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

Matthew R. Hipsey, Gideon Gal, George B. Arhonditsis, Cayelan C. Carey, J. Alex Elliott, Marieke A. Frassl, Jan H. Janse, Lee de Mora, Barbara J. Robson

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.

Keywords: AEM, CSPS, Ecological modelling, Uncertainty, Water quality, Freshwater, Marine
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 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
consequence has been that it is difficult to compare model performance between studies and approaches in a way that allows a clear definition of comparison etc. (Janssen et al., 2015). For any given application context, of model approaches, which is compounded by the fact that AEMs have models are fit for purpose. The reason can be found in the large diversity to approach aquatic ecosystem simulation and how to decide when Bennett et al. (2013) and previous discussions by others (Elliott et al., - available assessment approaches.

The limits of their predictions. was first noted by Arhonditsis and Brett (2004) and subsequent analyses was noted by Arhonditsis and Brett (2004) and subsequent analyses by others. This trend was first noted by Arhonditsis and Brett (2004) and subsequent analyses have pointed to a similar conclusion.

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: |
| BIAS, MAE, NMAE | Bias Mean Absolute Error | Normalised Mean Absolute Error |
| RMSE | Root Mean Square Error | |
| MEF, NSE, B | Model Efficiency, Nash-Sutcliffe Efficiency, Bardsley coefficient |
| d2 | Index of Agreement | Model Skill Score Willmott index |
| R, SR | Correlation coefficient | Spearman Rank Correlation |
| TD | Taylor/Target Diagram | |
| DF | Distribution Functions | |
| FFT, WT, WC | Fast Fourier Transform Wavelet Transform Wavelet Coherence |
| CCF, ACF | Cross-correlation Function Autocorrelation Function |

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 (Hellweg, 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 dete-
roration when key system functions are pushed over thresholds, resulting in the loss of resilience and the emergence of a regime shift (Schaefer 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 (Elliot 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.
4.4. Monitoring and evaluation of model performance

4.4.1. 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., Yeo 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., 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., DiSSayanake et al., 2019), and shear stresses, for example, at the sediment-water interface 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 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

4.4.2. 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., Sohna 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 | Typical range of data | Spatial scale | Assessment technique | Example references |
|-------------------------|-------------|------------|-----------------------|---------------|----------------------|--------------------|
| Water balance, waves & circulation | | | | | | |
| Water level | Water balance, waves & circulation | Time-series comparison | minutes-monthly | point | multiple points | E, VTS, V(XX), V(TS), V(TX) | Missaghi and Honda (2010) |
| Wave length and period | | | seconds-minutes | point | V(TS) | E, R, V(TS), V(XZ) | Ji (2017) |
| Frequency spectra | | | seconds-minutes | point | V(TS), V(XZ), V(YZ) | FFT, WT | Ji (2017) |
| Evaporation | | Time-series comparison | minutes-daily | point | V(TS), E, R | V(TS), E, R, V(XY) | Rimmer et al. (2009); Nussboim et al. (2017) |
| Temperature | | Time-series comparison | minutes-monthly | point | V(XX), V(YY), V(ZZ) | FFT, WT | Kara et al. (2012) |
| Heat & salt balance | | | | | | |
| Albedo | | Time-series comparison | minutes-monthly | point | multiple points | V(TS), E, R | E, R | Stadnyk et al. (2013) |
| Temperature or salinity | | Spatial comparison | daily-monthly | surface layer | V(XX), V(YY), V(ZZ) | V(XX), V(YY), V(ZZ) | E, R | Stadnyk et al. (2013) |
| Spatial variability | | | daily-monthly | surface layer | V(XX), V(YY), V(ZZ) | V(XX), V(YY), V(ZZ) | E, R | Stadnyk et al. (2013) |
| Spatial patchiness | | | daily-monthly | surface layer | V(XX), V(YY), V(ZZ) | V(XX), V(YY), V(ZZ) | E, R | Stadnyk et al. (2013) |
| Eddy structure | | | daily-monthly | water column | V(XX), V(YY), V(ZZ) | V(XX), V(YY), V(ZZ) | E, R | Stadnyk et al. (2013) |
| Benthic perimeter heat exchange | | | weekly-monthly | water column | V(XX), V(YY), V(ZZ) | V(XX), V(YY), V(ZZ) | E, R | Stadnyk et al. (2013) |
| Ice cover | | | weekly-monthly | point | | V(XX), V(YY), V(ZZ) | E, R | Stadnyk et al. (2013) |
| Property being assessed | Description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g. processing requirements) | Example references |
|-------------------------|-------------|------------------|---------------------------------------------|--------------|-----------------------------------|----------------------------------------|-------------------|
| Date of ice on/off      | Date of ice on/off | 1b | weekly-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) |
| Stratification          | Temperature, salinity, density | Depth comparison | 1a | minutes-monthly | vertical profile (continuous) or multiple depths (discrete) | V(TZ), R, d₂, MAE(ΔTZ), TZ error contour plot highlights errors in thermocline or pycnocline depth, by comparing interpolated observation and model profiles over time | Meineck et al. (2007) |
|                        | 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 | Meineck 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. | Woodward et al. (2017) |
|                        | Lateral gradient | 1b | hourly-daily | horizontal transect | V(XZ), MAR(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 | |
|                        | Internal wave frequency spectra | 1c | minutes-hourly | water column | FFT, WT, WC | Comparison of frequency spectra to demonstrate wave periods and modes are being reproduced | 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) |
|                        | 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 | Bruce et al. (2018) |
|                        | 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 | 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 | Chen et al. (2007) |
|                        | 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 | |
|                        | 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 change in depth measured using hydro-acoustic methods can be used to validate models | |
|                        | 3 | ad hoc | multiple points | V(other) | |

(continued on next page)
| Property being assessed | Description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g. processing requirements) | Example references |
|-------------------------|-------------|------------------|---------------------------------------------|--------------|------------------------------------|------------------------------------------|------------------|
| Variation in particle size distribution | | | | | | | |
| Spatial changes in bathymetry | | 3 | seasonal-decadal | bottom layer | V(XY), MAE(ΔXY), $d_2$ | Spatial differences in particle size composition of bottom sediment can be used to validate areas of differential deposition rates between particle size classes. 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 | |

* E = user determined combination of traditional error metrics, including BIAS, MAE, NMAE, MEF, NSE, B and/or $d_2$.
* 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.
* 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) | Comments (e.g. processing requirements) | Example references |
|-------------------------|-------------|------------------|---------------------------------------------|--------------|------------------------------------|------------------------------------------|------------------|
| 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 | |
| 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) Misaghi 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 | Webster et al. (2005) |
| Photosynthetic oxygen production | | 2b | minutes-hourly | point | R, DF | Comparison of simulated rate of oxygen production calculated from oxygen sensor logging at high frequency | |

(continued on next page)
| Property being assessed               | Description                                                                 | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g., processing requirements)                                                                 | Example references               |
|--------------------------------------|------------------------------------------------------------------------------|------------------|---------------------------------------------|---------------|-----------------------------------|----------------------------------------------------------------------------------------------------------|----------------------------------|
| Oxygen demand (net)                   |                                                                              | 2b               | minutes-hourly                              | point         | R, DF                             | Comparison of simulated rate and oxygen metabolism calculated from oxygen sensor logging at high frequency   | Hetland and DiMarco (2008)        |
| 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 | Brady et al. (2013)               |
| 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 | Chipman et al. (2012)             |
| 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 | Bryant et al. (2010)              |
| 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 wind speed and atmospheric gas flux can be used to ensure scaling | Huang et al. (2018)               |
| 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)               |
| 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)                  |
| Nutrients & organic matter           | Concentrations of SiO\(_2\), PO\(_4\), DOP, POP, TP, NO\(_3\), NH\(_4\), DON, PON, TN | 2a               | daily-monthly                                | vertical slice | VO(XZ), MAE (ΔXZ)               | Model assessment against data from grab samples                                                               | Ayata et al. (2013)              |
|                                      | Time-series comparison                                                      | 1a               | daily                                        | vertical slice | V(TS), E, R                     | Comparison with optical measurements, e.g. from autonomous underwater vessels                                | Bruce et al. (2006)              |
|                                      | Vertical transect                                                           | 1a               | daily                                        | vertical slice | V(TS), E, R                     | Comparison with optical measurements, e.g. from autonomous underwater vessels                                | Ménesguen et al. (2019)          |
|                                      | Multi-variate assessment                                                    | 3                | daily-monthly                                | multiple points | V(TS), E, R                     | Taylor Diagram comparing the relative error fit metrics of multiple interacting variables                    | Taylor (2001)                    |
|                                      | 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, DOC:TOC, DOC:POC, POC:TON                            | 1b               | daily-monthly                                | point          | V(TS), E, R, DF                 | Ratio of nutrient pools relevant to primary productivity and nutrient management                            | Li et al. (2013)                 |
|                                      | Inorganic nutrient: DIC vs. Organic nutrient:carbon                         | 3                | ad hoc                                       | basin          | V(other)                        | Ratio of partitioning of dissolved and total organic carbon                                                 | Martiny et al. (2013)            |
|                                      | Scattered data                                                              | 2a               | weekly-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 | de Mora et al. (2016)            |
| Carbon and nutrient metabolism       | Non-conservative fraction                                                   | 1b               | weekly-seasonal                              | horizontal transect | V(other)               | Scatter plot comparison showing sensitivity of changes in P associated with mineral/clay particles          | Chao et al. (2010)               |

(continued on next page)
| Property being assessed                        | Description                                      | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1)* | Comments (e.g. processing requirements)                                                                 | Example references                                   |
|-----------------------------------------------|--------------------------------------------------|------------------|---------------------------------------------|---------------|--------------------------------------|------------------------------------------------------------------------------------------------------|-----------------------------------------------------|
| 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 Schernicki (2008)                                                  |
| Rate of OM mineralisation                     | 2a, b                                           | weekly-monthly   | point                                       | R, DF, V(other)| Dark incubations measuring oxygen consumption (BOD) or CO₂ 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          | Fluxes may be obtained from benthic chambers or isotope pairing 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 |                                                                                                     |
| TCO₂ 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   | 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    | Carbon export as a function of depth in the ocean, fit to exponential curve. | Martin et al. (1987) Butenschon et al. (2012) Huang et al. (2019)                                      |
| Greenhouse gas dynamics                       | pCO₂, pH, N₂O time-series comparison            | 1a               | minutes-monthly                             | point          | Model assessment against data from grab samples or underway sampling | Schmid et al. (2017) Wells et al. (2018) Huang et al. (2019)                                         |
|                                             | pCO₂, pH, N₂O transect or profile               | 1a               | ad hoc                                      | horizontal transect; vertical profile | Transect along gradient, or vertical profile |                                                                                                     |
|                                             | CO₂, CH₄, N₂O surface flux                      | 2a               | minutes-monthly                             | point          | Floating chamber can provide estimate of point-scale CO₂ flux; relationship between windspeed and atmospheric gas flux can be used to ensure scaling | Schmid et al. (2017)                                   |
|                                             | CH₄ ebullition                                   | 2a               | ad hoc                                      | water column   | Direction measurement with capture chamber or indirectly through acoustic bubble monitoring |                                                                                                     |
|                                             | DOC, DON, NH₄ or NO₃ isotopic fraction           | 2b               | monthly-seasonal                            | point          | 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 | Comparison of pore-water (dissolved) concentration data with depth into the sediment |                                                                                                     |
|                                             | Sediment particulate concentration profiles      | 1a               | ad hoc                                      | sediment profile | Comparison of %C and N with depth into the sediment |                                                                                                     |
|                                             | Oxygen penetration depth                        | 1b               | ad hoc                                      | sediment column | Vertical depth into the sediment oxygen is predicted to penetrate |                                                                                                     |
|                                             | Denitrification efficiency                      | 3                | ad hoc                                      | sediment column | 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 | 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 | 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)
| Property being assessed | Description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g. processing requirements) | Example references |
|-------------------------|-------------|-----------------|--------------------------------------------|--------------|-----------------------------------|----------------------------------------|------------------|
| DIC, TFe, Fe(OH)$_3$, FeII, SO$_4^{2-}$, H$_2$S, TMn, MnII, major ions, metals, pH, Eh, alkalinity | 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) |
| Depth of redoxcline | Time-series comparison | 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. | | Hipsey et al. (2014) Shen et al. (2019) |
| Area of system crossing an acidification threshold | Time-series comparison | 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. | | |
| 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 | In situ flux measurement | 2a | ad hoc | V(TS), R, DF | Data collected from sediment traps can be used to validate simulated rates of particulate deposition. | | |
| Dissolved sediment flux | Rate of bottom water accumulation/depletion | 2b | weekly-monthly | V(TS), R | In situ determination of dissolved species (e.g., FeII, Al, SO$_4^{2-}$); can be compared as a function of oxygen or other variable. | | |
| Sediment quality | Pore-water concentration profiles | 1a | ad hoc point; sediment column | 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 | | | | | |
| 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 | 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) Miller et al. (2011) Margvelashvili et al. (2013) Margvelashvili et al. (2018) Chao et al. (2010) |
| Suspended particulates, turbidity | Spatial comparison | 1a | daily-monthly surface layer | V(XY), MAE($\Delta$XY), $\sigma_2$ | Satellite acquired turbidity data from MODIS etc compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match. | | |
| Particle composition | Chl-a or TOC/TSS | 3 | weekly-monthly point | R, DF | Scatter plot comparison of fraction of Chl-a or TOC fraction of suspended particulate pool. | | Hiremy et al. (2004) |
| 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 | V(TS), R, DF | Data from sediment traps can validate model sedimentation flux of particulate matter. | | Ostrovsky and Yacobi (2010) Fuji et al. (2007) Chao et al. (2007) Chao et al. (2010) |
| Extinction coefficient | Time-series comparison | 1b | weekly-monthly point | V(TS), R, DF | May need to specify wavelengths. | | |
| Secchi/euphotic depth | Time-series comparison | 1b | weekly-monthly point | V(TS), R, DF | Model assessment of light penetration against routine transparency measures. | | Robson et al. (2017) |

(continued on next page)
| Property being assessed | Description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g. processing requirements) | Example references |
|-------------------------|-------------|------------------|---------------------------------------------|--------------|-------------------------------------|------------------------------------------|-------------------|
| Relationship wet season river inputs and subsequent (e.g. dry season) average Secchi depth | Depth comparison | 1a | hourly-monthly | vertical profile | V(TZ), R, MAE (ΔTZ) | Assessment of seasonal lag between catchment inputs and water response | Fujii et al. (2007) |
| PAR: Photosynthetically Active Radiation intensity | Light quality | 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 | CDOM: Chromophoric Dissolved Organic Matter (a.k.a. gelvin, gelbstoff, colour) | 3 | weekly-monthly | point | V(TZ), E, R V(XY), MAE (ΔXY) | Model assessment against data from grab samples Satellite acquired CDOM data from MODIS etc compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match | Elliott et al. (2009) Bruce et al. (2009) Gal et al. (2009) Granger et al. (2009b) Ng et al. (2011) Guillaud et al. (2000) Meneguz et al. (2007) |
| Chlorophyll-a | Chlorophyll-a concentration | 1a | hourly-monthly | point | V(TS), E, R | Model assessment against data from samples or Chl-a sensors | Fujii et al. (2007) |
| Bloom peak biomass | 1b | weekly-monthly | point | R, DF, V(other) | Scatter plot of simulated vs observed bloom peak biomass | Guillaud et al. (2000) Fujii et al. (2007) Jones et al. (2016) |
| 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 | Meneguz et al. (2007) |
| Biomass (interannual) variability | Frequency spectra | 1b | weekly-monthly | point | FFT, WT, WC | Data measured from a chlorophyll-a sensor logging at high frequency | Meneguz et al. (2007) Doxey et al. (2009) Sinha et al. (2010) Jiang and Xia (2018) Quere et al. (2005) Doxey et al. (2009) Hilmer et al. (2008) Ng et al. (2011) |
| Spatial comparison | 1c | minutes-hourly | point | FFT, WT, WC | Satellite acquired Chl-a estimates compared pixel for pixel with the simulation. Model or data may require averaging to ensure spatial resolutions match | Meneguz et al. (2007) Rocha et al. (2019) |
| Spatial variability | 1c | daily-monthly | surface layer | CDF, V(TX), V(XY), V(XZ) | Compares distribution and range of Chl-a within the domain without conducting pixel by pixel comparison | Doxey et al. (2009) Sinha et al. (2010) Jiang and Xia (2018) Quere et al. (2005) Doxey et al. (2009) Hilmer et al. (2008) Ng et al. (2011) |
| Spatial patchiness | 3 | daily-monthly | surface layer | ACF, V(other) | Characteristic patch length scale comparison, including skewness, kurtosis. | Rocha et al. (2019) |
| Deep Chlorophyll Maximum (DCM) | 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; 14C uptake) compared with model. Could be reported as a function of environmental changes (e.g. light) | Granger 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 |
|-------------------------|-------------|------------------|---------------------------------------------|---------------|----------------------------------------|----------------------------------------|-------------------|
| PPB: Phytoplankton production per biomass | Eppley curve: Scaling relationship of net production as a function of 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 | Granger 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) Hipsey et al. (2006) |
| | 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 | |
| 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 \(d_0\).

\(^b\) 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 per hypoxia. Of comparison against EEMS studies. Fractory 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$_2$ 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 $^{14}$C-DIC, relatively to a conservative tracer. In this case, the use of isotopes in the calibration helped to reduce equifinality, since models can 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 is model-dependent, for example using in situ microprobe 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 (Salmén 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 when 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.
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 organisms 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...
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 potential 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 bio-accumulation 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 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 f) 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; Mieletiner 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 spatial-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 correspondence (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
**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) | Comments (e.g., data collection and processing requirements) | Example references |
|-------------------------|--------------------------------------------------|------------------|-------------------------------------------|--------------|------------------------------------|-------------------------------------------------------------|-------------------|
| Primary producers and lower trophic levels (bacteria, flagellates, ciliates, 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); Miehlsner and Reichert (2008); Gal et al. (2009); Rigosi et al. (2011); Chong et al. (2014) |
| Species or functional group biomass (including bacteria, micro zooplankton and phytoplankton) | Bloom magnitude | 1b | weekly-monthly point | R | Scatter plot of simulated vs observed bloom peak biomass | Peeters et al. (2007) |
| | Bloom duration and timing | 1b | weekly-monthly point | R, V(TS) | Visual comparison of bloom onset dates in model and observational time-series | Gal et al. (2009) |
| | Biomass (interannual) variability | 1c | weekly-monthly point | SR | Rank correlation of annual bloom magnitudes in multi-annual simulation | Galet 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. (2018) |
| | 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 | Bussielli et al. (2014) |
| | | Group specific attenuation coefficient | 1b | ad hoc | multiple points | V(other) | Comparison of model separation of group components contributing to the light extinction coefficient, \( K_d \) | Kromkamp and Walsby (1990) |
| | | 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 | Dugdale and Wilkerson (1986); Popendorf and Duhamel (2015); Morozov (2010) |
| | Vertical sinking rates | 2a | ad hoc | water column | V(other), R | Comparison of reported ranges from nutrient uptake experiment data with frequency distribution of model uptake rate | Buzzelli et al. (2014); Arora-Williams et al. (2018) |
| | Phytoplankton nutrient uptake rates, incl. nitrogen fixation | 2a | ad hoc | point | V(other), R | For models resolving size variation in the planktonic community, allocation of biomass across the size spectrum should occur consistent with the Sheldon spectrum | Bussielli et al. (2014) |
| | 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 | Dugdale and Wilkerson (1986); Popendorf and Duhamel (2015); Morozov (2010) |
| | Organism nutrition source partitioning | 2b | ad hoc | point | V(other) | 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) |
| | Gradient in microbial gene expression | 2a | ad hoc | horizontal transect or vertical profile | 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 | Li et al. (2013); Granger et al. (2009a); Sauterey et al. (2015) |
| | 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); Granger et al. (2009a); Sauterey et al. (2015) |
| | Canberra distance | 3 | ad hoc | point | V(other) | Comparison of the distance between cell-volume distributions | Li et al. (2013); Granger et al. (2009a); Sauterey et al. (2015) |
| Spatial niche formation | Horizontal patterns in species/group distribution | 1a | daily-monthly | surface layer | V(XY), R, MAE(\( \Delta X \)), MAE(\( \Delta Y \)) | Satellite or UAV acquired fluorescence data compared pixel for pixel with simulation. Model or data may require averaging to ensure spatial resolutions match | Menegus et al. (2007); Alexander and Imberger (2008); Down et al. (2009); Sinha et al. (2010); Sallée et al. (2013) |
| | Biomass of each species/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 | (continued on next page)
| Property being assessed | Approach and description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g., data collection and processing requirements) | Example references |
|-------------------------|--------------------------|------------------|---------------------------------------------|---------------|-----------------------------------|-------------------------------------------------------------|-------------------|
| Biomass – depth relationship | 3 hourly | horizontal transect | V(TX) | Biomass as a function of depth along a transect or depth limit of biomass observations | Chien et al. (2013) |
| 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 | Fujii et al. (2007) |
| Relative abundance of each species or PFT against the total community Chi-a | 3 weekly-monthly | surface layer | V(other), V(XY) | Relative abundance of each PFT against the total community Chi-a, presented spatially | Hirata et al. (2011) |
| 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) |
| 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. | Granger et al. (2009) |
| 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 | Granger et al. (2010) |
| 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) |
| Population size distribution | 1c ad hoc | point, basin | DF | Comparison of size distribution of organisms within a population against observed ranges/counts | Dijkstra and Uittenbogaard (2010) |
| 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) |
| Benthic metabolism | 2b ad hoc multiple points | V(other), R, DF | Difference between whole-system productivity (e.g. diurnal O2 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 | 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 | Testa et al. (2017) |
| 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 | Testa et al. (2017) |
| Spatial niche formation | Seed, larval dispersal extent | 2a ad hoc multiple points; bottom layer | V(XY), R | Comparison of measured seed or larval densities with modelled extent | Savina and Ménesguen (2008) |
| Horizontal patterns in species/group distribution within the community | 3 ad hoc bottom layer | V(TX), 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) |
Table 4 (continued)

| Property being assessed | Approach and description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g., data collection and processing requirements) | Example references |
|-------------------------|--------------------------|------------------|-------------------------------------------|--------------|------------------------------------|---------------------------------------------------------------|-------------------|
| Population structuring (population cohort) | Biomass – environment relationships | 3 | ad hoc | system | V(other) | Multi-variate statistical clustering of organism presence/biomass, as a function of environmental conditions | Le Goff et al. (2017) |
| 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 | |
| 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 Meneguen (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 | Gal et al. (2009) |
| | 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) | Model comparison with in situ metabolic loss rate determinations | |
| | 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 Fish or zooplankton biomass stoichiometry of N, P, HUFA etc. can be used to cross-check metabolic processes are balanced | Morozov (2010) |
| | Internal stoichiometry | 3 | ad hoc | point | V(TS), DF | 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 | Perhar et al. (2012) |
| | Vertical positioning of biomass (e.g. diel migration) | 3 | ad hoc | water column | V(TZ), MAE (ΔTZ) | Biomass as a function of depth along a transect, or depth limit of biomass observations | Greve and Krause-Jensen (2005) |
| 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 | 3 | weekly-annual | basin, system | V(other) | Process rates (growth, mortality, reproduction) should scale with age/size appropriately | Harfoot et al. (2014) |
| | Body mass: mortality | 3 | weekly-annual | basin, system | V(other) | Accurate depiction of the weight-length relationship of a fish population demonstrates the model is allocating resources to growth correctly | |
| | Body mass: reproductive success | 3 | weekly-annual | basin, system | V(other) | Accurate depiction of the weight-length relationship of a fish population demonstrates the model is allocating resources to growth correctly | |
| | Body mass: population | 3 | weekly-annual | basin, system | V(other) | Accurate depiction of the weight-length relationship of a fish population demonstrates the model is allocating resources to growth correctly | |
| 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 | Saikey et al. (2013) |
| | Biomass of each species/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 | Saikey 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)
| Property being assessed | Approach and description | Validation level | Spatial scale | Typical range of data observation | Assessment technique (see Table 1) | Example references |
|-------------------------|--------------------------|------------------|--------------|----------------------------------|-----------------------------------|-------------------|
| Habitat quality         | ad hoc                   | 3                | basin scale  | ad hoc                           | system                            | Link (2010)       |
| Recruitment and migration events | ad hoc                   | 3                | basin scale  | -                               | system                            | Heymann et al. (2016) |
| Food web structure     | ad hoc                   | 3                | basin scale  | ad hoc                           | system                            | Heymann et al. (2016) |
| Food web organisation  | ad hoc                   | 3                | basin scale  | ad hoc                           | system                            | Heymann et al. (2016) |

Note: Table 4 (continued)
Table 4 (continued)

| Property being assessed | Approach and description | Validation level | Typical range of data observation frequency | Spatial scale | Assessment technique (see Table 1) | Comments (e.g., data collection and processing requirements) | Example references |
|-------------------------|--------------------------|------------------|-------------------------------------------|---------------|-----------------------------------|-------------------------------------------------------------|-------------------|
| Foodweb linkages        | Dietary analysis of food sources | 2a | ad hoc | system | V(other), DF | Respiration per unit biomass, compared for autotrophs and heterotrophs, where the latter is computed by subtracting autotrophic respiration from the total. 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), Vorder Zenden et al. (1999), Makler-Pick et al. (2017) |
|                         | Intraguild predation       | 3   | ad hoc | system | V(other) | Ability of model to capture adaptation of pathways in response to changing predator presence and behaviour. Relationship between heterotrophic biomass and net primary production indicates upward transfer of energy. | Sailley et al. (2013) |
|                         | Trophic efficiency         | 3   | monthly-seasonal | system | V(other) | Food chain length is expected to increase with net primary production. | |
|                         | Relationship between food chain length and net primary production | 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), Robson et al. (2017) |
|                         | 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. Scatter plot comparison of relevant ecosystem variables (pH, O₂, Chl-a, nutrients or population abundance) to hydro-climatic forcing variable. | |
|                         | Response to climate changes (inflows, water level, temperature, CO₂) | 3 | monthly-annual | system | R, V(TS) | Identifying temporal (or spatial) coherence and strengths of relationships between linked variables (e.g. phytoplankton and zooplankton) under non-stationary forcing conditions. Changes in the strength of variable interaction following perturbation could be depicted by wavelet coherence, showing coupling and/or de-coupling of correlation between two or more variables in frequency space. | Huang et al. (2019) |
| Variable coherency      | Time-series analysis       | 1b | daily-decadal | system | CCF | 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. | |
| Spectral analysis       | hourly-decadal point      | 1b | WT, WC | | | Changes in the strength of variable interaction following perturbation could be depicted by wavelet coherence, showing coupling and/or de-coupling of correlation between two or more variables in frequency space. | |
| State space diagrams    | monthly-decadal system    | 3   | | | | | |
| 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. | Papparov 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) | Comments (e.g., data collection and processing requirements)                                                                 | Example references |
|----------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|------------------|---------------------------------------------|---------------|------------------------------------|---------------------------------------------------------------------------------------------------------------------------|-------------------|
| Response of biodiversity/richness indices to hydrodynamics and nutrient loading        |                                                                                          | 3                | monthly-annual                              | system        | V(other)                           |                                                                                                                           |                   |
| 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)  |
| Return time to "old" or "new" state                                                    |                                                                                          | 3                | daily-annual                                | system        | V(TS), V(other)                    |                                                                                                                           | O'Brien et al. (2018)  |
| Magnitude of return towards the "old" state                                            |                                                                                          | 3                | daily-annual                                | system        | V(TS), V(other)                    |                                                                                                                           | Müller et al. (2016)  |
| Potential for bistability                                                              |                                                                                          | 3                | daily-annual                                | system        | V(other)                           |                                                                                                                           | O'Brien et al. (2018)  |
| Critical slowing down                                                                  |                                                                                          | 3                | daily-monthly                               | system        | V(other), ACF, WT                  |                                                                                                                           | Janse et al. (2008)  |
|                                                                                       |                                                                                          | 3                | monthly-annual                              | system        | V(other)                           | Comparison of model’s ability to capture the lengthening response times to disturbance, prior to a catastrophic transition | Janse et al. (2010)  |
| 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)  |
|                                                                                       | Extent of hysteresis                                                                      | 3                | monthly-annual                              | system        | V(other)                           |                                                                                                                           | O'Brien et al. (2018)  |
|                                                                                       | 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) |                   |
| 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   | Dietze et al. (2013)  |

a E = user determined combination of traditional error metrics, including BIAS, MAE, NMAE, MEF, NSE, B and/or d².

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.
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 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 Menesguen, 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 Bocanivov 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 Höfker 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. Sailer 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: microzooplankton, 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 δ^{15}N signature from observed organism data, in the context of a marine ecosystem subject to trafficking problems.

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, 2014; Trolle et al., 2012; Arhonditis 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
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” (Kühnert, 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 modelling (Robson and Dourdet, 2015) or evolutionary algorithms 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.

Adyanty, S., Eyre, B.D., Maher, D.T., Santos, I., Golshby-Smith, L., Mangion, P., Hipsey, M. R., 2016. Stable isotopes reduce parameter uncertainty of an estuarine carbon cycling model. Environ. Model. Software 79, 253–265.

Ainsworth, C.H., Pitcher, T.J., 2006. Modifying Krymsk: a 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.

Aleshew, 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), 113–138.

Allen, J.J., Sommerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. J. Mar. Syst. 63, 3–14.

Allen, J.J., 2010. On the emergent properties of marine ecosystems. PNAS Dissertation, http://ediss.sub.uni-hamburg.de/dovolltexte/2010/4847/pdf/Emergent_PropsEcosys tem_Models-JI-Allen.pdf.

Anderson, K.H., Akanes, 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.

Andersson, 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-Valentini, R.A., Nielsen, L., Stubbins, A., 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 model fidelity 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.

Arthur, G.B., Qian, S.S., Stow, C.A., Lamon, E.C., Becklow, 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.

Arthur, G.B., Stow, C.A., Rao, Y.R., Perhar, G., 2014. What has been accomplished twenty years after the Othilie 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., Olsen, S.W., Scandella, B.P., Delwiche, K., Spencer, S.J., Myers, E. M., Abraham, S., Sookkil, 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., Scandia, 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., Cherkouros, N., Jones, K., Margavelathelli, N., Mangion, O., Oubelkhir, K., Ralph, P.J., Riizi, F., Robson, B.J., Schroder, T., Skerratt, J., 2014. 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 79, 101–112.

Baird, M.E., Adams, M.P., Babcock, R.C., Oubelkhir, K., Mangion, O., 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.

Barber, O., 2013. A good fit of measure related to r2 for model performance assessment. Hydrod. Process. 27 (19), 2851–2856.

Bayer, T.K., Burnz, 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.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newman, 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, 1–29.

Bert, F.E., Rovere, S.L., Macal, C.M., North, M.J., Podesta, G.P., 2014. Lessons from a comprehensive validation of an agent-based model: the experience of the Pampas Model of Argentinian agricultural systems. Ecol. Model. 273, 284–298.

Best, E.J., Buzzelli, C.P., Bartell, S.M., Wetzel, R.L., Boyle, W.A., Doyle, R.D., Campbell, K.R., 2001. Modeling submerged macrophyte growth in relation to underwater light climate: modeling approaches and application potential. Hydrobiologia 444 (1), 43–70.

Biemans, F., Freer, J., 2001. Emissivity, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249 (1–4), 11–29.
