Critique of model evaluation by the Washington Department of Ecology

Version 2.0

2 June 2020

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.

We thank Ecology for their clarifying statements in response to our initial critique. Those details that were lacking or unclear in the report are indeed important. This document now more clearly explains our understanding of the issues and how they might be addressed in a formal manner.

Model comparisons

We begin by defining the following terms:

- $O_{e,i}$ is the observed concentration of dissolved oxygen under existing conditions at location i
- $P_{e,i}$ is the predicted concentration of dissolved oxygen under existing conditions at location i
- N_e is the number of locations with both $O_{e,i}$ and $P_{e,i}$ for the same i
- $O_{r,i}$ is the observed concentration of dissolved oxygen under reference conditions at location i
- $P_{r,i}$ is the predicted concentration of dissolved oxygen under reference conditions at location i
- N_r is the number of locations with both $O_{r,i}$ and $P_{r,i}$ for the same i

We note here that the $O_{r,i}$ are *unknown quantities*. We further define the model error under each condition as the difference between the observed and predicted dissolved oxygen (DO) concentrations, such that

$$\begin{aligned} \epsilon_{e,i} &= O_{e,i} - P_{e,i} \\ \epsilon_{r,i} &= O_{r,i} - P_{r,i} \end{aligned} \tag{1}$$

We now define the vectors of the predictions errors to be

$$\boldsymbol{\epsilon}_e = \begin{bmatrix} \epsilon_{e,1} \\ \epsilon_{e,2} \\ \vdots \\ \epsilon_{e,N_e} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_r = \begin{bmatrix} \epsilon_{r,1} \\ \epsilon_{r,2} \\ \vdots \\ \epsilon_{r,N_r} \end{bmatrix} \tag{2}$$

The $\boldsymbol{\epsilon}$ are then assumed to come from multivariate normal distributions (MVN) with

$$\begin{aligned} \boldsymbol{\epsilon}_e &\sim \text{MVN}(\boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e) \\ \boldsymbol{\epsilon}_r &\sim \text{MVN}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) \end{aligned} \tag{3}$$

If the model is unbiased for the existing and reference conditions, then the mean vectors $\boldsymbol{\mu}_e$ and $\boldsymbol{\mu}_r$ will both be $\mathbf{0}$. In the most simple form, all of the errors are independent and identically distributed (IID), such that

$$\boldsymbol{\Sigma}_e = \boldsymbol{\Sigma}_r = \sigma^2 \mathbf{I} \tag{4}$$

where $\mathbf{I}$ is an $[n \times n]$ identity matrix. Here, however, it seems more reasonable to assume the following:

1) the prediction errors are not necessarily independent of one one another (i.e., locations near one another will be more similar than those farther away);
2) the variance of the prediction errors is not necessarily constant among the different sampling locations (i.e., the predictive ability of the model varies spatially); and
3) the variances and covariances are different in the existing and reference conditions (i.e., $\boldsymbol{\Sigma}_e \neq \boldsymbol{\Sigma}_r$.

This implies that

$$\begin{aligned} \boldsymbol{\Sigma}_e &= \sigma^2 \mathbf{C}_e \\ \boldsymbol{\Sigma}_r &= \sigma^2 \mathbf{C}_r \end{aligned} \tag{5}$$

where σ^2 is a scalar and $\mathbf{C}_e$ and $\mathbf{C}_r$ are $[N_e \times N_e]$ and $[N_r \times N_r]$ matrices, respectively, of (potentially) non-zero coefficients for the existing and reference conditions.

Differences in prediction errors

If we are interested in the differences between the prediction errors in the existing ($\boldsymbol{\epsilon}_e$) and reference conditions ($\boldsymbol{\epsilon}_r$), such that

$$\boldsymbol{\delta} = \boldsymbol{\epsilon}_e - \boldsymbol{\epsilon}_r \tag{6}$$

then

$$\boldsymbol{\delta} \sim \text{MVN}(\boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta) \tag{7}$$

with

$$\boldsymbol{\mu}_\delta = \boldsymbol{\mu}_e - \boldsymbol{\mu}_r \tag{8}$$

$$\boldsymbol{\Sigma}_\delta = \boldsymbol{\Sigma}_e + \boldsymbol{\Sigma}_r \tag{9}$$

However, because the $O_{r,i}$ are *unknown quantities*, we have no way of estimating $\boldsymbol{\epsilon}_r$, $\boldsymbol{\mu}_r$, or $\boldsymbol{\Sigma}_r$ and hence no way of calculating $\boldsymbol{\delta}$, $\boldsymbol{\mu}_\delta$, or $\boldsymbol{\Sigma}_\delta$. One possible solution to this problem is to assume that $\boldsymbol{\mu}_r = \boldsymbol{\mu}_e$ and $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_e$, such that

$$\begin{aligned} \boldsymbol{\mu}_\delta &= \boldsymbol{\mu}_e - \boldsymbol{\mu}_r \\ &= \boldsymbol{\mu}_e - \boldsymbol{\mu}_e \\ &= \mathbf{0} \end{aligned} \tag{10}$$

$$\begin{aligned} \boldsymbol{\Sigma}_\delta &= \boldsymbol{\Sigma}_e + \boldsymbol{\Sigma}_r \\ &= \boldsymbol{\Sigma}_e + \boldsymbol{\Sigma}_e \\ &= 2\boldsymbol{\Sigma}_e \end{aligned} \tag{11}$$

We can decompose the covariance matrix $\boldsymbol{\Sigma}_e$ in terms of the standard deviations of each of the $\epsilon_{e,i}$ (σ_i) and the correlations among the $\epsilon_{e,i}$ ($\rho_{i,j}$), such that

$$\boldsymbol{\Sigma}_e = \mathbf{SRS} \tag{12}$$

with

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \end{bmatrix} \tag{13}$$

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{1,2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{n-1,n} \\ \rho_{1,n} & \cdots & \rho_{n-1,n} & 1 \end{bmatrix} \tag{14}$$

As we understand it, Ecology assumes that $\sigma_i = \sigma \; \forall \; i$ and that $\hat{\sigma} \approx 1$ (Table 7 in Ahmed et al. 2019). If so, then

$$\mathbf{S} \approx \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{15}$$

and hence

$$\mathbf{\Sigma}_\delta \approx 2\mathbf{R} \tag{16}$$

At this point, however, we have no information on the $\rho_{i,j}$ within $\mathbf{R}$.

Questions

- Are the comparisons between observed and predicted [DO] based on the raw values or log-transformed [DO]?
- What are the sample sizes N_e and N_r?
- Are the N_e different $O_{e,i}$ data points novel in that they were not used in the model tuning process and instead held back for out-of-sample predictions $P_{e,i}$?
- Where are the N_e sampling locations?