Critique of model evaluation by the Washington Department of Ecology

Gordon Holtgrieve
School of Aquatic and Fishery Sciences
University of Washington
Seattle, WA
gholt@uw.edu

Mark Scheuerell
USGS Washington Cooperative Fish and Wildlife Research Unit
School of Aquatic and Fishery Sciences
University of Washington
Seattle, WA
scheuerl@uw.edu

Comparison of existing and reference scenarios

The focus of the modeling analysis is a comparison of results obtained with two scenarios: a "reference" case that represents a system without anthropogenic inputs, and an "existing" case that represents contemporary conditions. Specifically, Ecology is interested in the difference between the modeled concentration of dissolved oxygen estimated via the two models. In addition, Ecology would like to know the estimated uncertainty in that difference.

Variance of predictions

In the section titled "Uncertainty in Dissolved Oxygen Depletion Estimates" (p59), it states,

> The RMSE of differences is calculated to understand the uncertainty associated with the result of subtracting one model scenario from another model scenario (i.e., the difference between two model scenarios). In this case, we calculated the error associated with the DO depletions computed from the difference between the existing and reference model scenarios.

The section then goes on to describe how the calculations were made using the estimated root mean squared error (RMSE) between the predictions and observations, but there is a mistake in the assumed relationship between the standard deviation of the predictions and the RMSE.

Variance of predictions

To demonstrate this, consider this simple equation that relates individual observations (o_i) and predictions (p_i):

$$o_i = p_i + e_i,$$

where e_i are the model prediction errors (i.e., the difference between the observed and predicted values). From this relationship we know that the variance of the observations is a function of the variances of both the predictions and errors, and their covariance, such that

$$\text{Var}(o) = \text{Var}(p) + \text{Var}(e) + 2\ \text{Cov}(p, e)$$

We can rewrite the above equation to show that the variance of the predictions is

$$\text{Var}(p) = \text{Var}(o) - \text{Var}(e) + 2\ \text{Cov}(p, e).$$

Variance in the difference of predictions

In this case Ecology is interested in the uncertainty (variance) in the difference between the predictions from the two models representing existing and reference conditions, which we write as p_{ex} and p_{ref}, respectively. We then define the difference δ as

$$\delta = p_{ex} - p_{ref}$$

and hence

$$\begin{aligned} \text{Var}(\delta) &= \text{Var}(p_{ex}) + \text{Var}(p_{ref}) - 2\ \text{Cov}(p_{ex}, p_{ref}) \\ &= \text{Var}(p_{ex}) + \text{Var}(p_{ref}) - 2\ \text{Cor}(p_{ex}, p_{ref})\ \text{SD}(p_{ex})\ \text{SD}(p_{ref}) \end{aligned}$$

This is where Ecology gets their calculations wrong. In a forecasting context, the hope is that the predictions match the observations very closely and hence the errors are small. One measure of forecast skill is the root mean-squared error (RMSE), which equals the standard deviation of the errors. More specifically,

$$\text{RMSE}_{o,p} = \text{SD}(e) = \sqrt{\text{Var}(e)} = \sqrt{\frac{\sum(p_i - o_i)^2}{N}}.$$

Importantly, however, the $\text{RMSE}_{o,p}$ is not equal to the variance of the predictions, $\text{Var}(p)$, which is required for the calculations of the error in differences.

Re-analysis

The Ecology report does not provide estimates of the variance in the model predictions, but we can generate approximations from the information provided and a simple assumption. For most of the DO models, $\text{RMSE}_{ex} \approx 1$ (Table 7) and the correlation between the predicted and observed values is about 0.85 (Table 8). Recognizing that

$$\text{RMSE}_{ex} = \sqrt{(1 - R^2)}\ \text{SD}(o),$$

we can estimate the SD of the observations as

$$\mathrm{SD}(o) = \frac{\mathrm{RMSE}_{ex}}{\sqrt{(1-R^2)}} \approx \frac{1}{\sqrt{(1-0.85^2)}} \approx 1.9$$

and hence the variance of the observations is

$$\mathrm{Var}(o) = \mathrm{SD}(o)^2 \approx 1.9^2 = 3.61.$$

Now we can estimate the variance of the predictions for the model with existing conditions as above, with

$$\begin{aligned}\mathrm{Var}(p_{ex}) &= \mathrm{Var}(o) - \mathrm{Var}(e) + 2\ \mathrm{Cov}(p_{ex}, e)\\ &= \mathrm{Var}(o) - \mathrm{RMSE}_{ex}^2 + 2\ \mathrm{Cov}(p_{ex}, e)\\ &\approx 3.6 - 1^2 + 2\ \mathrm{Cov}(p_{ex}, e).\end{aligned}$$

Absent information on the covariance between the predicted values and the model errors, we will assume that the model is well behaved and $\mathrm{Cov}(p_{ex}, e) \approx 0$, such that

$$\mathrm{Var}(p_{ex}) \approx 3.6 - 1^2 + 2(0) = 2.6$$

To the extent that $\mathrm{Cov}(p_{ex}, e)$ is positive (negative), $\mathrm{Var}(p_{ex})$ will be larger (smaller) than this estimate.

If we also assume, as Ecology did, that $\mathrm{Var}(p_{ex}) = \mathrm{Var}(p_{ref})$, then we can estimate the variance in the difference (δ) between the predictions from the two models as above, such that

$$\begin{aligned}\mathrm{Var}(\delta) &= \mathrm{Var}(p_{ex}) + \mathrm{Var}(p_{ref}) - 2\ \mathrm{Cor}(p_{ex}, p_{ref})\ \mathrm{SD}(p_{ex})\ \mathrm{SD}(p_{ref})\\ &= \mathrm{Var}(p_{ex}) + \mathrm{Var}(p_{ex}) - 2\ \mathrm{Cor}(p_{ex}, p_{ref})\ \mathrm{SD}(p_{ex})\ \mathrm{SD}(p_{ex})\\ &= 2\ \mathrm{Var}(p_{ex}) - 2\ \mathrm{Cor}(p_{ex}, p_{ref})\ \mathrm{Var}(p_{ex})\\ &= 2\ \mathrm{Var}(p_{ex})\,(1 - \mathrm{Cor}(p_{ex}, p_{ref}))\\ &= 2(2.6)\,(1 - \mathrm{Cor}(p_{ex}, p_{ref}))\,.\end{aligned}$$

Thus, if $\mathrm{Cor}(p_{ex}, p_{ref}) = 0$, then $\mathrm{Var}(\delta) = 5.2 \Rightarrow \mathrm{SD}(\delta) \approx 2.3$; conversely, as $\mathrm{Cor}(p_{ex}, p_{ref}) \to 1$ then $\mathrm{Var}(\delta) \to 0$.

Although Ecology's report did not say what $\mathrm{Cor}(p_{ex}, p_{ref})$ was, but we can estimate it from the calculations on p59. For example, if we assume that $\mathrm{Var}(\delta) = 0.041$ as for Ecology's model in 2014, then analogous to above we have

$$\text{Var}(\delta) = \text{Var}(p_{ex}) + \text{Var}(p_{ref}) - 2\ \text{Cor}(p_{ex}, p_{ref})\ \text{SD}(p_{ex})\ \text{SD}(p_{ref})$$

$$\Downarrow$$

$$\begin{aligned}\text{Cor}(p_{ex}, p_{ref}) &= \frac{(\text{Var}(\delta) - \text{Var}(p_{ex}) - \text{Var}(p_{ref}))}{-2\ \text{SD}(p_{ex})\ \text{SD}(p_{ref})} \\ &\approx \frac{(0.041 - 1^2 - 1^2)}{-2(1)(1)} \\ &\approx 0.98\end{aligned}$$

This correlation is remarkably high, indicating that the two models produce nearly identical predictions of DO. Inserting this correlation coefficient into the equation for $\text{Var}(\delta)$ gives $\text{Var}(\delta) = 2(2.6)(1 - 0.98) = 0.104$, and hence $\text{SD}(\delta) \approx 0.32$. This value is about eight times greater than those reported in Ecology's document. Thus, if the treshhold concentration for DO depletion is 0.2 mg/L, then the estimated coefficient of variation (CV) around it is 160%.

Example of SD versus RMSE

Here is a simple example that shows how $\text{SD}(\hat{y})$ and $\text{RMSE}(\hat{y})$ are different. Consider a case where we had reason to believe that a variable y was a function of another variable x. In effort to undercover the nature of their relationship, we collected 20 samples of both y and x (Figure 1).

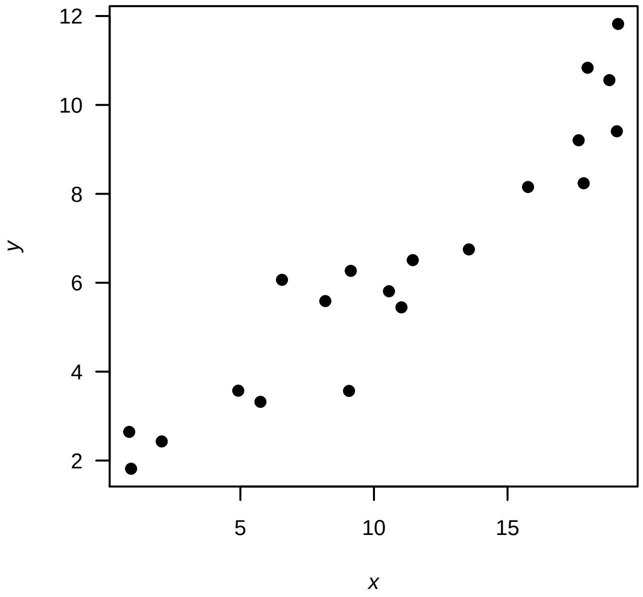

Figure 1. Plot of some hypothetical data.

Based on the apparent relationship between x and y, we might assume that each of the observed values y_i is a linear combination of an intercept β_0, the effect β_1 of a covariate x_i, and some random observation error ϵ_i, such that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

and $\epsilon_i \sim \mathrm{N}(0, \sigma)$. We could easily estimate the unknown parameters in this model (β_0, β_1, σ), and then use the deterministic portion of the model to make predictions to compare with each of the observed values. Specifically, the predictions ($\hat{y}_i$) would be given by a straight line, such that

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

We could then estimate the SD of these predictions and the model's RMSE (Figure 2). It turns out that the SD of $\hat{y}$ is ~2.82, but the RMSE is only ~0.94, which is about 3 times less.

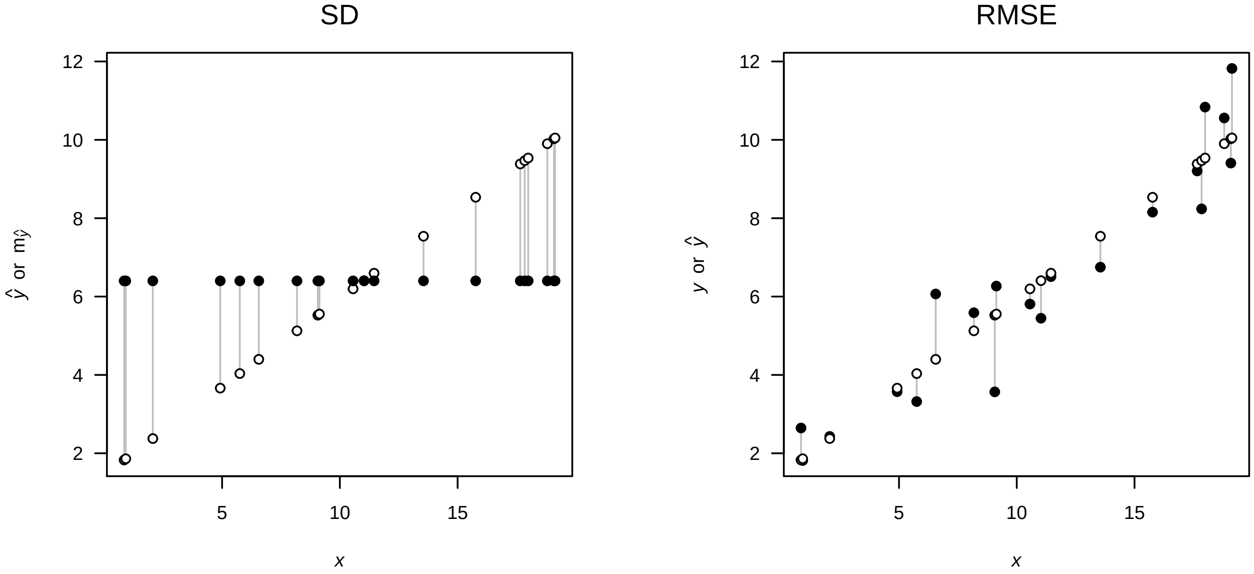

Figure 2. Graphical examples of the difference between the SD of the predictions (left) and the RMSE of the predictions (right). For the SD, the comparison is based upon the differences between the predictions (open circles) and their mean (filled circles). For the RMSE, the comparison is based upon differences between the predictions (open circles) and the observed data (filled circles). In both cases, one would square the length of each of the vertical gray lines, sum them up, and divide by the number of them before finally taking the square root.

Prediction errors

The above example dismisses an important aspect of RMSE: it should be used to compare "out of sample" predictions. Furthermore, RMSE give us an indication as to the predictive error, *on average*, rather than the uncertainty in a specific prediction.

Returning to our example above, we could estimate our uncertainty around the fitted relationship between x and y with a confidence interval (CI), which would give us an indication of the range of where the "true" fitted values would lie had we repeated our sampling exercise many times.

Specifically, a $(1-\alpha)100\%$ CI on the expected relationship between x and y at some value x_k is given by

$$\hat{y}_i \pm t_{\alpha/2,n-2}\sqrt{\sigma\left(\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}.$$

The interval increases as the distance between x_k and $\bar{x}$ increases (Figure 3).

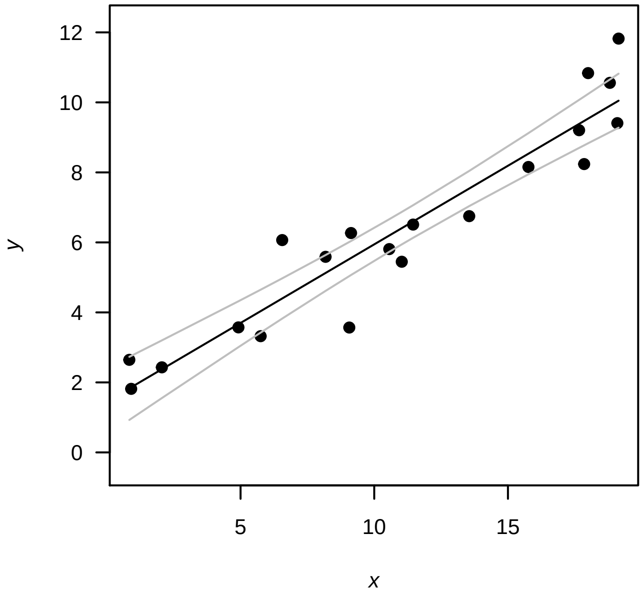

Figure 3. Example of a 95% confidence interval (gray lines) around the expected relationship between x and y (black line).

In a case like this, however, where we wish to make out-of-sample predictions about some new state of nature, our uncertainty around any single prediction will be necessarily greater. Specifically, a $(1-\alpha)100\%$ prediction interval (PI) around $\hat{y}$ at some value x_k is given by

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{\sigma\left(1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}.$$

Here the paranthetic multiplier on the residual variance σ has increased by 1, which means the prediction interval is wider (less certain) than the confidence interval (Figure 4). This is because the CI only needs to account for uncertainty in estimating the expected value of y whereas the PI needs to account for a random future value of y that tend to fall away from the mean.

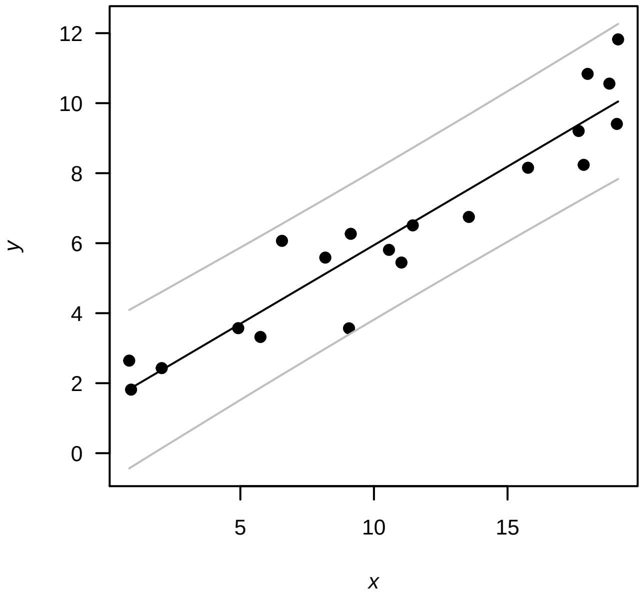

Figure 4. Example of a 95% prediction interval (gray lines) for future unobserved values of y.

So, for example, if we wanted to predict, with 95% certainty, what we would observe for y if $x = 10$, we would get 5.94 ± 2.13 (Figure 5). The relatively wide prediction interval suggests that it might be difficult to discern the prediction for y when $x = 10$ to the expected values for y if x were as low as 5 or as high as 15.

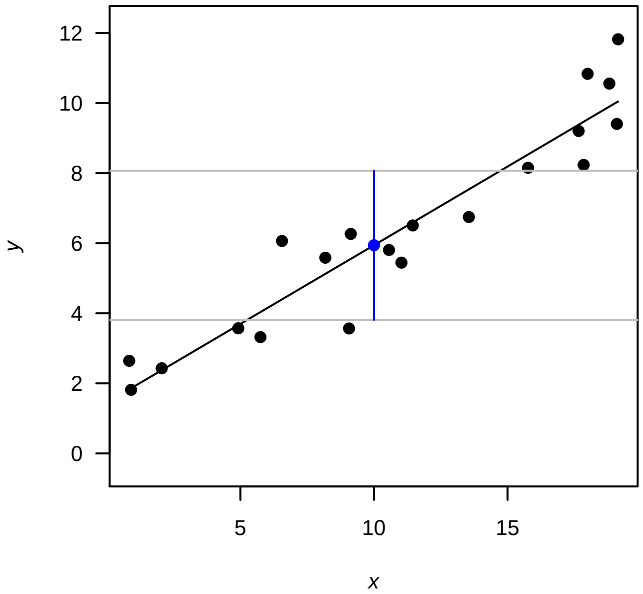

Figure 5. Example of the uncertainty around a new prediction for y when $x = 10$.