Virtual Beach 3.0.6 – Data Preparation for GBM model

In this module you will learn how to:

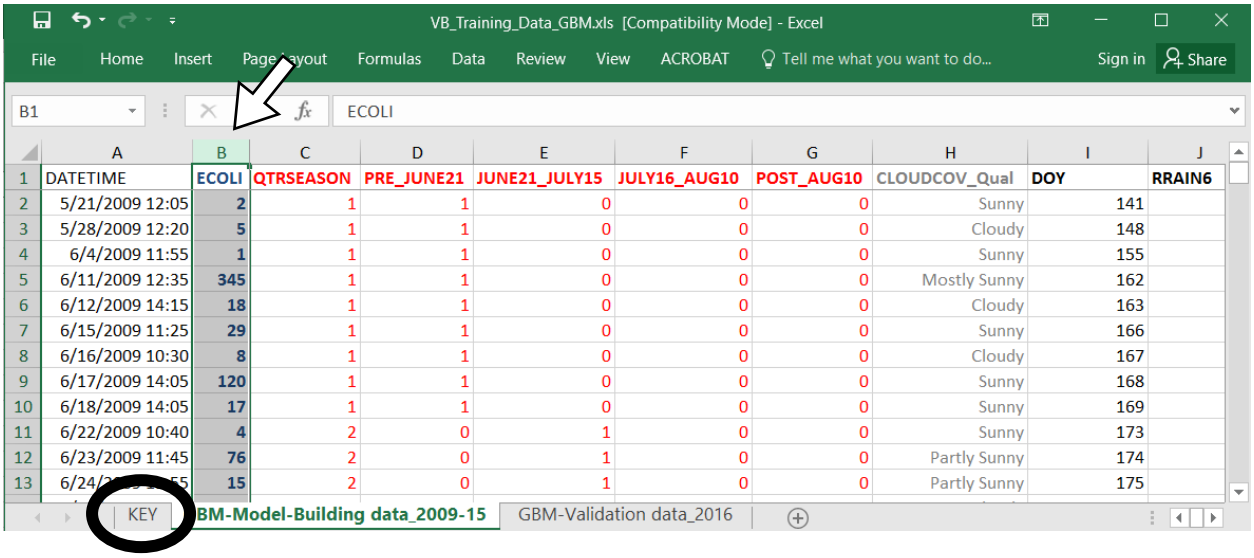
A. Import and clean-up model-building data for your beach
B. Process directional data (wind, currents, waves)
C. Combine two or more predictive variables
D. Transform the dependent bacteria variable

A. Import and clean-up model-building data for your beach

A.1. Open **Microsoft Excel** to preview the data you will be importing into **Virtual Beach 3**. Open the file "VB_Training_Data_GBM.xls".

Be sure to save your data as "*.xls" files. A plugin for **Virtual Beach** 3 is available for importing "*.xlsx" files, but there are still bugs to be worked out.

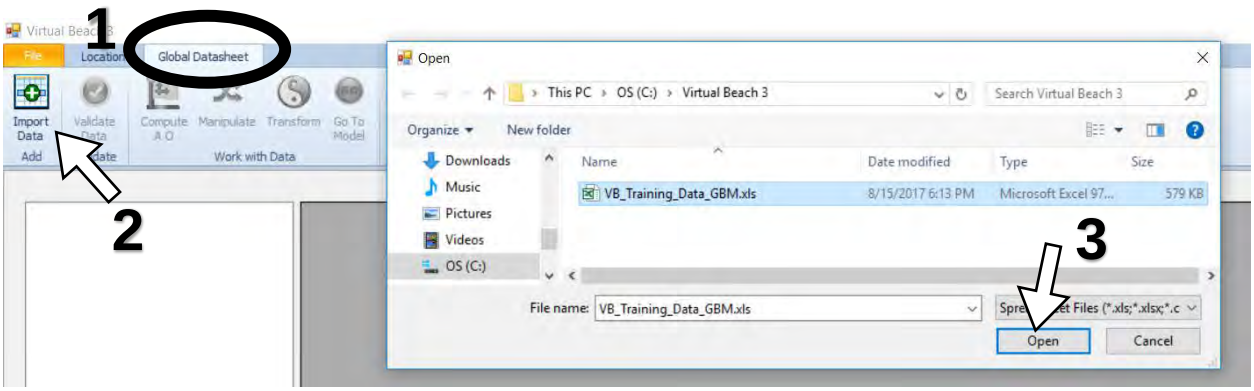
Column **B** is always the *response* variable, "ECOLI" in this example. All data to the right are potential *explanatory* variable. See the **KEY** tab of the **Excel** file for descriptions of variables used in this module. Close the **Excel** file before returning to **Virtual Beach 3**. Data cannot be imported from an open Excel file.



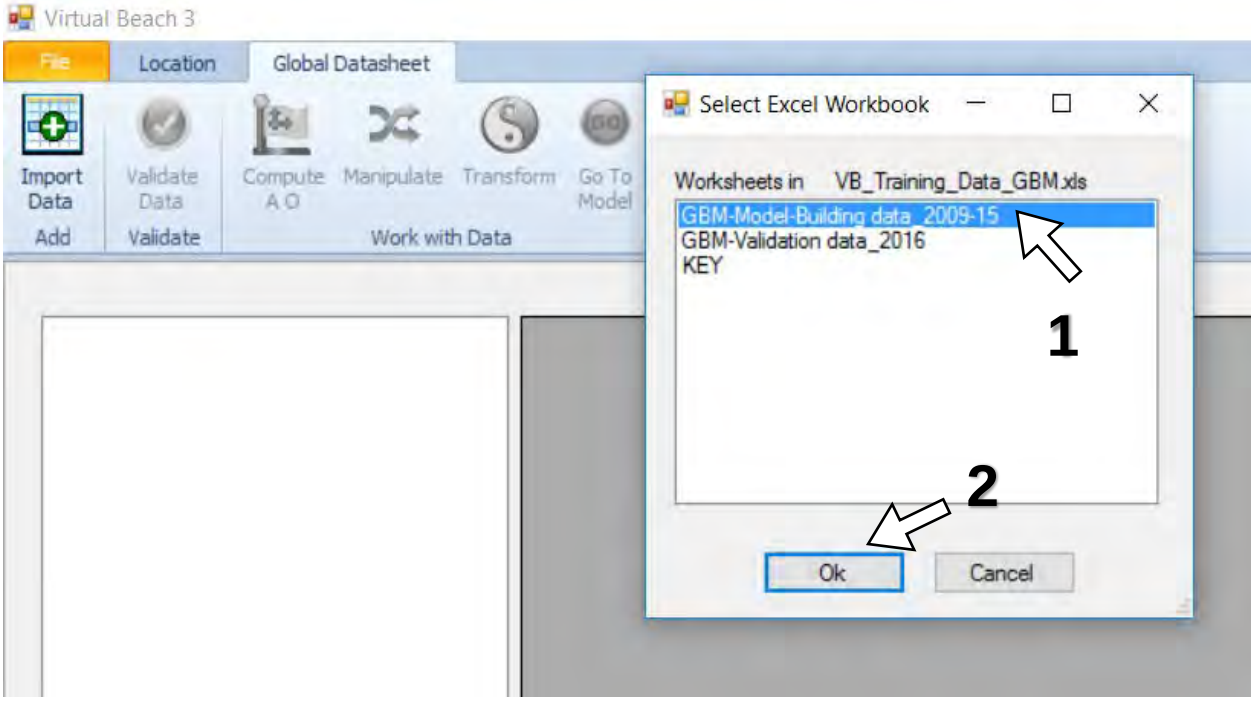
	A	B	C	D	E	F	G	H	I	J
1	DATETIME	ECOLI	QTRSEASON	PRE_JUNE21	JUNE21_JULY15	JULY16_AUG10	POST_AUG10	CLOUDCOV_Qual	DOY	RRAIN6
2	5/21/2009 12:05	2	1	1	0	0	0	Sunny	141	
3	5/28/2009 12:20	5	1	1	0	0	0	Cloudy	148	
4	6/4/2009 11:55	1	1	1	0	0	0	Sunny	155	
5	6/11/2009 12:35	345	1	1	0	0	0	Mostly Sunny	162	
6	6/12/2009 14:15	18	1	1	0	0	0	Cloudy	163	
7	6/15/2009 11:25	29	1	1	0	0	0	Sunny	166	
8	6/16/2009 10:30	8	1	1	0	0	0	Cloudy	167	
9	6/17/2009 14:05	120	1	1	0	0	0	Sunny	168	
10	6/18/2009 14:05	17	1	1	0	0	0	Sunny	169	
11	6/22/2009 10:40	4	2	0	1	0	0	Sunny	173	
12	6/23/2009 11:45	76	2	0	1	0	0	Partly Sunny	174	
13	6/24/2[illegible]	15	2	0	1	0	0	Partly Sunny	175	

This file can be used as a template for formatting beach-specific data.

A.2. Return to **Virtual Beach 3** project file created in the “Beach Orientation” module. **1.** Click the **Global Datasheet** tab. **2.** Click the **Import Data** icon and select the Excel file “VB_Training_Data_GBM.xls”. **3.** Click **Open**.

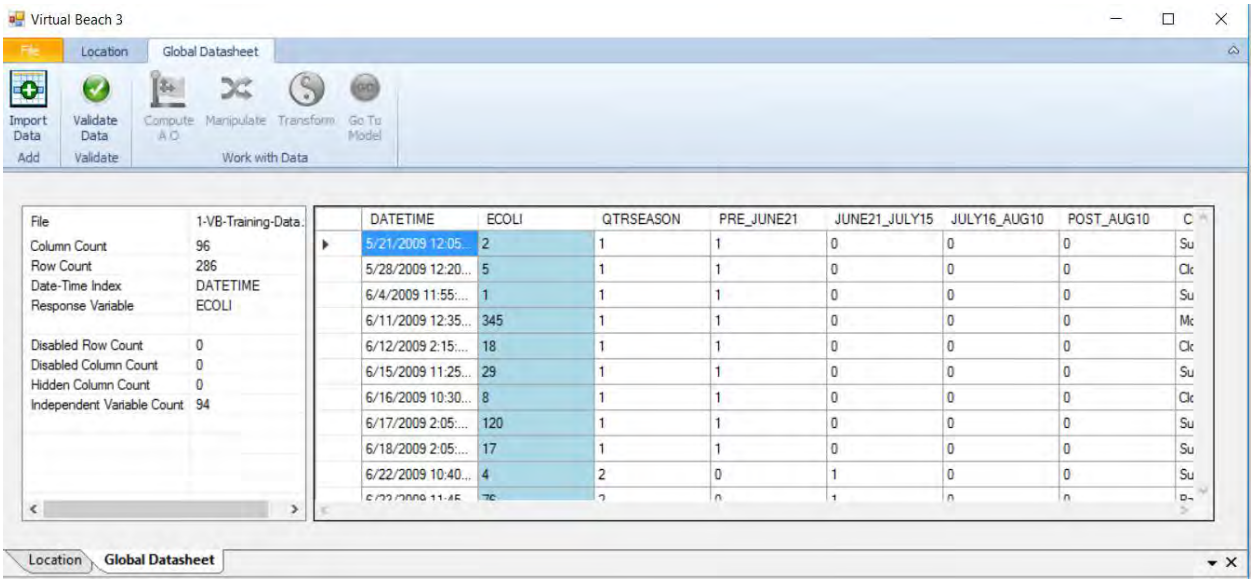


A.3. In this example there is more than one worksheet in the Excel file, so you must choose which one to import. **1.** Select the worksheet **GBM-Model-Building data_2009-15**. **2.** Click **OK**.

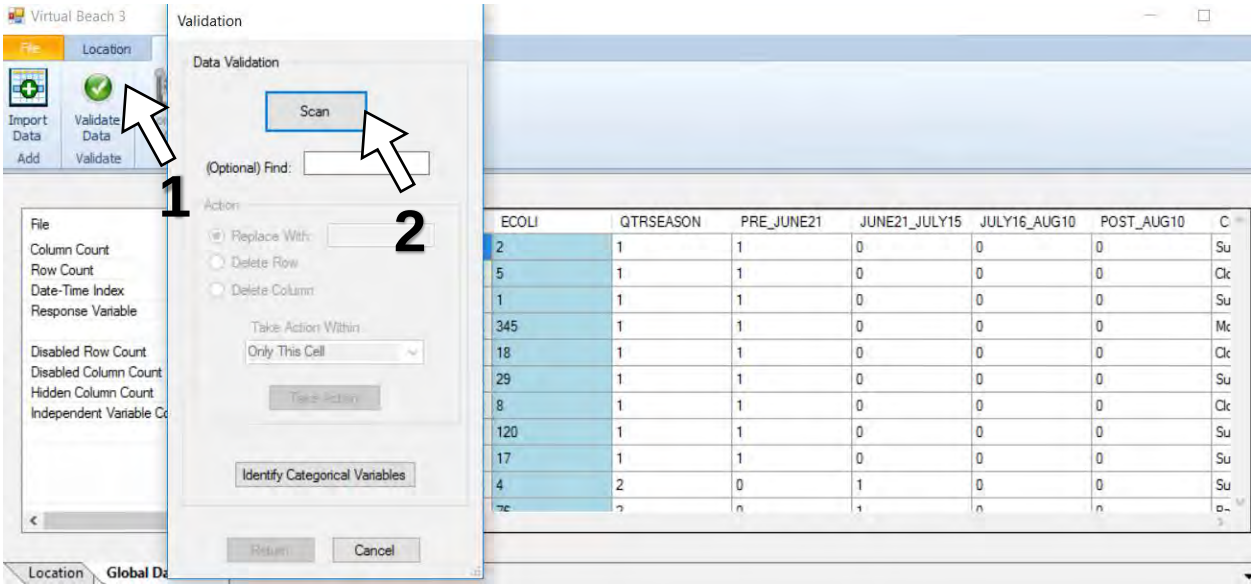


A.4. The data table will open in **Virtual Beach 3**.

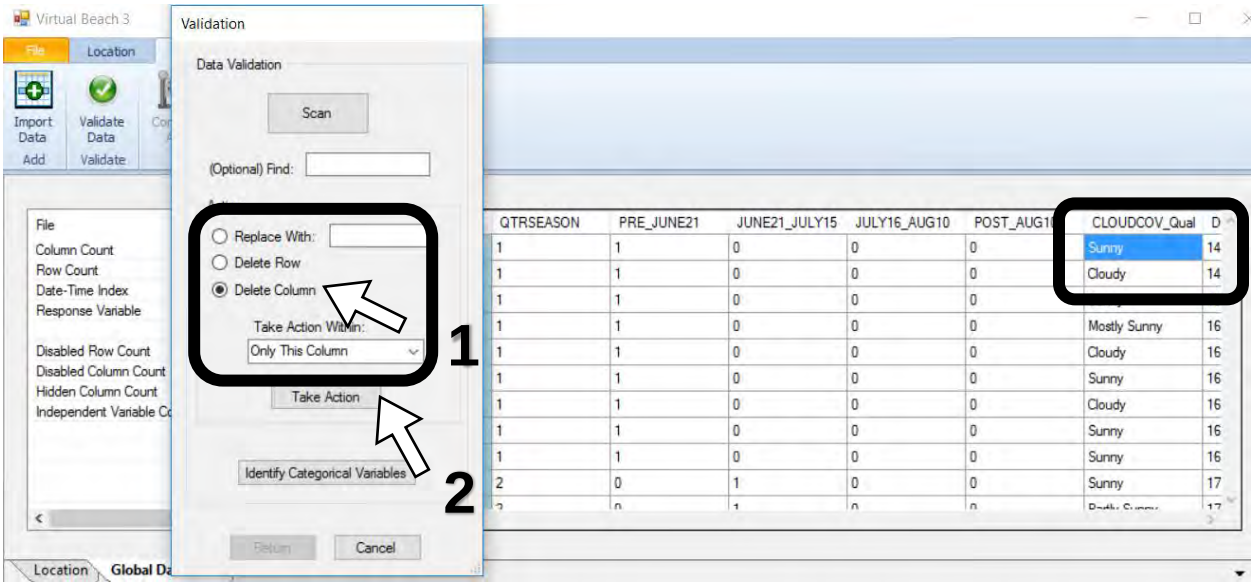
Virtual Beach 3 automatically highlights the second column of the datasheet as the *response* variable, “ECOLI”, in this example. The “Response Variable” is indicated in the left-hand panel, along with “Column Count”, “Row Count” and other descriptions of the data.



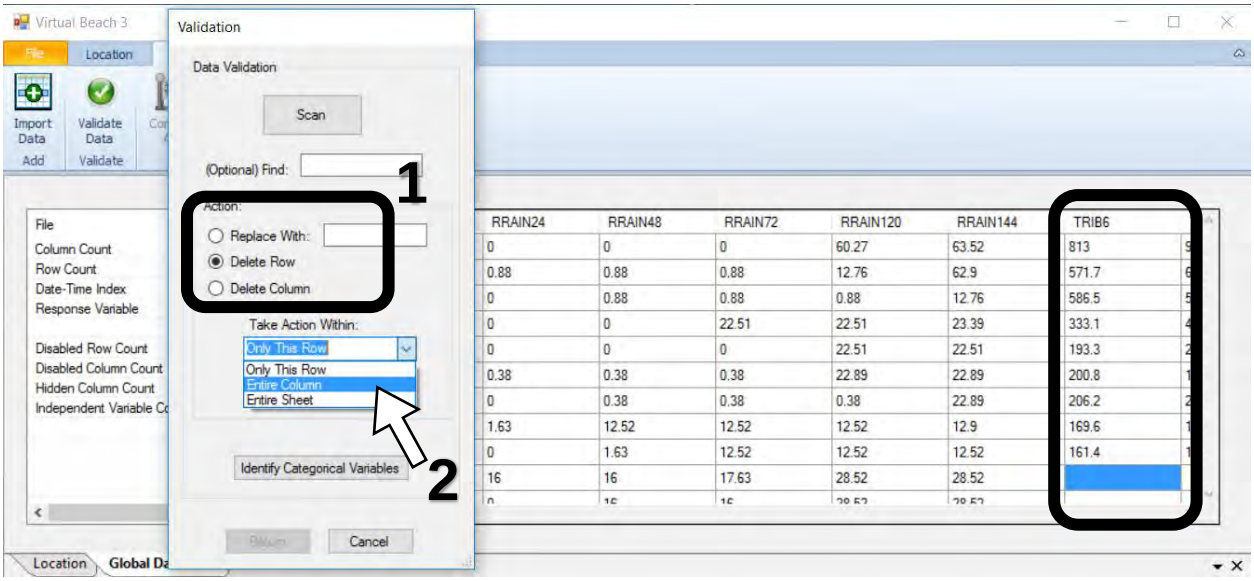
A.5. **Virtual Beach 3** will NOT build a model if any cells have null (missing), or non-numeric (text) values. **1.** Click the **Validate Data** icon to check your dataset. **2.** In the pop-up window, click **Scan**.



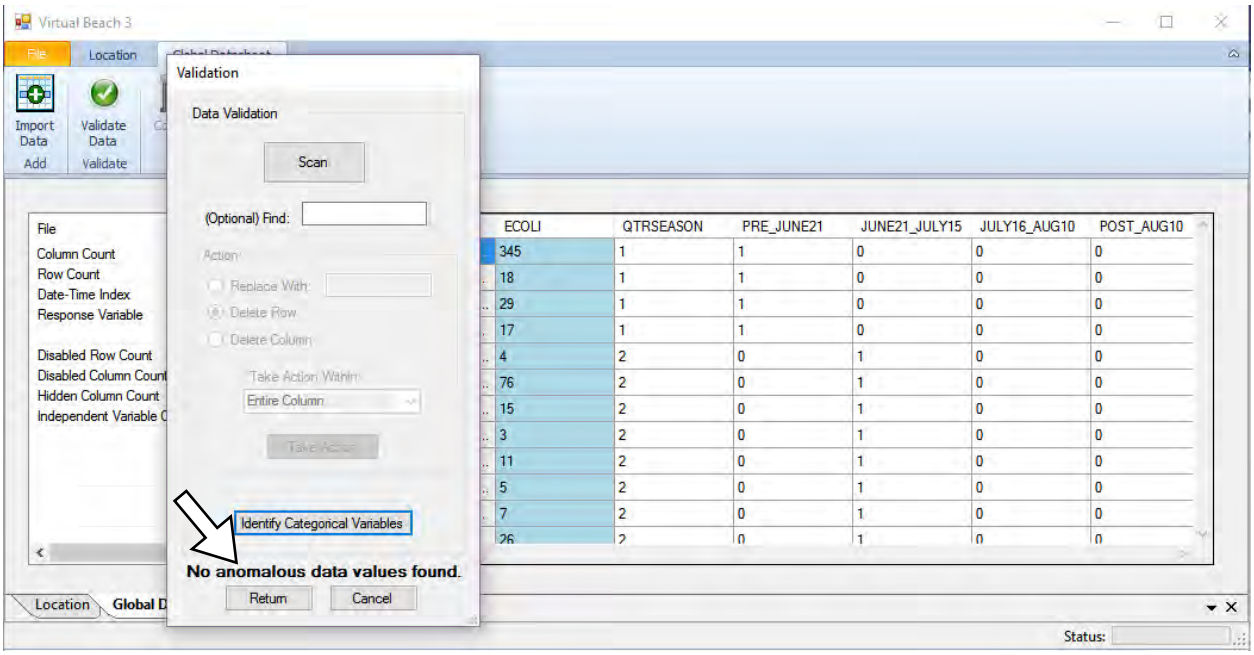
A.6. In this example, the “CLOUDCOV_qual” column is flagged because the values are text, or non-numeric. **1.** Click the radio button next to **Delete Column**. Under **Take Action Within** make sure **Only This Column** is selected. **2.** Click **Take Action**.



A.7. Repeat step A.6 until you come to the “TRIB6” column. The variable is numeric, but some cells are empty. Do not remove the entire column. **1.** Click the radio button for **Delete Row**. **2.** Select **Entire Column** and click **Take Action**.

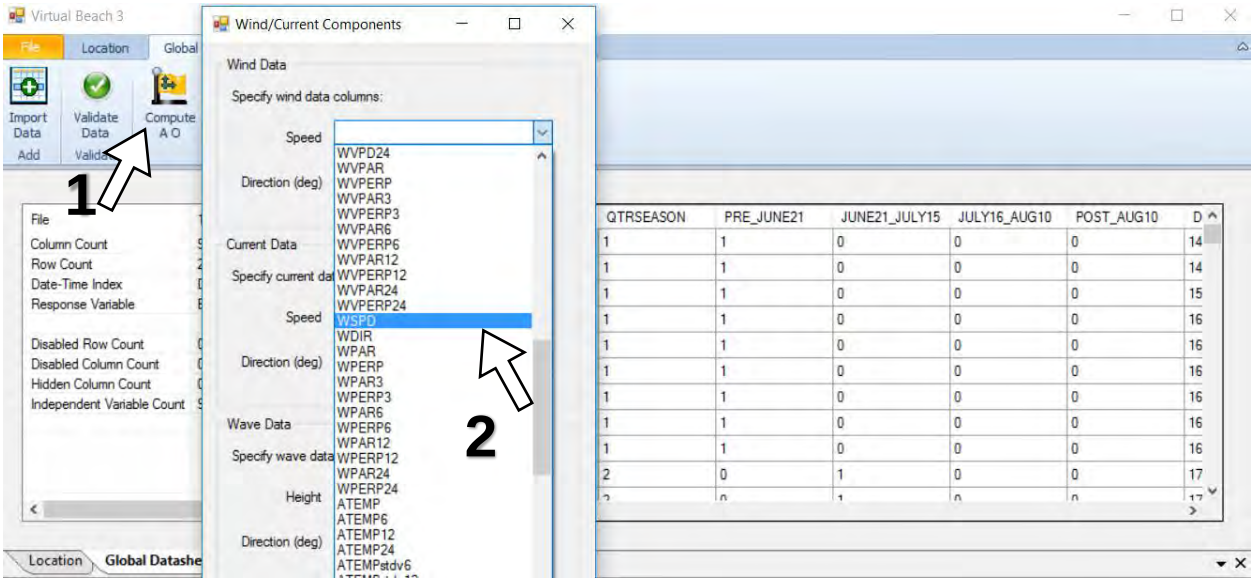


A.8. Repeat Step A.7 until a notice appears at the bottom of the pop-up window stating **No anomalous data values found**. Then click the **Return** button.

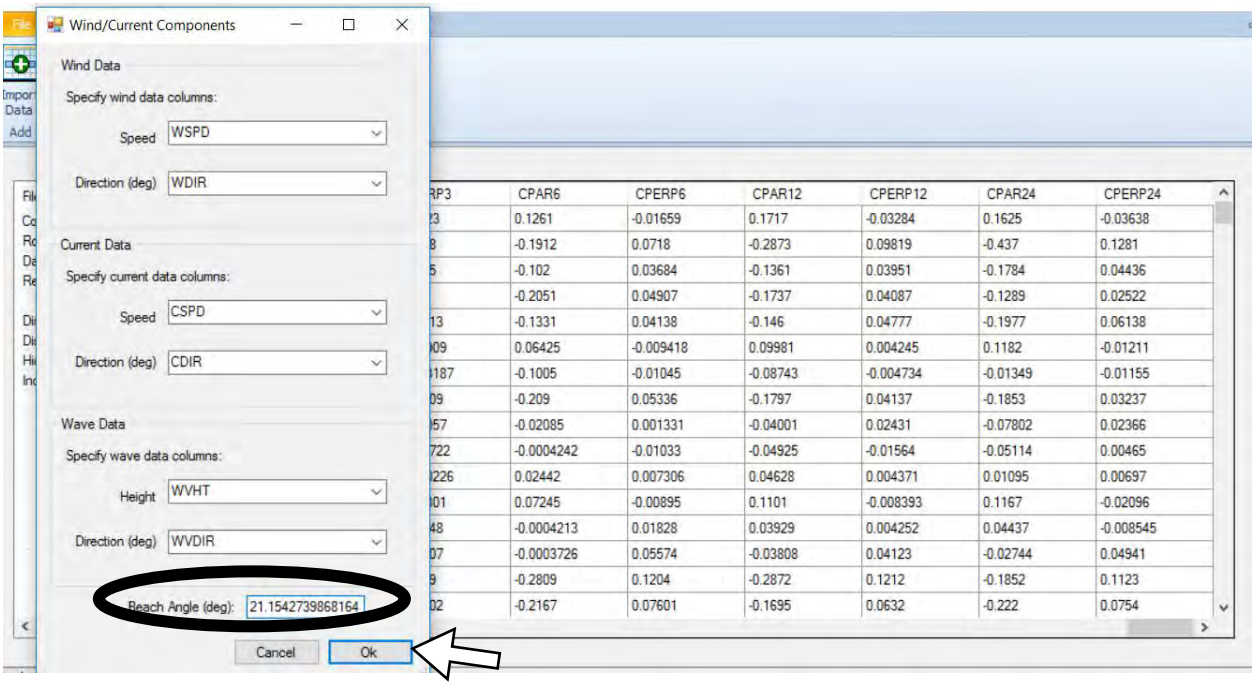


B. Process wind and current data

B.1. **1.** Click the **Compute A O** icon. **2.** In the pop-up window, under **Wind Data**, click the pull down arrow next to **Speed** and select WSPD. For **Direction**, select WDIR.

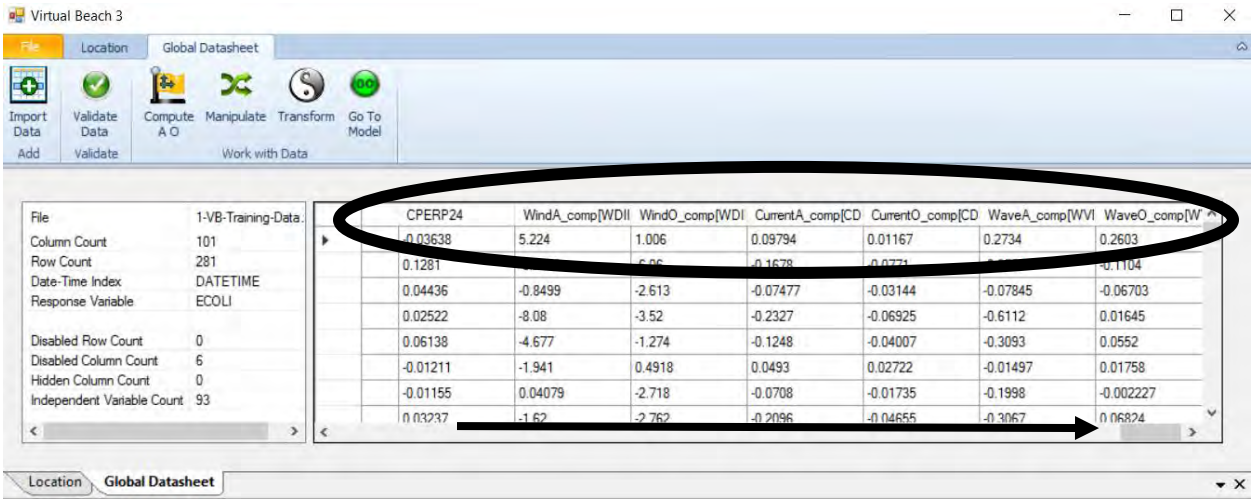


B.2. Repeat for **Current Data**, selecting CSPD and CDIR. Repeat for **Wave Data** selecting WVHT and WVDIR. The Beach Angle is automatically included. Click **OK**.



B.3. Scroll to the far-right end of the table. Six new columns have been added to the end of the global data sheet and that the unprocessed wind, current, and wave data columns are now inactive (red text):

Wind A_comp: along-shore wind speed
Wind O_comp: toward shore wind speed
Current A_comp: along-shore current speed
Current O_comp: toward shore current speed
Wave_A_comp: along-shore wave height
Wave_A_comp: on-shore wave height



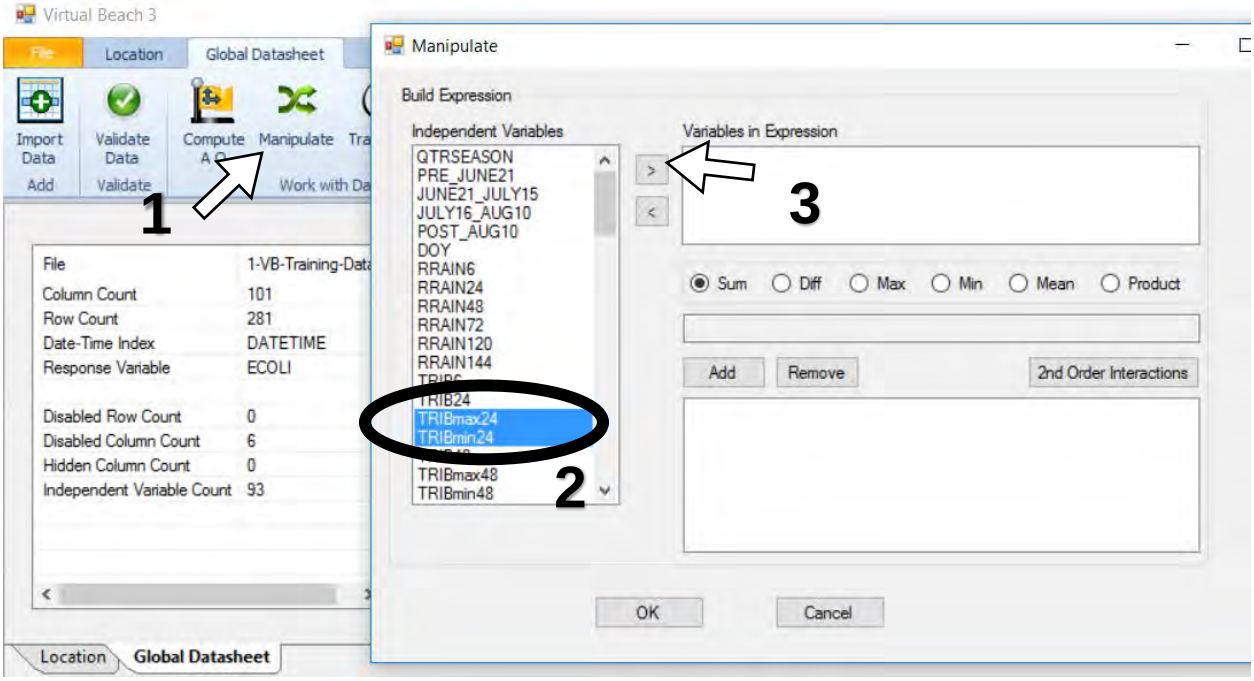
C. Combine two or more predictive variables

Interaction Terms: In situations where two predictive variables are themselves correlated, meaning they interact with one another in terms of how they influence water quality, it may be beneficial to combine them into a single interaction term by **multiplying** them together. Combined the two variables may be better predictors of water quality than if included individually.

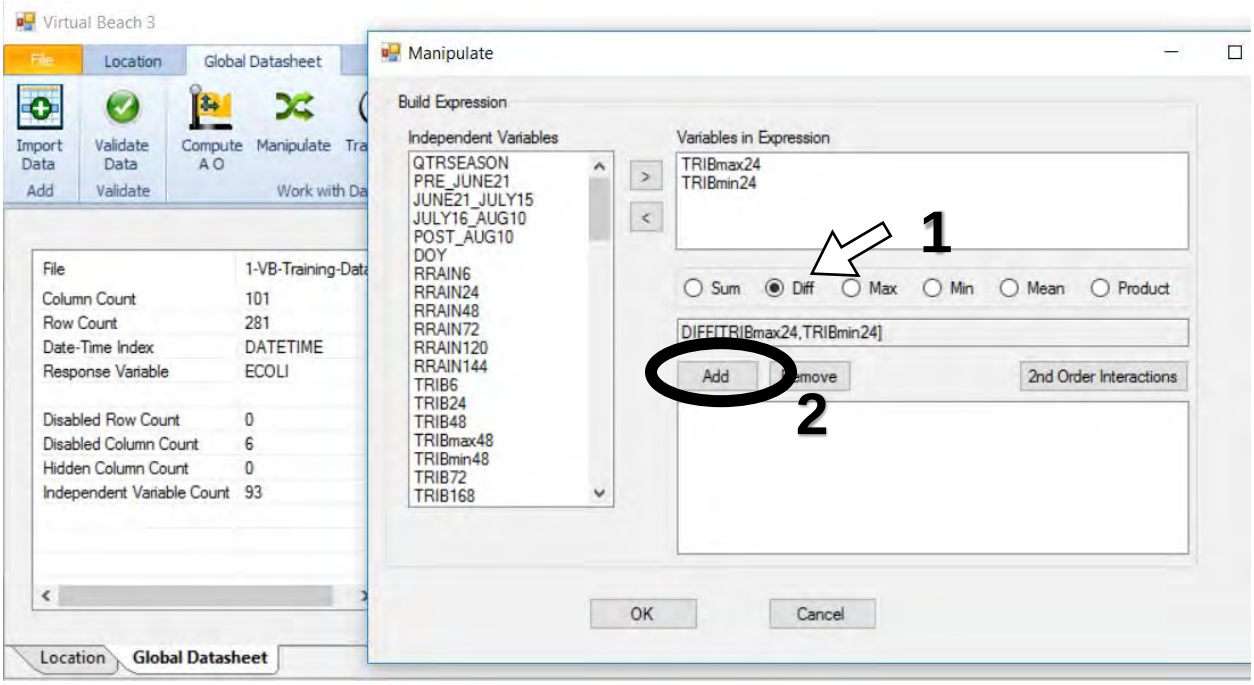
Combined Categories: Some variables are either yes or no. The 0 is "no" and the 1 is "yes". In situations where binary variables represent successive categories of some qualitative variable, like visually-observed water clarity, it may be useful to combine them into a single binary variable by **summing** them. The resulting variable will have a value of 1 when *either* of the two conditions is present. This can be especially helpful when there is little functional distinction between the categories or few cases in which one of the conditions is ever observed. In this example, the difference between TURBID and OPAQUE water is not very distinct; if the water is turbid, it was probably also opaque.

Change-in-Flow Variables. In situations where continuous stream flow data are pre-processed over different timeframes, **subtracting** one temporal snapshot from another can create proxy variables for *changes* in flow. The difference between 24-hour maximum and minimum flow rates indicates whether recent tributary discharge has been consistent or very different after a flash flood event.

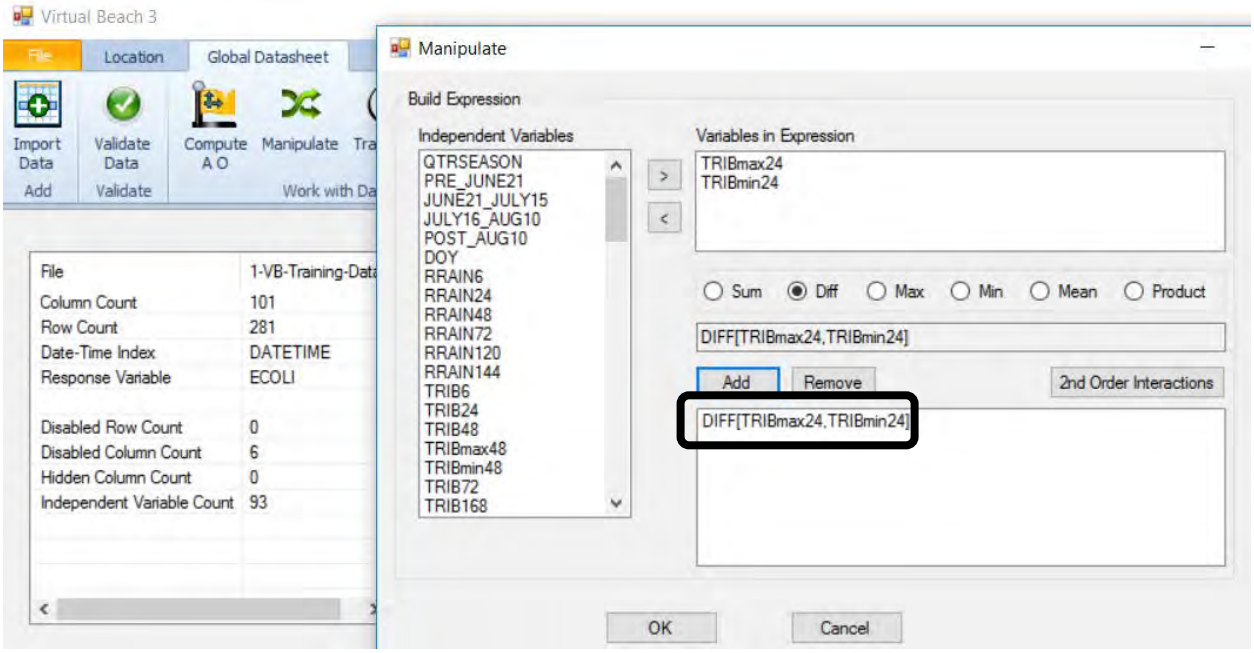
C.1. **1.** Click the **Manipulate** icon. **2.** In the pop-up window, ctrl-select TRIBmax24 and TRIBmin24. **3.** Click the right-arrow "**›**" button.



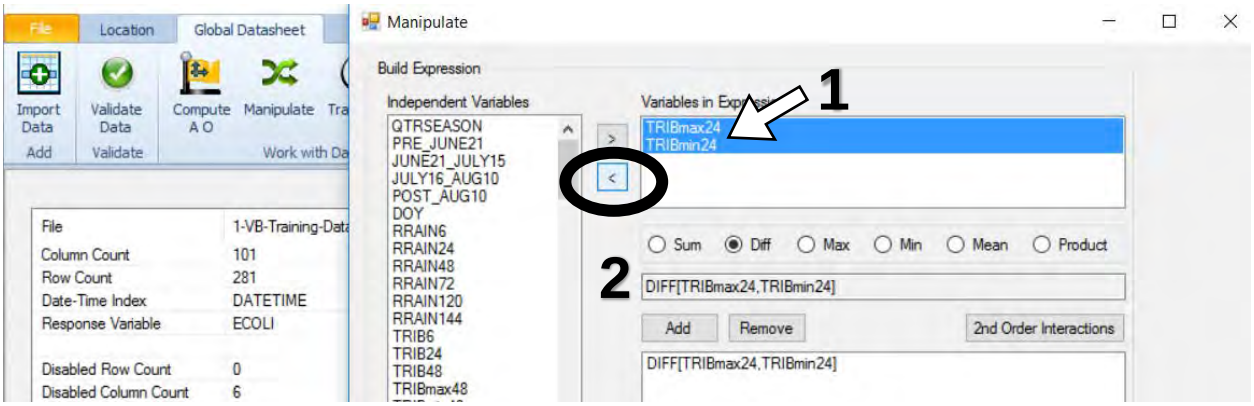
C.2. **1.** Click the radio button next to **Difference**. **2.** Click the **Add** button.



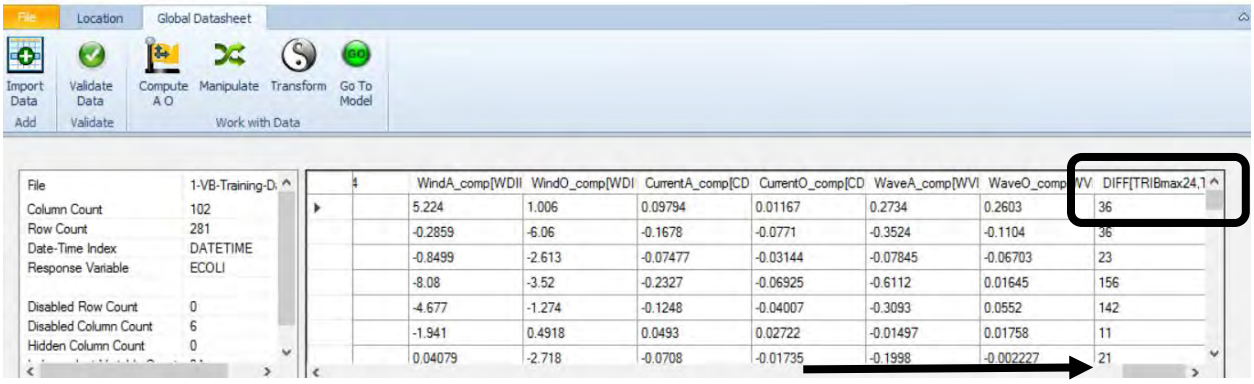
C.3. The change-in-flow variable **DIFF[TRIBmax24,TRIBmin24]** has been added. This approximates whether and to what extent the previous 24 hours of tributary discharge has been constant or varied a lot.



C.4. **1.** Since the expression has been added, remove the two variables by shift-selecting TRIBmax24 and TRIBmin24. **2.** Click the left-arrow "‹" button to move the variables back to the main list.



C.5. Repeat the steps in C.2 and C.3 again to create expressions of combined variables as needed and click **OK**. Scroll to the far-right end of the table to see any new columns added through this process.



D. Transform the dependent bacteria variable

D.1. To build a usable nowcast model, bacteria counts must be transformed. Log10 is a common transformation for microbial concentrations. Right-click on the "ECOLI" column header and select Transform > Log10. Save your project file. You can now move onto the next module, "Building a GBM Model".