Ecological Applications, 28(5), 2018, pp. 1260–1272

Dangerous relationships: biases in freshwater bioassessment based on observed to expected ratios

Heikki Hämäläinen,[1,5] Jukka Aroviita,[2] Jussi Jyväsjärvi,[3] and Salme Kärkkäinen[4]

[1]*Department of Biological and Environmental Sciences, University of Jyvaskyla, P.O. Box 35, Jyväskylä FI-40014 Finland*
[2]*Finnish Environment Institute, Freshwater Centre, PO Box 413, Oulu 90014 Finland*
[3]*Department of Ecology and Genetics, University of Oulu, P.O. Box 3000, Oulu FI-90014 Finland*
[4]*Department of Mathematics and Statistics, University of Jyvaskyla, P.O. Box 35, Jyväskylä FI-40014 Finland*

Abstract. The ecological assessment of freshwaters is currently primarily based on biological communities and the reference condition approach (RCA). In the RCA, the communities in streams and lakes disturbed by humans are compared with communities in reference conditions with no or minimal anthropogenic influence. The currently favored rationale is using selected community metrics for which the expected values (E) for each site are typically estimated from environmental variables using a predictive model based on the reference data. The proportional differences between the observed values (O) and E are then derived, and the decision rules for status assessment are based on fixed (typically 10th or 25th) percentiles of the O/E ratios among reference sites. Based on mathematical formulations, illustrations by simulated data and real case studies representing such an assessment approach, we demonstrate that the use of a common quantile of O/E ratios will, under certain conditions, cause severe bias in decision making even if the predictive model would be unbiased. This is because the variance of O/E under these conditions, which seem to be quite common among the published applications, varies systematically with E. We propose a correction method for the bias and compare the novel approach to the conventional one in our case studies, with data from both reference and impacted sites. The results highlight a conceptual issue of employing ratios in the status assessment. In some cases using the absolute deviations instead provides a simple solution for the bias identified and might also be more ecologically relevant and defensible.

Key words: bioassessment; classification error; ecological status; freshwaters; predictive models; reference condition approach.

Introduction

Modern bioassessment and monitoring of aquatic ecosystems, partially stipulated by legislation (e.g., European Commission 2000), is increasingly grounded on the Reference Condition Approach (RCA; Bailey et al. 2004). In the RCA, biotic communities of ecosystems disturbed by humans are compared with a range of communities expected in similar ecosystems undisturbed or minimally disturbed by humans. In practice, the assessments are, again partially driven by legislative guidelines, based on selected metrics typically measuring structural characteristics of particular communities or assemblages of organisms. An ecosystem is considered impacted, and its ecological quality deteriorated, if the observed community metric values are not within the estimated natural variation or fall beyond a specified range of values among reference systems.

Reliable estimation of the expected (reference) values and their variability is an essential requirement for the RCA. There are several associated problems, some of which are conceptual and others more technical in nature (Bowman and Somers 2005, Stoddard et al. 2006, Hawkins et al. 2010*a*, Cao and Hawkins 2011, Nichols and Dyer 2013). In this contribution, we will draw attention to an apparently previously neglected technical problem, which is inherent in RCA assessments using ratios of observed (O) to expected (E) values of metrics or relative differences for measuring the condition of biota. The problem can potentially lead to severe bias in decision making and hence losses of either economic or natural resources in the management of individual ecosystems and at a more global scale.

The use of ratios for measuring deviance between observed and expected values is explicitly required by European legislation (European Commission 2000) and is integral in the now globally widely used assessments of proportion of observed to expected species (Clarke et al. 1996) or "taxonomic completeness" as a measure of biotic integrity of freshwaters (Hawkins 2006). Typically the expected values for metrics are estimated by models based on data from a representative sample of reference sites and using environmental predictor variables that are not affected by anthropogenic activities, thereby allowing estimation of E for sites that are subject to human disturbance and that differ in their natural characteristics (e.g., Clarke et al. 1996, Hawkins 2006). A predictive model for a continuous variable can be partially validated or checked for bias by regressing (preferably a set of new and independent) observed values (O) on the predicted values (here E) as follows (e.g., Mayer and Butler 1993)

$$O = \alpha + \beta E + \varepsilon$$

An ideal model should yield unbiased predictions (E) of values observed (O) in undisturbed conditions ($\beta = 1$, $\alpha = 0$) and a random error (ε) independent of E:

Manuscript received 4 July 2017; revised 23 February 2018; accepted 16 March 2018. Corresponding Editor: Song S. Qian.
[5] E-mail: heikki.o.hamalainen@jyu.fi

$$O = E + \varepsilon$$

This outcome (assuming reasonable ε, encompassing observation error) should be fully acceptable. When applied to new cases representing ecosystems subject to human disturbance, deviation of observed values O from E beyond some percentile of distribution of ε could hence be justifiably considered a human effect. Indeed, Linke et al. (2005) used O on E regression to validate predictive models for assessing taxonomic completeness, and proposed acceptability ranges of 1.00 ± 0.15 and 0.00 ± 1.5 for the slope (β) and intercept (α), respectively. These criteria might well be reasonable for this particular case, although the exact ranges proposed were by no means justified and seem arbitrary. Even though these specific benchmarks have been used or referred to by several authors since then (e.g., Chessman et al. 2008, Feio et al. 2014a, b, Rose et al. 2016a, b), or the approach applied apparently independently (e.g., Hargett et al. 2007), the common rationale still has been largely to neglect the actual O on E relationship and to relying mainly on the overall mean (bias) and variance or standard deviation (random error) of the O/E ratio in model validation (e.g., Clarke et al. 1996, Aroviita et al. 2009).

Nevertheless, even if a model would fully satisfy the criteria suggested by Linke et al. (2005), or if the overall mean O/E ratio would be close to 1, or both, there remains a risk of bias, which stems from using proportions (O/E) instead of actual ε in decision making. For an ideal model (the latter equation) characterized above (Fig. 1a), average O to E ratio is (necessarily and as is desirable) constant ($O/E = 1$) across the full range of E. However, in the case of constant ε, dispersion of proportional error ($\varepsilon_{O/E}$) is no longer constant, but instead decreases with increasing E (Fig. 1b). This is simply because ε is proportionally smaller relative to E (and O) with increasing E. As a consequence, the present rationale of using a fixed (e.g., 10th or 25th) percentile of the overall $\varepsilon_{O/E}$ distribution as a common threshold independent of E (e.g., Clarke et al. 1996, Aroviita et al. 2010), would lead to a systematic error or bias in detecting deviance from natural variation. Sites that have naturally small metric values (small E) would be more likely than sites with large E to be (incorrectly) judged impaired, whereas for sites with large E, a deviation of O from E greater than the actual reference variation would be required to indicate impact (Fig. 1c). This bias would be fully avoided only if the relative error or O/E was constant for all sites, independent of E. This in turn would require variance of O given E to increase proportionally to E^2, that is, $\text{var}(O) = E^2\sigma^2$. Then the variance of O/E is $\text{var}(O/E) = 1/E^2 \times \text{var}(O) = \sigma^2$ (a constant) (Fig. 2a, b). Deviation from a slope of unity and/or zero intercept in the O on E regression could also lead to error varying with E and hence to differences between local and global error. These biases, alone or in variable combinations, might result in increased risk of mistaken decisions and thereby either to needless management actions and economical losses, or to unnecessary deprivation of natural resources, depending on the case.

In this contribution, we used simulated data together with selected real example data of our own case studies and from the literature to illustrate how the variability of proportional error (O/E) with E and the resulting bias in detecting deterioration vary among models, depending on their properties. We then propose a unifying solution to correct the bias, also encompassing the bias that concerned Linke et al. (2005) when it exists. Using data from sites disturbed by humans, we illustrate the differences between classification outcomes based on the conventional and suggested novel decision rules for each of our sample model. We also draw attention to a conceptual issue related to using ratios instead of actual differences in assessments.

Methods

To demonstrate how the described bias varies with the properties of O on E relationship and to test the suggested novel method, we used simulated data. For the data simulations and all analyses that follow, we used R software (R Core Team 2017). We first generated a realization from a O on E regression model with coefficients $\alpha = 0$ and $\beta = 1$, and a constant variance, $\text{var}(\varepsilon) = \sigma^2 = 3^2$, by using the R function rnorm. We selected the value 3 to represent a typical σ among the published models for the estimation of taxonomic completeness. We assumed that $O \sim N(\alpha + \beta E, \sigma^2)$. In practice, we used a set of fixed E values from the interval [10, 40] with the frequency of 0.50. For each E value, we simulated five observed values (O) from a normal distribution with expectation $\alpha + \beta E = E$ and variance $\text{var}(\varepsilon) = 3^2$, resulting in 305 observations. When the true parameters α, β, and σ^2 are known, as here, a new observation O/E given E as a linear transformation follows the normal distribution $N((\alpha + \beta E)/E, \sigma^2/E)$ (Davison 2003, Hocking 2013). Then, the lower limit of a classical $(1-2p)$ prediction interval, that is the p quantile, can be given by

$$(\alpha + \beta E)/E + z_p \cdot (1/E) \cdot \sqrt{[\text{var}(\varepsilon)]} = 1 + z_p \cdot (1/E) \cdot \sqrt{[\sigma^2]}$$

where z_p is the p quantile of a standard normal distribution (e.g., Hocking 2013, R function qnorm). Here the selected probability p is 0.10 or 0.25, corresponding to 10% and 25% decision curves, respectively. We consider this curve the theoretical (true) p decision curve (Fig. 1b), to serve in testing the correction method we suggest.

Second, to illustrate the (lack of) bias and to test the performance of the method with a different variance structure, we simulated similarly a realization from a regression model with coefficients $\alpha = 0$ and $\beta = 1$, but a variance $\text{var}(\varepsilon) = E^2\sigma^2 = E^2 0.2^2$, where the value 0.2 corresponds to values obtained for real data. The lower limit of a classical prediction interval is now a constant

$$(\alpha + \beta E)/E + z_p \cdot (1/E) \cdot \sqrt{[\text{var}(\varepsilon)]} = 1 + z_p \cdot \sqrt{[\sigma^2]}.$$

For this simulation set-up, the theoretical curves of the novel method are lines at y value 0.744, when $p = 0.10$, and at y value 0.865, when $p = 0.25$, respectively (Fig. 2b).

Furthermore, we used data from five representative case studies (CS) to illustrate the biases and then the effect of the suggested correction. The first data set (CS1) comprises the observed (O) numbers of macroinvertebrate taxa and expected numbers (E) estimated by a multivariate RIVPACS-type (Clarke et al. 1996) model predicting macroinvertebrate fauna

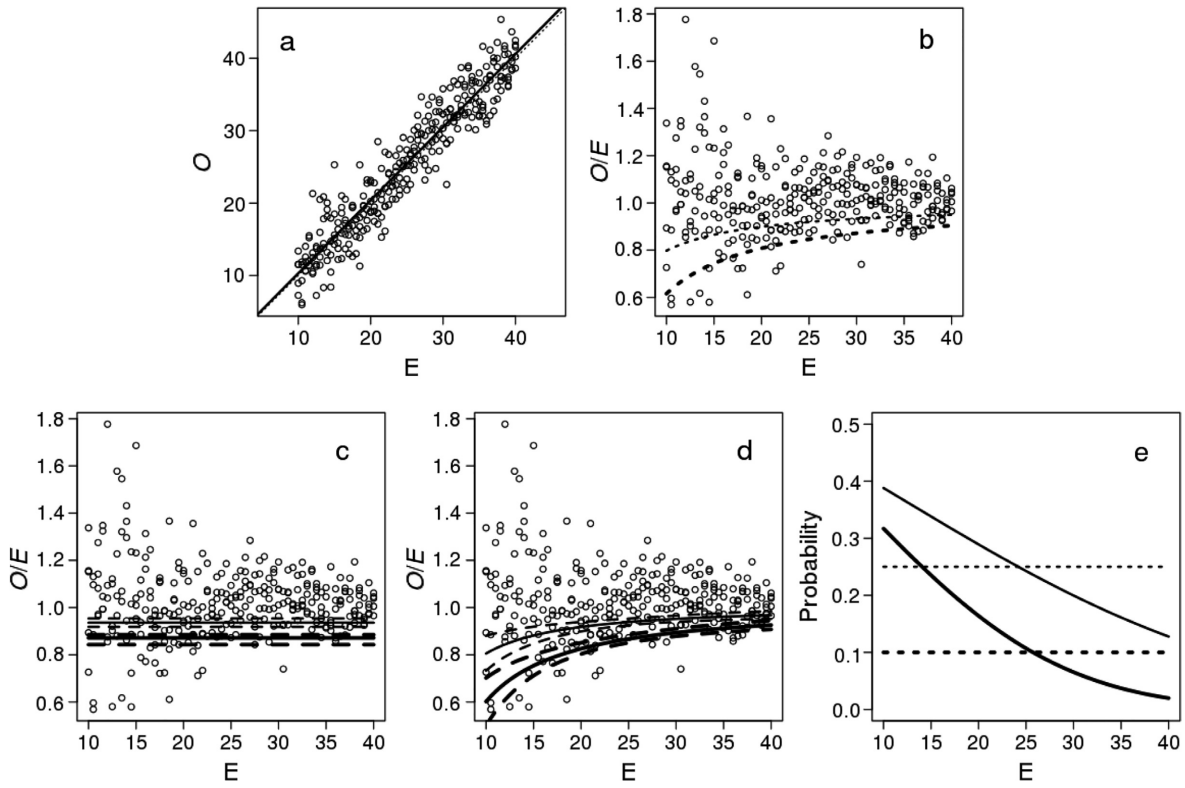

FIG. 1. (a) A simulated realization of a O (observed metric value) on E (expected metric value) regression model with coefficients $\alpha = 0$ and $\beta = 1$, and variance $\text{var}(\varepsilon) = 3^2$, and the estimated regression line with values $a = 0.219$, $b = 1.009$, $\text{var}(\varepsilon) = E^{-0.275}4.558^2$; (b) theoretical 10% (dashed lower) and 25% decision curves (dotted upper); (c,d) the 10% decision line at 0.872 (thick solid) and 25% decision line at 0.936 (thin solid) with 95% confidence intervals (dashed) when conventional method and (d) the estimated 10% and 25% decision curves with 95% confidence intervals when the proposed method are used; and (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

of Finnish streams (Aroviita et al. 2009). Of several alternative models we used the best ranking version (SD of $O/E = 0.17$) based on the greater spatial scale, catchment characters and location as the predictors, and the threshold probability of species capture Pt = 0.4 (see Aroviita et al. [2009] for further details). The data set contains 96 reference (REF) stream sites and 134 non-reference or impacted (IMP) sites, subject to various human disturbances. The second data set (CS2) includes observed values of a lake profundal macroinvertebrate assessment metric PICM_{DCA} and those predicted from sampling (lake maximum) depth by a simple linear regression model (Jyväsjärvi et al. 2014) for 79 Finnish REF lakes and 535 IMP lakes. For illustrative purposes, we polished these original data slightly by removing the most strongly deviating shallow (mean depth < 10 m) lakes (see Jyväsjärvi et al. 2012), retaining 68 and 431 REF and IMP lakes, respectively. The SD of O/E for these reference data is 0.21. The third example (CS3) is similar to CS1, but for littoral invertebrates of Finnish lakes (J. Aroviita, H. Hämäläinen, J. Jyväsjärvi, H. Mykrä, and K.T. Tolonen, *unpublished manuscript*). A RIVPACS type model using a random forest approach (Breiman 2001, Hawkins et al. 2010*a*) based on data from 118 REF lakes (SD of $O/E = 0.16$) was developed to predict fauna (with Pt = 0.5) from climate (annual maximum air temperature) and lake size (volume). There were also data from 142 IMP lakes. Furthermore, we selected two additional typical case studies from the literature on the taxonomic completeness of stream macroinvertebrates, with published statistics and graphs of the O on E regression. Based on the statistics and visual inspection of the graphs, these case studies, CS4 (Tsang et al. 2011) and CS5 (Hargett et al. 2007), represent near to constant and heteroscedastic variance of O relative to E, respectively. They both also had a reasonable number of observations to allow modeling of the variance structure. As we took the data from graphs (fig. 4c in Tsang et al. 2011, fig. 3 in Hargett et al. 2007), some overlying data points are missing, but the calculated statistics differ minimally from the published. These minor differences do not have any influence on the main results and conclusions. More generally, these selected models and their outputs represent realistic assessment applications comparable to others found from the literature, and are used to illustrate generalizable problems in interpreting O/E ratios; and hence their specific details are unimportant in the present context.

The conventional approach to differentiate impacted sites from the reference sites is based on a given (typically 25% or 10%) overall quantile of O/E values among the reference data (Clarke et al. 1996, Kilgour et al. 1998, Aroviita et al. 2010). In that case, and in order to treat all sites equally, the variance of O/E should be constant and independent of the E value. We expect this to be seldom, if ever, exactly true. To evaluate each example model in this respect, we plotted the O to E ratios and their 25% and 10% quantiles together with 95% confidence intervals on the E gradient. The confidence intervals for the quantiles were obtained by a non-parametric bootstrap technique (Efron and Tibshirani 1993), for

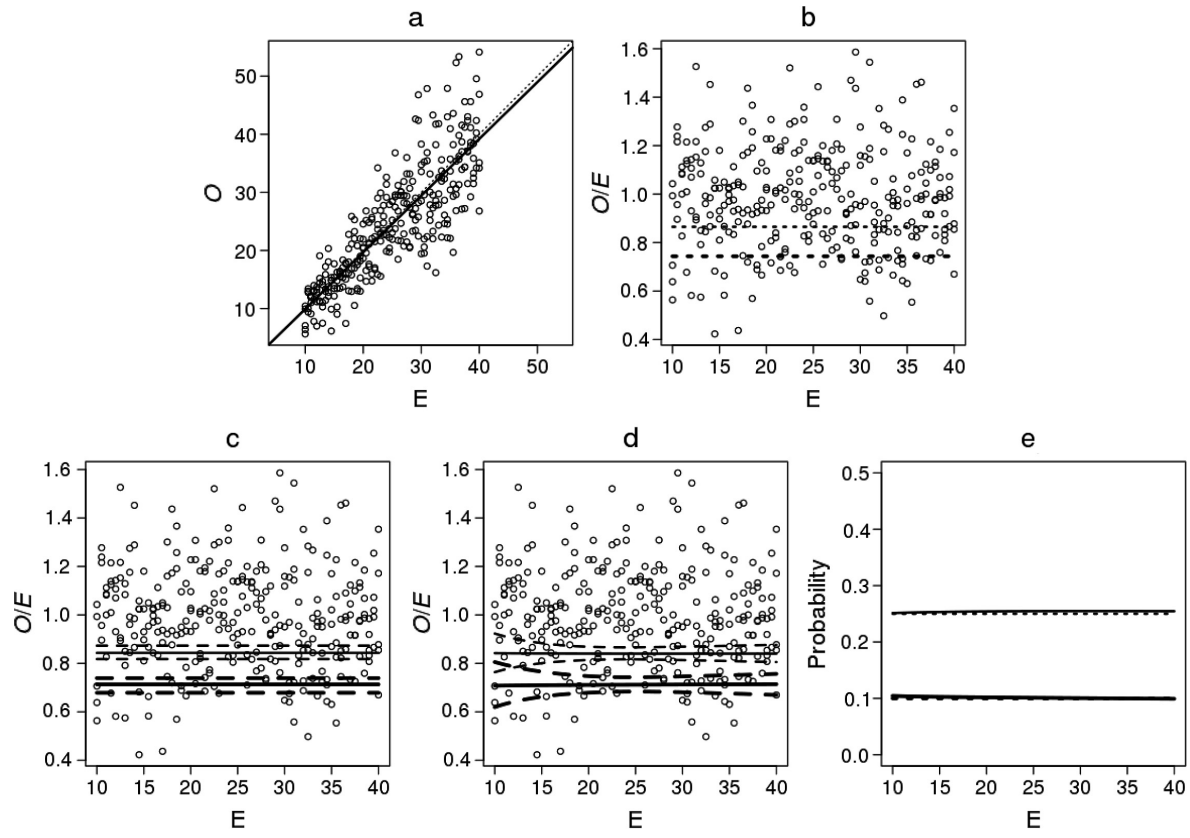

FIG. 2. (a) A simulated realization of a O (observed metric value) on E (expected metric value) regression model with coefficients $\alpha = 0$ and $\beta = 1$, and variance $\text{var}(\varepsilon) = E^2 \times 0.2^2$, and the estimated regression line with values $a = 0.152$, $b = 0.976$, $\text{var}(\varepsilon) = E^{1.920}0.240^2$; (b) theoretical 10% decision line at 0.744 (dashed lower) and 25% decision line at 0.865 (dotted upper); (c) estimated 10% decision line at 0.713 (thick solid) and 25% decision line at 0.843 (thin solid) with 95% confidence intervals (dashed) when conventional method and (d) similarly the 10% and 25% decision curves with 95% confidence intervals when the proposed method are used, and (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

which we drew 10,000 bootstrap replications from the rows of original data (R function sample).

For models with inconstant O/E variance, we propose a novel approach, which takes into account the variability in O/E ratio with E, and concomitantly the possible bias in O on E relationship, the concern of Linke et al. (2005). For simplicity we assume a linear regression of O on E such that it follows approximately a Gaussian distribution with mean $\alpha + \beta E$ and variance $\text{var}(\varepsilon)$, later being constant or depending on the E value. Furthermore, for simplicity, we assume each E value to be fixed, as has been the convention in the previous related studies and as is a common practice in the evaluation of predictive models (e.g., Mayer and Butler 1993).

In the simplest case, let us assume that $\alpha = 0$, $\beta = 1$, and $\text{var}(\varepsilon) = \sigma^2$, from which it follows that $\text{var}(O) = \sigma^2$. The distributions of O and O/E, given E, are then

$$N(\alpha + \beta \cdot E, \sigma^2) \tag{1a}$$

and

$$N\left((\alpha + \beta \cdot E)/E, (\sigma/E)^2\right) \tag{1b}$$

respectively. The latter distribution is based on the result of a linear transformation for a random Gaussian variable (e.g., Davison 2003).

For more complex cases, we consider the regression model with an inconstant variance given E, $\text{var}(O) = E^{2\delta} \times \sigma^2$, where δ can be fixed ($\text{var}(O) = E^2\sigma^2$ with $\delta = 1$, for example), or non-fixed and unknown. The corresponding models of O and O/E given E are Gaussian models

$$N(\alpha + \beta \cdot E, E^{2\delta} \cdot \sigma^2) \tag{2a}$$

and

$$N\left((\alpha + \beta \cdot E)/E, (E^{2\delta} \cdot \sigma^2)/E^2\right) \tag{2b}$$

respectively.

For $\delta = 1$, the theoretical variance of O/E is constant σ^2 and, with $\delta = 0$, we obtain the first models (1a and 1b). Therefore, the most important part is the structure of the variance of O/E values. In the models 1b and 2b, the variance of O/E depends on the E value if δ is not equal to 1. Furthermore, the mean of O/E is not always 1 but can vary with E.

In our solution to obtain a corrected decision rule, taking into account the dependence of O/E variation on E, we first estimate the model 1a or 2a, that is, the regression model from O and E values in order to have estimates to be used in forming the latter model in 1b or 2b, the model of O/E values.

For the simulated data and for each case study, we first fitted separately three models with fixed δ values (0, 0.5, and 1.0) by using the R function gls (based on generalized least squares function, e.g., Pinheiro and Bates 2000). The model with constant variance (1a, δ = 0) is obtained with gls function. The 2a models with fixed δ values are obtained by updating the gls object with an option weights = varFixed (~e) (δ = 0.5) or weights = varFixed(~$e \times e$) (δ = 1), respectively. We further used gls function with weights = varPower (~e), which also gives the estimate for the δ of the general model (2a) above. Due to the non-nestedness of the models with fixed δ, they were compared using the log-likelihood (loglik, the greater, the better) and the Akaike Information Criterion (AIC, the smaller, the better)

$$\mathrm{AIC} = -2\mathrm{Loglik} + 2\mathrm{npar}$$

where npar is the number of estimated parameters. We also compared the best model with fixed δ (three parameters) to the model with an estimated δ (four parameters) by using χ^2 test. In the diagnostics part, we further checked the plot of fitted values vs. standardized residuals by using plot function for the output object of gls function.

Second, using results of the selected, best fitting regression model, we formulated the p decision curve from the classical prediction interval formula constructed such that the variance-covariance structure of regression coefficients and variance of the error are taken into account (e.g., Davison 2003, Gelman and Hill 2007, Hocking 2013), as follows. To construct the classical prediction interval, the variance of the predicted observation $\hat{O}$ given E is needed, that is

$$\begin{aligned}\mathrm{var}(\hat{O}) &= \mathrm{var}(\varepsilon) + \mathrm{var}(a + b \cdot E) \\ &= \mathrm{var}(\varepsilon) + \mathbf{E}^T \cdot \mathbf{cov}(a,b) \cdot \mathbf{E}\end{aligned} \tag{3}$$

where a and b are the estimated regression coefficients, $\mathbf{E}^{\mathrm{T}} = [1E]$ is the vector of one and the value E, T stands for the transpose, and the matrix $\mathbf{cov}(a,b)$ is the covariance matrix of the estimated regression coefficients a and b. Consequently, the variance $\mathrm{var}(a + bE) = \mathbf{E}^T \times \mathbf{cov}(a,b) \times \mathbf{E}$ is needed due to estimation of α and β.

Then, in the case of constant variance (Eq. 1a), the lower limit of a classical (1−2p) prediction interval (p = 0.10 or p = 0.25), that is the p quantile (10% or 25%) for a new observation O given E can be formulated (e.g., Davison 2003) as

$$a + b \cdot E + t_{n-2}(p) \cdot \sqrt{[s^2 + \mathbf{E}^T \cdot \widehat{\mathbf{cov}}(a,b) \cdot \mathbf{E}]} \tag{4}$$

where $a + bE$ is the point prediction, s is the estimate of σ, and $\mathbf{cov}(a,b)$ is replaced by its estimate $\widehat{\mathbf{cov}}(a,b)$. The latter term forms the standard deviation of the prediction $\hat{O}$, $t_{n-2}(p)$ is the p quantile of the t distribution with degrees of freedom $n - 2$, and n is the number of observations in the data. Due to the estimation of σ, a t distribution (R function qt) is used instead of a normal distribution.

When considering the ratios O/E, instead of O, the variance of the predicted observation $\hat{O}/E$ given E is obtained from Eq. 3 after some small modifications

$$\begin{aligned}\mathrm{var}(\hat{O}/E) &= \mathrm{var}(\varepsilon/E) + \mathrm{var}((a + b \cdot E)/E) \\ &= (1/E^2) \cdot [\mathrm{var}(\varepsilon) + \mathbf{E}^T \cdot \mathbf{cov}(a,b) \cdot \mathbf{E}]\end{aligned} \tag{5}$$

and expression 4 is updated as follows

$$(a + b \cdot E)/E + t_{n-2}(p) \cdot (1/E) \cdot \sqrt{[s^2 + \mathbf{E}^T \cdot \widehat{\mathbf{cov}}(a,b) \cdot \mathbf{E}]}. \tag{6}$$

The latter formula forms the estimated p decision curve as a function of E values, when variance of ε is constant.

In the case of inconstant variance, $\mathrm{var}(\varepsilon) = E^{2\delta}\sigma^2$, equation 6 is slightly modified such that the term s^2 is replaced by $E^{2\delta}s^2$ as follows

$$(a + b \cdot E)/E + t_{n-2}(p) \cdot (1/E) \cdot \sqrt{\left[E^{2\delta}s^2 + \mathbf{E}^T \cdot \widehat{\mathbf{cov}}(a,b) \cdot \mathbf{E}\right]} \tag{7}$$

where δ is either fixed or estimated. If absolute differences ($O - E$) were used in the assessment, instead of O/E ratios, the formula 7 could be easily updated as follows

$$(a + b \cdot E - E) + t_{n-2}(p) \cdot \sqrt{\left[E^{2\delta}s^2 + \mathbf{E}^T \cdot \widehat{\mathbf{cov}}(a,b) \cdot \mathbf{E}\right]}. \tag{8}$$

For each simulation and CS, based on the fitted Gaussian model 2a with non-fixed δ, we plotted the p decision curve formed by p quantiles of O/E values given E, as calculated from the reference data. That is, we used equation 7, allowing the variance to be inconstant and we used both p = 0.10 and p = 0.25.

In addition, we plotted the 95% confidence intervals for the p decision curves, based on 10,000 bootstrap replicates, from the real data, as follows. For each bootstrap replicate, we first estimated the regression model (2a) and using the results of that model, p quantiles of O/E values were calculated for each E value with the equation 7. From the distributions of these percentage quantiles, we obtained their 2.5% and 97.5% quantiles, and thereby the 95% confidence intervals for each unknown p decision curve.

To quantify the potential bias or difference in decisions made by the conventional and the novel method, we calculated the proportion of IMP sites with differing classifications for the case studies 1–3. However, as these proportions are fully data specific and cannot be generalized or applied to any new observation or other data, we additionally evaluated and demonstrated the size of bias in a more generalizable way. For each simulation and CS, we first calculated the nominal p (overall 25% or 10%) quantile q_n from the data. Using the inverse of our approach, we were able to estimate and then plot the actual decision rule (quantile) given E, in comparison with the assumed (25% or 10%) quantile q_n. The actual q (and p) to be used, conditional to E, instead of q_n, (and p_n) is solved as follows. Utilizing the general formula 7, the quantile of the t distribution $t_{n-2}(p)$ is first solved from the expression

$$\begin{aligned}&(a + b \cdot E)/E + t_{n-2}(p) \cdot (1/E) \cdot \\ &\sqrt{\left[E^{2\delta}s^2 + \mathbf{E}^T \cdot \widehat{\mathbf{cov}}(a,b) \cdot \mathbf{E}\right]} = q_n,\end{aligned} \tag{9}$$

where q_n is the overall p (25% or 10%) quantile of the O/E values (conventional approach). Using $n - 2$ as the degrees

of freedom of t distribution, the corresponding probability p related to the quantile $t_{n-2}(p)$ can be obtained using the pt function of the R software.

Results

For the simulated data with constant variance, the estimated parameter values of the regression line were $a = 0.224$, $b = 1.009$, $\text{var}(\varepsilon) = 2.963^2$, when using known $\delta = 0$ (1a) and $a = 0.219$, $b = 1.009$, $\text{var}(\varepsilon) = E^{-0.275}4.558^2$, when using the estimated $\delta = -0.138$ (2a) (Fig. 1a). The model with four parameters (with the estimated δ) has only marginally (and not significantly, $P = 0.191$) greater log-likelihood, and as a more complex model, also higher AIC, when compared to the model with constant variance with $\delta = 0$ (Table 1). As mathematically necessary in these specific conditions, the variation in O/E varies with E, obviously increasing with decreasing E (Fig. 1b). We chose the model with estimated δ to calculate the decision curves, but the curves obtained from the model with fixed δ (not shown) are quite similar. The main difference is that the confidence intervals are wider when the δ parameter of the variance is estimated. The calculated overall 10th and 25th percentiles for O/E were 0.872 and 0.936, respectively (Fig. 1c). These values differ substantially from the theoretical and estimated decision curves given E (Fig. 1b, d), and are largely beyond their calculated 95% confidence intervals (Fig. 1d). As a result of these patterns, if the conventional method was used, the actual probability of judging reference sites as impacted would vary greatly with E and differ substantially from the nominal probability for both 10th and 25th percentiles, at the ends of the E gradient in particular (Fig. 1e).

For the simulated data with inconstant variance $\text{var}(\varepsilon) = E^2 0.2^2$, the estimated parameter values were $a = 0.138$, $b = 0.977$, $\text{var}(\varepsilon) = E^2 0.211^2$ when using known $\delta = 1$ and $a = 0.152$, $b = 0.976$, $\text{var}(\varepsilon) = E^{1.920}0.240^2$ when using estimated $\delta = 0.960$. The model with four parameters (with the estimated δ) had minimally (and not significantly) greater log-likelihood, but higher AIC when compared to the model with the fixed $\delta = 1$ (Table 1). We chose the model with the estimated δ for further calculations, but the decision curves are quite similar to the model with fixed δ (not shown). As theoretically expected for this error structure, variation of O/E values does not depend on E (Fig. 2b) and the estimated overall 10% and 25% tile decision lines (0.713 and 0.843; Fig. 2c) are almost identical to curves as estimated by the novel method (Fig. 2d). Accordingly, in this simulated case, there would be practically no difference in the probability of judging reference sites as impacted, independent of E (Fig. 2e).

The best-fitting regression model type or error structure varied among the five case studies, but with some consistency (Table 2). For each CS the best model with a fixed δ (0, 0.5 or 1) (Table 3) fitted almost as well as the option with an estimated δ (Table 4) according to the Loglik (Table 2).

Table 1. Akaike information criterion (AIC) and log-likelihood values (Loglik) of the estimated models with differing alternative error variance structures (Eqs 1a and 2a) for the two simulated data.

	Simulation 1		Simulation 2	
Model	AIC	Loglik	AIC	Loglik
σ^2	**1539.84**	**−766.92**	1917.11	−955.55
$E\sigma^2$	1575.82	−784.91	1864.44	−929.22
$E^2\sigma^2$	1657.45	−825.72	**1846.74**	**−920.37**
$E^{2\delta}\sigma^2$	1540.13	**−766.06**	1848.60	**−920.30**
δ	−0.138		0.960	
P		0.191		0.702

Notes: The fixed δ was used in the three first models and the non-fixed, estimated δ in the fourth model. Also shown are the estimated δ and the P of the χ^2 test for the comparison of the loglik values in boldface type.

Table 2. AIC and log-likelihood of the three different models with fixed δ and with estimated δ for the CS1–5 data sets.

	CS1		CS2		CS3		CS4		CS5	
Model	AIC	Loglik	AIC	Loglik	AIC	Loglik	AIC	Loglik	AIC	Loglik
σ^2	525.64	−259.82	**121.78**	**−57.89**	601.72	−297.86	**496.75**	−245.38	709.49	−351.75
$E\sigma^2$	518.15	−256.08	124.12	−59.06	594.89	−294.45	**496.17**	**−245.09**	689.54	−341.77
$E^2\sigma^2$	**513.84**	**−253.92**	128.48	−61.24	**593.01**	**−293.50**	519.70	−256.85	**679.75**	**−336.88**
$E^{2\delta}\sigma^2$	515.03	**−253.51**	123.34	**−57.67**	594.95	**−293.47**	**495.75**	**−243.88**	681.18	**−336.59**
δ	1.317		−0.355		0.926		0.269		1.147	
P		0.366		0.506		0.805		0.120		0.451

Note: Also shown are the estimated δ and the P values of the χ^2 test for the comparison of the loglik values in boldface type.

Table 3. The estimated regression parameter values (a and b) and variance, var(ε), for CS1–5 with fixed δ when the model with boldface AIC in Table 2 was chosen. Also shown are the number of reference sites (Nref) and impact sites (Nimp) for each CS.

Parameters	CS1	CS2	CS3	CS4	CS4	CS5
a	1.498	−0.456	−0.237	0.625	0.452	0.170
b	0.882	1.233	1.019	0.987	1.008	1.001
var(ε)	$E^2 0.179^2$	0.550^2	$E^2 0.161^2$	1.778^2	$E 0.655^2$	$E^2 0.183^2$
Nref	96	68	118	122	122	143
Nimp	134	431	142			

Note: CS4 has two columns for two competing models, since AICs (Table 2) are almost the same.

TABLE 4. The estimated parameter values for CS1–5, when the model with non-fixed δ was chosen.

Parameters	CS1	CS2	CS3	CS4	CS5
a	1.431	−0.420	−0.175	0.511	0.180
b	0.886	1.221	1.016	1.000	1.000
var(ε)	$E^{2.634}0.070^2$	$E^{-0.670}0.770^2$	$E^{1.852}0.199^2$	$E^{0.538}1.027^2$	$E^{2.294}0.124^2$

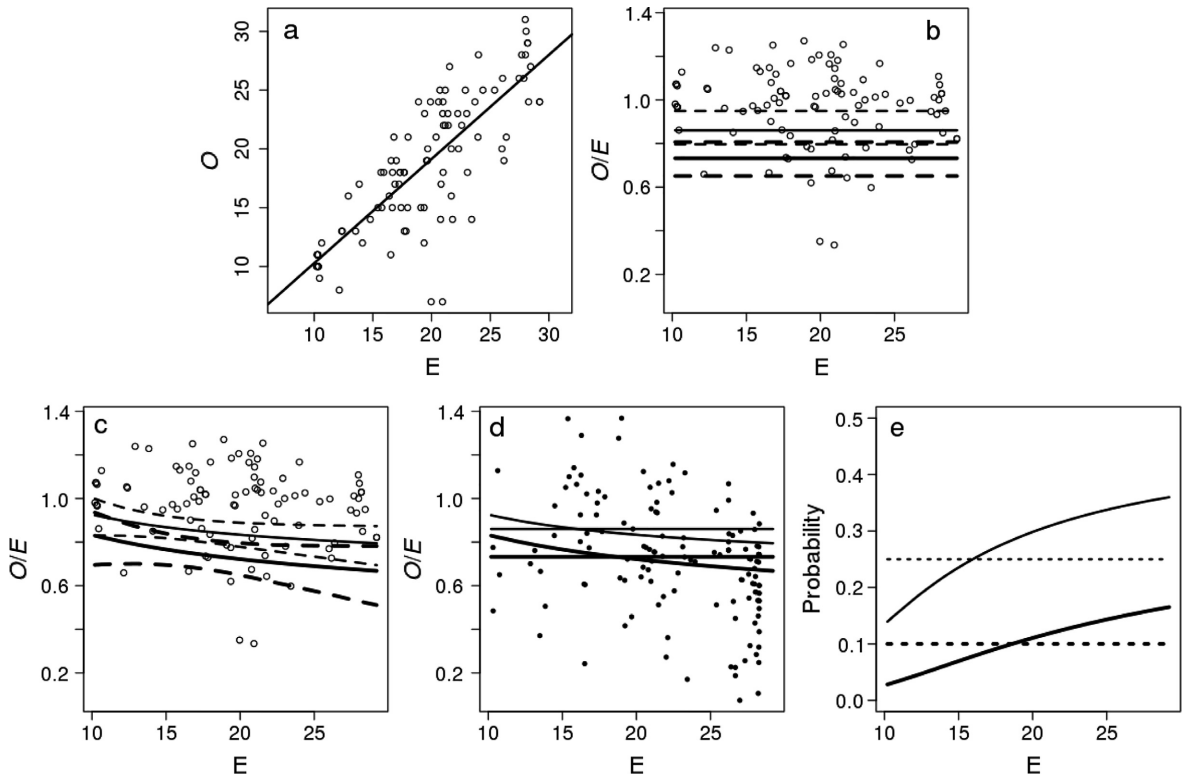

FIG. 3. For the first data set, CS1: (a) the estimated relationship of the observed metric value (*O*) to the expected value (*E*) with all points (upper), and without two outliers (lower); (b,c,d) the *O*/*E* ratio in relation to *E* for with decision curves of 25% (thin solid line) and 10% (thick solid line) quantiles, and their bootstrap 95% confidence intervals (dashed lines) for the reference data as based on the conventional method (b) and the novel method (c), and with the decision curves of both methods for the IMP data (d); (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

However, for CS4, the best fitting model had the estimated δ between the fixed δ:*s* of two alternatives, which were almost as good. From that, we noticed that the shape of the decision curves is sensitive to δ. Therefore, for each CS, we here selected the model with the greatest log-likelihood for plotting confidence intervals of decision curves, in order to take into account the uncertainty in the estimation of δ.

For the CS1, despite the apparently rather constant σ (Fig. 3a), the model with fixed δ = 1 fitted better than the other two options (greatest loglik) and was sufficient when compared to the model with an estimated δ = 1.317 (Table 2). Nonetheless, variance of *O*/*E* did not vary greatly with *E* (Fig. 3b). For the CS2 with apparently constant variance (Fig. 4a), the corresponding model with the fixed δ = 0 fitted the best among the three options, as expected, and was sufficient when compared to the model with estimated δ = −0.355 (Table 2). Also for the CS4 with apparently similar error pattern, a model with δ = 0 fitted well, but the model with δ = 0.5 was marginally better. The estimated δ = 0.269 was in between of these fixed values and by both AIC and Loglik, the corresponding model was the best (Table 2). For both CS2 and CS4, there was a strong trend of decreasing variance of *O*/*E* with increasing *E* (Figs 4b, 6b). For the CS3 and CS5 showing increasing error variance with increasing *E* (Figs. 3a, 5a), the model with δ = 1 fitted the best (Table 2), as could be expected, and for them, there was no apparent trend in *O*/*E* variation on the *E* gradient (Figs 3b, 5b).

For the CS1, there is only a relatively small difference in the decision curves between the conventional (Fig. 3b) and novel (Fig. 3c) approaches. For small *E*, the modeled values, conditional to *E*, however, are slightly greater than the overall values for both quantiles. In contrast, for the CS2 (Fig. 4b, c) and CS4 (Fig. 6b, c), there is a remarkable

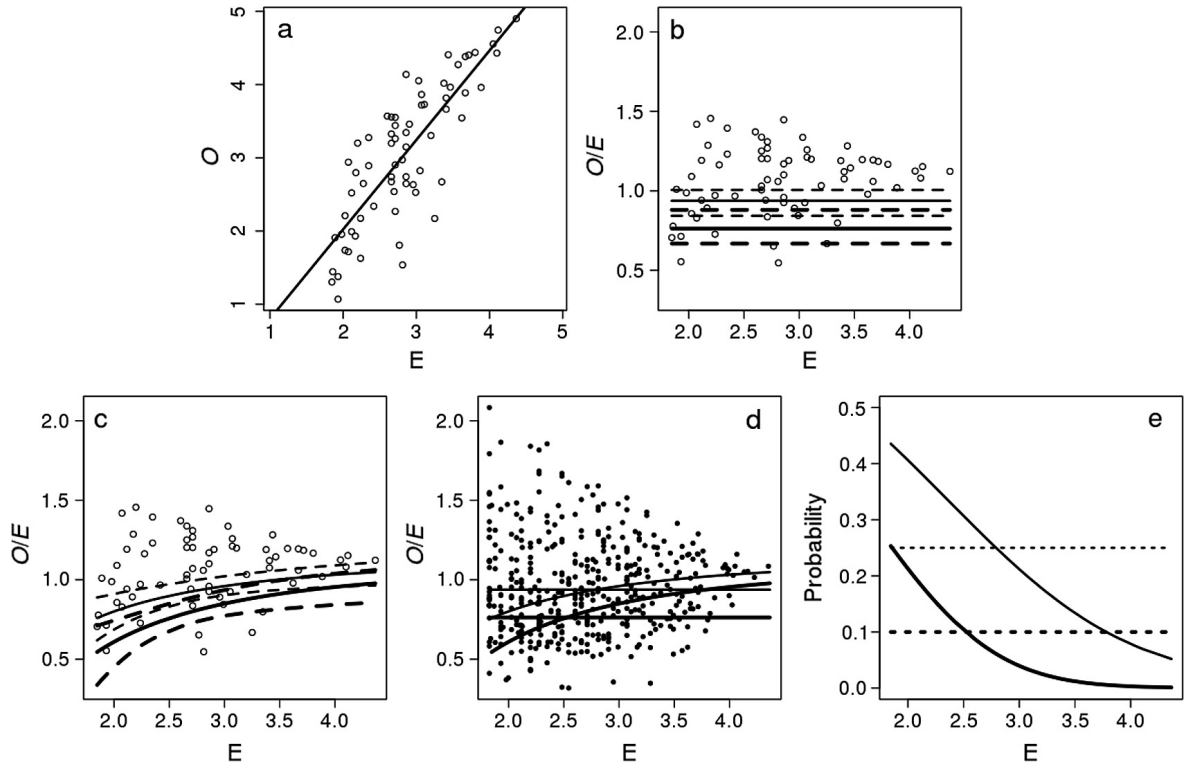

FIG. 4. For the second data set, CS2: (a) the estimated relationship of the observed metric value (O) to the expected value (E); (b,c,d) the O/E ratio in relation to E for with decision curves of 25% (thin solid line) and 10% (thick solid line) quantiles, and their bootstrap 95% confidence intervals (dashed lines) for the reference data as based on the conventional method (b) and the novel method (c), and with the decision curves of both methods for the IMP data (d); (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

difference in the decision curves. For both case studies, the modeled percentiles are substantially smaller than the overall percentiles for small E and in addition, the reverse is also true; for greater E, the modeled percentiles are greater than the conventional percentiles. For the CS3 and CS5 with more inconstant variance, the conventional method corresponds quite well with the novel approach for both the 25% and 10% quantiles (Figs 5b, c, 7b, c), even though, for the CS5, there is a similar type of slight difference as in the CS1. The precision of the classical prediction formula ($a + bE$) is known to be greatest at the mean of E values, and lower at small and high E values. Therefore the confidence intervals form a shape of an hourglass around the regression line. This is reflected in the confidence intervals of decision curves (Figs. 1d, 2d, 3–7c).

For each case study and percentile, the area between the two decision curves depicts the potential bias as within these regions the sites would be differently classified by the two approaches. For the particular set of IMP sites in the CS2 the proportion of sites differently classified was 12.5% and 9.5% for the 10% and 25% quantiles, respectively (Fig. 4d). For CS1 (Fig. 3d), the corresponding figures are 8% and 6% and for CS3 (Fig. 5d) only 2% and 1%.

Assuming the modeled O/E variance is even approximately correct as our simulations suggest, the actual 25% and 10% quantiles can differ greatly from the assumed quantiles in real case studies, depending on the E (Fig. 4e, 6d). For instance, for CS2 with $E = 2$ and using the traditional method and 25th percentile rule, broadly 45% of the reference sites (instead of the assumed 25%) would actually be judged impaired. In contrast, for the sites with high E greater than 4, the corresponding figure is even less than 10% (Fig. 4e). Thus, for such sites, a 10% decision rule would actually be used instead of the notional 25% rule. For the CS4, there is a similar difference for the 10th percentile, but less distinctive for the 25% percentile (Fig. 6d). For the other case studies the differences are much smaller, and for the 10% quantile for CS3 there is no difference at all (Fig. 5e).

DISCUSSION

We have shown that using predictive models to estimate expected values (E) for biotic variables and then a constant (e.g., 10th or 25th) percentile of overall observed (O) to expected ratio (O/E) distribution to differentiate impacted sites can lead to biased assessments. Even though we have here elaborated and exemplified the bias only for two percentiles commonly used as decision rules, the problem can be generalized to any others and to differing quality class boundaries based on percentiles of the overall O/E distribution. This bias will occur even for otherwise unbiased predictive models, if they show a constant prediction error ($O-E$) for all values of E, a condition that should normally be considered optimal. In such cases, however, the variance of O/E ratio will not be constant but will decrease with increasing

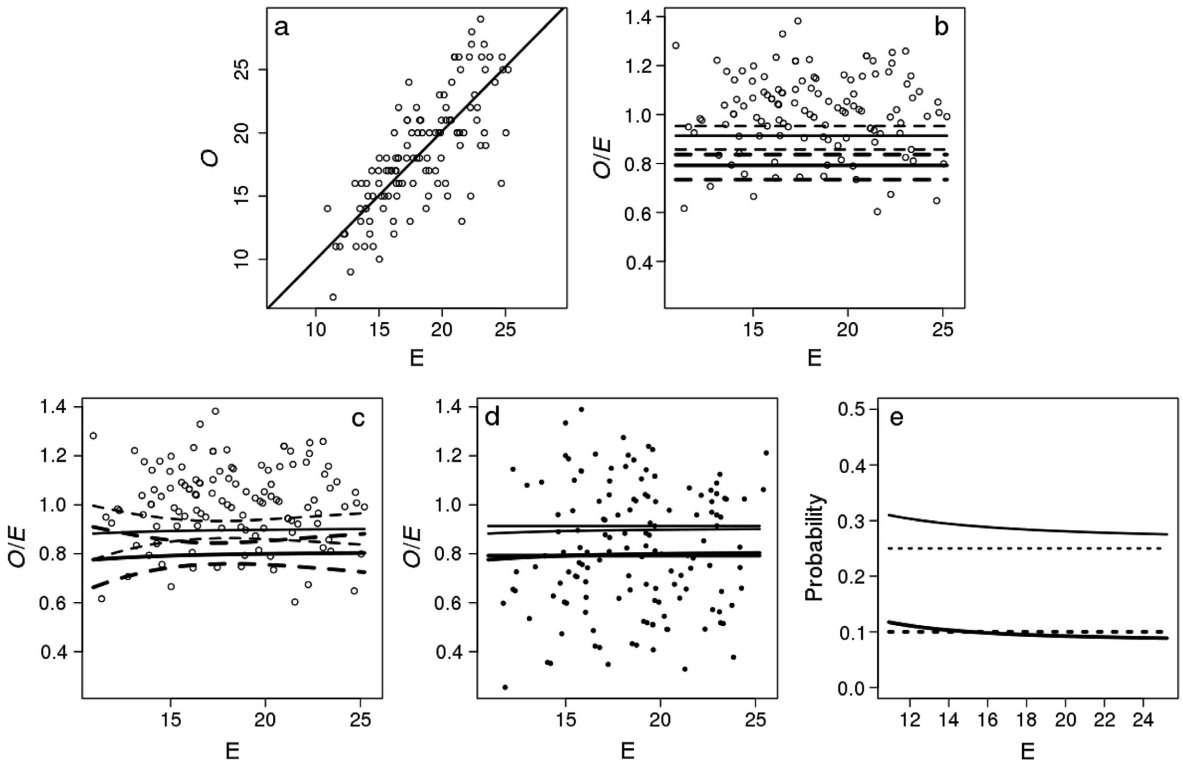

FIG. 5. For the third data set, CS3: (a) the estimated relationship of the observed metric value (O) to the expected value (E); (b,c,d) the O/E ratio in relation to E for with decision curves of 25% (thin solid line) and 10% (thick solid line) quantiles, and their bootstrap 95% confidence intervals (dashed lines) for the reference data as based on the conventional method (b) and the novel method (c), and with the decision curves of both methods for the IMP data (d); (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

E. This, in turn, will lead to a greater propensity of sites with small E to have O/E values falling beyond the critical percentile of overall O/E distribution, and hence being mistakenly judged as impacted. For sites with a large E in turn, there will mostly be a contrasting bias. The significance of this bias varies depending on the intercept, slope, and error parameters of the O on E relationship, but can be alarming, as shown for our CS2 and CS4 in particular. For the CS2 the intercept smaller than 0 and slope greater than 1, together making the small O to be over-predicted and high O under-predicted, actually strengthens the bias at the O/E scale. Contrasting bias in the O–E relationship ($\alpha > 1$ and $\beta < 1$) would alleviate the bias at the O/E scale, as perhaps for the CS1. In CS3 and CS5, error increased with E and retained the relative error quite stable with E. In such cases, the bias can even be negligible.

We suggested a relatively simple method to check and overcome the described bias stemming from variable distribution of O/E on E, and hence to produce more credible assessments of biotic condition in streams and lakes, when O/E is used in decision making. Actually, this approach corrects not only the bias of our original primary interest, but also the bias related to deviation from a 1:1 relationship in O on E regression, about which Linke et al. (2005) were concerned. Given that the O/E variance depends on E, the overall distribution of O/E, and thereby the critical percentiles of this distribution used in the traditional method, will be sensitive to the distribution of E in the modeling data. For instance, if in the simplest case of O on E relationship with constant error, observations with high E (and therefore smaller O/E variance) would be more frequent than those with low E in the data, it would increase the critical overall O/E percentiles. This, in turn, would increase the likelihood of sites with lower E values (and greater O/E variance), to be in the lower tail of the O/E distribution, and judged impaired, even if the original predictive model would suggest they were not. Hence, the severity of bias also depends on the sample distribution of E in the modeling data. Our suggested correction, based on the modeled error distribution conditional to E, at least alleviates this potential problem.

In effect, in the case of constant variance in the O on E relationship, our bias correction broadly equates to decision rules based on fixed overall percentiles of actual deviations from the expectation or $O-E$. In such a case, a simple and straightforward approximate solution to overcome the described bias is to use a fixed overall percentile of actual, rather than relative difference as a decision rule, and to replace values of E with the fitted values ($\hat{O} = a + bE$) from the estimated O on E regression model, when needed ($a \neq 0$ and/or $b \neq 1$). For instance, if this simple approach was applied to our simulation with constant error (Appendix S1: Fig. S1), and to CS4 with similar error pattern (Appendix S1: Fig. S3), nearly the same exact sites would fall beyond the 10th and 25th percentiles in the modeling data, as with the method we suggested for using O/E. The

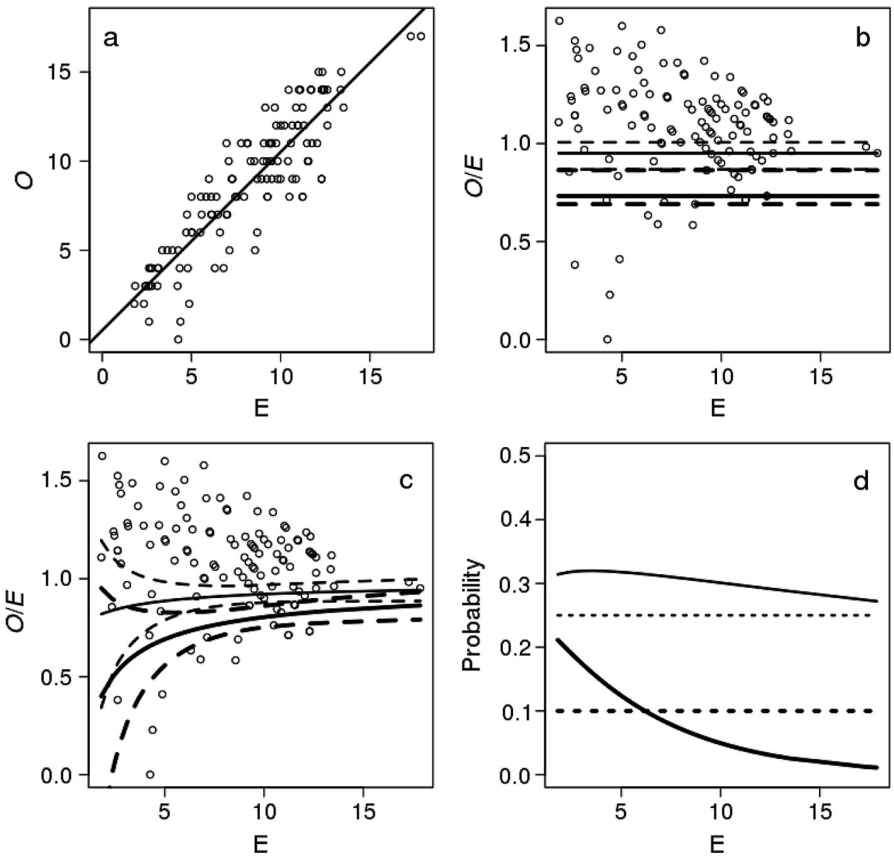

FIG. 6. For the fourth data set, CS4: (a) the estimated relationship of the observed metric value (O) to the expected value (E); (b,c) the O/E ratio in relation to E for with decision curves of 25% (thin solid line) and 10% (thick solid line) quantiles, and their bootstrap 95% confidence intervals (dashed lines) for the reference data as based on the conventional method (b) and the novel method (c), and (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

percentages of cases classified differently were 1.3 and 1.6 for the simulated data and 0.8 and 0.8 for CS4 for 10th and 25th percentiles, respectively. Hence, the classification outcomes for independent data would also likely to be quite similar with these two methods. For CS2 with error pattern of the same type, classification based on fixed percentiles of $O - \hat{O}$ (Appendix S1: Fig. S2) showed 94% and 98% match (for 10th and 25th percentiles, respectively) with our suggested general method for O/E ratio in the classification of the IMP sites. For the simulation data, the decision curves based on Eqs. 7 and 8 applied for O/E and $O - E$, respectively, yielded identical classifications, as theoretically expected (results not shown). However, it might be argued, that similar proportional species loss (ratio of observed to expected species) for example, instead of absolute decrease in species richness should have similar ecological or societal significance and hence should be acceptable for all sites, independent of the number of expected species (E). This position is intuitively justifiable, but, as we have shown, might come at the expense of increased risk of erroneously classifying sites that do not differ from undisturbed sites as impacted, and classifying others actually deviating from the reference range as non-impacted. Moreover, it remains uncertain how closely the O/E ratio or taxonomic completeness (sensu Hawkins 2006) correlates with the actual species loss (Hawkins et al. 2010*b*), and whether this depends on the E; or on the other hand, conceptually, whether indeed greater absolute (equal proportional) loss of species should be acceptable for more diverse ecosystems than for species poor. For instance, it is possible that a given absolute species loss reflects similar likelihood that a functionally important species is lost, independent of the total richness. The PICM metric of our second case study strongly and linearly correlates with the first axis of detrended correspondence analysis (DCA) on the macroinvertebrate community data among lakes (Jyväsjärvi et al. 2014). As the first DCA axis is proportional to the main gradient of β-diversity or species overturn (Gauch 1982), this suggests that each unit change in the metric value corresponds to a fixed degree of change in community composition. Thereby, to rely on the actual rather than proportional PICM change or on the corrections as we suggest, might be justified even on purely ecological grounds. This is likely to be true for many other biotic indices or metrics.

One should also recall that predictive modeling is not the only method to derive E values. Another, widely used approach is estimating the expected values for groups of water bodies sharing similar natural features (e.g., Aroviita et al. 2008). Establishing reference values for such categorical types of streams and lakes currently is the widely used

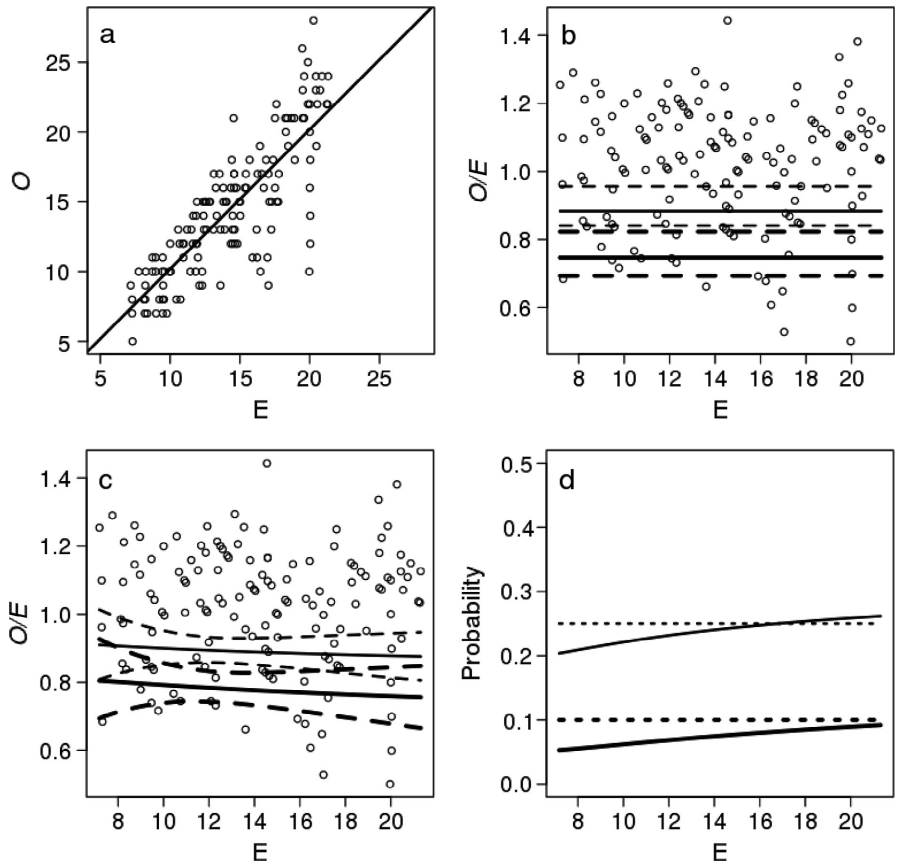

FIG. 7. For the fifth data set, CS5: (a) the estimated relationship of the observed metric value (O) to the expected value (E); (b,c) the O/E ratio in relation to E for with decision curves of 25% (thin solid line) and 10% (thick solid line) quantiles, and their bootstrap 95% confidence intervals (dashed lines) for the reference data as based on the conventional method (b) and the novel method (c), and (e) the actual 25% (thin solid curve) and 10% (thick solid curve) quantiles based on the conventional method, when the assumed ones are 25% (thin dashed line) and 10% (thick dashed line) quantiles.

default in the European legislation (e.g., Hering et al. 2010, Birk et al. 2012). In this approach, the O/E variation is treated and decision rules typically set group-wisely (instead of using global variance across all types) and thereby the potential dependence of variance of E, as here suggested for modeling approaches, is controlled for. However, this approach does not solve the conceptual issue of using ratios, has many other problems (Bowman and Somers 2005, Hawkins et al. 2010*a*, Hering et al. 2010) and when explicitly compared is inferior to modeling in performance (e.g., Aroviita et al. 2009).

A literature search showed that relatively few studies have reported the O on E regressions (Appendix S2: Table S1). Therefore it is impossible to assess how prevalent and severe the described biases might be in the proposed assessment systems and in those currently used. However, the reported estimates, including those in Linke et al. (2005), for the intercept and slope parameters are mostly close to and do not significantly deviate (when statistically tested) from 0 and 1, respectively, albeit there are some apparent exceptions (Appendix S1). None of the studies considered the error structure, but when the data were also presented graphically, the error variation appeared to increase, to be constant, or even to show a decreasing trend or curvature pattern along the E range. Hence, even though error increasing with E, a condition at least alleviating the bias, seems to be common for the taxonomic completeness or O/E metric in particular, other types of error structure and therefore some degree of bias are frequent, and should require careful attention. To our knowledge, only Mazor et al. (2016) have previously specifically addressed the dependence of O/E distribution on E. Interestingly, they reported a decreasing accuracy (proportion of reference sites with $O/E > 10\%$ percentile value) and decreasing precision (increasing SD of reference O/E) with decreasing E. These trends might actually be explainable by the phenomena we have described here.

For the developers of predictive models for the RCA assessments, we advocate at least, in addition to reporting the overall mean and variability of the O/E as presently habitual in model validation, also communicating statistics and preferably graphs of the O on E or O/E on E relationship, or both. This will allow both the authors and the audience to evaluate the risk of bias and need for correction. Even though the possible bias, as described, might often be relatively small compared to other sources of uncertainties, we consider it worth taking into account. Also, the use of either absolute or relative deviation from the expected to assess the biotic condition should preferably be explicitly

justified. However, when ratios are used for legislative requirements or for other reasons and the O/E ratio varies with E, we strongly recommend decision curves estimated by the approach suggested (Eqs. 6–7) to be compared with the conventional decision lines. Also alternative methods can be developed. A simple approximate approach for at least detecting the possible bias and also partially solving it, might be calculating and using piecewise O/E percentiles, separately for different intervals of E (see Mazor et al. 2016), or by using a "moving E window," when there are enough data to get credible estimates. More sophisticated (and perhaps more appropriate) statistical approaches directly modeling the uncertainty in E by using joint distribution models of O on E should also be considered.

Acknowledgments

Thoughtful suggestions of two anonymous reviewers and Song Qian as the editor greatly improved the manuscript. Roger Jones kindly helped in improving the English language. S. Kärkkäinen acknowledges support from the Academy of Finland project 289076.

Literature Cited

Aroviita, J., E. Koskenniemi, J. Kotanen, and H. Hämäläinen. 2008. A priori typology-based prediction of benthic macroinvertebrate fauna for ecological classification of rivers. Environmental Management 42:894–906.

Aroviita, J., H. Mykrä, T. Muotka, and H. Hämäläinen. 2009. Influence of geographical extent on typology- and model-based assessments of taxonomic completeness of river macroinvertebrates. Freshwater Biology 54:1774–1787.

Aroviita, J., H. Mykrä, and H. Hämäläinen. 2010. River bioassessment and the preservation of threatened species: Towards acceptable biological quality criteria. Ecological Indicators 10:789–795.

Bailey, R. C., R. Norris, and T. B. Reynoldson. 2004. Bioassessment of freshwater ecosystems: using the Reference Condition Approach. Kluwer Academic Publishers, Boston, Massachusetts, USA.

Birk, S., W. Bonne, A. Borja, S. Brucet, A. Courrat, S. Poikane, A. G. Solimini, W. van de Bund, N. Zampoukas, and D. Hering. 2012. Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. Ecological Indicators 18:31–41.

Bowman, M. F., and K. M. Somers. 2005. Considerations when using the reference condition approach for bioassessment of freshwater ecosystems. Water Quality Research Journal of Canada 40:347–360.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Cao, Y., and C. P. Hawkins. 2011. The comparability of bioassessments: a review of conceptual and methodological issues. Journal of the North American Benthological Society 30:680–701.

Chessman, B. C., M. Muschal, and M. J. Royal. 2008. Comparing apples with apples: use of limiting environmental differences to match reference and stressor-exposure sites for bioassessment of streams. River Research and Applications 8:103–117.

Clarke, R. T., M. T. Furse, J. F. Wright, and D. Moss. 1996. Derivation of a biological quality index for river sites: Comparison of the observed with the expected fauna. Journal of Applied Statistics 23:311–332.

Davison, A. C. 2003. Statistical models. Cambridge University Press, New York, New York, USA.

Efron, B., and R. Tibshirani. 1993. An introduction to bootstrap. Chapman & Hall, New York, New York, USA.

European Commission. 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. Official Journal L 327:1–72.

Feio, M. J., C. Viana-Ferreira, and C. Costa. 2014*a*. Combining multiple machine learning algorithms to predict taxa under reference conditions for streams bioassessment. River Research and Applications 30:1157–1165.

Feio, M. J., C. Viana-Ferreira, and C. Costa. 2014*b*. Testing a multiple machine learning tool (HYDRA) for the bioassessment of fresh waters. Freshwater Science 33:1286–1296.

Gauch, H. G., Jr. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, UK.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK.

Hargett, E. G., J. R. ZumBerge, C. P. Hawkins, and J. R. Olson. 2007. Development of a RIVPACS-type predictive model for bioassessment of wadable streams in Wyoming. Ecological Indicators 7:807–826.

Hawkins, C. P. 2006. Quantifying biological integrity by taxonomic completeness: It's utility in regional and global assessments. Ecological Applications 16:1277–1294.

Hawkins, C. P., Y. Cao, and B. Roper. 2010*a*. Method for predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology 55:1066–1085.

Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010*b*. The reference condition: predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society 29:312–343.

Hering, D., et al. 2010. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of the Total Environment 408:4007–4019.

Hocking, R. R. 2013. Methods and applications of linear models: regression and the analysis of variance. Third edition. John Wiley & Sons, Somerset, New Jersey, USA.

Jyväsjärvi, J., J. Aroviita, and H. Hämäläinen. 2012. Performance of profundal macroinvertebrate assessment in boreal lakes depends on lake depth. Fundamental and Applied Limnology 180:91–100.

Jyväsjärvi, J., J. Aroviita, and H. Hämäläinen. 2014. An extended Benthic Quality Index for assessment of lake profundal macroinvertebrates: addition of indicator taxa by multivariate ordination and weighted averaging. Freshwater Science 33:995–1007.

Kilgour, B. W., K. M. Somers, and D. E. Matthews. 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. Ecoscience 5:542–550.

Linke, S., R. H. Norris, D. P. Faith, and D. Stockwell. 2005. ANNA: A new prediction method for bioassessment programs. Freshwater Biology 50:147–158.

Mayer, D. G., and D. G. Butler. 1993. Statistical validation. Ecological Modelling 68:21–32.

Mazor, R. D., A. C. Rehn, P. R. Ode, M. Engeln, K. C. Schiff, E. D. Stein, D. J. Gillett, D. B. Herbst, and C. P. Hawkins. 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. Freshwater Science 35:249–271.

Nichols, S. J., and F. J. Dyer. 2013. Contribution of national bioassessment approaches for assessing ecological water security: an AUSRIVAS case study. Frontiers of Environmental Science and Engineering 7:669–687.

Pinheiro, J. C., and D. M. Bates. 2000. Mixed-effects models in S and S-plus. Springer, New York, New York, USA.

R Core Team. 2017. R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rose, P. M., M. J. Kennard, D. B. Moffatt, G. L. Butler, and F. Sheldon. 2016*a*. Incorporating species losses and gains into a fish-based index for stream bioassessment increase the detection of anthropogenic disturbances. Ecological Indicators 69:677–685.

Rose, P. M., M. J. Kennard, D. B. Moffatt, G. L. Butler, and F. Sheldon. 2016*b*. Testing three species distribution modeling strategies to define fish assemblage reference conditions for stream bioassessment and related applications. PLoS ONE 11:e0146728.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological Applications 16:1267–1276.

Tsang, Y.-P., G. K. Felton, G. E. Moglen, and M. Paul. 2011. Region of influence method improves macroinvertebrate predictive models in Maryland. Ecological Modelling 222:3473–3485.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.1725/full

Data Availability

Data are available from the Jyväskylä University Digital Archive: http://urn.fi/URN:NBN:fi:jyu-201803221810