

Virtual Beach 3: User's Guide



Virtual Beach 3: User's Guide

Mike Cyterski¹, Wesley Brooks², Mike Galvin¹, Kurt Wolfe¹, Rebecca Carvin²,
Tonia Roddick², Mike Fienen², Steve Corsi²

¹National Exposure Research Laboratory
USEPA
960 College Station Road
Athens, GA 30605

²U. S. Geological Survey
Wisconsin Water Science Center
8505 Research Way
Middleton, WI 53562



Table of Contents

1. Introduction.....	6
1.1 On Predictive Modeling.....	6
1.2 Recommended User Background	7
1.3 General Overview	7
1.3 History of VB.....	8
2. Composition and installation	10
3. Operational Overview	11
4. Project Management	12
5. Location Interface	13
5.1 Finding a Beach	13
5.2 Defining the Beach Boundaries for Orientation Calculation.....	14
5.3 Finding nearby Data Sources	15
5.4 Saving Beach Information	17
6. Global Datasheet.....	18
6.1 Data Requirements and Considerations	18
6.2 Importing a Dataset.....	19
6.3 Validating the Imported Data.....	20
6.4 Working with a Dataset after Validation	24
Scatter Plot Interpretation	25
6.5 Computing Wind, Wave and Current Components	27
Notes on Component Calculations.....	28
6.6 Creation of New Independent Variables	31
6.7 Transforming the Independent Variables.....	33
Plotting Transformed IVs	35
6.8 Singular Matrices and Nominal Variables	36
6.9 Saving Processed Data.....	37
6.10 Proceeding to Modeling.....	37
7. Multiple Linear Regression Modeling	38
7.1 Selecting Variables for Model Building	38
7.2 Modeling Control Options	39
7.3 Linear Regression Modeling Methods.....	40
7.4 Using the Genetic Algorithm	42
7.5 Evaluating Model Output.....	43
7.6 Viewing X-Y Scatter plots.....	48
7.7 ROC Curves	49
7.8 Residual Analysis.....	49
Viewing the Data Table	53
7.9 Cross-Validation	55
7.10 Report Generation	55
8. Partial Least Squares.....	58
8.1 Data Manipulation	58
8.2 Selecting Variables for Model Building	59
8.3 The Regulatory Standard	60
8.4 Modeling Control Options	60
Dropping Unimportant Variables	60

Setting the Decision Threshold	61
8.5 Diagnostics.....	62
9. Gradient Boosting Machine	64
9.1 Data Manipulation	65
9.2 Selecting Variables for Model Building	65
9.3 The Regulatory Standard	66
9.4 Modeling Control Options	67
Dropping Unimportant Variables	67
Setting the Decision Threshold	68
9.5 Diagnostics.....	69
10. Prediction	71
10.1 Model Statement	71
10.2 Model Evaluation Thresholds	71
10.3 Prediction Form	72
10.4 Column Mapping of Imported Data.....	72
10.5 Viewing Plots.....	76
10.6 Prediction Form Manipulation.....	77
10.7 Importation of EnDDaT Data	77
11. User Feedback.....	79
12. References.....	80
13. Acknowledgments.....	81
Appendix.....	82
A.1 Transformations	82
A.2 Singular Matrices and Nominal Variables	84
A.3 MLR Model Evaluation Criteria	86

List of Figures

Figure 1. The major components of VB ₃	8
Figure 2. Location interface	13
Figure 3. Location controls and their function.....	14
Figure 4. Adding shoreline and water markers to define beach orientation	15
Figure 8. Importing a dataset into the Data Processing tab	20
Figure 9. Data validation required to begin data processing	21
Figure 10. Context-sensitive choices for the “Take Action Within” drop-down menu.....	22
Figure 11. Pop-up window for identifying categorical variables.	23
Figure 12. Post-validation enabling of the Global Datasheet functionality	24
Figure 13. Right-click options on columns that are not the response variable	24
Figure 14. Four different plots available for evaluation of IVs	25
Figure 15. Identifying an observation from within the XY scatter plot.....	26
Figure 16. Available choices when right-clicking the response variable	27
Figure 17. Window for computation of alongshore and offshore/onshore components	28
Figure 18. A- and O-component definitions for wind, current, and wave data	29
Figure 19. Principal beach orientations given in degrees	30
Figure 20. Window for the formulation of “Manipulates”	31
Figure 21. Creation of a new IV defined as the mean of two existent IVs	32
Figure 22. Formation of two-way cross-products of a set of four IVs	33
Figure 23. The choices for IV transformations.....	34
Figure 24. Pearson correlation coefficient scores for judging the efficacy of IV transformations	35
Figure 25. Scatter plots (Response vs. IV) for six different data transformations of a single IV	36
Figure 26. Selecting variables for MLR processing within the Modeling tab.....	38
Figure 27. Setting modeling options within the Modeling interface	39
Figure 28. Setting evaluation thresholds and threshold transformation information.....	40
Figure 29. Model building interface	41
Figure 30. Using the IV filter to select a subset of variables from the best-fit models	42
Figure 31. Genetic algorithm options within the modeling interface	43
Figure 32. Modeling results after completion of a run using the genetic algorithm.....	45
Figure 33. Modeling Interface showing variable statistics for the selected Best-Fit model.....	46
Figure 34. Modeling interface showing model evaluation metrics for the selected Best-Fit model.....	46
Figure 35. Modeling interface showing a time series plot for the selected model	47
Figure 36. A scatter plot of fitted values versus observations of the selected model.....	47
Figure 37. The ROC curves and AUC table for the Best Fit models.....	48
Figure 38. Information available on the Residuals sub-tab	50
Figure 39. A table of the DFFITS scores of the residuals	51
Figure 40. A plot of the DFFITS scores of the residuals	51

Figure 41. DFFITS/Cook's Distance controls for removing highly influential data points.....	52
Figure 42. Residuals interface showing a list of rebuilt models.....	53
Figure 43. "View Data Table" window	54
Figure 44. Fitted vs Observed plot on the Residual sub-tab.....	54
Figure 45. Cross-validation results for each of the 10 best-fit models	55
Figure 46. A text report generated on the modeling results.....	56
Figure 47. Plots of various model evaluation metrics for the 10 best-fit models	57
Figure 48. Scaled versus un-scaled views of selected model evaluation criteria	57
Figure 49. Data Manipulation: the first sub-tab on each of the method tabs.....	59
Figure 50. Selecting variables for PLS processing within the modeling module.....	59
Figure 51. Setting the Regulatory Standard and running models for PLS.....	60
Figure 52. Results after completion of a PLS model run.....	61
Figure 53. Summary of PLS model performance metrics	62
Figure 54. PLS Diagnostic plots to help evaluate model fit and influential outliers.....	63
Figure 55. Data Manipulation: the first sub-tab on each of the Method tabs	65
Figure 56. Selecting variables for GBM processing within the modeling module.....	66
Figure 57. Setting the Regulatory Standard and running models for GBM	67
Figure 58. Results after completion of a GBM model run	68
Figure 59. Summary of GBM model performance metrics	69
Figure 60. GBM diagnostic plots to help evaluate model fit and influential outliers.....	70
Figure 61. The VB ₃ Prediction interface.....	71
Figure 62. Importation of IV data using the "Column Mapper" window.....	73
Figure 63. Importation of observational data using the "Column Mapper" window.....	73
Figure 64. The scan IV window on the MLR Prediction tab.....	74
Figure 65. A prediction grid after IVs and observational data have been imported.....	75
Figure 66. Prediction interface plotting of the observations versus predictions.....	76
Figure 67. The three EnDDaT-related buttons on the prediction tab	77
Figure 68. Setting URL options for retrieval of data from EnDDaT.....	78
Figure 69. Choosing a previous day for EnDDaT data retrieval	78
Figure 70. Pop-up window indicating that data have been requested from EnDDaT	78

1. INTRODUCTION

Virtual Beach version 3 (VB₃) is a decision support tool that constructs site-specific statistical models to predict fecal indicator bacteria (FIB) concentrations at recreational beaches. VB₃ is primarily designed for beach managers responsible for making decisions regarding beach closures or the issuance of swimming advisories due to pathogen contamination. However, researchers, scientists, engineers, and students interested in studying relationships between water quality indicators and ambient environmental conditions will find VB₃ useful. VB₃ reads input data from a text file or Excel document, assists the user in preparing the data for analysis, enables automated model selection using a wide array of possible model evaluation criteria, and provides predictions using a chosen model parameterized with new data. With an integrated mapping component to determine the geographic orientation of the beach, the software can automatically decompose wind/current/wave speed and magnitude information into along-shore and onshore/offshore components for use in subsequent analyses. Data can be examined using simple scatter plots to evaluate relationships between the response and independent variables (IVs). VB₃ can produce interaction terms between the primary IVs, and it can also test an array of transformations to maximize the linearity of the relationship between the response variable and IVs. The software includes search routines for finding the "best" models from an array of possible choices. Automated censoring of statistical models with highly correlated IVs occurs during the selection process. Models can be constructed either using previously collected data or forecasted environmental information. VB₃ has residual diagnostics for regression models, including automated outlier identification and removal using DFFITs or Cook's Distances.

1.1 On Predictive Modeling

Empirical/statistical modeling outperforms persistence models (using the most recent FIB concentration as the sole predictor of the next FIB concentrations) at beaches where conditions such as weather, water characteristics, and human/animal density levels change significantly day to day (Frick et al. 2008, Brooks et al. 2013). Virtual Beach constructs models that can predict a dependent or response variable (i.e., FIB) by using variables to describe current environmental conditions that can be measured or estimated in a timely manner. These are referred to as independent variables (IVs) and often include beach water parameters such as turbidity, water temperature, specific conductance, or wave height; parameters monitored and made available via the web such as rainfall, stream flow, and stream water quality; and parameters estimated by environmental models such as water currents, wave height and direction, and radar rainfall.

In any predictive modeling endeavor, variability and uncertainty associated with model output arise for a variety of reasons that are impossible to eradicate completely. VB₃ attempts to examine this variability and uncertainty in a transparent manner using a probability of exceedance for any regulatory standard the user wishes to investigate. Even so, there is no guarantee that every model prediction will be correct, and a situation may arise in which the model predicts acceptable water quality for public recreation that could be erroneous. Decisions to allow or disallow swimming at beaches must be made,

however, and in the best case scenarios, regression models developed with VB₃ will outperform traditional persistence models based on just the previous day's FIB concentrations.

1.2 Recommended User Background

For those using VB₃, some experience with spreadsheet data manipulation programs like Microsoft Excel is recommended, but not necessary. A familiarity with multiple linear regression analysis is also helpful, but again not mandatory. Without this background, VB₃ will take longer to master, but it should not prohibit users from producing and using models.

1.3 General Overview

VB₃ has four major components:

- Beach location map interface where users can locate sites, define the orientation of the beach, and examine nearby potential data sources.
- Interface that facilitates initial import and manipulation of data.
- Multiple “method” tabs where the statistical modeling is done. Each tab has some features identical to those seen in other method tabs and some that are unique. For example, the multiple linear regression (MLR) tab allows examination of regression residuals, elimination of highly influential data records, and viewing of receiver operating characteristic (ROC) curves.
- Prediction interface allowing entry of new data and subsequent estimation of pathogen indicator concentrations with a selected model from any of the statistical methods.

Each component is accessible from the application's main window via tabs at the top and bottom of the main screen (Figure 1). The Location and Global Datasheet tabs are always visible, while the statistical method tabs only become visible once data pre-processing has been completed (i.e., clicking the “Go to Model” button on the Global Datasheet ribbon). The Prediction tab appears when model-building on any method tab is complete and a model is selected.

Lastly, we note that statistical models are only as effective as the data used to develop them. No statistician, however skilled, can turn a dataset of low-quality independent variables (IVs) into a useful predictive device.

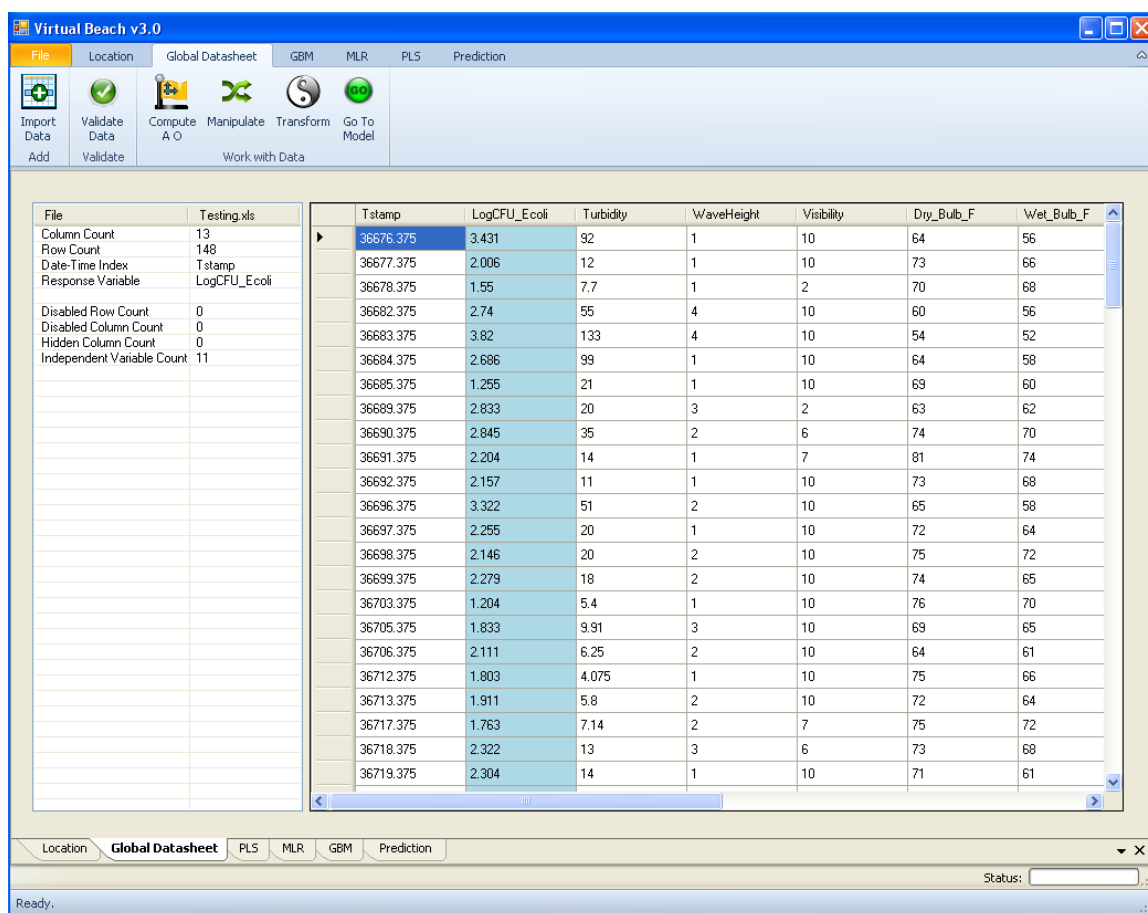


Figure 1. The major components of VB₃: “Location,” “Global Datasheet,” three “Method” tabs (GBM, MLR, and PLS), and the “Prediction” interface. The Global Datasheet is currently active.

1.3 History of VB

VB₃ is a direct descendant of Virtual Beach version 2, whose most recent release is VB_{2.4}. The original Virtual Beach Model Builder application (Virtual Beach version 1) was developed by Walter Frick and Zhongfu Ge at the USEPA in Athens, Ga (Frick et al. 2008). VB₁ can be characterized as a linear regression model-building tool that supports primarily manual analyses of datasets via visual inspection of data plots and manipulation of variables (e.g., transformations, creating interaction terms), followed by an iterative process of testing, comparing and evaluating models. The fitness of developed models is computed and tracked, allowing comparison and eventual selection of a “best” model for the dataset under consideration. This model then produces estimates of pathogen indicator concentrations using current or forecasted environmental data from the site.

VB₂ (Cyterski et al. 2012) enhanced the functionality of its predecessor by performing similar functions (visual inspection of univariate data plots, manual transformations of individual variables, MLR model building, prediction, etc.), but also automated and extended functionality in several ways:

- The Map component provided information on the location and availability of nearby data sources through the map interface. These sources include the USGS National Water Information System (NWIS) and the National Climatic Data Center (NCDC)

which provide recently collected and/or forecasted data to generate predictions by a chosen model.

- The Map component provided a convenient method for defining beach orientation by overlaying the beach on current shoreline layers (satellite images, Google Maps, MS Virtual Earth, etc). Given the orientation, VB₂ could calculate wind, wave, or current components (the A-component is parallel to shore and the O-component is perpendicular to shore) which can be important predictor variables.
- Although manual processing and analysis of imported data (visual inspection of univariate data plots and the transformations/interactions of variables) was retained, the data-processing component of VB₂ automated generation of all possible second-order interaction terms among a set of IVs, formed more complex functions of multiple columns, and automated testing of a suite of variable transformations that improved model linearity. This functionality increased the number of models to evaluate during later selection routines and removed the burden of manual assessment that users of VB₁ encountered.
- Within the linear regression analysis component, multi-collinearity among predictor variables was handled automatically. Any model containing an IV with a high degree of correlation with others (as measured by a large Variance Inflation Factor [VIF]) was removed from consideration during model selection.
- During MLR model selection, models were ranked by a user-selected evaluation criterion: R^2 , Adjusted R^2 , Akaike Information Criterion (AIC), Corrected AIC, Predicted Error Sum of Squares (PRESS), Bayesian Information Criterion (BIC), Accuracy, Sensitivity, Specificity, or the model's Root Mean Square Error (RMSE). See Section A.3 for definitions of these criteria. Regardless of which criterion is chosen, the software records the ten best models in terms of it. In comparison, VB₁ had a single criterion choice, Mallows's Cp.
- As the number of IVs in a dataset increases, possible MLR models increase exponentially (considering transforms/interactions), resulting in trillions of possible models from a modest number (12-13) of IVs. VB₂ implemented a genetic algorithm (GA) that efficiently searched for the best possible MLR model. Alternatively, VB₂ users could perform exhaustive calculations in which all possible combinations of IVs were tested if the number of possible models was reasonably small (< 500,000). Both the GA and exhaustive approaches greatly expanded the model-building capabilities of VB₂, compared to VB₁.
- Users no longer had to enter data values in transformed, interacted, or component-decomposed form to make a prediction with the selected MLR model. On the VB₂ MLR Prediction tab, a user-selected model is coded into an input grid with data entry columns matching main effects of the model. Any mathematical manipulation of these IVs is then performed automatically prior to making predictions.

VB₃ primarily builds on VB₂ by adding additional statistical methods that give users more flexibility in modeling their datasets. In addition to MLR, users can now use Partial Least Squares (PLS) regression and a Gradient Boosting Machine (GBM) to fit their data and make predictions. The redesigned software architecture (using DotSpatial libraries) easily accommodates future expansions of the suite of modeling tools. Possible future additions could be Binary Logistic Regression, Least-Absolute Shrinkage (LASSO) and Neural Networks. The Prediction tab of VB₃ also has a button to allow direct interaction with the USGS's data acquisition system, EnDDaT (<http://cida.usgs.gov/enddat/>), for automated dataset construction and ease of FIB prediction from web-accessible data.

2. COMPOSITION AND INSTALLATION

VB₃ was developed with MS Visual Studio and written in C#, and uses multiple public domain system components:

- FLEE equation parser (<http://flee.codeplex.com/>)
- Accord.Net math libraries (<http://accord-framework.net/>)
- R statistical libraries (<http://cran.r-project.org/web/packages/>)
- DotSpatial mapping libraries (<http://dotspatial.codeplex.com/>)
- Weifen Luo Docking UI (<http://sourceforge.net/projects/dockpanelsuite/>)
- ZedGraph (<http://sourceforge.net/projects/zedgraph/>)
- GMap.Net (<http://greatmaps.codeplex.com/>)

No license or software purchase is required to install and run VB₃, but an internet connection is needed to display Geographical Information System (GIS) information. Users must have Windows XP or 7 with DotNet Framework 4.0 to assure proper installation and operation. Other versions of Windows (e.g., Vista) have caused various errors to occur, thus are not recommended for use with VB₃. Certain VB₃ data manipulation and model-building operations are computationally intensive, so faster CPUs are better, but laptop or desktop systems with at least 2 GB RAM will be adequate. Disk space requirements are about 140 MB for VB₃ and 170 MB for the DotNet Framework 4. The VB₃ application installer will attempt to download and install the DotNet Framework 4.0 if it is not already installed on the target system; this also requires a network connection. If necessary, a user can obtain the DotNet Framework 4 installer at no cost at:

<http://www.microsoft.com/download/en/details.aspx?id=17851>

The EPA's Center for Exposure Assessment Modeling (CEAM) web site distributes VB at:

<http://www2.epa.gov/exposure-assessment-models/virtual-beach-vb>

Obtain and run the VB₃ application installer and follow the on-screen instructions. After installation, a shortcut will appear on the desktop.

3. OPERATIONAL OVERVIEW

To make VB₃ straightforward to operate, it has four functions, each with its own interface:

Location – an optional mapping/GIS screen for calculating a beach orientation used for later computation of orthogonal (alongshore and offshore/onshore) wind, current, and/or wave components for the beach under consideration. Such components can be powerful predictors of pathogen indicator concentrations at the beach, so defining the beach orientation is recommended if the dataset under consideration contains wind, wave or current data. This tab is also useful for locating nearby NWIS/NCDC data sources for a specific location, although VB₃ does not directly acquire data from these sites.

Global Datasheet – a way to support data manipulation on an imported dataset. In addition to wind/current/wave component generation, users can generate new independent variables that represent the products, means, sums, differences, minimums, and maximums of other IVs, as well as investigate data transformations for the IVs.

Methods – there are three Method tabs – Multiple Linear Regression (MLR), Partial Least Squares regression (PLS), and a Gradient Boosting Machine (GBM). Each has its own unique interface, but shares common elements. One common element is a “variable selection” tab where the user chooses from a list of eligible IVs for consideration in model-building and model-generation. Another common element is a “Data Manipulation” tab which is initially populated with data from the Global Datasheet. After initialization, however, the user can then modify “local” data for the chosen statistical technique.

Prediction -- this tab is comprised of three spreadsheets/grids where users can enter or import the IVs needed for the chosen model (left grid), enter or import the values of the response/dependent variable that will be compared to model predictions (middle grid), and examine model predictions and exceedance probabilities (right grid). Time series and scatter plots of the measured dependent variable values versus predictions help users gauge model effectiveness.

The following list attempts to provide an overall context for how a general, basic modeling session using VB₃ would be conducted (optional actions in green, required actions in red):

- Open the Software
- Location Tab is Visible
 - Use the GIS map to find beach of interest
 - Delineate beach shoreline
 - VB₃ calculates the beach orientation angle
- Click on the Global Datasheet Tab
 - Import data from a file
 - Validate the imported data
 - Compute wind/wave/current A/O components
 - Create new IVs
 - Investigate transformations to the IVs
 - Click the “Go To Model” button
- Click the MLR, PLS, or GBM Tabs
 - Set the method-specific modeling options
 - Run the model
 - Look at fitted results and choose a model to use
 - PLS/GBM – only a single model produced*
 - MLR – returns the “best” ten models; user must choose one*
- Take the Chosen Model to the Prediction Tab
 - Import data file, or manually enter new data
 - Make predictions using new data and the chosen model
 - Evaluate predictive model performance

4. PROJECT MANAGEMENT

The user will often perform a number of pre-processing steps on an imported dataset to prepare it for analysis, and then develop models from the resulting data. To avoid repeating all of this work, a file can be saved (termed a “project” file) and re-opened via the File → Save and File → Open menu selection. Project files have a “.vb3p” extension. Opening a saved project file will load the saved data into the Global Datasheet and re-populate the methods tabs with the local data, as well as any modeling results generated prior to the save. The beach orientation defined by the user on the Location tab is also saved inside a project file. We suggest giving Project files a descriptive name of the beach/site being modeled for later easy identification.

In addition to project files, “model” files can be saved by using “Save As (prediction only)” under the “File” menu at the top of the VB₃ interface. These files have a “.vb3m” file extension. A model file contains information on the IVs, model parameters, and other metadata for the currently selected models on each method tab. When users open a saved model file within VB₃, they are taken directly to the Prediction

tab (the only accessible tab) where they can use the model to generate predictions. Model files allow the user to construct models and choose a “best” one for a site, save a model file, and deliver this file to a beach manager. With this approach, a manager will not need VB₃ for full-scale model development, but only to input new data, generate predictions, and make decisions about issuing swimming advisories.

If the user clicks the red “X” in the upper-right corner of the main VB₃ window (Figure 1), a prompt will ask if they wish to save their project before closing.

5. LOCATION INTERFACE

On VB₃ application startup, the “Location” tab is shown first (Figure 2). Because use of this tab is optional, users can go directly to the “Global Datasheet” interface by clicking that tab at the top or bottom of the screen.

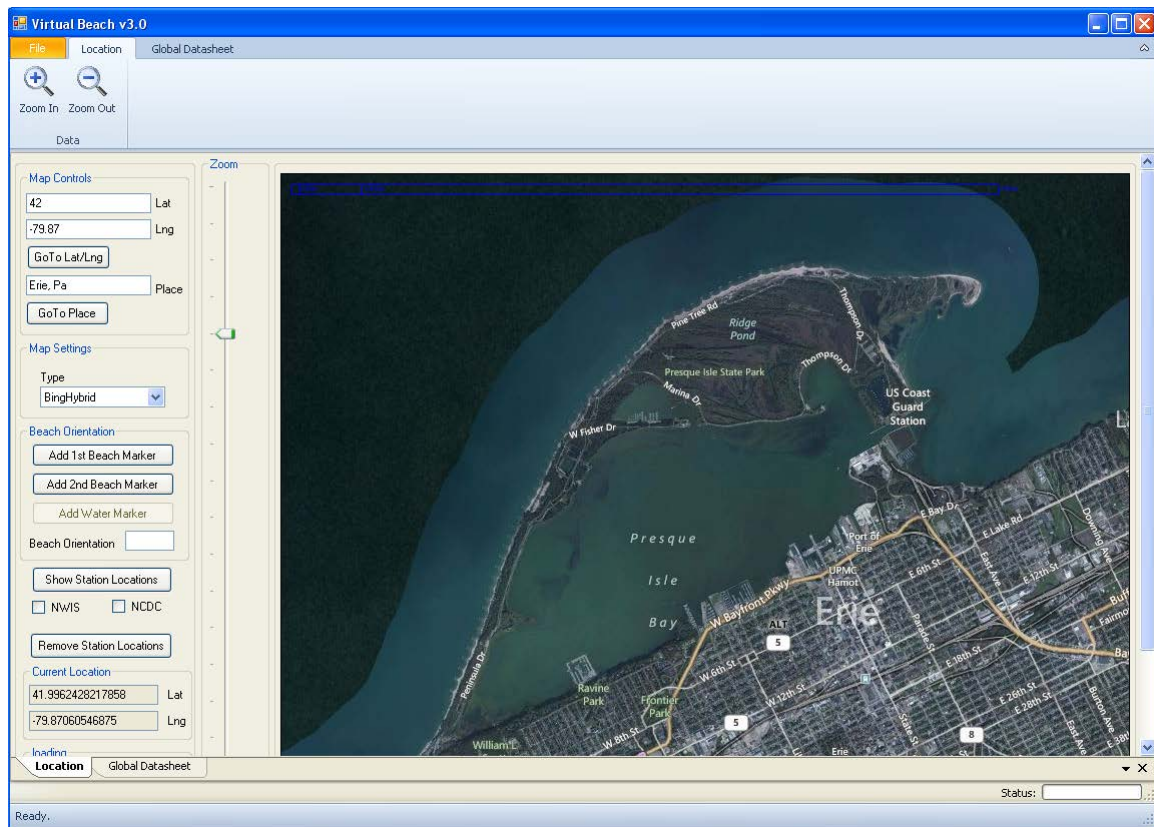


Figure 2. Location interface - the default map type is Yahoo Map, but users have many mapping options.

5.1 Finding a Beach

The location interface provides map controls (Figure 3) that let users navigate to a beach site by panning and zooming (right-click and drag mouse to pan; use mouse wheel, slider at the left of the map, or the two buttons in the top ribbon for zoom). Alternately, a

latitude/longitude or “city, state” name can be entered at the top left, followed by a click on “GoToLat/Lng” or “GoToPlace” buttons. The control uses Google Maps’ reverse geo-coding network service to find locations.

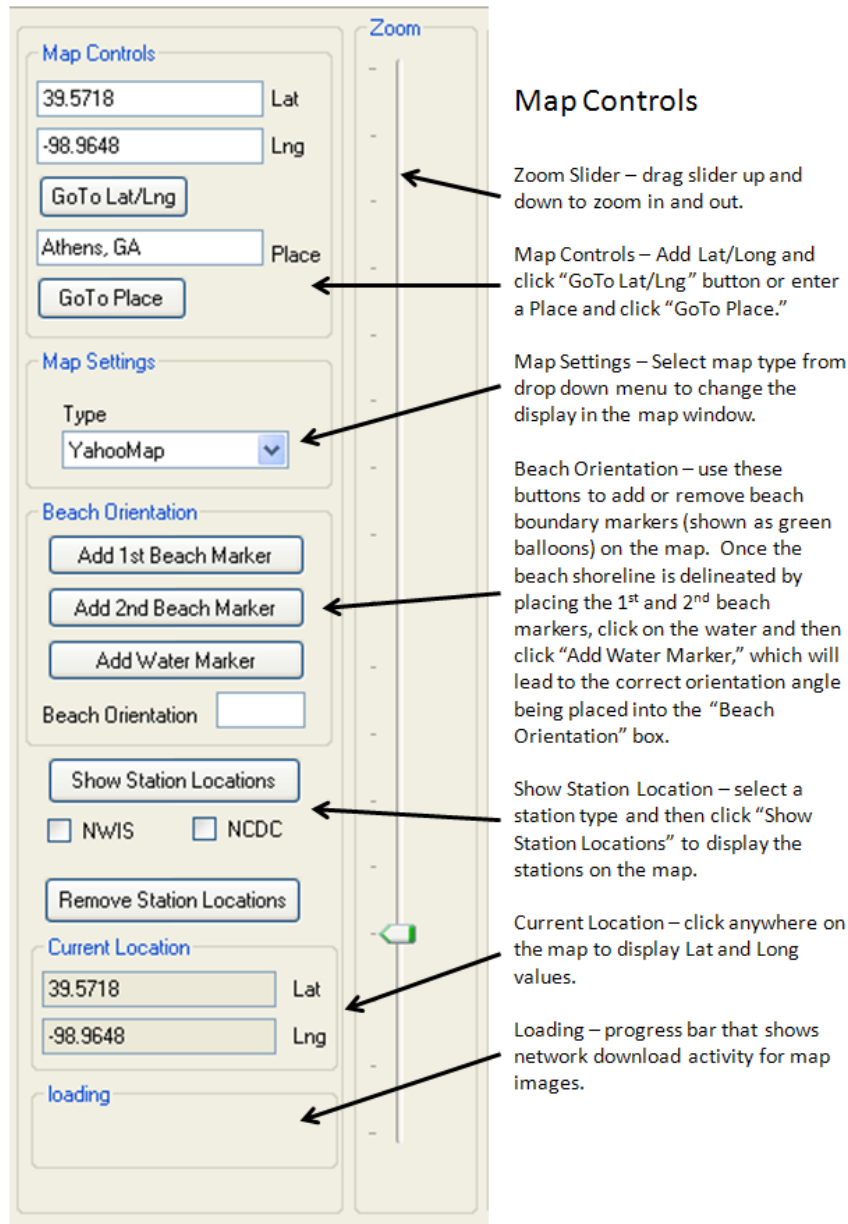


Figure 3. Location controls and their function

5.2 Defining the Beach Boundaries for Orientation Calculation

The map control allows delineation of a beach’s boundaries so that VB₃ can calculate its orientation (Figure 4), which is useful if wind, wave, and/or current flow components are used in model-building. Maps provide less shoreline detail, so it is recommended that a hybrid or satellite image be selected prior to adding point locations that define beach boundaries. Once the beach of interest is found and the swimming area

is located, left-click on the map (a red marker will appear) and click the “Add 1st Beach Marker” button; this represents one endpoint of the beach shoreline/swimming area. Now left-click the other end of the beach on the map and click the “Add 2nd Beach Marker” button. Finally, left-click on the map to indicate where the water is, relative to the shoreline, and click the “Add Water Marker” button. Marker points will turn from red to green as they are identified. Once the water marker is added, a shaded box appears and the beach orientation angle is displayed to the left of the map at the bottom of the “Beach Orientation” box (Figure 4).

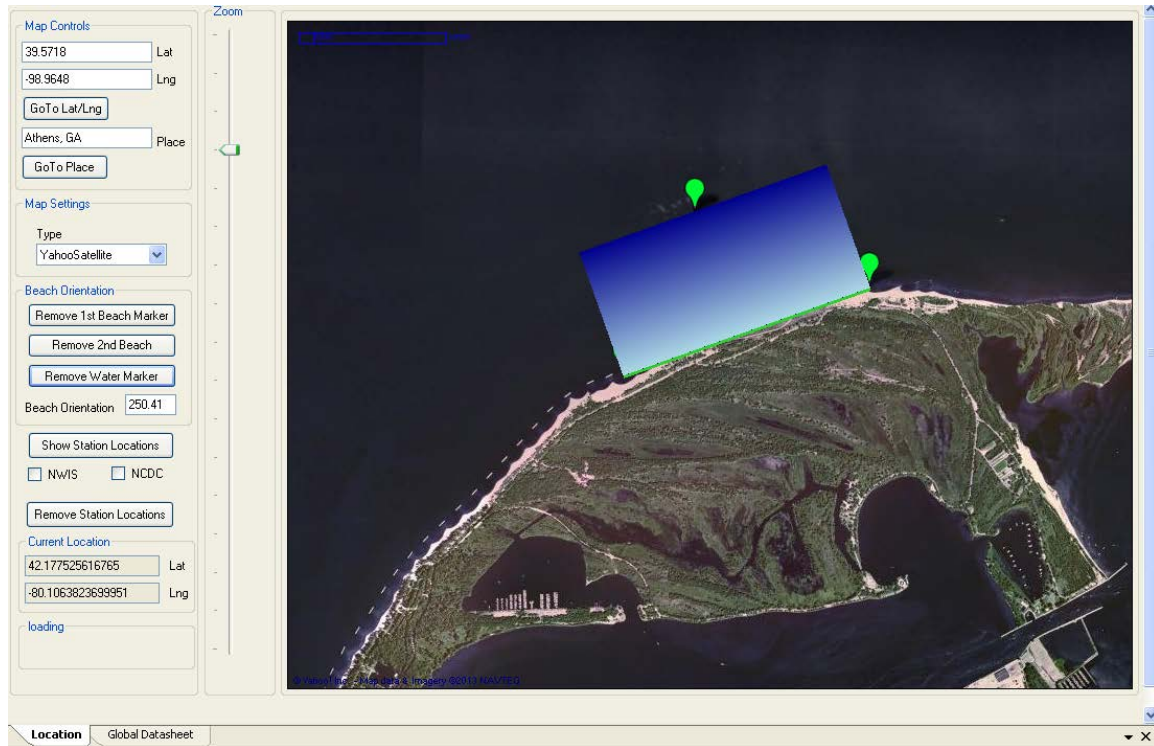


Figure 4. Adding shoreline and water markers to define beach orientation

These boundary points can be added or removed until the user is satisfied with the beach representation. VB₃ will pass the calculated beach orientation angle to the global datasheet for wind/current/wave component calculations.

5.3 Finding nearby Data Sources

Possible nearby data sources for beach of interest may be located and displayed on the map (Figure 5). USGS NWIS and NOAA NCDC station markers in the mapped area can be displayed by checking appropriate items in the map window and clicking the “Show Station Locations” button in the left panel or the “Show Stations” button on the top ribbon of the tab. Clicking these buttons will issue an error message if the zoom level (the vertically-oriented slider located just to the left of the map as shown in Figure 4) is not at least at level 13 (5th hash mark from the top of the zoom slider). If any of the selected station categories are present within the map display area, they will appear. Once station location markers are displayed on the map, hovering over the top-left corner

of any station marker will display station ID information (Figures 6 and 7). With that information, users can visit the appropriate web address to gather water/weather data for the area of interest:

USGS NWIS web site: <http://waterdata.usgs.gov/nwis/inventory>

NOAA NCDC web site: <http://www.ncdc.noaa.gov/oa/climate/stationlocator.html>

The station icons can be cleared by hitting either the “Remove Station Locations” button in the left panel or the “Clear Stations” button on the top ribbon.

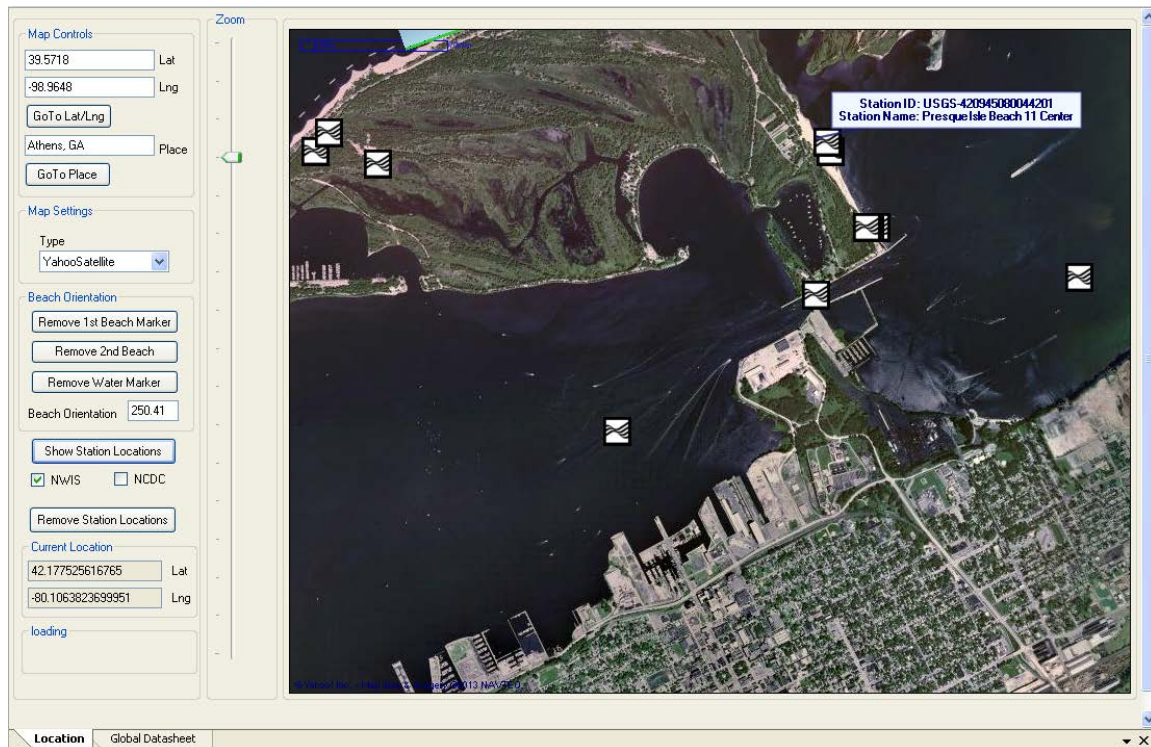


Figure 5. Location interface showing NWIS station markers near Erie, PA.

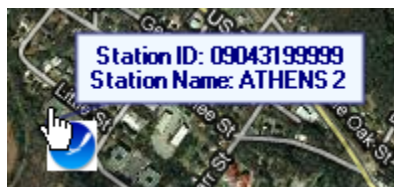


Figure 6. NOAA/NCDC station marker showing station ID information

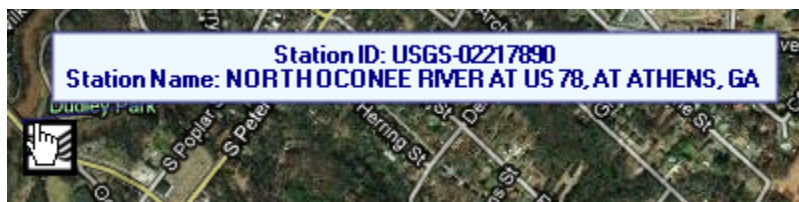


Figure 7. USGS/NWIS station marker showing station ID information

5.4 Saving Beach Information

As covered in Section 4, the File→Save menu selection will open a window that allows the user to save the project information (such as placement of the beach/water boundary markers and the calculated beach orientation) inside a VB₃ project file.

6. GLOBAL DATASHEET

6.1 Data Requirements and Considerations

VB₃ can import .xls, .xlsx, and .csv files, but input data must conform to certain standards:

- The first row of any column must be a header specifying the column's name.
- For error-free operation of the software, column names should be composed only of letters, numbers, and/or underscores (“_”).
- Do not begin a column name with a number.
- VB₃ will issue an error statement if a dataset with spaces in a column name is imported.
- The left (first) column of the dataset must be an identifier for the observations -- typically a date, time, or serial number that indicates when or where that row of data was collected.
- Each row MUST have a unique ID value (left-most column). If VB₃ finds duplicate IDs, it will issue an error statement.
- If the ID column specifies a collection date or time, time series plots in VB₃ will be most interpretable if the rows are in chronological order, from the earliest to the most recent data. VB₃ will not re-arrange the data in chronological order on its own.
- The second column of the dataset will initially be set as the response variable; however, this can be changed after data are imported. Other columns will be considered as IVs (besides the first ID column).
- Variable measurement units are not considered by VB₃, but certainly affect predictions. Ensure that any data used for predictions are in the same units as those used to build the models; for example, do not build a model with water temperature in degrees Fahrenheit, then import water temperature in degrees Celsius for predictions. It is prudent to include unit information in the column names (e.g., “WaterTemp_C”) to remind the user of the proper unit when entering data to make predictions.
- Missing data (blank cells) are permitted upon import, but must be dealt with (either deleted or values filled in) prior to modeling.
- If Excel data files are imported, cells with non-numeric values (i.e., symbols or text) are converted to empty cells. Exceptions are the column names and the first column of IDs. If such non-numeric characters are present in an imported .csv file, they will be imported into VB₃'s datasheet. However, they will be flagged as anomalous during the validation scan and they must be dealt with (deleted or populated) at that time.
- When the required validation scan is launched, VB₃ will identify any column in the dataset containing only a single value and ask the user to delete the column (because such data columns are useless for predictive purposes).
- There is no hard-coded limit on the number of IVs one can import; however, the VB₃ datasheet is designed for a maximum of 300 columns. Beyond that number, the application's performance will degrade significantly. Investigating 250+ IVs results in over 2×10^{20} possible IV combinations for MLR processing. The MLR genetic

algorithm can handle this modeling task, but choosing “Run all combinations” would likely take months or years to complete. Depending on how many additional IVs will be created by the user, importing a dataset with less than 100 IVs should be acceptable.

We note here that VB₃ can be used as a powerful exploratory research tool, allowing the user to investigate a great many IVs concurrently. However, this approach can lead to models with spurious response/IV relationships (i.e., the association is only a random statistical artifact, not a “real” phenomenon). To avoid this, the user could restrict their analyses to only those IVs for which they have a prior, process-based, theoretical expectation of influence on pathogen concentrations. A criticism of this approach is that the researcher will never discover a relationship between the response and a truly influential IV if they don’t already expect it to exist. Discovery of unexpectedly influential IVs can lead to process insight and advancements in understanding of the physical system. If an exploratory approach is taken, there are mechanisms within the statistical modules of VB₃ (primarily cross-validation to ensure that predictions on future data points are nearly as good as the model fits) to protect against over-fitting a model using too many IVs and finding spurious correlations that don’t hold up when the model is used for prediction of future events.

6.2 Importing a Dataset

When users first click on the Global Datasheet tab, they can import a data file using the “Import Data” button in the top ribbon (Figure 8). This opens a dialog screen where a directory explorer can be used to find the data file. If the file is an Excel workbook with multiple worksheets, the dialog box asks which worksheet to import.

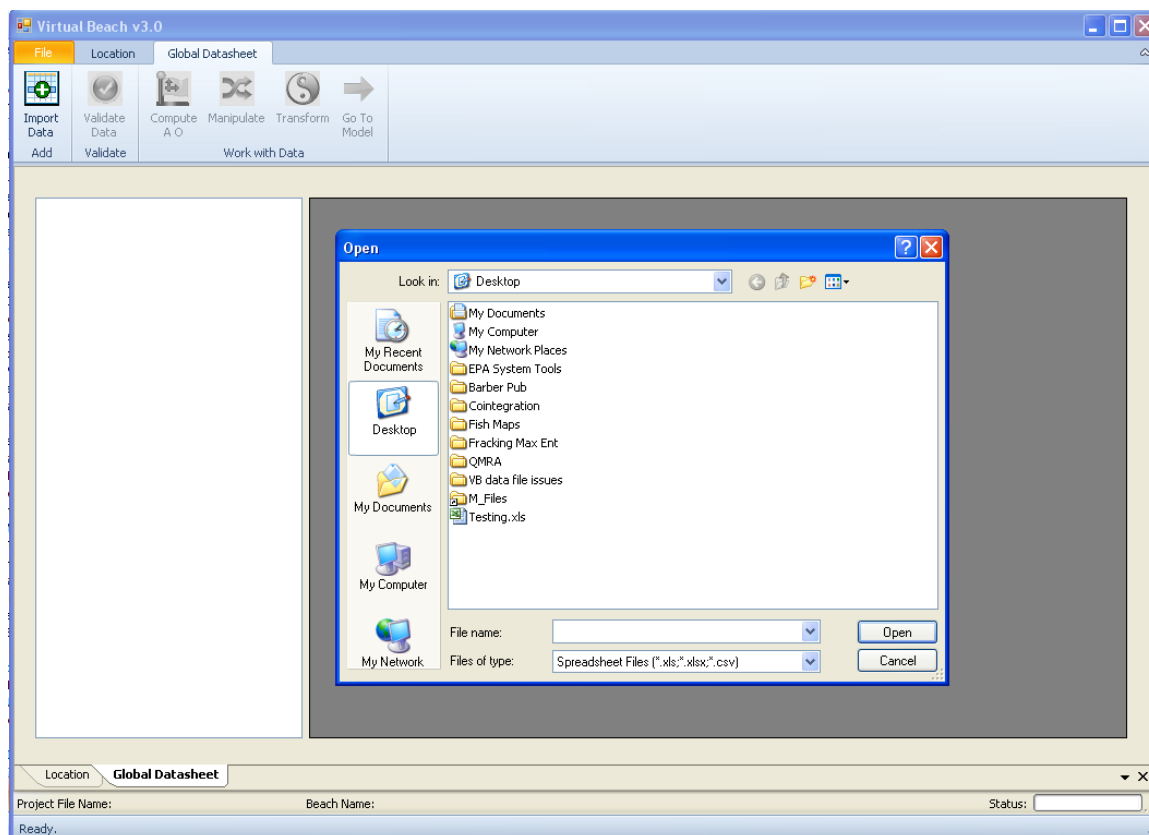


Figure 8. Importing a dataset into the Data Processing tab

Once imported, the data are shown in a datasheet. The second column of this datasheet will be highlighted in blue to indicate its status as the current response variable. Information about the dataset, such as number of rows and columns, name of the ID column and name of the response variable, appear at the left of the datasheet. At this point, the datasheet cannot be edited or interacted with in any manner; to access additional processing functionality, the data must be validated.

6.3 Validating the Imported Data

Validation options can be accessed by clicking the “Validate Data” button in the top button ribbon. Validating the data launches a required scan to identify blank and non-numeric cells in the imported spreadsheet (Figure 9). One can also find and replace other specified values (e.g., a missing data tag like -999) in the dataset, using the “(Optional) Find:” input box.

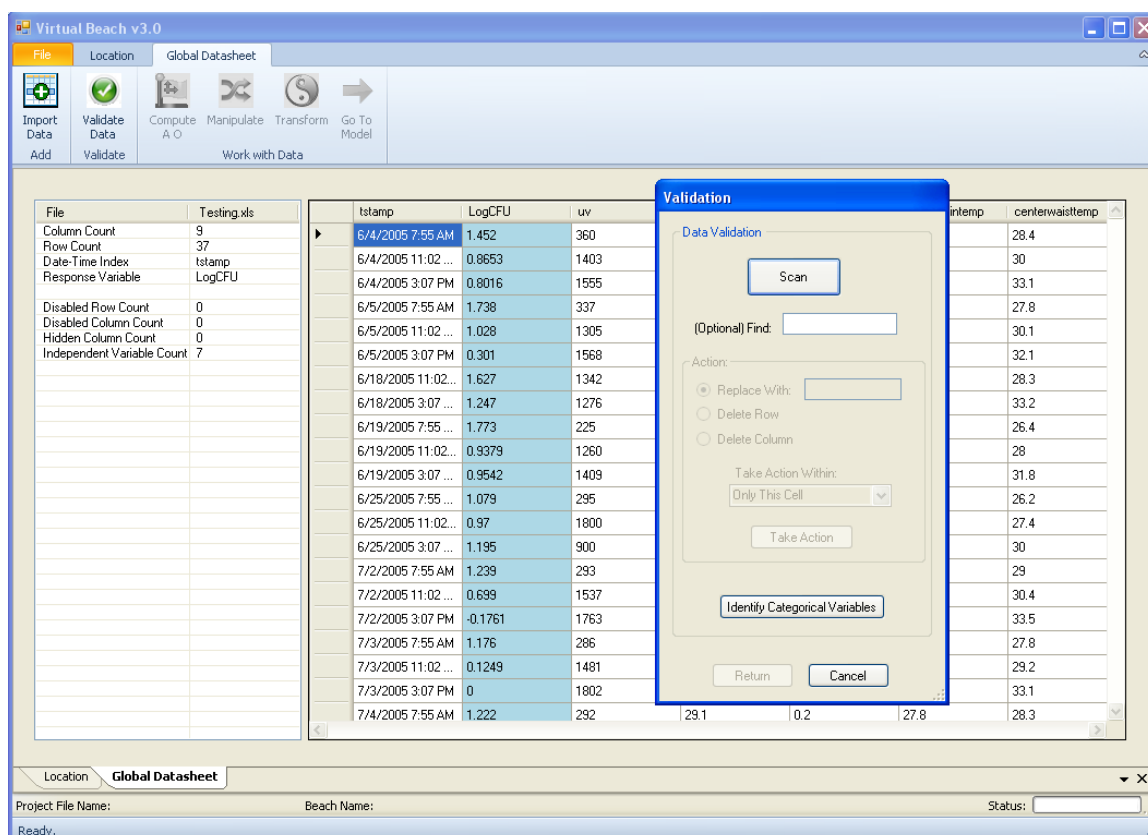


Figure 9. Data validation required to begin data processing

Clicking “Scan” begins the validation process. VB₃ goes through the datasheet, cell by cell, looking for blanks, non-numeric, or user-specified values entered in the “(Optional) Find:” input box. If such a cell is found, the scan will stop and highlight it. Users must then decide how to deal with that cell from choices in the “Action” section (Figure 10): replace the cell with a specified value, using the “Replace With:” input box, or delete the row or column containing the cell. The user must decide where to implement the chosen action with the “Take Action Within” dropdown menu. Possible choices are “Only this Cell,” “Entire Row,” “Entire Column,” and “Entire Sheet.” Items in this menu are context-sensitive, i.e., they change with the Action selected. After setting the “Take Action Within” menu, the user clicks the “Take Action” button, VB₃ makes the specified changes to the datasheet, and the scan continues. Even if no cell errors are found, VB₃ may still report that a “Column has no distinct values” and prompt the user to delete the column (see the second-to-last bulleted item in Section 6.1). When the entire datasheet has passed inspection, VB₃ reports “no anomalous data values found” at the bottom of the Validation window.

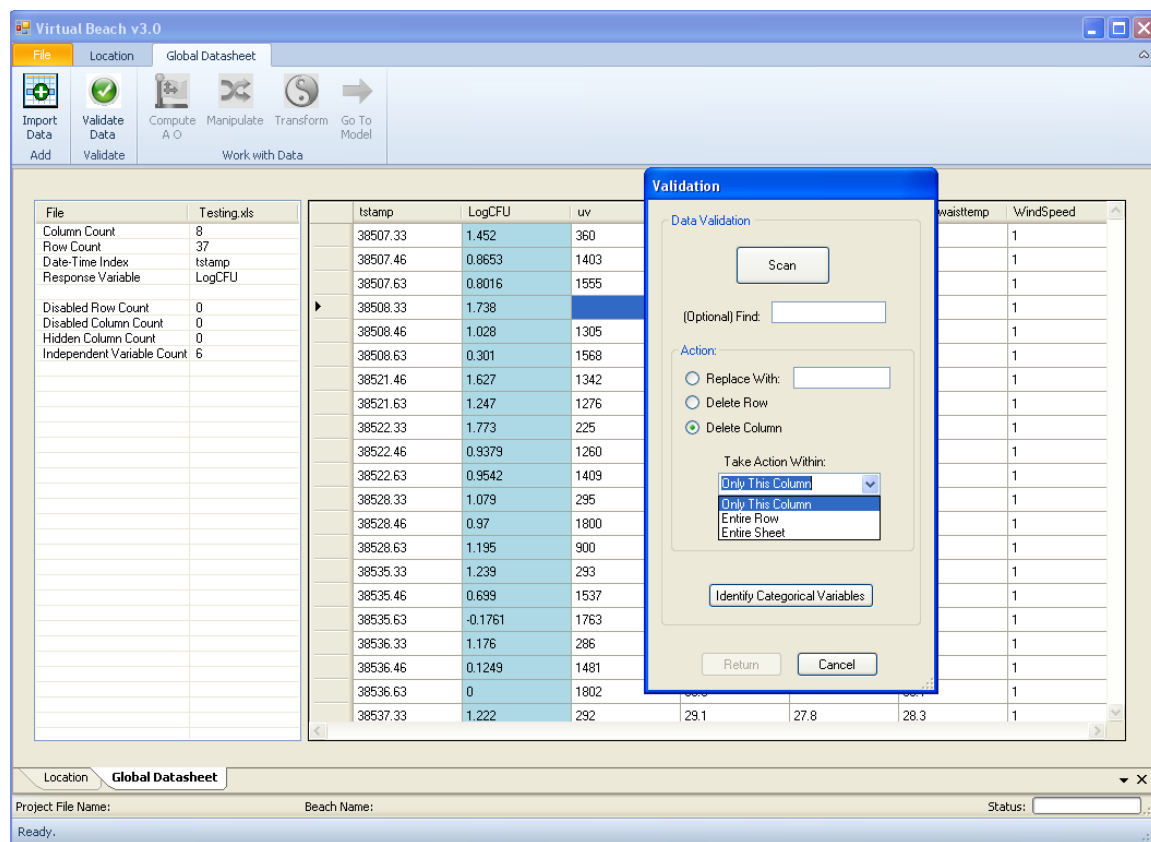


Figure 10. Context-sensitive choices for the “Take Action Within” drop-down menu

After the data have been validated, but prior to clicking the “Return” button on the Validation window, the user has the option to specify which columns in the dataset are categorical variables. Why do this? VB₃ will not attempt to transform categorical data columns (transformations discussed later), because it generally does not make sense to do so. Thus, identifying IV columns as categorical saves time later when transformations are investigated. If the user clicks on the “Identify Categorical Variables” button (Figure 10), a window pops up (Figure 11). A list of the datasheet’s independent variables is shown in the right-hand section of this window. VB₃ automatically identifies columns with only two unique values as categorical variables (i.e., they will already be in the left section of this window); if the user has other categorical IVs with more than two categories, those should be moved from the right to the left section using the  button. The user can also move any currently-identified categorical IV back to the right list using the  button.

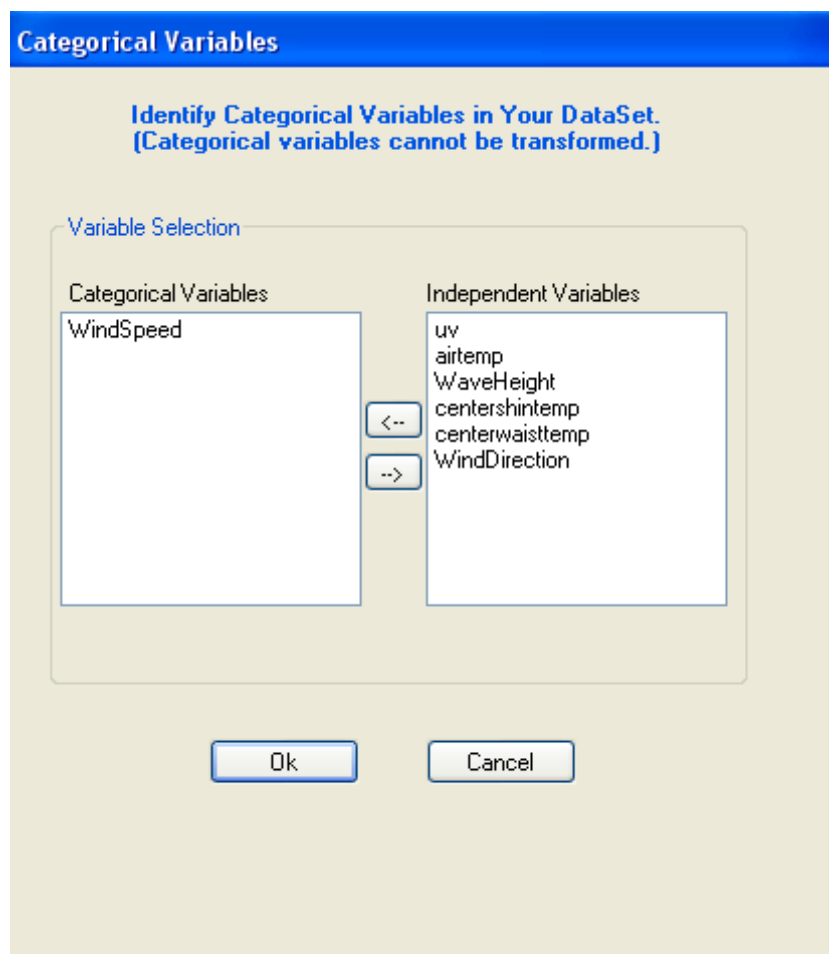


Figure 11. Pop-up window for identifying categorical variables.

6.4 Working with a Dataset after Validation

After the dataset has passed the validation scan, the function buttons across the top of the Global Datasheet tab ribbon are enabled (Figure 12).

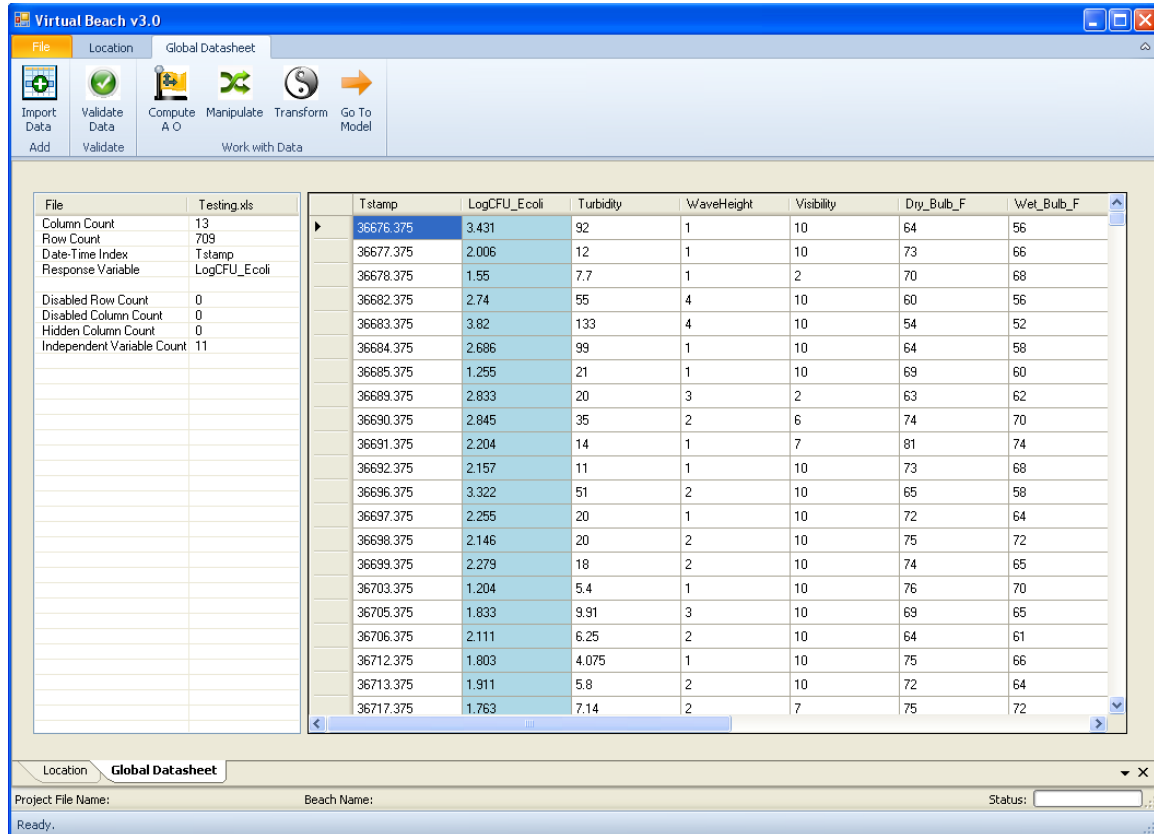


Figure 12. Post-validation enabling of the Global Datasheet functionality

At this point, grid cells (other than the ID column) are editable – that is, users can manually enter new numeric data with a left-double-click on a cell and typing in a new value. VB₃ does not allow a cell to be made blank or non-numeric. A right-click on an IV column header presents additional options (Figure 13):

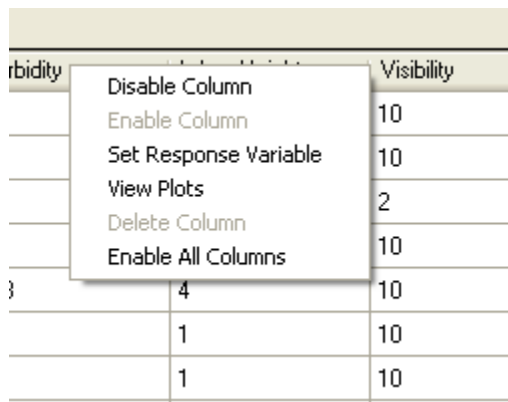


Figure 13. Right-click options on columns that are not the response variable

“Disable Column” turns the text red and prevents the column from being passed to the method tabs. Previously-disabled columns can be activated with “Enable Column.” “Set Response Variable” makes the chosen IV the new response variable (the column becomes blue to indicate this change). “View Plots” shows a new screen with column statistics at the far left and four plots for the chosen column (Figure 14): (1) a scatter plot of the IV versus the response variable in the lower left panel; (2) a plot of the IV values versus the ID column at the upper left (a time series plot if the ID is an observation date); (3) a box-and-whiskers plot at the top right; and (4) a histogram for IV values at the bottom right.

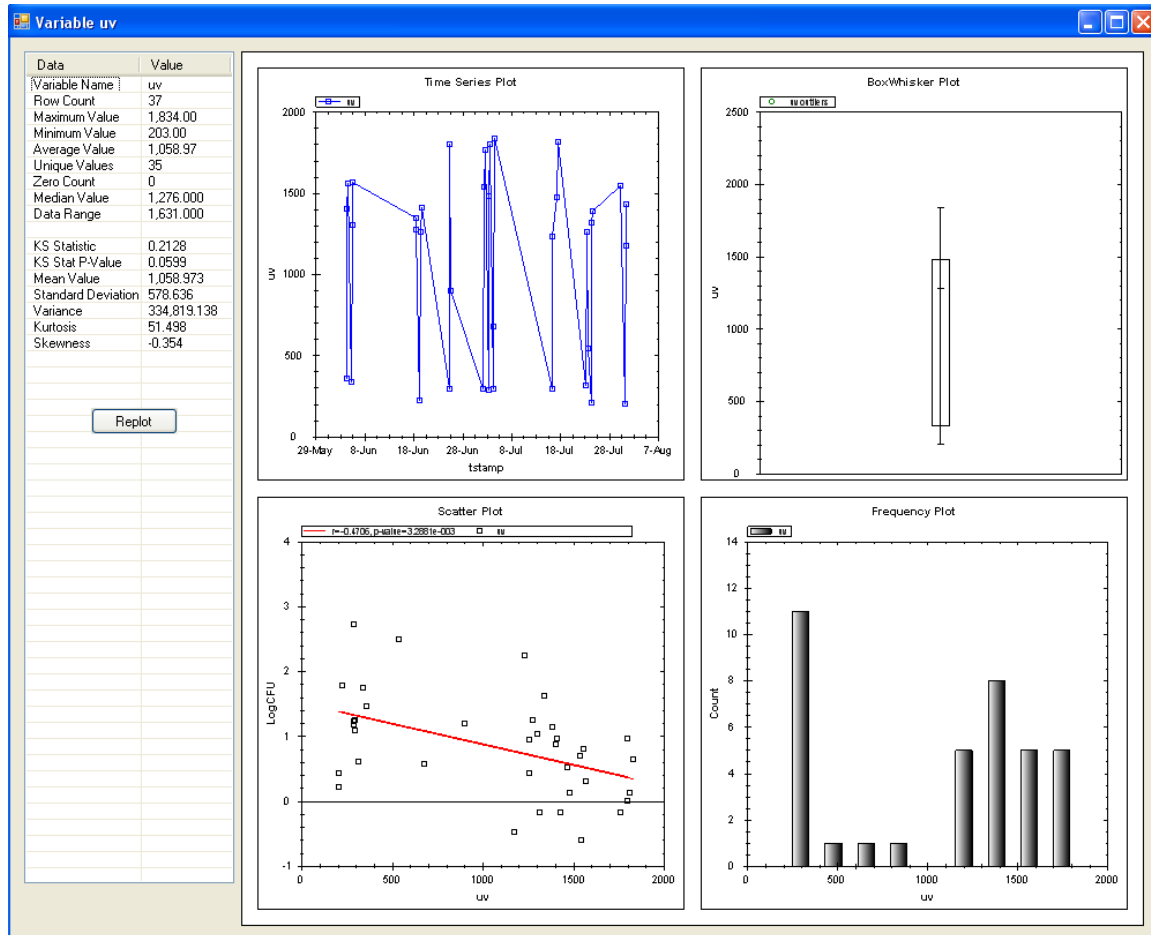


Figure 14. Four different plots available for evaluation of IVs

Scatter Plot Interpretation

Curvature in the scatter plot (lower left) can indicate a non-linear relationship between the IV and the response variable, problems with homogeneity of variance across the range of the IV, or outliers. Ensuring that the IVs are linearly related to the response variable raises the probability of producing a robust, meaningful MLR and PLS analysis (GBM does not need linearity). If the relationship between the response and the IV is not well-approximated by a straight line (a fundamental assumption of MLR and PLS), it may be beneficial to transform the IV. Using VB₃ to accomplish this will be explained later (Section 6.7). The scatter plot also shows the best-fit linear regression line in red,

along with the correlation coefficient (r) and the significance (p -value) of the correlation coefficient at the top of the plot. In general, p -values below 0.05 are considered statistically significant. While VB₃ does not provide a plot of the residuals of the regression line depicted in the scatter plot, this important diagnostic is given much attention on the MLR tab (see Section 7.8).

Identifying odd values (potential outliers or bad data) of any IV can often be done by visual inspection. If users move the mouse cursor over a data point in any plot (other than the histogram), they will see the ID value of that observation (Figure 15). They can then go back to the datasheet, find the outlying observation (data row), and disable that row (described below) if justifiable.

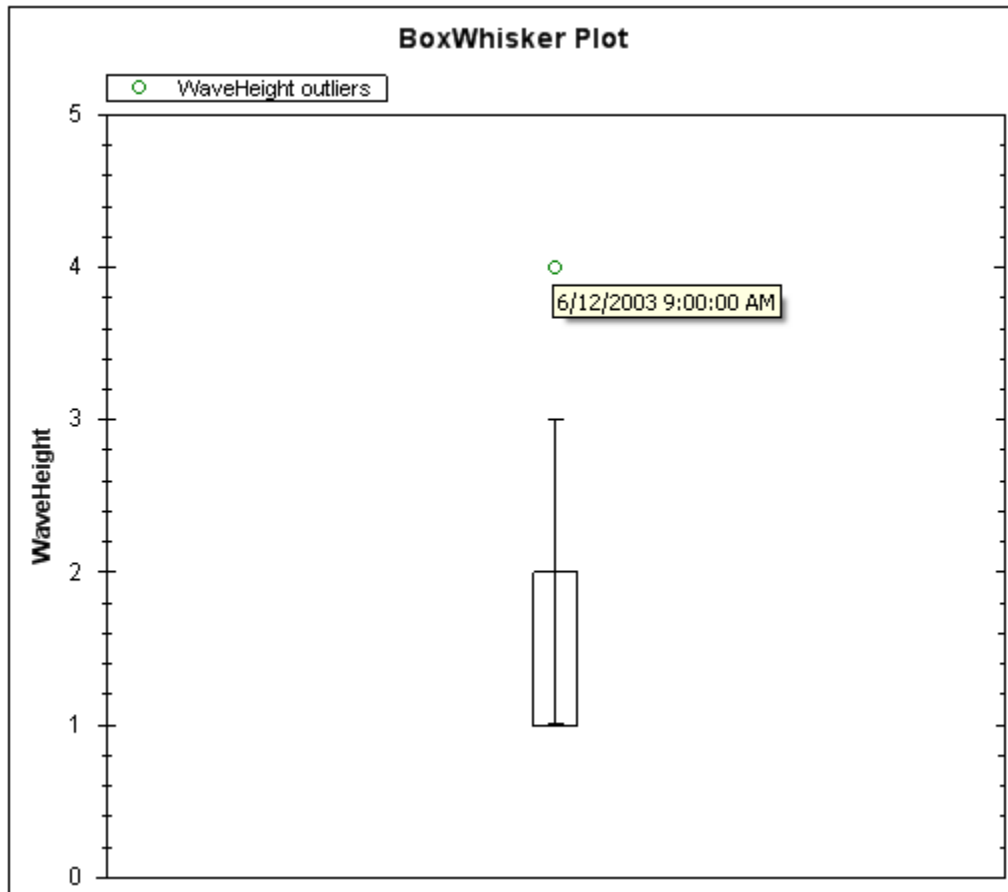


Figure 15. Identifying an observation from within the XY scatter plot

The “Delete Column” right-click column header option deletes a column from the VB₃ datasheet. Note that original columns of the imported data sheet (VB₃ defines these as “main effects”) cannot be deleted. Rows can be disabled and enabled, but not deleted, from the datasheet by right-clicking the row header (far left of each row) and making the desired choice. Changes that the user makes can be undone and redone using the “Undo” and “Redo” options under the VB₃ “File” menu.

If the user right-clicks on the column header of the response variable, a different set of choices is shown (Figure 16).

	Tstamp	LogCFU_E		WaveHeight
	36676.375	3.431	Transform	1
	36677.375	2.006	View Plots	1
	36678.375	1.55	UnTransform	1
	36682.375	2.74	Define Transform	4
	36683.375	3.82		4
	36684.375	2.686		1
	36685.375	1.255		1
	36689.375	2.833		3

Figure 16. Available choices when right-clicking the response variable

Users can transform the response variable in three ways: $\log_{10}$, $\log_e$, or a power transformation (raising the response to an exponent: y^λ). They can also un-transform the response, view the plots shown previously for the IVs, or define a transformation of the response variable. This last option is used when a datasheet is imported with an already-transformed response variable. For example, users could import a datasheet with $\log_{10}$ -transformed fecal indicator bacteria concentrations and should define the response as $\log_{10}$ -transformed. Doing this facilitates later comparisons with the fitted response variable values, decision criteria, and regulatory standards. If this is not done, then later plots and comparisons of model predictions to response variable values will be strange and misleading. When users transform the response variable within VB₃ using the “Transform” option, VB₃ automatically defines the response as having the chosen transformation and, in doing so, synchronizes the units of measurement for later comparisons.

6.5 Computing Wind, Wave and Current Components

Orthogonal wind, current, and wave components can be powerful predictors of beach bacterial concentrations. Depending on the orientation of the beach, wind and currents can influence the movement of bacteria from a nearby source to the beach, and wave action can re-suspend bacteria buried in beach sediment. To make more sense of this information, researchers typically decompose wind/current/wave magnitude and direction data into A (alongshore) and O (offshore/onshore) components for analysis (see equations at the end of this section).

If direction and magnitude (speed/height) data are available, A and O components can be calculated with the “Compute A O” button in the ribbon (Figure 12). Clicking it brings up a window with drop-down menus for users to specify which columns of the datasheet contain the relevant magnitude and directional data (Figure 17). There is also an input box at the bottom of the form for the beach orientation angle. If the user defined the beach angle on the “Location” tab, that value will be seen. After clicking “OK,” new data columns are added to the far right of the grid, representing the A and O components of the specified wind, current, or wave data. Unlike the originally-imported IVs, these components can be deleted from the grid after creation. Names of these new columns are: WindA_comp(X,Y,Z), CurrentO_comp(X,Y,Z), WaveA_comp(X,Y,Z), etc., where

X is the name of the column of data used for direction, Y is the name of the column used for magnitude, and Z is the beach orientation angle. Note that the IVs used to create the A and O components are automatically disabled by VB₃ once the components are created. These columns can be re-enabled by right-clicking on their column header in the datasheet and choosing “Enable Column.” The “Compute A O” function is repeatable as many times as the user wishes.

The image shows a software dialog box titled "Wind/Current/Wave Components". It contains three main sections for specifying data columns:

- Wind Data:** Includes "Specify wind data columns:" with dropdowns for "Speed" and "Direction (deg)".
- Current Data:** Includes "Specify current data columns:" with dropdowns for "Speed" and "Direction (deg)".
- Wave Data:** Includes "Specify wave data columns:" with dropdowns for "Wave Height" and "Direction (deg)".

At the bottom of the dialog, there is a "Beach Angle (deg):" field with the value "0.00" and two buttons: "Ok" and "Cancel".

Figure 17. Window for computation of alongshore and offshore/onshore components

Notes on Component Calculations

Direction is an angular degree measure. Moving in a clockwise direction from north (0 degrees), values are positive, and negative while moving counter-clockwise. Wind and current speed (as well as wave height) can be measured in any unit. VB₃ adheres to scientific convention: wind direction is specified as the direction from which the wind

blows and current and wave directions are specified as the direction towards which the current or waves move. Thus, wind blowing west to east has a direction of 270 degrees (or equivalently -90) degrees, while a current/wave also moving west to east has a direction of 90 (or -270) degrees.

The A-component measures the force of the wind/current/wave moving parallel to the shoreline (Figure 18). A positive A-component means winds/currents/waves are moving from right to left as an observer looks out onto the water. A negative A-component means winds/currents/waves are moving left to right as an observer looks out onto the water. The O-component measures force perpendicular to the shoreline. A negative O value indicates movement from the land surface directly offshore (unlikely to be seen with wave action). A positive O indicates waves/wind/currents from the water to the shore. These relationships apply no matter how the beach is oriented (Figure 19).

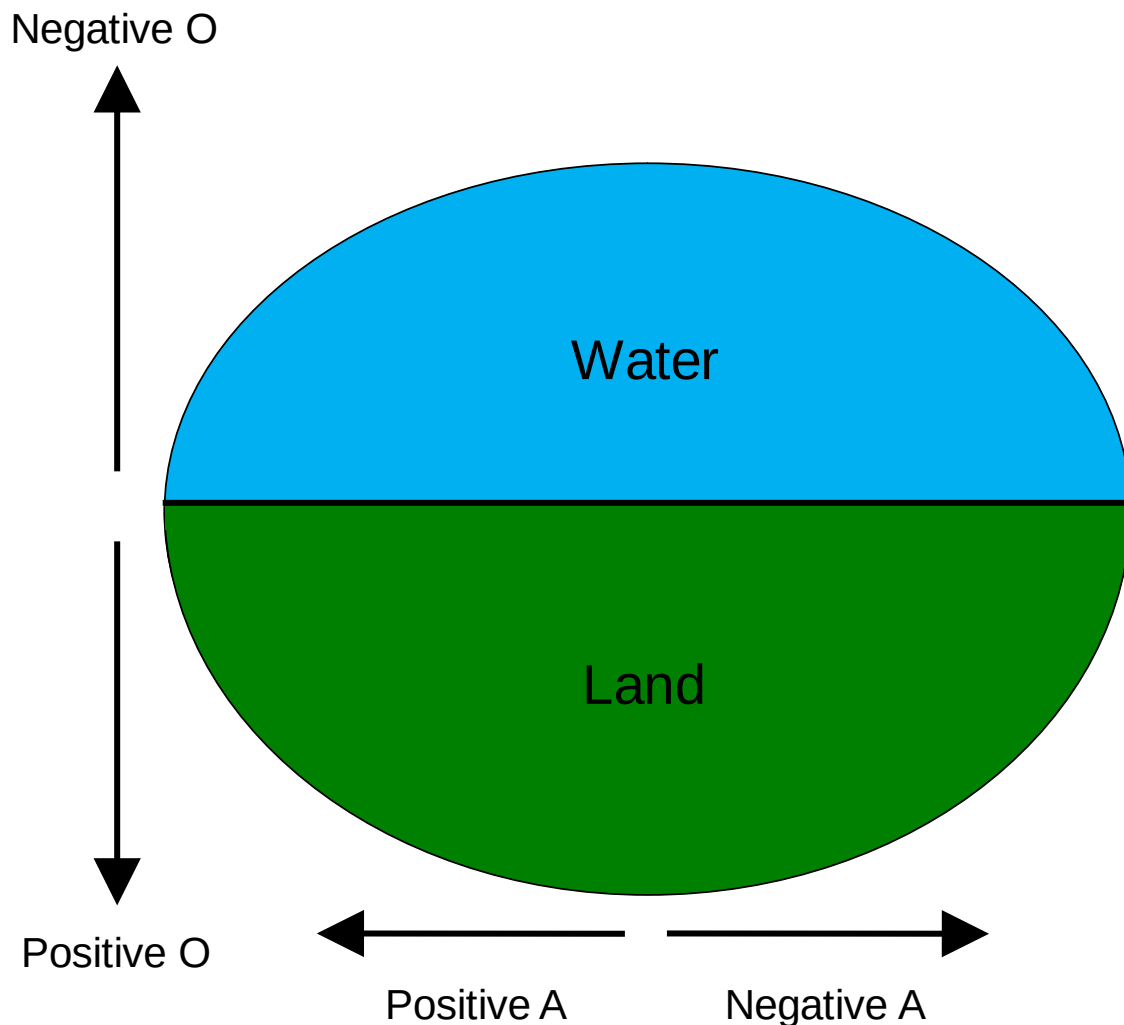


Figure 18. A- and O-component definitions for wind, current, and wave data

Beach Orientation for Component Calculations

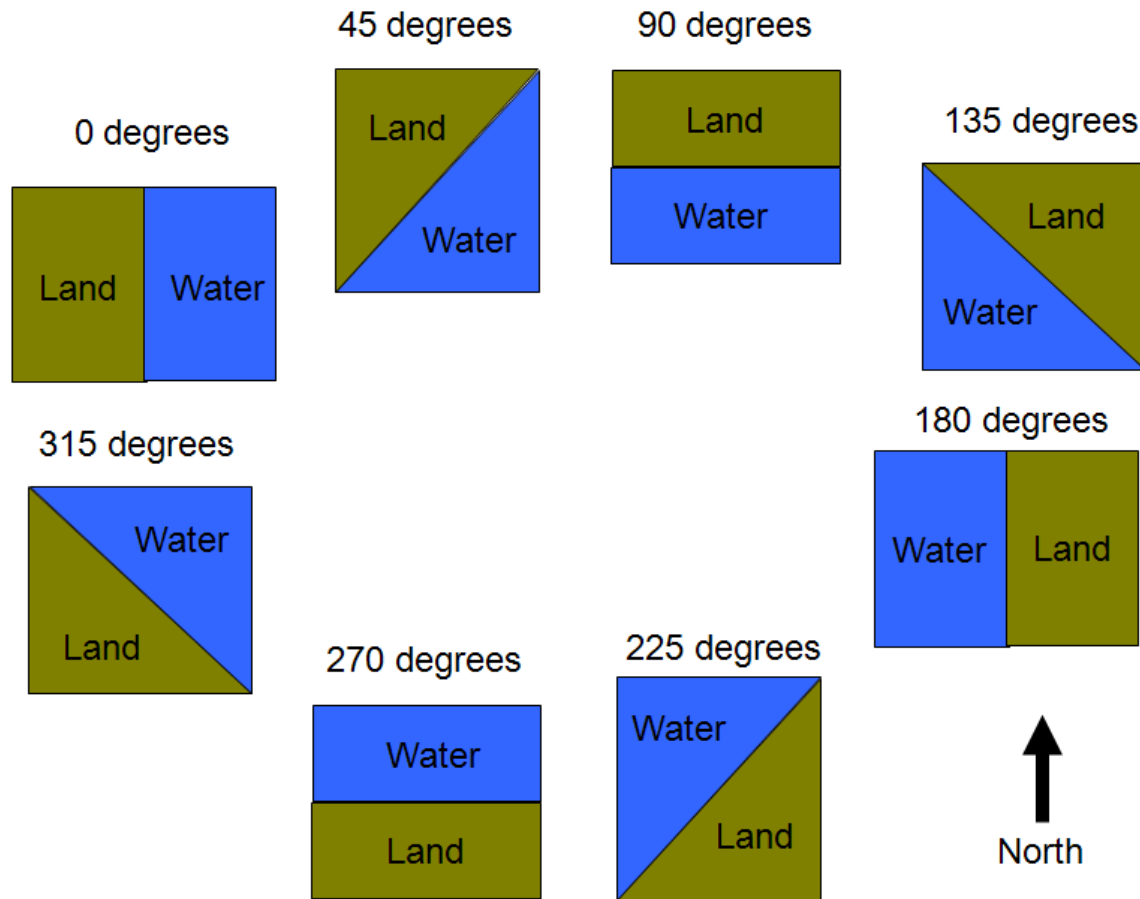


Figure 19. Principal beach orientations given in degrees

The equations for calculation of Wind A/O components:

$$\text{Wind A: } -S * \cosine((D-B) * \pi/180)$$

$$\text{Wind O: } S * \text{sine}((D-B) * \pi/180)$$

where S is wind speed, D is wind direction, B is the beach orientation (in degrees) and $\pi \approx 3.1416$. Current A/O and Wave A/O are the same equations multiplied by -1 to account for the difference in how these data are measured.

6.6 Creation of New Independent Variables

Users may click the “Manipulate” button (Figure 11) to create new columns of data (as functions of existing IVs) that might be useful IVs. On the pop-up screen (Figure 20), there is a list (automatically populated by VB₃ from the imported spreadsheet) of available IVs on the far left under “Independent Variables.” If users wish to create a new term, they add the desired existing IVs to the “Variables in Expression” box by selecting the IV and clicking the “>” button. Clicking and dragging, shift-clicking and control-clicking in the “Independent Variables” list allow multiple IVs to be added at once.

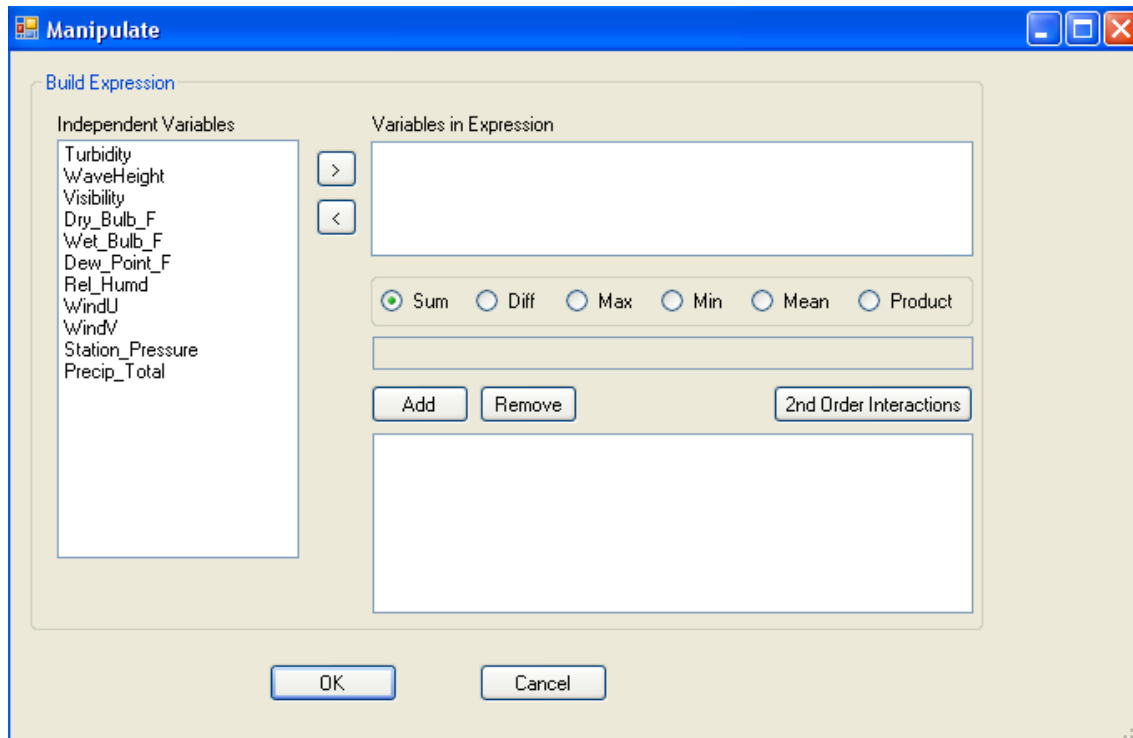


Figure 20. Window for the formulation of “Manipulates” - arithmetic combinations of existing columns within the datasheet

For example, if users wish to create a new IV that is a row-by-row mean value of the “Dry_Bulb_F” and “Wet_Bulb_F” variables, they add those two IVs to the “Variables in Expression” box (Figure 21), choose the “Mean” function, “Add” that expression to the lower box, then click “OK.” A new column of data representing a row-by-row average of those two IVs is then added to the end of the datasheet.

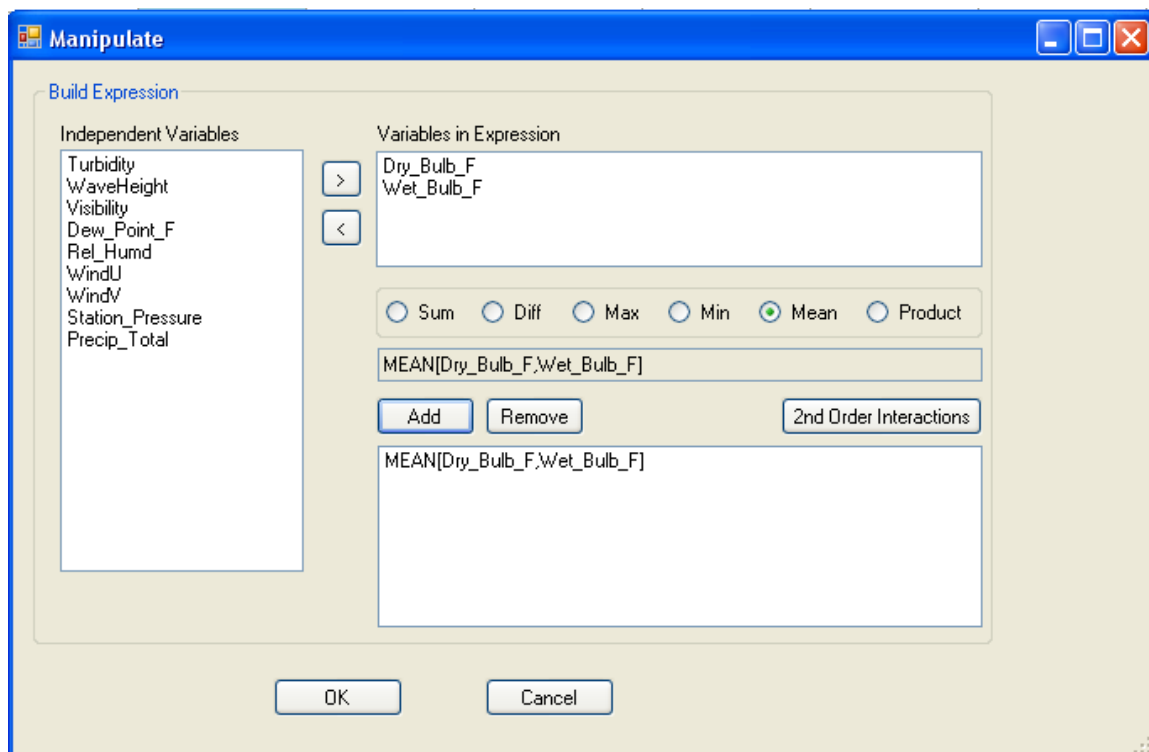


Figure 21. Creation of a new IV defined as the mean of two existent IVs

Users can create a row-by-row sum, difference, maximum, minimum, mean, or product from any number of IVs added to the “Variables in Expression” box. More than one expression can be created before the “OK” button is clicked and IVs can be easily moved in and out of the “Variables in Expression” box using “<” and “>” keys. Note that creating a difference of more than two columns (e.g., X1, X2, X3, and X4) would lead to this quantity:

$$\text{Diff}(X1,X2,X3,X4) = X1 - X2 - X3 - X4$$

Created expressions can be removed from the lower box with the “Remove” button. No matter how many IVs are added to the “Variables in Expression” box, clicking “2nd Order Interactions” will add the cross-products for all possible pairings of those IVs (Figure 22). Thus, four IVs in the “Variables in Expression” box will produce six 2nd second-order interactions; five IVs will produce ten interactions, and so on. Note that the names of the columns used to create any new data columns are inside the parentheses of those columns’ names.

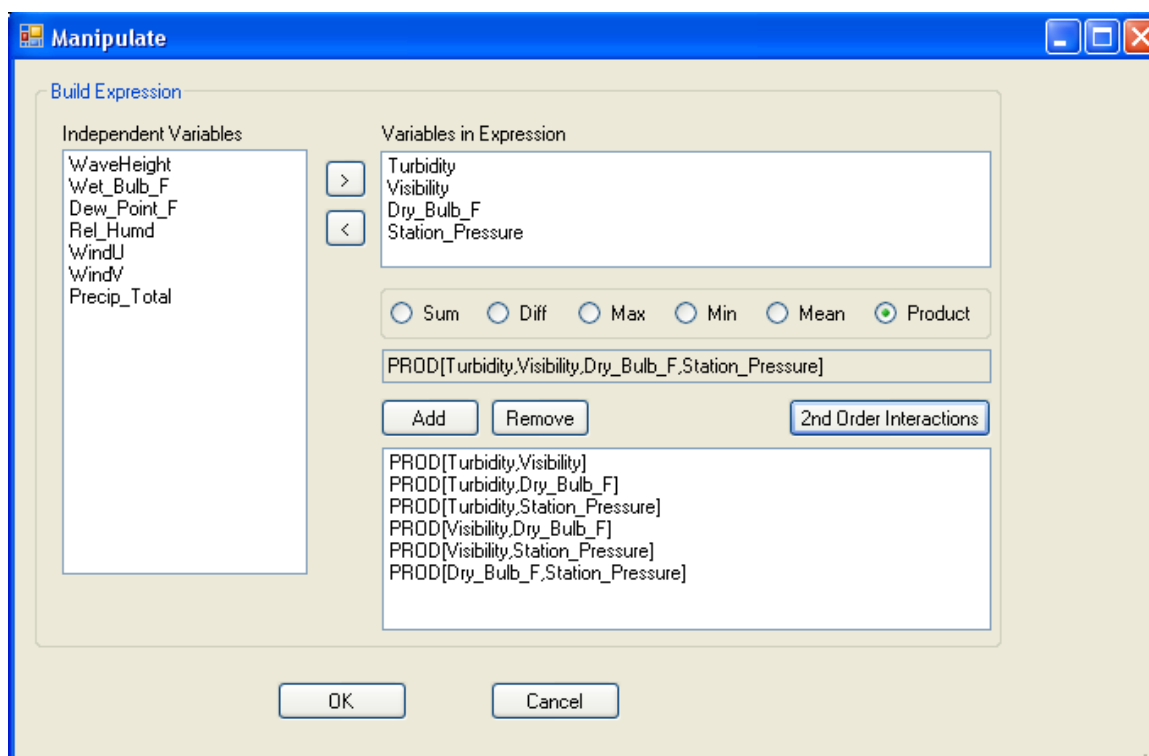


Figure 22. Formation of two-way cross-products of a set of four IVs

VB₃ does not allow previously created “manipulates” -- new columns of data created through the “Manipulate” button -- to be further manipulated. Previously created manipulates will not appear in the “Independent Variables” section at the left. They can, however, be chosen as the response variable or deleted from the datasheet, using the appropriate menu choices accessed by a right-click of the column header.

6.7 Transforming the Independent Variables

VB₃ gives users the ability to transform non-categorical IVs to assist in linearizing the relationship between the IVs and the response variable, a fundamental assumption of an MLR/PLS analysis. VB₃ transformations are described in section A.1. When users click the “Transform” button (Figure 12) in the Global Datasheet ribbon, they are presented with the window seen in Figure 23:

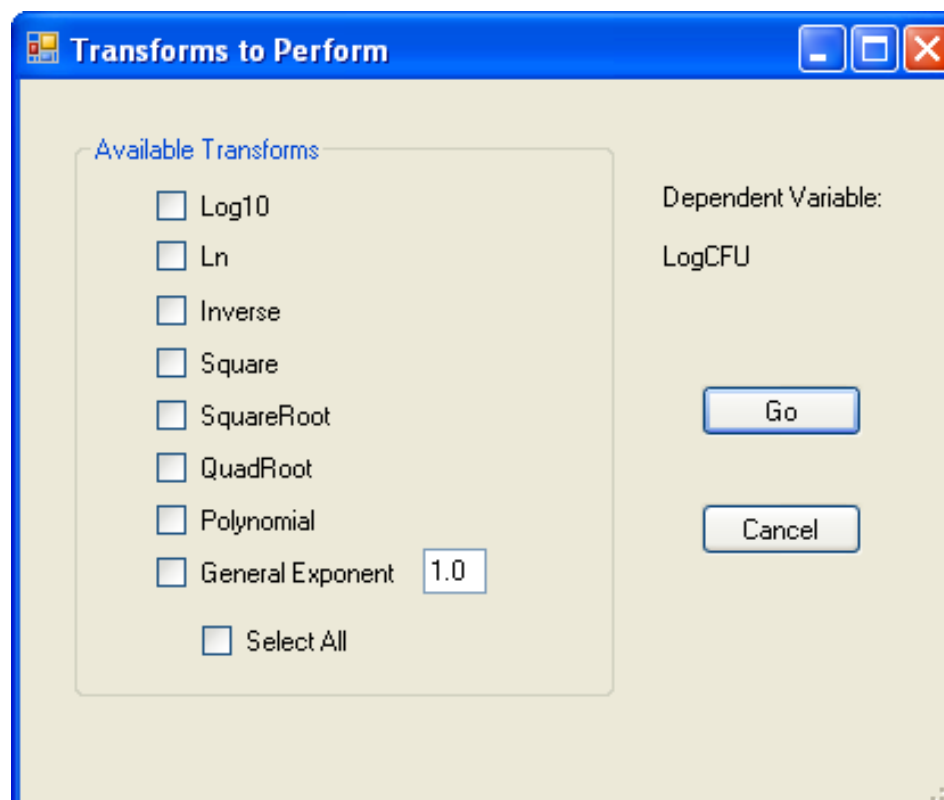


Figure 23. The choices for IV transformations

When users click “Go,” the chosen transformations are applied to each and every non-categorical IV (there is not an option to ignore transformation for particular IVs). VB₃ then opens a table (Figure 24) that compares the success of each transformation using a Pearson correlation coefficient which is a measure of linear dependence between the response variable and the IVs.

The table created by VB₃ groups all transformed versions of each IV and specifies type of transformation, the Pearson coefficient, and its statistical significance (p-value). This includes the un-transformed version of the IV, denoted by “none.” By default, the transformation with the largest absolute value of the Pearson coefficient is highlighted in black text. Users may override the default selection by left-clicking on the row header of a transformed IV. They may also override the default by setting a percentage and clicking “Go” under the “Threshold Select” box on the left side of the window. This will select the un-transformed version of every IV unless the transformed IV with the highest absolute value Pearson coefficient exceeds the un-transformed IV Pearson coefficient by the specified percentage. In essence, the user is saying, “Unless the Pearson coefficient of the transformed IV is some % greater than the Pearson coefficient of the un-transformed IV, use the un-transformed IV.” This can be useful because transforming IVs makes interpreting model coefficients more difficult; unless a major improvement is seen, transformation simply may not be worth the trouble. Users can also revert to the default (selecting the transform with the largest absolute value Pearson coefficient) by clicking “Go” under “Auto Select.”

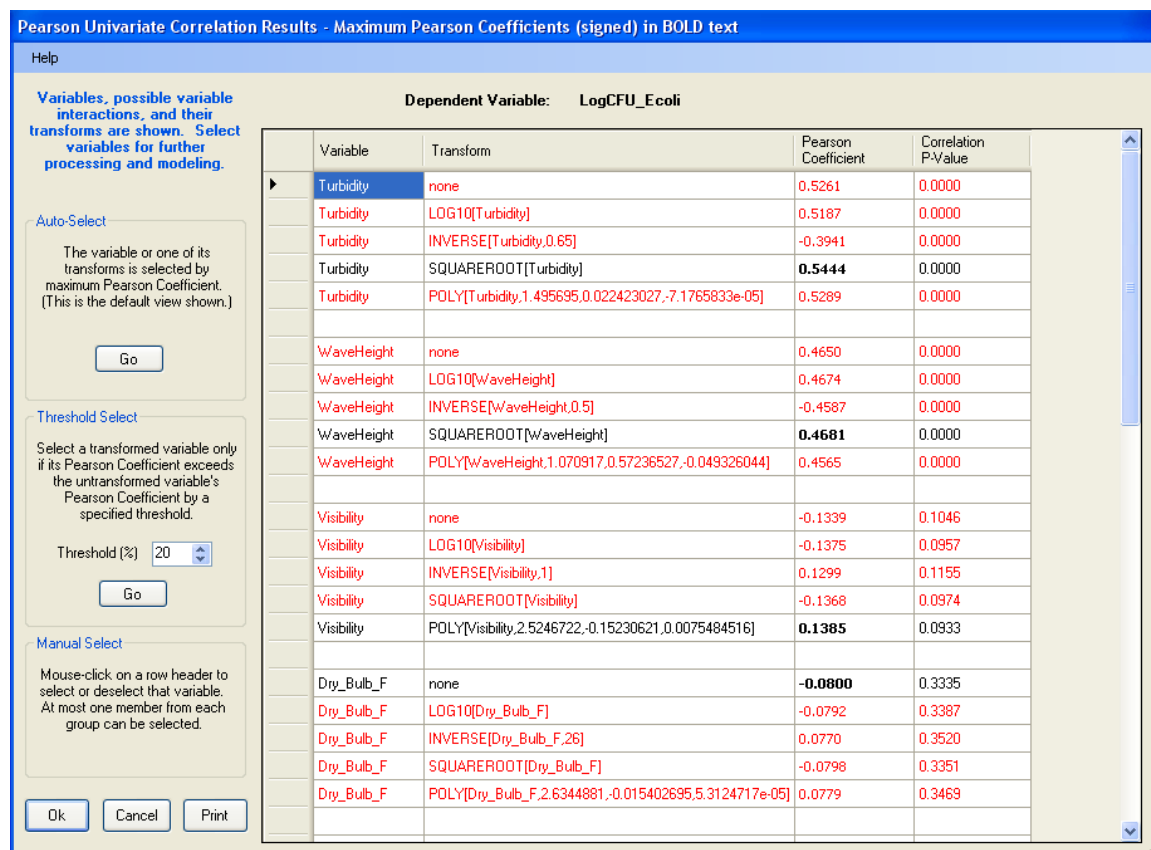


Figure 24. Pearson correlation coefficient scores for judging the efficacy of IV transformations

Plotting Transformed IVs

Users may prefer to examine plots visually in determining which transformation of IV to choose. Right-clicking on a row header in the correlation table provides an array of scatter plots, time series plots, or frequency plots for each transformation of that IV (Figure 25). Scatter plots show the best-fit regression line, the correlation coefficient, and the p-value for that correlation coefficient.

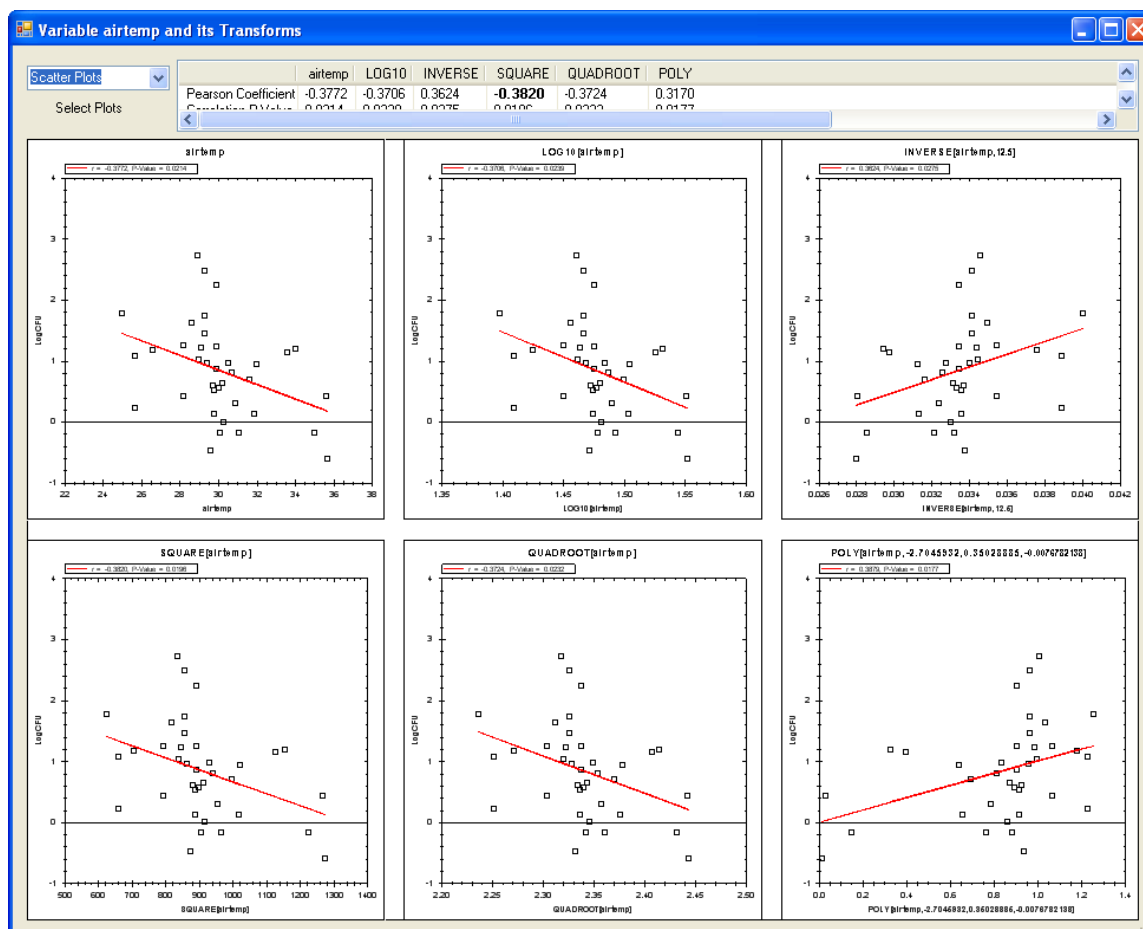


Figure 25. Scatter plots (Response vs. IV) for six different data transformations of a single IV

After choosing a transformation for each IV, users click “OK.” This populates the datasheet with new columns representing transformed versions of the IVs. Notice two things: if a transformation was chosen for an IV, the column representing the untransformed version of that IV is disabled in the datasheet (it can be re-enabled by using the right-click column header menu option) and the transformed versions of an IV are put into the datasheet immediately after the original, un-transformed IV. Any transformations put into the datasheet can be deleted with the “Delete Column” choice (right-click on their column header). Transformed IVs will appear in the list of IVs on the “Manipulate” screen, however, transformed IVs cannot be further transformed and will not appear in the transform table if the user returns to the “Transform” window. Also, transformed IVs cannot be the response variable. Finally, because transformations are determined from the current response variable, all transformed IVs in the datasheet are erased (a warning appears) when users change the response variable in the datasheet. For the interested reader, further discussion of VB₃ transformations can be found in section A.1.

6.8 Singular Matrices and Nominal Variables

Advice on avoiding singularities within the data matrix and handling nominal categorical variables can be found in section A.2.

6.9 Saving Processed Data

Changes made to the imported spreadsheet can be saved in a project file (File→Save). When it is re-opened, the datasheet will appear as it did when the project was saved. Users also may highlight the entire datasheet or sections of the datasheet and use Control-C and Control-V to copy and paste it into a word processing or spreadsheet application.

6.10 Proceeding to Modeling

After data processing is complete, users must click the “Go to Model” button to open the statistical method tabs. If they have already done some modeling and return to the global datasheet to make changes, they will receive a message that the datasheet has changed and any prior modeling results will be erased.

7. MULTIPLE LINEAR REGRESSION MODELING

The MLR tab finds the best multiple linear regression model based on criteria selected by the user. As the number of IVs increases, the number of possible models in the solution space increases exponentially. Users may select all or a subset of the IVs for consideration in the model to reduce the size of the solution space.

Notice that the MLR tab (as well as the PLS and GBM tabs) has its own datasheet on the “Data Manipulation” sub-tab. When the user first moves over to the MLR tab from the Global Datasheet, the data in the MLR Data Manipulation sub-tab is identical to the data on the Global Datasheet. Once inside the MLR tab, the user can change the “local” data to suit the MLR analysis. The local datasheet has all of the functionality of the Global Datasheet discussed in Section 6. Changing the local data has no effect on the Global Datasheet, however, going back to the Global Datasheet and making changes causes local datasheets on the MLR, PLS, and GBM tabs to be overwritten.

7.1 Selecting Variables for Model Building

Under the “Model” sub-tab, two additional sub-tabs are found (Figure 26). On the “Variable Selection” sub-tab, all eligible IVs are listed in the left column (“Available Variables”). Any variable users wish to consider for model inclusion must be moved to the right column list (“Indep. Variables”) by highlighting the IV and clicking the “>” button. IVs currently under consideration (in the right list) can be ignored by highlighting them and clicking the “<” button. The user can hold down shift while left-clicking or control while left-clicking to select multiple IVs at once.

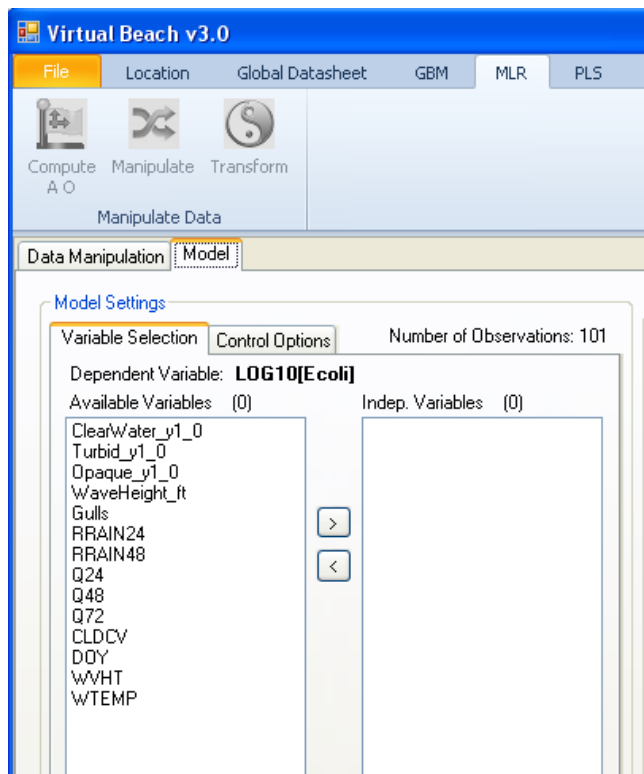
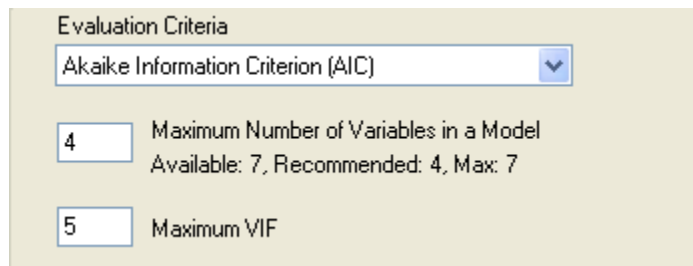


Figure 26. Selecting variables for MLR processing within the Modeling tab

7.2 Modeling Control Options

After choosing the set of IVs to investigate, the user should click the “Control Options” sub-tab. The first decision to be made involves which evaluation criterion will be used to judge model fitness (Figure 27). There are ten choices in the drop-down menu:

- Akaike Information Criterion (AIC)
- Corrected Akaike Information Criterion (AICC)
- R^2
- Adjusted R^2
- Predicted Error Sum of Squares (PRESS)
- Bayesian Information Criterion (BIC)
- RMSE
- Sensitivity
- Specificity
- Accuracy



The screenshot shows a dialog box titled "Evaluation Criteria". It contains a drop-down menu with "Akaike Information Criterion (AIC)" selected. Below the menu are two input fields: the first is labeled "Maximum Number of Variables in a Model" with the value "4" entered, and the second is labeled "Maximum VIF" with the value "5" entered. Below the first input field, there is a note: "Available: 7, Recommended: 4, Max: 7".

Figure 27. Setting modeling options within the modeling interface

Depending on the evaluation criteria, VB₃ searches for a minimum or maximum value. The minimum value for AIC, AICC, BIC, RMSE, and PRESS is used to choose a model, while the maximum is used for R^2 , Adjusted R^2 , accuracy, specificity, and sensitivity. A more detailed description of each criterion can be found in section A.3.

Sensitivity, specificity and accuracy are special cases requiring users to enter both a Decision Criterion (DC) and Regulatory Standard (RS) so that true/false positives and true/false negatives can be defined (Figure 28). The user chooses the DC value. Model predictions above this threshold are considered exceedances/positives, and model predictions below this value are considered non-exceedances/negatives. The RS is typically a safety limit on fecal indicator bacteria (FIB) concentrations set by a state or federal agency. The “Threshold Transform” radio buttons tell VB₃ the units of DC and RS to ensure a proper comparison to model predictions and observations. For example, if “235” is entered into the DC box (representing the EPA standard for freshwater *E.coli*), then “none” should be chosen. If 2.371 (= $\log_{10}(235)$) is entered as the DC, then “Log10” is used. The DC and RS should always use the same units. Improper setting of this button choice will lead to problems later when comparing modeling predictions to observations.

Model Evaluation Thresholds

235 Decision Criterion (Horizontal)

235 Regulatory Standard (Vertical)

Threshold Transform

☒ None

☐ Log10

☐ Ln

☐ Power

Current US Regulatory Standards

E. coli, Freshwater:	235
Enterococci, Freshwater:	61
Enterococci, Saltwater:	104

Figure 28. Setting evaluation thresholds and threshold transformation information within the modeling interface

The “Maximum Number of Variables in a Model” parameter tells VB₃ the maximum allowable size for any tested models. In general, one should have about 10 observations per estimated parameter in a model, otherwise model over-fitting and poor estimation of regression parameters can occur. VB₃ recommends this limit be set to $(1 + n/10)$ parameters, where n is the number of observations in the dataset. The maximum allowable limit is $n/5$. The total number of available parameters is also shown.

The “Maximum VIF” (Variance Inflation Factor) is used to discard models containing variables with a high degree of multi-collinearity, i.e., IVs that are highly correlated with other IVs in the model. If any IV in a model has a VIF exceeding the VIF threshold, that model will be ignored. The default VIF is 5, which means that 80% ($1 - 1/VIF = 1 - 1/5 = 4/5$) of the variability in an IV can be explained by the other IVs in the model. A VIF of 10 means that 90% ($1 - 1/10 = 9/10$) of the IVs variability can be explained, and so on. Raising the Maximum VIF means a higher degree of multi-collinearity will be tolerated, but this can lead to poorly estimated regression coefficients (i.e., large standard deviations of these coefficients).

7.3 Linear Regression Modeling Methods

Two buttons are at the bottom of the “Control Options” sub-tab to provide different ways of exploring the regression solution space (Figure 29).

- The Manual button is for a directed model search. If the ‘Run all combinations’ box is not checked, only a single model that includes every IV that was added to the “Indep. Variables” column will be evaluated. If the number of available IVs exceeds the “Maximum Number of Variables in a Model” value, however, VB₃ will show an error. If ‘Run all combinations’ is checked, an exhaustive search is performed, testing every model that can be constructed with the selected IVs, but does not evaluate models with more parameters than the “Maximum Number of Variables in a Model.” For example, if there are 24 available IVs and the maximum number of IVs is 8, the exhaustive routine will examine every 1-, 2-, 3-, 4-, 5-, 6-, 7- and 8-parameter model. VB₃ shows the total possible number of combinations below the “Model Settings” box. As the number of IVs rises, the number of possible models gets so large that the time needed to compute regression fits for each of them becomes unreasonable. We advise switching to the genetic algorithm in this case.

- The genetic algorithm (GA) button explores solution spaces too large to handle exhaustively. Genetic algorithms are loosely based on natural evolution in which individuals in a population reproduce and mutate (Fogel 1998). Individuals with high fitness (regression models that produce small residuals) are more likely to reproduce and pass their genes (IVs) to the next generation. The goal is to find a good solution without having to examine every possible option. The GA balances random and directed searching.

The figure shows two side-by-side screenshots of a software interface for model building. Both panels are titled 'Model Settings' and have tabs for 'Variable Selection' and 'Control Options'. The 'Number of Observations' is 101.

Left Panel (Manual Search):

- Evaluation Criteria:** Akaike Information Criterion (AIC) (dropdown menu).
- Maximum Number of Variables in a Model:** 10 (input field). Available: 17, Recommended: 11, Max: 17.
- Maximum VIF:** 5 (input field).
- Model Evaluation Thresholds:**
 - Decision Criterion (Horizontal): 235 (input field).
 - Regulatory Standard (Vertical): 235 (input field).
- Threshold Transform:**
 - ☒ None
 - ☐ Log10
 - ☐ Ln
 - ☐ Power
- Current US Regulatory Standards:**
 - E. coli, Freshwater: 235
 - Enterococci, Freshwater: 61
 - Enterococci, Saltwater: 104
- Buttons:** Manual, Genetic Algorithm, Run.
- Footer:** There are 109294 possible variable combinations.

Right Panel (Genetic Algorithm):

- Evaluation Criteria:** Akaike Information Criterion (AIC) (dropdown menu).
- Maximum Number of Variables in a Model:** 3 (input field). Available: 14, Recommended: 11, Max: 14.
- Maximum VIF:** 5 (input field).
- Model Evaluation Thresholds:**
 - Decision Criterion (Horizontal): 235 (input field).
 - Regulatory Standard (Vertical): 235 (input field).
- Threshold Transform:**
 - ☒ None
 - ☐ Log10
 - ☐ Ln
 - ☐ Power
- Current US Regulatory Standards:**
 - E. coli, Freshwater: 235
 - Enterococci, Freshwater: 61
 - Enterococci, Saltwater: 104
- Buttons:** Manual, Genetic Algorithm, Run.
- Genetic Algorithm Parameters:**
 - ☐ Set Seed Value: (input field)
 - Population Size: 100 (input field)
 - Number of Generations: 100 (input field)
 - Mutation Rate: 0.05 (input field)
 - Crossover Rate: 0.50 (input field)

Figure 29. Model building interface using a manual search (left panel) or the genetic algorithm (right panel)

Choosing between the exhaustive and the GA searches depends on the dataset, the computer's available random access memory (RAM), and time constraints. On a dataset of 101 observations and ten IVs, the exhaustive search was completed in approximately 6 seconds, using a Dell Precision T5400 (WinXP; dual Xeon 2.66 GHz processors; 4 GB RAM). Every additional IV doubles the number of models to examine and, thus, approximately doubles necessary computational time (Table 1).

Table 1. Relationship between the number of IVs, number of possible models, and time required to execute an exhaustive search using VB₃.

Exhaustive Search – Run All Combinations		
Number of IVs	Number of MLR models	Approximate Time Required to Generate and Filter Models (seconds)
10	1023	6
11	2047	13
12	4095	27

In contrast, running the GA with 10 IVs, using a population of 100 for 100 generations, took 90 seconds to complete ($90/6 = 15$ times slower than the exhaustive routine for this number of IVs); the GA with 12 IVs takes about the same amount of time - 90 seconds. So, as computational time of the exhaustive routine doubles every time an IV is added, the time required to run the GA stays approximately the same. As the number of IVs rises (here, to 14 or 15), the GA would be expected to save time and provide a solution very close to optimal.

An alternative modeling strategy with a large number of IVs would be to run the GA on the entire list of IVs initially, then switch to the exhaustive search on a subset of initial IVs – any IV that appears in one of the best ten models found by the GA. This two-step process is facilitated with the “IV Filter” list control (Figure 30).

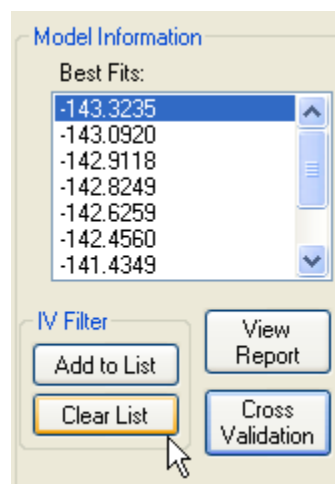


Figure 30. Using the IV filter to select a subset of variables from the best-fit models

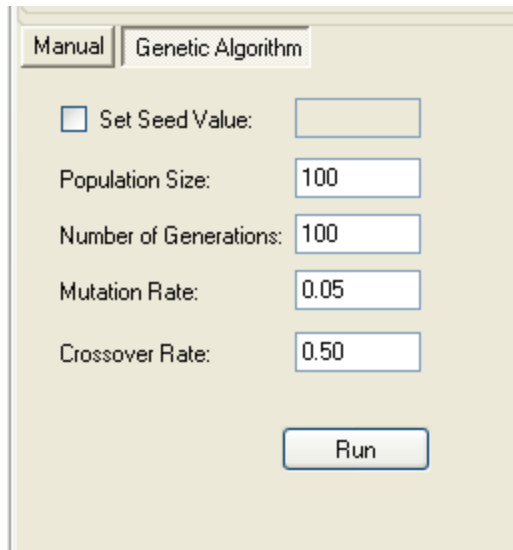
When the GA finishes and the 10 best models are shown in the Model Information box “Best Fits” window, clicking the “Clear List” button removes all IVs from the selection list. Select a model from the “Best Fits” list and click “Add to List” which adds any IVs in the selected model to the “Indep. Variable” list in the Model Settings box. After doing this for each of the ten best models, users will have a more manageable IV list and can run an exhaustive search to find the best combination of IVs. Regardless of the method chosen to build models, the “Best Fits” window shows the top ten models found, based on user-specified evaluation criterion.

7.4 Using the Genetic Algorithm

Users can set five parameters to adjust the performance (i.e., the amount of parameter space searched and the likelihood of finding a near-optimal solution) of the GA (Figure 31):

- Seed value: VB₃ uses an internal random number generator to produce random values. Setting the seed to a previously-used value will produce results identical to that earlier run, allowing the analysis to be reproduced by other parties. Changing the seed creates a new series of random values, possibly returning a different set of identified regression models.
- Population size: number of individuals in the population of each generation. A larger population broadens the search at each generation, but slows processing time.
- Number of generations: because individuals can reproduce and mutate once each generation, the question is how long to run the search. Fitness of every individual in the population is evaluated at the end of each generation.
- Mutation rate: chance each individual has of undergoing random mutation in each generation. The higher the mutation rate, the more random (less directed) the search of parameter space is.
- Crossover rate: probability that two selected individuals in the population will exchange genome parts. Exchanging genes creates new individuals in the population.

The best GA parameter values depend on the dataset being investigated, but typical values of the mutation rate are between 0.001 and 0.1 (0.1 and 10%) and typical values of the crossover rate are between 0.4 and 0.75 (40 and 75%). For most datasets, a population size and generation number of 100 are sufficient. Larger datasets may require increased numbers for optimal solutions. The user must invoke an experimental approach for changing these parameters and examining the results.



The screenshot shows a software window titled 'Genetic Algorithm' with a 'Manual' tab selected. It contains five input fields: 'Set Seed Value' (with an unchecked checkbox), 'Population Size' (100), 'Number of Generations' (100), 'Mutation Rate' (0.05), and 'Crossover Rate' (0.50). A 'Run' button is located at the bottom right of the window.

Figure 31. Genetic algorithm options within the modeling interface

7.5 Evaluating Model Output

After selecting a method to build models (GA or Exhaustive) and an evaluation criterion, click the “Run” button at the bottom of the “Control Options” sub-tab (Figure 28). Progress is displayed on the “Progress” sub-tab at the lower left of the MLR screen. Note that the “Run” button changes to “Cancel” if the user desires to terminate the process. Once model-building is completed, the ten best models are displayed in the “Best Fits” window (Figure 32). Selecting a model from the list results in:

- A list of selected IVs for the model, with associated regression coefficients and statistics displayed on the “Variable Statistics” sub-tab (Figure 33).
- A list of evaluation metrics for the selected model shown on the “Model Statistics” sub-tab (Figure 34).
- The “Results” sub-tab shows two data series - model fits and observations versus observations (Figure 35). Observations that are chronologically ordered are similar to a time series plot of the two data series, but ignore the possibility that time steps between data points are not equally spaced.
- The “Fitted vs Observed” sub-tab shows plots and tables based on fitted model values versus the observations (Figure 36).
- The “ROC Curves” sub-tab shows a plot of the Receiver Operating Characteristic curve of each “Best Fits” model (Figure 37), as well as a table showing the computed AUC (area-under-the-curve) for each ROC curve (see Section 7.7).
- The “View Report” generates a text report of model and variable statistics for the selected model.
- The “Residuals” sub-tab allows access to residual analysis functions in VB₃ (see Section 7.8).
- The “Prediction” tab appears at the top and bottom of the VB₃ screen, allowing users to proceed to the prediction component (Figure 32).

Note that selecting a different model from the “Best Fits” list will update the Variable and Model Statistics tables, as well as the information displayed on the “Results,” “Fitted vs Observed,” “ROC Curves,” and “Residuals” sub-tabs.

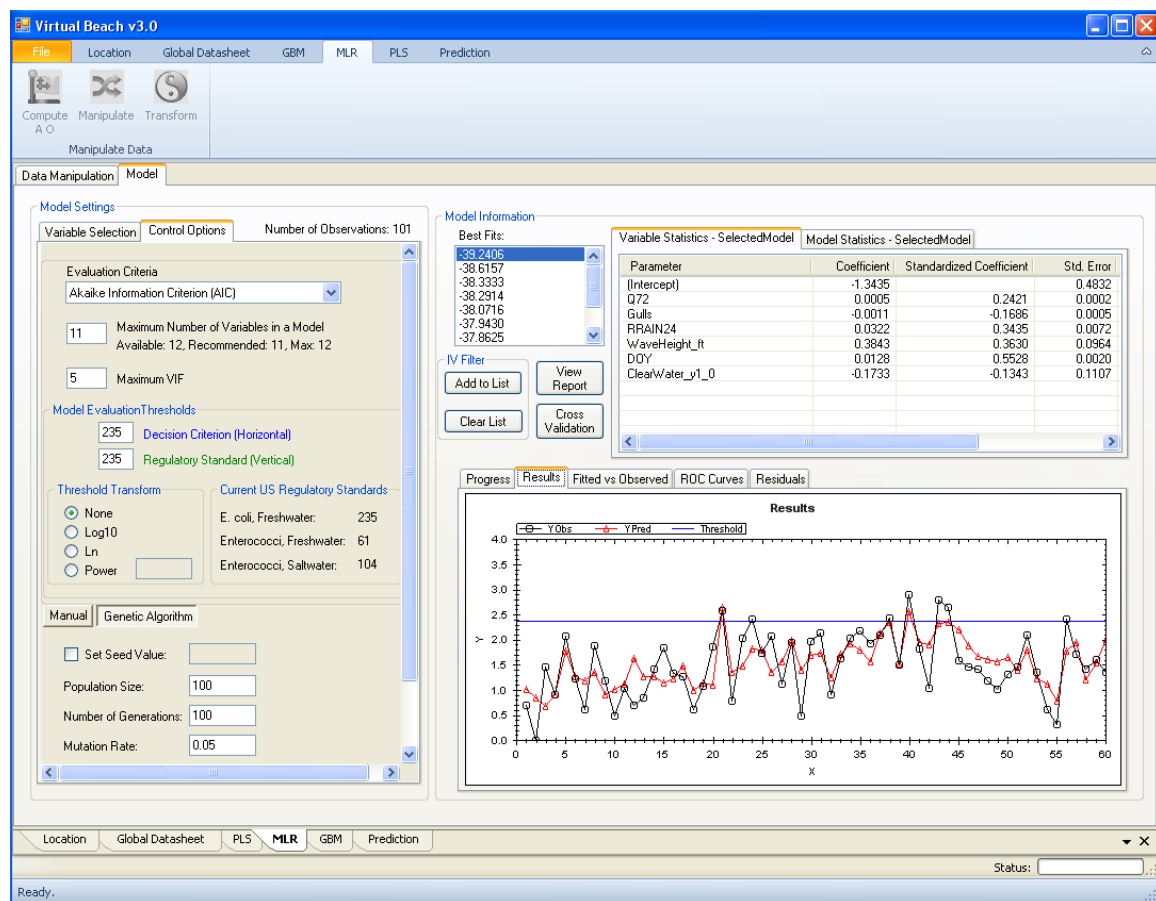


Figure 32. Modeling results after completion of a run using the genetic algorithm

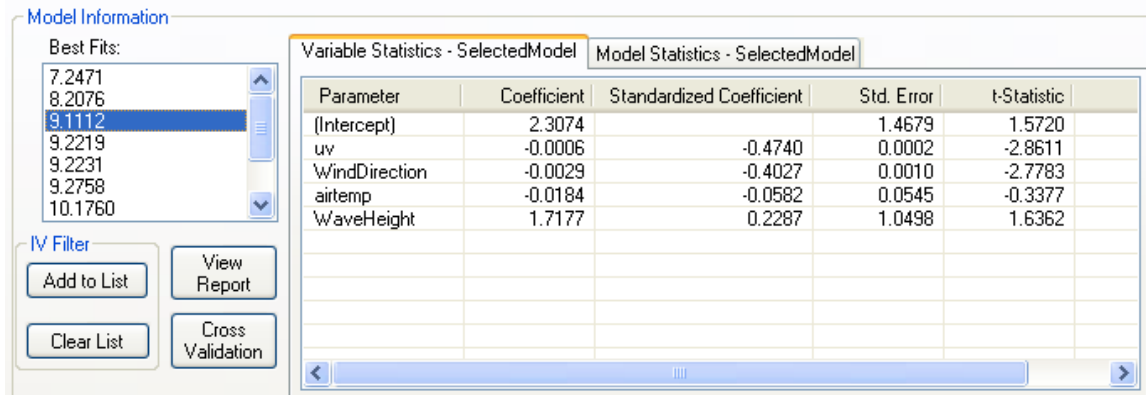


Figure 33. Modeling Interface showing variable statistics for the selected model

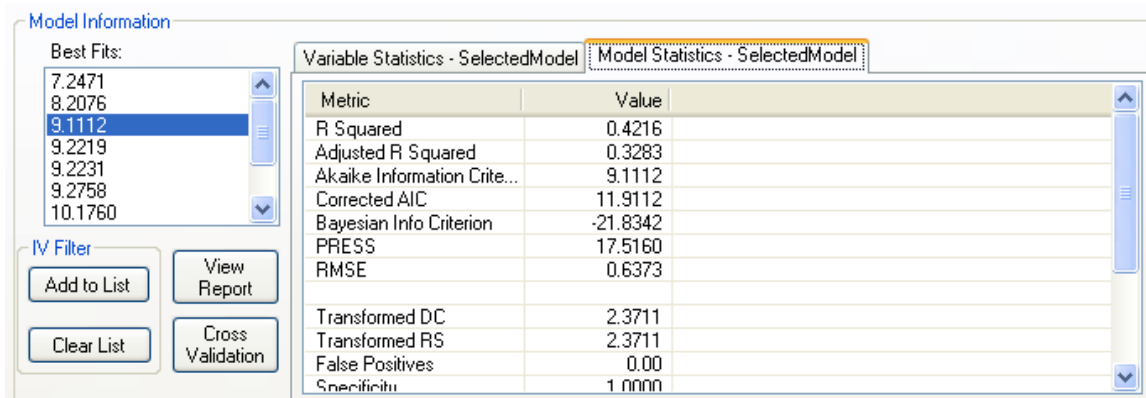


Figure 34. Modeling interface showing model evaluation metrics for the selected model

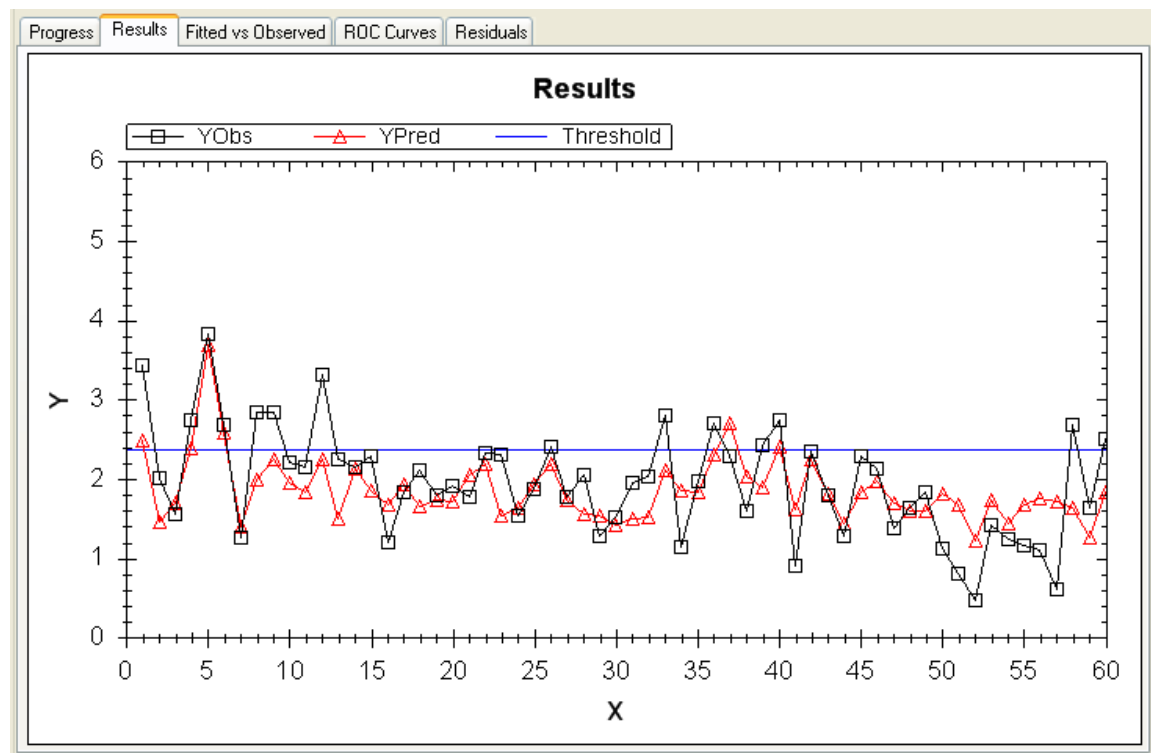


Figure 35. Modeling interface showing a time series plot for the selected model

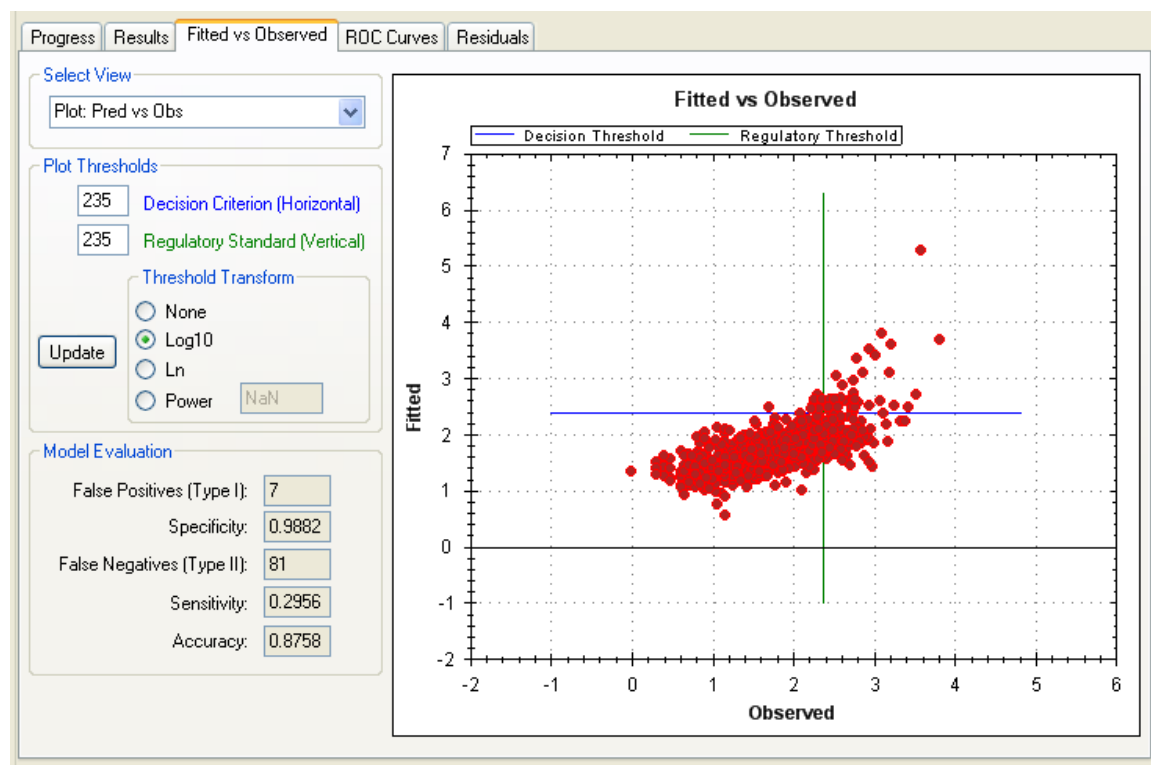


Figure 36. A scatter plot of fitted values versus observations of the selected model

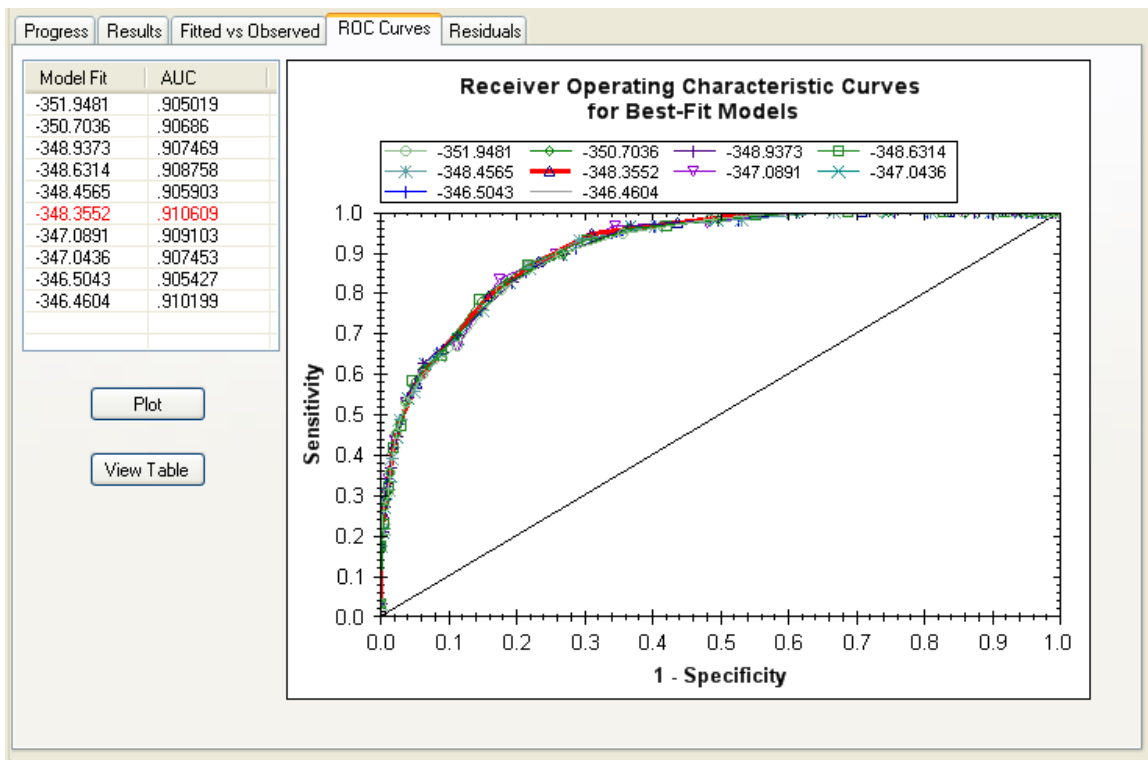


Figure 37. The ROC curves and AUC table for the model chosen from the “Best Fits” window

7.6 Viewing X-Y Scatter plots

On the MLR “Fitted vs Observed” and the MLR “Residuals” sub-tabs in the Model Information box, users are shown a graph to compare observations to fitted values from the model (Figure 36). Users can view different results from the pull-down tab from the “Select View” box:

- A plot of fitted values versus observations: “Pred vs. Obs”
- A table summarizing model errors (false negatives/false positives) as the decision criterion (DC) varies across the range of the response variable: “Error Table: DC as CFU”
- A plot of the percent of probability of exceedance (based on the current DC) versus observations: “% Exc vs. Obs”
- A table summarizing model errors as the percent of probability of exceedance is varied: “Error Table: DC as % Exc”

On the two plots, a right-click in the plot area shows a menu of functions for saving, copying, printing or manipulating the plot view. The plot area can be zoomed and un-zoomed: the left-click on the mouse drags an area for zooming in; the right-click selects “Un-Zoom” or “Set Scale to Default” to see the entire data set. To pan to a plot area not in view, hold the Shift key down and use the left mouse button to drag the view. Hovering the cursor over a data point shows the ID of the selected data point; if the information does not appear, right-click on the graph and select “Show Point Values.”

Regarding interpretation of these plots, the green (Regulatory Standard or RS) and blue (Decision Criterion or DC) lines allow model evaluation and provide information for choosing a DC for later predictive purposes. On the plots, false positives represent data points in the upper left quadrant of the graph, where the model fits/predictions exceed the DC, but observations are below the RS. In such cases, a beach advisory would be incorrectly issued based on the model's prediction, potentially leading to, for example, economic losses. False negatives (points in the lower right quadrant) represent a more serious scenario: model fits/predictions below the DC and observations that exceed the RS. In other words, swimming at the beach may have been allowed when it should have been prohibited due to elevated FIB concentrations.

A model that produces no false positives or false negatives would be an ideal decision tool, but this is often unattainable with real data. Examining the two tables from the "Fitted vs Observed" select view tab should allow users to set a robust DC, by using units of the actual response variable or a percentage probability of exceedance that minimizes both errors. In most cases, the RS is set by federal or state law and should not be adjusted by the user; however, users are free to adjust the DC to minimize false negatives and false positives.

7.7 ROC Curves

In addition to time series and scatter plots which show results for an individual model, users may also compare all the "Best Fits" models using the ROC Curves tab (Figure 37). A Receiver Operating Characteristic curve shows the true positive rate (sensitivity) plotted against its false positive rate (1 - specificity) for a model, as the Decision Criterion (DC) varies between its minimum and maximum predicted values. Models can then be compared using the area under their ROC curves (AUC). Models having the largest AUC values perform best over the entire decision space.

The model with the largest AUC appears in red text in the ROC tab's model list. A single ROC may be plotted by selecting a model in the list and clicking the "Plot" button. Multiple models can be selected in the usual Windows fashion with Shift-Click (select all items between the first and second selection) or Control-Click (select only the clicked items). The background cell color of models *not* selected for plot display will be gray after "Plot" button is clicked.

Clicking the "View Table" button will replace the ROC plot with a table showing false positives, false negatives, sensitivity, and specificity at every evaluated value of the Decision Criterion for a single model. Users need only click on a model in the list at the left of this table to see its results. The ROC plot returns to view after clicking the "View Plot" button.

AUC calculations are performed and curves are plotted when the "ROC Curve" sub-tab is selected. If this tab is active and new models are subsequently built, leaving this tab and returning will generate the new plots and AUC values.

7.8 Residual Analysis

Users may click the "Residuals" sub-tab to view information about the residuals of the selected model (Figure 38). There are three additional tabs on Residuals: "Residuals vs Fitted," "Fitted vs Observed," and "DFFITS/Cooks" (DF/C).

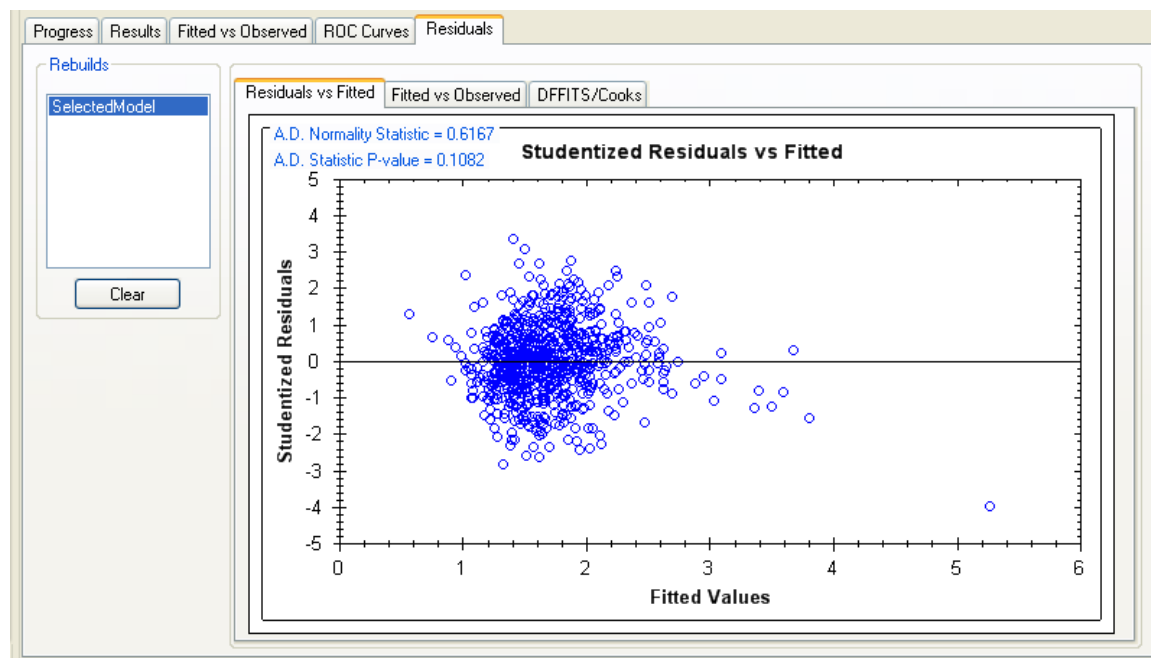


Figure 38. Information available on the Residuals sub-tab, including a plot of externally-studentized residuals versus model fits that shows results of the Anderson-Darling normality test

The Residuals vs Fitted tab shows a plot of externally-studentized residuals (Cook and Weisberg 1982) versus their fitted model values (Figure 38). In the upper-left corner of the plot, the Anderson-Darling normality statistic (Anderson and Darling 1952) is shown with its statistical significance (p-value). Linear regression assumes normally-distributed residuals, so that if this A-D normality test fails (i.e., the p-value is less than 0.05), the user can transform the response variable, transform some of the IVs, or delete high leverage observations, using the DF/C tab.

On the DF/C tab, observations are sorted by the largest (absolute value) measure in a table (Figure 39). At the lower left, radio buttons can be used to toggle between DFFITS and Cook's values, as well as change the view from a table of sorted values to a plot of the DF/C values versus the Record ID (Figure 40). Data points with very large DF/C values (i.e., lying outside the horizontal red boundaries on the plot) distort the estimates and standard deviations of the regression coefficients. They are essentially "outliers" and some thought to their removal from the dataset should be given.

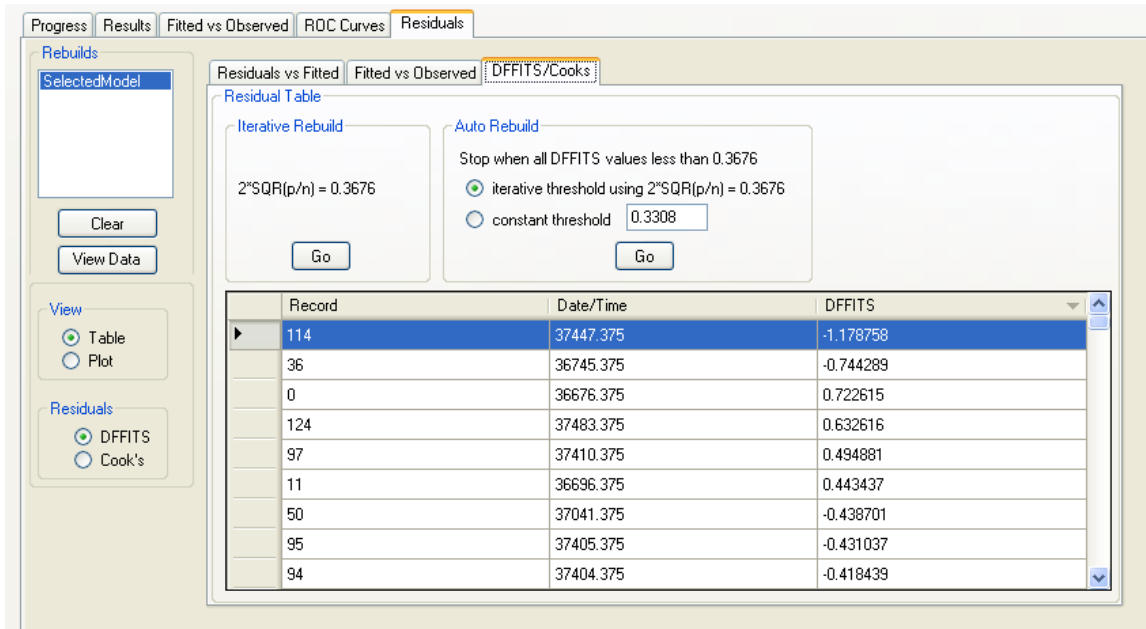


Figure 39. A table of the DFFITS scores of the residuals

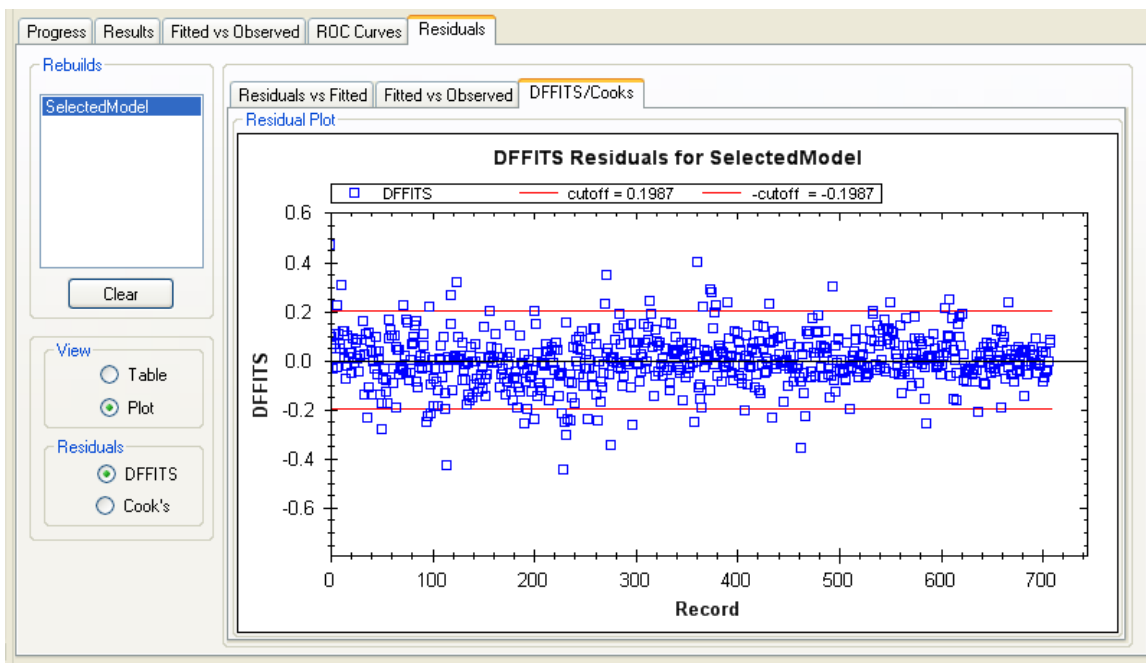


Figure 40. A plot of the DFFITS scores of the residuals

When the grid of DF/C values is visible, clicking the “Go” button in the Iterative Rebuild section removes the observation with the largest absolute value DF/C, re-fits the regression, and calculates new DF/C values for the remaining observations (Figure 41). This model is named Rebuild1 and added to the “Rebuilds” window at the top left of the sub-screen. Clicking the Iterative Rebuild “Go” button again produces a model called Rebuild2 which is calculated after removing the observation with the largest absolute

value DF/C remaining in the dataset. The user can continue to click “Go” and remove observations with the largest remaining DF/C, creating Rebuild3, Rebuild4, Rebuild5, etc. VB₃ will not allow users to delete any observations if 10 or fewer remain in the dataset.

Whenever a rebuild model is created by pressing the “Go” button, the information displayed in the Variable and Model Statistics tables, as well as the plots and information on the “Residuals” sub-tab, is automatically updated to reflect it, even if another model is highlighted in the “Best Fits” window. The user can select any model in the “Best Fits” window list, however, to view its associated data and plots.

The user has freedom to remove outliers while toggling between DF/C measures. For example, the first removal can be based on a DFFITS value, the next removal on a Cook’s Distance, the next two removals on DFFITS, etc. Users may clear models from the “Rebuilds” window by clicking the “Clear” button.

Rather than using Iterative Rebuild, there are two other choices under the “Auto Rebuild” box, both of which remove all observations above some threshold. The “iterative threshold” radio button bases removals on a threshold that is updated whenever an observation is deleted. For DFFITS, this threshold is $2 \cdot (p/n)^{0.5}$, where p is the number of IVs in the model and n is the current number of observations in the dataset. For Cook’s Distance, the threshold is $4/n$.

The screenshot shows a software window titled "Residual Table". It is divided into two main sections. The left section, titled "Iterative Rebuild", displays the formula $2 \cdot \text{SQR}(p/n) = 0.3676$ and a "Go" button. The right section, titled "Auto Rebuild", contains the instruction "Stop when all DFFITS values less than 0.3676". It features two radio buttons; the first, labeled "iterative threshold using $2 \cdot \text{SQR}(p/n) = 0.3676$ ", is selected. The second radio button is labeled "constant threshold" and is followed by an input field containing the value "0.3308". A "Go" button is also present at the bottom of the "Auto Rebuild" section.

Figure 41. DFFITS/Cook’s Distance controls for removing highly influential data points

When the “iterative threshold” radio button is invoked inside the “Auto Rebuild” box, VB₃ first checks if any DF/C values are above the threshold; if so, VB₃ removes the observation with the largest absolute DF/C and recalculates the regression model, the DF/C values, and the threshold because n has been reduced by 1. VB₃ then checks if any of these new DF/C values are above the recalculated threshold. If so, the process repeats. VB₃ continues until no remaining DF/C values exceed the current threshold or until half of the dataset has been removed, whichever comes first. For example, if a dataset has 100 observations, VB₃ will allow 50 to be removed before it breaks the Auto Rebuild removal loop. The user can then click the Auto Rebuild “Go” button again to remove another 25 observations of the remaining 50. In practice, one should not remove more than about 5% of the original dataset as outliers; removing more observations than this indicates a poor regression fit and warrants a different analytical technique. Indeed, under the assumption of normally distributed data, we expect 5% of the observations to fit relatively poorly.

The “constant threshold” radio button option differs from the “iterative threshold” only in that the threshold entered by the user to the input box remains the same regardless of how many observations are deleted. Updated DF/C values are still calculated after every removal. VB₃ will also stop this process if half the number of

starting observations has been deleted. There is an upper limit to the number that can be entered into the “constant threshold” input box (DFFITS = 3; Cook’s Distance = 16/n).

Upon completion of the Auto Rebuild process, multiple models may have been added to the “Rebuilds” window (Figure 42). For example, if 10 observations were removed, Rebuild1 through Rebuild10 will appear in that window.

When the user wants to move from the MLR tab to the Prediction tab, the model carried forward is the one highlighted blue in the “Best Fits” window or “Rebuilds” window. It is easy to confirm that the model selected will be carried forward by checking the numbers shown within the “Variable Statistics” and “Model Statistics” sub-tabs (Figures 33 and 34). Note that observations removed from the dataset using the “Residuals” sub-tab are not removed from the local dataset shown on the MLR “Data Manipulation” tab.

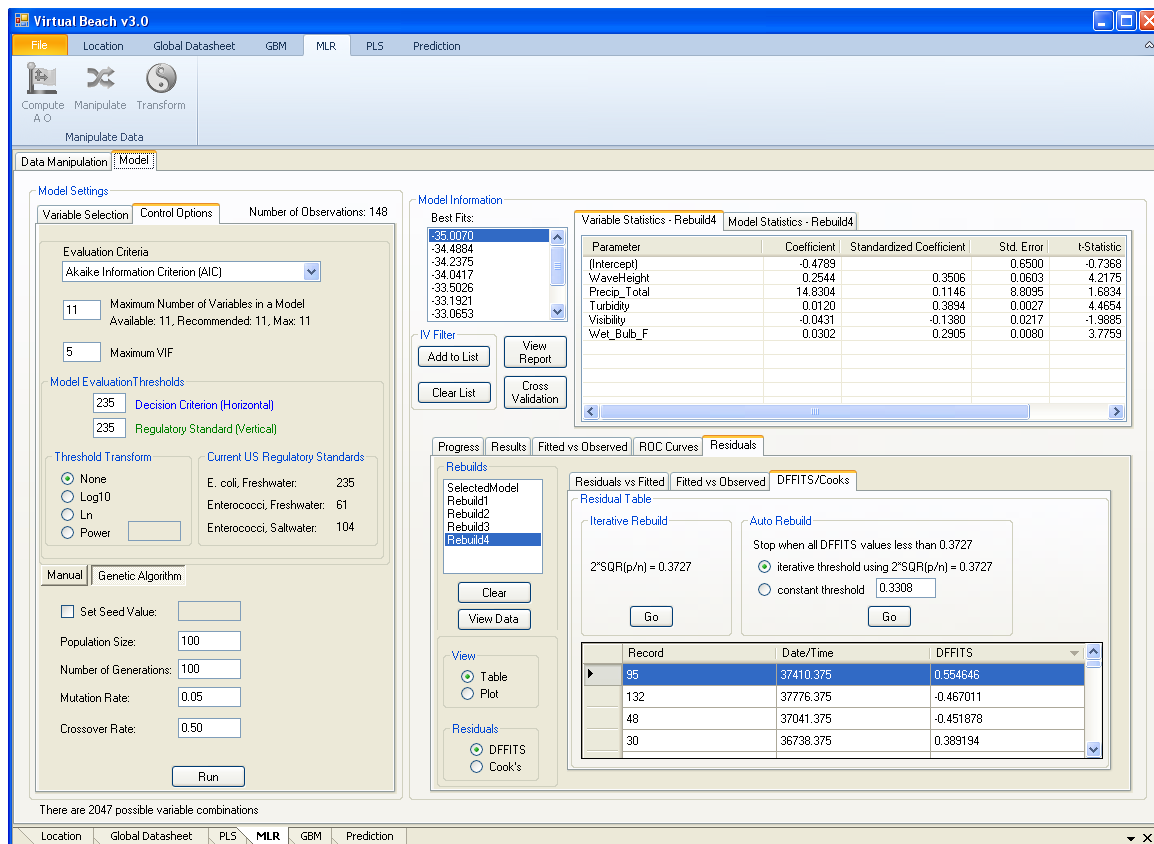


Figure 42. Residuals interface showing a list of rebuilt models resulting from observation deletions, and their associated statistics and residual plots

Viewing the Data Table

From the DFFITS/Cooks sub-tab, users can click the “View Data” button to display a history of observation removal for the selected model. From this window, users may export the dataset for external use or re-importation into VB₃ (Figure 43).

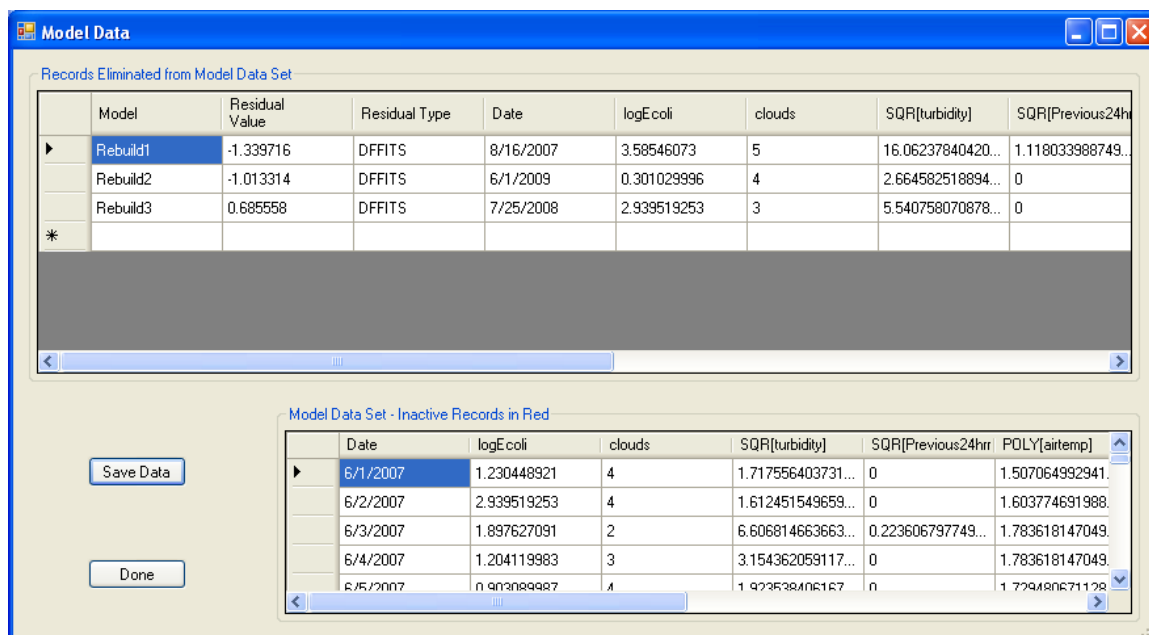


Figure 43. “View Data Table” window for examining the dataset after removal of influential data points

The “Fitted vs Observed” plot on the “Residuals” sub-tab is the same as that introduced in Section 7.6 (Figure 44). There are two plots and two tables to examine, along with controls to modify the Decision Criterion (blue horizontal line) and Regulatory Standard (green vertical line).

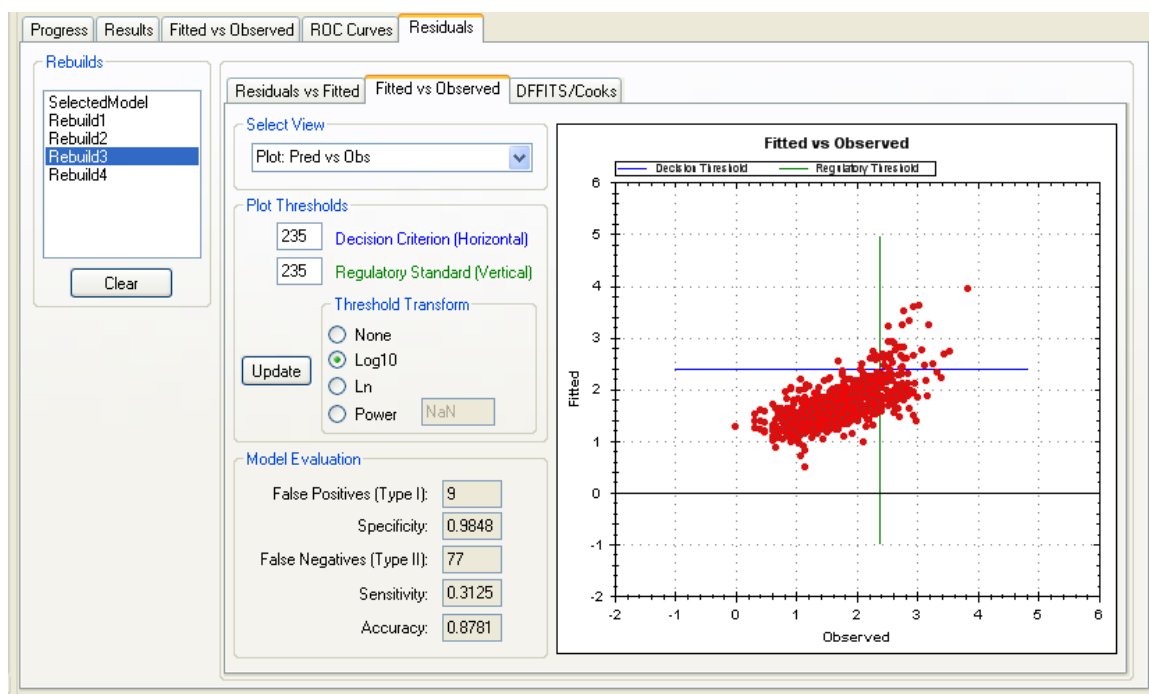


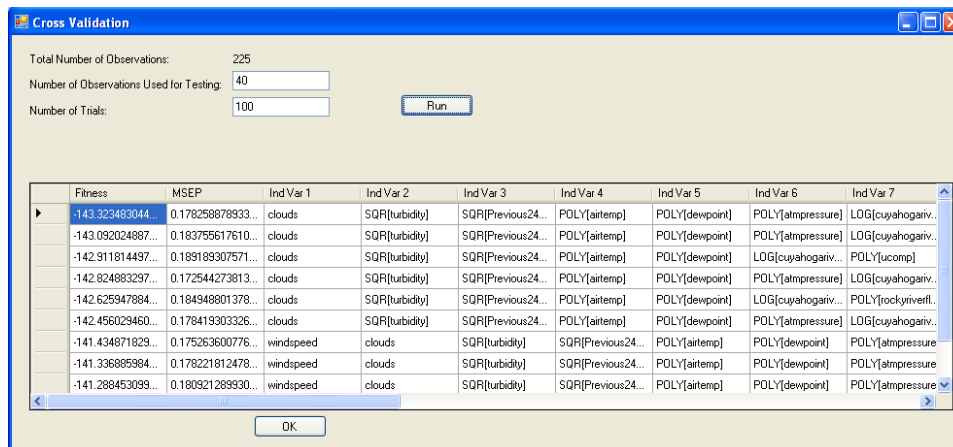
Figure 44. Fitted vs Observed plot on the Residual sub-tab with model evaluation threshold control and model evaluation statistics

7.9 Cross-Validation

Clicking the “Cross-Validation” button in the “Model Information” box brings up another window where the user can set two parameters: sample size for the *testing* data (N_E) and number of random samples (N_R) taken (Figure 45). When the “Run” button is clicked, a random sample of size N_E is taken from the modeling dataset and set aside. Each “Best Fits” model is then re-fit to the remaining *training* data. The IVs in each model stay the same, but the regression coefficients are adjusted to reflect the least-squares fit to the *training* data. The Mean Squared Error of Prediction (MSEP) is then calculated based on the N_E *testing* data points for each candidate model. This process is done N_R times. A table then appears to show the average MSEP values for each of the 10 “Best-Fit” models.

Cross-validation is useful for examining the predictive power of models, i.e., ability to make predictions for data they have not seen before. For users wishing to emphasize predictive ability of a potential model, cross-validation allows evaluation of which candidate model consistently makes the best predictions, i.e., has the lowest MSEP. Note that the PRESS statistic VB_3 provides as a model evaluation criterion is a cross-validation statistic with N_E , set to 1. The PRESS algorithm removes one observation at a time from the dataset, re-fits the model regression coefficients, and calculates the squared residual for the removed observation. It does this once for every observation in the dataset to compute the model’s PRESS value -- a somewhat cursory look at a model’s predictive potential.

We recommend that approximately 25% of the total number of observations be used for testing, and that at least 1000 trials be performed.



The screenshot shows a window titled "Cross Validation" with the following input fields: "Total Number of Observations:" set to 225, "Number of Observations Used for Testing:" set to 40, and "Number of Trials:" set to 100. A "Run" button is located to the right of these fields. Below the input fields is a table with 9 columns: Fitness, MSEP, Ind Var 1, Ind Var 2, Ind Var 3, Ind Var 4, Ind Var 5, Ind Var 6, and Ind Var 7. The table contains 10 rows of data, each representing a different model. The first row is highlighted with a mouse cursor.

Fitness	MSEP	Ind Var 1	Ind Var 2	Ind Var 3	Ind Var 4	Ind Var 5	Ind Var 6	Ind Var 7
-143.323483044...	0.178258878933...	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]	LOG[cuyahogariv...
-143.092024887...	0.183755617610...	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]	LOG[cuyahogariv...
-142.911814497...	0.189189307571...	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	LOG[cuyahogariv...	POLY[lucomp]
-142.824883297...	0.172544273813...	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]	LOG[cuyahogariv...
-142.625947884...	0.184948801378...	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	LOG[cuyahogariv...	POLY[rockyriverrf...
-142.456029460...	0.178419303326...	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]	LOG[cuyahogariv...
-141.434871829...	0.175263600776...	wind speed	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]
-141.336885984...	0.178221812478...	wind speed	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]
-141.288453099...	0.180921289930...	wind speed	clouds	SQR[turbidity]	SQR[Previous24...	POLY[airtemp]	POLY[dewpoint]	POLY[atmpressure]

Figure 45. Cross-validation results for each of the 10 best-fit models

7.10 Report Generation

A text report of modeling results can be generated, copied to the system clipboard, or saved to a text file using the “View Report” button in the middle of the MLR-Model screen. From here (Figure 46), users can view the report by selecting the

desired models and clicking the “Generate Report for Selected Models” button. The report contains descriptive statistics for each model variable and model evaluation statistics. Any number of best-fit models can be selected for reporting.

A recommended approach to saving the information in an external application is to copy the report to the clipboard with the “CopyToClipboard” button and paste it into an application such as Microsoft Word or WordPad. NotePad or other simple text editors will also work, but column formats will likely be lost, making the report difficult to interpret.

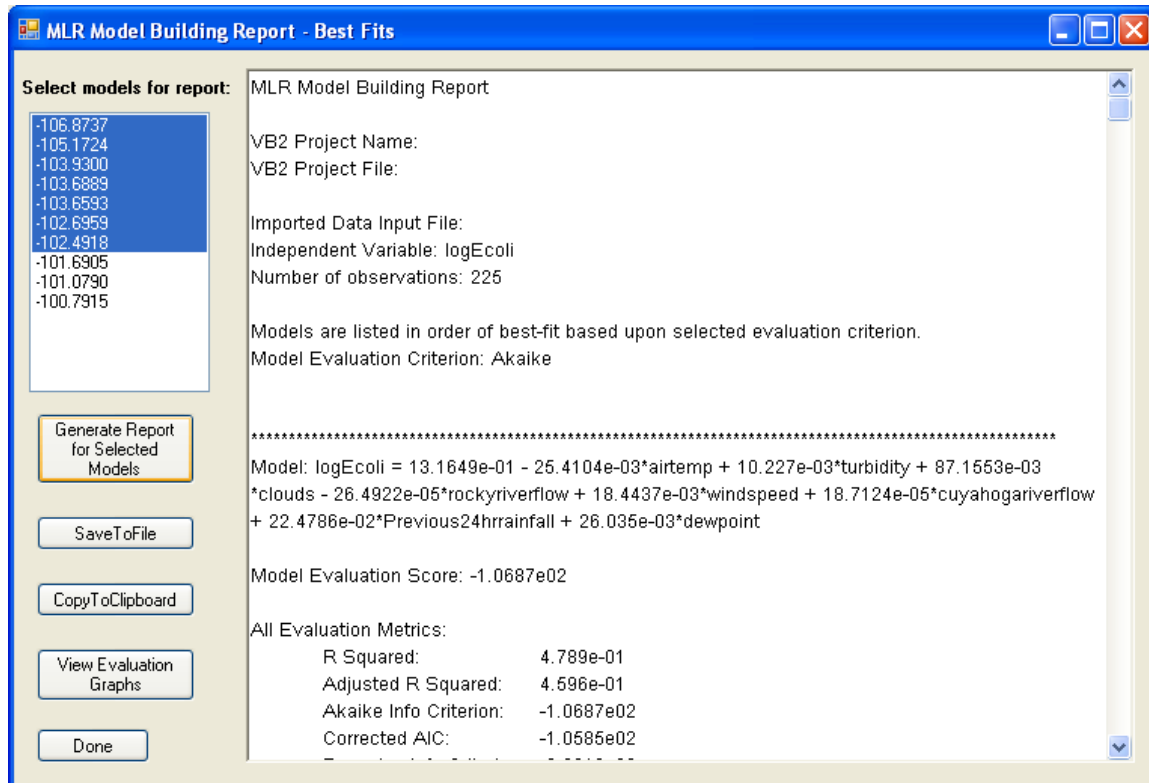


Figure 46. A text report generated on the modeling results

Comparative bar graphs can be displayed (Figure 47) to view evaluation criteria for all top models by left-clicking and dragging the mouse to highlight selection and clicking the “View Evaluation Graphs” button (Figure 46). Hover the mouse over any plot to display the model evaluation criteria at the very top of the screen. Moving the mouse over a bar on a plot will show that model’s coefficients under the title at the top, and a label will appear with that same information. Note that evaluation criteria graphs are initially scaled to emphasize differences between model scores although those differences may, in fact, be quite small on an absolute scale (Figure 48). With the cursor over any graph, right-click the mouse and select “Set Scale to Default” to view the un-scaled graph.

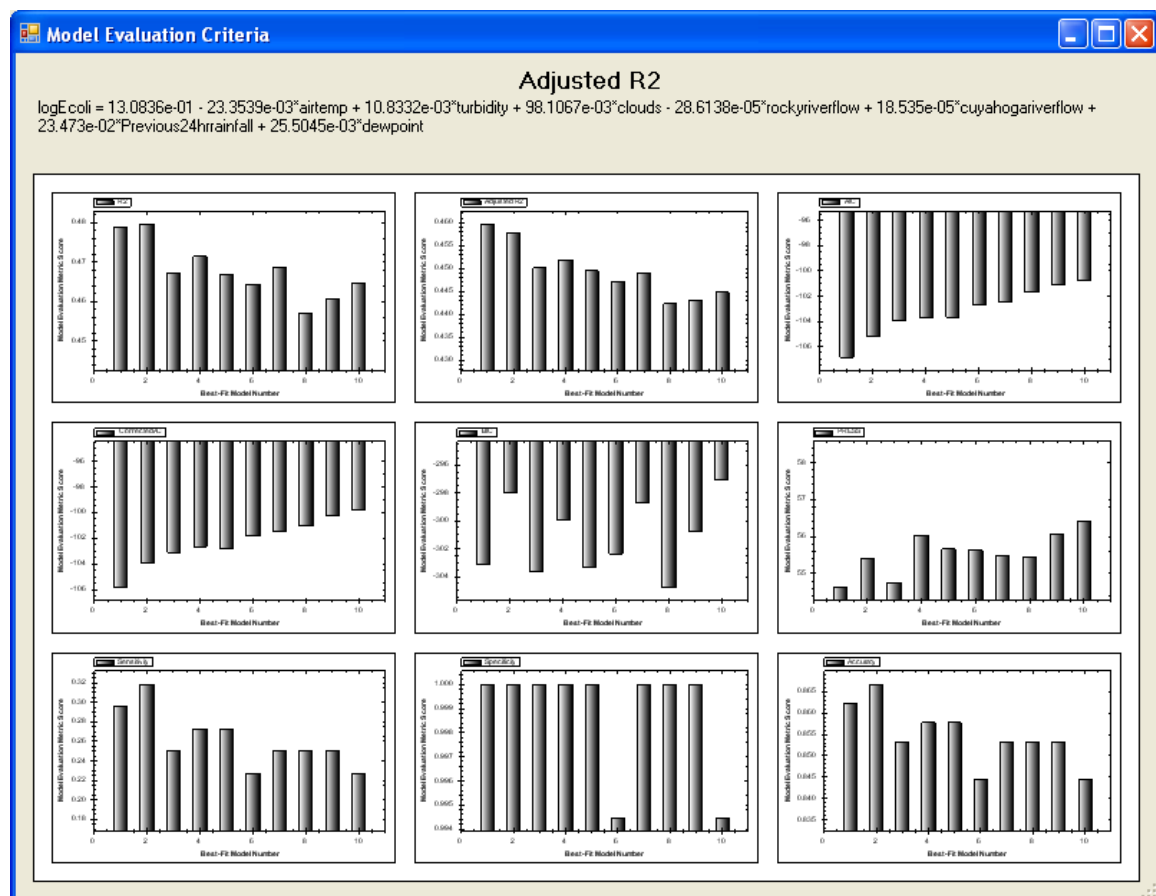


Figure 47. Plots of various model evaluation metrics for the 10 best-fit models

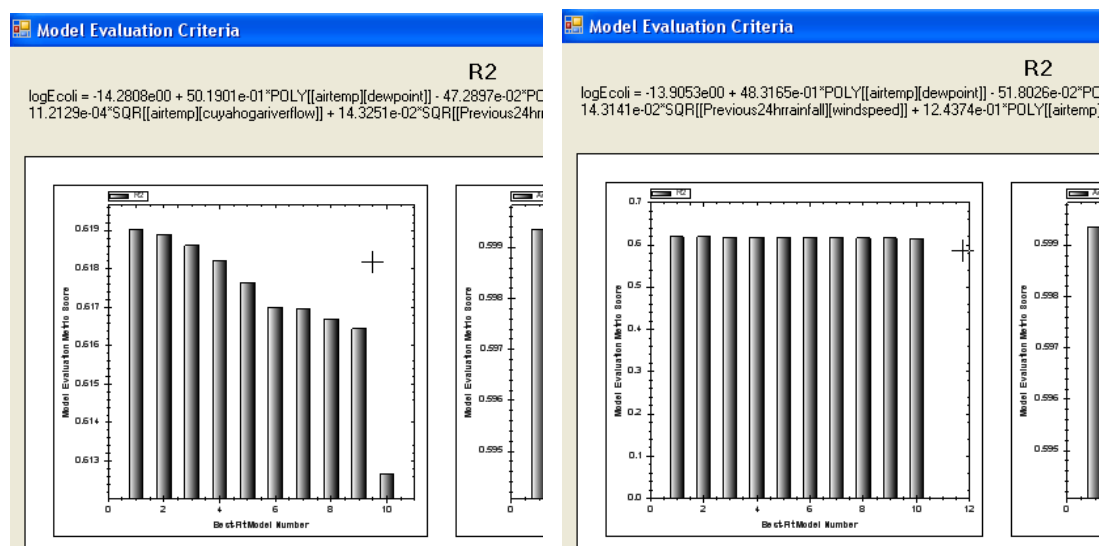


Figure 48. Scaled versus un-scaled views of selected model evaluation criteria

8. PARTIAL LEAST SQUARES

Partial Least Squares (PLS) regression minimizes a problem that can arise in MLR modeling: over-fitting in the presence of correlated predictors. To over-fit is to match past data more closely than the real-world process being modeled. MLR is prone to over-fitting because it makes the closest possible linear match to past data, even at the cost of accuracy in predicting future observations.

As opposed to requiring the MLR user to be vigilant and proactive, PLS regression (Brooks et al. 2013) inherently accounts for collinearity to suppress over-fitting, and ranks the IVs by their influence in variable selection. Using PLS regression, the user can include all available IVs in the model and let the algorithm sort out which IVs are most useful, simplifying the sometimes laborious processes of variable selection and comparing interactions.

A key feature of PLS (and GBM) modeling is the use of cross-validation to assess real-world prediction accuracy. Model selection and threshold setting (section 8.4) are done with reference to the true positive, true negative, false positive and false negative counts, which are calculated by 5-fold cross validation. This means that the data are split randomly and evenly into five subsets and five models are built to predict exceedances on each of the five subsets. For each of these models, the subset predicted is left out of model building, so the counts reflect prediction of novel observations, not accuracy in fitting past observations. Greater detail about the PLS modeling method is available in Brooks et al., 2013 and Hastie et al. 2009.

8.1 Data Manipulation

The MLR, PLS, and GBM modules all have “Data Manipulation” sub-tabs (Figure 49). When the user first clicks on the PLS tab from the Global Datasheet, data in the PLS Data Manipulation sub-tab is identical to data on the Global Datasheet. From the PLS data tab, the user can change the “local” data to suit the PLS analysis. The local datasheet has all of the functionality of the Global Datasheet discussed in Section 6. Changing local data has no effect on the Global Datasheet; however, going back to the Global Datasheet and making changes will overwrite local datasheets on each of the modeling tabs.

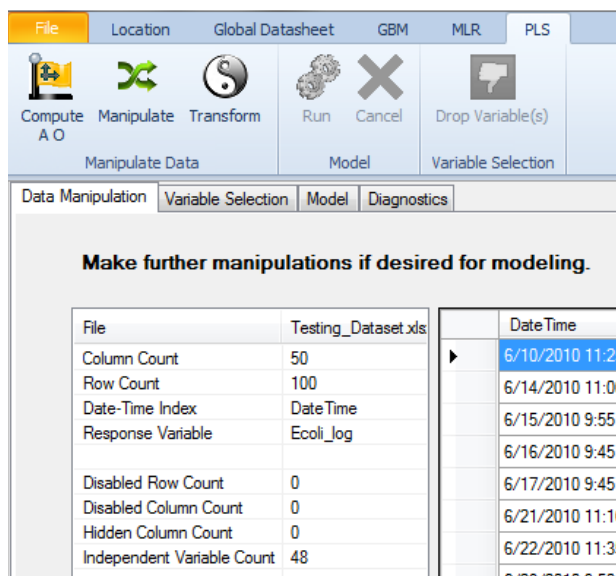


Figure 49. Data Manipulation: the first sub-tab on each of the method tabs

8.2 Selecting Variables for Model Building

The “Variable Selection” tab is where IVs for model development are chosen (Figure 50). Users may select all or a subset of the IVs for consideration in the model. All eligible IVs are listed in the “Available Variables” window (left column). Any IVs that users wish to include in the model must then be moved to the “Independent Variables” window by highlighting the IV and clicking the “>” key. Any number of IVs can be added or removed from this list. Once the desired IVs have been selected, click the “Model” sub-tab.

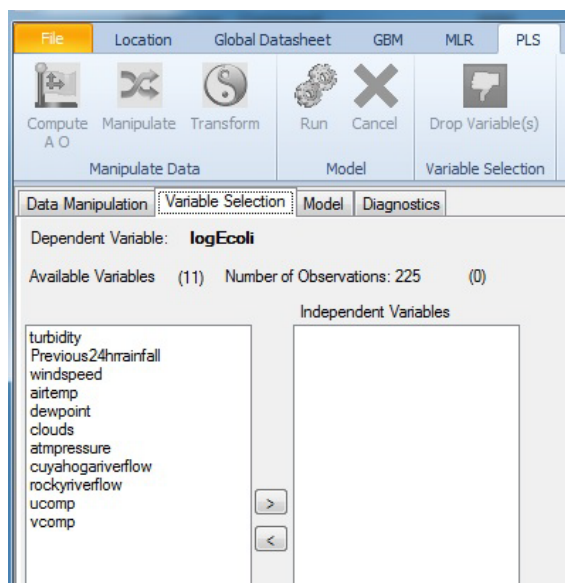


Figure 50. Selecting variables for PLS processing within the modeling module

8.3 The Regulatory Standard

To build a PLS ribbon, observations must be defined as exceedances or non-exceedances; PLS and GBM models will not run if the dataset has no exceedances. This is done by setting the Regulatory Standard (RS) at the top of the “Model” tab and then specifying, using the radio buttons, units to enter into the RS. The default RS is the USEPA’s federal standard for *E. coli* in freshwater, 235 CFU per 100 mL. Because these are raw units of measurement, the radio button transformation choice should be set to “Value.” However, users may be thinking of bacteria concentrations in logarithmic units; if so, the RS is 2.371 ($= \log_{10}(235)$). To communicate this to VB₃, enter 2.371 in the “Regulatory Standard” box and click the “Log10 (value)” radio button (Figure 51).

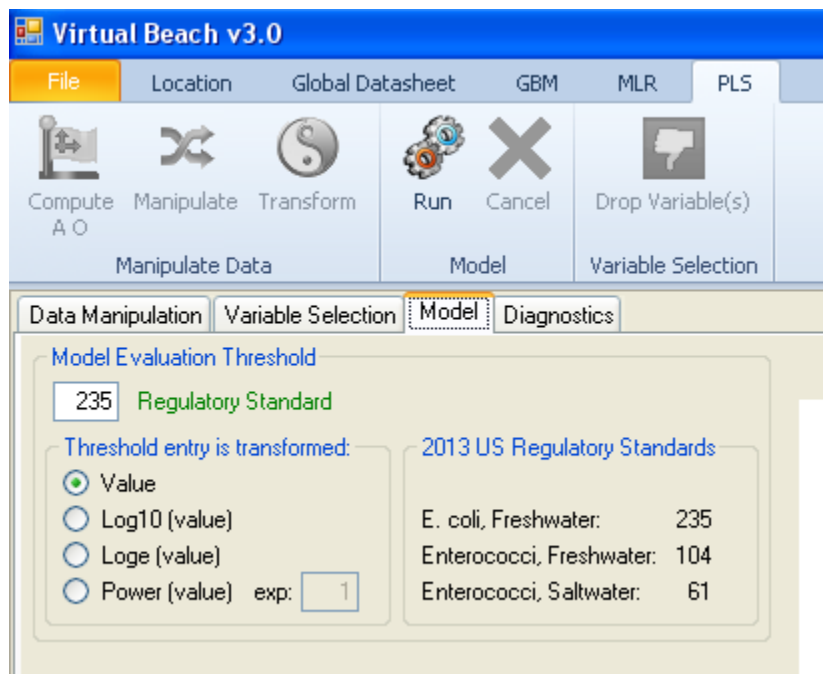


Figure 51. Setting the Regulatory Standard and running models for PLS

8.4 Modeling Control Options

Clicking “Run” on the PLS Model tab (Figure 50) will start model development. Depending on the number of IVs and observations in the dataset, it could take several minutes to compute the model; GBM typically takes longer than PLS. When modeling is finished, results are displayed (Figure 52). The “Drop Variable(s)” option on the PLS ribbon is now available and a Decision Criterion (DC) for the model can be chosen.

Dropping Unimportant Variables

The “Model Summary” window (left side of Figure 52) lists IVs in descending order of influence. For a PLS model, a variable’s influence is its model coefficient multiplied by

its standard deviation. The influence measurements are then adjusted to sum to one. The larger the influence of a variable (global sensitivity), the more its variation drives the response. Low-influence variables (highlighted in red text) can be dropped from the model by clicking on the variable's name in the list, then clicking the “Drop Variable(s)” button on the ribbon. If any variables are dropped at this stage, the model must be rebuilt by clicking the “Run” button.

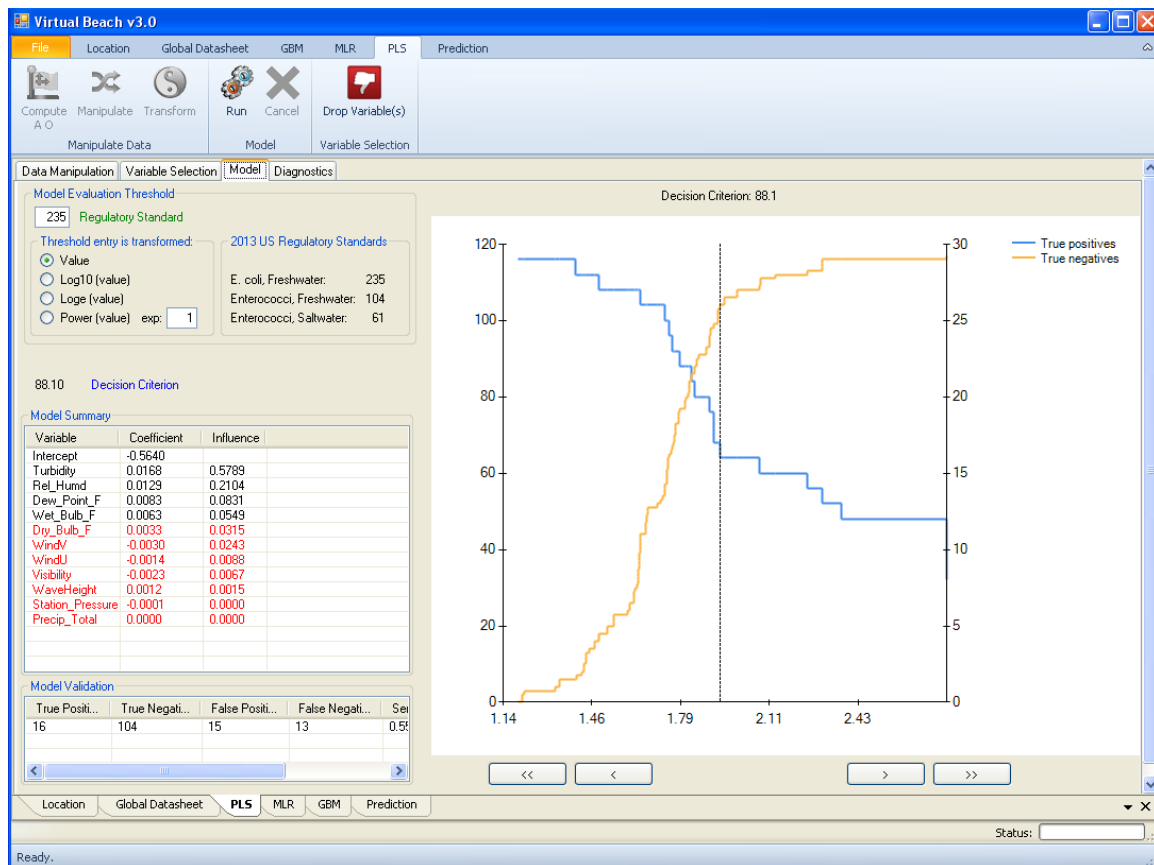


Figure 52. Results after completion of a PLS model run

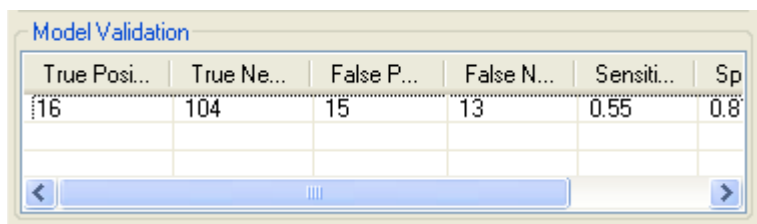
Setting the Decision Threshold

Once the user has selected a model, the Decision Criterion (DC) is chosen (see Section 7.2 for a description of the DC). The graph on the right side of the “Model” tab is used for this purpose (Figure 52). To understand the plot, consider that a lower DC will correctly identify more exceedances of the RS threshold, but also produce more “false positives” by flagging predicted values when the actual water quality is below the RS. Raising the DC has the opposite tradeoff: reducing false positives at the expense of identifying fewer true exceedances.

The blue line on the graph indicates true positives and the yellow line indicates true negatives. Current model performance is indicated by the following:

- The vertical dotted line indicates the current location of the DC. The arrow buttons at the bottom are used to lower/raise the threshold by a small ($<$, $>$) or large ($<<$, $>>$) amount.
- The “Model Validation” window (Figure 53) indicates the number of true positives, true negatives, false positives, false negatives, sensitivity, specificity, and total accuracy of the model using the current DC. These results are based on cross-validation of the training data. These numbers will change after a short computational delay as the DC is moved. Care should be taken when comparing PLS/GBM model performance (in terms of false/true positives/negatives) with MLR models. MLR model performance is based on fitted, not cross-validated results. Cross-validation results are commonly thought to be more realistic in how well the model will do in future predictions, while fitted values better indicate how well the model fits previously-collected data. Cross-validation results are generated by developing models with partial data sets and making predictions for data left out of model development. For example, 5-fold cross validation would result in five different sets of IV coefficients for a single model by using 4/5 of the data to develop each set of IV coefficients, then predicting the remaining 1/5 of the data using those coefficients. MLR, on the other hand, uses all available data points to fit the model coefficients and then predicts the same data points. Look at cross-validated performance of MLR models using the “Cross-Validation” button described in Section 7.9.
- The current numeric value of the DC is shown above the “Model Summary” window (Figure 52).

The user can change the DC, drop variables, and re-run the model to fine-tune it. After the model and DC have been chosen, the user can advance to the Prediction tab (Section 10) to make predictions with the most recently computed model.



True Posi...	True Ne...	False P...	False N...	Sensiti...	Sp
16	104	15	13	0.55	0.8

Figure 53. Summary of PLS model performance metrics

8.5 Diagnostics

There are four plots are offered on the “Diagnostics” sub-tab (Figure 54):

- The Time Series plot (upper left) displays predicted and observed values of the response variable. This is a time-series plot if the ID values for the observations are

chronologically-ordered dates/times. If they are not, then this plot will look rather messy and strange, and be of little interest to the user.

- The Residuals vs. Fitted plot (upper right) shows the externally-studentized residuals versus model-fitted values. The externally-studentized residuals are a way to flag influential outliers. A common benchmark for a data point with undue influence on the regression model is an externally-studentized residual (absolute value) greater than 3.0.
- The Residuals vs. Observed plot (lower left) graphs the externally-studentized residuals against the observations.
- The Fitted vs. Observed plot (lower right) shows observations versus model fits and depicts the RS (green horizontal line) and current DC (blue vertical line).

Note that the fitted values plotted here are not cross-validated fits; rather they are the model fits based on all the data. For this reason, model performance in this plot (numbers of true negatives/positives) will likely be better than the model performance metrics given in the “Model Validation” window on the “Model” tab. A perfect model will fall along the 1:1 line. The more scatter in this plot, the worse the model fit.

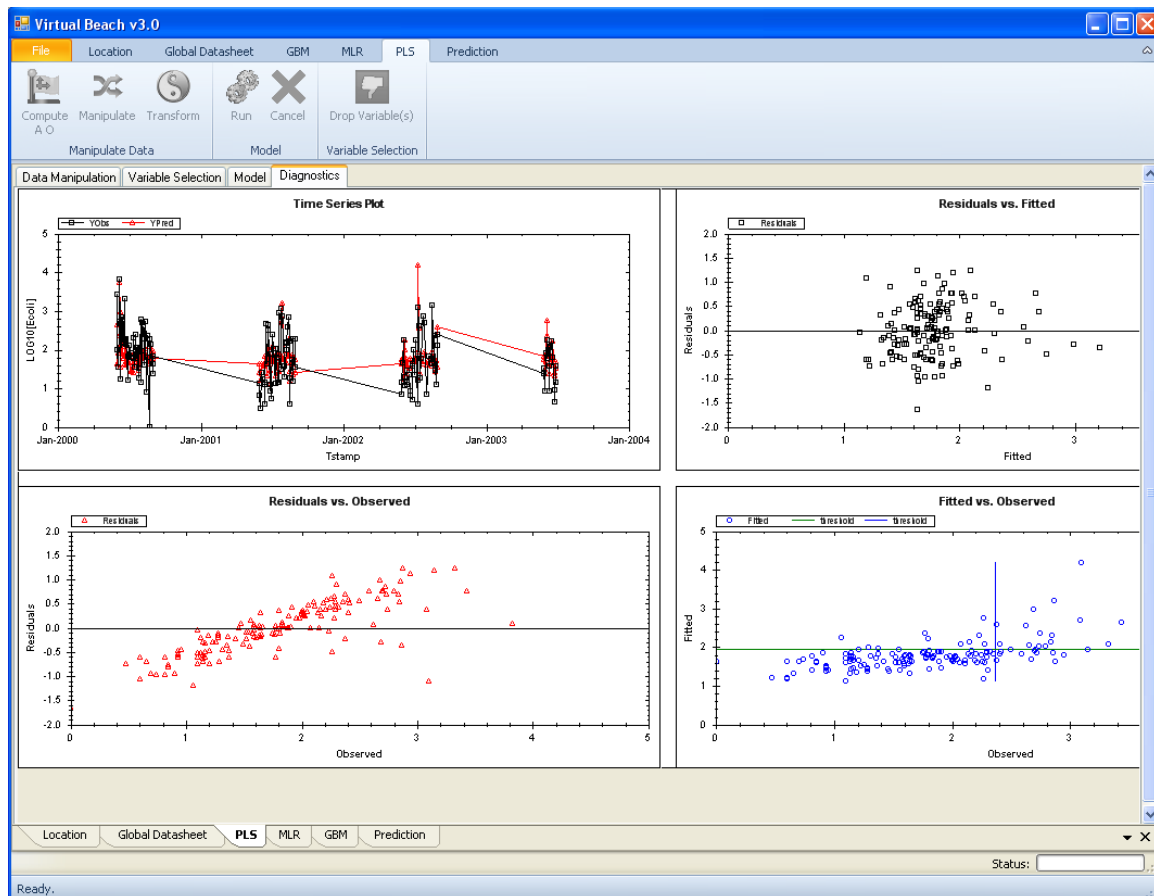


Figure 54. PLS Diagnostic plots to help evaluate model fit and influential outliers

9. GRADIENT BOOSTING MACHINE

The Gradient Boosting Machine (GBM) is a machine learning method that uses decision/regression trees instead of linear equations (Friedman, 2001). A decision/regression tree is a set of binary decision rules. For example, “if turbidity is less than 15 NTU, go down the right branch, otherwise go left.” A “node” is the end of any branch and designates a continuous or categorical predictive value for the response variable. The innovative aspect of GBM is that it doesn’t build a single, complex tree: it builds a hierarchical set of many simple trees, with each subsequent tree fit to the remaining residual error in the data after previous trees have all been fit. The default maximum number of trees in VB₃ is 10,000. Each tree is determined using a random 50% of the residual values from the dataset. This, along with the fact that it sensibly weights the data to learn more about the most difficult-to-predict cases, means GBM can make accurate predictions for new observations without over-fitting the training data. A “shrinkage” parameter (= 0.01) also limits the influence of each successive tree to an ever-increasing degree.

While each tree is a simple structure, the long, linear combination of regression trees is very complicated. A negative aspect of a GBM model is that the model cannot easily be inspected graphically or expressed mathematically – it’s something of a “black box.” But what it lacks in interpretability and transparency can often be made up in terms of prediction accuracy. Another noted aspect of GBM, unlike MLR and PLS, is that it handles non-linear relationships between the response and IVs without having to transform the IVs.

In a GBM, variable selection (identifying and dropping unimportant IV’s from the model) is less important, compared to MLR. Even so, the “Drop Variables” button (Figure 55) performs as described in Section 8.4. For a GBM model, an IV’s influence is the percentage of branches across all of the decision trees involving that variable, i.e., the most important variables are those that are most often used to create the branches.

For the GBM analysis, VB₃ implements the “gbm” package in R. Details of the algorithm are provided in Hastie (2009). Despite very different underlying mathematics, the GBM modeling interface in VB₃ is almost identical to the PLS interface (Section 8).

A key feature of GBM and PLS modeling is the use of cross-validation to assess real-world prediction accuracy. Model selection and threshold setting (Section 8.4) are done with reference to true positive, true negative, false positive and false negative counts which are calculated by 5-fold cross-validation. This means the data are split randomly and evenly into five sections and five models are built to predict exceedances on each of the five sections. For each, the section being predicted is left out of model building, so the counts reflect prediction of novel observations, not accuracy in fitting to past observations.

9.1 Data Manipulation

The MLR, PLS, and GBM modules all have “Data Manipulation” sub-tabs (Figure 55). When the user first clicks the GBM tab, the data in the GBM Data Manipulation sub-tab are identical to data on the Global Datasheet. From here, the user can change the “local” data to suit the GBM analysis. The local datasheet has all of the functionality of the Global Datasheet discussed in Section 6. Changing the local data has no effect on the Global Datasheet; however, going back to the Global Datasheet and making changes will overwrite the local datasheets on each of the modeling tabs.

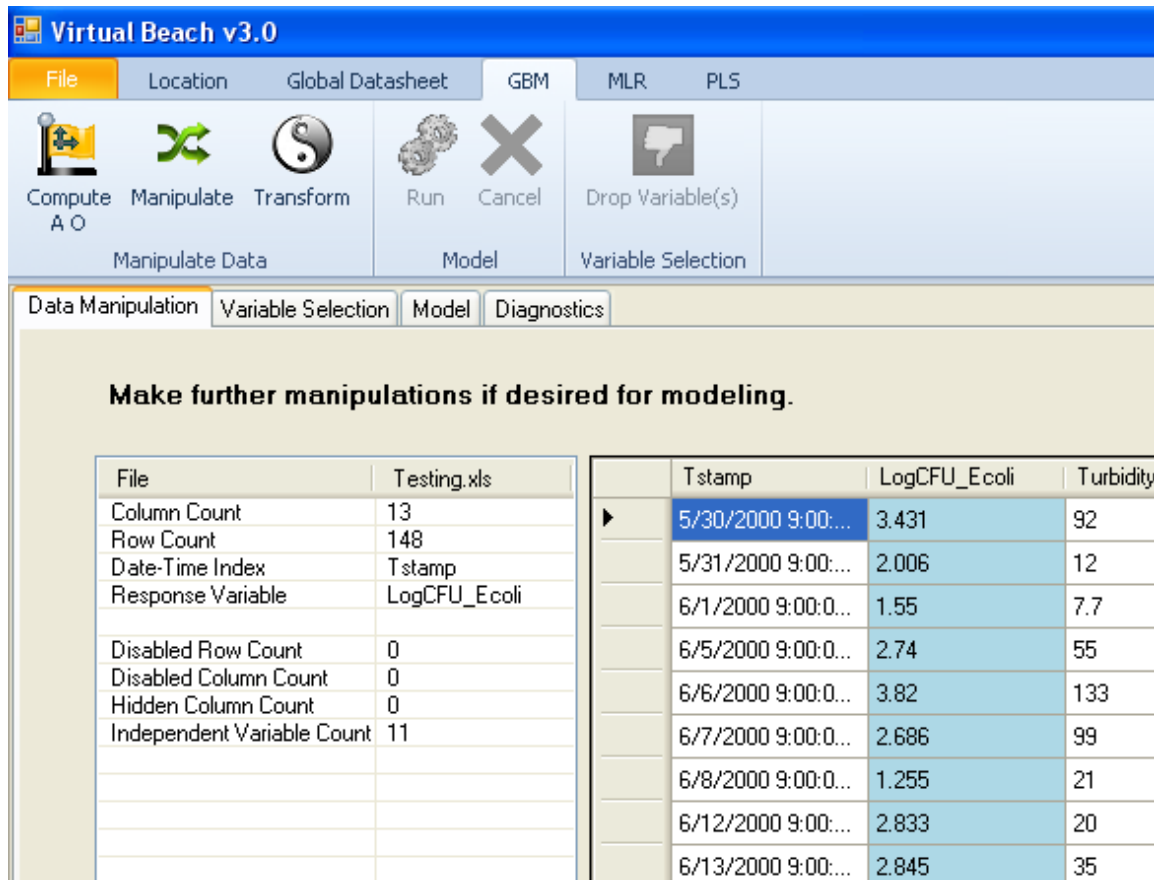


Figure 55. Data Manipulation: the first sub-tab on each of the Method tabs

9.2 Selecting Variables for Model Building

The “Variable Selection” sub-tab is where IVs for model development are chosen (Figure 56). Users may select all or a subset of IVs for the model. All eligible IVs are listed in the “Available Variables” window (left column). Any IVs that users wish to include in the model must be moved to the “Independent Variables” window by highlighting the IV and clicking the “>” key. Any number of IVs can be added or removed from this list. Once the desired IVs have been selected, click the “Model” sub-tab.

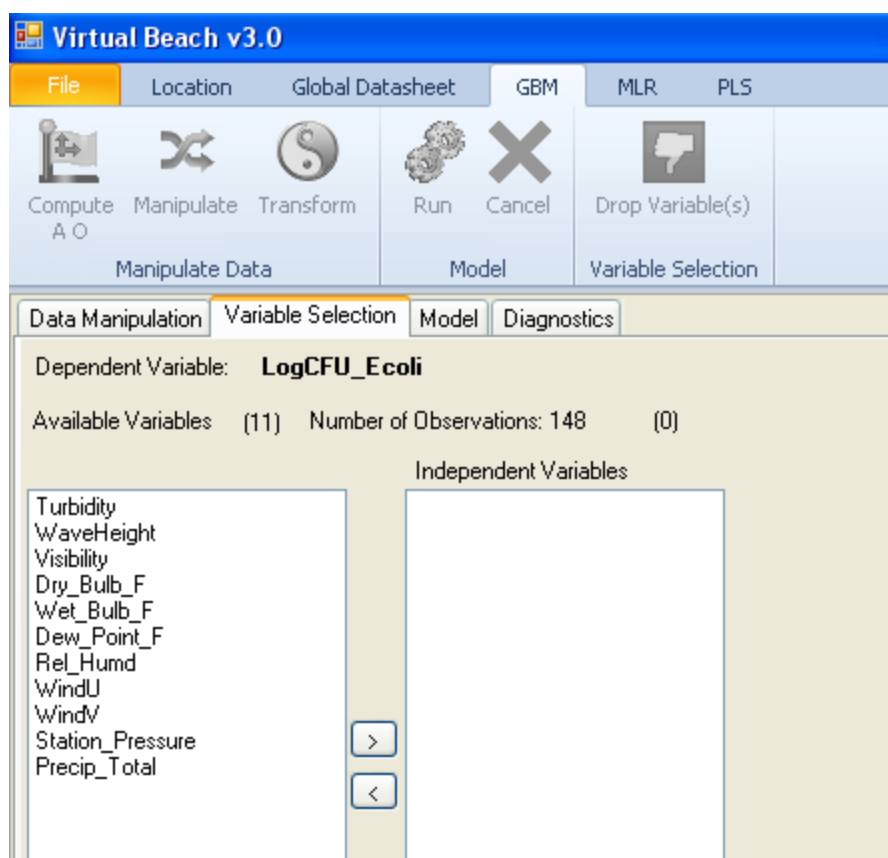


Figure 56. Selecting variables for GBM processing within the modeling module

9.3 The Regulatory Standard

To build a GBM model, observations must be defined as exceedances or non-exceedances; GBM and PLS models will not run if the dataset has no exceedances. This is done by setting the Regulatory Standard (RS) at the top of the “Model” sub-tab (Figure 57) and then specifying with the radio buttons units to enter for the RS. The default RS is the USEPA’s federal standard for *E. coli* in freshwater, 235 CFU per 100 mL. Because these are the raw units of measurement, the radio button transformation choice should be set to “Value.” When thinking of bacteria concentrations in logarithmic units, think of the RS as 2.371 ($= \log_{10}(235)$). To communicate this to VB₃, enter 2.371 in the “Regulatory Standard” box and click the “Log10 (value)” radio button (Figure 57).

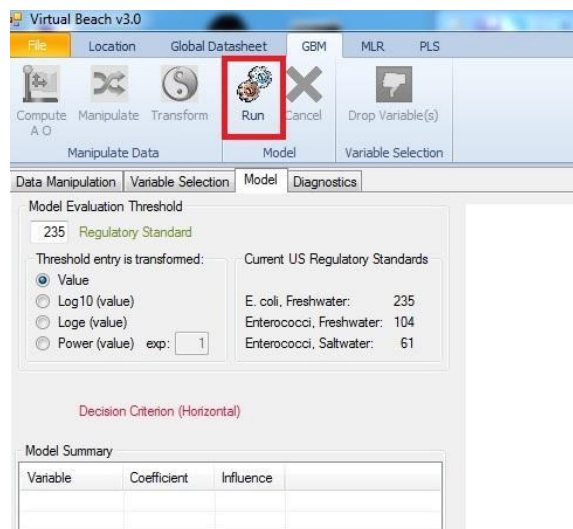


Figure 57. Setting the Regulatory Standard and running models for GBM

9.4 Modeling Control Options

Clicking the “Run” button on the GBM ribbon (Figure 57) will start model development. Depending on the number of IVs and observations in the dataset, it could take several minutes to compute the model; GBM typically takes longer than PLS. When modeling is finished, results are displayed (Figure 58). The “Drop Variable(s)” option is now available on the ribbon and a Decision Criterion (DC) for the model can be chosen.

Dropping Unimportant Variables

The “Model Summary” window (left side of Figure 58) lists the IVs in descending order of influence. For a GBM model, a variable’s influence is the percentage of the model’s total error reduction contributed by branches on the given variable. The larger the influence of a variable (global sensitivity), the more its variation drives the response. Low-influence variables (highlighted in red text) can be dropped from the model by clicking on the variable’s name in the list, then clicking the “Drop Variable(s)” button. If any variables are dropped at this stage, the model must be rebuilt by clicking the “Run” button.

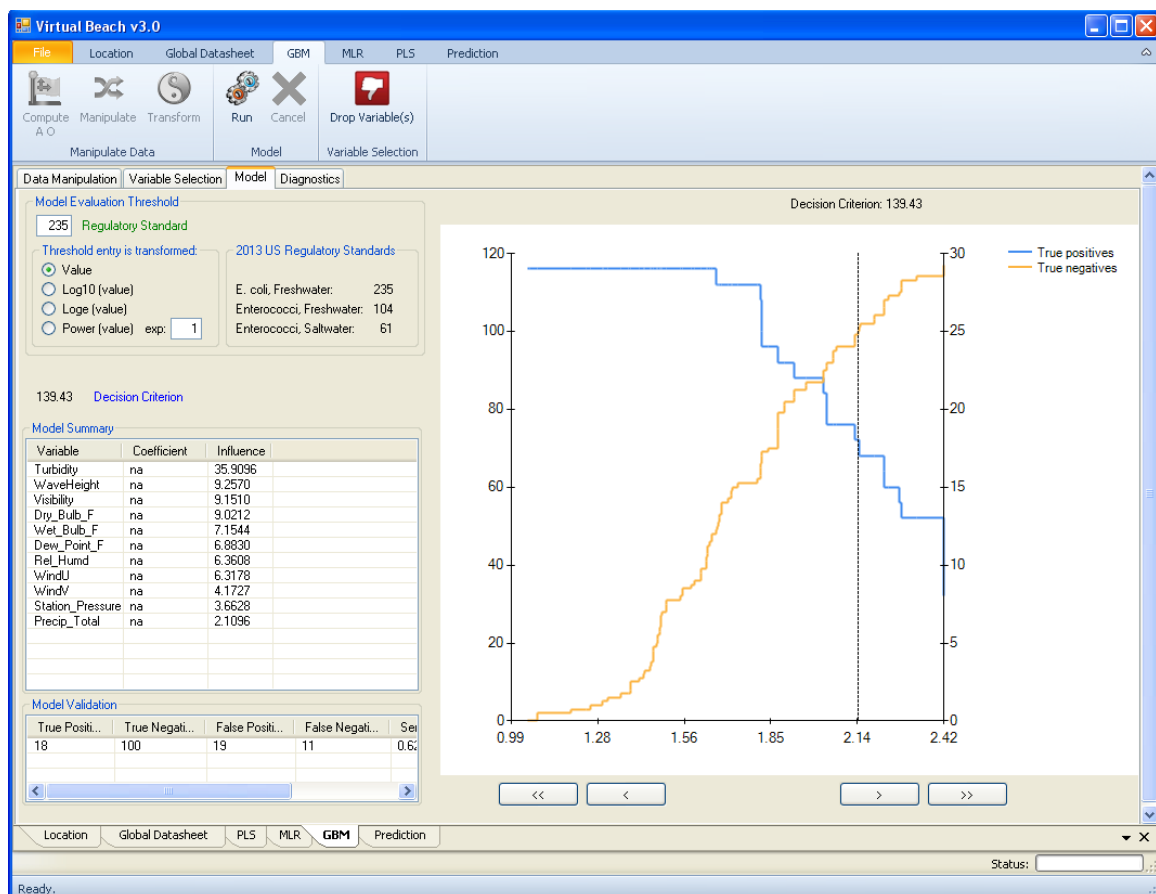


Figure 58. Results after completion of a GBM model run

Setting the Decision Threshold

Once the user selects a model, the Decision Criterion (DC) can be chosen (see Section 7.2 for a description of the DC). The graph on the right side of the Model tab is used for this purpose (Figure 58). To understand the plot, consider that a lower DC correctly identifies more exceedances of the RS threshold, but also produces more “false positives” by flagging predicted values when the actual water quality is below the RS. Raising the DC has the opposite tradeoff: reducing false positives at the expense of indentifying fewer true exceedances.

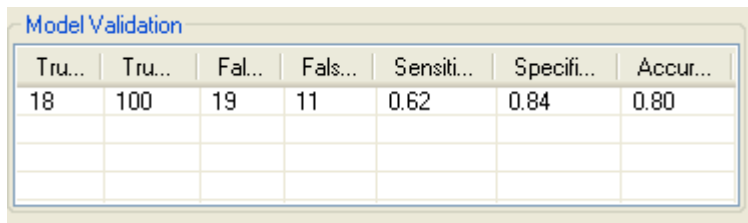
The blue line on the graph indicates true positives and the yellow line indicates true negatives. Current model performance is indicated by the following:

- The vertical dotted line indicates the current location of the DC. The arrow buttons at the bottom are used to lower/raise the threshold by a small (< , >) or large (<< , >>) amount.
- The “Model Validation” window (Figure 59) indicates the number of true positives, true negatives, false positives, false negatives, sensitivity, specificity, and total accuracy of the model using the current DC. These results are based on cross-validation of the training data. These numbers will change after a short

computational delay as the DC is moved. Comparing GBM (and PLS) model performance, in terms of false/true positives/negatives, with MLR models must be done carefully. MLR model performance is based on fitted results, not cross-validated results. Cross-validation results are commonly thought to indicate more realistically how well the model will do in future predictions and fitted values better indicate how well the model fits previously collected data. Cross-validation results are generated by developing models with partial data sets and subsequently making predictions for data that were left out. For example, 5-fold cross validation would result in 5 different sets of IV coefficients for a single model by using 4/5 of the data to develop each set of IV coefficients, and then predicting the remaining 1/5 of the data using those coefficients. MLR, on the other hand, uses all available data points to fit the model coefficients and then predicts the same data points. Look at cross-validated performance of MLR models using the “Cross-Validation” button described in Section 7.9.

- The current numeric value of the DC is shown above the “Model Summary” window (Figure 58).

The user can change the DC, drop variables and re-run the model to fine-tune it. After the model and DC have been chosen, the user can advance to the Prediction tab (Section 10) to make predictions with the most recently computed model.



Tru...	Tru...	Fal...	Fals...	Sensiti...	Specifi...	Accur...
18	100	19	11	0.62	0.84	0.80

Figure 59. Summary of GBM model performance metrics

9.5 Diagnostics

There are four plots offered on the “Diagnostics” sub-tab (Figure 60):

- The Time Series plot (upper left) displays predicted and observed values of the response variable over time if the ID values for the observations are dates/times.
- The Residuals vs. Fitted plot (upper right) shows the externally-studentized residuals versus model-fitted values. The externally-studentized residuals are a way to flag influential outliers. A common benchmark for a data point with undue influence on the regression model is an externally-studentized residual (absolute value) greater than 3.0. Certain patterns seen in this residual plot can indicate the need for a transformation of the response variable or model IVs. Refer to Meyers (1990) for details.

- The Residuals vs. Observed plot (lower left) graphs the externally-studentized residuals against the observations.
- The Fitted vs. Observed plot (lower right) shows observations versus model fits and depicts the RS (green horizontal line) and current DC (blue vertical line).

Note that the fitted values plotted here are not cross-validated fits; rather they are the model fits based on all the data. For this reason, model performance in this plot (numbers of true negatives/positives) will likely be better than the model performance metrics given in the Model Validation table on the Model tab. A perfect model will fall along the 1:1 line. The more scatter in this plot, the worse the model fit.

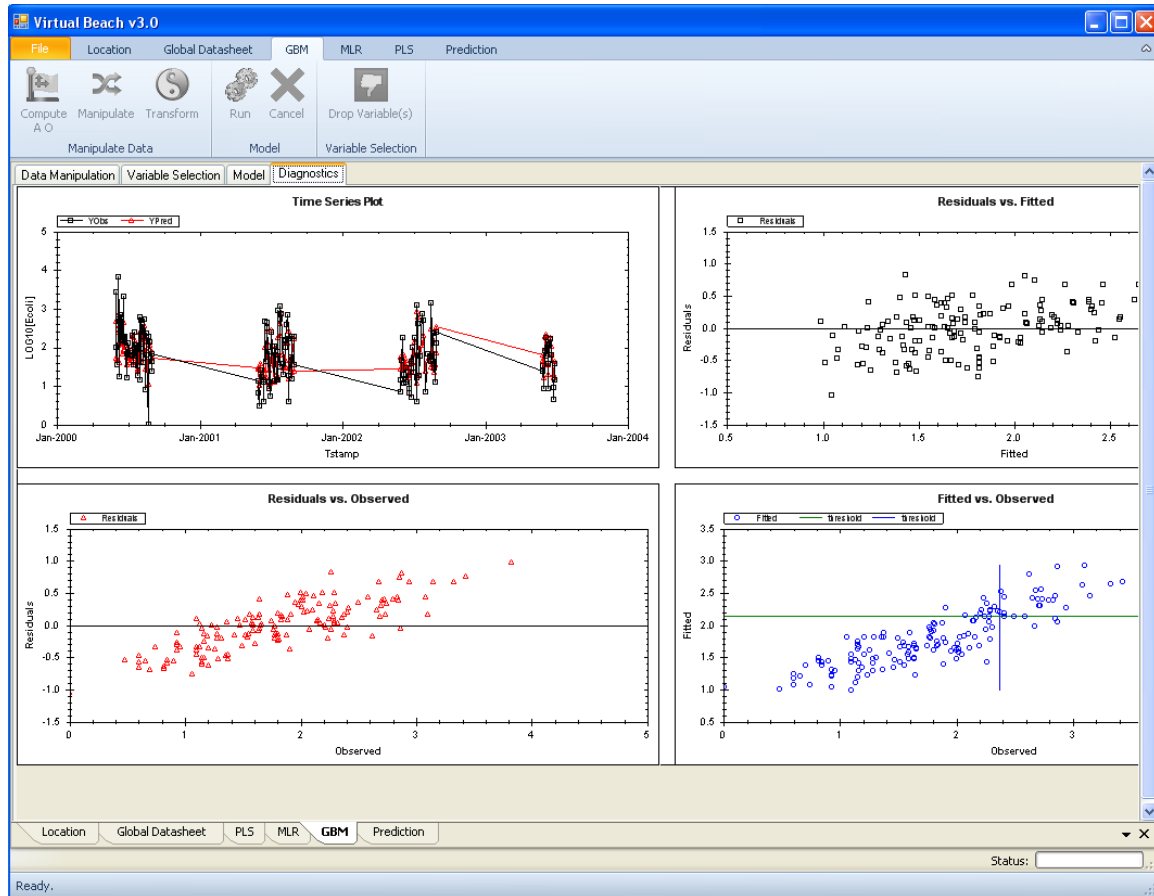


Figure 60. GBM diagnostic plots to help evaluate model fit and influential outliers

10. PREDICTION

VB₃'s Prediction interface allows users to select a model from the PLS, GBM, or MLR tabs and make predictions with it, but the prediction tab is hidden until a model is chosen.

10.1 Model Statement

At the top left of the Prediction tab is the “Available Models” window. Depending on how many statistical methods were performed on the data, the user could see “MLR,” “PLS,” and/or “GBM” in this area. Once a model is chosen, an expression with the IVs and coefficients in that model is shown in the “Model” window to the right (Figure 61).

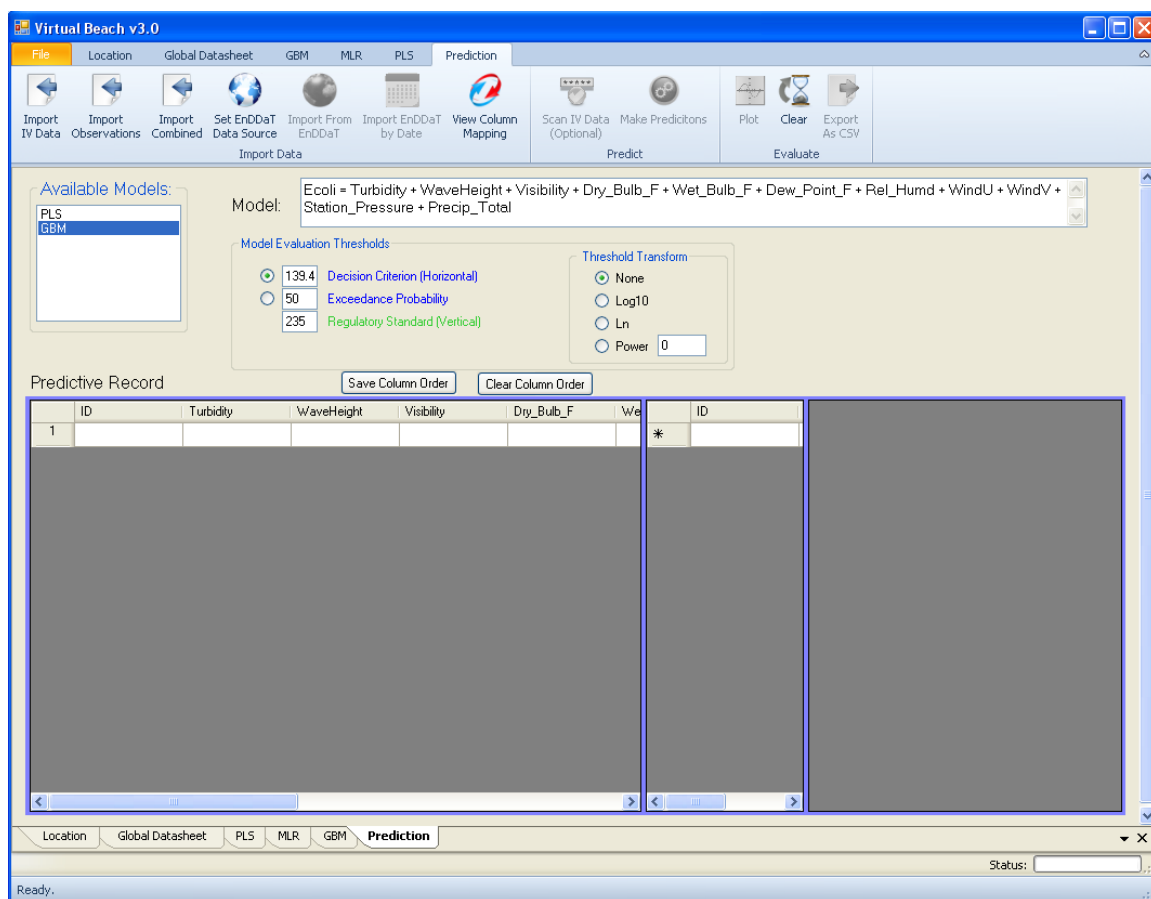


Figure 61. The VB₃ Prediction interface

10.2 Model Evaluation Thresholds

In the “Model Evaluation Thresholds” box, there are input boxes for the Decision Criterion (DC), Exceedance Probability, and Regulatory Standard (RS). Setting these allows model predictions to be evaluated and model specificity, sensitivity, and accuracy to be calculated. The radio buttons inside the “Threshold Transform” box tells VB₃ how

to transform the DC and RS to compare to model predictions and observations (see Section 7.2 for further guidance). Note that we define the “observations” as the measured values of the model’s response variable (e.g., *E.coli* CFU measurements). If the threshold transform definition is set improperly, there can be problems when comparing modeling predictions to observations, so exercise caution.

10.3 Prediction Form

The bottom half of the Prediction interface is occupied by three data panels (the empty gray sections separated by blue vertical bars at the bottom of Figure 61): the left holds IV data; the middle is for observations; and the right shows model predictions and evaluation metrics. Each panel also contains a column for a unique ID for each row of data, e.g., the date that data were collected. The panels have separate horizontal and vertical scroll bars that become visible if the number of rows or columns exceeds the viewable area. The three panels independently scroll horizontally, but as a group vertically. Panels can be re-sized by clicking and dragging the blue vertical partitions. The order of the columns in the left (IV) and right (Model Predictions) panels can be changed by clicking and dragging the column headers left or right. If it is important to save a re-arranged column order for the selected model, click on the “Save Column Order” button just above the IV panel.

Users can import data from files using the “Import IV Data,” “Import Observations” and “Import Combined” (both IVs and observations) buttons on the top ribbon, or type data directly into the left and middle grids. It is the user’s responsibility to ensure that IV data are in the same units as those used to construct the model. Depending on the model selected for prediction, the left panel will contain one column for every unique model IV plus a column for an ID. The middle panel has two columns: one for the ID and one for the observations (note that the name of the observation column is identical to the name of the model’s response variable).

10.4 Column Mapping of Imported Data

When data are imported via one of the three import buttons (Figure 61), a “Column Mapper” window opens (Figure 62). This allows users to tell VB₃ which columns in the imported datasheet should be used to fill in the row IDs, IVs and the observations. By default, the first column of the imported file is mapped to the ID field, but this can be overridden. If a column in the imported spreadsheet has a name identical to a model IV or the response variable, VB₃ will select it as the appropriate column for that IV or the observations. If no identically-named column is found, the user must specify which column of the imported file should be used for the IV and observations.

Once a user has gone through the mapping process for a model, that configuration is saved. If another data file with the same column names is imported, the column mapper will not appear. If a model has a saved mapping configuration, it can be viewed and cleared by clicking “View Column Mapping” on the ribbon (Figure 61).

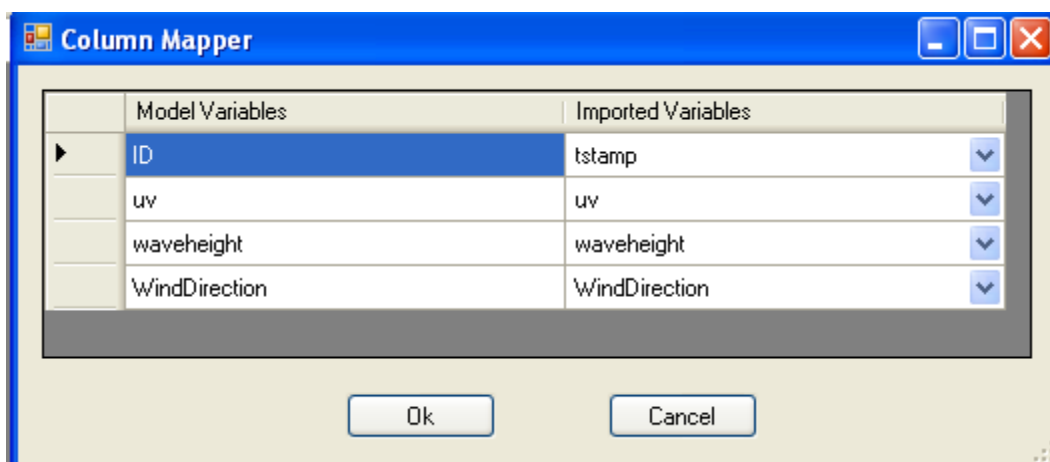


Figure 62. Importation of IV data using the “Column Mapper” window

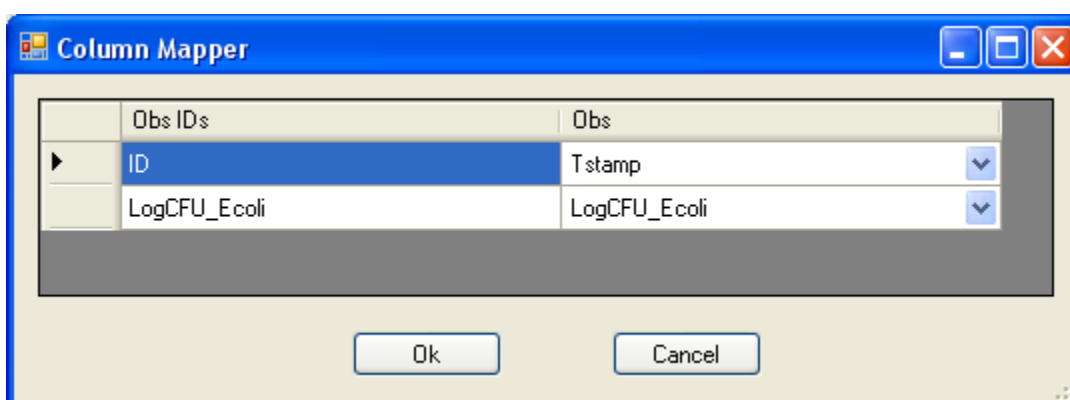


Figure 63. Importation of observational data using the “Column Mapper” window

After observations have been imported or manually entered, users specify the correct data transformation to ensure proper comparison to model predictions. This is done by right-clicking on the observation column header (the right column of the middle panel) and choosing an option from the “Define Transform” drop-down menu: none, $\log_{10}$, $\log_e$, or a power transformation; “none” is the default choice. For example, if $\log_{10}$ observations are imported, the user must change the “Define Transform” menu option to “Log10.” If untransformed (raw) values of the observations are entered/imported, then the appropriate “Define Transform” menu choice would be “none.”

The IV data are automatically scanned for errors (e.g., blank or non-numeric cells) when “Make Predictions” is clicked on the ribbon (however, this button is not enabled until data are entered into the IV data panel). If bad data cells are found, VB₃ will tell the user to run an IV data scan by clicking the “Scan IV Data” button on the ribbon (Figure 64). The IV scan pop-up window is very similar to the one seen on the Global Datasheet; however, “Delete Column” is not a choice. “Replace With” and “Delete Row” are the only options for dealing with problems in the IV data grid.

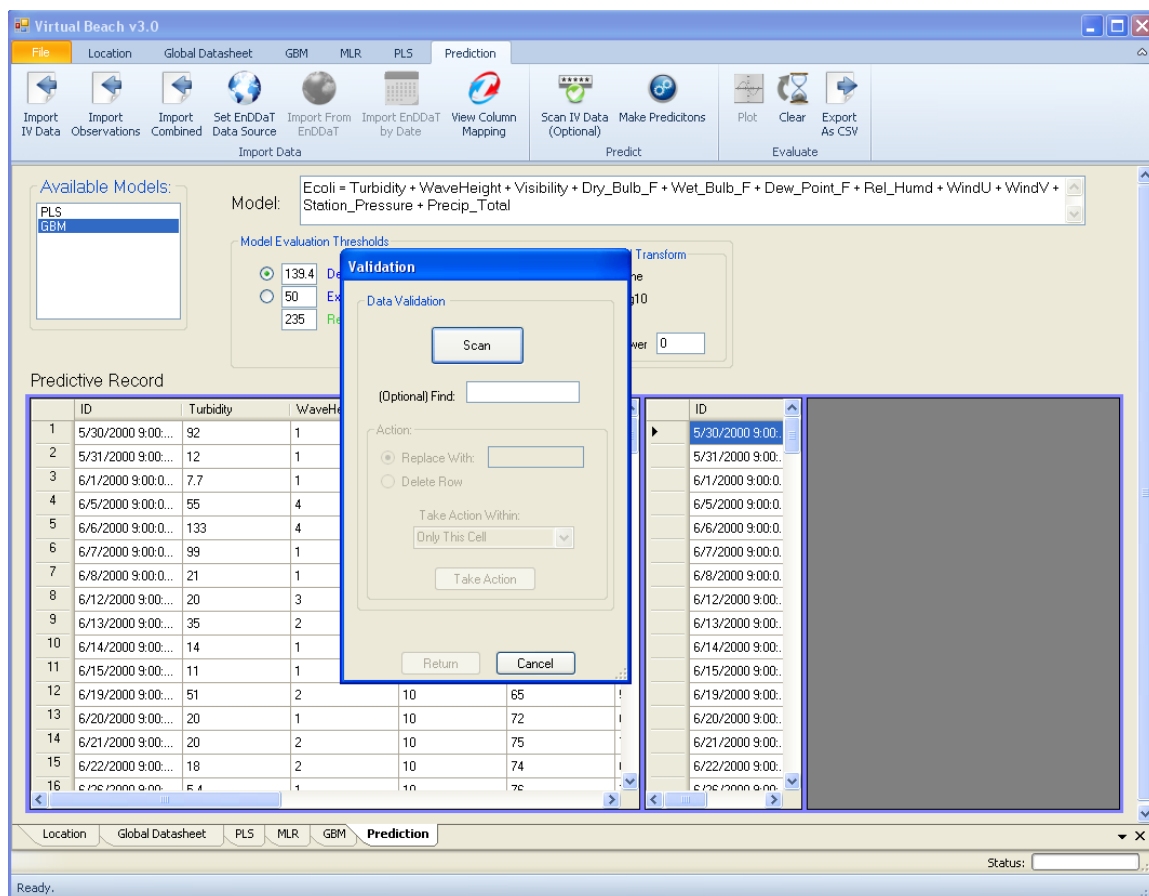


Figure 64. The scan IV window on the MLR Prediction tab

Observational data need not be present to make predictions, but they are needed for model evaluation (sensitivity, specificity, false negatives, false positives, accuracy). After clicking “Make Predictions” on the ribbon, VB₃ uses the model, IV data, and observational data to fill the right panel with these data columns: ID, Model Prediction, Decision Criterion, Exceedance Probability, Regulatory Standard, and Error Type (Figure 65).

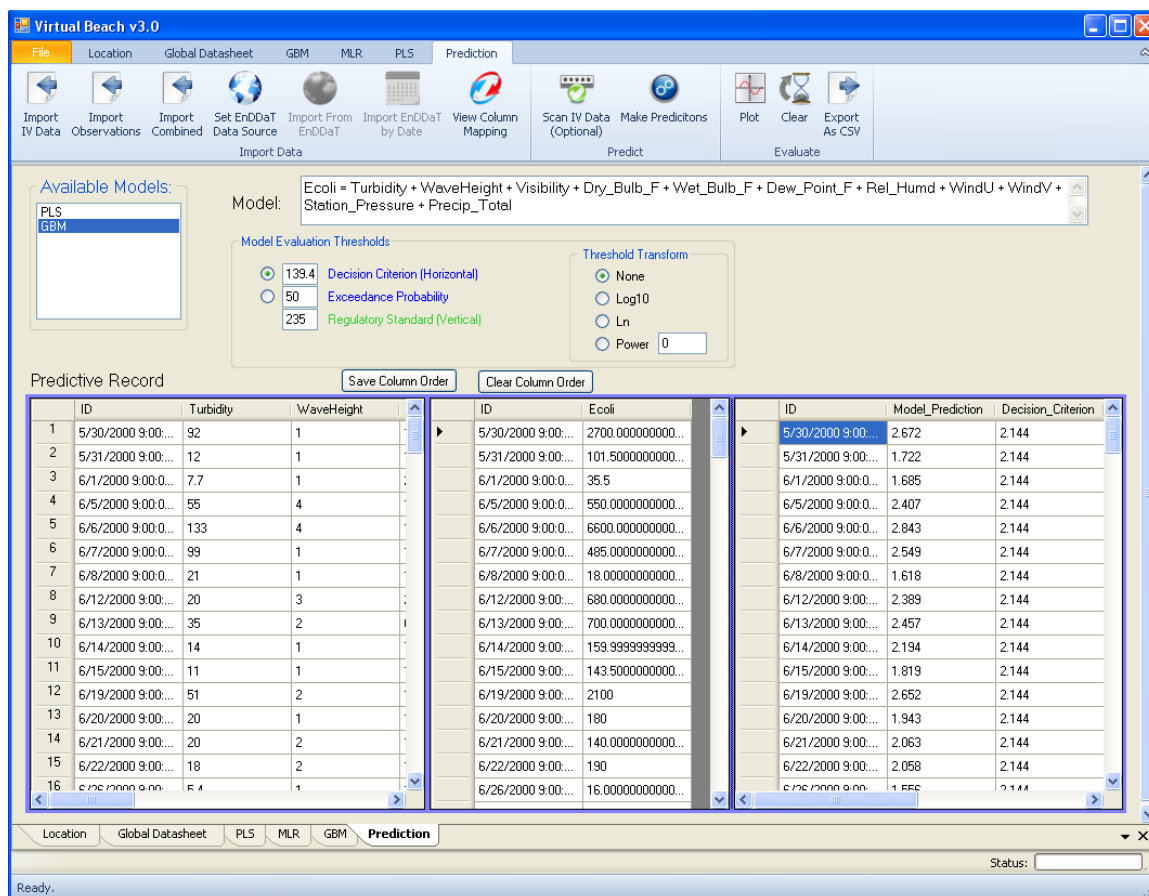


Figure 65. A prediction grid after IVs and observational data have been imported, and model predictions made

The ID column of the model output panel is taken directly from the ID column of the IV panel, not IDs in the middle panel. VB₃ will make one model prediction per row in the IV data panel, regardless of how many observations are entered in the middle panel.

The Model Prediction column contains predicted values of the response variable, initially displayed in the same units as the model's response variable. Right-clicking on this column header changes how predictions are displayed in the table (raw, log, or power units). The Decision Criterion and Regulatory Standard are set by the user. They are displayed in the same units as the Model Predictions, and their column headers can be right-clicked to change the displayed units. The Exceedance Probability (displayed as a percentage, or 100 times the probability) is defined as the probability that the model's prediction will be larger than the Decision Criterion, based on uncertainty bounds (confidence intervals) of the model's predictions.

To compare model predictions to observations, VB₃ looks at the prediction ID and attempts to find an observation in the middle panel with the same ID. It does not require unique IDs for each row in the observation panel, but a model prediction is compared to the first observation found with the same ID. When comparing model predictions to observations, an error ("False Negative" or "False Positive") will be reported in the "Error Type" column.

We again emphasize that assessing model output correctly depends on the synchronization of units of the Decision Criterion (DC), Regulatory Standard (RS),

model predictions, and observations. VB₃ will ensure this happens if the user correctly specifies the units for the observations (using the right-click column header menu of the right column of the middle panel) and for the DC and RS (using the radio buttons in the “Threshold Transform” box of the prediction window).

10.5 Viewing Plots

After predictions are made, a scatterplot of observations versus predictions, or observations versus the probability of exceedance, can be viewed by clicking “Plot” on the ribbon (Figures 65 and 66). If no observational data were entered, a message asking for observational data appears. The features of this plot are similar to those described in Section 7.6. Plotted points are based on comparing model predictions (right pane of the Prediction Form) with observations (middle pane) that share the same unique row ID. Note that the plotted exceedance probabilities are not automatically re-computed because the Decision Criterion is changed in this plotting window. To see updated exceedance probabilities for a new Decision Criterion, users must close this plotting window, change the DC in the “Model Evaluation Thresholds” box, re-click the “Make Predictions” button on the ribbon, and then click the “Plot” button again.

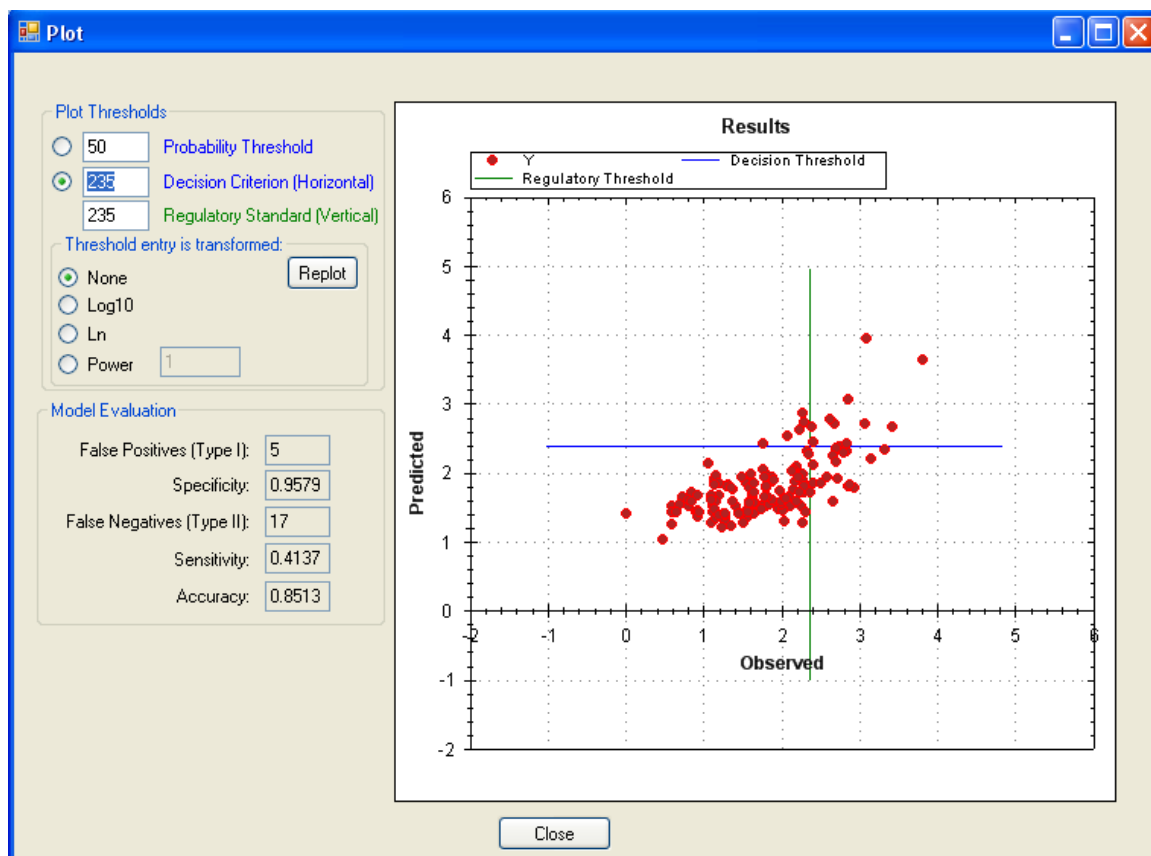


Figure 66. Prediction interface plotting of the observations versus predictions, with model evaluation threshold controls

10.6 Prediction Form Manipulation

Two other buttons found in the “Evaluate” section of the ribbon are “Clear” and “Export as CSV.” To view the table in a spreadsheet or word processing program, “Export as CSV” saves the contents of the entire table (all three panels) in .csv format. “Clear” deletes all information in every panel of the table. As with most tabular information in VB₃, data in individual panels can be selected with a left-click and drag. Control-c and Control-v are then used to copy and paste the data into another application such as Excel.

10.7 Importation of EnDDaT Data

The Environmental Data Discovery and Transformation (EnDDaT; <http://cida.usgs.gov/enddat/>) service accesses data from a variety of sources, compiles and processes it, and performs common transformations. The result is environmental data from multiple sources sorted into a single table. EnDDaT is a tool for compiling datasets prior to model development. Once models are developed, EnDDaT can create datasets for the VB₃ Prediction tab. The “Set EnDDaT Data Source,” “Import from EnDDaT” and “Import EnDDaT by Date” buttons on the ribbon (Figure 67) allow users to import data directly from the EnDDaT web service to the prediction tab of VB₃, avoiding manual entry. See the EnDDaT user guide (available from the EnDDaT website link above) for step-by-step instructions on obtaining data, specifying transforms, processing data and developing a URL.

To import EnDDaT data to the IV panel of the prediction grid, click the “Set EnDDaT Data Source” button and insert an EnDDaT-generated URL that calls for the IVs needed to make predictions (Figure 68). Choose and activate the radio button for whether to collect data from a specific time (e.g., the time the beach was visited) or from the most recently available time. Users must also choose the desired time zone from the dropdown list. After clicking “OK,” the “Import from EnDDaT” and “Import EnDDaT by Date” buttons are enabled on the ribbon. To import data for the current day, use the former button. Clicking the latter button opens a calendar for retrieval of data from a previous day (Figure 69). Whichever button is used, afterwards a pop-up window will indicate EnDDaT is being accessed (Figure 70). Once data have been retrieved, the “column mapper” window will open, allowing the user to specify which columns in the imported EnDDaT data should be matched to each IV in the selected model (see Section 10.4 for more details on column mapping).



Figure 67. The three EnDDaT-related buttons on the prediction tab

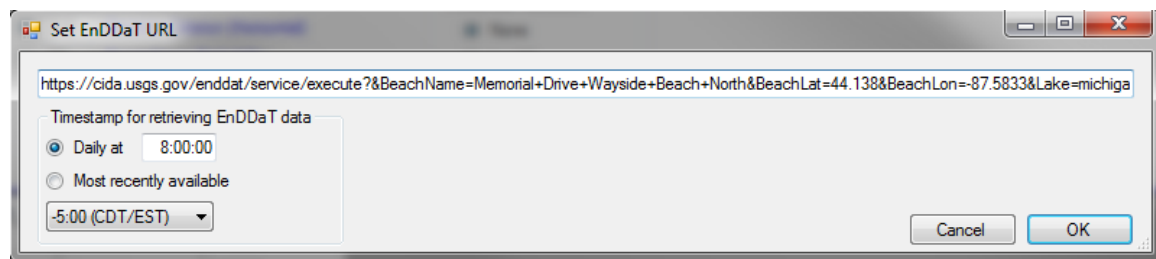


Figure 68. Setting URL options for retrieval of data from EnDDaT

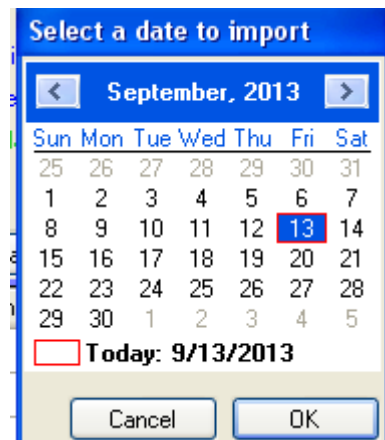


Figure 69. Choosing a previous day for EnDDaT data retrieval

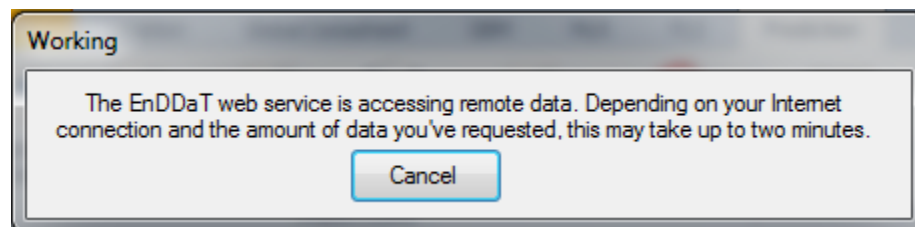


Figure 70. Pop-up window indicating that data have been requested from EnDDaT

11. USER FEEDBACK

The USEPA and USGS provide no warranty, expressed or implied, as to the correctness of the furnished software or the suitability for any purpose. The software has been tested, but as with any complex software, there could be undetected errors. Suggestions and experiences from the user community are welcomed by the Virtual Beach design/development team, and users are encouraged to report problems, issues and likes/dislikes to:

Mike Cyterski, USEPA: 706.355.8142 (cyterski.mike@epa.gov)

Steve Corsi, USGS: 608.821.3835 (srcorsi@usgs.gov)

The USEPA has limited resources to assist users; however, we make an attempt to fix reported problems and help whenever possible.

12. REFERENCES

- Anderson, T.W., Darling, D.A., 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics* 23: 193–212.
- Brooks, W.R., Fienen, M.N., Corsi, S.R., 2013. Partial least squares for efficient models of fecal indicator bacteria on Great Lakes beaches. *J Environ Manage* 114:470-5. doi: 10.1016/j.jenvman.2012.09.033.
- Cook, R., Weisberg, S., 1982. *Residuals and Influence in Regression*. Chapman and Hall, New York.
- Cyterski, M., Galvin, M., Parmar, R., Wolfe, K., 2012. *Virtual Beach User's Manual – version 2.2*. USEPA/600/R-12/024/.
- Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1 (1), 54e77.
- Fogel, D. (editor), 1998. *Evolutionary Computation: The Fossil Record*. New York: IEEE Press.
- Frick, W.E., Ge, Z., Zepp, R.G., 2008. Nowcasting and forecasting concentrations of biological contaminants at beaches: a feasibility and case study. *Environmental Science and Technology* 42, 4818-4824.
- Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:5, 1189-1232.
- Hastie, T., Friedman, J., Tibshirani, R., 2009. *The Elements of Statistical Learning*. Springer-Verlag, New York.
- Myers, R., 1990. *Classical and Modern Regression with Applications*, 2nd Edition. Duxbury Press, Belmont, California.

13. ACKNOWLEDGMENTS

We would like to thank the following people, who generously donated their time and expertise for software testing and review of this document, as well as general support for the continued development of VB:

Adam Mednick
Wisconsin Department of Natural Resources
Madison, WI

David Rockwell
Cooperative Institute for Limnology and Ecosystems Research, University of Michigan
Center of Excellence for Great Lakes and Human Health
NOAA Great Lakes Environmental Research Laboratory
Ann Arbor, MI

Gerry Laniak
USEPA National Exposure Laboratory
Ecosystems Research Division
Athens, GA

Brett Hayhurst
USGS New York Water Science Center
Ithaca, NY

Diane Mas
Fuss and O'Neill, Inc.
Manchester, CT

Amie Brady
USGS Ohio Water Science Center
Columbus, OH

Donna Francy
USGS Ohio Water Science Center
Columbus, OH

Richard Zepp
USEPA National Exposure Laboratory
Ecosystems Research Division
Athens, GA

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

APPENDICES

A.1 Transformations

VB₃ provides the following transformations, where X_t is the transformed IV and X is the original IV:

- Log₁₀: $X_t = \log_{10}(X)$
- Log_e: $X_t = \log_e(X)$
- Inverse: $X_t = 1/X$
- Square: $X_t = X^2$
- Square Root: $X_t = X^{0.5}$
- Quad Root: $X_t = X^{0.25}$
- Polynomial: $X_t = a + bX + cX^2$
- General Exponent: $X_t = X^z$ where the user specifies the value of z

For the polynomial transformation, the Pearson coefficient is calculated as the square root of the adjusted R^2 value derived from the regression of the response on X_t . Because this adjusted R^2 value can be negative, an empirically-derived formula is applied when adjusted R^2 values fall below 0.1:

$$\text{Polynomial Pearson Coefficient} = (-6.67*RE_1^2 + 13.9*RE_1 - 6.24)*(R^2)^{0.5}$$

where

$$RE_1 = 1.015 - 1.856*R^2 + 1.862*adjR^2 - 0.000153*N.$$

R^2 and $adjR^2$ are defined by the regression of the response on X_t , and N = number of observations.

VB₃ transformations (primarily converting x into x^b) have specific processing for certain data values and are not pure mathematical transformations -- they were designed to maintain data order while helping to linearize the response-IV relationship. For the SQUARE ($b = 2$), SQUAREROOT ($b = 0.5$), QUADROOT ($b = 0.25$), INVERSE ($b = -1$) and GENERAL EXPONENT (user-defined b) transformations, VB₃ uses the signed equivalent of the mathematical function:

$$x^b == \text{sign}(x)*(abs(x))^b$$

For example: $(-2)^2 = -4$ $(-9)^{0.5} = -3$ $(-4)^{-0.5} = -0.5$ $(-2)^{-2} = -0.25$

To avoid potentially undefined values (e.g., $1/x$ when $x = 0$), the INVERSE and GENERAL EXPONENT (if the user sets $b < 0$) transformations have special processing:

If $x = 0$, VB₃ will find the minimum value of $abs(z)$ where z is the set of all non-zero values for the IV in question. To compute the transformation after z is defined, VB₃

substitutes $z/2$ for x . From this definition, note that z can be a positive or negative number.

LOG_{10} and LOG_e transforms are also the signed equivalent of their mathematical functions:

$$\begin{aligned}\log_e(x) &== \log_e(x) \\ \log_e(-x) &== -\log_e(x) \\ \log_{10}(x) &== \log_{10}(x) \\ \log_{10}(-x) &== -\log_{10}(x)\end{aligned}$$

In addition, if $(-1 \leq x \leq 1)$, then $\log_e(x) = 0$ and $\log_{10}(x) = 0$

VB_3 will not compute the INVERSE, GENERAL EXPONENT (with a negative b), LOG_{10} and LOG_e transformations for data columns if more than 10% of the IV values are zero. Programmatically, zero is defined as any number whose absolute value is less than $1.0\text{e-}21$.

POLYNOMIAL transformations are the result of a linear regression of the response variable on the IV and the square of the IV:

$$\text{Poly}(x) = a + b*x + c*x^2$$

where a , b , and c are determined by a multiple linear regression of x and x^2 on the response variable.

In general, the name of the transformed column of data that VB_3 creates is simply the type of transformation, with the original data column name in parentheses. For example, the $\log_{10}$ of WaterTemp becomes $\text{LOG}(\text{WaterTemp})$; however, there are some exceptions:

$\text{INVERSE}(x,y)$: x is the original data column name and y is the $z/2$ value discussed in the last paragraph on page 80.

$\text{POWER}(x,y)$: when y is positive, x is the original data column name and y is the exponent specified by the user.

$\text{POWER}(x,y,z)$: when y is negative, x is the original data column name, y is the exponent specified by the user, and z is the $z/2$ value discussed earlier in this section.

$\text{POLY}(x, a,b,c)$: x is the original data column name and a , b , and c are the values of the polynomial regression coefficients.

A.2 Singular Matrices and Nominal Variables

The solution to least squares regression (MLR modeling is discussed in Section 7) involves computing the inverse of the $X'X$ matrix (the X matrix contains the IV values for the model). When one IV is a linear combination of other IVs, the $X'X$ matrix is *singular*, and trying to invert it produces a mathematical quandary (i.e., division by zero). Examples of variables that are linear combinations of other IVs:

$$\begin{aligned}X_1 &= 3.5 + 4.2 * X_2 \\X_1 &= 1 - X_2 - X_3 - X_4 \\X_3 &= X_1 + X_2\end{aligned}$$

In these examples, it doesn't matter if the IVs are continuous (real numbers) or categorical (0/1 values). In fact, VB₃ allows the user to produce, using the “manipulate” button described in section 6.6, IVs that are linear combinations of others (like example c above). When VB₃ evaluates MLR models, it checks each model for highly correlated IVs because perfectly correlated IVs lead to a matrix singularity and throws out any model with this condition (as measured by the Variance Inflation Factor, explained in Section 7.2). Using example equation c: attempting to compute a regression model involving X_1 , X_2 , and X_3 , VB₃ will issue an error message.

Singularities are often produced if an IV with several categories is being defined using multiple indicator variables. Let's say there is an IV for cloud cover. One could make this categorical measure a continuous variable by using a single column with values ranging from 1 (no clouds) to 5 (completely overcast). This is acceptable because this IV is “ordinal” -- there is a natural order to its values. As values increase from 1 to 5, it implies more clouds.

There may be other categorical IVs that are “nominal,” meaning there is no real order to their values. An example is the species of bird most abundant at the beach on a given day. If there are four possible species (A, B, C, D), it would be incorrect to code this IV in a single column with values 1, 2, 3, and 4. A value of 2 doesn't imply any larger mathematical quantity than a value of 1 or a smaller quantity than a value of 4. So the bird species should be coded as a series of indicator variables, using 0's and 1's (Table A.1):

Table A.1. Example of using 0/1 indicator variables for a multi-category IV

ID	Species_A	Species_B	Species_C	Species_D
Day 1	1	0	0	0
Day 2	1	0	0	0
Day 3	0	1	0	0
Day 4	0	0	0	1
Day 5	1	0	0	0
Day 6	0	0	1	0
Day 7	0	0	1	0
Day 8	0	1	0	0
Day 9	0	1	0	0
Day 10	0	0	0	1
Day 11	0	1	0	0
Day 12	1	0	0	0
Day 13	0	0	1	0
Day 14	0	0	0	1
Day 15	0	1	0	0
Day 16	0	1	0	0

A “1” denotes when a species is dominant and “0” when it isn’t. Looking closely, we see that the four columns form a linear combination:

$$\text{Species_D} = 1 - \text{Species_A} - \text{Species_B} - \text{Species_C}$$

Given this relationship, VB₃ cannot evaluate a MLR model that includes all four columns (mathematically impossible due to a matrix singularity), but a model that contains three or fewer of the columns is acceptable, as is including all four columns in the dataset (but they will never occur together in a model). An advantage of PLS (Section 8) and GBM (Section 9) modeling is that they are not constrained by the collinearity of IVs and can compute solutions for models that include all four columns.

A.3 MLR Model Evaluation Criteria

If p is defined as the number of parameters in a model, n as the number of observations in the dataset, RSS as the residual sum of squares for a model, and TSS as the total sum of squares for a model, then the evaluation criteria for any model can be defined as:

- Akaike Information Criterion (AIC): $2p + n \ln(\text{RSS})$
- Corrected Akaike Information Criterion (AICC): $\ln(\text{RSS}/n) + (n+p)/(n-p-2)$
- R^2 : $1 - \text{RSS}/\text{TSS}$
- Adjusted R^2 : $1 - (1-R^2)(n-1)/(n-p-1)$
- Bayesian (Schwarz) Information Criterion (BIC): $= n \ln(\text{RSS}/n) + p \ln(n)$
- Root Mean Squared Error (RMSE): $(\text{RSS}/n)^{1/2}$
- Predicted Error Sum of Squares (PRESS): $1 - \sum (y_i - \hat{y}_{-i})^2 / \sum (y_i - y_m)^2$
where y_i is the i_{th} observation, $\hat{y}_{-i}$ is the model estimate of the i_{th} observation when the model coefficients are fitted with the i_{th} observation removed from the dataset, and y_m is the mean value of y in the dataset
- Accuracy: $(\text{true positives} + \text{true negatives}) / \text{number of total observations}$
- Specificity: $\text{true negatives} / (\text{true negatives} + \text{false positives})$
- Sensitivity: $\text{true positives} / (\text{true positives} + \text{false negatives})$



Printed with vegetable-based ink on paper that contains a minimum of 50% post-consumer fiber content and processed chlorine free.

SCIENCE
IN
ACTION



EPA

United States
Environmental Protection
Agency

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT NO. G-35

Office of Research and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
\$300