, to highlight active areas of erosion and deposition.

4.2. Water quality and biogeochemistry

A common goal for AEMs is to understand the controls and dynamics of chemical and biological variables relevant to water quality. What constitutes a ‘water quality variable’ can vary depending on the application and context, however, for the purpose of this analysis, the literature is categorised according to several areas of focus pertaining to prediction of a) oxygen and the extent of hypoxia/anoxia, b) the cycling of inorganic nutrients and organic matter, c) geochemistry, d) water colour and clarity e) chlorophyll-a, and f) other chemical and biological contaminant dynamics (Table 3). Across these categories most variables being assessed are dissolved or particulate concentrations that are routinely sampled via traditional monitoring programs and subsequent time-series assessments (Level 1), though, a more detailed exploration reveals a broader range of examples that can support a diversity of potential approaches for capturing water column and sediment biogeochemistry.

Oxygen has often been a focus of water quality models as it plays a pivotal role in nutrient cycling and sediment processes. Given the increasing availability of sensor data for measuring oxygen concentrations, it also provides an interesting case study for how a system of metrics can assist in model assessment. For this context, a range of more rigorous Level 1 metrics has emerged such as wavelet analysis of high-frequency *in situ* data series from a stratified lake (Kara et al., 2012), longitudinal analysis of surface and bottom oxygen in estuaries where strong lateral gradients exist (Xu and Hood, 2006), and the spatiotemporal extent of anoxia in a riverine estuary (Bruce et al., 2014) and Lake Erie (Bocaniov et al., 2016). Direct *in situ* Level 2 sediment flux measurements from benthic chambers or eddy-correlation instruments have provided useful validation of sediment oxygen demand (e.g., Sohma et al., 2008). In another example, Hetland and DiMarco (2008) adopted apparent oxygen utilisation (AOU), the difference between the oxygen concentration and the saturation value, as an indirect process validation metric to demonstrate the model was capturing the combination of oxygen consumption mechanisms. Of increasing interest in lacustrine and marine environments is the application of high-frequency oxygen sensors to estimate free water metabolism (Hanson et al., 2008), where an inverse modelling technique is used to extract hourly to daily estimates of primary productivity, community respiration and atmospheric exchange from diel changes in oxygen concentration (e.g. Lovato et al., 2013; Webster et al., 2005; Wikner et al., 2013; Winslow et al., 2016). While this method provides relatively coarse estimates due to confounding factors of advection and mixing (Villamizar et al., 2014), repeated estimates over a range of environmental conditions provide an *in situ* view of water column net productivity and respiration that can be

Table 2
Summary of validation metrics for physical models of aquatic systems (refer to Table 1 for assessment technique abbreviations).

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
Water balance, waves & circulation							
Water level	Time-series comparison	1a	minutes-monthly	point; multiple points	E, R, V(TS), V(XY)	Direct observation or calculation from logged pressure gauge sensors, or from remote sensing approaches (e.g., satellite altimetry, radar)	Missaghi and Hondzo (2010)
	Tidal propagation	2b	minutes-hourly	horizontal transect	V(TX)	Magnitude of attenuation or amplification of tidal range within an estuary or coastal embayment, plotted as a function of distance	
Surface waves	Significant wave height	1a	seconds-minutes	point	V(TS), E, R	For models simulating surface waves the comparison of wave properties can be undertaken	Ji (2017)
	Wave length and period	1b	seconds-minutes	point	V(TS)		
	Frequency spectra	1c	seconds-minutes	point	FFT, WT		
Evaporation	Time-series comparison	2a	minutes-daily	point	V(TS), E, R	Evaporative mass flux data can be collected from an evaporation pan, or flux anemometer	Rimmer et al. (2009)
	Time-series comparison	2b	minutes-hourly	point	V(TS), E, R	Comparison against latent heat fluxes derived from energy balance fitting to surface meterological data	Nussboim et al. (2017)
	H_2O isotopes	2b	*ad hoc*	point	V(other)	Fitting isotopic data can help source identification and compute evaporation rates based on deviation of meteoric water line	Stadnyk et al. (2013)
Velocity	Time-series comparison	1a	hourly-weekly	point; horizontal transect	V(TS), E, R V(XZ), MAE(ΔXZ),	Use of point ADCP measurements for point scale or cross section	
	Variance ellipse	1c	hourly-weekly	point	V(other)	Summary of magnitude and direction of current field that can be compared with point data	Hetland and DiMarco (2012)
	Residual currents	3	weekly-seasonal	surface layer; horizontal transect	V(XY), V(XZ)	Particle trajectories from model simulations can be compared with tracks from drogues and/or drifters released in the field.	Dissanayake et al. (2019)
Mixing	Mixing intensity	2a	*ad hoc*	vertical profile	V(other)	Turbulent diffusivities derived from SCAMP data can guide turbulence parameterisation	Rueda and MacIntyre (2010)
	Tracer dilution	2b	*ad hoc*	surface layer; horizontal transect	V(TS)	Capturing the horizontal and vertical dispersion of a conservative tracer (e.g., rhodamine or chloride) can ensure diffusion is being accurately captured	
Retention characteristics	Water age variation	3	*ad hoc*	multiple sites	E, R	The use of radioisotopes could be used to correlate simulated water age with observed estimates from geochemical tracers	
	Water source apportionment	3	*ad hoc*	multiple sites	V(other)	Use of conservative tracers indicating water source from specific surface or groundwater inputs or rainfall, e.g., caffeine, radon, etc.	
Heat & salt balance							
Temperature or salinity	Time-series comparison	1a	minutes-monthly	point	V(TS), E, R	Data measured from an *in situ* thermistor or salinity sensor, or *ad hoc* measurement	Most papers present this
	Frequency spectra	1c	minutes-hourly	point	FFT, WT	Data measured from a thermistor or salinity sensor logging at high frequency	Kara et al. (2012)
	Spatial comparison	1a	daily-monthly	surface layer	V(XY), MAE(ΔXY), d_2	Satellite acquired temp data (e.g., LANDSAT, MODIS etc) compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match	Spillman et al. (2007) Ménesguen et al. (2007)
	Spatial variability	1c	daily-monthly	surface layer	DF	Compares distribution and range of T or S variation within the simulated domain without conducting pixel by pixel comparison	
	Spatial patchiness	1c	daily-monthly	surface layer	CCF	Can assess similarity in spatial coherence of T or S	
	Eddy structure	3	hourly-monthly	water column	V(XY)	Visual comparison of the emergence of complex eddy structures and gyre formation in T or S fields	Holt et al. (2014)
Temperature	Albedo	2a	*ad hoc*	surface layer	V(TS), E, R	Models simulating spatiotemporal variability in albedo can validate against estimates computed via upwelling and downwelling pyranometer	
	Radiative heat flux	2a	*ad hoc*	surface layer or benthic layer	V(TS), E, R	Radiative heat flux across the surface of the water or at the sediment-water interface measured using eddy-correlation, microprofiles, IR measurements.	
	Benthic perimeter heat exchange	2b	daily-monthly	water column	V(other)	Rate of change of bottom (hyoplimnion) temperature	Salmon et al. (2017)
Ice cover	Ice thickness	1a	weeky-monthly	point	E, R	Change in ice thickness over time	

(*continued on next page*)

Table 2 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
	Date of ice on/off	1b	weeky-monthly	surface layer	BIAS, R, SR, DF	Capturing the date on which ice is formed or disappears from the lake is important for spring and autumn thermal dynamics and climate change impacts	Hipsey et al. (2019) Yao et al. (2014) DeStasio et al. (2015)
Stratification							
Temperature, salinity, density	Depth comparison	1a	minutes-monthly	vertical profile (continuous) or multiple depths (discrete)	V(TZ), R, d_2 MAE(ΔTZ),	TZ error contour plot highlights errors in thermocline or pycnocline depth, by comparing interpolated observation and model profiles over time	Ménesguen et al. (2007) Missaghi and Hondzo (2010)
	Duration of stratification	1b	hourly-seasonal	water column	BIAS, R, SR, DF	Capturing the total length of time a waterbody experiences stratification can be useful for understanding water quality and/or the impacts of climate change	Frassl et al. (2018)
	Date of water column mixing/over turn	1b	hourly-monthly	water column	BIAS, R, SR, DF	Capturing the specific date of water column overturn may be important when forecasting water quality in reservoirs, for example.	
	Lateral gradient	1b	hourly-daily	horizontal transect	V(XZ), MAE(ΔXZ)	Comparison on lateral gradient in stratification can be used to diagnose model performance in capturing density currents associated with differential surface forcing or boundary inputs	Woodward et al. (2017)
	Internal wave frequency spectra	1c	minutes-hourly	water column	FFT, WT, WC	Comparison of frequency spectra to demonstrate wave periods and modes are being reporeduced	Hodges et al. (2000)
Velocity	Depth comparison	1a	minutes-hourly	vertical profile	V(TZ), MAE(ΔTZ)	TZ error contour plot highlights mixing errors, by comparing ADCP data and modelled velocity profiles over time	
Layer structure	Surface mixed-layer depth	1b	daily-monthly	water column	V(TS), E, R	Capturing the mixed layer depth can aid in diagnosing mixing and heat balance problems	Bruce et al. (2018) Steyn and Oke (1982) Acreman and Jeffery (2007) Bayer et al. (2013)
	Metalimnion thickness	1b	daily-monthly	water column	V(TS), E, R	As above, the thickness of the thermocline (or pycnocline) region may assist in validating mixing in lake or ocean models	
	Bottom vs surface difference	1b	daily-monthly	2 layer	V(TS), E, R, V(TX)	Time-distance contour plot highlights errors in stratification horizontally, e.g., for assessing seasonal salt-wedge propagation in an estuary	Huang et al. (2018)
Layer stability	Richardson (Ri) number	1b	daily-monthly	2 layer	V(TS)	The (bulk) Richardson number can be estimated from surface and bottom densities and velocities, to give a quantitative measure of the buoyancy vs inertia forces controlling layer stability	Bruce et al. (2018)
	Schmidt stability	1b	daily-monthly	water column	V(TS), E, R	As above, the Schmidt stability parameter is useful for diagnosing the strength of lake stratification	
Bottom morphometry & sediment transport [c]							
Bottom stress	Time-series comparison	1b	minutes-hourly	point	V(TS), E, R	Stress derived from velocity profile measurements can be used to validate model the bottom stress impacting the rate of resuspension	
	Wave attenuation	2b	*ad hoc*	multiple points	V(other)	Wave driven resuspension is important in shallow systems and model validation could consider wave attenuation with depth and depending on the character of the benthic substrate	Chen et al. (2007)
Sediment movement	Resuspension rate	2a	*ad hoc*	point	V(TS), R	*In situ* experiments measuring resuspension rate can be compared under different hydrodynamic conditions to validate model rates	Sun et al. (2010)
	Rate of accumulation or erosion of benthic sediments	2b	monthly-decadal	point	V(TS)	For models simulating the change in bottom depth due to sedimentation or erosion, the relative rate of changed in depth measured using hydro-acoustic methods can be used to validate models	
		3	*ad hoc*	multiple points	V(other)		

(*continued on next page*)

Table 2 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
	Variation in particle size distribution					Spatial differences in particle size composition of bottom sediment can be used to validate areas of differential deposition rates between particle size classes.	
	Spatial changes in bathymetry	3	seasonal-decadal	bottom layer	V(XY), MAE(ΔXY), d_2	Can be used to compare model performance capturing spatial patterns in areas of net accumulation and erosion.	
	Wave length, height in sediment undulations	3	*ad hoc*	bottom layer	V(XY)	Complex patterns that emerge in fine-scale simulations of hydrodynamics and bottom sediment movement	

[a] E = user determined combination of traditional error metrics, including BIAS, MAE, NMAE, MEF, NSE, B and/or d_2.
[b] Where possible references provided indicate an example of the metric being applied, otherwise references relevant to give context to use of the metric are listed.
[c] Water clarity and light metrics are covered in Table 3.

Table 3
Summary of validation metrics relevant to models simulating water quality and biogeochemistry of aquatic systems (refer to Table 1 for assessment technique abbreviations).

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
Dissolved oxygen							
Dissolved oxygen concentration, or saturation	Time-series comparison	1a	minutes-monthly	point; multiple depths	V(TS), E, R	Data measured from an oxygen sensor, or *ad hoc* measurement	Carraro et al. (2012) Lovato et al. (2013) Zhu et al. (2016) Fig. 3
	Temporal variability	1c	daily-monthly	point	DF	Captures the exceedance likelihood of (low) oxygen concentrations	Fig. 3
	Frequency spectra	1c	minutes-hourly	point	FFT, WT	Data measured from an oxygen sensor logging at high frequency	Kara et al. (2012)
	Time averaged longitudinal plot	1b	weekly-monthly	horizontal transect	V(other)	Seasonal average of individual stations along transect to demonstrate spatial gradient; box-whisker plots of station data at each location can be used instead of average to indicate the range	Xu and Hood (2006)
	Cross section	1a	daily-monthly	vertical slice	V(XZ), MAE(ΔXZ)	Interpolated contour of profile or glider data along a transect compared to simulated cross section	Von Westernhagen et al. (2010) Missaghi and Hondzo (2010) Fig. 3
Oxygen metabolism	Biological Oxygen Demand (BOD) measurement	2a	*ad hoc*	point	R, DF	Data from BOD jar tests can be used to validate simulated rates of oxygen demand	
	Photosynthetic oxygen production	2b	minutes-hourly	point	R, DF	Comparison of simulated rate of oxygen production calculated from oxygen sensor logging at high frequency	Webster et al. (2005)

(*continued on next page*)

Table 3 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
	Oxygen demand (net)	2b	minutes-hourly	point	R, DF	Comparison of simulated rate and oxygen metabolism calculated from oxygen sensor logging at high frequency	
	Time-series comparison	2b	daily-weekly	point	V(TS)	Comparing the daily average GPP and respiration over time can highlight model performance in capturing changes in productivity	
	Apparent Oxygen Utilisation (AOU)	2b	daily-monthly	point; horizontal transect	V(TS) V(other)	AOU is an indicator of the oxygen sag relative to saturation which is a derived indicator of net oxygen consumption	Hetland and DiMarco (2008) Fig. 3
Benthic oxygen exchange	Sediment Oxygen Demand measurement	2a	*ad hoc*	point	V(TS), R, DF, V(other)	Use of data from a Lander or *in situ* benthic chamber experiments, or eddy correlation flux measurements can provide a direct estimate of oxygen flux into the benthos or sediment for comparison with simulated rates	Brady et al. (2013) Chipman et al. (2012) Bryant et al. (2010) Gantzer et al. (2009) Fig. 3
	Volumetrically derived	2b	weekly-monthly	basin	V(other)	Fitting the rate of change of oxygen data in a water volume below thermocline/pycnocline	Snortheim et al. (2017)
Atmospheric oxygen exchange	Air-water oxygen gas flux	2a	*ad hoc*	point	V(TS), R, DF V(other)	Floating chamber can provide estimate of point-scale oxygen flux; relationship between windspeed and atmospheric gas flux can be used to ensure scaling	
Extent of low oxygen	Bottom vs surface difference	1b	daily-monthly	2 layer	V(TS), E, R V(TX)	Time-distance contour plot highlights errors in stratification horizontally, e.g., for assessing seasonal salt-wedge propagation in an estuary	Huang et al. (2018) Fig. 3
	Area/volume of benthic hypoxia/anoxia	3	daily-monthly	bottom layer	V(TS), E, R, V(other)	Area of benthos below critical oxygen threshold (e.g. <2 mg L^{-1}), compared against interpolated profile data	Li et al. (2016) Huang et al. (2018) Fig. 3
Nutrients & organic matter							
Concentrations of SiO_2, PO_4, DOP, POP, TP NO_3, NH_4, DON, PON, TN DIC, DOC, TOC	Time-series comparison	1a	daily-monthly	point; multiple depths	V(TS), E, R	Model assessment against data from grab samples	Ayata et al. (2013) Bruce et al. (2006)
	Vertical transect	1a	daily	vertical slice	V(XZ), MAE (ΔXZ)	Comparison with optical measurements, e.g. from autonomous underwater vehicles	Ménesguen et al. (2019)
	Multi-variate assessment	3	daily-monthly	point; multiple points	TD	Taylor Diagram comparing the relative error fit metrics of multiple interacting variables	Taylor (2001)
Relationship between water colour and concentrations of DIN, DIP	Optical plume class vs DIN, DIP, etc	3	*ad hoc*, monthly	surface layer	DF	Relationship between optical plume class calculated from simulated/measured reflectance at various wavelengths and aggregated water column constituent concentrations in each class	Robson et al. (2017)
Stoichiometric indicators	TN:TP DIN:TP DIN:DIP OC:ON:OP	1b	daily-monthly	point	V(TS), E, R, DF	Ratio of nutrient pools relevant to primary productivity and nutrient management	Li et al. (2013)
	DOC:TOC	1b	daily-monthly	point	V(TS), E, R, DF	Ratio of partitioning of dissolved and total organic carbon	
	POC:PON	1b	daily-monthly	point	V(TS), E, R, DF	Comparison of PON to POC under given temperature, salinity, depth, and inorganic nutrient regimes confirms oceanic nutrient partitioning	Martiny et al. (2013) de Mora et al. (2016)
	Inorganic nutrient: DIC vs. Organic nutrient:carbon	3	*ad hoc*	basin	V(other)	Comparison of the ratio of each modelled nutrient to carbon ratio in organic matter against the dissolved inorganic nutrient to carbon ratios ensures sensible partitioning across nutrient pools	de Mora et al. (2016)
	Sorbed/desorbed PO_4	2a	weekly-monthly	point	V(TS), E, R, DF	Scatter plot comparison showing sensitivity of changes in P associated with mineral/clay particles	Chao et al. (2010)
Carbon and nutrient metabolism	Non-conservative fraction	1b	weekly -seasonal	horizontal transect	V(other)	The deviation from the conservative mixing trend (as indicated by salinity or other tracer) is an indicator of net internal biogeochemical fluxes along a mixing gradient (e.g. estuary)	Adiyanti et al. (2016)

(*continued on next page*)

Table 3 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
	Nitrification/Denitrification rate	2a	*ad hoc*	point; multiple points	R, DF, V(other)	Estimates of nitrification or denitrification (e.g. by isotope pairing method), either *ex situ* or *in situ*, across locations or along an oxygen gradient can validate model process sensitivity	Han et al. (2016)
	Nitrogen fixation rate	2a	*ad hoc*	point	R, DF, V(other)	Data derived from acetate reduction with labelled nitrogen can be used to confirm nitrogen addition by fixation in models capturing this process	Hood et al. (2004) Neumann and Schernewski (2008)
	Rate of OM mineralisation	2a,b	weekly-monthly	point	R, DF, V(other)	Dark incubations measuring oxygen consumption (BOD) or CO_2 production can confirm bulk mineralisation rates	
	Inorganic nutrient uptake rates	2a	seconds-hours	point	R, DF, V(other)	*In situ* determination of uptake by labelled nutrient addition experiments	
Dissolved sediment flux	In situ flux measurement	2a	*ad hoc*	point	V(TS), R, DF	Fluxes may be obtained from benthic chambers or isotope paring technique in the case of N	Chao et al. (2010) Brady et al. (2013) Clark et al. (2017)
	Rate of bottom water accumulation/depletion	2b	weekly-monthly	basin scale	V(other)	Measurements with optical nitrate sensors or grab samples showing change over time used to derive flux rate	
	TCO_2 vs N or P species relationship	3	*ad hoc*	point	R	Nutrient release rate relative to overall sediment metabolism indicates sensitivity to oxygen variability	Zhu et al. (2016)
Particulate sedimentation	Organic matter sedimentation flux	2a	*ad hoc*	water column	R2, MAE	Data collected from sediment traps can be used to validate simulated rates of particulate deposition	
	Organic carbon export as a function of depth/distance offshore	3	annual	basin scale	V(other)	Carbon export as a function of depth in the ocean, fit to exponential curve.	Martin et al. (1987) Butenschön et al. (2012)
Greenhouse gas dynamics	pCO2, pCH4, N2O time-series comparison	1a	minutes-monthly	point	V(TS), E, R	Model assessment against data from grab samples or underway sampling	Huang et al. (2019)
	pCO2, pCH4, N2O transect or profile	1a	*ad hoc*	horizontal transect; vertical profile	V(other), V(TX), V(TZ)	Transect along gradient, or vertical profile	Schmid et al. (2017) Wells et al. (2018) Huang et al. (2019)
	CO_2, CH_4, N_2O surface flux	2a	minutes-monthly	point	V(TS), R, DF, V(other)	Floating chamber can provide estimate of point-scale CO2 flux; relationship between windspeed and atmospheric gas flux can be used to ensure scaling	
	CH4 ebullition	2a	*ad hoc*	water column	V(TS), R, DF	Direction measurement with capture chamber or indirectly through acoustic bubble monitoring	Schmid et al. (2017)
C or N isotopic signature	DOC, DON, NH_4 or NO_3 isotopic fraction	2b	monthly-seasonal	point	V(TS) V(TX)	Correct isotopic signature of simulated nutrient or organic matter variables implies correct flux pathways	Sugimoto et al. (2010) van Engeland et al. (2012) Adiyanti et al. (2016)
Sediment dynamics	Pore-water concentration profiles	1a	*ad hoc*	sediment profile	R, V(TS), V(TZ)	Comparison of pore-water (dissolved) concentration data with depth into the sediment	
	Sediment particulate concentration profiles	1a				Comparison of %C and N with depth into the sediment	
	Oxygen penetration depth	1b	*ad hoc*	sediment column	V(TS), R, DF	Vertical depth into the sediment oxygen is predicted to penetrate	
	Denitrification efficiency	3	*ad hoc*	sediment column	V(TS), R, DF	Ratio of denitrification relative to total nitrogen release from the sediment, compared with data from benthic chamber experiments	
	Carbon burial efficiency	3	*ad hoc*	sediment column	V(TS), R, DF	Comparison of net difference between incoming particulate carbon flux and that released via respiration and methanogenesis at the sediment-water interface	
	Middleburg curve (reactivity-age/depth relationship)	3	*ad hoc*	horizontal transect	V(TS), R, DF	Comparison of models to resolve the drop in organic matter reactivity with increasing depth/distance from the coast	
Redox condition & geochemistry							

(*continued on next page*)

Table 3 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
DIC, TFe, $Fe(OH)_3$, FeII, SO_4, H_2S, TMn, MnII, major ions, metals	Time-series comparison	1a	weekly-monthly	point; multiple depths	V(TS), E, R	Comparison against grab sample data from a fixed monitoring station	Salmon et al. (2017) Shen et al. (2019)
pH, Eh, alkalinity	Time-series comparison	1a	minutes-monthly	point; multiple depths	V(TS), E, R	Comparison against grab sample data from a fixed monitoring station, or *in situ* sensor	
	Depth of redoxcline	1b	weekly-monthly	water column	V(TS), R	In stratified systems, the depth of the redox-cline can be compared due to its significance in influencing deep water conditions.	
	Area of system crossing an acidification threshold	3	*ad hoc*	multiple points	R, V(other)	The emergence of critical areas of low pH in models capturing effects of acidification can be compared with observed locations experiencing low pH	Hipsey et al. (2014) Shen et al. (2019)
Mineral dynamics (e.g. calcite, iron hydroxide, gibbsite)	Mineral saturation state	1b	weekly-monthly	multiple points	V(TS)	Comparison of simulated saturation state against saturation state derived from observed solution properties	Salmon et al. (2017)
	Particulate sedimentation flux	2a	*ad hoc*	water column	R, DF	Data collected from sediment traps can be used to validate simulated rates of particulate deposition	
Dissolved sediment flux	In situ flux measurement	2a	*ad hoc*	point; multiple points	V(TS), R, DF	In situ determination of dissolved species (e.g., FeII, Al, SO_4); can be compared as a function of oxygen or other variable	
	Rate of bottom water accumulation/depletion	2b	weekly-monthly	bottom layer	V(TS), R	Comparison of the rate of change of concentration over time in water that has negligible mixing with the surrounding environment can be used to guide sediment flux rate accuracy	
Sediment quality	Pore-water concentration profiles	1a	*ad hoc*	point; sediment column	R, V(TS), V(TZ)	Comparison of pore-water (dissolved) concentration data with depth into the sediment	Couture et al. (2009)
	Sediment total concentrations	1a	*ad hoc*			Comparison of total concentration data with depth into the sediment	
	Sediment pH	1a	*ad hoc*			Comparison with pH data with depth into the sediment	
Metal bioavailability	Dissolved fraction	1b	weekly-monthly	point	R, DF, V(other)	Comparison to confirm adsorption/desorption, redox, and/or mineral solubility is correctly partitioning metal bioavailability	
	Bioaccumulated fraction	2b	*ad hoc*	point	R, DF, V(other)	Comparison to confirm metal uptake and accumulation into microbes and biota	
Water colour & clarity							
Suspended particulates, turbidity	Time-series comparison	1a	minutes-monthly	point	V(TS), E, R	Data measured from a turbidity sensor, or *ad hoc* suspended solids measurement	Margvelashvili et al. (2008) Margvelashvili et al. (2016)
	Spatial comparison	1a	daily-monthly	surface layer	V(XY), MAE(ΔXY), d_2	Satellite acquired turbidity data from MODIS etc compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match	Miller et al. (2011) Margvelashvili et al. (2013) Margvelashvili et al. (2018)
Particle composition	Chl-*a*:SS or TOC:TSS	3	weekly-monthly	point	R, DF	Scatter plot comparison of fraction of Chl-*a* or TOC fraction of suspended particulate pool	Chao et al. (2010)
Particle size distribution	Frequency distribution of particle size	3	*ad hoc*	point	DF	For models with multiple particle sizes resolved, comparison of particle size distribution (e.g. measured *in situ* by LISST, or using data from discrete samples)	Hipsey et al. (2004)
Sedimentation	Sedimentation flux	2a	*ad hoc*	water column	V(TS), R, DF	Data from sediment traps can validate model sedimentation flux of particulate matter	Ostrovsky and Yacobi (2010)
Extinction coefficient	Time-series comparison	1b	weekly-monthly	point	V(TS), R, DF	May need to specify wavelengths	Fujii et al. (2007) Chao et al. (2007) Chao et al. (2010)
Secchi/euphotic depth	Time-series comparison	1b	weekly-monthly	point	V(TS), R, DF	Model assessment of light penetration against routine transparency measures	
		3	Interannual	system	V(other), R		Robson et al. (2017)

(*continued on next page*)

Table 3 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
	Relationship wet season river inputs and subsequent (eg dry season) average Secchi depth					Assessment of seasonal lag between catchment inputs and water response	
PAR: Photosynthetically Active Radiation intensity	Depth comparison	1a	hourly-monthly	vertical profile	V(TZ), R, MAE (ΔTZ)	Model assessment against data from vertical light profiles	Fujii et al. (2007)
Light quality	Reflectance or irradiance in a specified wavelength range	1b	daily-weekly	surface layer	V(XY), MAE (ΔXY)	Comparison of simulated ocean colour with satellite-observed ocean colour for the optical surface layer	Baird et al. (2016a) Jones et al. (2016)
Water colour	Spatial comparison	3					
CDOM: Chromophoric Dissolved Organic Matter (a.k.a. gilven, gelbstoff, colour)	Time-series comparison	1a	weekly-monthly	point	V(TS), E, R	Model assessment against data from grab samples	
	Spatial comparison	1a	daily-monthly	surface layer	V(XY), MAE (ΔXY)	Satellite acquired CDOM data from MODIS etc compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match	
Chlorophyll-*a*							
Chlorophyll-*a* concentration	Time-series comparison	1a	hourly-monthly	point	V(TS), E, R	Model assessment against data from samples or Chl-*a* sensors	Elliott et al. (2000) Bruce et al. (2006) Gal et al. (2009) Grangeré et al. (2009b) Ng et al. (2011)
	Bloom peak biomass	1b	weekly-monthly	point	R, DF, V(other)	Scatter plot of simulated vs observed bloom peak biomass	Guillaud et al. (2000)
	Bloom peak time and recovery timescale	1b	weekly-monthly	point	R, DF, ACF	Scatter plot of simulated vs observed time of peak biomass; Visual comparison of bloom onset and end dates in model and observational time-series	Ménesguen et al. (2007)
	Biomass (interannual) variability	1b	weekly-monthly	point	R, SR	Rank correlation of annual bloom magnitudes	Hearn and Robson (2000)
	Frequency spectra	1c	minutes-hourly	point	FFT, WT, WC	Data measured from a chlorophyll-*a* sensor logging at high frequency	Kara et al. (2012)
	Spatial comparison	1a	daily-monthly	surface layer	V(XY), MAE(ΔXY)	Satellite acquired Chl-*a* estimates compared pixel for pixel with the simulation. Model or data may require averaging to ensure spatial resolutions match	Ménesguen et al. (2007) Doney et al. (2009) Sinha et al. (2010) Jiang and Xia (2018)
	Spatial variability	1c	daily-monthly	surface layer	CDF, V(TX)	Compares distribution and range of Chl-*a* within the domain without conducting pixel by pixel comparison	Quere et al. (2005) Doney et al. (2009)
	Spatial patchiness	3	daily-monthly	surface layer	ACF, V(other)	Assesses similarity in spatial coherence of Chl- *a*; Characteristic patch length scale comparison, including skewness, kurtosis.	Hillmer et al. (2008) Ng et al. (2011)
	Deep Chlorophyll Maximum (DCM)	3		water column; vertical trasect	V(TZ), V(XZ)	Vertical profile data collected via fluorometry at a single location or along a transect can be compared with simulated Chl-*a* contours	Varela et al. (1992) Jones et al. (2016)
	Modes of variance	3	daily-monthly	surface layer	EOF	Comparison of EOF decomposition results for remote sensing image and model spatial output	Rocha et al. (2019)
Primary productivity	Photosynthesis rate	2a	*ad hoc*	point	V(TS), R, MAE, DF, V(other)	*In situ* productivity measures (e.g. PhytoPAM; ^{14}C uptake) compared with model. Could be reported as a function of environmental changes (e.g light)	Grangeré et al. (2009a) Ayata et al. (2013) Saba et al. (2010) Brush and Nixon (2017)
	Photosynthesis rate	2b	minutes-hourly	point	V(TS), R, MAE, DF, V(other)	Oxygen metabolism calculated from oxygen sensor logging at high frequency (community photosynthesis); labelled carbon experiments (combined with PAM); *in situ* incubations (pelagic vs. benthic photosynthesis)	Brush and Nixon (2017)
		2a	weekly-monthly	point	R, DF, V(other)	Photosynthesis rate relative to biomass	Fulton et al. (2004)

(*continued on next page*)

Table 3 (*continued*)

Property being assessed	Description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g. processing requirements)	Example references[b]
	PPB: Phytoplankton production per biomass						
	Eppley curve: Scaling relationship of net production as a function temperature	3	weekly-monthly	point	V(other)	Scatter plot comparison of *in situ* measured productivity with temperature, compared against model equivalent	Brush et al. (2002) Mark et al. (2002)
	P-I relationship: *in situ* bulk productivity compared as a function of irradiance	3	hourly-daily	point	V(other)	Scatter plot comparison of *in situ* measured productivity with irradiance, compared against model equivalent	Grangeré et al. (2009a)
Monbet or Vollenweider relationship	Scaling relationship between nutrient concentration and chlorophyll-*a*	3	monthly-annual	system	V(other)	Scatter plot comparison showing slope of relationship between ambient nutrient levels and Chl-*a* is consistent with data	Fulton et al. (2004) Jones and Lee (1988)
Other pollutants							
Microbial contaminant organism concentrations	Time-series comparison	1a	daily-weekly	point	V(TS), E, R	Model assessment against data from grab samples on coliforms, viruses and other pathogens	Hipsey et al. (2008)
	Viable fraction	2b	*ad hoc*	point	V(other), R, DF	Measure viable but not culturable (VBNC) vs total counts	Hipsey et al. (2004)
	Sedimentation rate of attached fraction	2b	*ad hoc*	point	V(other), R	Indirect estimation of organism sedimentation rate from profile measurements of particle size distribution	Hipsey et al. (2006)
Organic chemical contaminant (e.g. PAH, POP) concentrations	Time-series comparison	1a	weekly-monthly	point	V(TS), E, R	Comparison of temporal variability of concentrations at a monitoring site. For example, assessment against data from sediment cores, backdated via ^{210}Pb and ^{137}Cs radionuclide	Kong et al. (2007)
	Sorbed/desorbed fraction	2a	*ad hoc*	multiple points	DF	Demonstration of pollutant partitioning between dissolved and particulate phase	
	Bioaccumulation	2b	*ad hoc*	multiple points	V(other)	Concentration in microbial biomass relative to the water concentration	
Number of particles of plastics	Time-series comparison	1a	weekly-monthly	point	V(TS), E, R	Comparison of temporal variability of plastics particle density at a monitoring site	
	Spatial variability	1a	*ad hoc*	multiple points	V(other), DF V(XY)	Assessment of spatial variability in plastic particle density, showing accumulation areas	

[a] E = user determined combination of traditional error metrics, including BIAS, MAE, NMAE, MEF, NSE, B and/or d_2.
[b] Where possible references provided indicate an example of the metric being applied, otherwise references relevant to give context to use of the metric are listed.

used to assess model equivalents. A similar approach for estimating basin-scale average sediment oxygen demand has been shown to be useful in stratified lakes, allowing quantification of the rates of hypolimnetic oxygen drawdown where other consumption mechanisms may be assumed to be relatively minor (Snortheim et al., 2017). Following these examples, Fig. 3 illustrates the utility of combining relevant metrics across the assessment levels for an estuary experiencing frequent hypoxia.

Many aquatic modelling studies have had their roots in predicting the impacts of eutrophication and have demonstrated the Level 1 performance of their models against data on dissolved and total nutrients, with a focus on P in freshwaters and N in marine waters. An increasing trend towards simulating the complete N, P, Si and C cycles has provided opportunity to validate models using other Level 1 indicators that capture more nuanced aspects of nutrient cycling, such as partitioning between organic and inorganic phases and stoichiometric variability (Li et al., 2013). Other potential Level 1 metrics relate to organic matter composition, such as POC:DOC, OC:ON, or potentially the labile:refractory ratio of the simulated organic pool. To date this has rarely been the subject of model assessment, but it may be possible by comparing with increasingly reported data from Excitation-Emission Mass Spectroscopy (EEMS) studies.

Relevant Level 2 process validation efforts can be applied in the form of comparison against *in situ* estimates of nitrification/denitrification, organic matter mineralisation, and community respiration or BOD data. In estuarine environments, dilution curves have been applied to estimate sources and sinks of materials by comparing concentrations relative to salinity. For example, a sink of NO_3 along the length of a system can be used as an indirect measure of denitrification intensity (e.g. Eyre and Balls, 1999). Fig. 4 demonstrates this approach for an estuarine carbon cycle investigation, showing validation of the along-stream predictions of DOC, DIC and ^{13}C-DIC and ^{13}C-DOC, relative to a conservative tracer. In this case, the use of isotopes in the calibration helped to reduce equifinality, since models able to correctly capture patterns in stable isotope cycling are more likely to be resolving flux pathways that are imprinting distinct signatures during isotope fractionation processes (Sugimoto et al., 2010; van Engeland et al., 2012; Adiyanti et al., 2016). Where sediment-water interaction is an important driver of nutrient cycling, the flux of dissolved constituents as estimated from *in situ* benthic chambers or from eddy correlation can be used as a powerful Level 2 approach to reduce uncertainty. Two examples comparing modelled against measured NH_4 release are applications in Chesapeake Bay (Brady et al., 2013) and Tokyo Bay (Sohma et al., 2008).

The net rates of carbon and/or organic matter sedimentation are important drivers of water and sediment condition, and are important fluxes to test models against, since they can vary substantially through time and between sites. In oceanic systems, the rate of organic carbon export shows a logarithmic decline with depth, and can be plotted as a Martin curve (Martin et al., 1987), which was tested as a validation metric of the European Regional Seas Ecosystem Model (ERSEM) (Butenschön et al., 2012). Studies employing depth-resolved sediment models themselves require a significant validation effort (e.g. Paraska et al., 2014), depending on availability of pore-water and solid phase concentrations, and can benefit from Level 2 validation metrics, such as denitrification efficiency, oxygen penetration depth and oxygen exposure time, which are known to be important determinants of carbon cycling and burial. Capturing vertical gradients in oxygen and other constituents in sediment, for example using *in situ* microprofile data, may mean models implicitly capture these variables, but further assessment against *in situ* flux rate determinations (e.g., denitrification) can help test sediment model function. The significance of bioturbation was indirectly validated as an important process by Zhu et al. (2016), by ensuring the DO vs. PO_4 flux rate matched *in situ* observations. Additional Level 2 metrics for assessing sediment biogeochemical predictions may include summaries of the O_2:CO_2 sediment-water flux ratio, reflecting the models ability to correctly capture the balance of aerobic and anaerobic respiration that is occurring.

Particularly for spatially resolved models, ensuring models capture empirically established scaling relationships between oxygen exposure time and carbon burial efficiency or organic matter reactivity and age (as depicted by the Middleburg curve) may prove to be a particularly useful test. However, to date this has yet to be reported (Paraska et al., 2014).

Simulating other aspects of aquatic geochemistry is increasingly being undertaken to capture acidity and risks associated with heavy metals. Examples include model applications in an acidic environment such as mining impacted landscapes (Salmon et al., 2017) or coastal sites impacted by acid sulfate soils (Hipsey et al., 2014). Other applications include reservoir management whereby seasonal anoxia and the accumulation of metals in the hypolimnion requires models to capture the redox sensitivity of Mn and Fe.

Despite its importance in driving productivity and shaping biogeochemistry, the light climate has not always featured during model validation. Light profile data and light quality (specific bandwidth attenuation), including the relative shift in extinction coefficient in response to suspended sediment concentrations, can be a useful exercise. When included with Chl-a model predictions, it is useful to show that predictions of Chl-a on average scale correctly with the suspended solids concentration (Chao et al., 2007, 2010), which may be considered a system-level property. The light climate is also influenced by dissolved substances, though even where dissolved organic matter (DOM) is simulated, few models separate coloured dissolved organic matter (CDOM) from other DOM or specifically validate the contribution of CDOM contribution to light attenuation. Validation of wavelength-specific light attenuation profiles concurrently with CDOM and suspended sediment concentrations offers the potential for us to better resolve the light climate. This is likely to become more important as models aim to resolve ultra-violet (UV), photosynthetically active (PAR) and near infra-red (NIR), due to their different effects on water thermal structure and also organism growth and mortality. Recently, the validation of a coastal ocean model against true colour from satellite imagery also demonstrated the utility of capturing the specific contributors of light reflectance at multiple wavelengths to capture the overall light climate (Baird et al., 2016a).

Total chlorophyll *a* (Chl-a) is the most common biological variable simulated in aquatic models ranging from small ponds to the global ocean. Elliott et al. (2000) identified the relative merits of a range of error calculation methods for algae time-series. Many of these metrics suffered when the magnitude of the data values is large and they are unforgiving of temporal misalignments between modelled and observed data. For example, if a simulated bloom is of the correct magnitude but one week earlier/later than the observed bloom, is the model still fit for purpose? Further Level 1 comparisons, such as bloom magnitude (e.g. for the spring and/or summer) may be compared separately from the bloom's timing allowing discrepancy in the latter to be isolated. Thus, if capturing bloom size over several years was of more value than predicting its exact timing, a typical error metric for assessing bloom size alone would be of greater use to the study. When assessing modelled Chl-a time-series data, traditional Level 1 assessment metrics can be supplemented by derived metrics such as bloom peak magnitude, duration and time of onset. Another example is the application of wavelet analysis of a high-frequency Chl-a time-series (Kara et al., 2012) to test model performance at predicting scales of variability from days to seasons (Fig. 5). This approach can give an improved view over simple time-series comparisons about the scales of variability that a model can capably reproduce. Spatial comparisons of Chl-a from remote sensing are becoming more common, particularly in coastal and marine models, and wavelet-based comparison may be an effective way to evaluate model Chl-a against satellite data (Saux-Picart et al., 2012).

At Level 2, rates of algal productivity – i.e., carbon fixation – determined directly from *in situ* experiments using isotopically labelled carbon, or estimates from instruments measuring photosynthetic

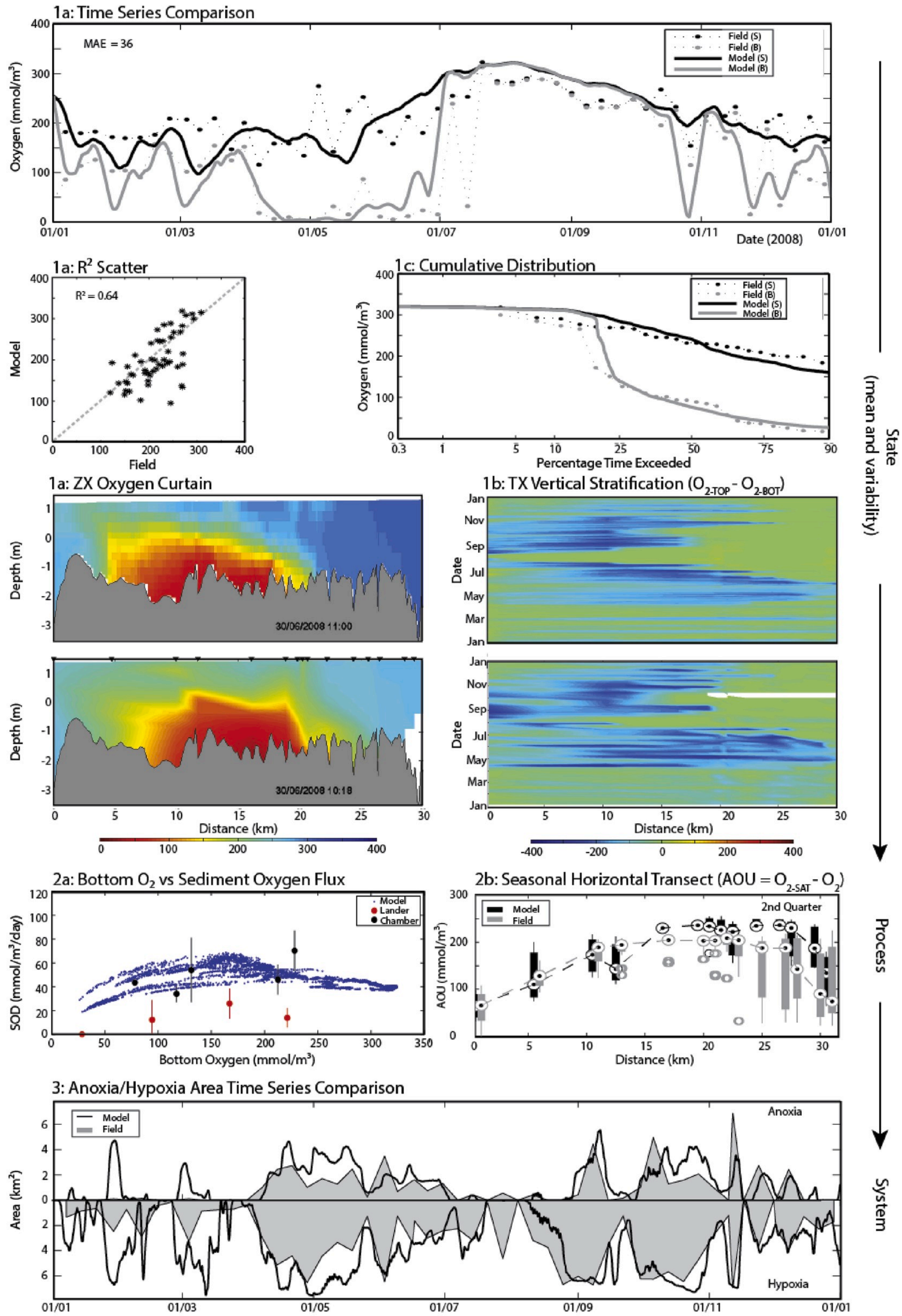

Fig. 3. An example of multiple assessments of a 3D estuary model of the Swan River Estuary demonstrating performance of oxygen and hypoxia/anoxia prediction; the assessment level is indicated in the top-left corner of each panel. In this example, the level of predictability based on time-series analysis alone is modest (MAE = 36 mmol m^{-3}, $R^2 = 0.64$), but assessment of the temporal and spatial variability in predictions, comparison with available sediment flux rates, and assessment of system-scale anoxia extent together allow a more complete picture of model suitability, leading to a more positive conclusion about the overall model behaviour and its suitability for scenario assessment. Refer to Huang et al. (2018) for model and data details.

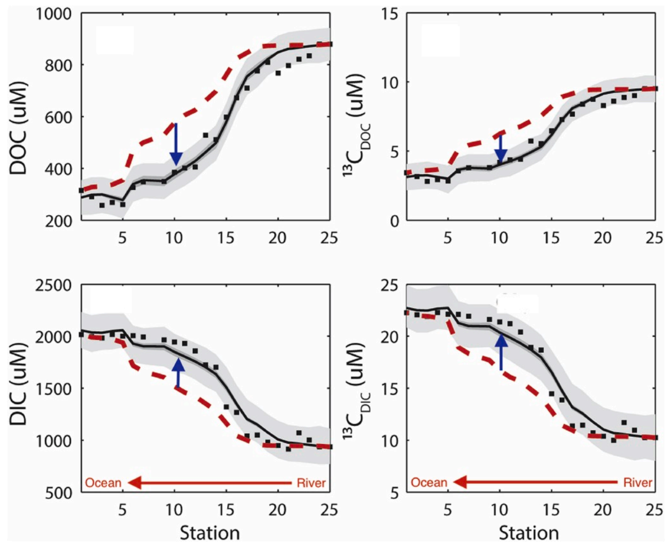

Fig. 4. An assessment of a carbon cycling model of the Caboolture River estuary demonstrating performance of simulations capturing DOC and DIC concentration changes along a salinity gradient from the river to the ocean mouth. The redline indicates concentrations of conservative tracers, subject to hydrodynamic mixing only, whilst the black line indicates the most likely model predictions based on a MCMC calibration algorithm. The blue arrow depicts the degree of concentration change associated with organic carbon mineralisation (Level 2b), and correct validation of the appropriate parameters in the model was constrained by including variations in the isotopic fractions of both DOC and DIC. Plots adapted from Adiyanti et al. (2016) with permission from Elsevier. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

activity based on fluorescence, can provide an indication of gross primary production (GPP) that can be compared with modelled rates of photosynthesis; though we found this was surprisingly absent from the reviewed literature. For Level 3, deep Chl-a maxima (DCMs) are an emergent feature of complex model dynamics that manifest in response to physical, chemical and biological interactions within stratified systems and can focus model validation (e.g., Carraro et al., 2012; Ayata et al., 2013). Where model application spans a wide range of trophic conditions, general scaling relationships relating system average Chl-a concentration to the degree of external loading may be useful. For example, relatively simple assessments of models of coastal domains can compare performance against the Monbet relationship, or in the case of freshwater lakes, against the Vollenweider model (Vollenweider and Kerekes, 1982; Reckhow and Chapra, 1983). These comparisons may not be relevant for short-term simulations (e.g., of an individual phytoplankton bloom), but may be useful to ensure that complex water quality models are able to meet expectations of cross-system empirical data (see also Chang et al., 2019).

The simulation of biological and chemical contaminants has also been identified as an important area of aquatic ecosystem simulation for inland and coastal waters, particularly for the purposes of informing public health risk assessments. Models of microbial pollutants have been coupled with hydrodynamic models to simulate pathogens and faecal indicator organisms and assess risk in waters used for drinking and recreation (e.g. Hipsey et al., 2008; Sokolova et al., 2013). These studies have tended to validate models by assessing organism counts from available monitoring data (Hipsey et al., 2004; Sokolova et al., 2013). However, this approach is complicated since viable but non-culturable cells (VBNC) make determining viable and inactivated fractions of organism populations uncertain. The increasing adoption of molecular approaches for organism enumeration, in conjunction with adaptive agent-based models has been applied to resolve strains of different pathogenicity (Bucci et al., 2012). Prioritisation of validation at key

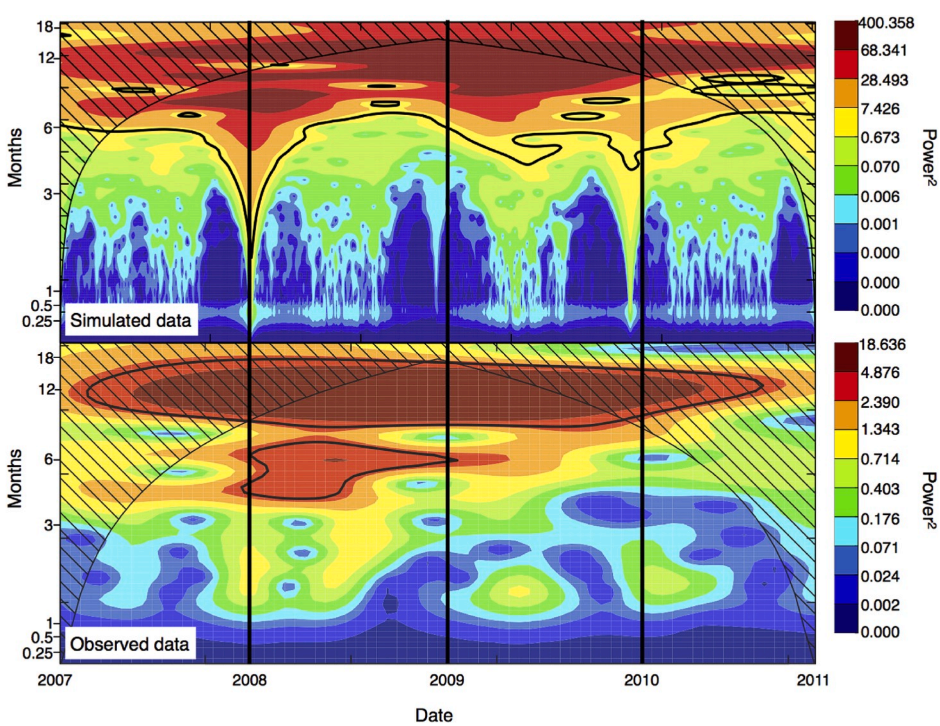

Fig. 5. Continuous wavelet power spectra showing the periodicity of the simulated hourly chlorophyll-a data (top panel) and observed weekly data (bottom panel) from Lake Mendota, Wisconsin, USA during 2007–2011. The continuous wavelet spectrum shows how the strength of the periodicities of the data changed over time, with the colours highlighting the intensity, or power of a particular frequency at a point in time (dark red = high power; dark blue = low power; colour scale is power squared). The diagonal lines show the cone of influence, where edge effects at the beginning and end of the time series may compromise the interpretation of the power spectrum. The thick black contour line shows the 5% significance level of the power spectrum in comparison to a null red noise spectrum (wavelet methods described by Carey et al., 2016). The simulated data were output from a General Lake Model-Aquatic Eco-Dynamics (GLM-AED) model calibrated for Lake Mendota (see Snortheim et al., 2017). Given the differences in the temporal resolution of the simulated and observed data, the wavelet transform provides a useful approach for comparing the two datasets' dominant temporal scales of variability (Level 1c). As evident from the dark red colour at the 12-month frequency, the annual scale emerges as the most important scale of variability throughout the time series for both the simulated and observed data. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

exposure points (e.g., drinking water offtakes or beaches) using exceedance probability plots can also be used to demonstrate a model is suited to risk assessments. Simulating organic chemical contamination is still scarce, likely because of the limited data available for model validation and process identification. A recent study used data derived from sediment cores to validate a 60-year model simulation of polycyclic aromatic hydrocarbons (PAHs) in a large shallow lake (Kong et al., 2017), though further work in this area is warranted.

4.3. Community dynamics and ecosystem function

Aquatic ecologists have a long history of developing metrics to describe the structure and function of aquatic populations and communities. Models that aim to resolve these more complex ecological dynamics build on those listed in the previous sub-section, but increasingly are including a wider range of theoretically relevant indicators of ecosystem interactions and function. For these studies we broadly categorised the focus of various AEM applications, and identified 6 broad areas of a) microbial lower trophic level communities, b) benthic populations and habitats, c) pelagic populations and habitats, d) community structure, e) community function, and e) ecosystem response to disturbance (Table 4). Within each category we initially considered Level 1 metrics describing temporal variability of species and their interactions, and the development of spatial niches. We then focused more specifically on literature where models have considered measures of trophic structure and complexity, inter-relationships between simulated ecological groups, and system level responses to ecosystem perturbation. Within this context, the diversity of metrics did appear to be less consistent and differ more notably across the freshwater and marine community.

In modelling the phytoplankton, zooplankton, and bacterial dynamics, the challenges of capturing changes at the level of species or even genus are considerable (e.g. see Lignell et al., 2013; Li et al., 2014; Andersen et al., 2015). Of the published AEMs that simulate Chl-a, only a few aim to simulate succession at the species or genus level (e.g. Elliott et al., 2006; Mieleitner and Reichert, 2008; Gal et al., 2009), with most designed to simulate at the level of taxonomical (e.g. Elliott et al., 2005; Trolle et al., 2011; Chan et al., 2002) or functional groups (PFT's; e.g. Elliott et al., 2000; Segura et al., 2012; Quere et al., 2005; Baird et al., 2016a). For these applications, specific use of even Level 1 validation metrics relevant to species/genus/group partitioning and their spatio-temporal dynamics has been limited, with a reliance on validation against Chl-a observations (discussed above). A reason for this discrepancy is thought to be simply a lack of corresponding observational data (where phytoplankton taxonomic data are available at all, they are usually at a much lower frequency and/or spatial resolution than total chlorophyll estimates) and/or difficulty in converting species counts into the unit used by the model (e.g. biovolume, intracellular carbon, nitrogen or Chl-a concentrations). If the observed species data can be converted into comparable units, the metrics used for total chlorophyll validation can also suffice at the phytoplankton community level (e.g. Elliott et al., 2000).

Planktonic community structure, which may be a Level 1 or Level 3 metric depending on the model structure, can also be examined in terms of relative contributions of various species to the total community biomass. Gal et al. (2009), for example, validated the relative contribution of cyanobacterial species to the total phytoplankton biomass as a function of nutrient loading into the lake. This comparison provided a multi-tier assessment of the model as a good fit that indicated not only successful simulation of phytoplankton biomass and the relative contribution (and hence succession) of the various species, but also an accurate reproduction of the interactions between forcing conditions (nutrient loading) and the food web. The latter metric ensures not only the accurate simulation of phytoplankton in the model but also the assembly of dynamics and processes occurring in the ecosystem.

At larger oceanic scales, Holt et al. (2014) similarly used the emergence of correct biomass partitioning between functional groups to support validation by plotting diatoms as a fraction of total Chl-a, and comparing with the observed scaling between the diatom fraction and total Chl-a as estimated from satellite observations. The model data were also shown as a two-dimensional density histogram, as a running average and least squares regression fit to the continuum function used in Brewin et al. (2012). At the global scale, de Mora et al. (2016) used the known scaling relationships between abundance of phytoplankton functional types and total chlorophyll concentrations to demonstrate the community structure was correctly reproduced (Fig. 6). They further supported model validation with tests of N, P, Si and Fe stoichiometric ratios, and carbon:Chl-a partitioning to demonstrate the correct emergence of expected patterns from the simulated nutrient flux pathways.

Further assessment can be undertaken to characterise model performance capturing ecological species succession (Level 3). True ecological succession, whereby the presence of one functional group creates the conditions required for the emergence of the next group (e.g. Hearn and Robson, 2000), should not be confused in this context with the pattern of replacement of one functional group by another due to unrelated changes in environmental conditions, sometimes also referred to as succession (e.g. Chan et al., 2002). The former can be considered a genuine emergent property of the system, while the latter allows only a more detailed state validation of different simulated size or functional groups. Graphical examination of simulated phytoplankton succession based on visual comparison of a time series has been used in a number of studies (see Rigosi et al., 2010, for a partial list). The use of quantitative metrics, however, has been limited and when applied, it typically focuses on goodness of fit and correlation metrics. Successful modelling of phytoplankton succession requires accurate simulation of numerous processes and food-web interactions, which even a tool like a Taylor Diagram may not be able to confidently describe.

The use of tracers to evaluate food-web dynamics is another efficient Level 3 option for assessing a model's ability to capture the food-web dynamics, and in particular the trophic interactions. The use of stable isotopes has become increasingly popular in empirical food-web studies, but also more recently as a means for assessing models. The signature of stable isotopes in organisms integrates over time and provides validation of the mapping of the predator-prey interactions resulting from prey preference and availability in the system. Thus, the successful match between observed stable isotopes and model trophic levels improves confidence in the strength of the simulated trophic pathways occurring within the food web (e.g., Nilsen et al., 2008; Dame and Christian, 2008). Adopting a similar approach, Carrer et al. (2000) used bioaccumulation of dioxins in the food web to evaluate model performance of trophic linkages.

A further high-level series of indicators is based on food quality, both in terms of stoichiometry and fatty acid concentrations. The use of dynamic intracellular stoichiometric ratios in a model allows the user to evaluate model performance not only in the form of changes in the stoichiometric ratio of a certain species over time but also the change in the ratios of organisms of higher trophic levels affected by the lower level organisms. Gaedke et al. (2002), for example, examined C:P ratios in egested vs ingested food in the model as part of mass-balance models of the food-web. Li et al. (2013) compared a probability density function of five phytoplankton group internal N:P ratios with sporadic observations of minimum, maximum and mean internal nutrient concentrations. Thingstad et al. (2007) used model validation and the mismatch between simulated and observed bacterial production in a mesocosm to identify the weakness in model simulation of varying stoichiometric ratios. Mitra and Flynn (2006) incorporated stoichiometry into a multi-species predator-prey model with varying elemental composition and selectivity. Their model successfully simulated the switch in predator diet from predation on the prey to cannibalism when the prey suffered from nutrient limitation and decreased in food quality.

Mulder and Bowden (2007) examined whether zooplankton can alter their internal stoichiometry under nutrient poor conditions and whether

Table 4
Summary of validation metrics relevant to models simulating ecological function of aquatic systems (refer to Table 1 for assessment technique abbreviations).

Property being assessed	Approach and description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g., data collection and processing requirements)	Example references [b]
Primary producers and lower trophic levels (bacteria, flagellates, ciliates, phytoplankton)							
Species or functional group biomass (including bacteria, micro zooplankton and phytoplankton)	Time-series comparison	1a	weekly-monthly	point	V(TS), E, R	Validation assessment of cell or organism counts, converted to biomass, or inferred from pigment analyses or optical proxies	Elliott et al. (2006) Mieleitner and Reichert (2008) Gal et al. (2009) Rigosi et al. (2011) Chung et al. (2014)
	Bloom magnitude	1b	weekly-monthly	point	R	Scatter plot of simulated vs observed bloom peak biomass	
	Bloom duration and timing	1b	weekly-monthly	point	R, V(TS)	Scatter plot of simulated vs observed time of peak biomass; Visual comparison of bloom onset dates in model and observational time-series	Peeters et al. (2007)
	Biomass (interannual) variability	1c	weekly-monthly	point	SR	Rank correlation of annual bloom magnitudes in multi-annual simulation	Gal et al. (2009)
Species/group relative biomass fraction and succession	Species succession and relative amount of a particular species or PFT relative to the total biomass over time	1b	weekly-monthly	point	V(other), DF	Time-series of species fraction (e.g. fraction of cyanobacteria relative to total algal biomass), and/or visual comparison of relative distribution plot as a function of time.	Gal et al. (2009) Robson et al. (2008)
Species/group characteristics	Cell size distribution, colony presence and size	1c	*ad hoc*	point	V(other), DF	For models resolving size variation in the planktonic community, allocation of biomass across the size spectrum should occur consistent with the Sheldon spectrum	
	Group specific attenuation coefficient	1b	*ad hoc*	multiple points	V(other)	Comparison of model separation of group components contributing to the light extinction coefficient, K_d	Buzzelli et al. (2014)
	Cell density, buoyancy	2b	*ad hoc*	point	V(other)	Changes in density could be directly measured or calculated from observed rising/sinking rates measured using, for example, laser spectroscopy	Kromkamp and Walsby (1990)
	Vertical sinking rates	2a	*ad hoc*	water column	V(other), R	Sediment traps can be used to estimate vertical mass export form the surface mixed layer or water column	
	Phytoplankton nutrient uptake rates, incl. nitrogen fixation	2a	*ad hoc*	point	V(other), R	Comparison of reported ranges from nutrient uptake experiment data with frequency distribution of model uptake rate	Dugdale and Wilkerson (1986) Popendorf and Duhamel (2015)
	Zooplankton prey grazing rates	2a	*ad hoc*	point	V(other)	Time-series comparison of organism food ingestion rates, compared against available data on observed ingestion rates	Morozov (2010)
	Organism nutrition source partitioning	2b	*ad hoc*	point	V(other)	Model outputs of nutrient or grazing food sources, integrated over time, can be cross-checked against known preferences (see also foodweb linkages)	
	Gradient in microbial gene expression	2a	*ad hoc*	horizontal transect or vertical profile	V(other) V(TZ) V(XY)	Targeted measurements indicating the expression of specific microbial genes (e.g., denitrification or nitrogen fixation) over distance, can validate the changing presence of particular process pathways	Coles et al. (2017) Arora-Williams et al. (2018)
	Internal stoichiometry (e.g., IN:IP or Chl *a*:C)	3	*ad hoc*	point	V(other), DF	Comparison of reported ranges from observed data with frequency distribution of model ratio	Li et al. (2013) Grangeré et al. (2009a)
	Canberra distance	3	*ad hoc*	point	V(other)	Comparison of the distance between cell-volume distributions	Sauterey et al. (2015)
Spatial niche formation	Horizontal patterns in species/group distribution	1a	daily-monthly	surface layer	V(XY), R, MAE(ΔXY)	Satellite or UAV acquired fluorescence data compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match	Ménesguen et al. (2007) Alexander and Imberger (2008) Doney et al. (2009) Sinha et al. (2010)
	Biomass of each species/functional group as a function of water surface temperature or salinity	1c	monthly-seasonal	surface layer	CCF	The example given is for a cross-model comparison rather than a model validation against data	Sailley et al. (2013)

(*continued on next page*)

Table 4 (*continued*)

Property being assessed	Approach and description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g., data collection and processing requirements)	Example references [b]
	Biomass – depth relationship	3	hourly	horizontal transect	V(TX)	Biomass as a function of depth along a transect or depth limit of biomass observations	
	Formation of surface or deep biomass maximum	3	hourly-monthly	vertical profile; vertical slice	R, V(TZ), MAE (ΔTZ)	Vertical profile data collected via multi-spectral fluorometry or discrete samples via Niskin bottle collections; Hourly data required for validation of diel vertical migration behaviour	Chien et al. (2013) Fujii et al. (2007)
	Relative abundance of each species or PFT against the total community Chl-*a*	3	weekly-monthly	surface layer	V(other), V(XY)	Relative abundance of each PFT against the total community Chl-*a*, presented spatially	Hirata et al. (2011) Brewin et al. (2012) Devred et al. (2011)
Primary production and respiration	*PP: Refer to* Table 3 *Chl-a metrics* *CR: Refer to* Table 4 *Community Function metrics*						
Benthic populations and habitats (macrophytes, macroalgae, coral, invertebrates)							
Benthic population biomass/ density	Time series comparison	1a	monthly-annual	point	V(TS), E, R	*In situ* sampling organism counts per area or areal biomass from quadrats over time	Best et al. (2001) Brigolin et al. (2009) Grangeré et al. (2009b) Grangeré et al. (2010) Robson et al. (2010)
	Spatial distribution and extent	1a	monthly-annual	horizontal transect; basin scale	V(XY), R, MAE (ΔXY)	*In situ* surveyed benthic coverage mapping or remotely sensed data can be used. Model or data may require averaging to ensure spatial resolutions match.	Baird et al. (2016b)
	Total area of coverage	3	monthly-annual	system	V(TS), MAE	Integrated area of benthic organism coverage from surveys or remotely sensed data, showing expansion and contraction of extent over time	
Benthic species/group characteristics	Tissue stoichiometry	3	monthly	point	V(TS), R, DF	Internal stoichiometry (e.g. of N or P) of plant or invertebrate tissue can support validation of nutrient uptake and assimilation	Higgins et al. (2005)
	Ratio of macrophyte's above-ground and below-ground biomass	1b	weekly-monthly	multiple points	V(other), R, DF	Partitioning of biomass at the plant, meadow or population scale	Verhagen and Nienhuis (1983)
	Macrophyte canopy attributes (height, deflection extent)	1b	*ad hoc*	multiple points	V(other), R	Models with flexible vegetation can be validated against measured canopy properties, incl meadow height, deflection height, etc.	Nakayama et al. (2019) Dijkstra and Uittenbogaard (2010)
	Population size distribution	1c	*ad hoc*	point, basin	DF	Comparison of size distribution of organisms within a population against observed ranges/counts	Spillman et al. (2008)
Benthic primary productivity and respiration	Photosynthesis rate	2a	*ad hoc*	multiple points	V(other), R, DF	*In situ* productivity measures (e.g. benthic chamber) can be used to validate model rates	Webster et al. (2005) Testa et al. (2017)
	Benthic metabolism	2b	*ad hoc*	multiple points	V(other), R, DF	Difference between whole-system productivity (e.g. diurnal O_2 and DIC variations) and estimated pelagic productivity (pelagic incubations)	Webster et al. (2005)
Benthic filtration	Water filtration rate	2a	*ad hoc*	multiple points	V(other), R, DF	*In situ* filtration estimates of mussels, oysters etc (e.g. from a benthic chamber)	
Reproduction and recruitment	Timing of reproduction events (e.g. spawning, flowering and seed-set, seed germination, sprouting, eggs)	2a	*ad hoc*	point; basin	V(TS), R, SR	Checking model timing of critical life-history events such as vegetation flowering or coral spawning can ensure correct model formulation and response to environmental cues	
	Seed, larvae dispersal extent	2a	*ad hoc*	multiple points; bottom layer	V(XY), R	Comparison of measured seed or larvae densities with modelled extent	
Spatial niche formation	Biomass – depth relationship	2b	*ad hoc*	horizontal transects	V(TX)	Biomass structuring along a depth gradient	
	Horizontal patterns in species/ group distribution within the community	3	*ad hoc*	bottom layer	V(XY), DF	Survey data used to demonstrate emergence of different suitable areas for competing functional species/groups	Savina and Ménesguen (2008)
	Spatial variability in habitat quality	3	*ad hoc*	basin scale	V(other), R	Habitat index (e.g. HSI) correlation with observed survey data	Collier et al. (2017)

(*continued on next page*)

Table 4 (*continued*)

Property being assessed	Approach and description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g., data collection and processing requirements)	Example references [b]
	Biomass – environment relationships	3	*ad hoc*	system	V(other)	Multi-variate statistical clustering of organism presence/ biomass, as a function of environmental conditions	
Population structuring (population cohort)	Age-length relationship	3	*ad hoc*	basin, system	V(other)	Accurate depiction of the length-age relationship of a fish population demonstrates the model is allocating resources to growth correctly	Le Goff et al. (2017)
	Individuals vs age class	3	*ad hoc*	basin, system	V(other)	Partitioning of biomass is correct age class categories emerges in response to fishing pressure, changing metabolic rates and food availability, incl competition	Savina and Ménesguen (2008)
Higher trophic level populations and habitats (zooplankton, invertebrates, fish, sea mammals and birds)							
Pelagic population concentration/number	Time series comparison	1a	weekly-monthly	point	V(TS), E, R	Comparison with organism counts or biomass per volume over time	
	Magnitude of biomass, or population, peak	1b	weekly-monthly	point	V(TS), R, SR	Scatter plot of simulated vs observed peak (seasonal) biomass	
	Timing of peak biomass or population	1b	weekly-monthly	point	V(TS), R	Scatter plot of simulated vs observed time of peak biomass; Visual comparison of bloom onset dates in model and observational time-series	
	Interannual variability in peak magnitude	1c	weekly-monthly	point	SR	Rank correlation of annual peak magnitudes in multi-annual simulation	Gal et al. (2009)
Pelagic species/group relative biomass fraction	Succession pattern and relative amount of a particular species relative to the total biomass over time	1b	weekly-monthly	point	V(TS), DF	Time-series of species/group fraction, and/or visual comparison of relative distribution plot as a function of time (e.g. fraction of copepods relative to total zooplankton biomass, or YOY relative to total fish population)	Gal et al. (2009)
Trait-specific species/group characteristics (bulk population)	Consumption and predation rates	2a	*ad hoc*	point	V(TS), R, V (other)	Model comparison with *in situ* consumption rate determinations	
	Consumption and predation timing	2a	*ad hoc*	point	V(TS), V(other)	Validation of model feeding behaviour	
	Egestion, excretion and respiration rates	2a	*ad hoc*	point	V(TS), V(other)	Model comparison with *in situ* metabolic loss rate determinations	
	Prey grazing fraction	2b	*ad hoc*	point	V(TS), DF	Time-series comparison of organism food ingestion rates, compared against available data on observed ingestion rates	Morozov (2010)
	Internal stoichiometry	3	*ad hoc*	point	V(TS), DF	Fish or zooplankton biomass stoichiometry of N, P, HUFA etc. can be used to cross-check metabolic processes are balanced	Perhar et al. (2012)
	Vertical positioning of biomass (e.g. diel migration)	3	*ad hoc*	water column	V(TZ), MAE (ΔTZ)	Correctly capturing spatial localisation of population biomass in the water column can demonstrate behaviours such as diel migration for feeding, predator avoidance or reaction to light and/or temperature	Gal et al. (2004)
Population structuring (population cohort)	Size range and distribution	1c	*ad hoc*	basin, system	DF	For individual-based models, users can demonstrate accurate spread of biomass across a size distribution using a histogram comparison	Makler-Pick et al. (2011)
	Weight-length relationship	3	*ad hoc*	basin, system	V(other)	Accurate depiction of the weight-length relationship of a fish population demonstrates the model is allocating resources to growth correctly	Makler-Pick et al. (2011)
Relative body size relationships (population cohort)	Body mass: net growth rate Body mass: mortality Body mass: reproductive success Body mass: population	3 3 3 3	weekly-annual	basin, system	V(other)	Process rates (growth, mortality, reproduction) should scale with age/size appropriately	Harfoot et al. (2014)
Spatial niche formation	Spatial distribution of species/ group	1a	daily-monthly	multiple points	V(XY), R, V (other)	Spatial survey data compared with modelled presence and abundance	
	Biomass of each species/ functional group as a function of water surface temperature or salinity	1c	monthly-seasonal	multiple points; surface layer	CCF	The example given is for a cross-model comparison rather than a model validation against data	Sailley et al. (2013)
	Biomass – depth relationship	2	monthly-seasonal	horizontal transect	V(TX)	Biomass as a function of depth along a transect, or depth limit of biomass observations	Greve and Krause-Jensen (2005)

(*continued on next page*)

Table 4 (*continued*)

Property being assessed	Approach and description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g., data collection and processing requirements)	Example references [b]
	Spatial variability in habitat quality	3	*ad hoc*	basin scale	V(other), R	Suitability of conditions for presence of species of interest – presence and activity of those species Habitat index (e.g. HSI) correlation with observed survey data	
	Biomass – environment relationships	3	*ad hoc*	system	V(other)	Multi-variate statistical clustering of organism presence/biomass, as a function of environmental conditions	
Recruitment and migration events	Timing of reproduction events (e.g., spawning, egg-hatching)	1b	annual	system	V(TS), V(other), R	Checking model timing of critical life-history events such as spawning can ensure correct model formulation and response to environmental cues	
	Magnitude	2a	annual	system	V(TS), V(other), R	Accurate depiction of biomass allocated to recruitment (e.g., YOY)	
	Larvae dispersal	2a	hourly-weekly	system	V(XY)	Comparison of measured larvae densities with modelled extent	
	Timing of migration events	2a	annual	system	V(TS), V(other), R	Checking model timing of migration events	
Community structure							
Foodweb organisation	PREBAL diagnostics	0	–	system	V(other)	For complex ecosystems models "PREBAL diagnostics" include biomasses, biomass ratios, vital rates, vital rate ratios, total production, and total removals (and slopes thereof) across the taxa and trophic levels"	Link (2010) Heymans et al. (2016)
	Food chain length	3	*ad hoc*	system	V(other)	For models resolving a variable food-web structure 13C and 15N measurements or dietary analysis can be used to confirm food chain length and structure	Sprules and Bowerman (1988) Cabana and Rasmussen (1996) Vander Zanden et al. (1999) Dame and Christian (2008)
	Multi-variate state-space relationship	3	weekly-seasonal	multiple points	TD	Taylor diagram of relevant groups to allow comparison of model performance of each trophic level	Raick et al. (2006) Doney et al. (2009)
Trophic level biomass distribution	Time series comparison	1b	weekly-monthly	point	V(TS)	Time-series of species fraction and/or visual comparison of biomass of a trophic level relative to the community biomass	Gal et al. (2009)
	Autotroph:heterotroph	3	weekly-monthly	point	V(other)	Scaling of autotrophic biomass relative to heterotrophic groups	Harfoot et al. (2014)
	Trophic structure	3	*ad hoc*	system	V(other)	Scaling ratios of autotroph:herbivore:omnivore:carnivore biomass	Harfoot et al. (2014)
Reference ecological indicators	ECOIND	3	*ad hoc*	system	V(other)	Assessment of changes to key food-web indicators over time; these indicators could serve as high level validation metrics	Coll and Steenbeek (2017)
	Species trophic position	3	*ad hoc*	point, basin	V(other)	Comparison of organism food source ingestion rates with estimates of trophic position from stable isotopes	Post (2002) Corrales et al. (2017b)
	Relationship between food chain length and net primary production	3	*ad hoc*	system	V(other)	Food chain length is expected to increase with net primary production	Harfoot et al. (2014)
Diversity and richness indices	Biodiversity indices	1b	monthly-annual	system	V(other)	Most biodiversity and diversity indices are only calculable for complex multispecies models. In such cases, rather than directly comparing index values, relationships between indices and environmental variables can be compared, e.g. over multiple sites	Washington (1984) Simpson and Norris (2000)
	Kempton's species diversity index	3	*ad hoc*	system	V(other)		Ainsworth and Pitcher (2006)
	Species richness Shannon index	3	weekly-seasonal	point	V(other)		Sauterey et al. (2015)
Community function							
Community respiration	Total respiration	2b	minutes-hourly	Point or reach	V(TS), R, DF	Oxygen metabolism calculated from oxygen sensor logging at high frequency during darkness in the surface layer	Townsend et al. (2011)
	Planktonic community respiration	2b	minutes-hourly	Point	V(TS), R, DF	Comparison with results from labelled carbon isotope uptake assays	Testa et al. (2017) Brush and Nixon (2017)
	Turnover rates of autotrophs and heterotrophs	2b	daily-annual	system	V(other), DF		Harfoot et al. (2014)

(*continued on next page*)

Table 4 (*continued*)

Property being assessed	Approach and description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g., data collection and processing requirements)	Example references [b]
						Respiration per unit biomass, compared for autotrophs and heterotrophs, where the latter is computed by subtracting autotrophic respiration from the total	
Foodweb linkages	Dietary analysis of food sources	2a	*ad hoc*	system	V(other), DF	Intended food web structure can be discerned from model equations, but actual structures emerging from model runs may be simpler, as some interactions may be very weak in simulated conditions.	Sailley et al. (2013)
	Isotope based food source identification	2b	*ad hoc*	system	V(other)	Confirmation of pathways in food webs by using 13C and 15N measurements, and/or dietary analysis can validate the strength and stability of trophic linkages	Cabana and Rasmussen (1996) Vander Zanden et al. (1999)
	Intraguild predation	3	*ad hoc*	system	V(other)	Ability of model to capture adaptation of pathways in response to changing predator presence and behaviour	Makler-Pick et al. (2017)
	Trophic efficiency	3	monthly-seasonal	system	V(other)	Relationship between heterotrophic biomass and net primary production indicates upward transfer of energy	Sailley et al. (2013)
	Relationship between food chain length and net primary production	3	monthly-seasonal	system	V(other)	Food chain length is expected to increase with net primary production	Harfoot et al. (2014)
	Brown vs green trophic partitioning	3	monthly-seasonal	system	V(other)	Models resolving microbial loop pathways can demonstrate carbon energy flow pathways from detrital vs algal sources match expected	Li et al. (2014)
Catch rates	Total yield or CPUE	2a	monthly-annual	basin; system	V(TS), R	Models predicting the catch of a species can validate the harvest amount against records	Corrales et al. (2017a)
	Species partitioning within the catch	3	monthly-annual	basin; system	V(other), DF	For end-to-end models, several metrics pertaining to the catch composition can be validated with observations, (e.g. trophic level of catch etc., as per the ECOIND plug-in for EwE).	Coll and Steenbeek (2017)
Ecosystem response to stress, non-linearity and thresholds							
Ecosystem state sensitivity to external forcing	Response to external nutrient loading	3	monthly-annual	system	R	Scatter plot comparison of mean nutrient, oxygen, Chl-*a* or population abundance relative to N and/or P loading	Jones and Lee (1988) Reynolds (2006) Gal et al. (2009)
	Response to external sediment loading	3	monthly-annual	system	R	Scatter plot comparison of mean turbidity, Chl-*a* or population abundance relative to sediment loading	Robson et al. (2017)
	Response to climate changes (inflows, water level, temperature, CO_2)	3	monthly-annual	system	R, V(TS)	Scatter plot comparison of relevant ecosystem variables (pH, O_2, Chl-*a*, nutrients or population abundance) to hydro-climatic forcing variable	
Variable coherency	Time-series analysis	1b	daily-decadal	system	CCF	Identifying temporal (or spatial) coherence and strengths of relationships between linked variables (e.g. phytoplankton and zooplankton) under non-stationary forcing conditions	
	Spectral analysis	1b	hourly-decadal	point	WT, WC	Changes in the strength of variable interaction following perturbation could be depicted by wavelet coherency, showing coupling and/or de-coupling of correlation between two or more variables in frequency space	Huang et al. (2019)
	State-space diagrams	3	monthly-decadal	system	V(other)	State-space representations of linked variables can be used to compare model with observations in systems with complex time-histories. e.g. to depict Lorenz attractor type relationships between variables	
Ecosystem index sensitivity to external forcing	Response of water quality to hydrodynamics	3	monthly-annual	system	V(other)	Comparison of model sensitivity in water quality to externally driven changes in hydrodynamics, e.g., water quality index changes in response to water level manipulation	Parparov and Gal (2012)
	Response of habitat to hydrodynamics and nutrient loading	3	monthly-annual	system	V(other)	Comparison of model sensitivity in habitat quality to externally driven changes in hydrodynamics and nutrient delivery, e.g., seagrass habitat suitability in response to different flow regimes	Collier et al. (2017)

(*continued on next page*)

Table 4 (*continued*)

Property being assessed	Approach and description	Validation level	Typical range of data observation frequency	Spatial scale	Assessment technique (see Table 1)[a]	Comments (e.g., data collection and processing requirements)	Example references [b]
	Response of biodiversity/ richness indices to hydrodynamics and nutrient loading	3	monthly-annual	system	V(other)	Comparison of model sensitivity in ecosystem indices to externally driven changes in hydrodynamics and nutrient delivery, e.g., fish diversity index in response to different flow regimes	
State variable or population response to chronic and/or acute perturbations	Magnitude of threshold change in state	3	daily-annual	system	V(TS), V(other)	Degree of change in state variable or aggregate index following perturbation ("old" - "new" attractor state)	Dai et al. (2012)
	Timing and abruptness of threshold change	3	daily-annual	system	V(TS), V(other)	Assessment of when a critical transition occurs (e.g. turbidification), and the rapidity of state transition	Bayley et al. (2007)
	Critical external forcing (e.g. nutrient loading) to induce abrupt regime shift	3	daily-annual	system	V(other)	Measure of resistance or ability of state (e.g. macrophyte density) to resist stress	Janse et al. (2008) Janse et al. (2010) O'Brien et al. (2018)
	Return time to "old" or "new" state	3	daily-annual	system	V(TS), V(other)	Recovery time to original state (full recovery) or new state (partial recovery	Müller et al. (2016) O'Brien et al. (2018)
	Magnitude of return towards the "old" state	3	daily-annual	system	V(TS), V(other)	Comparison of model ability to capture the magnitude of return of ecosystem state	Janse et al. (2008) Janse et al. (2010)
	Potential for bistability	3	daily-annual	system	V(other)	Comparison of a model's ability to capture bistability, where evidence for bistability exists	Dai et al. (2012) Kuiper et al. (2015) Adams et al. (2016)
	Critical slowing down	3	daily-monthly	system	V(other), ACF, WT	Comparison of a model's ability to capture the lengthening response times to disturbance, prior to a catastrophic transition	Dai et al. (2012) Dakos et al. (2012a) Dakos et al. (2012b) Kuiper et al. (2015)
Response to restoration effort	Restoration effort (e.g. nutrient load reduction) required to reverse state transition	3	monthly-annual	system	V(other)	Demonstration of a model's ability to resolve the critical recovery point and presence of hysteresis in ecosystem state (e.g. in response to oligotrophication)	Janse et al. (2008) Janse et al. (2010)
	Extent of hysteresis	3	monthly-annual	system	V(other)		
	Recovery time	3	monthly-annual	system	V(TS), V(other)	Time frame until recovery (e.g. of a seagrass ecosystem) following restoration interventions (such as oligotrophication, replanting)	O'Brien et al. (2018)
Cross-site coherency	Time series comparison	3	monthly-annual	system	V(TS), R	Validation of multiple variables concurrently across multiple sites provides confidence in the model setup configuration	Dietzel et al. (2013)

[a] E = user determined combination of traditional error metrics, including BIAS, MAE, NMAE, MEF, NSE, B and/or d_2.

[b] Where possible references provided indicate an example of the metric being applied, otherwise references relevant to give context to use of the metric are listed.

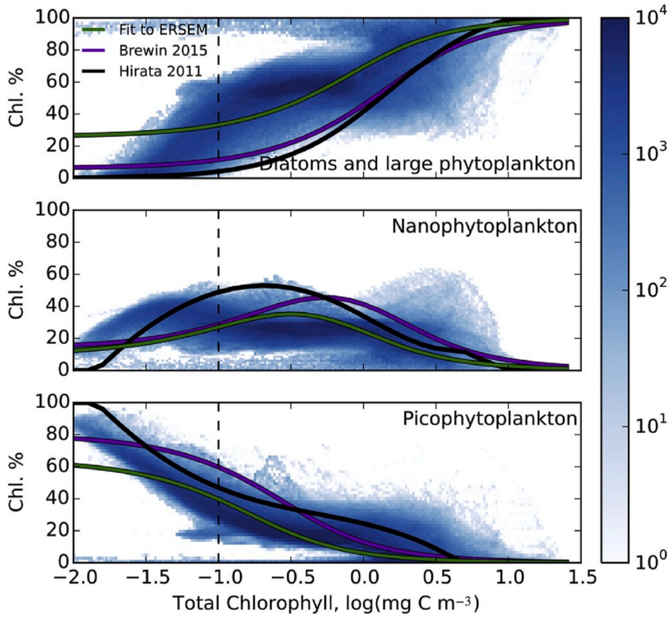

Fig. 6. The phytoplankton community structure of the NEMO-ERSEM global biogeochemical model. The relative abundance at the ocean surface layer of each phytoplankton group is shown in blue-scale as a function of chlorophyll-a concentration. The green line is the least-squares fit of the model data to the three-population absorption model of Brewin et al. (2014), and the purple line is the fit of historic *in situ* data to the three-population absorption model from Brewin et al. (2014). A fit to data from Hirata et al. (2011) is shown in a black line. The dashed vertical line indicates a typical detection limited of HPLC and SFF methods. Reproduced from de Mora et al. (2016) without modification under license CC BY 4.0. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

production efficiency of limiting nutrients increases when food quality is poor. To validate their models, they compared growth rates under varying nutrient use efficiency across reported food quantity/quality gradients. Simulating growth rates with varying production efficiency under varying food quantity and quality is only possible when correct simulation of the various processes that link food quantity, quality and growth rates is achieved. Additional unique markers such as fatty acid concentrations have been used for high level model assessment. Perhar et al. (2012) used biochemical control on zooplankton growth by N, P and Highly Unsaturated Fatty Acids (HUFA) as a means for testing accuracy of model output. As with stable isotopes and stoichiometry, correct simulation of HUFA concentrations require accurate model description of multiple processes within the food-web.

As model structures and the number of species interactions becomes more complex, the validation challenge becomes considerably more difficult. Sauterey et al. (2015) simulated planktonic diversity (as indicated by the Shannon index, which is a measure of the distribution of biomass among species) and evolving cell sizes of dominant plankton species in a global ocean model. They used the Canberra distance to compare the distance between cell-volume distributions of different model outcomes. Although Sauterey et al. did not directly compare their model results with real-world Shannon and Canberra indices, their work outlined the possibility for how the emergent feature of planktonic community diversity could be used as an additional performance metric for ecological models. Goebel et al. (2010) were able to demonstrate how the expected patterns of plankton distribution in a coastal environment emerged in the model from the interaction of a highly diverse population.

Much of the above discussion about pelagic plankton communities similarly applies to benthic communities, bearing in mind that generally their position is fixed for most of their life history. For individual macrophyte, macroalgal, or invertebrate species simulations, biomass or organism density in coastal and lake models can be assessed across space and time to ensure habitats are accurately represented (Savina and Ménesguen, 2008). *In situ* rates of detritus production (assessed by traps in flowing environments) and rates of decay of benthic plant detritus (assessed with incubation chambers) can be used as a Level 2 metric. Benthic plant productivity can be assessed by measuring changes in plant biomass when grazers are experimentally excluded, and grazing rates can be assessed through food-web studies supported by stable isotope measurements. Similarly, filtration and clearance rates of filter feeders can be used for species such as mussels. The 3D model study by Bocaniov et al. (2014) simulated zebra and quagga mussels and whilst there was limited *in situ* data for direct mussel filtration rate validation, the effect of the mussels on improving the Chl-a prediction in the overlying water column was used as a proxy indicator to help justify the filtration rate predictions. In a study by Renton et al. (2011), the development of a restored seagrass bed was simulated using a functional-structural plant model. Results were assessed against total rhizome length, length of the longest rhizome axis and total number of live buds (apices/axes) and internodes, based on a snapshot of data taken two years after the restoration began. At the community level, benthic plant succession can be assessed. Often it is the variability in benthic communities and biomass along a gradient of light/depth and their relationship with patterns of benthic substrates that is important. Li et al. (2010), tested a multi-agent systems model of two macrophyte species in Lake Veluwe and illustrated performance using a map of occurrence indicating regions of model over and under-prediction.

Evaluating model simulations of populations of fish and 'higher' biotic populations is somewhat more complex than variables described above due to lower sampling resolution, species mobility and different indices used to characterise fish populations. Unless regular stock assessment data for fisheries (e.g., Savina et al., 2013) or other organism counts are available, it is difficult to assess model performance using traditional Level 1 indicators (Lehuta et al., 2013), and population models should be assessed using an array of alternative measures. Models of populations often adopt individual-based approaches and the spatial context of predictions needs to be considered in light of the sampling regime used to collect observations. For fish, observed data are often based on catch data. Metrics such as catch per unit effort must therefore be translated to match simulation variables or used to qualitatively assess model spatio-temporal patterns of fish abundance (e.g., Holt et al., 2014; Savina et al., 2013). Higher level indicators often used to characterise fish populations include length-weight (L-W) relationships, length/weight at age and size distribution histograms (Makler-Pick et al., 2011; Megrey et al., 2007; Rose et al., 2007). All three indicators emerge dependent on a large number of individual and population-specific interactions with environmental conditions, thus accurate predictions serve as an indication of model robustness. However, the predictions usually also integrate over large spatial and/or temporal scales, and thus are not truly assessing population response at finer scales. Breckling et al. (2005) and Hölker and Breckling (2005) discuss Level 3 emergent properties relevant to fish population modelling: self-sorting age groups, trophic bottlenecks, size-dependent winter mortality, spatial organisation measures, and the influence of lake morphology on phenotype. Assessing changing population size distributions in response to fishing pressure (e.g. Makler-Pick et al., 2011), can also be a means to qualitatively assess ecosystem response to external forcing.

Assessment of food webs with a high degree of variable interaction may take the form of a multivariate assessment tool (e.g., TD), or more *ad hoc* tests of theoretically relevant patterns, trends and relationships. An example of the latter is the approach by Fulton et al. (2004), where a range of semi-quantitative tests and qualitative comparisons of population and ecosystem level metrics for coastal embayment models were used. Predictions of the trophic structure were put in context of the Sheldon spectra, to demonstrate mass partitioning between trophic levels was appropriate. Sailley et al. (2013) also compared trophic efficiency metrics that could be used to assess how complex food webs emerge, including comparison of bulk community heterotroph to autotroph ratios, and variation of zooplankton predator: micozooplankton, and predator:prey scaling relationships. These metrics were used in the context of comparing model structures, however, they may also serve as a way for modellers to compare food-web predictions with data. Dynamic relationships that emerge within food-webs such as intraguild predation, and competition between species, may also be used to assess models, however, specific Level 3 metrics quantifying these measures of system-organisation are difficult to define. Reynolds and Elliott (2012) explore the predictability of several emergent properties of freshwater ecosystems including carrying capacity, exergy accumulation, carbon processing capacity and habitat templates (i.e. functional zones). They conclude, that "*while species composition may remain quite unpredictable, except on the basis of probabilities and hindsight, the characteristic traits of the successful contestants can be anticipated with considerable certainty*" (p. 87). For more complex food web analyses, Deehr et al. (2014), demonstrated a novel approach to validation of a complex EcoPath food-web through integration with N-isotope data. They demonstrate the strong relationship between effective trophic level (ETL) from the model and the $\delta^{15}N$ signature from observed organism data, in the context of a marine ecosystem subject to trawling pressures.

Ultimately, a large number of modellers are seeking to elucidate fundamental ecological relationships and forecast potentially complex response pathways of ecosystems to changes in external and internal drivers. Metrics described in the above sections specifically support assessing sub-model components (e.g., hydrodynamics, nutrients, phytoplankton etc.), but there are range of more specific metrics that can be used to holistically assess model predictions. An area of increasing interest is demonstrating models are able to capture system-scale Level 3 metrics such as resilience to perturbation, thresholds and stable state transitions, and hysteresis effects (Hipsey et al., 2015; Müller et al., 2016). As yet there remain limited examples where models have been confronted with empirical data that display these trends. Challenges exist in terms of computing compound indices that can be used as indicators of ecosystem "state", though indices of water quality or ecosystem diversity are increasingly being used. In a shallow lake example, Janse et al. (2008, 2010), were able to demonstrate the ability of their model to capture the threshold shift from macrophyte to algal dominance, validated by an assessment across multiple lakes. In doing so they were able to identify the threshold P loading level required for "turbidification", and subsequent "restoration" including demonstration of hysteresis effects, which agreed well with empirical work (Fig. 7). In a stability analysis based on the same model, Kuiper et al. (2015) showed that the food web and system stability gradually decreases with the distance from the critical loading in the bistability range, for both directions, thereby highlighting the potential for correctly formulated models to inform users on phenomena such as critical slowing down, ecosystem flickering, and system resilience. These studies hold great promise for informing assessment of the next generation of AEMs that can be more confidently applied to predict ecosystem collapse, recovery and restoration strategies.

5. Discussion

The four levels of model validation may be challenging, bearing in mind in the past often inadequate data have been limiting the extent to which assessment could be advanced. However, we are seeing an ever increasing range of data streams from new monitoring technologies such as optical nutrient loggers (Rode et al., 2016; Claustre et al., 2019), improved processing of existing sources such as satellite observations (Jouini et al., 2013), citizen science initiatives (Dickinson et al., 2010), open-access and long-term monitoring initiatives, and real-time data portals (Reed et al., 2010). All these data hold the potential to improve the way we run and assess environmental models. Indeed, aquatic ecosystems modellers are beginning to take up these data streams (Li et al., 2010; Johnson and Needoba, 2008; Turuncoglu et al., 2013) and it is timely to reconsider the ways we can use this data to improve our model formulations and to describe model uncertainty.

5.1. Improving models through improved assessment

Several recent commentaries have discussed the challenges and issues in application of complex environmental models in general (Nordstrom, 2012) and AEMs in particular (Robson, 2014a; Trolle et al., 2012; Arhonditsis et al., 2014; Frassl et al., 2019). The emergence of community models and flexible modelling frameworks to reduce duplication of effort (e.g. Bruggeman and Bolding, 2014; Mooij et al., 2014; Hipsey et al., 2019), and the application of advanced techniques for assessment of model error and sensitivity, go some way towards addressing these challenges. The CSPS framework and examples we provide is an attempt to help modellers move from relying on Level 1 metrics, to more robust and insightful Level 2 and 3 metrics in order to more thoroughly challenge our models and assess their capabilities and limitations. As these metrics become more widely used and reported, they will facilitate an increased depth of analysis in comparative studies (e.g., Salihoglu et al., 2013; Kim et al., 2014; Kwiatkowski et al., 2014; Tittensor et al., 2018) and help us to assess how different model structures, parameterisations, and algorithms perform across a diverse range of applications and simulation contexts.

For the foreseeable future, there may be insufficient data to complete assessment at all four levels in every case. Nonetheless, being aware of the diversity and suitability of metrics at multiple levels and being explicit about the level (0–3) at which assessment has been conducted will help modellers to communicate the type of assessment that is being performed and the implications for model uncertainty. This awareness may also facilitate prioritisation of observational studies and monitoring programs that consider not only state variables, but also fluxes and emergent properties (in other words, ecological "states, rates, and traits"). Undertaking assessment using this structured approach can help modellers to further pinpoint where models are fundamentally weak and communicate to stakeholders where further investment in data collection and monitoring will support prediction. Conversely, modelling that leads to discovery of new or interesting phenomena can motivate new experiments or monitoring to support post-hoc validation efforts.

5.2. Model purpose and selection of appropriate metrics

Technical assessment of models is varied and requires the development of workflows that bring together several methods, tailored to the specific application (Bennett et al., 2013) and taking into account the intended purpose of the model (e.g. Harmel et al., 2014). The examples in Tables 2–4 are intended to provide an expandable library that can serve as the basis of a common reference for assessment of aquatic system models. Some examples may serve to cross-fertilise ideas across sub-disciplinary divides. For example, Target or Taylor diagrams are widely used in oceanography, but not routinely used for freshwater models.

Note that not all are relevant in any given case; for example, a specific relationship may be used to parameterise the model directly, in which case it is less suitable as a validation tool. Others may only be applicable to specific spatial or temporal model resolutions. For example, a metric quantifying stratification and mixing requires a minimal vertical disaggregation of the water column. The selected metrics will also depend on the model purpose. Models used for operational forecasting may prefer different assessment metrics than those designed to explore algal seasonal dynamics or long-term nutrient load assessments. In either case, a combination of Level 1 metrics may demonstrate the model's potential, and support with Level 2 metrics would help to ensure the model was not over-fitted and introduce a higher level of credibility into the assessment. However, if the intention

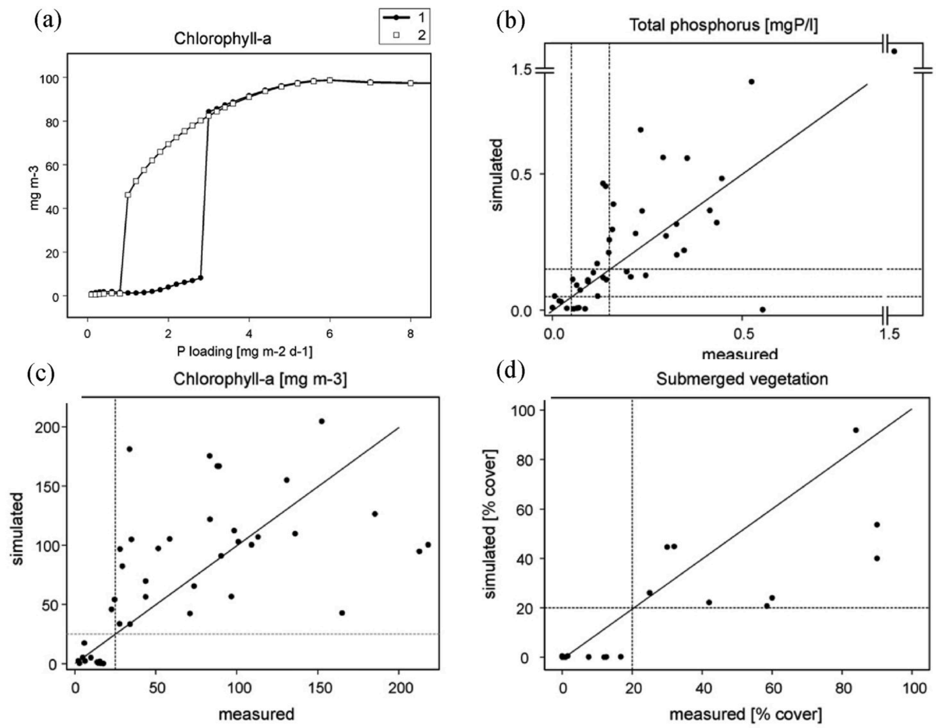

Fig. 7. Analysis using the model PCLake to (a) demonstrate the threshold loading rate of P to transition from a clear macrophyte-dominated state to a turbid, phytoplankton dominated state, and hysteresis effect (based on sequential simulations undertaken increasing load (series 1, black circles) and decreasing load (2, hollow squares)). Validation of the model states against a multi-lake data set of (b) TP, (c) Chl-a and (d) submerged vegetation, spanning the loading range, confirms the model captures ecosystem organisation over a wide range of conditions. Images reprinted from Janse et al. (2010), with permission from Elsevier.

is to then apply the model outside the bounds of historical conditions to explore the impact of major system changes on ecosystem dynamics, then including Level 3 validation becomes an important step. Well-validated modelling studies in different disciplinary areas (e.g., ponds, lakes, rivers and marine systems) demonstrating use of metrics across multiple levels are encouraged to serve as benchmarks that can guide practitioners.

5.3. Using multiple metrics to enrich the calibration and uncertainty assessment process

Calibration of AEMs has historically been a relatively manual 'trial and error' process, though formal calibration methods are beginning to be used in both marine (e.g. Parslow et al., 2013) and freshwater ecosystem modelling (e.g. Ramin and Arhonditsis, 2013; Dietzal and Reichert, 2012). The adoption of calibration-validation pairing (e.g. Trolle et al., 2008), where validation relies on a data set independent from that used in calibration, remains the exception rather than the rule in aquatic ecosystem modelling, though it is common in other fields and widely considered best practise (Robson, 2014b). Adopting a calibration and validation period can avoid overfitting model parameters, and identify the predictive capability of a model, particularly where patterns during the validation period differ from those in the calibration period. Widening the range of metrics models are assessed against may assist in achieving a broader application of calibration-validation pairing.

In recognition of model uncertainty and equifinality problems, there has been a shift in some areas of the AEM community to change model calibration practice from seeking a single "optimal" value for each model parameter, to seeking a range of parameter sets that all meet a pre-specified standard of agreement with the data (Aldenberg et al., 1995; Stow et al., 2007; Arhonditsis et al., 2007; Chiu and Gould, 2010; Janse et al., 2010). Running an ensemble of simulations using values from amongst these acceptable parameter sets provides a basis for estimating the uncertainty associated with model predictions. This practice, termed "physical-statistical modelling" (Kuhnert, 2014), relies on Bayesian probability to combine existing (prior) information with observations to project the (posterior) likelihood of ecosystem response. The effective characterisation of model uncertainty using Bayesian approaches depends upon two critical steps: i) selection of a sampling scheme to generate input vectors (e.g., Latin hypercube, Markov Chain Monte Carlo etc.), and ii) selection of a likelihood measure to quantify model misfit. In complex models, the choice of likelihood measures for assessment leads to conceptual dilemmas for modellers such as the selection of likelihood functions that can meaningfully change the inference drawn (Beven and Freer, 2001; Hong et al., 2005; Arhonditsis et al., 2008b). By further tailoring the adopted likelihood measures to consider the assessment ideas introduced in this paper the model calibration and uncertainty process can be enriched to focus on more diagnostically powerful likelihood measures.

Finally, recognising that there is no true model of an ecological system, but rather several adequate descriptions of different conceptual basis and structure, ensemble modelling is a means to obtain better predictions and a better understanding of uncertainty by combining the results of 'competing' models (Trolle et al., 2014). Several methods exist to synthesise predictions across ensembles, including sequential data assimilation approaches (such as the ensemble Kalman filter and ensemble particle filters; Moradkhani et al., 2006; Vrugt and Robinson, 2007), and post-hoc ensemble integration strategies such as the Bayesian Model Averaging (BMA) (Ramin et al., 2012). Including models of differing complexity and with varied structures in the ensemble allows structural uncertainty to be addressed alongside uncertainty from input data and parameter selection. When combined with the use of more specific and nuanced assessment metrics, modellers can better decide which model structures perform best, without an overt reliance on diagnostically weak Level 1 error metrics, thus motivating reconsideration of their Level 0 validation.

6. Conclusions

In evaluating the performance of a model, we want to know the answers to several questions: Is the model capable of reproducing observations? If so, is it getting it right for the right reasons (or conversely, is it over-fitted or does it have one error cancelling another)? Can we trust the model to make predictions? If so, in what range of circumstances can we trust it? To answer these questions in the case of mechanistic AEMs, we need to go beyond simply comparing simulated and observed concentrations of state variables. Here, we present a way forward; the hierarchical CSPS framework to encourage evaluation of models at four levels: conceptual accuracy (Level 0), state accuracy (Level 1), process accuracy (Level 2), and accuracy in capturing system behaviour (Level 3). Assessment at Level 2 can improve confidence in the biogeochemical basis of model formulations, while assessment at Level 3 allows modellers to critically assess model predictions against spatial and temporal scales of change, stoichiometric indices, and a range of trophic relationships, all of which are based on theoretically-informed indicators of ecosystem function. Arguably, only applications that perform well at highest level of assessment justify the implementation of complex, process-based models. Short-to mid-term forecasting predictions, after all, can often be less expensively and more accurately produced through simpler approaches such as regression modelling (Robson and Dourdet, 2015) or evolutionary algorithms (Recknagel et al., 2014), while if the aim is to shed light on system function, a model that fails to perform at Level 3 may be producing the "right answer for the wrong reason". Though it may not always be possible to rigorously assess a model at all levels, modellers should strive to ensure the level of assessment is suited to the purpose of the model and the severity of consequences of an incorrect prediction. A model that is successfully validated for a specific system at all four levels is one that can be applied with confidence to forecast future trajectories of the system. A model that has been successfully validated across several systems at all four levels is one that can more generally be applied with confidence. Over time, it is envisioned that the community–driven adoption of these metrics will accelerate advances in model structure and function and provide an improved foundation for model assessment on which developments in model-data fusion can be built.

Software and data availability

The system of metrics reported in the present analysis is an evolving collection and can be accessed online at: https://aquaticecodynamics.github.io/aem-metrics/.

Author contributions

MRH, GG and BJR designed the framework outline and prepared the main manuscript body with GA; All authors contributed to metric identification, literature review, preparation of the tables and figures, and manuscript development and revision. Funding: MRH received support from the Australian Research Council [DP130104078, DP170104832, LP130100756, LP150100451]. CCC received support from the National Science Foundation [NSF1753639].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge discussions with members of the Aquatic Ecosystem Modelling Network (AEMON), including Louise Bruce, Karsten Rinke, Paul Hanson, David Hamilton and Wolf Mooij, and the suggestions and improvements made by two anonymous reviewers.

References

Acreman, D.M., Jeffery, C.D., 2007. The use of Argo for validation and tuning of mixed layer models. Ocean Model. 19 (1–2), 53–69.
Adams, M.P., Hovey, R.K, Hipsey, M.R., Bruce, L.C., Ghisalberti, M., Lowe, R.J., Gruber, R.K., Ruiz-Montoya, L., Maxwell, P.S., Callaghan, D.P., Kendrick, G.A., O'Brien, K.R., 2016. Feedback between sediment and light for seagrass: Where is it important? Limnol. Oceanogr. 61 (6), 1937–1955. https://doi.org/10.1002/lno.10319.
Adiyanti, S., Eyre, B.D., Maher, D.T., Santos, I., Golsby-Smith, L., Mangion, P., Hipsey, M.R., 2016. Stable isotopes reduce parameter uncertainty of an estuarine carbon cycling model. Environ. Model. Software 79, 233–255.
Ainsworth, C.H., Pitcher, T.J., 2006. Modifying Kempton's species diversity index for use with ecosystem simulation models. Ecol. Indicat. 6 (3), 623–630.
Aldenberg, T., Janse, J.H., Kramer, P.R.G., 1995. Fitting the dynamic lake model PCLake to a multi-lake survey through Bayesian statistics. Ecol. Model. 78, 83–99.
Alewell, C., Manderscheid, B., 1998. Use of objective criteria for the assessment of biogeochemical ecosystem models. Ecol. Model. 107, 213–224.
Alexander, R., Imberger, J., 2008. Spatial distribution of motile phytoplankton in a stratified reservoir: the physical controls on patch formation. J. Plankton Res. 31 (1), 101–118.
Allen, J.I., Somerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. J. Mar. Syst. 64, 3–14.
Allen, J.I., 2010. On the emergent properties of marine ecosystems. PhD Dissertation. http://ediss.sub.uni-hamburg.de/volltexte/2010/4847/pdf/Emergent_Props_Ecosystem_Models-JI-Allen.pdf.
Andersen, K.H., Aksnes, D.L., Berge, T., Fiksen, Ø., Visser, A., 2015. Modelling emergent trophic strategies in plankton. J. Plankton Res. 37 (5), 862–868.
Anderson, T.R., Gentleman, W.C., Sinha, B., 2010. Influence of grazing formulations on the emergent properties of a complex ecosystem model in a global ocean general circulation model. Prog. Oceanogr. 87 (1–4), 201–213.
Anderson, T.R., Mitra, A., 2010. Dysfunctionality in ecosystem models: an underrated pitfall? Prog. Oceanogr. 84 (1–2), 66–68.
Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Mar. Ecol. Prog. Ser. 271, 13–26.
Arhonditsis, G.B., Adams-VanHarn, B.A., Nielsen, L., Stow, C.A., Reckhow, K.H., 2006. Evaluation of the current state of mechanistic aquatic biogeochemical modeling: citation analysis and future perspectives. Environ. Sci. Technol. 40 (21), 6547–6554.
Arhonditsis, G.B., Perhar, G., Zhang, W., Massos, E., Shi, M., Das, A., 2008a. Addressing equifinality and uncertainty in eutrophication models. Water Resour. Res. 44 (1).
Arhonditsis, G.B., Papantou, D., Zhang, W., Perhar, G., Massos, E., Shi, M., 2008b. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. J. Mar. Syst. 73 (1-2), 8–30.
Arhonditsis, G.B., Qian, S.S., Stow, C.A., Lamon, E.C., Reckhow, K.H., 2007. Eutrophication risk assessment using Bayesian calibration of process-based models: application to a mesotrophic lake. Ecol. Model. 208 (2–4), 215–229.
Arhonditsis, G.B., Stow, C.A., Rao, Y.R., Perhar, G., 2014. What has been accomplished twenty years after the Oreskes et al.(1994) critique? Current state and future perspectives of environmental modeling in the Great Lakes. J. Great Lake. Res. 40, 1–7.
Arora-Williams, K., Olesen, S.W., Scandella, B.P., Delwiche, K., Spencer, S.J., Myers, E. M., Abraham, S., Sooklal, A., Preheim, S.P., 2018. Dynamics of microbial populations mediating biogeochemical cycling in a freshwater lake. Microbiome 6 (1), 165.
Ayata, S.D., Lévy, M., Aumont, O., Sciandra, A., Sainte-Marie, J., Tagliabue, A., Bernard, O., 2013. Phytoplankton growth formulation in marine ecosystem models: should we take into account photo-acclimation and variable stoichiometry in oligotrophic areas? J. Mar. Syst. 125, 29–40.
Baird, M.E., Cherukuru, N., Jones, E., Margvelashvili, N., Mongin, M., Oubelkheir, K., Ralph, P.J., Rizwi, F., Robson, B.J., Schroeder, T., Skerratt, J., 2016a. Remote-sensing reflectance and true colour produced by a coupled hydrodynamic, optical, sediment, biogeochemical model of the Great Barrier Reef, Australia: comparison with satellite data. Environ. Model. Software 78, 79–96.
Baird, M.E., Adams, M.P., Babcock, R.C., Oubelkheir, K., Mongin, M., Wild-Allen, K.A., Skerratt, J., Robson, B.J., Petrou, K., Ralph, P.J., O'Brien, K.R., 2016b. A biophysical representation of seagrass growth for application in a complex shallow-water biogeochemical model. Ecol. Model. 325, 13–27.
Bardsley, W.E., 2013. A goodness of fit measure related to r2 for model performance assessment. Hydrol. Process. 27 (19), 2851–2856.
Bayer, T.K., Burns, C.W., Schallenberg, M., 2013. Application of a numerical model to predict impacts of climate change on water temperatures in two deep, oligotrophic lakes in New Zealand. Hydrobiologia 713 (1), 53–71.
Bayley, S.E., Creed, I.F., Sass, G.Z., Wong, A.S., 2007. Frequent regime shifts in trophic states in shallow lakes on the Boreal Plain: alternative "unstable" states? Limnol. Oceanogr. 52 (5), 2002–2012.
Beck, M.B., 1987. Water quality modelling: a review of the analysis of uncertainty. Water Resour. Res. 23 (8), 1393–1442.
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. Software 40, 1–20.
Bert, F.E., Rovere, S.L., Macal, C.M., North, M.J., Podestá, G.P., 2014. Lessons from a comprehensive validation of an agent based-model: the experience of the Pampas Model of Argentinean agricultural systems. Ecol. Model. 273, 284–298.
Best, E.P., Buzzelli, C.P., Bartell, S.M., Wetzel, R.L., Boyd, W.A., Doyle, R.D., Campbell, K.R., 2001. Modeling submersed macrophyte growth in relation to underwater light climate: modeling approaches and application potential. Hydrobiologia 444 (1), 43–70.
Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249 (1–4), 11–29.

Bocaniov, S.A., Leon, L.F., Rao, Y.R., Schwab, D.J., Scavia, D., 2016. Simulating the effect of nutrient reduction on hypoxia in a large lake (Lake Erie, USA-Canada) with a three-dimensional lake model. J. Great Lake. Res. 42 (6), 1228–1240.
Bocaniov, S.A., Smith, R.E., Spillman, C.M., Hipsey, M.R., Leon, L.F., 2014. The nearshore shunt and the decline of the phytoplankton spring bloom in the Laurentian Great Lakes: insights from a three-dimensional lake model. Hydrobiologia 731 (1), 151–172.
Brady, D.C., Testa, J.M., Di Toro, D.M., Boynton, W.R., Kemp, W.M., 2013. Sediment flux modeling: calibration and application for coastal systems. Estuar. Coast Shelf Sci. 117, 107–124.
Breckling, B., Müller, F., Reuter, H., Hölker, F., Fränzle, O., 2005. Emergent properties in individual-based ecological models—introducing case studies in an ecosystem research context. Ecol. Model. 186 (4), 376–388.
Brigolin, D., Dal Maschio, G., Rampazzo, F., Giani, M., Pastres, R., 2009. An individual-based population dynamic model for estimating biomass yield and nutrient fluxes through an off-shore mussel (Mytilus galloprovincialis) farm. Estuar. Coast Shelf Sci. 82 (3), 365–376.
Brewin, R.J.W., Hirata, T., Hardman-Mountford, N.J., Lavender, S.J., Sathyendranath, S., Barlow, R., 2012. The influence of the Indian Ocean Dipole on interannual variations in phytoplankton size structure as revealed by Earth Observation. Deep Sea Res. Part II Top. Stud. Oceanogr. 77–80, 117–127.
Brewin, R.J., Sathyendranath, S., Tilstone, G., Lange, P.K., Platt, T., 2014. A multicomponent model of phytoplankton size structure. J. Geophys. Res.: Oceans 119 (6), 3478–3496.
Bruggeman, J., Bolding, K., 2014. A general framework for aquatic biogeochemical models. Environ. Model. Software 61, 249–265.
Bruce, L.C., Hamilton, D., Imberger, J., Gal, G., Gophen, M., Zohary, T., Hambright, K.D., 2006. A numerical simulation of the role of zooplankton in C, N and P cycling in Lake Kinneret, Israel. Ecol. Model. 193 (3–4), 412–436.
Bruce, L.C., Cook, P.L.M., Teakle, I., Hipsey, M.R., 2014. Hydrodynamic controls on oxygen dynamics in a riverine salt-wedge estuary, the Yarra River estuary, Australia. Hydrol. Earth Syst. Sci. 18, 1397–1411.
Bruce, L.C., Frassl, M.A., Arhonditsis, G.B., Gal, G., Hamilton, D.P., Hanson, P.C., Hetherington, A.L., Melack, J.M., Read, J.S., Rinke, K., Rigosi, A., Trolle, D., Winslow, L., Adrian, R., Ayala, A.I., Bocaniov, S.A., Boehrer, B., Boon, C., Brookes, J. D., Bueche, T., Busch, B.D., Copetti, D., Cortés, A., de Eyto, E., Elliott, J.A., Gallina, N., Gilboa, Y., Guyennon, N., Huang, L., Kerimoglu, O., Lenters, J.D., MacIntyre, S., Makler-Pick, V., McBride, C.G., Moreira, S., Özkundakci, D., Pilotti, M., Rueda, F.J., Rusak, J.A., Samal, N.R., Schmid, M., Shatwell, T., Snorthheim, C., Soulignac, F., Valerio, G., van der Linden, L., Vetter, M., Vinçon-Leite, B., Wang, J., Weber, M., Wickramaratne, C., Woolway, R.I., Yao, H., Hipsey, M.R., 2018. A multi-lake comparative analysis of the General Lake Model (GLM): stress-testing across a global observatory network. Environ. Model. Software 102, 274–291.
Brush, M.J., Nixon, S.W., 2017. A reduced complexity, hybrid empirical-mechanistic model of eutrophication and hypoxia in shallow marine ecosystems. In: Modeling Coastal Hypoxia. Springer International Publishing, pp. 61–93.
Brush, M.J., Brawley, J.W., Nixon, S.W., Kremer, J.N., 2002. Modeling phytoplankton production: problems with the Eppley curve and an empirical alternative. Mar. Ecol. Prog. Ser. 238, 31–45.
Bryant, L.D., McGinnis, D.F., Lorrai, C., Brand, A., Little, J.C., Wüest, A., 2010. Evaluating oxygen fluxes using microprofiles from both sides of the sediment-water interface. Limnol Oceanogr. Methods 8 (11), 610–627.
Bucci, V., Nunez-Milland, D., Twining, B.S., Hellweger, F.L., 2012. Microscale patchiness leads to large and important intraspecific internal nutrient heterogeneity in phytoplankton. Aquat. Ecol. 46 (1), 101–118.
Butenschön, M., Zavatarelli, M., Vichi, M., 2012. Sensitivity of a marine coupled physical biogeochemical model to time resolution, integration scheme and time splitting method. Ocean Model. 52, 36–53.
Buzzelli, C., Doering, P.H., Wan, Y., Sun, D., Fugate, D., 2014. Modeling ecosystem processes with variable freshwater inflow to the Caloosahatchee River Estuary, southwest Florida. I. Model development. Estuar. Coast Shelf Sci. 151, 256–271.
Cabana, G., Rasmussen, J.B., 1996. Comparison of aquatic food chains using nitrogen isotopes. Proc. Natl. Acad. Sci. Unit. States Am. 93 (20), 10844–10847.
Cantwell, M.G., Katz, D.R., Sullivan, J.C., Borci, T., Chen, R.F., 2016. Caffeine in Boston Harbor past and present, assessing its utility as a tracer of wastewater contamination in an urban estuary. Mar. Pollut. Bull. 108 (1–2), 321–324.
Carey, C.C., Hanson, P.C., Lathrop, R.C., St Amand, A.L., 2016. Using wavelet analyses to examine variability in phytoplankton seasonal succession and annual periodicity. J. Plankton Res. 38, 27–40.
Carraro, E., Guyennon, N., Hamilton, D., Valsecchi, L., Manfredi, E.C., Viviano, G., Salerno, F., Tartari, G., Copetti, D., 2012. Coupling high-resolution measurements to a three-dimensional lake model to assess the spatial and temporal dynamics of the cyanobacterium Planktothrix rubescens in a medium-sized lake. In: Phytoplankton Responses to Human Impacts at Different Scales. Springer, Dordrecht, pp. 77–95.
Carrer, S., Halling-Sørensen, B., Bendoricchio, G., 2000. Modelling the fate of dioxins in a trophic network by coupling an ecotoxicological and an Ecopath model. Ecol. Model. 126 (2–3), 201–223.
Chan, T.U., Hamilton, D.P., Robson, B.J., Hodges, B.R., Dallimore, C., 2002. Impacts of hydrological changes on phytoplankton succession in the Swan River, Western Australia. Estuaries 25 (6), 1406–1415.
Chang, M., Teurlincx, S., DeAngelis, D.L., Janse, J.H., Troost, T.A., van Wijk, D., Mooij, W.M., Janssen, A.B., 2019. A Generically Parameterized model of Lake eutrophication (GPLake) that links field-, lab-and model-based knowledge. Sci. Total Environ. 695, 133887.
Chao, X., Jia, Y., Shields Jr., F.D., Wang, S.S., Cooper, C.M., 2007. Numerical modeling of water quality and sediment related processes. Ecol. Model. 201 (3–4), 385–397.
Chao, X., Jia, Y., Shields Jr., F.D., Wang, S.S., Cooper, C.M., 2010. Three-dimensional numerical simulation of water quality and sediment-associated processes with application to a Mississippi Delta lake. J. Environ. Manag. 91 (7), 1456–1466.
Chapra, S.C., Canale, R.P., 2010. Numerical Methods for Engineers. McGraw-Hill Higher Education, Boston.
Chen, S.N., Sanford, L.P., Koch, E.W., Shi, F., North, E.W., 2007. A nearshore model to investigate the effects of seagrass bed geometry on wave attenuation and suspended sediment transport. Estuar. Coast 30 (2), 296–310.
Chien, Y.C., Wu, S.C., Chen, W.C., Chou, C.C., 2013. Model simulation of diurnal vertical migration patterns of different-sized colonies of Microcystis employing a particle trajectory approach. Environ. Eng. Sci. 30 (4), 179–186.
Chipman, L., Huettel, M., Berg, P., Meyer, V., Klimant, I., Glud, R., Wenzhoefer, F., 2012. Oxygen optodes as fast sensors for eddy correlation measurements in aquatic systems. Limnol Oceanogr. Methods 10 (5), 304–316.
Chiu, G.S., Gould, J.M., 2010. Statistical inference for food webs with emphasis on ecological networks via Bayesian melding. Environmetrics 21 (7–8), 728–740.
Chung, S.W., Imberger, J., Hipsey, M.R., Lee, H.S., 2014. The influence of physical and physiological processes on the spatial heterogeneity of a Microcystis bloom in a stratified reservoir. Ecol. Model. 289, 133–149.
Claustre, H., Johnson, K.S., Takeshita, Y., 2019. Observing the global ocean with biogeochemical-argo. Annu. Rev. Mar. Sci. 12.
Clark, J.B., Long, W., Hood, R.R., 2017. Estuarine sediment dissolved organic matter dynamics in an enhanced sediment flux model. J. Geophys. Res.: Biogeosciences 122 (10), 2669–2682.
Coles, V.J., Stukel, M.R., Brooks, M.T., Burd, A., Crump, B.C., Moran, M.A., Paul, J.H., Satinsky, B.M., Yager, P.L., Zielinski, B.L., Hood, R.R., 2017. Ocean biogeochemistry modeled with emergent trait-based genomics. Science 358 (6367), 1149–1154.
Coletti, J.Z., Vogwill, R., Hipsey, M.R., 2017. Water management can reinforce plant competition in salt-affected semi-arid wetlands. J. Hydrol. 552, 121–140.
Coll, M., Steenbeek, J., 2017. Standardized ecological indicators to assess aquatic food webs: the ECOIND software plug-in for Ecopath with Ecosim models. Environ. Model. Software 89, 120–130.
Collier, C., van Dijk, K., Ertemeijer, P., Foster, N., Hipsey, M., O'Loughlin, E., Ticli, K., Collier, M., 2017. Optimising Coorong Ruppia Habitat. Strategies to Improve Habitat Conditions for Ruppia Tuberosa in the Coorong (South Australia) Based on Literature Review, Manipulative Experiments and Predictive Modelling. Report by University of Adelaide, Department of Environment Water and Natural Resources, University of Western Australia and DAMCO Consulting, Adelaide, Australia.
Corrales, X., Coll, M., Ofir, E., Piroddi, C., Goren, M., Edelist, D., Heymans, J.J., Steenbeek, J., Christensen, V., Gal, G., 2017a. Hindcasting the dynamics of an Eastern Mediterranean marine ecosystem under the impacts of multiple stressors. Mar. Ecol. Prog. Ser. 580, 17–36.
Corrales, X., Ofir, E., Coll, M., Goren, M., Edelist, D., Heymans, J.J., Gal, G., 2017b. Modeling the role and impact of alien species and fisheries on the Israeli marine continental shelf ecosystem. J. Mar. Syst. 170, 88–102.
Couture, R.M., Shafei, B., Van Cappellen, P., Tessier, A., Gobeil, C., 2009. Non-steady state modeling of arsenic diagenesis in lake sediments. Environ. Sci. Technol. 44 (1), 197–203.
Dai, L., Vorselen, D., Korolev, K.S., Gore, J., 2012. Generic indicators for loss of resilience before a tipping point leading to population collapse. Science 336 (6085), 1175–1177.
Dakos, V., Carpenter, S.R., Brock, W.A., Ellison, A.M., Guttal, V., Ives, A.R., Kefi, S., Livina, V., Seekell, D.A., van Nes, E.H., Scheffer, M., 2012a. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. PloS One 7 (7), e41010.
Dakos, V., Van Nes, E.H., D'Odorico, P., Scheffer, M., 2012b. Robustness of variance and autocorrelation as indicators of critical slowing down. Ecology 93 (2), 264–271.
Dame, J.K., Christian, R.R., 2008. Evaluation of ecological network analysis: validation of output. Ecol. Model. 210 (3), 327–338.
de Mora, L., Butenschön, M., Allen, J.I., 2013. How should sparse marine in situ measurements be compared to a continuous model: an example. Geosci. Model Dev. (GMD) 6, 533–548.
de Mora, L., Butenschön, M., Allen, J.I., 2016. The assessment of a global marine ecosystem model on the basis of emergent properties and ecosystem function: a case study with ERSEM. Geosci. Model Dev. 9, 59–76.
DeStasio, B., Joice, A., Prescott, K., Gal, G., Hamilton, D., Rudstam, L., Mills, E., Rudstam, L., Jackson, J., Stewart, D., 2015. Interactions between water clarity and climate warming on hydrodynamics of Oneida Lake: applications of a dynamic reservoir model. In: Oneida Lake: Long-Term Dynamics of a Managed Ecosystem and its Fisheries. American Fisheries Society, Bethesda, Maryland.
Deehr, R.A., Luczkovich, J.J., Hart, K.J., Clough, L.M., Johnson, B.J., Johnson, J.C., 2014. Using stable isotope analysis to validate effective trophic levels from Ecopath models of areas closed and open to shrimp trawling in Core Sound, NC, USA. Ecol. Model. 282, 1–17.
Devred, E., Sathyendranath, S., Stuart, V., Platt, T., 2011. A three component classification of phytoplankton absorption spectra: applications to ocean colour data. Remote Sens. Environ. 115 (9), 2255–2266.
Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: challenges and benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172.
Dietzal, A., Reichert, P., 2012. Calibration of computationally demanding and structurally uncertain models with an application to a lake water quality model. Environ. Model. Software 38, 129–146.

Dietzel, A., Mieleitner, J., Kardaetz, S., Reichert, P., 2013. Effects of changes in the driving forces on water quality and plankton dynamics in three Swiss lakes–long-term simulations with BELAMO. Freshw. Biol. 58 (1), 10–35.
Dijkstra, J.T., Uittenbogaard, R.E., 2010. Modeling the interaction between flow and highly flexible aquatic vegetation. Water Resour. Res. 46 (12), W12547.
Dissanayake, P., Hofmann, H., Peeters, F., 2019. Comparison of results from two 3D hydrodynamic models with field data: internal seiches and horizontal currents. Inland Waters 1–22.
Doney, S.C., Lima, I., Moore, J.K., Lindsay, K., Behrenfeld, M.J., Westberry, T.K., Mahowald, N., Glover, D.M., Takahashi, T., 2009. Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data. J. Mar. Syst. 76 (1–2), 95–112.
Dugdale, R.C., Wilkerson, F.P., 1986. The use of 15N to measure nitrogen uptake in eutrophic oceans; experimental considerations. Limnol. Oceanogr. 31 (4), 673–689.
Elliott, J.A., Irish, A.E., Reynolds, C.S., Tett, P., 2000. Modelling freshwater phytoplankton communities; an exercise in validation. Ecol. Model. 128, 19–26.
Elliott, J.A., Jones, I.D., Thackeray, S.J., 2006. Testing the sensitivity of phytoplankton communities to changes in water temperature and nutrient load, in a temperate lake. Hydrobiologia 559, 401–411.
Elliott, J.A., Thackeray, S.J., Huntingford, C., Jones, R.G., 2005. Combining a Regional Climate Model with a phytoplankton community model to predict future changes in phytoplankton in lakes. Freshw. Biol. 50, 1404–1411.
Eyre, B., Balls, P., 1999. A comparative study of nutrient behavior along the salinity gradient of tropical and temperate estuaries. Estuaries 22 (2), 313–326.
Flynn, K.J., 2005. Castles built on sand: dysfunctionality in plankton models and the inadequacy of dialogue between biologists and modellers. J. Plankton Res. 27 (12), 1205–1210.
Fujii, M., Boss, E., Chai, F., 2007. The value of adding optics to ecosystem models: a case study. Biogeosciences 4, 817–835.
Fulton, E.A., Smith, A.D., Johnson, C.R., 2004. Biogeochemical marine ecosystem models I: IGBEM—a model of marine bay ecosystems. Ecol. Model. 174 (3), 267–307.
Fulton, E.A., 2010. Approaches to end-to-end ecosystem models. J. Mar. Syst. 81 (1), 171–183.
Franks, P.J., 2009. Planktonic ecosystem models: perplexing parameterizations and a failure to fail. J. Plankton Res. 31 (11), 1299–1306.
Frassl, M.A., Abell, J.M., Botelho, D.A., Cinque, K., Gibbes, B.R., Jöhnk, K.D., Muraoka, K., Robson, B.J., Wolski, M., Xiao, M., Hamilton, D.P., 2019. A short review of contemporary developments in aquatic ecosystem modelling of lakes and reservoirs. Environ. Model. Software 117, 181–187.
Frassl, M., Boehrer, B., Holtermann, P., Hu, W., Klingbeil, K., Peng, Z., Zhu, J., Rinke, K., 2018. Opportunities and limits of using meteorological reanalysis data for simulating seasonal to sub-daily water temperature dynamics in a large shallow lake. Water 10 (5), 594.
Frassl, M.A., Rothhaupt, K.O., Rinke, K., 2014. Algal internal nutrient stores feedback on vertical phosphorus distribution in large lakes. J. Great Lake. Res. 40, 162–172.
Gaedke, U., Hochstädter, S., Straile, D., 2002. Interplay between energy limitation and nutritional deficiency: empirical data and food web models. Ecol. Monogr. 72 (2), 251–270.
Gal, G., Hipsey, M.R., Paparov, A., Makler, V., Zohary, T., 2009. Implementation of ecological modelling as an effective management and investigation tool. Ecol. Model. 220, 1697–1718.
Gal, G., Rudstam, L.G., Johannsson, O.E., 2004. Predicting Mysis relicta vertical distribution in Lake Ontario. Arch. Hydrobiol. 159, 1–23.
Gantzer, P.A., Bryant, L.D., Little, J.C., 2009. Effect of hypolimnetic oxygenation on oxygen depletion rates in two water-supply reservoirs. Water Res. 43, 1700–1710.
Gentleman, W., 2002. A chronology of plankton dynamics in silico: how computer models have been used to study marine ecosystems. Hydrobiologia 480 (1–3), 69–85.
Glibert, P.M., Allen, J.I., Bouwman, A.F., Brown, C.W., Flynn, K.J., Lewitus, A.J., Madden, C.J., 2010. Modeling of HABs and eutrophication: status, advances, challenges. J. Mar. Syst. 83 (3–4), 262–275.
Goebel, N., Edwards, C.A., Zehr, J.P., Follows, M.J., 2010. An emergent community ecosystem model applied to the California Current System. J. Marine Sys. 83, 221–241.
Grangeré, K., Lefebvre, S., Ménesguen, A., Jouenne, F., 2009a. On the interest of using field primary production data to calibrate phytoplankton rate processes in ecosystem models. Estuar. Coast Shelf Sci. 81, 169–178.
Grangeré, K., Ménesguen, A., Lefebvre, S., Bacher, C., Pouvreau, S., 2009b. Modelling the influence of environmental factors on the physiological status of the Pacific oyster Crassostrea gigas in an estuarine embayment; the Baie des Veys (France). J. Sea Res. 62, 147–158.
Grangeré, K., Lefebvre, S., Bacher, C., Cugier, P., Ménesguen, A., 2010. Modelling the spatial heterogeneity of ecological processes in an intertidal estuarine bay: dynamic interactions between bivalves and phytoplankton. Mar. Ecol. Prog. Ser. 415, 141–158.
Greve, T.M., Krause-Jensen, D., 2005. Predictive modelling of eelgrass (Zostera marina) depth limits. Mar. Biol. 146 (5), 849–858.
Guillaud, J.F., Andrieux, F., Menesguen, A., 2000. Biogeochemical modelling in the Bay of Seine (France): an improvement by introducing phosphorus in nutrient cycles. J. Mar. Syst. 25 (3–4), 369–386.
Hamilton, D.P., Carey, C.C., Arvola, L., Arzberger, P., Brewer, C., Cole, J.J., Gaiser, E., Hanson, P.C., Ibelings, B.W., et al., 2015. A Global Lake Ecological Observatory Network (GLEON) for synthesising high-frequency sensor data for validation of deterministic ecological models. Inland Waters 5 (1), 49–56. https://doi.org/10.5268/IW-5.1.566.
Han, H.J., Los, F.J., Burger, D.F., Lu, X.X., 2016. A modelling approach to determine systematic nitrogen transformations in a tropical reservoir. Ecol. Eng. 94, 37–49.
Hanson, P.C., Carpenter, S.R., Kimura, N., Wu, C., Cornelius, S.P., Kratz, T.K., 2008. Evaluation of metabolism models for free-water dissolved oxygen methods in lakes. Limnol Oceanogr. Methods 6 (9), 454–465.
Harfoot, M.B., Newbold, T., Tittensor, D.P., Emmott, S., Hutton, J., Lyutsarev, V., Smith, M.J., Scharlemann, J.P., Purves, D.W., 2014. Emergent global patterns of ecosystem structure and function from a mechanistic general ecosystem model. PLoS Biol. 12 (4), e1001841.
Harmel, R.D., Smith, P.K., Migliaccio, K.W., Chaubey, I., Douglas-Mankin, K.R., Benham, B., Shukla, S., Muñoz-Carpena, R., Robson, B.J., 2014. Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: a review and recommendations. Environ. Model. Software 57, 40–51.
Hearn, C.J., Robson, B.J., 2000. Modelling a bottom diurnal boundary layer and its control of massive alga blooms in an estuary. Appl. Math. Model. 24 (11), 843–859.
Hellweger, F.L., 2017. 75 years since Monod: it is time to increase the complexity of our predictive ecosystem models (opinion). Ecol. Model. 346, 77–87.
Hetland, R.D., DiMarco, S.F., 2008. How does the character of oxygen demand control the structure of hypoxia on the Texas–Louisiana continental shelf? J. Mar. Syst. 70 (1–2), 49–62.
Hetland, R.D., DiMarco, S.F., 2012. Skill assessment of a hydrodynamic model of circulation over the Texas–Louisiana continental shelf. Ocean Model. 43, 64–76.
Heymans, J.J., Coll, M., Link, J.S., Mackinson, S., Steenbeek, J., Walters, C., Christensen, V., 2016. Best practice in Ecopath with Ecosim food-web models for ecosystem-based management. Ecol. Model. 331, 173–184.
Higgins, S.N., Hecky, R.E., Guildford, S.J., 2005. Modeling the growth, biomass, and tissue phosphorus concentration of Cladophora glomerata in eastern Lake Erie: model description and field testing. J. Great Lake. Res. 31 (4), 439–455.
Hillmer, I., van Reenen, P., Imberger, J., Zohary, T., 2008. Phytoplankton patchiness and their role in the modelled productivity of a large, seasonally stratified lake. Ecol. Model. 218 (1–2), 49–59.
Hipsey, M.R., Antenucci, J.P., Brookes, J.D., Burch, M.D., Regel, R.H., Linden, L., 2004. A three dimensional model of Cryptosporidium dynamics in lakes and reservoirs: a new tool for risk management. Int. J. River Basin Manag. 2 (3), 181–197.
Hipsey, M.R., Brookes, J.D., Regel, R.H., Antenucci, J.P., Burch, M.D., 2006. In situ evidence for the association of total coliforms and Escherichia coli with suspended inorganic particles in an Australian reservoir. Water Air Soil Pollut. 170 (1–4), 191–209.
Hipsey, M.R., Antenucci, J.P., Brookes, J.D., 2008. A generic, process-based model of microbial pollution in aquatic systems. Water Resour. Res. 44 (7), W07408.
Hipsey, M.R., Salmon, S.U., Mosley, L.M., 2014. A three-dimensional hydro-geochemical model to assess lake acidification risk. Environ. Model. Software 61, 433–457.
Hipsey, M.R., Hamilton, D.P., Hanson, P.C., Carey, C.C., Coletti, J.Z., Read, J.S., Ibelings, B.W., Valesini, F.J., Brookes, J.D., 2015. Predicting the resilience and recovery of aquatic systems: a framework for model evolution within environmental observatories. Water Resour. Res. 51 (9), 7023–7043.
Hipsey, M.R., Bruce, L.C., Boon, C., Busch, B., Carey, C.C., Hamilton, D.P., Hanson, P.C., Read, J.S., De Sousa, E., Weber, M., Winslow, L.A., 2019. A General Lake model (GLM 3.0) for linking with high-frequency sensor data from the global lake ecological observatory network (GLEON). Geosci. Model Dev. (GMD) 12 (1), 473–523.
Hirata, T., Hardman-Mountford, N.J., Brewin, R.J.W., Aiken, J., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., Yamanaka, Y., 2011. Synoptic relationships between surface Chlorophyll-a and diagnostic pigments specific to phytoplankton functional types. Biogeosciences 8 (2), 311–327.
Hodges, B.R., Imberger, J., Saggio, A., Winters, K.B., 2000. Modeling basin-scale internal waves in a stratified lake. Limnol. Oceanogr. 45 (7), 1603–1620.
Hölker, F., Breckling, B., 2005. A spatiotemporal individual-based fish model to investigate emergent properties at the organismal and the population level. Ecol. Model. 186 (4), 406–426.
Holt, J., Allen, J.I., Anderson, T.R., Brewin, R., Butenschön, M., Harle, J., Huse, G., Lehodey, P., Lindemann, C., Memery, L., Salihoglu, B., 2014. Challenges in integrative approaches to modelling the marine ecosystems of the North Atlantic: physics to fish and coasts to ocean. Prog. Oceanogr. 129, 285–313.
Hong, B., Strawderman, R.L., Swaney, D.P., Weinstein, D.A., 2005. Bayesian estimation of input parameters of a nitrogen cycle model applied to a forested reference watershed, Hubbard Brook Watershed Six. Water Resour. Res. 41 (3).
Hood, R.R., Coles, V.J., Capone, D.G., 2004. Modeling the distribution of trichodesmium and nitrogen fixation in the Atlantic Ocean. J. Geophys. Res.: Oceans 109 (C6).
Huang, P., Kilminster, K., Larsen, S., Hipsey, M.R., 2018. Assessing artificial oxygenation in a riverine salt-wedge estuary with a three-dimensional finite-volume model. Ecol. Eng. 118, 111–125.
Huang, P., Trayler, K., Wang, B., Saeed, A., Oldham, C.E., Busch, B., Hipsey, M.R., 2019. An integrated modelling system for water quality forecasting in an urban eutrophic estuary: the Swan-Canning Estuary Virtual Observatory. J. Mar. Syst. 199, 103218.
Janse, J.H., De Senerpont Domis, L.N., Scheffer, M., Lijklema, L., Van Liere, L., Klinge, M., Mooij, W.M., 2008. Critical phosphorus loading of different types of shallow lakes and the consequences for management estimated with the ecosystem model PCLake. Limnologica 38, 203–219.
Janse, J.H., Scheffer, M., Lijklema, L., Van Liere, L., Sloot, J.S., Mooij, W.M., 2010. Estimating the critical phosphorus loading of shallow lakes with the ecosystem model PCLake: sensitivity, calibration and uncertainty. Ecol. Model. 221, 654–665.
Janssen, A.B., Arhonditsis, G.B., Beusen, A., Bolding, K., Bruce, L., Bruggeman, J., Couture, R.M., Downing, A.S., Elliott, J.A., Frassl, M.A., Gal, G., 2015. Exploring,

exploiting and evolving diversity of aquatic ecosystem models: a community perspective. Aquat. Ecol. 49 (4), 513–548.
Ji, Z.G., 2017. Hydrodynamics and Water Quality: Modeling Rivers, Lakes, and Estuaries. John Wiley & Sons.
Jiang, L., Xia, M., 2018. Modeling investigation of the nutrient and phytoplankton variability in the Chesapeake Bay outflow plume. Prog. Oceanogr. 162, 290–302.
Johnson, K.S., Needoba, J.A., 2008. Mapping the spatial variability of plankton metabolism using nitrate and oxygen sensors on an autonomous underwater vehicle. Limnol. Oceanogr. 53, 2237–2250.
Jolliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A., Helber, R., Arnone, R. A., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. J. Mar. Syst. 76 (1–2), 64–82.
Jones, R.A., Lee, G.F., 1988. Use of Vollenweider-OECD modeling to evaluate aquatic ecosystem functioning. In: Cairns Jr., J., Pratt, J.R. (Eds.), Functional Testing of Aquatic Biota for Estimating Hazards of Chemicals, ASTM STP 988. American Society for Testing and Materials, Philadelphia, pp. 17–27.
Jones, E.M., Baird, M.E., Mongin, M., Parslow, J., Skerratt, J., Lovell, J., Margvelashvili, N., Matear, R.J., Wild-Allen, K., Robson, B., Rizwi, F., 2016. Use of remote-sensing reflectance to constrain a data assimilating marine biogeochemical model of the Great Barrier Reef. Biogeosciences 13 (23), 6441.
Jouini, M., Lévy, M., Crépon, M., Thiria, S., 2013. Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method. Remote Sens. Environ. 131, 232–246.
Kara, E.L., Hanson, P., Hamilton, D., Hipsey, M.R., McMahon, K.D., Read, J.S., Winslow, L., Dedrick, J., Rose, K., Carey, C.C., Bertilsson, S., 2012. Time-scale dependence in numerical simulations: assessment of physical, chemical, and biological predictions in a stratified lake at temporal scales of hours to months. Environ. Model. Software 35, 104–121.
Kim, D.K., Zhang, W., Watson, S., Arhonditsis, G.B., 2014. A commentary on the modelling of the causal linkages among nutrient loading, harmful algal blooms, and hypoxia patterns in Lake Erie. J. Great Lake. Res. 40, 117–129.
Kong, X., He, W., Qin, N., Liu, W., Yang, B., Yang, C., Xu, F., Mooij, W.M., Koelmans, A. A., 2017. Integrated ecological and chemical food web accumulation modeling explains PAH temporal trends during regime shifts in a shallow lake. Water Res. 119, 73–82.
Kromkamp, J., Walsby, A.E., 1990. A computer model of buoyancy and vertical migration in cyanobacteria. J. Plankton Res. 12 (1), 161–183.
Kubicek, A., Jopp, F., Breckling, B., Lange, C., Reuter, H., 2015. Context-oriented model validation of individual-based models in ecology: a hierarchically structured approach to validate qualitative, compositional and quantitative characteristics. Ecol. Complex. 22, 178–191.
Kuhnert, P., 2014. Physical-statistical modelling. Environmetrics 25, 201–202.
Kuiper, J.J., Van Altena, C., De Ruiter, P.C., Van Gerven, L.P., Janse, J.H., Mooij, W.M., 2015. Food-web stability signals critical transitions in temperate shallow lakes. Nat. Commun. 6, 7727.
Kwiatkowski, L., Yool, A., Allen, J.I., Anderson, T.R., Barciela, R., Buitenhuis, E.T., Butenschön, M., Enright, C., Halloran, P.R., Le Quéré, C., De Mora, L., 2014. iMarNet: an ocean biogeochemistry model intercomparison project within a common physical ocean modelling framework. Biogeosciences 11, 7291–7304.
Le Goff, C., Lavaud, R., Cugier, P., Jean, F., Flye-Sainte-Marie, J., Foucher, E., Desroy, N., Fifas, S., Foveau, A., 2017. A coupled biophysical model for the distribution of the great scallop Pecten maximus in the English Channel. J. Mar. Syst. 167, 55–67.
Lehuta, S., Petitgas, P., Mahévas, S., Huret, M., Vermard, Y., Uriarte, A., Record, N.R., 2013. Selection and validation of a complex fishery model using an uncertainty hierarchy. Fish. Res. 143, 57–66.
Li, H., Mynett, A., Penning, E., Qi, H., 2010. Revealing spatial pattern dynamics in aquatic ecosystem modelling with Multi-Agent Systems in Lake Veluwe. Ecol. Inf. 5 (2), 97–107.
Li, M., Lee, Y.J., Testa, J.M., Li, Y., Ni, W., Kemp, W.M., Di Toro, D.M., 2016. What drives interannual variability of hypoxia in Chesapeake Bay: climate forcing versus nutrient loading? Geophys. Res. Lett. 43 (5), 2127–2134.
Li, Y., Waite, A.M., Gal, G., Hipsey, M.R., 2013. An analysis of the relationship between phytoplankton internal stoichiometry and water column N: P ratios in a dynamic lake environment. Ecol. Model. 252, 196–213.
Li, Y., Gal, G., Makler-Pick, V., Waite, A.M., Bruce, L.C., Hipsey, M.R., 2014. Examination of the role of the microbial loop in regulating lake nutrient stoichiometry and phytoplankton dynamics. Biogeosciences 11 (11), 2939–2960.
Lignell, R., Haario, H., Laine, M., Thingstad, T.F., 2013. Getting the "right" parameter values for models of the pelagic microbial food web. Limnol. Oceanogr. 58 (1), 301–313.
Link, J.S., 2010. Adding rigor to ecological network models by evaluating a set of pre-balance diagnostics: a plea for PREBAL. Ecol. Model. 221 (12), 1580–1591.
Lovato, T., Ciavatta, S., Brigolin, D., Rubino, A., Pastres, R., 2013. Modelling dissolved oxygen and benthic algae dynamics in a coastal ecosystem by exploiting real-time monitoring data. Estuar. Coast Shelf Sci. 119, 17–30.
Makler-Pick, V., Gal, G., Shapiro, J., Hipsey, M.R., 2011. Exploring the role of fish in a lake ecosystem (Lake Kinneret, Israel) by coupling an individual-based fish population model to a dynamic ecosystem model. Can. J. Fish. Aquat. Sci. 68 (7), 1265–1284.
Makler-Pick, V., Hipsey, M.R., Zohary, T., Carmel, Y., Gal, G., 2017. Intraguild predation dynamics in a lake ecosystem based on a coupled hydrodynamic-ecological model: the example of lake Kinneret (Israel). Biology 6 (2), 22.
Margvelashvili, N., Saint-Cast, F., Condie, S., 2008. Numerical modelling of the suspended sediment transport in Torres Strait. Continent. Shelf Res. 28 (16), 2241–2256.
Margvelashvili, N., Andrewartha, J., Herzfeld, M., Robson, B.J., Brando, V.E., 2013. Satellite data assimilation and estimation of a 3D coastal sediment transport model using error-subspace emulators. Environ. Model. Software 40, 191–201.
Margvelashvili, N.Y., Herzfeld, M., Rizwi, F., Mongin, M., Baird, M.E., Jones, E., Schaffelke, B., King, E., Schroeder, T., 2016. Emulator-assisted data assimilation in complex models. Ocean Dynam. 66 (9), 1109–1124.
Margvelashvili, N., Andrewartha, J., Baird, M., Herzfeld, M., Jones, E., Mongin, M., Rizwi, F., Robson, B.J., Skerratt, J., Wild-Allen, K., Steven, A., 2018. Simulated fate of catchment-derived sediment on the Great Barrier Reef shelf. Mar. Pollut. Bull. 135, 954–962.
Mark, J.B., John, W.B., Scott, W.N., James, N.K., 2002. Modeling phytoplankton production: problems with the Eppley curve and an empirical alternative. Mar. Ecol. Prog. Ser. 238, 31–45.
Martin, J.H., Knauer, G.A., Karl, D.M., Broenkow, W.W., 1987. VERTEX: carbon cycling in the northeast Pacific. Deep-Sea Res. Part A Oceanogr. Res. Pap. 34 (2), 267–285.
Martiny, A.C., Vrugt, J.A., Primeau, F.W., Lomas, M.W., 2013. Regional variation in the particulate organic carbon to nitrogen ratio in the surface ocean. Global Biogeochem. Cycles 27 (3), 723–731.
Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecol. Model. 68 (1–2), 21–32.
Megrey, B.A., Rose, K.A., Klumb, R.A., Hay, D.E., Werner, F.E., Eslinger, D.L., Smith, S.L., 2007. A bioenergetics-based population dynamics model of Pacific herring (Clupea harengus pallasi) coupled to a lower trophic level nutrient–phytoplankton–zooplankton model: description, calibration, and sensitivity analysis. Ecol. Model. 202 (1–2), 144–164.
Ménesguen, A., Cugier, P., Loyer, S., Vanhoutte-Brunier, A., Hoch, T., Guillaud, J.F., Gohin, F., 2007. Two-or three-layered box-models versus fine 3D models for coastal ecological modelling? A comparative study in the English Channel (Western Europe). J. Mar. Syst. 64 (1–4), 47–65.
Ménesguen, A., Dussauze, M., Dumas, F., Thouvenin, B., Garnier, V., Lecornu, F., Répécaud, M., 2019. Ecological model of the Bay of Biscay and English Channel shelf for environmental status assessment part 1: nutrients, phytoplankton and oxygen. Ocean Model. 133, 56–78.
Mieleitner, J., Reichert, P., 2008. Modelling functional groups of phytoplankton in three lakes of different trophic state. Ecol. Model. 211 (3–4), 279–291.
Miller, R.L., Liu, C.C., Buonassissi, C.J., Wu, A.M., 2011. A multi-sensor approach to examining the distribution of total suspended matter (TSM) in the Albemarle-Pamlico estuarine system, NC, USA. Rem. Sens. 3 (5), 962–974.
Missaghi, S., Hondzo, M., 2010. Evaluation and application of a three-dimensional water quality model in a shallow lake with complex morphometry. Ecol. Model. 221 (11), 1512–1525.
Mitra, A., Flynn, K.J., 2006. Accounting for variation in prey selectivity by zooplankton. Ecol. Model. 199 (1), 82–92.
Mooij, W.M., Trolle, D., Jeppesen, E., Arhonditsis, G., Belolipetsky, P.V., Chitamwebwa, D.B.R., Degermendzhy, A.G., DeAngelis, D.L., De Senerpont Domis, L.N., Downing, A.S., Elliott, A.E., Fragoso Jr., C.R., Gaedke, U., Genova, S.N., Gulati, R.D., Håkanson, L., Hamilton, D.P., Hipsey, M.R., Hoen, J., Hülsmann, S., Los, F.J., Makler-Pick, V., Petzoldt, T., Prokopkin, I.G., Rinke, K., Schep, S.A., Tominaga, K., Van Dam, A.A., Van Nes, E.H., Wells, S.A., Janse, J.H., 2010. Challenges and opportunities for integrating lake ecosystem modelling approaches. Aquat. Ecol. 44 (3), 633–667.
Mooij, W.M., Brederveld, B., DeAngelis, D., Downing, A., Faber, A., Gerla, D., van Gerven, L., Hipsey, M.R., Hoen, J., Janse, J.H., Janssen, A.B.G., Jeuken, M., de Klein, J., Kooi, B., Lischke, B., Postma, L., Petzoldt, T., Schep, S., Thiange, C., Trolle, D., Kuiper, J., 2014. Serving many at once: how a database approach can create unity in dynamical ecosystem modelling. Environ. Model. Software 61, 266–273.
Moradkhani, H., Hsu, K., Hong, Y., Sorooshian, S., 2006. Investigating the impact of remotely sensed precipitation and hydrologic model uncertainties on the ensemble streamflow forecasting. Geophys. Res. Lett. 33 (12).
Morozov, A.Y., 2010. Emergence of Holling type III zooplankton functional response: bringing together field evidence and mathematical modelling. J. Theor. Biol. 265 (1), 45–54.
Müller, F., Bergmann, M., Dannowski, R., Dippner, J.W., Gnauck, A., Haase, P., Jochimsen, M.C., Kasprzak, P., Kröncke, I., Kümmerlin, R., Küster, M., 2016. Assessing resilience in long-term ecological data sets. Ecol. Indicat. 65, 10–43.
Mulder, K., Bowden, W.B., 2007. Organismal stoichiometry and the adaptive advantage of variable nutrient use and production efficiency in Daphnia. Ecol. Model. 202 (3–4), 427–440.
Murphy, A.H., Epstein, E.S., 1989. Skill scores and correlation coefficients in model verification. Mon. Weather Rev. 117 (3), 572–582.
Nakayama, K., Nakagawa, Y., Nakanishi, Y., Kuwae, T., Watanabe, K., Moki, H., Komai, K., Tada, K., Tsai, J.W., Hipsey, M.R., 2019. A Dynamically Coupled Hydrodynamic-Submerged Aquatic Vegetation Model for Aquatic Systems. Submitted to Water Resources Research.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. 10 (3), 282–290.
Neumann, T., Schernewski, G., 2008. Eutrophication in the Baltic Sea and shifts in nitrogen fixation analyzed with a 3D ecosystem model. J. Mar. Syst. 74 (1), 592–602.
Ng, S., Antenucci, J.P., Hipsey, M.R., Tibor, G., Zohary, T., 2011. Physical controls on the spatial evolution of a dinoflagellate bloom in a large lake. Limnol. Oceanogr. 56, 2265–2281.
Nilsen, M., Pedersen, T., Nilssen, E.M., Fredriksen, S., 2008. Trophic studies in a high-latitude fjord ecosystem—a comparison of stable isotope analyses (δ13C and δ15N) and trophic-level estimates from a mass-balance model. Can. J. Fish. Aquat. Sci. 65 (12), 2791–2806.

Nordstrom, K.D., 2012. Models, validation, and applied geochemistry: issues in science, communication, and philosophy. Appl. Geochem. 27, 1899–1919.
Nussboim, S., Rimmer, A., Lechinsky, Y., Gutman, P.O., Broday, D., 2017. Improving the estimation of Lake Kinneret's heat balance and surface fluxes using the Kalman Filter algorithm. Limnol Oceanogr. Methods 15 (5), 467–479.
O'Brien, K.R., Waycott, M., Maxwell, P., Kendrick, G.A., Udy, J.W., Ferguson, A.J., Kilminster, K., Scanes, P., McKenzie, L.J., McMahon, K., Adams, M.P., 2018. Seagrass ecosystem trajectory depends on the relative timescales of resistance, recovery and disturbance. Mar. Pollut. Bull. 134, 166–176.
Oliver, R.L., Hamilton, D.P., Brookes, J.D., Ganf, G.G., 2012. Physiology, blooms and prediction of planktonic cyanobacteria. In: Ecology of Cyanobacteria II. Springer, Dordrecht, pp. 155–194.
Omlin, M., Reichert, P., 1999. A comparison of techniques for the estimation of model prediction uncertainty. Ecol. Model. 115 (1), 45–59.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263 (5147), 641–646.
Ostrovsky, I., Yacobi, Y.Z., 2010. Sedimentation flux in a large subtropical lake: spatiotemporal variations and relation to primary productivity. Limnol. Oceanogr. 55 (5), 1918–1931.
Paraska, D.W., Hipsey, M.R., Salmon, S.U., 2014. Sediment diagenesis models: review of approaches, challenges and opportunities. Environ. Model. Software 61, 297–325.
Parparov, A., Gal, G., 2012. Assessment and implementation of a methodological framework for sustainable management: Lake Kinneret as a case study. J. Environ. Manag. 101, 111–117.
Parslow, J., Cressie, N., Campbell, E.P., Jones, E., Murray, L., 2013. Bayesian learning and predictability in a stochastic nonlinear dynamical model. Ecol. Appl. 23 (4), 679–698.
Peeters, F., Straile, D., Lorke, A., Livingstone, D.M., 2007. Earlier onset of the spring phytoplankton bloom in lakes of the temperate zone in a warmer climate. Global Change Biol. 13 (9), 1898–1909.
Perhar, G., Arhonditsis, G.B., Brett, M.T., 2012. Modelling the role of highly unsaturated fatty acids in planktonic food web processes: a mechanistic approach. Environ. Rev. 20 (3), 155–172.
Pohjola, M.V., Pohjola, P., Tainio, M., Tuomisto, J.T., 2013. Perspectives to performance of environment and health assessments and models—from outputs to outcomes? Int. J. Environ. Res. Publ. Health 10 (7), 2621–2642.
Popendorf, K.J., Duhamel, S., 2015. Variable phosphorus uptake rates and allocation across microbial groups in the oligotrophic Gulf of Mexico. Environ. Microbiol. 17 (10), 3992–4006.
Post, D.M., 2002. Using stable isotopes to estimate trophic position: models, methods, and assumptions. Ecology 83 (3), 703–718.
Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68 (1–2), 33–50.
Quere, C.L., Harrison, S.P., Colin Prentice, I., Buitenhuis, E.T., Aumont, O., Bopp, L., Claustre, H., Cotrim Da Cunha, L., Geider, R., Giraud, X., Klaas, C., 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biol. 11 (11), 2016–2040.
Raick, C., Soetaert, K., Grégoire, M., 2006. Model complexity and performance: how far can we simplify? Prog. Oceanogr. 70 (1), 27–57.
Ramin, M., Arhonditsis, G.B., 2013. Bayesian calibration of mathematical models: optimization of model structure and examination of the role of process error covariance. Ecol. Inf. 18, 107–116.
Ramin, M., Labencki, T., Boyd, D., Trolle, D., Arhonditsis, G.B., 2012. A Bayesian synthesis of predictions from different models for setting water quality criteria. Ecol. Model. 242, 127–145.
Reckhow, K.H., Chapra, S.C., 1983. Engineering Approaches for Lake Management. Vol. 1, Data Analysis and Empirical Modelling. Butterworth Publishers.
Recknagel, F., Ostrovsky, I., Cao, H., 2014. Model ensemble for the simulation of plankton community dynamics of Lake Kinneret (Israel) induced from in situ predictor variables by evolutionary computation. Environ. Model. Software 61, 380–392.
Reed, G., Keeley, R., Belov, S., Mikhailov, N., 2010. Ocean Data Portal: a standards approach to data access and dissemination. Proc. Asia Ocean 21–25.
Renton, M., Airey, M., Cambridge, M.L., Kendrick, G.A., 2011. Modelling seagrass growth and development to evaluate transplanting strategies for restoration. Ann. Bot. 108 (6), 1213–1223.
Reynolds, C.S., 2006. The Ecology of Phytoplankton. Cambridge University Press.
Reynolds, C.S., Elliott, J.A., 2012. Complexity and emergent properties in aquatic ecosystems: predictability of ecosystem responses. Freshw. Biol. 57, 74–90.
Rigosi, A., Fleenor, W., Rueda, F., 2010. State-of-the-art and recent progress in phytoplankton succession modelling. Environ. Rev. 18, 423–440.
Rigosi, A., Marcé, R., Escot, C., Rueda, F.J., 2011. A calibration strategy for dynamic succession models including several phytoplankton groups. Environ. Model. Software 26 (6), 697–710.
Rimmer, A., Samuels, R., Lechinsky, Y., 2009. A comprehensive study across methods and time scales to estimate surface fluxes from Lake Kinneret, Israel. J. Hydrol. 379 (1–2), 181–192.
Robson, B.J., Hamilton, D.P., Webster, I.T., Chan, T., 2008. Ten steps applied to development and evaluation of process-based biogeochemical models of estuaries. Environ. Model. Software 23, 369–384.
Robson, B.J., 2010. A dynamic model of primary production and plant coverage in an oligotrophic tropical river. In: Swayne, D.A., Yang, W., Voinov, A.A., Rizzoli, A., Filatova, T. (Eds.), 2010 International Congress on Environmental Modelling and Software: Modelling for Environment's Sake. International Environmental Modelling and Software Society (iEMSs), Ottawa, Canada.
Robson, B.J., 2014a. When do aquatic systems models provide useful predictions, what is changing, and what is next? Environ. Model. Software 61, 287–296.
Robson, B.J., 2014b. State of the art in modelling of phosphorus in aquatic systems: review, criticisms and commentary. Environ. Model. Software 61, 339–359.
Robson, B.J., Dourdet, V., 2015. Prediction of sediment, particulate nutrient and dissolved nutrient concentrations in a dry tropical river to provide input to a mechanistic coastal water quality model. Environ. Model. Software 63, 97–108.
Robson, B.J., Andrewartha, J., Baird, M.E., Herzfeld, M., Jones, E.M., Margvelashvili, N., Mongin, M., Rizwi, F., Skerratt, J., Wild-Allen, K., 2017. Evaluating the eReefs Great Barrier Reef marine model against observed emergent properties. In: Syme, G., Hatton MacDonald, D., Fulton, B., Piantadosi, J. (Eds.), MODSIM2017, 22nd International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp. 1976–1982.
Robson, B.J., Arhonditsis, G.B., Baird, M.E., Brebion, J., Edwards, K.F., Geoffroy, L., Hébert, M.P., van Dongen-Vogels, V., Jones, E.M., Kruk, C., Mongin, M., 2018. Towards evidence-based parameter values and priors for aquatic ecosystem modelling. Environ. Model. Software 100, 74–81.
Rocha, C., Edwards, C.A., Roughan, M., Cetina-Heredia, P., Kerry, C., 2019. A high-resolution biogeochemical model (ROMS 3.4+ bio_Fennel) of the East Australian Current system. Geosci. Model Dev. (GMD) 12 (1), 441–456.
Rode, M., Arhonditsis, G., Balin, D., Kebede, T., Krysanova, V., van Griensven, A., van der Zee, S., 2010. New challenges in integrated water quality modelling. Hydrol. Proced. 24, 3447–3461.
Rode, M., Wade, A.J., Cohen, M.J., Hensley, R.T., Bowes, M.J., Kirchner, J.W., Arhonditsis, G.B., Jordan, P., Kronvang, B., Halliday, S.J., Skeffington, R.A., 2016. Sensors in the stream: the high-frequency wave of the present. Environ. Sci. Technol. 50, 1910297–1910307.
Rose, K.A., Werner, F.E., Megrey, B.A., Aita, M.N., Yamanaka, Y., Hay, D.E., Schweigert, J.F., Foster, M.B., 2007. Simulated herring growth responses in the Northeastern Pacific to historic temperature and zooplankton conditions generated by the 3-dimensional NEMURO nutrient–phytoplankton–zooplankton model. Ecol. Model. 202 (1–2), 184–195.
Rose, K.A., Allen, J.I., Artioli, Y., Barange, M., Blackford, J., Carlotti, F., Cropp, R., Daewel, U., Edwards, K., Flynn, K., Hill, S.L., 2010. End-to-end models for the analysis of marine ecosystems: challenges, issues, and next steps. Mar. Coast. Fish 2 (1), 115–130.
Rueda, F.J., MacIntyre, S., 2010. Modelling the fate and transport of negatively buoyant storm–river water in small multi-basin lakes. Environ. Model. Software 25 (1), 146–157.
Saba, V.S., Friedrichs, M.A., Carr, M.E., Antoine, D., Armstrong, R.A., Asanuma, I., Aumont, O., Bates, N.R., Behrenfeld, M.J., Bennington, V., Bopp, L., 2010. Challenges of modeling depth-integrated marine primary productivity over multiple decades: a case study at BATS and HOT. Global Biogeochem. Cycles 24 (3).
Sailley, S.F., Vogt, M., Doney, S.C., Aita, M.N., Bopp, L., Buitenhuis, E.T., Hashioka, T., Lima, I., Le Quéré, C., Yamanaka, Y., 2013. Comparing food web structures and dynamics across a suite of global marine ecosystem models. Ecol. Model. 261, 43–57.
Salihoglu, B., Neuer, S., Painting, S., Murtugudde, R., Hofmann, E.E., Steele, J.H., Hood, R.R., Legendre, L., Lomas, M.W., Wiggert, J.D., Ito, S., 2013. Bridging marine ecosystem and biogeochemistry research: lessons and recommendations from comparative studies. J. Mar. Syst. 109, 161–175.
Salmon, S.U., Hipsey, M.R., Wake, G.W., Ivey, G.N., Oldham, C.E., 2017. Quantifying lake water quality evolution: coupled geochemistry, hydrodynamics, and aquatic ecology in an acidic pit lake. Environ. Sci. Technol. 51 (17), 9864–9875.
Sauterey, B., Ward, B.A., Follows, M.J., Bowler, C., Claessen, D., 2015. When everything is not everywhere but species evolve: an alternative method to model adaptive properties of marine ecosystems. J. Plankton Res. 37 (1), 28–47.
Saux Picart, S., Butenschön, M., Shutler, J.D., 2012. Wavelet-based spatial comparison technique for analysing and evaluating two-dimensional geophysical model fields. Geosci. Model Dev. (GMD) 5 (1), 223–230.
Savina, M., Ménesguen, A., 2008. A deterministic population dynamics model to study the distribution of a benthic bivalve with planktonic larvae (Paphia rhomboides) in the English Channel (NW Europe). J. Mar. Syst. 70 (1–2), 63–76.
Savina, M., Forrest, R.E., Fulton, E.A., Condie, S.A., 2013. Ecological effects of trawling fisheries on the eastern Australian continental shelf: a modelling study. Mar. Freshw. Res. 64 (11), 1068–1086.
Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., Held, H., Van Nes, E.H., Rietkerk, M., Sugihara, G., 2009. Early-warning signals for critical transitions. Nature 461 (7260), 53.
Schmid, M., Ostrovsky, I., McGinnis, D.F., 2017. Role of gas ebullition in the methane budget of a deep subtropical lake: what can we learn from process-based modeling? Limnol. Oceanogr. 62 (6), 2674–2698.
Segura, A.M., Kruk, C., Calliari, D., Fort, H., 2012. Use of a morphology-based functional approach to model phytoplankton community succession in a shallow subtropical lake. Freshw. Biol. 58, 504–512.
Shen, C., Testa, J.M., Li, M., Cai, W.J., Waldbusser, G.G., Ni, W., Kemp, W.M., Cornwell, J., Chen, B., Brodeur, J., Su, J., 2019. Controls on carbonate system dynamics in a coastal plain estuary: a modeling study. J. Geophys. Res.: Biogeosciences 124 (1), 61–78.
Simpson, J.C., Norris, R.H., 2000. Biological assessment of river quality: development of AUSRIVAS models and outputs. In: Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques. Proceedings of an International Workshop Held in Oxford, UK. Freshwater Biological Association (FBA), pp. 125–142 on 16-18 September 1997.
Sinha, B., Buitenhuis, E.T., Le Quéré, C., Anderson, T.R., 2010. Comparison of the emergent behavior of a complex ecosystem model in two ocean general circulation models. Prog. Oceanogr. 84 (3–4), 204–224.

Skerratt, J., Wild-Allen, K., Rizwi, F., Whitehead, J., Coughanowr, C., 2013. Use of a high resolution 3D fully coupled hydrodynamic, sediment and biogeochemical model to understand estuarine nutrient dynamics under various water quality scenarios. Ocean Coast Manag. 83, 52–66.
Snortheim, C.A., Hanson, P.C., McMahon, K.D., Read, J.S., Carey, C.C., Dugan, H.A., 2017. Meteorological drivers of hypolimnetic anoxia in a eutrophic, north temperate lake. Ecol. Model. 343, 39–53.
Sohma, A., Sekiguchi, Y., Kuwae, T., Nakamura, Y., 2008. A benthic–pelagic coupled ecosystem model to estimate the hypoxic estuary including tidal flat—model description and validation of seasonal/daily dynamics. Ecol. Model. 215 (1–3), 10–39.
Sokolova, E., Pettersson, T.J., Bergstedt, O., Hermansson, M., 2013. Hydrodynamic modelling of the microbial water quality in a drinking water source as input for risk reduction management. J. Hydrol. 497, 15–23.
Spillman, C.M., Hamilton, D.P., Hipsey, M.R., Imberger, J., 2008. A spatially resolved model of seasonal variations in phytoplankton and clam (Tapes philippinarum) biomass in Barbamarco Lagoon, Italy. Estuarine Coastal Shelf Sci. 79 (2), 187–203.
Spillman, C.M., Imberger, J., Hamilton, D.P., Hipsey, M.R., Romero, J.R., 2007. Modelling the effects of Po River discharge, internal nutrient cycling and hydrodynamics on biogeochemistry of the Northern Adriatic Sea. J. Mar. Syst. 68 (1–2), 167–200.
Sprules, W.G., Bowerman, J.E., 1988. Omnivory and food chain length in zooplankton food webs. Ecology 418–426.
Stadnyk, T.A., Delavau, C., Kouwen, N., Edwards, T.W.D., 2013. Towards hydrological model calibration and validation: simulation of stable water isotopes using the isoWATFLOOD model. Hydrol. Process. 27 (25), 3791–3810.
Steele, J.H., Aydin, K., Gifford, D.J., Hofmann, E.E., 2013. Construction kits or virtual worlds; Management applications of E2E models. J. Mar. Syst. 109, 103–108.
Steyn, D.G., Oke, T.R., 1982. The depth of the daytime mixed layer at two coastal sites: a model and its validation. Boundary-Layer Meteorol. 24 (2), 161–180.
Stow, C.A., Roessler, C., Borsuk, M.E., Bowen, J.D., Reckhow, K.H., 2003. Comparison of estuarine water quality models for total maximum daily load development in Neuse River Estuary. J. Water Resour. Plann. Manag. 129 (4), 307–314.
Stow, C.A., Reckhow, K.H., Qian, S.S., Lamon III, E.C., Arhonditsis, G.B., Borsuk, M.E., Seo, D., 2007. Approaches to evaluate water quality model parameter uncertainty for adaptive TMDL implementation 1. J. Am. Water Resour. Assoc. 43 (6), 1499–1507.
Stow, C.A., Jolliff, J., McGillicuddy Jr., D.J., Doney, S.C., Allen, J.I., Friedrichs, M.A., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. J. Mar. Syst. 76 (1–2), 4–15.
Sugimoto, R., Kasai, A., Miyajima, T., Fujita, K., 2010. Modeling phytoplankton production in Ise Bay, Japan: use of nitrogen isotopes to identify dissolved inorganic nitrogen sources. Estuar. Coast Shelf Sci. 86 (3), 450–466.
Sun, G.Q., Li, L., Jin, Z., Li, B.L., 2010. Pattern formation in a spatial plant-wrack model with tide effect on the wrack. J. Biol. Phys. 36 (2), 161–174.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106 (D7), 7183–7192.
Testa, J.M., Li, Y., Lee, Y.J., Li, M., Brady, D.C., Di Toro, D.M., Kemp, W.M., 2017. Modeling physical and biogeochemical controls on dissolved oxygen in Chesapeake Bay: lessons learned from simple and complex approaches. In: Modeling Coastal Hypoxia. Springer International Publishing, pp. 95–118.
Thingstad, T.F., Havskum, H., Zweifel, U.L., Berdalet, E., Sala, M.M., Peters, F., Alcaraz, M., Scharek, R., Perez, M., Jacquet, S., Flaten, G.A.F., 2007. Ability of a "minimum" microbial food web model to reproduce response patterns observed in mesocosms manipulated with N and P, glucose, and Si. J. Mar. Syst. 64 (1–4), 15–34.
Thomann, R.V., Fitzpatrick, J.J., 1982. Calibration and Verification of a Mathematical Model of the Eutrophication of the Potomac Estuary. DC Department of Environmental Sciences.
Tittensor, D.P., Eddy, T.D., Lotze, H.K., Galbraith, E.D., Cheung, W., Barange, M., Blanchard, J.L., Bopp, L., Bryndum-Buchholz, A., Büchner, M., Bulman, C., Carozza, D.A., Christensen, V., Coll, M., Dunne, J.P., Fernandes, J.A., Fulton, E.A., Hobday, A.J., Huber, V., Jennings, S., Jones, M., Lehodey, P., Link, J.S., Mackinson, S., Maury, O., Niiranen, S., Oliveros-Ramos, R., Roy, T., Schewe, J., Shin, Y.-J., Silva, T., Stock, C.A., Steenbeek, J., Underwood, P.J., Volkholz, J., Watson, J.R., Walker, N.D., 2018. A protocol for the intercomparison of marine fishery and ecosystem models: fish-MIP v1.0. Geosci. Model Dev. 11, 1421–1442.
Tomasky-Holmes, G., Valiela, I., Charette, M.A., 2013. Determination of water mass ages using radium isotopes as tracers: implications for phytoplankton dynamics in estuaries. Mar. Chem. 156, 18–26.
Townsend, S.A., Webster, I.T., Schult, J.H., 2011. Metabolism in a groundwater-fed river system in the Australian wet/dry tropics: tight coupling of photosynthesis and respiration. J. North Am. Benthol. Soc. 30 (3), 603–620.
Trolle, D., Hamilton, D.P., Pilditch, C.A., Duggan, I.C., Jeppesen, E., 2011. Predicting the effects of climate change on trophic status of three morphologically varying lakes: implications for lake restoration and management. Environ. Model. Software 26, 354–370.
Trolle, D., Skovgaard, H., Jeppesen, E., 2008. The Water Framework Directive: setting the phosphorus loading target for a deep lake in Denmark using the 1D lake ecosystem model DYRESM–CAEDYM. Ecol. Model. 219 (1–2), 138–152.
Trolle, D., Hamilton, D.P., Hipsey, M.R., Bolding, K., Bruggeman, J., Mooij, W.M., Janse, J.H., Nielsen, A., Jeppesen, E., Elliott, J.E., Makler-Pick, V., Petzoldt, T., Rinke, K., Flindt, M.R., Arhonditsis, G.B., Gal, G., Bjerring, R., Tominaga, K., Hoen, J., Downing, A.S., Marques, D.M., Fragoso Jr., C.R., Søndergaard, M., Hanson, P.C., 2012. A community-based framework for aquatic ecosystem models. Hydrobiologia 683 (1), 25–34.
Trolle, D., Elliott, J.A., Mooij, W.M., Janse, J.H., Bolding, K., Hamilton, D.P., Jeppesen, E., 2014. Advancing projections of phytoplankton responses to climate change through ensemble modelling. Environ. Model. Software 61, 371–379.
Turuncoglu, U.U., Dalfes, N., Murphy, S., Deluca, C., 2013. Toward self-describing and workflow integrated Earth system models: a coupled atmosphere-ocean modeling system application. Environ. Model. Software 39, 247–262.
van Engeland, T.V., Kluijver, A.D., Soetaert, K., Meysman, F.J.R., Middelburg, J.J., 2012. Isotope data improve the predictive capabilities of a marine biogeochemical model. Biogeosci. Discuss. 9 (7), 9453–9486.
Vander Zanden, M.J., Shuter, B.J., Lester, N., Rasmussen, J.B., 1999. Patterns of food chain length in lakes: a stable isotope study. Am. Nat. 154 (4), 406–416.
Varela, R.A., Cruzado, A., Tintore, J., Garda Ladona, E., 1992. Modelling the deep-chlorophyll maximum: a coupled physical-biological approach. J. Mar. Res. 50 (3), 441–463.
Verhagen, J.H.G., Nienhuis, P.H., 1983. A simulation model of production, seasonal changes in biomass and distribution of eelgrass (Zostera marina) in Lake Grevelingen. Mar. Ecol. Prog. Ser. 1 (2), 187–195.
Villamizar, S.R., Pai, H., Butler, C.A., Harmon, T.C., 2014. Transverse spatiotemporal variability of lowland river properties and effects on metabolic rate estimates. Water Resour. Res. 50 (1), 482–493.
Vollenweider, R., Kerekes, J., 1982. Eutrophication of Waters; Monitoring, Assessment and Control. OECD, Brussels.
Von Westernhagen, N., Hamilton, D.P., Pilditch, C.A., 2010. Temporal and spatial variations in phytoplankton productivity in surface waters of a warm-temperate, monomictic lake in New Zealand. Hydrobiologia 652 (1), 57–70.
Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43 (1), W01411.
Washington, H.G., 1984. Diversity, biotic and similarity indices: a review with special relevance to aquatic ecosystems. Water Res. 18 (6), 653–694.
Ward, N.K., Fitchett, L., Hart, J.A., Shu, L., Stachelek, J., Weng, W., Zhang, Y., Dugan, H., Hetherington, A., Boyle, K., Carey, C.C., Cobourn, K.M., Hanson, P.C., Kemanian, A. R., Sorice, M.G., Weathers, K.C., 2019. Integrating fast and slow processes is essential for simulating human–freshwater interactions. Ambio 48, 1169–1182.
Webster, I.T., Rea, N., Padovan, A.V., Dostine, P., Townsend, S.A., Cook, S., 2005. An analysis of primary production in the Daly River, a relatively unimpacted tropical river in northern Australia. Mar. Freshw. Res. 56 (3), 303–316.
Wells, N.S., Maher, D.T., Erler, D.V., Hipsey, M.R., Rosentreter, J.A., Eyre, B.D., 2018. Estuaries as sources and sinks of N_2O across a land-use gradient in subtropical Australia. Global Biogeochem. Cycles 32, 877–894.
Wikner, J., Panigrahi, S., Nydahl, A., Lundberg, E., Båmstedt, U., Tengberg, A., 2013. Precise continuous measurements of pelagic respiration in coastal waters with Oxygen Optodes. Limnol Oceanogr. Methods 11, 1–15.
Williams, R.N., de Souza Jr., P.A., Jones, E., 2014. Analysing coastal ocean model outputs using competitive-learning pattern recognition techniques. Environ. Model. Software 57, 165–176.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Winslow, L.A., Zwart, J.A., Batt, R.D., Dugan, H.A., Woolway, R.I., Corman, J.R., Hanson, P.C., Read, J.S., 2016. LakeMetabolizer: an R package for estimating lake metabolism from free-water oxygen using diverse statistical models. Inland Waters 6 (4), 622–636.
Woodward, B.L., Marti, C.L., Imberger, J., Hipsey, M.R., Oldham, C.E., 2017. Wind and buoyancy driven horizontal exchange in shallow embayments of a tropical reservoir: lake Argyle, Western Australia. Limnol. Oceanogr. 62 (4), 1636–1657.
Xu, J., Hood, R.R., 2006. Modeling biogeochemical cycles in Chesapeake Bay with a coupled physical–biological model. Estuar. Coast Shelf Sci. 69 (1–2), 19–46.
Yao, H., Samal, N.R., Joehnk, K.D., Fang, X., Bruce, L.C., Pierson, D.C., Rusak, J.A., James, A., 2014. Comparing ice and temperature simulations by four dynamic lake models in Harp Lake: past performance and future predictions. Hydrol. Process. 28 (16), 4587–4601. https://doi.org/10.1002/hyp.10180.
Zhu, Y., Hipsey, M.R., McCowan, A., Beardall, J., Cook, P.L.M., 2016. The role of bioirrigation in sediment phosphorus dynamics and blooms of toxic cyanobacteria in a temperate lagoon. Environ. Model. Software 86, 277–304.