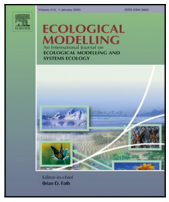
Ecological Modelling 487 (2024) 110562

Contents lists available at ScienceDirect

Ecological Modelling

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



Linking error measures to model questions

Bas Jacobs *, Hilde Tobi, Geerten M. Hengeveld

Mathematical and statistical methods (Biometris), Wageningen University & Research, Wageningen, 6708 PB, The Netherlands

ARTICLE INFO

Keywords:
Error measures
Model evaluation
Model fit
Goodness of fit
Validation
Research methodology
Cyanobacterial blooms

ABSTRACT

Models for forecasting various ecosystem properties have great potential that comes with a need for model validation. Before we can perform such validation, we need to define what it means for the model to perform well, which depends on the question being asked. Often, it seems easy to ignore the model question and take a standard well-known error measure for comparing the model to the available data. The question is whether this practice is adequate. Here, we defined different types of model-data mismatches that may be more or less relevant to different types of questions. We show that error measures differ in their sensitivity to the type of mismatch and robustness to sparse and noisy data. The results imply that a careful selection of error measures, using a clearly defined ecological question as a starting point, is vital to proper model evaluation. While we present our results as generally applicable to the validation of any type of forecasting model, we also illustrate them using cyanobacterial bloom modelling as a detailed example of a case where different questions could be asked of the same model.

1. Introduction

Models that predict the future state of ecosystem properties have great potential, both for answering a wide variety of fundamental questions, and for practical applications in managing ecosystems (Lewis et al., 2023; Dietze et al., 2018; Payne et al., 2017; Petrovskii and Petrovskaya, 2012). Many of these models attempt to predict how certain continuous variables (e.g., nutrient concentrations and species densities) change over time, depending on parameter values, initial conditions, and sometimes input variables (e.g., weather conditions) (Dietze, 2017). A wide variety of such models exist, ranging from process-based (white box) models, for example using differential equations, to data-driven (black box) models, e.g., machine learning (Rousso et al., 2020; Dietze, 2017). The output of such models may range from simple time-series for each variable to entire probability distributions taking into account uncertainty through, e.g., ensemble modelling and may or may not be updated through data assimilation, e.g., Kalman filtering (Luo et al., 2011; Carey et al., 2022; Woelmer et al., 2022). Even within a given modelling approach, many model variants are possible, and often available, for any given ecological system (Rousso et al., 2020; Janssen et al., 2019a; Lewis et al., 2022). Such ecosystem models are often constructed with multiple purposes in mind and re-used to answer a variety of questions. When we want to use a model to answer a specific ecological question, we are, therefore, faced with the question: Which of these models are "better" and which are "worse"?

To answer this question, we need a way to compare our model output to measured ecological data. Such data is usually sparse (available only at intermittent time points) and noisy, and may be gathered as part of ongoing monitoring or an active experiment. Inevitably, the modelled and measured time-series will not match perfectly. The severity of this mismatch can be quantified with an error measure. Many different error measures (or, inversely, measures for goodness of fit) exist (Koutsandreas et al., 2022; Mehdiyev et al., 2016). However, before we blindly pick one of these, we should first define precisely what we mean by "better" and "worse". This definition will not be universal, as it depends on the model application, i.e., the question that we want to answer with the model. It is generally considered good modelling practice to both design and evaluate a model with respect to a specific well-defined goal or question (Jakeman et al.,

Box 1. Terminology

Model: Set of equations that generates a prediction (calculated state) of certain output variables, given certain parameters and input variables.

Data: Experimentally or routinely (e.g., through monitoring) measured values of variables of interest.

Time-series: A set of measured or modelled values for a variable of interest at different time points.

Mismatch: Local difference between a modelled and measured time-series.

Error measure: Method of quantifying the severity of a mismatch as a single value.

* Corresponding author.
E-mail address: bas.jacobs@wur.nl (B. Jacobs).

https://doi.org/10.1016/j.ecolmodel.2023.110562
Received 21 July 2023; Received in revised form 30 October 2023; Accepted 30 October 2023
Available online 18 November 2023

2006; Parker, 2020; Hamilton et al., 2022). Although model questions and evaluation are often mentioned explicitly in discussions of good modelling practice (e.g., Jakeman et al. (2006), steps 1, 2 and 6), most focus has been on model design and the algorithm for estimation. This focus, while important, leaves open the question how to link the choice for a relevant error measure to the ecological question at hand.

Potential questions that we may want an ecological model to answer include:

1. What is the value of an ecological property (e.g., biomass (g/m^2), concentration (g/l), number of individuals (#/ha)) at any given time?
2. What is the qualitative state of an ecosystem attribute (e.g., endangered, dangerous, healthy) at any given time?
3. At what time can we expect a change in a qualitative ecosystem attribute?

These are distinct types of questions and models may be better or worse at answering them, especially when considering ecosystems where sudden changes in properties of interest may occur. Ideally, therefore, the chosen error measure would reflect the specific question under consideration.

However, rather than tailoring the error measure to the model purpose, the vast majority of forecasting models are evaluated using familiar measures such as the root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), Nash–Sutcliffe efficiency (NSE), Kling-Gupta efficiency, or continuous ranked probability score (CRPS), whether in ecology (Lewis et al., 2022), or other environmental sciences, such as hydrology (Clark et al., 2021; Jackson et al., 2019), geosciences (Hodson, 2022), or climate sciences (Gleckler et al., 2008). These measures are ultimately all based on a composite of the differences between model output values ($\hat{y}$) and measurements of those values (y) at every time point where measurements are available. Therefore, these measures seem specifically suitable when the aim of the model is to estimate values (question 1). In many cases where ecological models are being used, however, this may not be the relevant question.

However, the appropriateness of a measure is not the only concern. In ecology and other biological and environmental sciences, data are often sparse and noisy, making robustness of the error measure a relevant concern (Clark et al., 2021). When the goal is to estimate values, well-established answers may exist (for instance, RMSE is better for normally distributed noise, and MAE for Laplacian noise Hodson, 2022). When it is not, there may be a trade-off between robustness and appropriateness that should at least be considered.

Here, we explore ways in which we may better link error measures to our model questions, taking the three questions listed above as examples. First, we will define a range of characteristic mismatches between modelled and observed time-series, focusing on sudden changes in the property of interest. Then, we will select several representatives from a wide range of error measures, both commonly and less commonly used, and compare their outcomes for the different mismatches. We will also examine the robustness of these error measures in the face of sparse and noisy data points. We will use the outcomes to discuss the suitability of the different types of error measures when asking each of the three questions of interest. To facilitate the interpretation of this process with the rather abstractly phrased questions, we provide a detailed example of cyanobacterial bloom modelling in a set of boxes separate from the main text (box 2–4) (see Burford et al. (2020), He et al. (2016), Huisman et al. (2018), Ibelings et al. (2003), Janse and van Liere (1995), Janssen et al. (2019b), Korppoo et al. (2017), Lürling and Mucci (2020), Paerl and Huisman (2008), Paerl and Otten (2013), Page et al. (2018), Recknagel et al. (2008), Saloranta and Andersen (2007), Trolle et al. (2014), van Basshuysen (2023) and Schets et al. (2020)).

Box 2. An example: Models for cyanobacterial bloom prediction

A good example of a case where error measures like RMSE may be less appropriate can be found in the short-term prediction of cyanobacterial blooms in lakes. Cyanobacteria are a group of photosynthetic bacteria that can become a plague by sudden intense blooms (periods of high cyanobacterial densities) that block light from submerged macrophytes (Huisman et al., 2018; Burford et al., 2020). In addition, many species can produce toxic compounds that are harmful to humans and other animals, so that blooms may require more expensive water treatment and the closing of recreational sites (He et al., 2016; Paerl and Otten, 2013). Blooms are particularly frequent under eutrophic and warm conditions, and are therefore expected to become an increasing problem with the changing climate (Paerl and Huisman, 2008). In the ideal world, blooms would be prevented by limiting the nutrient loading of water bodies, but it may take a long time to achieve this and for this change to show an effect, making mitigation measures unavoidable (Lürling and Mucci, 2020).

Models of aquatic ecosystems may help assist in making decisions on how to control or manage blooms. These models may be used to answer a variety of questions, often relating to the occurrence of blooms, with bloom presence and timing of appearance and disappearance events being more relevant than the cyanobacterial densities at any given time point. Since these questions do not relate directly to the value of the output variables at any given time point, RMSE-like error measures are less appropriate here. However, they are still widely used to evaluate model success (e.g., Trolle et al. (2014) and Page et al. (2018)). Depending on the exact goal of the study, alternatives to RMSE-like error measures may therefore be desirable. Many modelling studies focus on long-term scenarios, e.g., to study the impact of climate change, or various management scenarios, making the precise quantitative errors less interesting than the overall trends (Janse and van Liere, 1995; Janssen et al., 2019b; Korppoo et al., 2017; Recknagel et al., 2008; Saloranta and Andersen, 2007). Short-term quantitative, even adaptive, forecasts have also been attempted (Ibelings et al., 2003; Trolle et al., 2014; Page et al., 2018) Short-term forecasts might, for instance, be interesting for the goal of adaptive control, initiating mitigation measures based on model predictions. Choosing a proper error measure for such a case may get convoluted, as the model itself will end up changing the behaviour of the system (van Basshuysen, 2023). This means that the model should be evaluated with respect to the desired system state, i.e., “no blooms”, but if there are no blooms, it becomes hard to determine if the model contributed to this absence or if blooms would have been absent either way. A simpler use-case would be to employ the model as an early warning system for posting warning signs on time, as this would not affect the ecosystem dynamics. Alternatively, we may consider the goal of making an app that will inform people whether a bloom is currently present or absent at a recreational location.

For the specific example of cyanobacterial bloom modelling, the three types of questions that we considered for general ecological models would translate to: (1) “What will the water quality be at any given time?”, (2) “Is it safe to go swimming in this lake tomorrow?”, and (3) “When should we place and remove warning signs for blooms?” The first question relates directly to the cyanobacterial density, the second question relates whether or not a bloom is present (whether the density

exceeds a certain value), and the third relates to the start and end times of a bloom.

2. Methods

2.1. Characteristic mismatches

When deciding on an appropriate error measure for a given question, it can be useful to first examine the behaviour of different types of error measures when confronted with the different types of mismatches between model and data that one might expect. Therefore, we constructed several characteristic mismatches related to sudden changes over time t in a measured ecological variable of interest y and its modelled counterpart $\hat{y}$. The mismatches were chosen to have different relevance when asking each of our three different questions. To make mismatches that capture aspects of quantity, quality, and timing, we considered peaks in y and $\hat{y}$ that differ in timing ('small' and 'large' delays), magnitude, duration, shape, and presence ('missed' prediction), along with a threshold Y that denotes a critical value of y at which we consider some qualitative ecosystem attribute to be either 'present' or 'absent' (Fig. 1). For the precise equations used to construct these mismatches see Appendix A.

Box 3. Characteristic mismatches in the cyanobacterial bloom example

Cyanobacterial blooms are characterised by sudden surges in cyanobacterial density. The characteristic mismatches we consider here could, therefore, be interpreted as differences in timing, magnitude, and duration of blooms (Fig. 1). The 'small delay' then implies that the surge in cyanobacterial density is predicted to start and end later than the actual surge, while a 'large' delay can be considered a 'missed' prediction of such a surge alongside a spurious one. A magnitude mismatch occurs when there is a difference in the cyanobacterial densities reached during a predicted and actual surge, when both surges reach sufficient densities to be considered 'blooms'.

The question if the water is safe for swimming is commonly addressed in protocols for lake management with a fixed threshold value for the cyanobacterial density (or chlorophyll concentration as a proxy) above which swimming should be banned (e.g., Schets et al. (2020)). This threshold could be taken as one example of the threshold that we use in our characteristic mismatches, with exceedance of this threshold representing the presence of a bloom. Error measures can then be defined with respect to the severity of the bloom, the presence of the bloom, or the timing of events where blooms appear or disappear.

2.2. Candidate error measures for comparison

We selected several different types of error measures as examples, aiming for a wide diversity in the types of errors that these might emphasise. We selected these to have representatives of different groups of error measures that each seemed *a priori* more appropriate to one of our three different questions. For clarity, we will focus on error measures for models that produce a single time-continuous output of $\hat{y}$ (i.e., deterministic models rather than probabilistic models). Many of these error measures can be extended to probabilistic models that generate entire probability density functions (e.g., ensemble models) (Simonis et al., 2021), for which we will provide a few examples. Other adjustments may be required for, e.g., spatial models or Kalman filtering. Note that the different error measures cannot be normalised such that the values of different measures can be compared directly. As such, the error measure values in the results can only be compared directly between mismatch types and not between error measures.

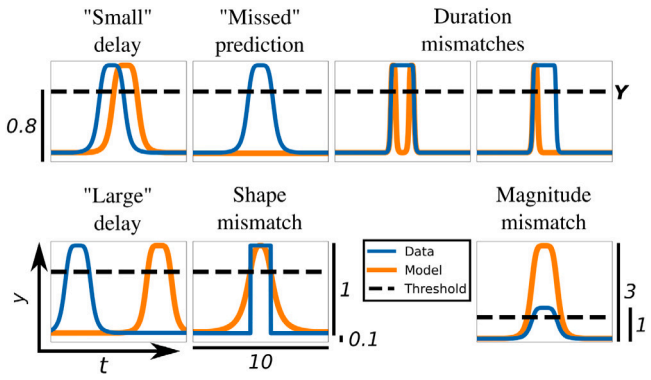


Fig. 1. Overview of the characteristic mismatches between data and modelled values of ecological quantity y. Blue lines indicate (perfect, time-continuous) measured data, and orange lines indicate modelled values. A threshold Y (dashed line) indicates the point where an ecosystem attribute changes from 'absent' to 'present'. Numbers indicate base levels, peak magnitudes and duration. Units are arbitrary.

2.2.1. Root mean squared error and related measures

The root mean squared error (RMSE) is one of the most widely used model error measures (Lewis et al., 2022; Rousso et al., 2020). It is an easy to compute and interpret error measures that imposes for each time point i a cost of $(y_i - \hat{y}_i)^2$, where y_i and $\hat{y}_i$ are, respectively, the measured and modelled output value at time point i. These costs are then averaged over all N time points:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2}{N}}. \tag{1}$$

Many variants of measures in this category exist (e.g., mean absolute error, root mean squared percentage error, mean absolute percentage error, mean absolute scaled error, etc.), which have been elaborately evaluated elsewhere (Jackson et al., 2019; Bennett et al., 2013; Koutsandreas et al., 2022; Mehdiyev et al., 2016; Chen et al., 2017; Hyndman and Koehler, 2006; Morley et al., 2018). We will use the RMSE as an archetype of this category of measures that evaluate how well the model output values match the data at every given time point, and as such seem most appropriate when the goal is to predict these values (question 1).

2.2.2. Critical threshold exceedance measures

If we consider a single threshold value Y for the value of y above which we consider an ecological attribute to be 'present', we can turn both the measured and modelled output values (y and $\hat{y}$) into binary values (x and $\hat{x}$):

$$\begin{aligned} x &= 0 \quad y < Y \qquad \hat{x} = 0 \quad \hat{y} < Y \\ x &= 1 \quad y \geq Y \qquad \hat{x} = 1 \quad \hat{y} \geq Y \end{aligned} \tag{2}$$

We can then evaluate our models on how well they predict if the attribute is present at any given time point, rather than on the precise value of $\hat{y}$, which may be more appropriate when the aim of the model is to estimate the presence of this attribute (question 2). One way to do this would be to turn the cost of each error into $(x_i - \hat{x}_i)^2$ and thus calculate an RMSE on the binarised variables instead of the original ones:

$$\mathrm{RMSE}_{\mathrm{bin}} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}. \tag{3}$$

Without the square root, this is identical to the Brier score, which can also be used in probabilistic forecasts by replacing $\hat{x}_i$ with the probability $\hat{p}_i$ of exceeding Y (Brier, 1950; Wilks, 2011; Taylor and Yu, 2016).

Alternatively, with the binarised values, the model could be evaluated as a classification problem, allowing for the construction of a confusion or error matrix based on the agreement at each measured time point, as is sometimes done in environmental modelling applications with important events that trigger above a certain threshold (Bennett et al., 2013). To capture the essence of a confusion matrix in a single number, many different performance measures for classification problems exist, each with their own advantages and disadvantages (Bennett et al., 2013; Stehman, 1997; Sokolova and Lapalme, 2009; Mehdiyev et al., 2016). Here, we will include two of these in our comparisons: the overall accuracy, and the F_1-score. The overall accuracy is given by the percentage of time points at which the classification is correct, i.e.:

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{fp} + \text{tn}}, \tag{4}$$

where tp is the number of true positives ($x_i = \hat{x}_i = 1$), tn is the number of true negatives ($x_i = \hat{x}_i = 0$), fn is the number of false negatives ($x_i = 1$, $\hat{x}_i = 0$), and fp is the number of false positives ($x_i = 0$, $\hat{x}_i = 1$). The F_1-score is given by:

$$F_1 = \frac{2\text{tp}}{2\text{tp} + \text{fn} + \text{fp}}, \tag{5}$$

which is the harmonic mean of the precision ($\text{tp}/(\text{tp} + \text{fp})$) and recall ($\text{tp}/(\text{tp} + \text{fn})$).

To keep all our measures yielding higher values for worse errors, we subtracted the accuracy and F_1-score from 1 when displaying our results.

2.2.3. Event time errors

If we consider events as changes in an ecosystem attribute for which we desire a timely warning (question 3), then there are four types of errors in event times that could each have their own associated cost: (1) issuing a warning at a wrong time, (2) withdrawing a warning at the wrong time, (3) failing to warn at the right time, (4) failing to withdraw a warning at the right time. In the case where each measured event has a corresponding modelled event, the last two types completely overlap with the first two, making them redundant. However, if there are spurious or missed predictions, all four types need to be taken into account. For each type we might choose a different cost function if desired, but here we will consider the simple case where all types have the same symmetric, saturating cost function, resulting in an overall event time error of:

$$\text{Event time error} = \sum_{i=1}^{4} \sum_{j=1}^{N_i} \frac{|t_{ij} - \hat{t}_{ij}|^n}{|t_{ij} - \hat{t}_{ij}|^n + t_{\frac{1}{2}}^n} \Big/ \sum_{i=1}^{4} N_i, \tag{6}$$

where t_{ij} and $\hat{t}_{ij}$ are, respectively, the measured and modelled time corresponding to the jth error of the ith type, N_i is the number of timing errors of the ith type, $t_{\frac{1}{2}}$ is the time difference at which the cost of an individual timing error is half its maximum value, and n is the hill function coefficient determining the steepness of the cost function. For an elaboration on the underlying cost functions, see Appendix B.

2.2.4. Moving average errors

A pragmatic option to at least tolerate some small temporal mismatches when interested in the timing of events (question 3), may be to compare moving averages of the measured and modelled values of y. The moving average m of output y at time point t is given by:

$$m(t) = \frac{\sum_{j \in k} y(t_j)}{\#k} \qquad k = \{i | t_i \in [t - \Delta t, t + \Delta t]\}, \tag{7}$$

where k is the set of indices for which the associated time points fall within the moving average window and Δt is a tolerance parameter specifying the size of the window. An RMSE can than be calculated for these moving averages rather than for the exact values at each specific time point. This procedure can be applied either directly to the variable y or to the binarised variable x. The Δt should be chosen to represent the window for which temporal mismatches in the prediction of events where an ecosystem attribute changes are considered acceptable.

2.3. Noise and robustness

When choosing an error measure for model evaluation, there may be more considerations than solely its appropriateness to the question of interest. For example, we suspect that, when the available data are sparse and noisy, some error measures may be more robust than others, i.e., they are better at providing the same value on average regardless of the amount of noise ('unbiased') and doing so with minimal deviations from that average ('consistent'). Sometimes a more robust measure may be preferable, even though in the ideal world it would be less appropriate. To study the effect of sparse and noisy data, we recomputed the error measures for the characteristic mismatches from Fig. 1 for various measurement frequencies in the presence of noise. The noise was drawn from a normal distribution with a standard deviation equal to 10% of the "real" value. In the unlikely event that this produced negative numbers, these were set to zero.

For each measurement frequency, this was repeated 1000 times, with uniform spacing between the data points, but a random offset of the first data point. To examine the robustness, we computed for each measure the difference between the average error resulting from these runs and the error computed with perfect data ("bias"), as well as the standard deviation of the errors calculated from the noisy data ("inconsistency").

2.4. Probabilistic models

In many cases, a similar approach can be applied to probabilistic models. We selected several of these and applied them to our mismatches. For details, see Supplementary methods.

3. Results

3.1. Different error measures reflect different types of mismatches

When we compared different types of mismatches using each of the error types mentioned above, we found that different mismatches were reflected differently by the different error measures (Fig. 2). For obvious reasons, large differences between y and $\hat{y}$, with both exceeding threshold Y (magnitude mismatch), yielded far larger error values for RMSE-like methods than for measures that respond to the presence or timing of a qualitative attribute. Also, a 'missed' prediction was considered particularly bad by measures based on event times (all events are incorrect) and the F_1 score, which punishes the lack of true positives, even when the majority of time points have (correct) true negatives.

Furthermore, both the measures based on event times and moving averages were less strict on 'small' delays (Fig. 3), though for long durations of threshold exceedance where the same fraction of the duration is missed, the event time error was actually stricter (Supplemental Figs. S.1 and S.2). 'Large' delays, which were essentially a missed event and a spurious event prediction combined, were considered universally bad. Most measures considered this to be worse than only missing an event, except the event time and F_1 measures, which gave their maximum error to both.

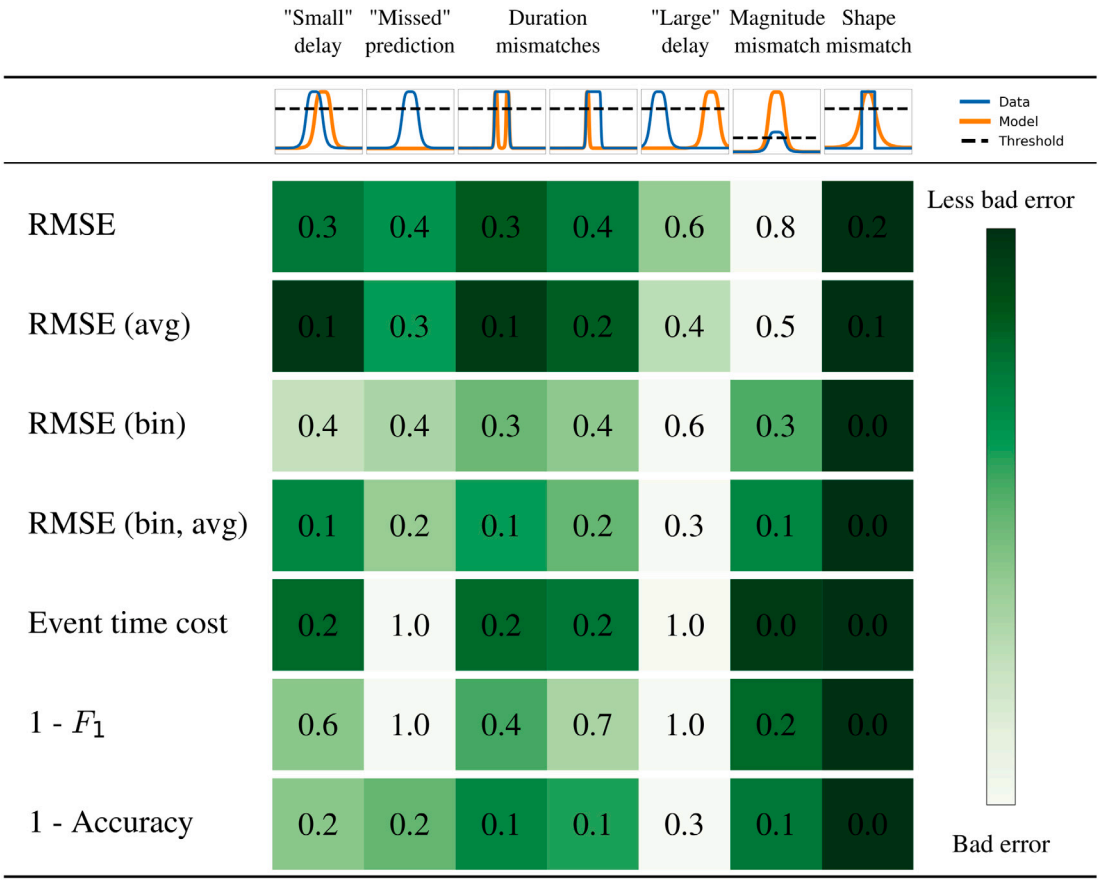


Fig. 2. Values of error measures for various types of characteristic mismatches between data and modelled output values. Blue lines indicate perfect, time-continuous measured data, and orange lines indicate modelled values. A threshold (dashed line) indicates the point where an ecosystem attribute changes from 'absent' to 'present'. The colour scale was applied per error measure, with the mismatch that was considered 'worst' in white and the mismatch considered 'least bad' in dark green. Note that the values can only be compared within rows and not between rows. Different colours for the same values in the same row are due to rounding. RMSE = Root Mean Squared Error, avg = calculated on moving average of data and model values, bin = calculated on binarised data and model values (bloom or no bloom).

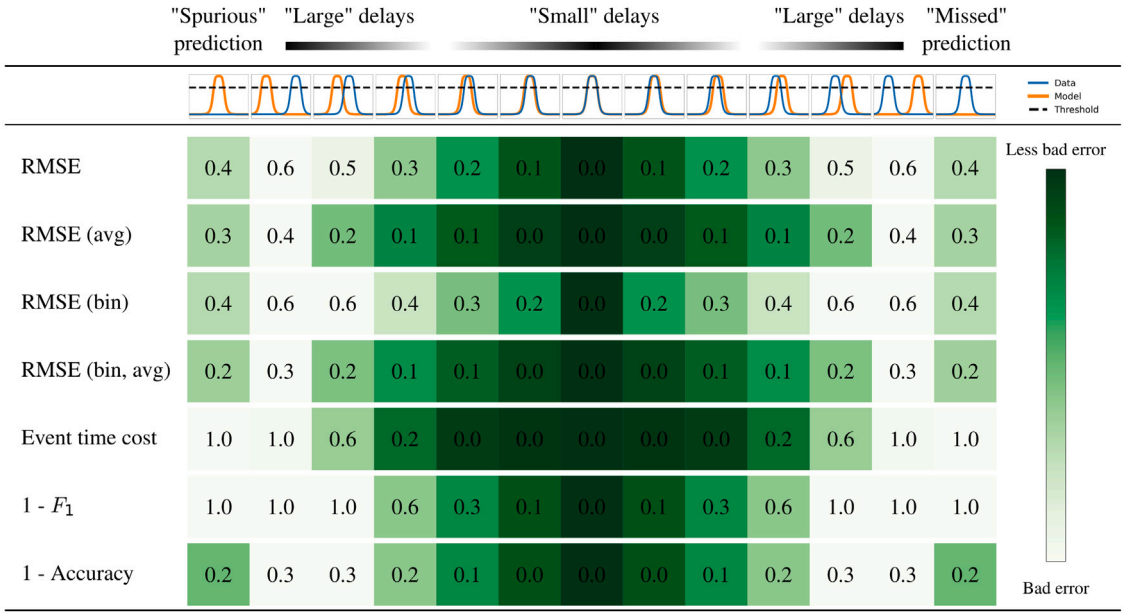


Fig. 3. Values of error measures for a range of delays between peaks in data and modelled output values. Blue lines indicate perfect, time-continuous measured data, and orange lines indicate modelled values. A threshold (dashed line) indicates the point where an ecosystem attribute changes from 'absent' to 'present'. The colour scale was applied per error measure, with the mismatch that was considered 'worst' in white and the mismatch considered 'least bad' in dark green. Note that the values can only be compared within rows and not between rows. Different colours for the same values in the same row are due to rounding. RMSE = Root Mean Squared Error, avg = calculated on moving average of data and model values, bin = calculated on binarised data and model values (presence or absence).

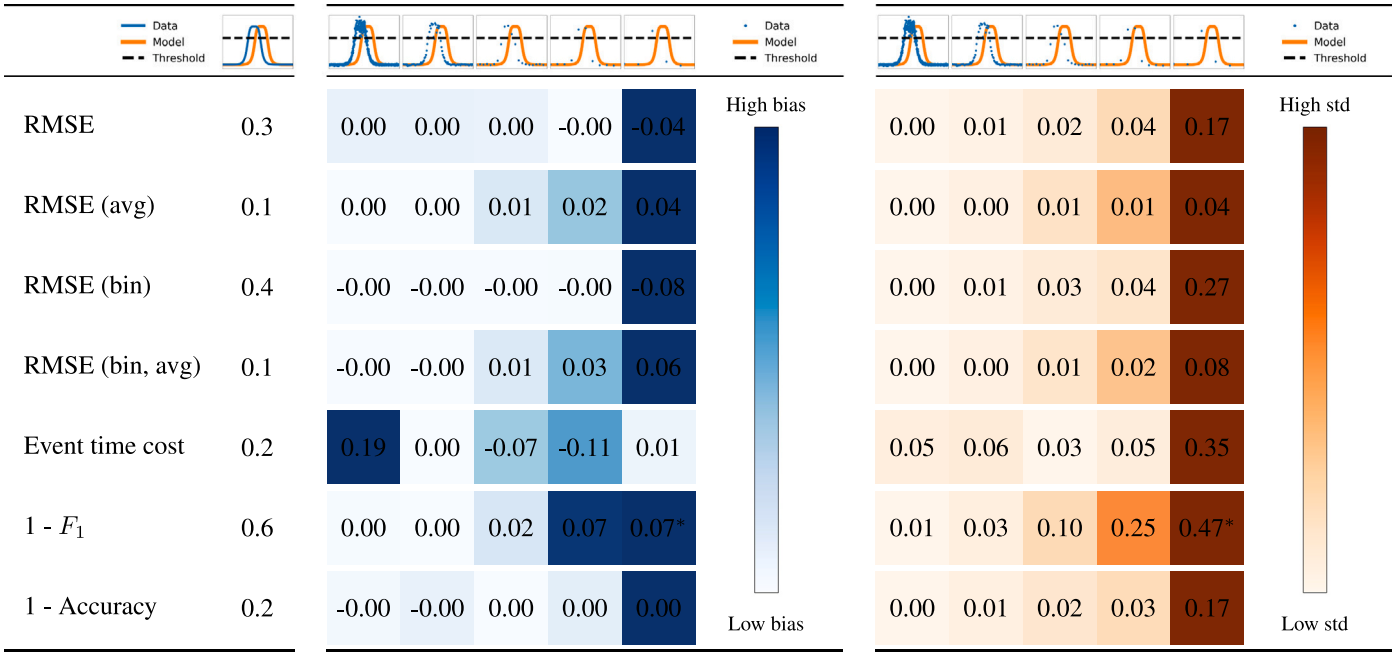


Fig. 4. Bias and consistency of error measures under sparse and noisy data for a small delay in the event prediction. Left table shows values of the error measures for perfect data as a reference. Middle table shows bias of the error measure for different data frequencies, defined as the difference between the error for perfect data and the average error measure over 1000 samples with a given data frequency. Right table shows the standard deviation (std) of the error measure calculated over 1000 samples for different data frequencies, as a measure of the consistency at that data frequency. Note that the values can only be compared within rows and not between rows. Data frequencies used are (from left to right): 1000, 100, 20, 10, and 5 data points. The colour scale was applied per error measure, with biases coloured by their absolute values. Different colours for the same values in the same row are due to rounding. Stars indicate that samples were left out because no error measure could be calculated (i.e., only true negatives for F_1 score).

3.2. Robustness to noise and data frequency varies between error measures

Not all error measures under consideration were equally robust to sparse and noisy data. In particular the event time error measure was highly sensitive to noise, especially with high frequency data, since any random deviation crossing the threshold value would create spurious events (Fig. 4 and Supplemental Figs. S.3–S.8). This sensitivity may be reduced by noise filtering prior to calling events. For sparse data, however, events can easily be missed, resulting in inflated errors. Similarly, F_1-scores cannot be determined when there are only true negatives (Eq. (5)), making them less reliable when events were missed entirely. While other measures are not as sensitive to missing events, evaluating a model's ability to predict events may be better done with high-frequent data that is likely to capture all events anyway.

Of all the measures tested, the RMSE and overall accuracy appeared to be the most robust against noise and sparse data (Fig. 4 and Supplemental Figs. S.3–S.8), as these measures treat all points similarly, even when the available points are not very informative.

3.3. Probabilistic models

The measures that can be extended to the probabilistic case behave highly similarly to their deterministic counterparts (Supplemental Fig. S.9). The continuous ranked probability score (CRPS) matches closely to the mean absolute error (MAE), which behaves like the RMSE, but is relatively less sensitive to large absolute differences. Likewise, measures based on threshold exceedance behave similarly, whether they are based on binary values or probabilities. The log score, on the other hand, behaves very differently, as it creates extremely large error scores at time points where the observations are far outside the predicted range. In our examples with sudden peaks that can be missed, this effect is particularly large. For binarised data, logarithms of numerical zero probabilities even created numerical infinities.

Box 4. Relating error measures to questions on cyanobacterial blooms

When asking the question what the water quality will be at any given time (question 1), RMSE-like methods are clearly preferable, as these respond to bloom magnitude. However, when asking if it is safe to go swimming tomorrow (question 2), the safety threshold from the relevant swim water protocol becomes important, making methods based on a binary 'safe' or 'unsafe' more appropriate. Finally, when the goal of the model is to aid a lake manager asking when to place and remove warning signs (question 3), event-time based methods are preferable, though care should be taken that the data is of sufficient quality to avoid the bias that noisy data can introduce. This may prove unfeasible, because accurate high-frequency data on cyanobacterial densities can be difficult to obtain. Therefore, a moving window average of a (binarised) RMSE, which also allows some control over the acceptable timing error, could be considered as an alternative.

The strong differences between these types of error measures emphasise the importance of carefully considering which question a model for bloom prediction is supposed to answer before attempting to evaluate the model's quality.

4. Discussion

Our results show that the choice of error measure can make a strong difference in how a model is appraised. When evaluating a model,

we should therefore take care to select the most appropriate one for the question that we are asking. Ideally, a rigorous definition of the question that the model aims to answer would translate directly to a definition of an error measure, guaranteeing its appropriateness. In practice, however, goals are often hard to define with this degree of mathematical rigour. Goals may be subject to new insights, e.g., one may initially think that they are after an absolute concentration at every time point, only to realise later that their actual interest is subtly different. Studying the response of various error measures to the kind of characteristic mismatches we show here could help in choosing a reasonably appropriate measure beforehand, thereby avoiding the pitfall of cherry-picking an error measure after one has already calculated it for their actual model and data. The three types of questions considered here, with associated mismatch types, can serve as an example.

The first question related entirely to the value of an ecological quantity at any given time point. The threshold value Y can be ignored for this case, as it is irrelevant. The RMSE and related measures were designed for evaluating models with respect to this type of question. For choosing one of the many specific RMSE-related measures, we refer to the extensive comparisons of these measures made by others (Jackson et al., 2019; Bennett et al., 2013; Koutsandreas et al., 2022; Mehdiyev et al., 2016; Chen et al., 2017; Hyndman and Koehler, 2006; Morley et al., 2018). If the goal is to predict the precise quantity, whereas the timing does not need to be precise, some leniency in timing error is provided by using time-averaging. This is conceptually similar to looking at an RMSE-like metric for different time-lags in one of the time series as proposed for certain hydrological models (Jackson et al., 2019), but with the advantage that it provides a single number to compare.

When the question relates to an ecological attribute that is present when an ecologically relevant threshold value is exceeded (question 2), it becomes important to take this threshold into account. For answering this type of question, classification metrics and an RMSE on the binarised values perform similarly. The F_1-score does tend to max out earlier when there are no true positives, though arguably the distinction between a 'very bad' model and a 'terrible' one is usually not the most relevant anyway. As before, time-averaging can be employed to increase the tolerance for small delays.

When the question of interest is the timing of events where a qualitative ecological attribute changes (question 3), the event time cost becomes the natural choice. This measure is most sensitive to completely incorrectly estimated start and end times. It can also be tuned to be more or less forgiving of small delays. However, it responds rather poorly to the presence of noise. This lack of robustness is the result of 'fake' events appearing due to noise when the data is around the threshold level. This problem may be addressed by first smoothing the data if sufficient data is available. When data is sparse enough that events may be missed completely, this method loses its usefulness, though arguably such sparse data are themselves not appropriate when evaluating a model's ability to predict sudden events. Applying the RMSE to time-averaged binarised data may be considered as an imperfect alternative when data of sufficiently high frequency cannot be obtained.

The high robustness of the RMSE may in part result from the use of normally distributed noise, which makes the RMSE a good choice compared to other point-by-point value comparing metrics, such as the MAE (Hodson, 2022). However, the overall accuracy showed a similar robustness, so it need not be avoided in cases where it is more appropriate. Still, the relative robustness of the RMSE may in part, explain the popularity of applying the RMSE in cases where it is less appropriate for the question being asked.

Many error measures can be extended for the evaluation of probabilistic models (Simonis et al., 2021). In our examples, these error measures behaved similarly to their deterministic counterparts when compared across different types of mismatches. Some probabilistic measures, have no deterministic counterpart and behave differently. The log score, for example, inherently shows an extreme response to observations that fall far outside the predicted probability distribution. Such behaviour is probably unpractical for evaluating models of systems with sudden peaks that could be missed, like the ones in our examples. For other measures, like the event time error, there may not be obvious extensions to probabilistic models. When the goal is to estimate the timing of events to, for example, place warning signs on time, a probabilistic output must at some point be converted to a "yes or no" decision, as it makes little sense to place half a sign. This decision could be based on the average, but also, for example, on the ninetieth percentile, to be on the safe side. The error measure, then, should also be set less sensitive to placing signs too early than to placing them too late.

Of course, many other formulations of error measures are possible beyond the few that we considered. Even among the established RMSE-related measures (Jackson et al., 2019; Bennett et al., 2013; Koutsandreas et al., 2022; Mehdiyev et al., 2016; Chen et al., 2017; Hyndman and Koehler, 2006; Morley et al., 2018) and classification measures (Bennett et al., 2013; Sokolova and Lapalme, 2009; Mehdiyev et al., 2016), there are a great many options to choose from. When it remains unclear exactly which measure is most appropriate, it may be useful to select several reasonable ones, to test if there is a model that performs best on all of them.

In practice, existing models are often re-used to answer new questions. Indeed, availability, usability, and existing expertise can be important factors in the decision to use a specific model (Melsen, 2022; Hamilton et al., 2022). Larger models generally are developed over longer periods of time, with the explicit purpose of answering a variety of questions (for an example on cyanobacterial blooms see those listed in Janssen et al. (2019a)). Just like the performance of such a model should be re-evaluated when applied at a different location (with the same error measure), it should also be re-evaluated when answering a different type of question, using a different error measure (i.e., one that is appropriate for the new question).

The example of cyanobacterial bloom modelling presented in boxes 2–4 can be seen as exemplary for (management) questions about events in (socio-)ecological or (bio-)physical systems for which general descriptive models are developed. Other examples could include the forecast of epidemiological outbreaks, wildfires or river discharge (e.g., Dietze et al. (2018) and van Kempen et al. (2021)). Specific questions in these cases could focus on the total discharge during an extreme rainfall event (in line with question 1), the classification of the risk posed by an infection (in line with questions 2), or the expected moment of the outbreak of wildfires (in line with questions 3). As illustrated here, in each of these cases different mismatches have a different effect on the evaluation of the model prediction. The choice of error measure should reflect this. These choices are to be made in each specific situation by people knowledgeable of the modelled systems and the implication of specific research or management questions.

As a final note, error measures are used not only for model validation, but also for model calibration, e.g., parameter estimation for process-based models, or training machine learning models and AIs. As with model validation, these error measures should reflect the question we want to ask these models, or we may inadvertently fit the models to answer questions we did not mean to ask.

Box 5. Key points

- When evaluating a model, the chosen error measure(s) should be appropriate for the kind of question being asked.
- For questions about the value of an ecological quantity, measures that compare modelled and measured values at each time point (such as the RMSE) are suitable.
- For questions about the qualitative state of an ecosystem attribute, classification metrics may be more appropriate.
- For questions about the timing of changes in the state of an ecosystem attribute, event-time related error measures are more appropriate, though a moving average of metrics that consider each time-point could be a reasonable compromise.
- Given the reality of sparse and noisy data, there can sometimes be trade-offs between the appropriateness and the robustness of an error measure.

5. Conclusion

When selecting an error measure to evaluate a model, it is important to consider the question being asked. Different error measures may be more or less appropriate for different questions and they may yield vastly different appraisals for how well a model describes the data. What constitutes a good description of the data can only be meaningfully defined in relation to a specific question. Therefore, a well-defined (ecological) question is always required to be able to determine if a model is a good description of the observations and should be reported as part of model validation.

CRediT authorship contribution statement

Bas Jacobs: Conceptualization, Methodology, Formal analysis, Investigation, Visualisation, Writing – original draft, Writing – review & editing. **Hilde Tobi:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Geerten M. Hengeveld:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

Python scripts used to generate the results are provided in the supplementary materials.

Acknowledgements

We gratefully acknowledge useful discussions with Miguel Dionisio Pires and Tineke Troost from Deltares, as well as Sven Teurlincx, and Lilith Kramer from NIOO-KNAW. B.J. was funded by project 645.002.002 in the NWO-complexity program. The authors also acknowledge support from the Biometris investment projects.

Appendix A. Construction of characteristic mismatches

The peaks in the characteristic mismatch examples (Fig. 1) were constructed from mirrored hill functions of the type:

$$y = y_{\text{base}} + A\frac{(t - t_{\text{start}})^n}{(t - t_{\text{start}})^n + K^n}, \tag{A.1}$$

where y is the ecological output variable, y_{base} is a background value of y, A is the maximum size of the peak above background level, t_{start} is the time point at which the hill function begins, and K and n are parameters that determine the shape of the hill function. Outside the peaks, the y is kept at y_{base} and where the peaks form plateaus y is kept constant. The default values are $y_{\text{base}} = 0.1$, $A = 1$, and 10 for the total time duration (all in arbitrary units). The critical threshold value Y is set at 0.8.

Appendix B. Derivation of event time cost functions

If we want to issue a warning for events where an ecological attribute becomes 'present' (start events) and remove it for events where it becomes 'absent' (end events), then there are four types of error that could in principle each carry its own cost C_{ij} as a function of the relevant timing difference. For $i = 1$ (issuing a warning at a wrong time), this difference is between the time of the warning ($\hat{t}_{1j}$) and the closest time point where a start event was measured (t_{1j}). For $i = 2$ (removing a warning at a wrong time), this difference is between the time of removing the warning ($\hat{t}_{2j}$) and the closest time point where an end event was measured (t_{2j}). For $i = 3$ (failing to issue a warning at the right time), this difference is between the time where a start event is measured (t_{3j}) and the closest time point where a warning would be issued ($\hat{t}_{3j}$). For $i = 4$ (failing to remove a warning at the right time), this difference is between the time where an end event is measured (t_{4j}) and the closest time point where a warning would be removed ($\hat{t}_{4j}$).

If the error in timing becomes large enough, the cost should become similar to that of a missed event since at some point, someone is likely to notice the missed event and a warning will be issued or removed anyway. Therefore, the cost function should saturate to a maximum cost $C_{m,i}$, which may depend on the type of error i. For this reason, we chose a generic hill function to describe the relation between the cost and the time difference. Assuming a symmetric cost function, i.e., equal costs for being too early and too late, this means:

$$C_{ij} = C_{m,i}\frac{|t_{ij} - \hat{t}_{ij}|^n}{|t_{ij} - \hat{t}_{ij}|^n + t^n_{\frac{1}{2},i}}, \tag{B.1}$$

where $t_{\frac{1}{2},i}$ is the time error at which the cost of type i is half its maximum and n is the hill coefficient that determines the shape of the slope. The value of $t_{\frac{1}{2},i}$ should be chosen such that it represents a time error that is starting to become costly. If small timing errors are to be tolerated, it makes sense to choose a value of $n > 1$, which gives the hill function more of a step.

We can then add up all of the costs into a single error measure:

$$\text{Total cost} = \sum_{i=1}^{4}\sum_{j=1}^{N_i} C_{ij} \tag{B.2}$$

If we take the simplest case where $C_{m,i} = C_m$ and $t_{\frac{1}{2},i} = t_{\frac{1}{2}}$ for all i, and normalise by the maximum cost per error C_m and the total number of error cases, we obtain Eq. (6).

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecolmodel.2023.110562.

References

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. http://dx.doi.org/10.1016/j.envsoft.2012.09.011.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78 (1), 1–3.

Burford, M., Carey, C., Hamilton, D., Huisman, J., Paerl, H., Wood, S., Wulff, A., 2020. Perspective: Advancing the research agenda for improving understanding of cyanobacteria in a future of global change. Harmful Algae 91, 101601. http://dx.doi.org/10.1016/j.hal.2019.04.004.

Carey, C.C., Woelmer, W.M., Lofton, M.E., Figueiredo, R.J., Bookout, B.J., Corrigan, R.S., Daneshmand, V., Hounshell, A.G., Howard, D.W., Lewis, A.S.L., McClure, R.P., Wander, H.L., Ward, N.K., Thomas, R.Q., 2022. Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. Inland Waters 12 (1), 107–120. http://dx.doi.org/10.1080/20442041.2020.1816421.

Chen, C.A., Twycross, J.A., Garibaldi, J.M., 2017. A new accuracy measure based on bounded relative error for time series forecasting. PLoS One 12 (3), 1–23. http://dx.doi.org/10.1371/journal.pone.0174202.

Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J.M., Tang, G., Gharari, S., Freer, J.E., Whitfield, P.H., Shook, K.R., Papalexiou, S.M., 2021. The abuse of popular performance metrics in hydrologic modeling. Water Resour. Res. 57 (9), e2020WR029001. http://dx.doi.org/10.1029/2020WR029001.

Dietze, M.C., 2017. Prediction in ecology: a first-principles framework. Ecol. Appl. 27 (7), 2048–2060. http://dx.doi.org/10.1002/eap.1589.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., Keitt, T.H., Kenney, M.A., Laney, C.M., Larsen, L.G., Loescher, H.W., Lunch, C.K., Pijanowski, B.C., Randerson, J.T., Read, E.K., Tredennick, A.T., Vargas, R., Weathers, K.C., White, E.P., 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proc. Natl. Acad. Sci. 115 (7), 1424–1432. http://dx.doi.org/10.1073/pnas.1710231115.

Gleckler, P.J., Taylor, K.E., Doutriaux, C., 2008. Performance metrics for climate models. J. Geophys. Res.: Atmos. 113 (D6), http://dx.doi.org/10.1029/2007JD008972.

Hamilton, S.H., Pollino, C.A., Stratford, D.S., Fu, B., Jakeman, A.J., 2022. Fit-for-purpose environmental modeling: Targeting the intersection of usability, reliability and feasibility. Environ. Model. Softw. 148, 105278. http://dx.doi.org/10.1016/j.envsoft.2021.105278.

He, X., Liu, Y.-L., Conklin, A., Westrick, J., Weavers, L.K., Dionysiou, D.D., Lenhart, J.J., Mouser, P.J., Szlag, D., Walker, H.W., 2016. Toxic cyanobacteria and drinking water: Impacts, detection, and treatment. Harmful Algae 54, 174–193. http://dx.doi.org/10.1016/j.hal.2016.01.001.

Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. 15 (14), 5481–5487. http://dx.doi.org/10.5194/gmd-15-5481-2022.

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M.H., Visser, P.M., 2018. Cyanobacterial blooms. Nat. Rev. Microbiol. 16 (8), 471–483. http://dx.doi.org/10.1038/s41579-018-0040-1.

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

Ibelings, B.W., Vonk, M., Los, H.F.J., van der Molen, D.T., Mooij, W.M., 2003. Fuzzy modeling of cyanobacterial surface waterblooms: Validation with noaa-avhrr satellite images. Ecol. Appl. 13 (5), 1456–1472. http://dx.doi.org/10.1890/01-5345.

Jackson, E.K., Roberts, W., Nelsen, B., Williams, G.P., Nelson, E.J., Ames, D.P., 2019. Introductory overview: Error metrics for hydrologic modelling – a review of common practices and an open source library to facilitate use and adoption. Environ. Model. Softw. 119, 32–48. http://dx.doi.org/10.1016/j.envsoft.2019.05.001.

Jakeman, A., Letcher, R., Norton, J., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21 (5), 602–614. http://dx.doi.org/10.1016/j.envsoft.2006.01.004.

Janse, J.H., van Liere, L., 1995. PCLake: A modelling tool for the evaluation of lake restoration scenarios. Water Sci. Technol. 31 (8), 371–374. http://dx.doi.org/10.1016/0273-1223(95)00392-Z.

Janssen, A.B., Janse, J.H., Beusen, A.H., Chang, M., Harrison, J.A., Huttunen, I., Kong, X., Rost, J., Teurlincx, S., Troost, T.A., van Wijk, D., Mooij, W.M., 2019a. How to model algal blooms in any lake on earth. Curr. Opin. Environ. Sustain. 36, 1–10. http://dx.doi.org/10.1016/j.cosust.2018.09.001.

Janssen, A.B., Teurlincx, S., Beusen, A.H., Huijbregts, M.A., Rost, J., Schipper, A.M., Seelen, L.M., Mooij, W.M., Janse, J.H., 2019b. PCLake+: A process-based ecological model to assess the trophic state of stratified and non-stratified freshwater lakes worldwide. Ecol. Model. 396, 23–32. http://dx.doi.org/10.1016/j.ecolmodel.2019.01.006.

Korppoo, M., Huttunen, M., Huttunen, I., Piirainen, V., Vehviläinen, B., 2017. Simulation of bioavailable phosphorus and nitrogen loading in an agricultural river basin in Finland using VEMALA v.3. J. Hydrol. 549, 363–373. http://dx.doi.org/10.1016/j.jhydrol.2017.03.050.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., Assimakopoulos, V., 2022. On the selection of forecasting accuracy measures. J. Oper. Res. Soc. 73 (5), 937–954. http://dx.doi.org/10.1080/01605682.2021.1892464.

Lewis, A.S.L., Rollinson, C.R., Allyn, A.J., Ashander, J., Brodie, S., Brookson, C.B., Collins, E., Dietze, M.C., Gallinat, A.S., Juvigny-Khenafou, N., Koren, G., McGlinn, D.J., Moustahfid, H., Peters, J.A., Record, N.R., Robbins, C.J., Tonkin, J., Wardle, G.M., 2023. The power of forecasts to advance ecological theory. Methods Ecol. Evol. 14 (3), 746–756. http://dx.doi.org/10.1111/2041-210X.13955.

Lewis, A.S.L., Woelmer, W.M., Wander, H.L., Howard, D.W., Smith, J.W., McClure, R.P., Lofton, M.E., Hammond, N.W., Corrigan, R.S., Thomas, R.Q., Carey, C.C., 2022. Increased adoption of best practices in ecological forecasting enables comparisons of forecastability. Ecol. Appl. 32 (2), e2500. http://dx.doi.org/10.1002/eap.2500.

Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J.S., Schimel, D.S., 2011. Ecological forecasting and data assimilation in a data-rich era. Ecol. Appl. 21 (5), 1429–1442. http://dx.doi.org/10.1890/09-1275.1.

Lürling, M., Mucci, M., 2020. Mitigating eutrophication nuisance: in-lake measures are becoming inevitable in eutrophic waters in the netherlands. Hydrobiologia 847 (21), 4447–4467. http://dx.doi.org/10.1007/s10750-020-04297-9.

Mehdiyev, N., Enke, D., Fettke, P., Loos, P., 2016. Evaluating forecasting methods by considering different accuracy measures. Procedia Comput. Sci. 95, 264–271. http://dx.doi.org/10.1016/j.procs.2016.09.332.

Melsen, L.A., 2022. It takes a village to run a model — The social practices of hydrological modeling. Water Resour. Res. 58 (2), e2021WR030600. http://dx.doi.org/10.1029/2021WR030600.

Morley, S.K., Brito, T.V., Welling, D.T., 2018. Measures of model performance based on the log accuracy ratio. Space Weather 16 (1), 69–88. http://dx.doi.org/10.1002/2017SW001669.

Paerl, H.W., Huisman, J., 2008. Blooms like it hot. Science 320 (5872), 57–58. http://dx.doi.org/10.1126/science.1155398.

Paerl, H.W., Otten, T.G., 2013. Harmful cyanobacterial blooms: Causes, consequences, and controls. Microb. Ecol. 65 (4), 995–1010. http://dx.doi.org/10.1007/s00248-012-0159-y.

Page, T., Smith, P.J., Beven, K.J., Jones, I.D., Elliott, J.A., Maberly, S.C., Mackay, E.B., De Ville, M., Feuchtmayr, H., 2018. Adaptive forecasting of phytoplankton communities. Water Res. 134, 74–85. http://dx.doi.org/10.1016/j.watres.2018.01.046.

Parker, W.S., 2020. Model evaluation: An adequacy-for-purpose view. Philos. Sci. 87 (3), 457—477. http://dx.doi.org/10.1086/708691.

Payne, M.R., Hobday, A.J., MacKenzie, B.R., Tommasi, D., Dempsey, D.P., Fässler, S.M.M., Haynie, A.C., Ji, R., Liu, G., Lynch, P.D., Matei, D., Miesner, A.K., Mills, K.E., Strand, K.O., Villarino, E., 2017. Lessons from the first generation of marine ecological forecast products. Front. Mar. Sci. 4, 289. http://dx.doi.org/10.3389/fmars.2017.00289.

Petrovskii, S., Petrovskaya, N., 2012. Computational ecology as an emerging science. Interface Focus 2 (2), 241–254. http://dx.doi.org/10.1098/rsfs.2011.0083.

Recknagel, F., Cetin, L., Zhang, B., 2008. Process-based simulation library SALMO-OO for lake ecosystems. Part 1: Object-oriented implementation and validation. Ecol. Inform. 3 (2), 170–180. http://dx.doi.org/10.1016/j.ecoinf.2008.04.002.

Rousso, B.Z., Bertone, E., Stewart, R., Hamilton, D.P., 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Res. 182, 115959. http://dx.doi.org/10.1016/j.watres.2020.115959.

Saloranta, T.M., Andersen, T., 2007. MyLake—A multi-year lake simulation model code suitable for uncertainty and sensitivity analysis simulations. Ecol. Model. 207 (1), 45–60. http://dx.doi.org/10.1016/j.ecolmodel.2007.03.018.

Schets, F., van der Oost, R., van de Waal, D., Lammertink, M., Slot, D., van Druten, G., 2020. Blauwalgenprotocol 2020. http://dx.doi.org/10.21945/RIVM-2020-0107.

Simonis, J.L., White, E.P., Ernest, S.K.M., 2021. Evaluating probabilistic ecological forecasts. Ecology 102 (8), e03431. http://dx.doi.org/10.1002/ecy.3431.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45 (4), 427–437. http://dx.doi.org/10.1016/j.ipm.2009.03.002.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 62 (1), 77–89. http://dx.doi.org/10.1016/S0034-4257(97)00083-7.

Taylor, J.W., Yu, K., 2016. Using auto-regressive logit models to forecast the exceedance probability for financial risk management. J. R. Stat. Soc. A (Stat. Soc.) 179 (4), 1069–1092, URL http://www.jstor.org/stable/44682197.

Trolle, D., Elliott, J.A., Mooij, W.M., Janse, J.H., Bolding, K., Hamilton, D.P., Jeppesen, E., 2014. Advancing projections of phytoplankton responses to climate change through ensemble modelling. Environ. Model. Softw. 61, 371–379. http://dx.doi.org/10.1016/j.envsoft.2014.01.032.

van Basshuysen, P., 2023. Austinian model evaluation. Philos. Sci. 1–14. http://dx.doi.org/10.1017/psa.2023.24.

van Kempen, G., van der Wiel, K., Melsen, L.A., 2021. The impact of hydrological model structure on the simulation of extreme runoff events. Nat. Hazards Earth Syst. Sci. 21 (3), 961–976. http://dx.doi.org/10.5194/nhess-21-961-2021.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. vol. 100, Academic Press.

Woelmer, W.M., Thomas, R.Q., Lofton, M.E., McClure, R.P., Wander, H.L., Carey, C.C., 2022. Near-term phytoplankton forecasts reveal the effects of model time step and forecast horizon on predictability. Ecol. Appl. 32 (7), e2642. http://dx.doi.org/10.1002/eap.2642.