Goodness-of-Fit Statistics

Root Mean Square Error (RMSE):

The Root Mean Square Error (RMSE) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

The **RMSE** of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}}$$

where X_{obs} is observed values and X_{model} is modelled values at time/place i.

The calculated **RMSE** values will have units, and RMSE for phosphorus concentrations can for this reason not be directly compared to RMSE values for chlorophyll a concentrations etc. However, the RMSE values can be used to distinguish model performance in a calibration period with that of a validation period as well as to compare the individual model performance to that of other predictive models.

Correlation Coefficient, *r*:

The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

The mathematical formula for computing r is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}},$$

where n is the number of pairs of data.

The value of r is such that $-1 < r < +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative.

A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Coefficient of Determination, r^2 or R^2:

The coefficient of determination, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable.

It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
The coefficient of determination is the ratio of the explained variation to the total variation. The coefficient of determination is such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y.

The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if r = 0.922, then r^2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.