Virtual Beach 3.0.6 - Building and Evaluating an MLR Model

In this module you will learn how to:

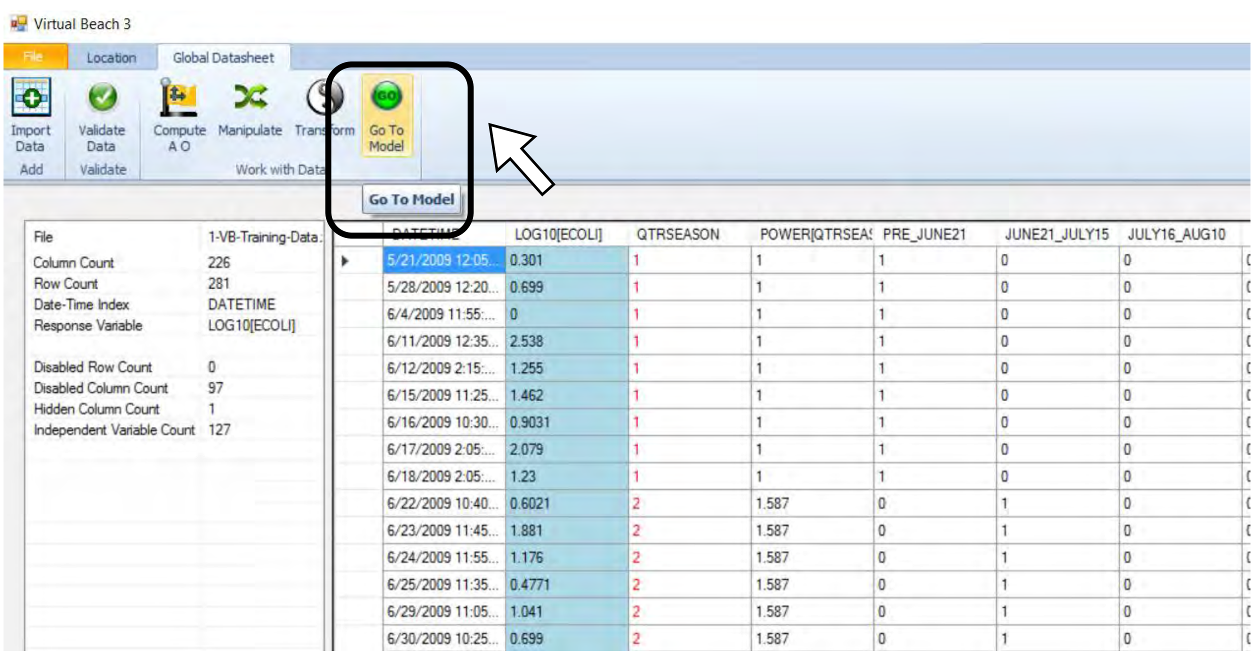
A. Set-up and run an 'MLR' model-building and optimization routine
B. Evaluate top-rated MLR models
C. Set MLR decision criteria
D. Evaluate MLR residuals and search for influential outliers

Multiple Linear Regression (MLR) is the traditional method for developing and operating Nowcast models. MLR is especially well-suited to create nowcast water quality models on days when beach monitoring personnel go into the field to collect samples and conduct routine sanitary surveys.

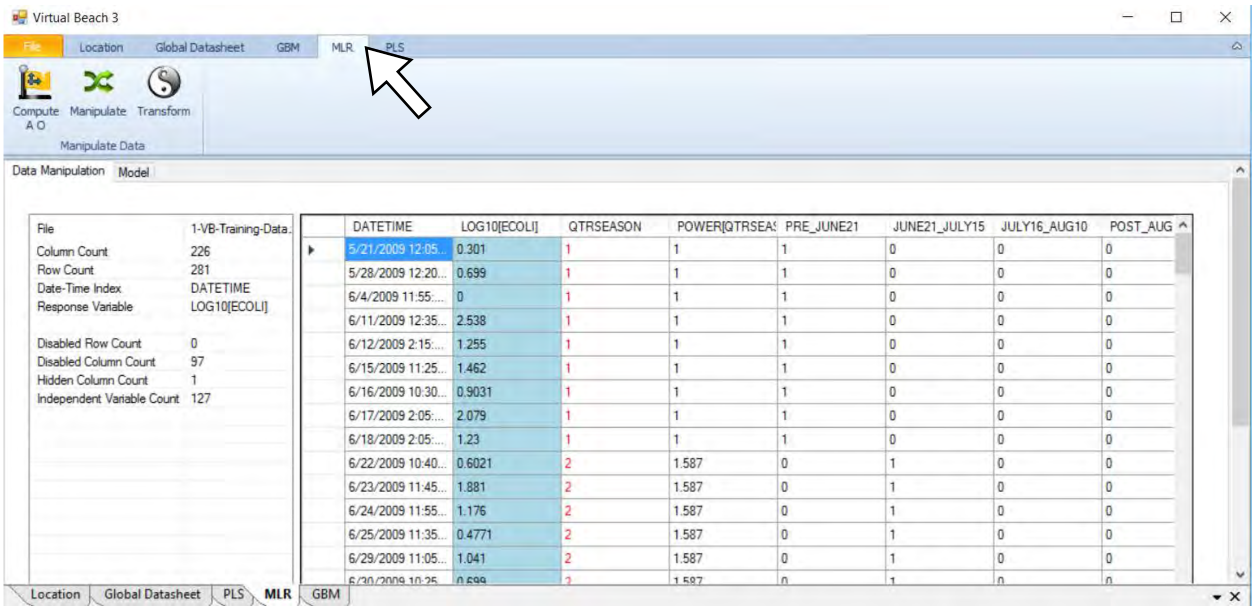
MLR models have the advantage of producing models with only a few independent variables that are easy to interpret. With a limited number of variables, it is easier to determine what factor is effecting water quality on a given day. One disadvantage is that MLR entails many more process-steps and decision-making along the way, compared to PLS and GBM. Virtual Beach has several tools that make the process as efficient as possible.

A. Set-up and run an 'MLR' model-building and optimization routine

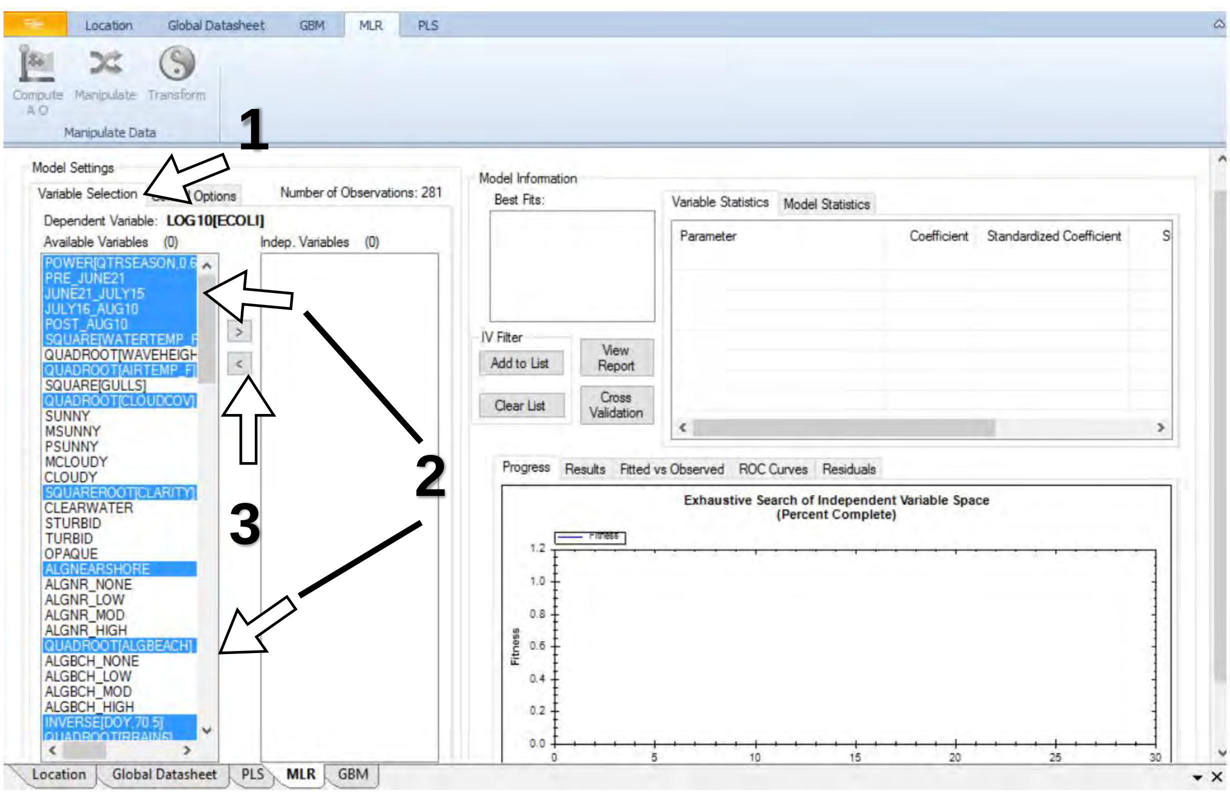
A.1. Open the file saved at the end of the "Virtual_Beach_3.0.6_Data_Prep-MLR" module. In the Global Datasheet tab, click the "Go to Model" button



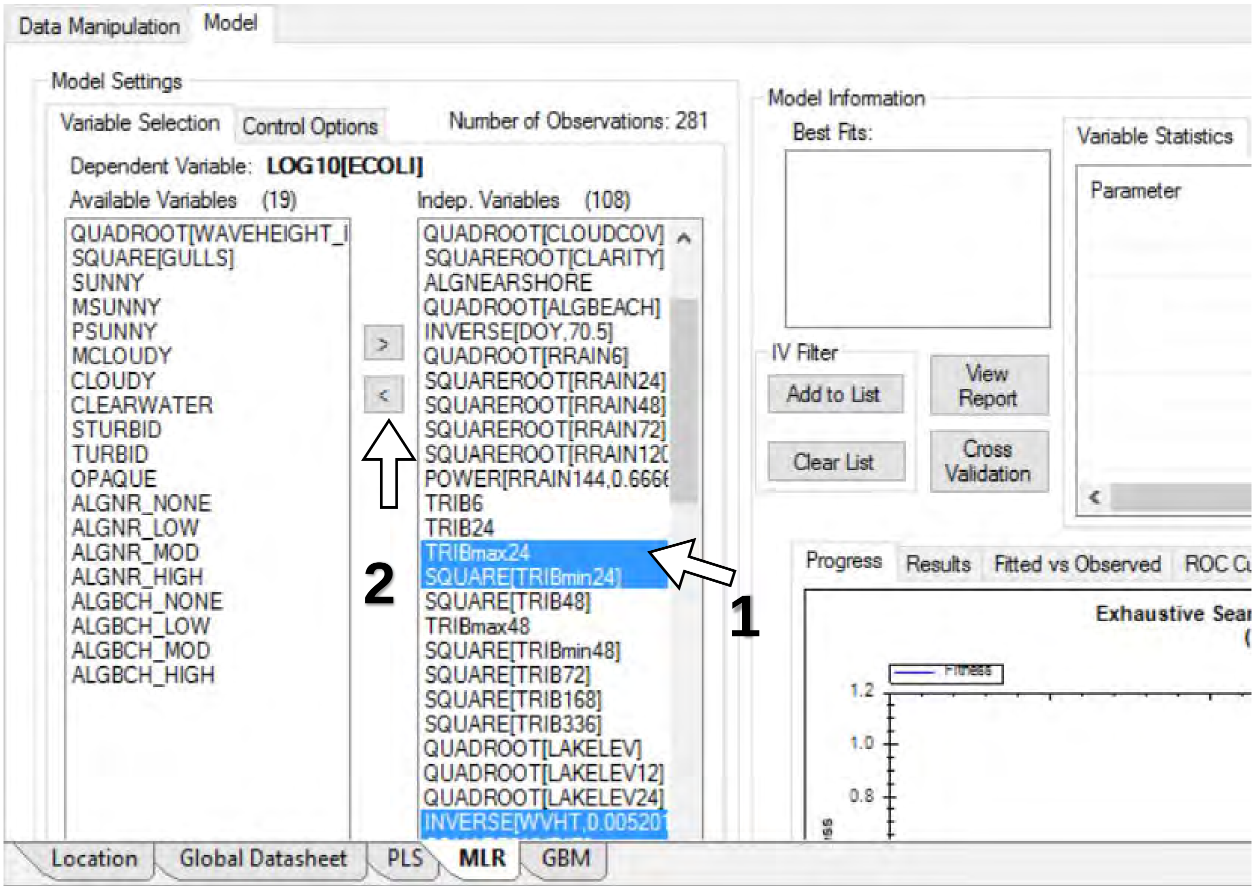
A.2. Click on the 'MLR' tab. A copy of the main data table will open.



A.3. **1.** Click the "Model" sub-tab. **2.** Under "Available Variables" select, (Control -click) potential variables to use for building the model. Select the summary variables: CLOUDCOV, CLARITY, ALGNEARSHORE, ALGBEACH. **3.** Click the right-arrow "**›**" button to move the selected variables to the right-hand panel.



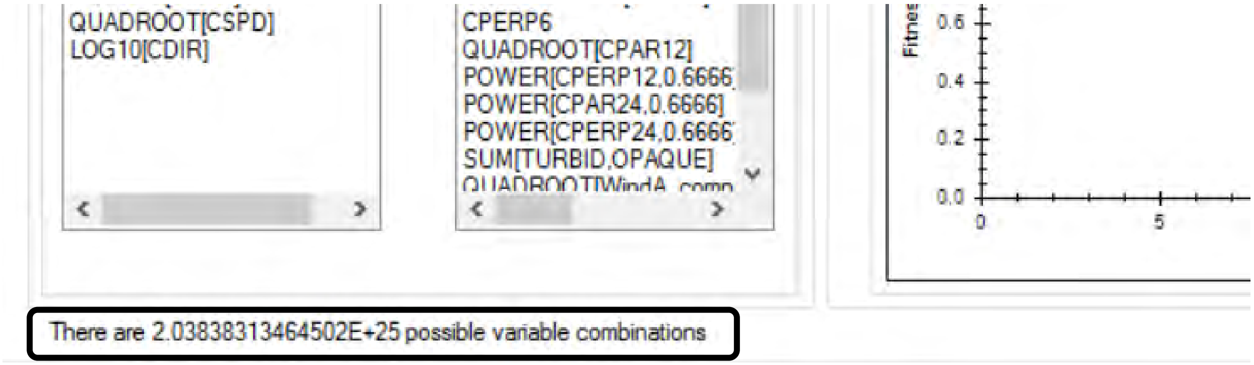
A.4. Do *not* select any variables from the “Available Variable” box if used for alongshore and offshore vector calculations or manipulations such as “compute A-O”. These variables used include WAVEHEIGHT_FT, GULLS, OPAQUE, TURBID, WVHT, WVDIR, WSPD, WDIR, CDIR, CSPD, TRIBMIN24, and TRIBMAX24. **1.** Highlight any of these if they are in the “Indep. Variable” box. **2.** Click the left-arrow “**‹**” button to return them to the “Available Variable” box. In this example, there will be 100 Independent Variables.



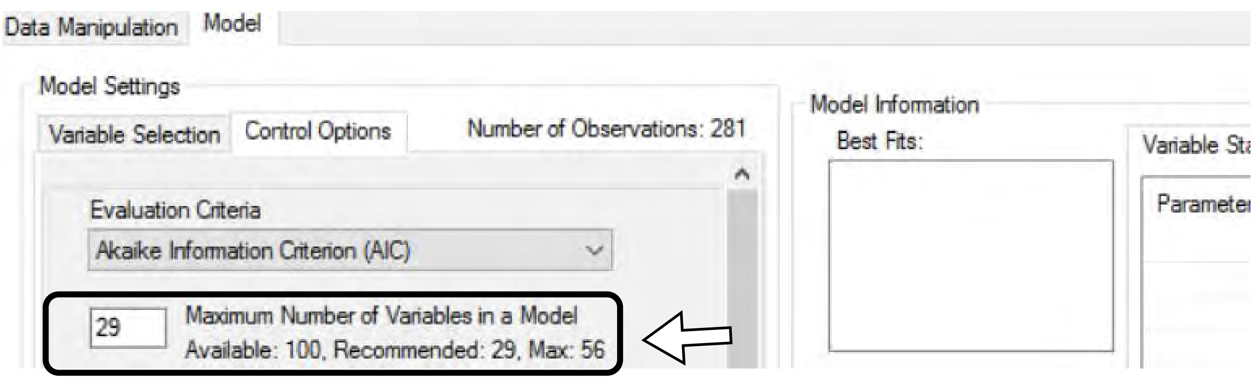
One of the assumptions of building models is that the variables are truly independent, meaning one variable does not influence, or is not correlated with, any other variable. Since in the real world, variables do influence each other, only one of the correlated variables should be chosen when constructing a model.

The Virtual Beach software does conduct a “variance inflation factor screen” to catch and remove any model that has highly correlated independent variables. In that case, any model with both TURBID and TURBID+OPAQUE as independent variables would be removed since those two variables are highly correlated. It is not critical to catch every case of correlated variables, but trying to remove as many as possible is a good practice.

A.5. Virtual Beach calculates how many MLR models can be generated with 100 variables and displays that number below the variable selection boxes. In this example, 20 septillion models!

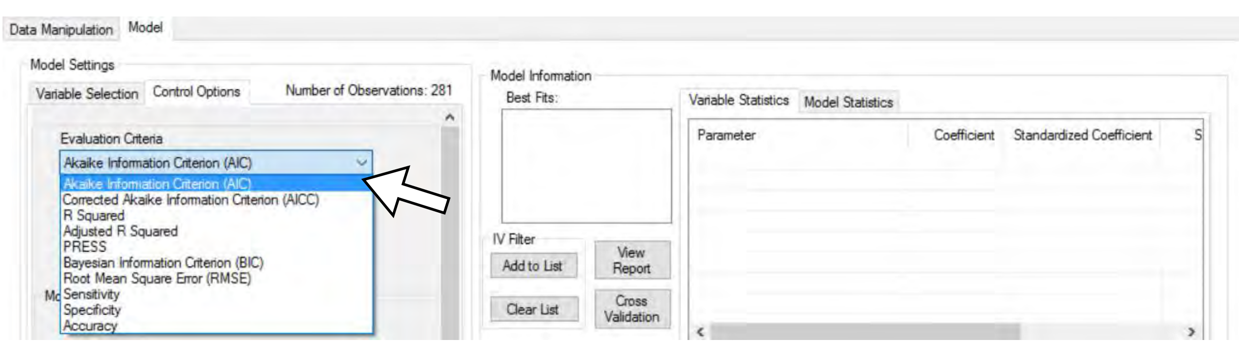


A.7. Under the Model Settings, click the “Control Options” sub-tab to view options for the model-building/ optimization routine. Note that the recommended maximum number of variables for an MLR model with 281 observations is **<u>29</u>.**

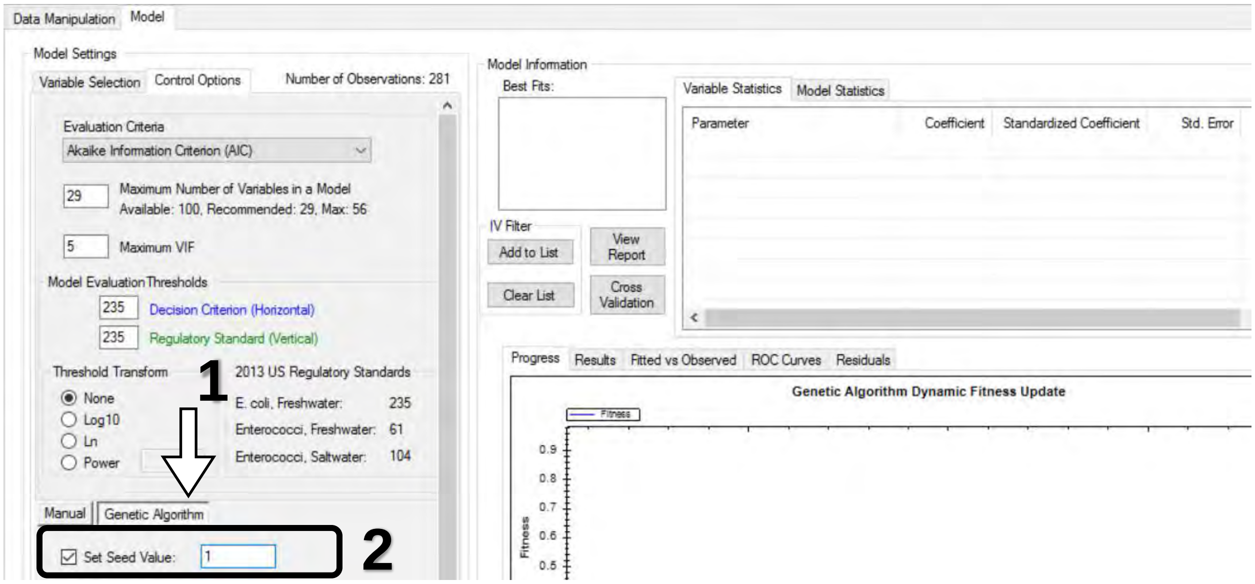


B. Remove Extraneous and Insignificant Variables

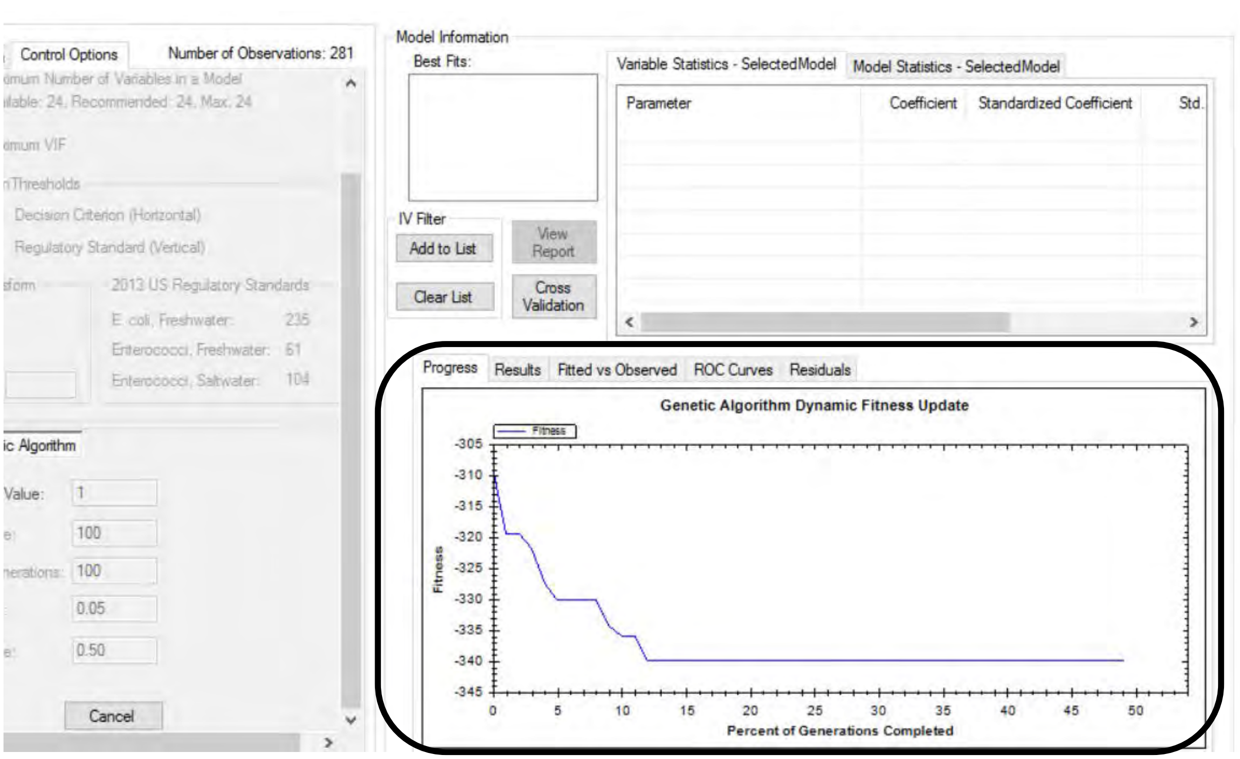
B.1. Click the menu under “Evaluation Criteria” to view the options for evaluating and ranking models. These are various statistical approaches for identifying variables considered insignificant for predicting E. coli concentrations. Keep the default choice, Akaike Information Criteria (AIC). AIC is moderately restrictive in terms of weeding out insignificant variables.



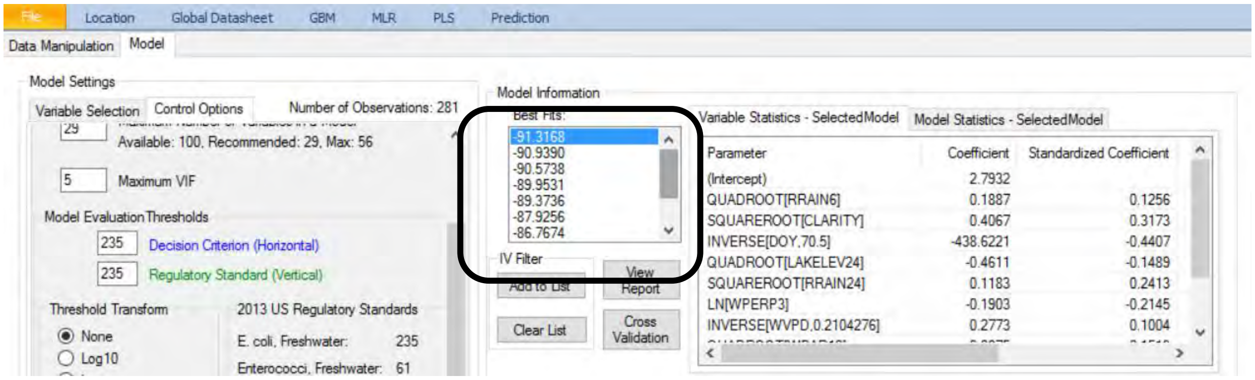
B.2. **1.** Click the “Genetic Algorithm” (GA) button. GA simulates the evolution of a “population” of models over multiple “generations.” **2.** Click “Set Seed Value” and use the default of 1 which makes the algorithm reproducible.



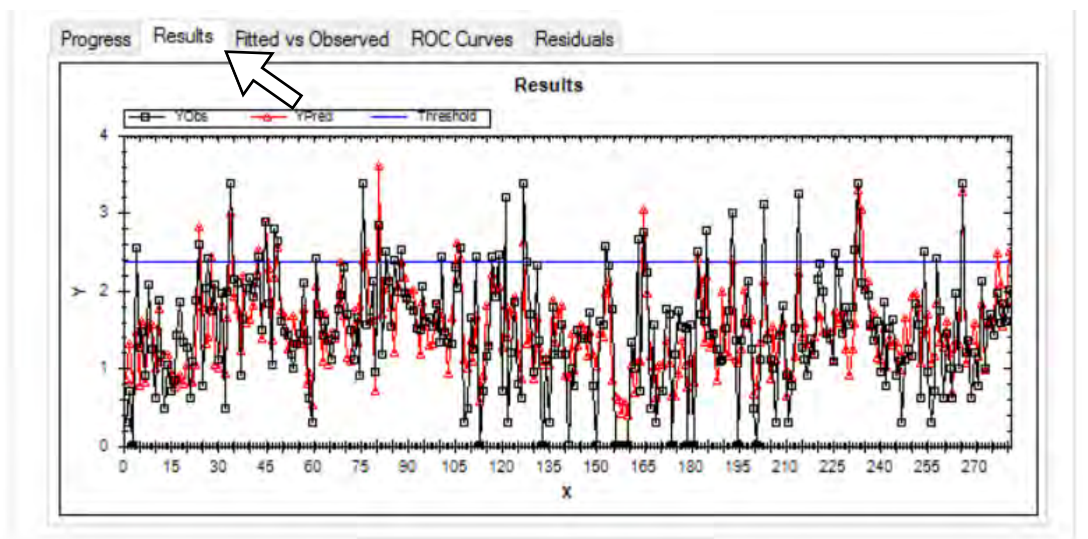
B.3. Click “Run.” VB will now begin analyzing various combinations of variables and calculating the corresponding equations. While the GA optimization is running, you will see a blue line progressing across a graph. The Y Axis (“Fitness”) shows the successive models’ values with respect to the selected evaluation criteria; in this example, AIC.



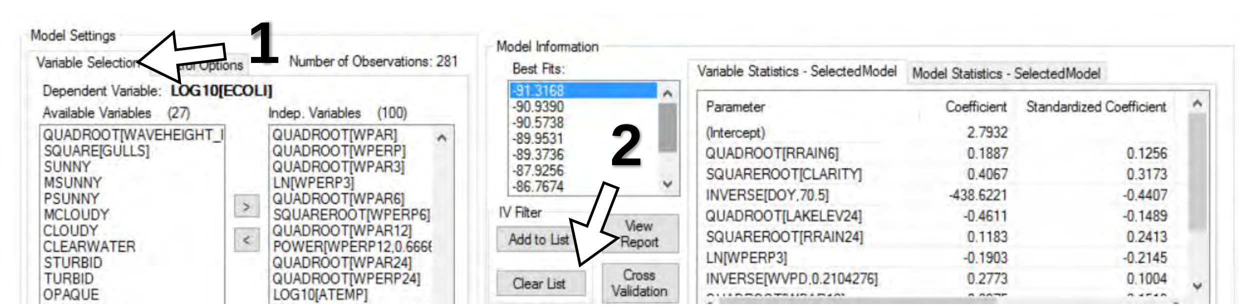
B.4. When the model-building/optimization routine is completed, a “Top 10” list of the models with the “Best Fit” (lowest AIC’s) will appear, ranked by AIC value.



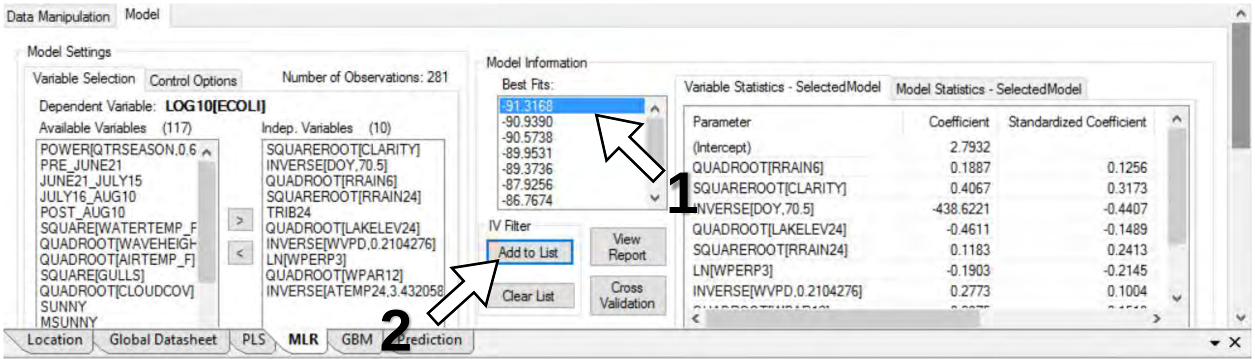
B.5. Near the center of the screen click on the “Results” sub-tab, to view a comparative plot of predicted- versus observed *E. coli* (Y), in log scale over time (X). The horizontal blue line corresponds to 235 CFU/ 100 mL.



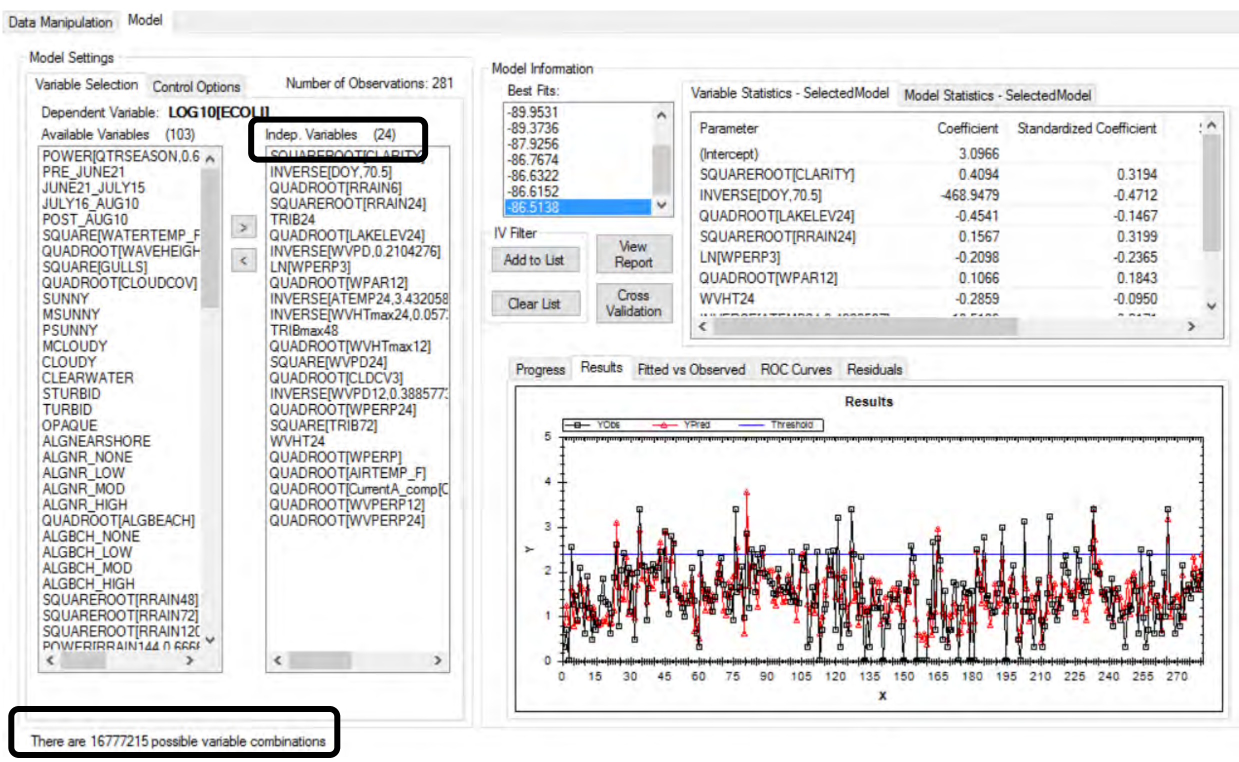
B.6. **1.** Click the “Variable Selection” sub-tab to return to the list of potential variables. Typically, a variable set of 15 or fewer potential variables will allow you to exhaustively evaluate all potential models. Now you will use the “IV Filter” tool to remove less-significant variables from the list. **2.** Near the center of the screen, click “Clear List”. The 100 previously selected independent variables will be cleared.



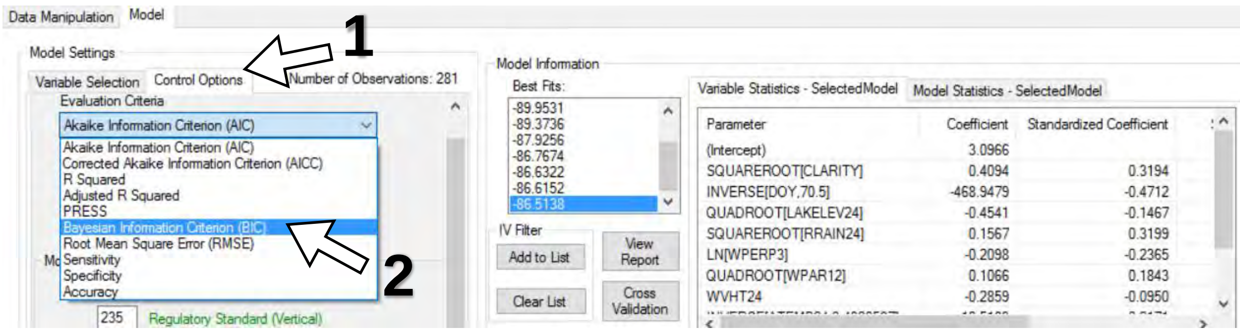
B.7. **1.** With the #1 “Best Fit” model selected, highlighted blue, **2.** Click the button “Add to List.” All of the variables included in that model will be re-added to the list of selected variables. In this example, 10 variables are re-added.



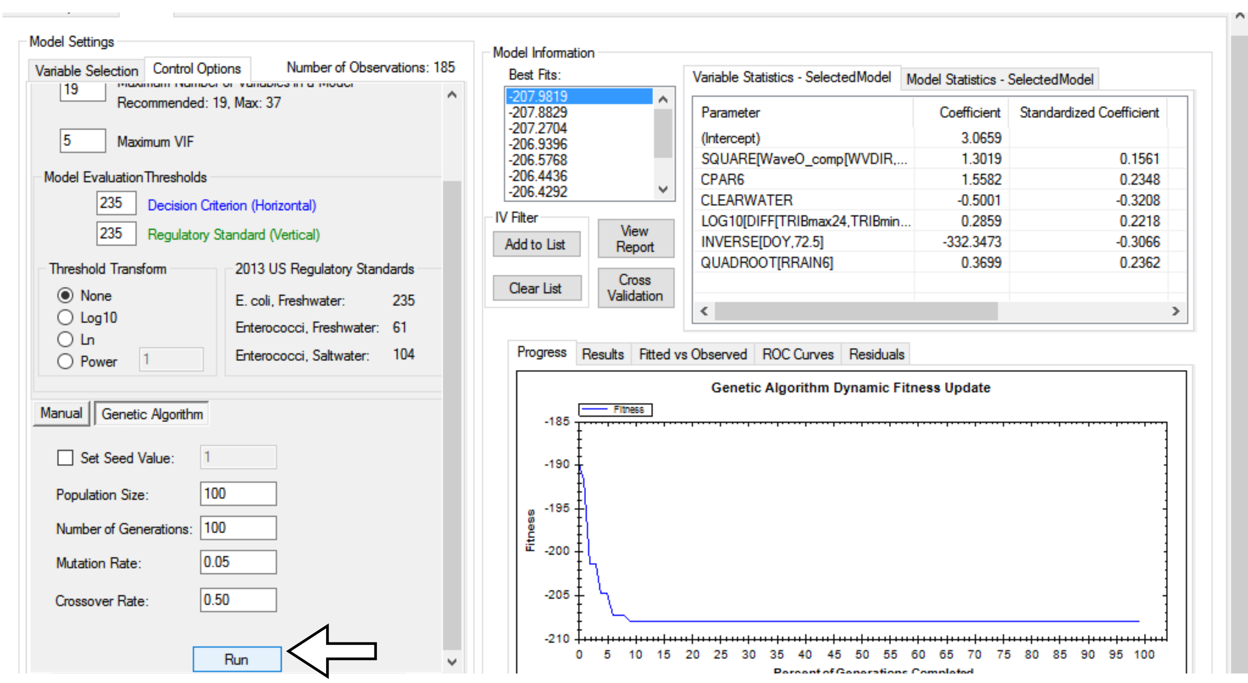
B.8. Add back in the variables for each of the remaining top 10 “Best Fit” models by clicking on the next model and then clicking on the “Add to List” button. At the end, this process results in a “filtered” set of 24 potential variables. Note that the number of potential MLR models is now approximately 16 million– still too many for an exhaustive evaluation of all possible models.



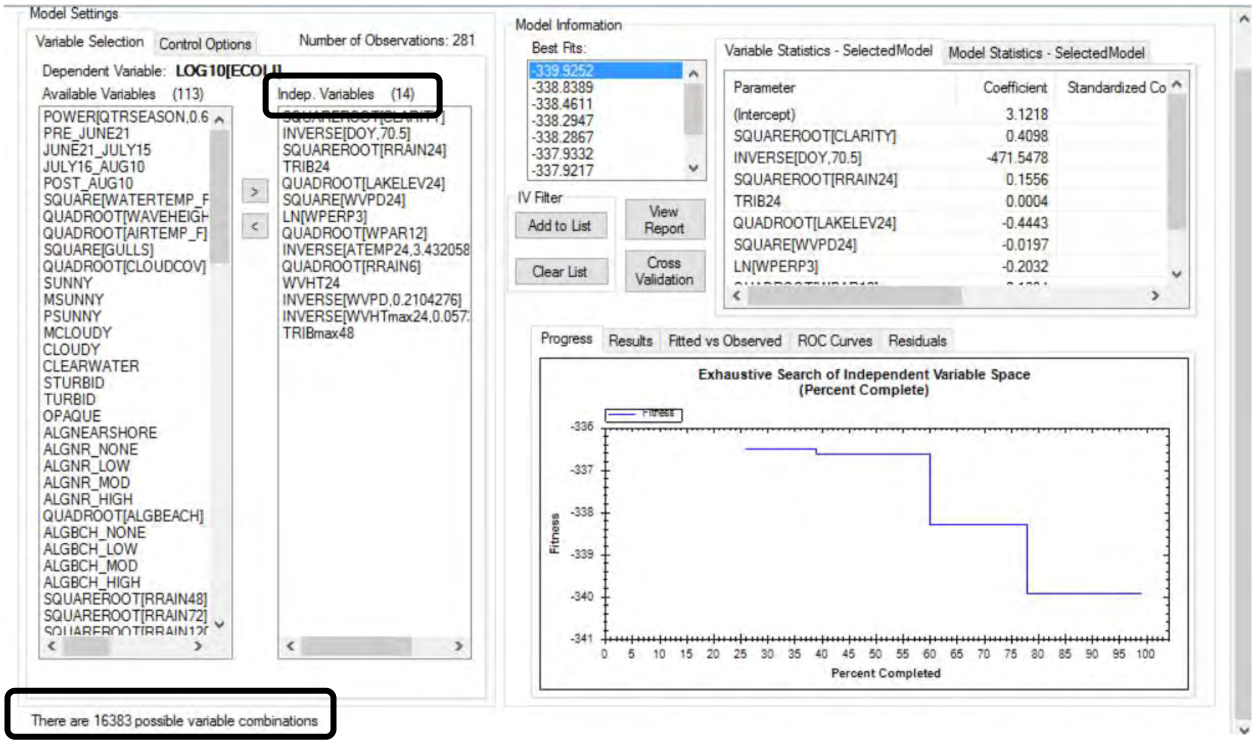
B.9. **1.** Return to the “Control Options” sub-tab. **2.** This time, select the Bayesian Information Criteria (BIC) under the “Evaluation Criterion” drop-down menu. BIC is more restrictive than AIC in terms of weeding-out insignificant variables.



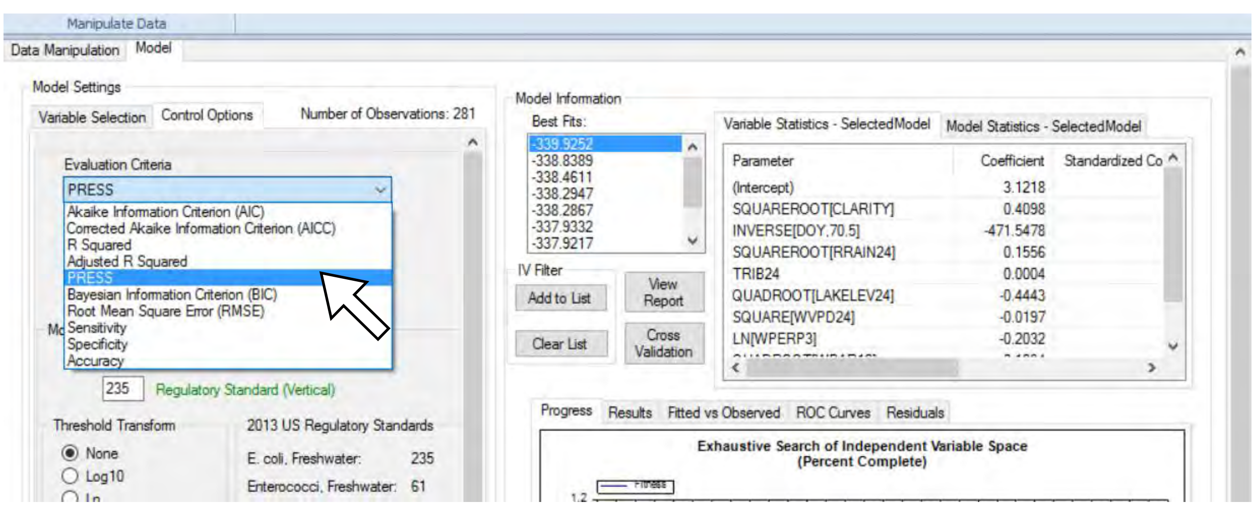
B.10. On the bottom of the “Control Options” subtab, click “Run”. When the Genetic Algorithm optimization completes, repeat the “IV Filter” process (in Steps B.7 – B.9) to reduce the number of potential MLR models.



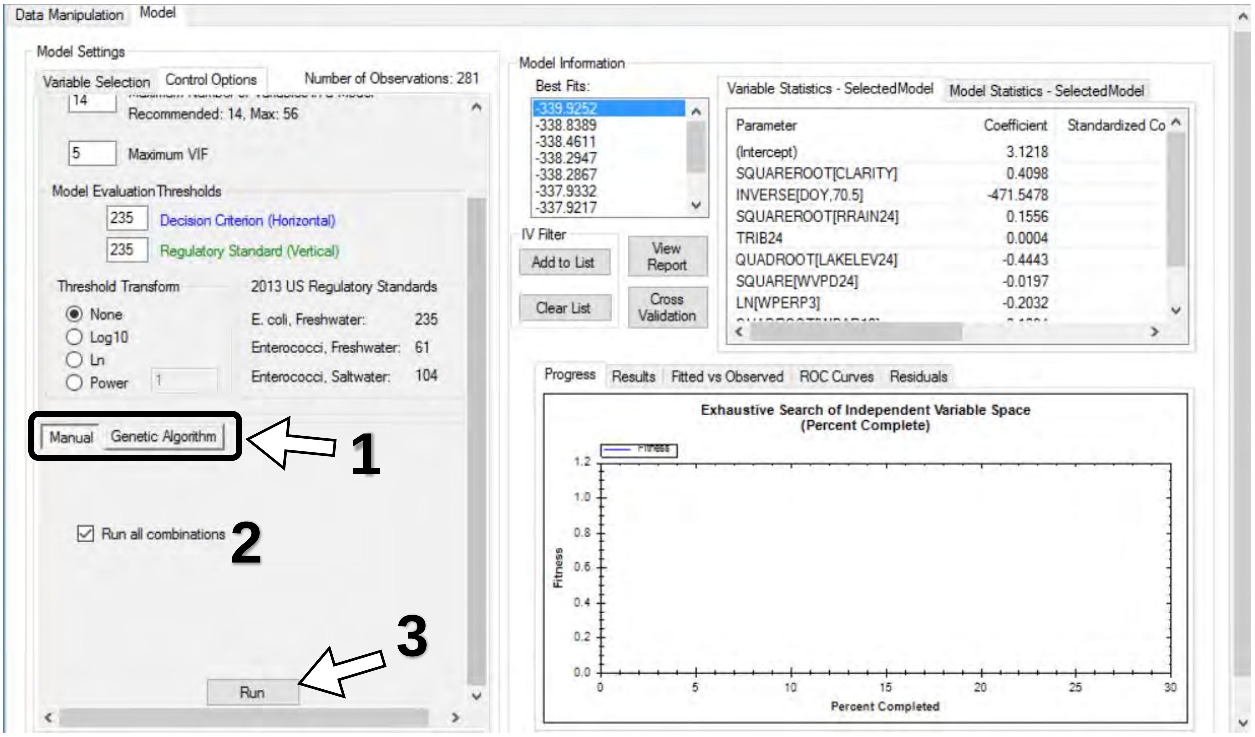
B.12. After completing the second "Independent Variable (IV) Filter" using the more restrictive BIC criteria, we are down to 14 potential variables, or 16,383 potential MLR models. That is a small enough number for an exhaustive evaluation, so now we can focus on improving the model's predictive power.



B.13. Return to the "Control Options" sub-tab. This time select "PRESS". PRESS is the sum of squared prediction-errors generated by removing 1 observation at a time and re-fitting to predict that observation. It is less restrictive in terms of model size and statistical significance but is more focused on predictive power.



B.14. Click the "Manual" evaluation button. **2.** Check "Run all combinations", that is an exhaustive evaluation of all possible models. **3.** Click "Run".



C. Evaluate top-rated MLR models

When building and evaluating potential models for operational use as water-quality nowcasts or forecasts, it is important to understand the difference between a model's **fit** and its **predictive power**.

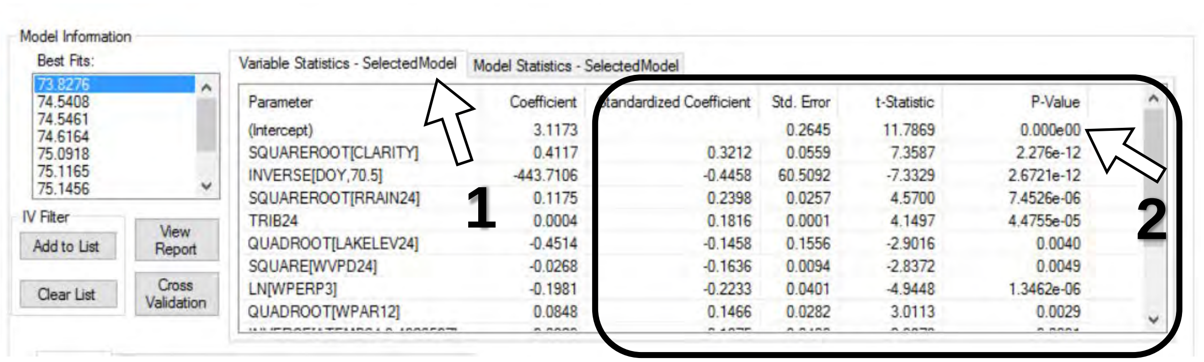
Fit refers to how well a model estimates the response variable, such as the $\log_{10}$ value of *E. coli* over the model's training period. That is, how well it retroactively predicts the observations that were used to build the model.

Predictive power refers to how well a model predicts the response variable on days falling outside of the training period. Cross-validation (C.5-7) measures predictive power, but does so retroactively. The ultimate measure of a model's predictive power is how well it performs when used in the real world.

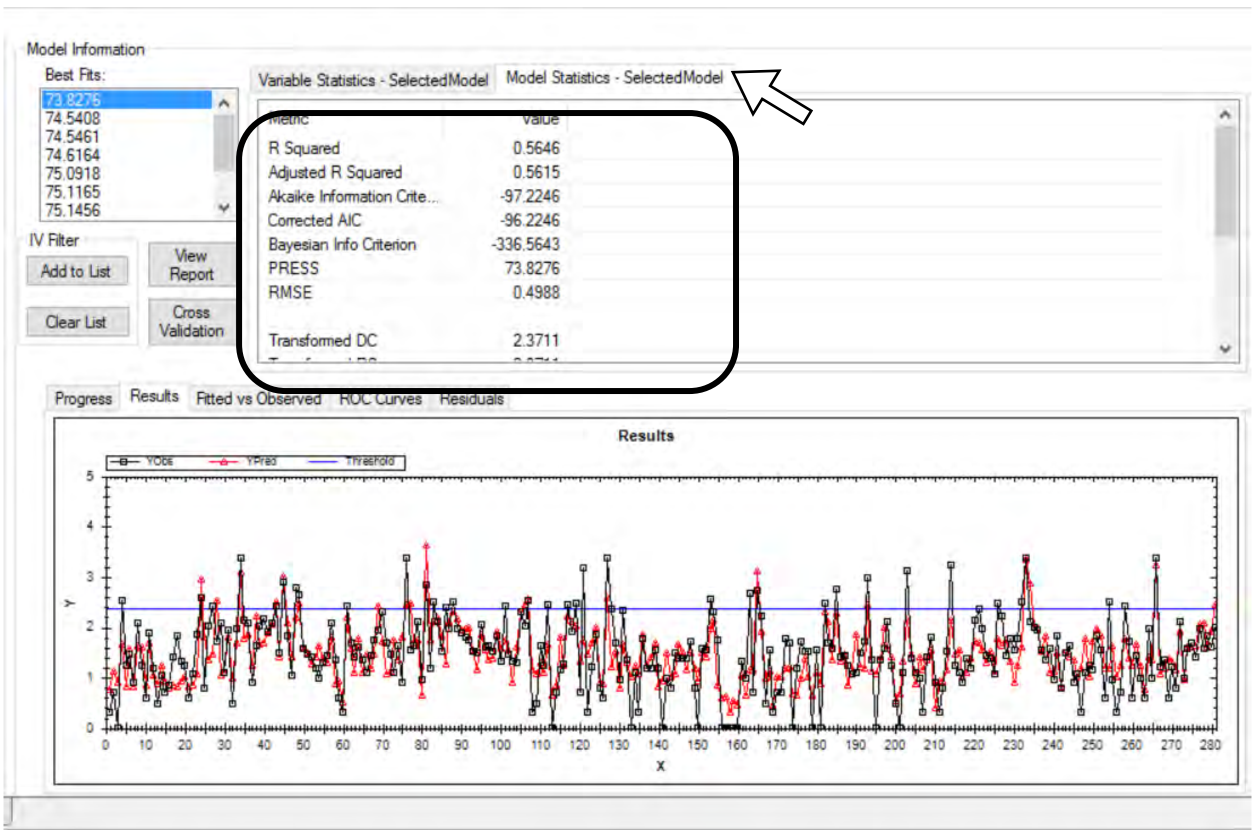
There is also a critical distinction between statistical significance and influence.

Some variables may be statistically significant, as indicated by a P-Value below 0.05, but have relatively little influence over *E. coli* (as indicated by their Standardized Coefficient). In other words, you can have a variable that varies linearly with E. coli but does not actually influence E. coli levels. **The opposite may also be true...**

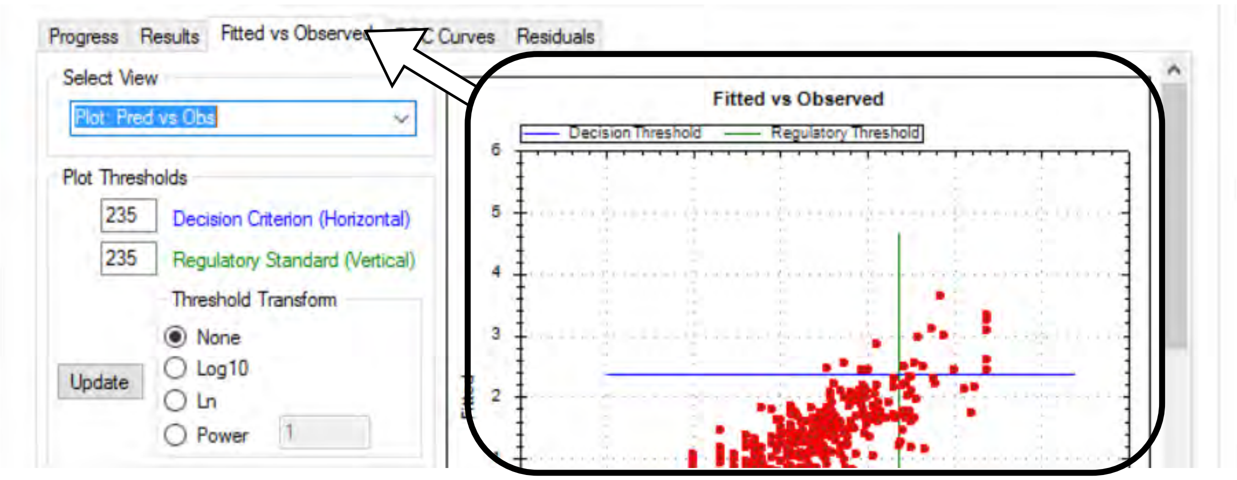
C.1. Select the first “Best Fit” model. **2.** Under “Variable Statistics”, you can adjust column widths to show the Standardized Coefficients (relative influence) and P-Values of the variables included in the model.



C.2. Click the “Model Statistics” sub-tab to view different measures of selected models “fit” (e.g., R-square, AIC, BIC) and potential predictive power (PRESS).

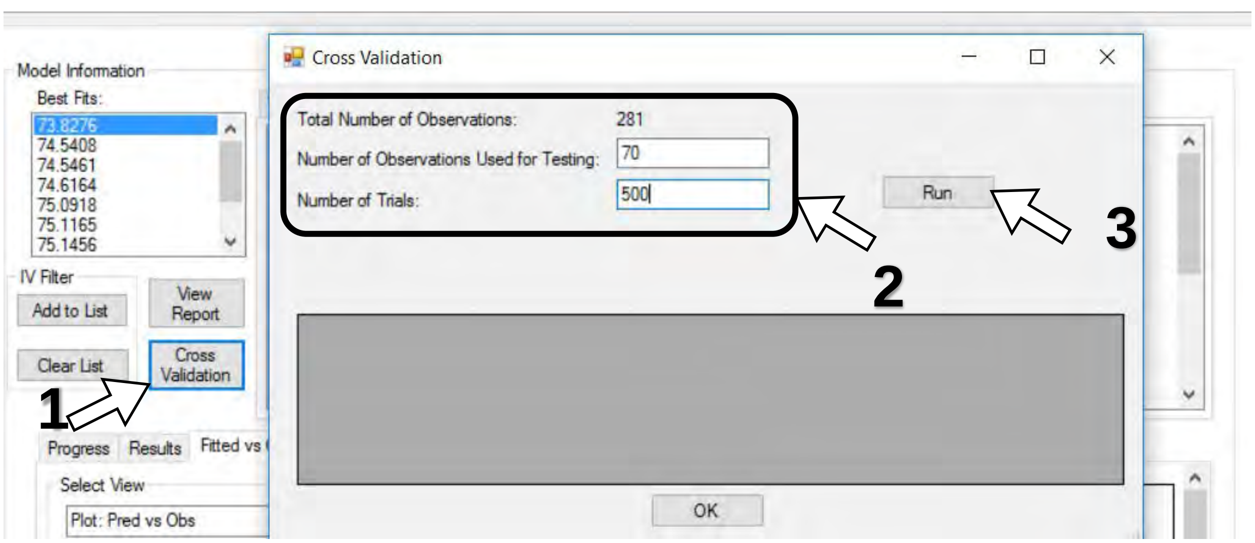


C.3. Click on “Fitted vs. Observed” to see a four-quadrant plot of **False Positives**, false exceedances of the decision threshold and **False Negatives**, false non-exceedances. Scroll down the “Best Fit” list to compare the various models on relative fit, potential predictive power, false +/-, accuracy, statistical significance, etc.

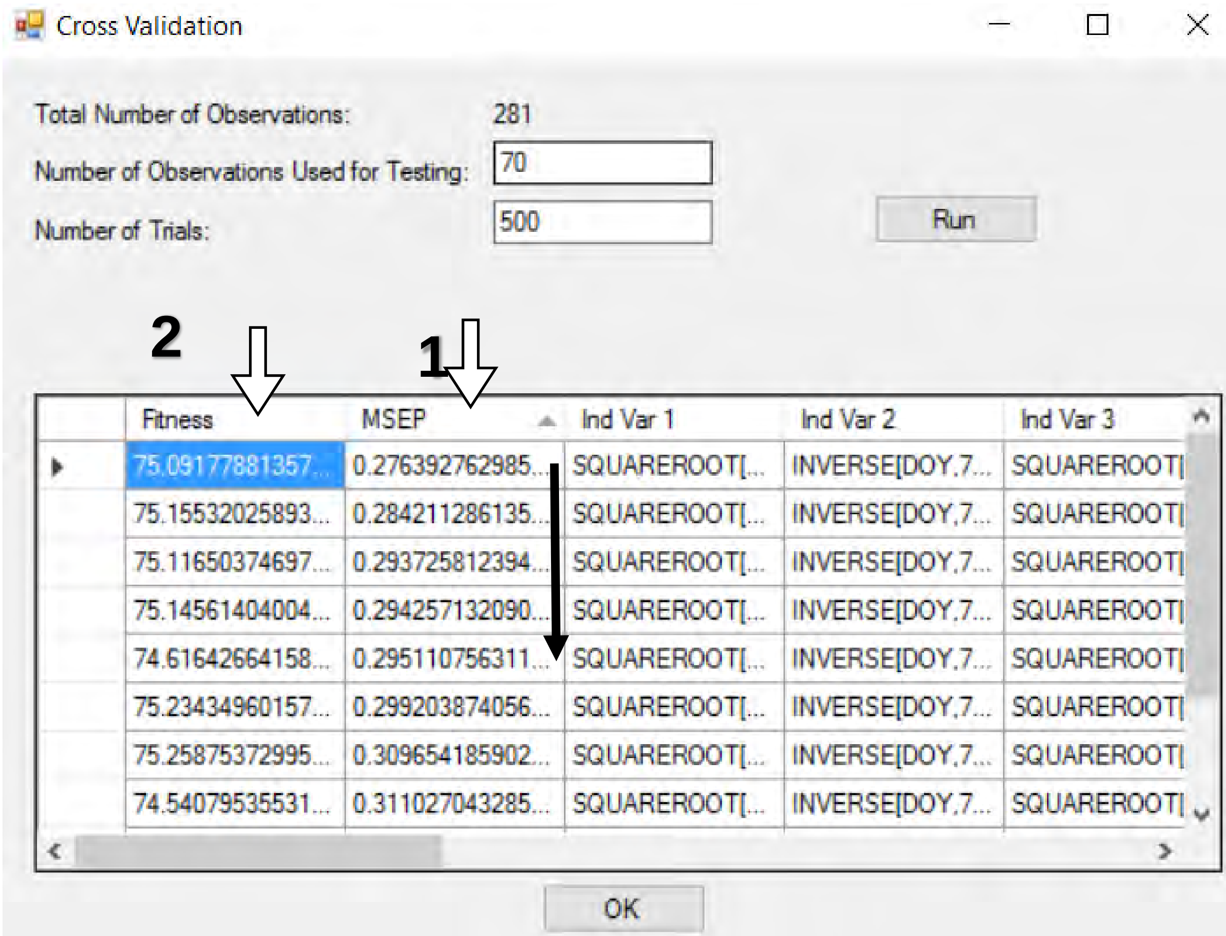


C.4. For additional evaluation, click “Cross Validation.” In Cross Validation, models are run and re-run a set number of times. In each iteration, a set number of data points (records) are removed from your data set to validate the models’ predictions. Prediction errors are averaged over the different model runs.

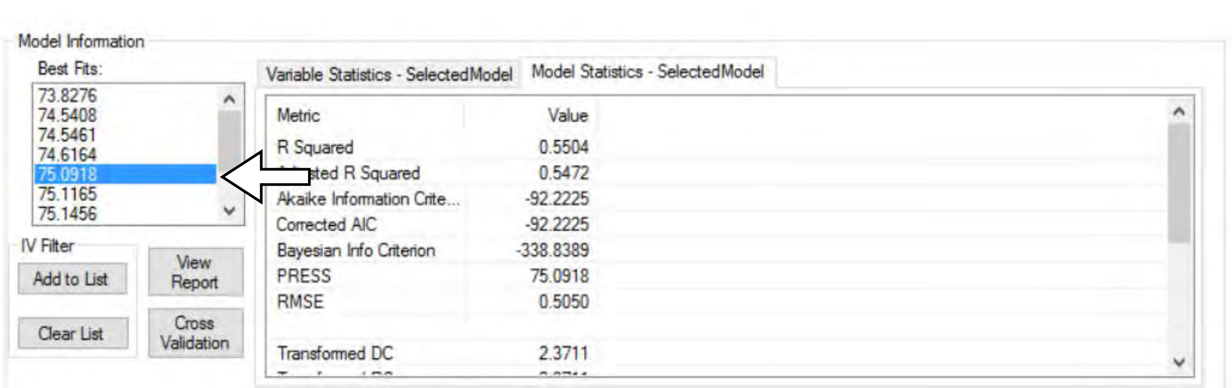
1. Click the box titled “Cross Validation”. **2.** In the pop-up window, set the “Number of Observations Used for Testing” to 70 (approximately 1/4 of the total 281 observations). Set the “Number of Trials” to 500. **3.** Click “Run.”
Now the top 10 “Best Fit” models will each be re-run 500 times. In each run, 70 of the 281 observations are randomly removed and used as the “validation dataset” to which the models’ results are compared.



C.5. **1.** Click on the MSEP column. This will sort the models by their mean square of predicted errors. **2.** Identify the model with the lowest MSEP by looking at its corresponding “Fitness” number and then click “OK.”



C.6. When you return to the “Best Fit” list, click the model corresponding to the lowest mean square of predicted errors found during cross-validation in the previous step. Sometimes the model with the lowest MSEP is the same as the original best-fit model; however, the two will not always correspond - as in this example.



D. Set MLR decision criteria

The predictive power of MLR models can be greatly improved by adjusting the **decision criterion**, the threshold value of predicted *E. coli* above which there is a better than 50% probability that an actual exceedance, over 235 CFU, will occur at the beach.

The important metrics of model performance are not the common statistical measures of ‘fit’ (like R-square), nor are they measures of ‘precision’ (like mean absolute error). Rather, they are **sensitivity** and **specificity**. These key measures, in turn, are related to the model-specific, and adjustable, **decision criteria**.

<u>KEY TERMS</u>

Decision Criteria: The prediction thresholds that determine whether an actual exceedance of a regulatory standard. In GBM, when Virtual Beach has finished developing a model, it automatically recommends a **decision criterion (DC)**.

In this example, the **decision criterion** has is set to the same as the regulatory *E. coli* standard of 235 CFU/100 mL, or 2.371 when transformed by taking the $\log_{10}$ of 235.

Particularly on those days with very high levels of *E. coli* at the beach, model-predicted concentrations will typically be lower than the actual values. In effect, most nowcast models are “muted.” That is, the predicted extremes are not as high as the actual extremes. The optimal **decision criterion** will typically be much lower than 235 CFU.

While the concept of using decision criteria that are different from 235 CFU may seem confusing at first, it is critical that you *not* simply insert 235 or some other common threshold in place of the optimal threshold as identified through the process highlighted above. Using a sub-optimal threshold for simplicity sake will result in increased decision errors; i.e., more missed or unnecessary advisories.

Sensitivity: The percentage of correctly predicted water-quality exceedances (true positives) out of all measured, or observed, exceedances. As a general rule-of-thumb, over 0.50 [50%] is considered good. In this example, 36 observations were actual exceedances.

Model example using 235 as decision criteria: 12/ (12+24) = 0.33 [33%]
Model example using 120 as decision criteria: 20/ (20+16) = 0.55 [55%]

Specificity: The percentage of correctly predicted non-exceedances out of all measured, or observed, non-exceedances. As a general rule-of-thumb, over 0.90 [90%] is considered good. In this example, 245 observations were actual non-exceedances.

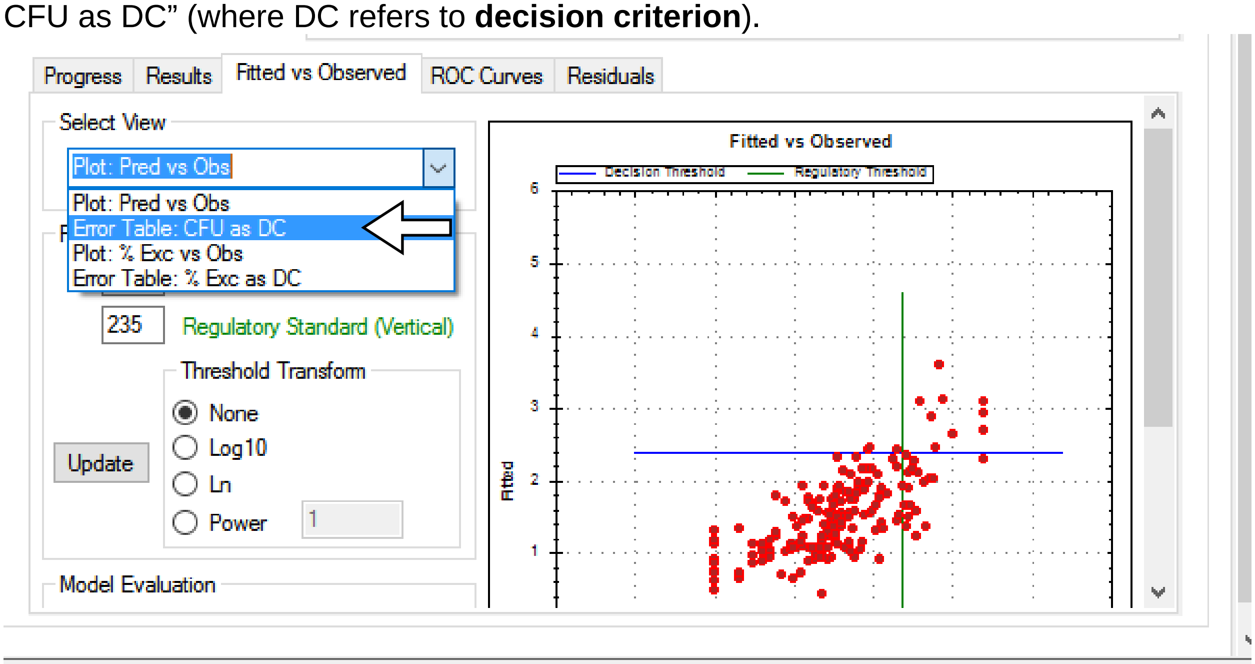
Model example using 235 as decision criteria: 240/ (240+5) = 0.98 [98%]
Model example using 120 as decision criteria: 231/ (231+14) = 0.94 [94%]

Accuracy: The percentage of correctly predicted exceedances and non-exceedances out of all results. Do *not* use accuracy as the sole basis for setting Decision Criteria. Often the Decision Criterion corresponding to highest Accuracy has an unacceptably low Sensitivity. The goal is not to maximize accuracy, but to find an optimal balance of Sensitivity and Specificity, using the 50% - 90% rule-of-thumb, or whatever balance makes the most sense from the local managers' perspective.

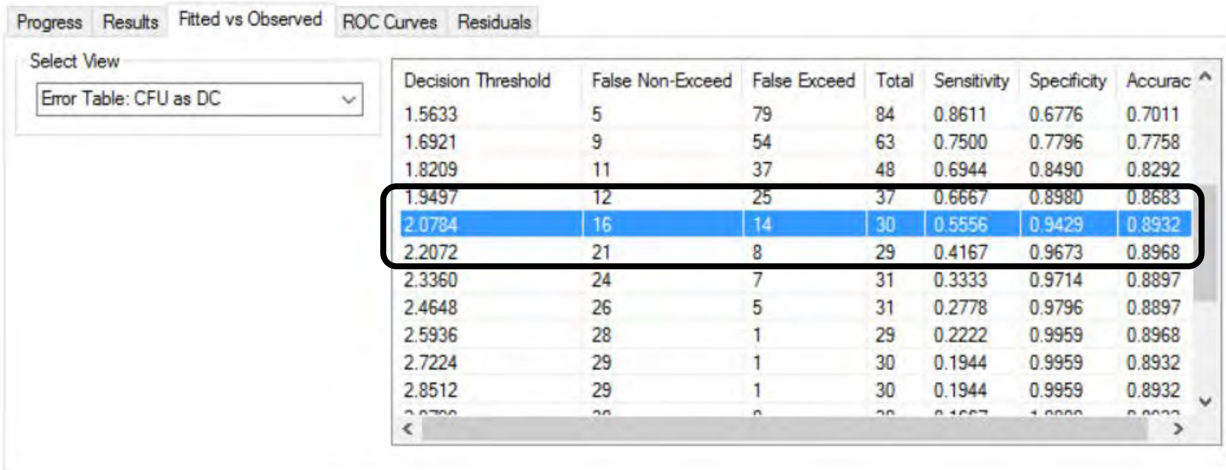
Model example using 235 as decision criteria: (12+240)/ (281) = 0.90 [90%]
Model example using 120 as decision criteria: (20+231)/ (281) = 0.89 [89%]

Comparison of using 235 or 120 as the Decision Criteria (DC) for the example model		
	TRUE RIGHT Prediction	**FALSE** WRONG Prediction
POSITIVES As predicted by model	**Points really OVER standard** DC 235: 12 DC 120: 20	**Points really UNDER standard** DC 235: 5 DC 120: 14
NEGATIVES As predicted by model	**Points really UNDER standard** DC 235: 240 DC 120: 231	**Points really OVER standard** DC 235: 24 DC 120:16

D.1. Once you have selected a preferred model (C.6), return to the "Fitted vs. Observed" plot, and under the "Select View" drop-down menu, choose "Error Table: CFU as DC" (where DC refers to **decision criterion**).

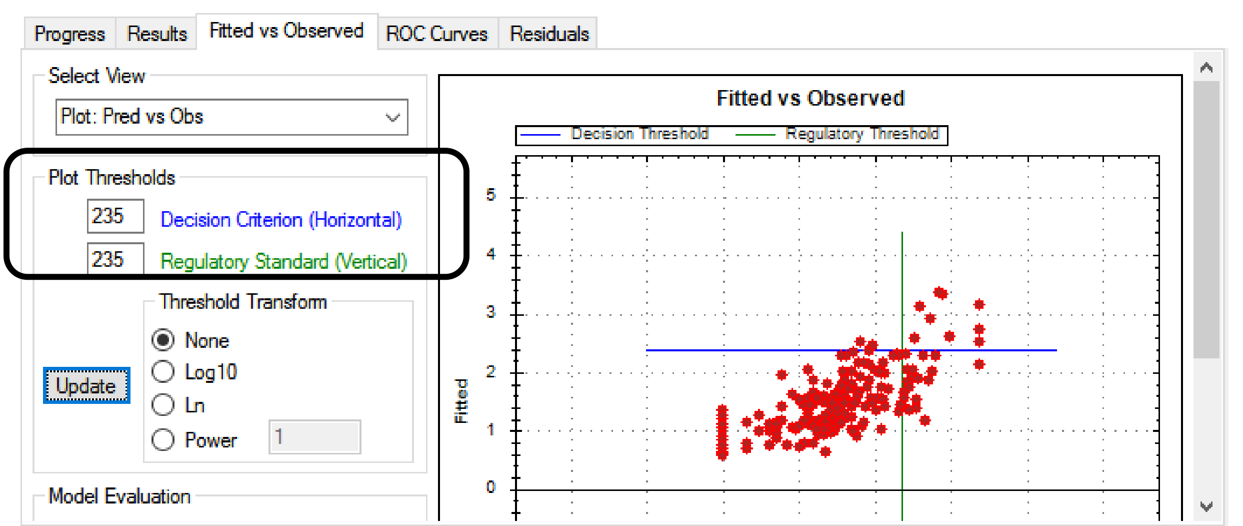


D.2. In the table that opens, re-size the columns so you can easily see the **sensitivity**, **specificity**, and **accuracy** values. Search this selected list to see whether there are any **decision thresholds** likely to produce an optimal balance of greater than 0.50 **sensitivity** and greater than 0.90 **specificity**.

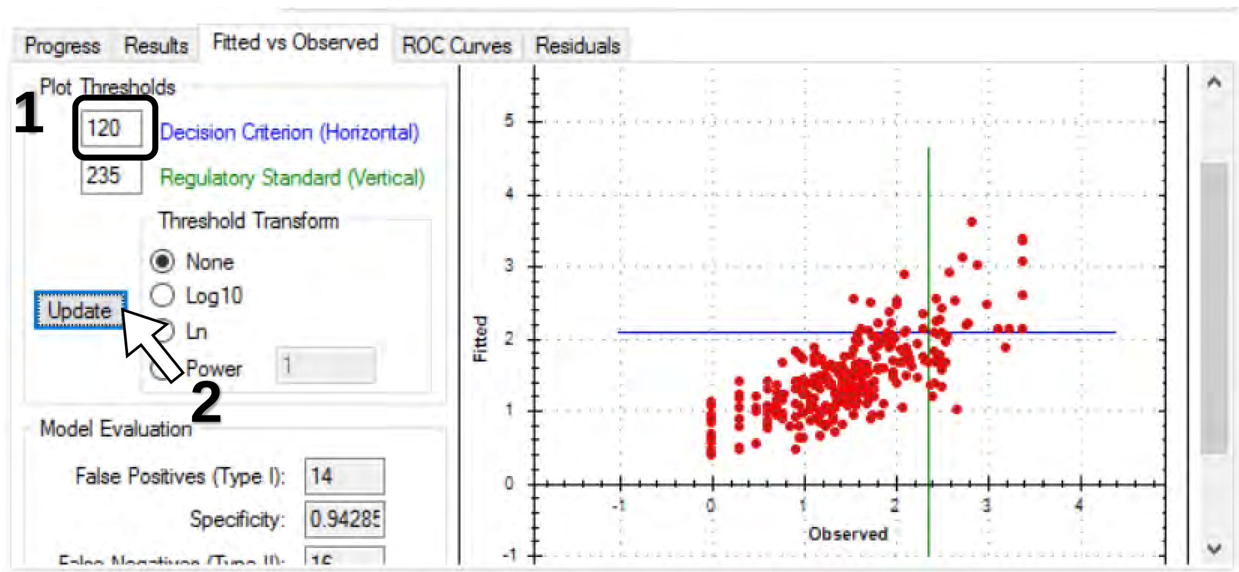


Decision Threshold	False Non-Exceed	False Exceed	Total	Sensitivity	Specificity	Accurac
1.5633	5	79	84	0.8611	0.6776	0.7011
1.6921	9	54	63	0.7500	0.7796	0.7758
1.8209	11	37	48	0.6944	0.8490	0.8292
1.9497	12	25	37	0.6667	0.8980	0.8683
2.0784	16	14	30	0.5556	0.9429	0.8932
2.2072	21	8	29	0.4167	0.9673	0.8968
2.3360	24	7	31	0.3333	0.9714	0.8897
2.4648	26	5	31	0.2778	0.9796	0.8897
2.5936	28	1	29	0.2222	0.9959	0.8968
2.7224	29	1	30	0.1944	0.9959	0.8932
2.8512	29	1	30	0.1944	0.9959	0.8932

D.3. Return to the “Plot: Pred vs Obs” graph by selecting that choice under the “Select View” pull-down menu. We will use this graph to set the **decision criterion**.
To the left of the plot, under “Plot Thresholds,” note that you can change the value for the “Decision Criterion”, the blue horizontal line, and the “Regulatory Standard”, the green vertical line. Both default to 235, which may not give the optimal **sensitivity** and **specificity**.

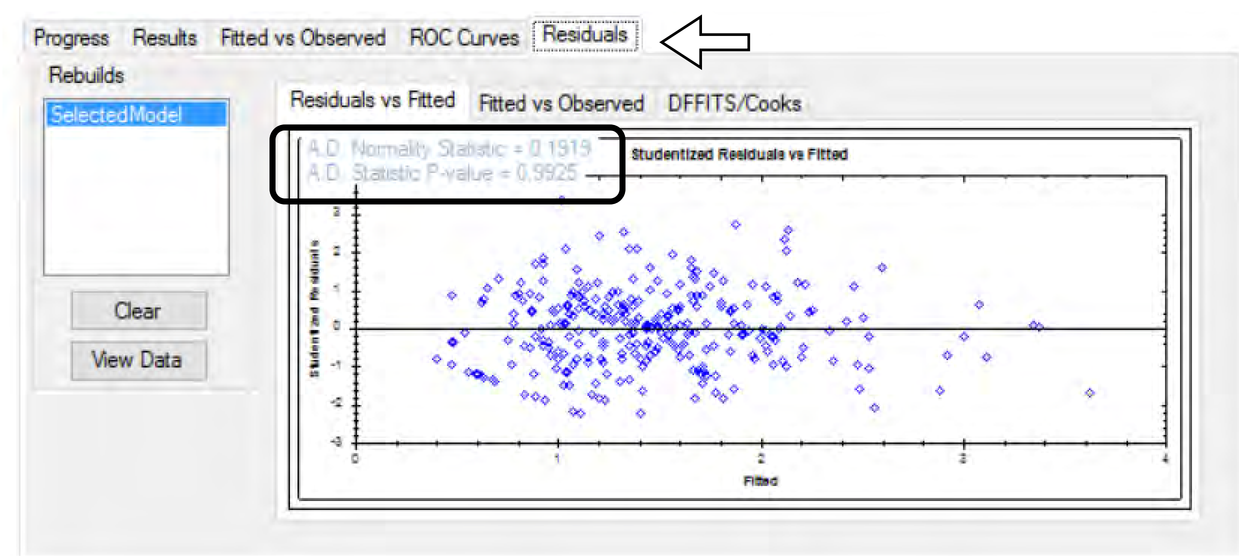


D.4. **1.** Change the value to the left of “Decision Criterion,” to 120 (That is 10 raised to 2.0784, the $\log_{10}$ value given in the “Error Table: CFU as DC). Alternatively, you could transform the regulatory standard (235) to Log10 and select the corresponding button under “Threshold Transform”. **2.** Click the “Update” button. The value of 120 CFU’s achieves a result close the optimal balance of 0.50 / 0.90. You can experiment with other decision criterion to see how **sensitivity** and **specificity** change

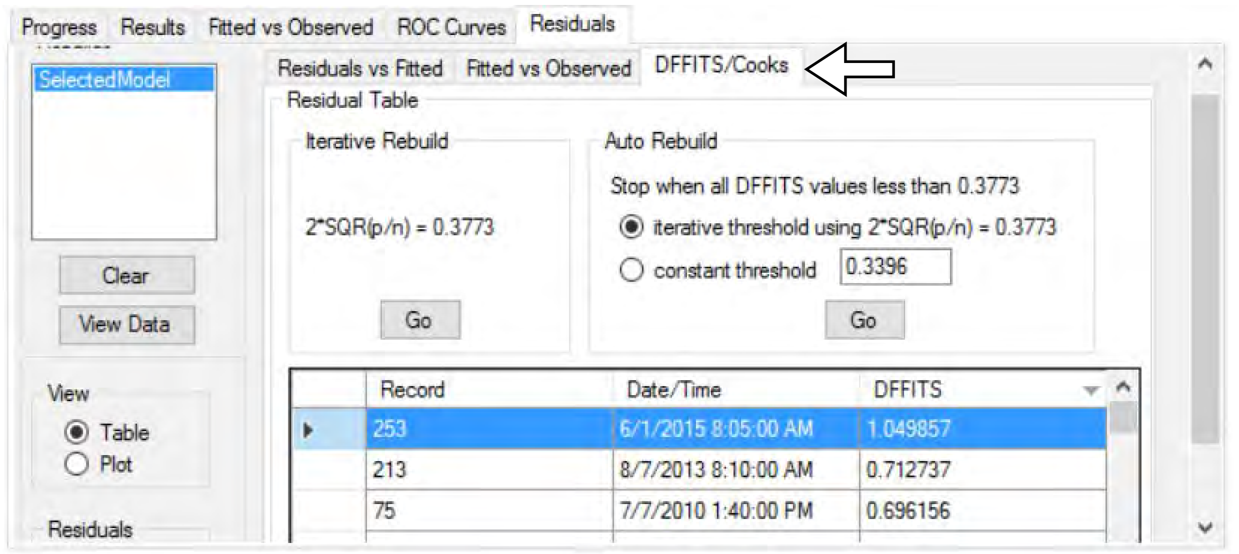


E. Evaluate MLR residuals and search for influential outliers

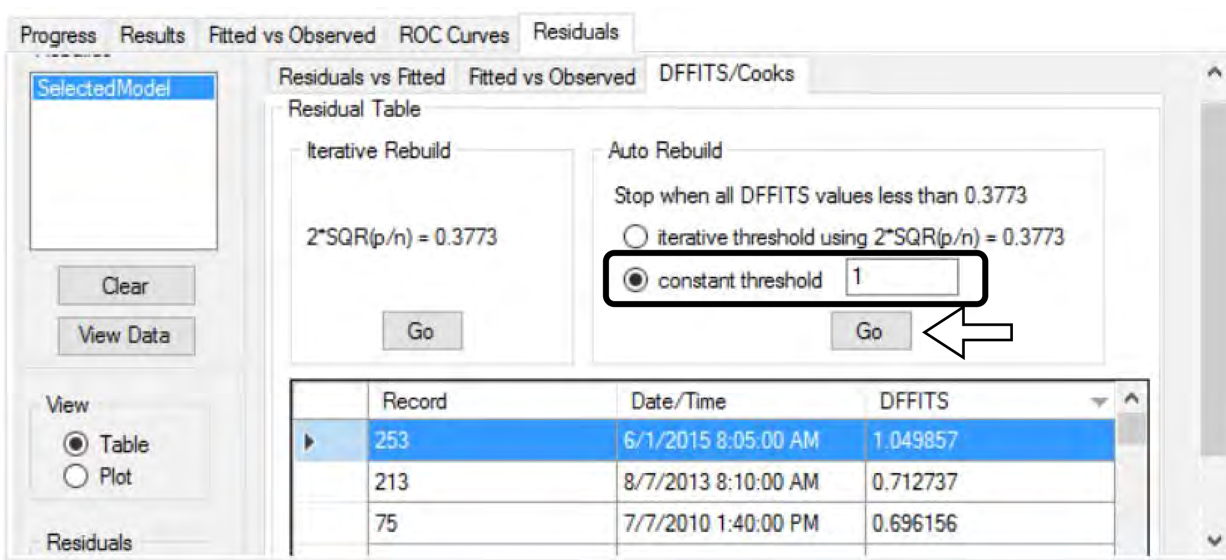
E.1. Click the “Residuals” sub-tab. The shape of the residuals vs. model predictions plot that appears, can sometimes show when the OLS assumption of normally distributed residuals has been violated. If A-D Normality Statistic has a P-value less than 0.05, this assumption has been violated.



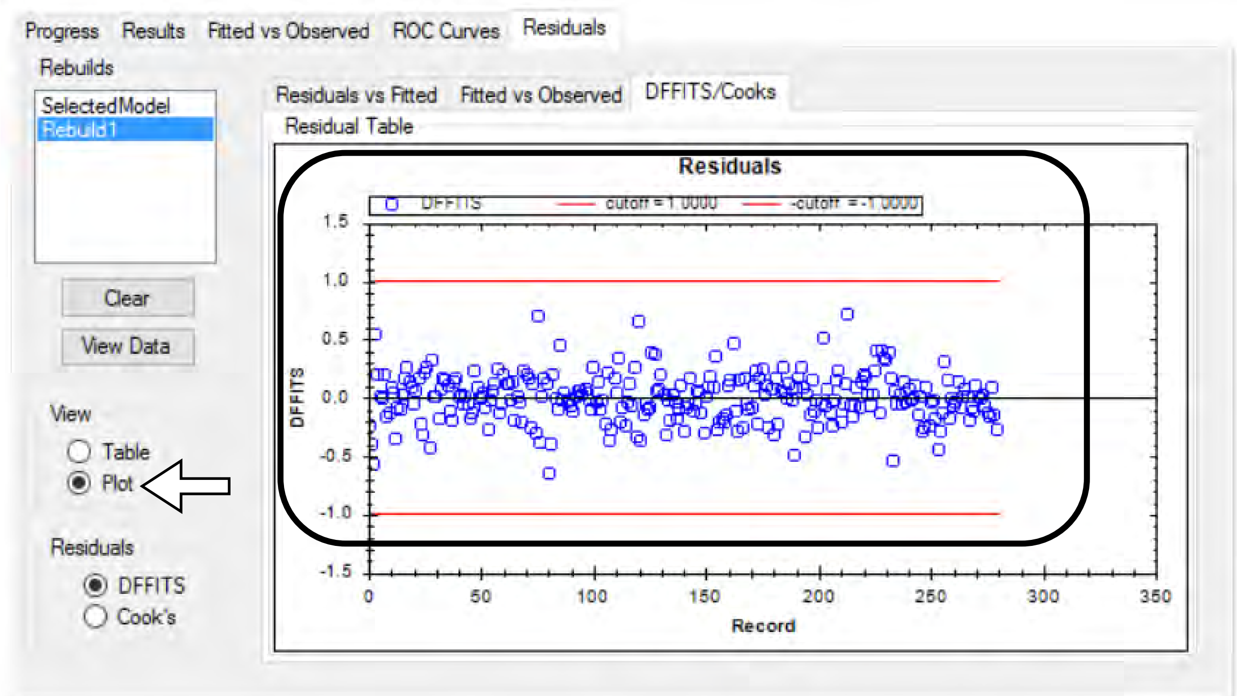
E.2. Next, click on the “DFFITS/Cooks” sub-tab.



E.3. Under “Auto Rebuild,” check the radio button next to “constant threshold” and set the value to 1. Click “Go”. **DFFITS** is a measure of how influential a given observation is on the overall model. A conservative rule of thumb is that any observation with an absolute (+ or -) DFFITS value > 1.0 is a potentially influential outlier and should be removed from the dataset. If that is the case, the model should be re-run.



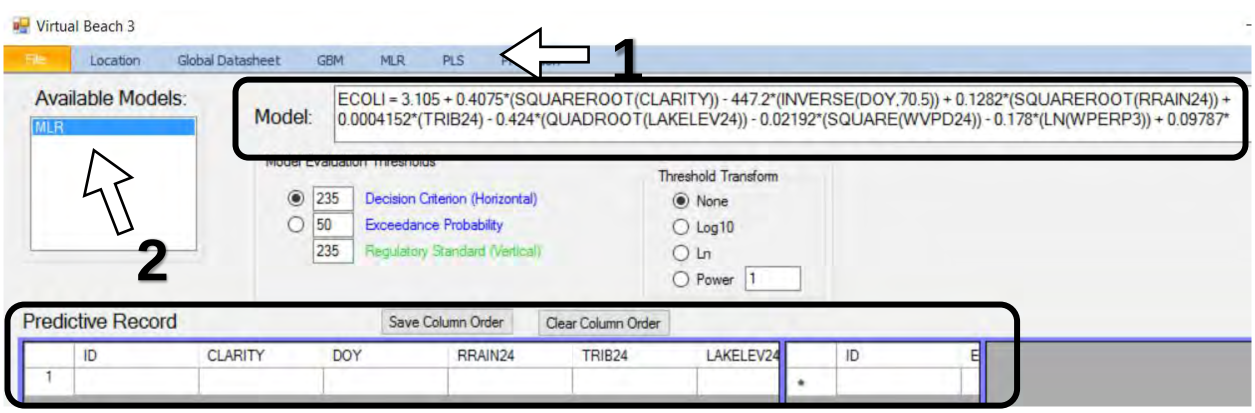
E.4. Click the radio button next to “Plot” to confirm that there are no outliers.



F. View an MLR model within the Virtual Beach Prediction tab

The Virtual Beach “Prediction” tab shows a model in the format that the eventual Nowcast operator will use to make routine water-quality predictions. It is here that the daily observations of explanatory variables like antecedent rainfall, wave height, and gull counts will be manually entered or downloaded via EnDDaT.

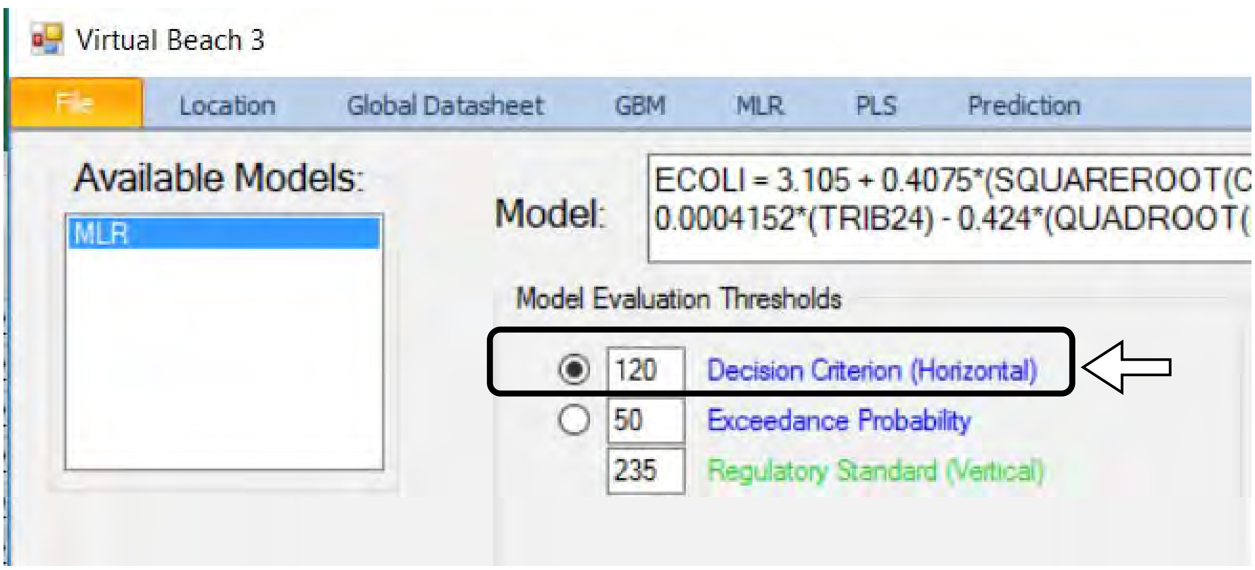
F.1. **1.** Click on the “Prediction” tab at the top of the page. **2.** Under “Available Models” click ‘MLR’. This will display a model equation, plus a row of blank cells under “Predictive Record.”



Model Equation: The text box at the top-center of the Prediction tab contains the mathematical expression of the selected model. In the case of MLR models, this equation includes numeric coefficients that define the independent relationship with each explanatory variable and the response variable; e.g., 'ECOLI'.

Predictive Record: The bottom half of the Prediction tab is the "Predictive Record." Here, each row represents a unique date/time for which field observations and/or remotely-measured data will be entered or downloaded for each of the 'native' (i.e., un-transformed) explanatory variables in the model. From these, the response variable (e.g., 'ECOLI') can be predicted, as well as the probability of exceeding the established Decision Criterion.

F.2. Change the Decision Criterion (from the default value of to 235) to the value identified in Step C.3 – in this example, 120. If you do not remember the value you identified, simply return to the 'MLR' tab to view the plot showing the threshold value.



F.3. Be sure to save your model! From the 'File' tab (pull-down menu) select "Save As." Navigate to the VB3Training directory – or any folder where you plan to keep your models – and save the project as something like "[beachname]_project_MLR". This will capture all of the work that you have completed to this point.

A note on saving VB files

Virtual Beach project (.vb3p) files allow users to save their work at any stage of the model building, evaluation, or refinement process. Project files are completely self-contained and portable. Imported data are saved within the project. Collaborators with whom you share these files will only need to have Virtual Beach and an Internet connection to use the files.