

Virtual Beach 3.0.6 – Data Preparation for MLR model

In this module you will learn how to:

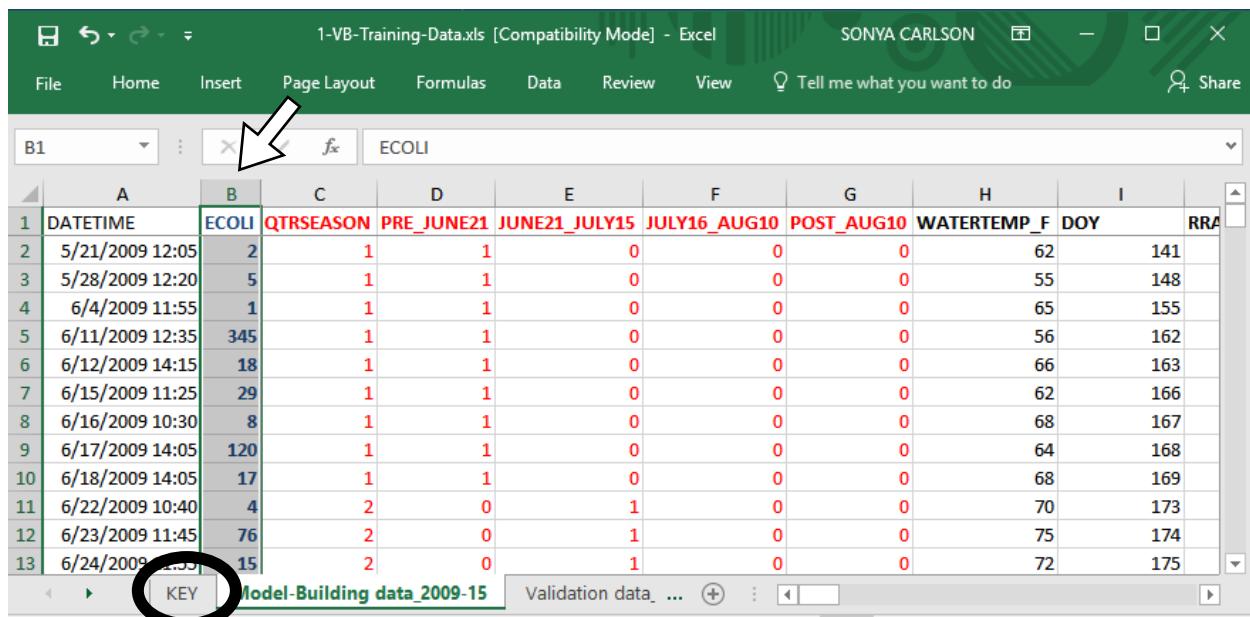
- A. Import and clean-up model-building data for your beach
- B. Process directional data (wind, currents, waves)
- C. Combine two or more predictive variables
- D. Transform variables and explore potential relationships

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_MLR.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.



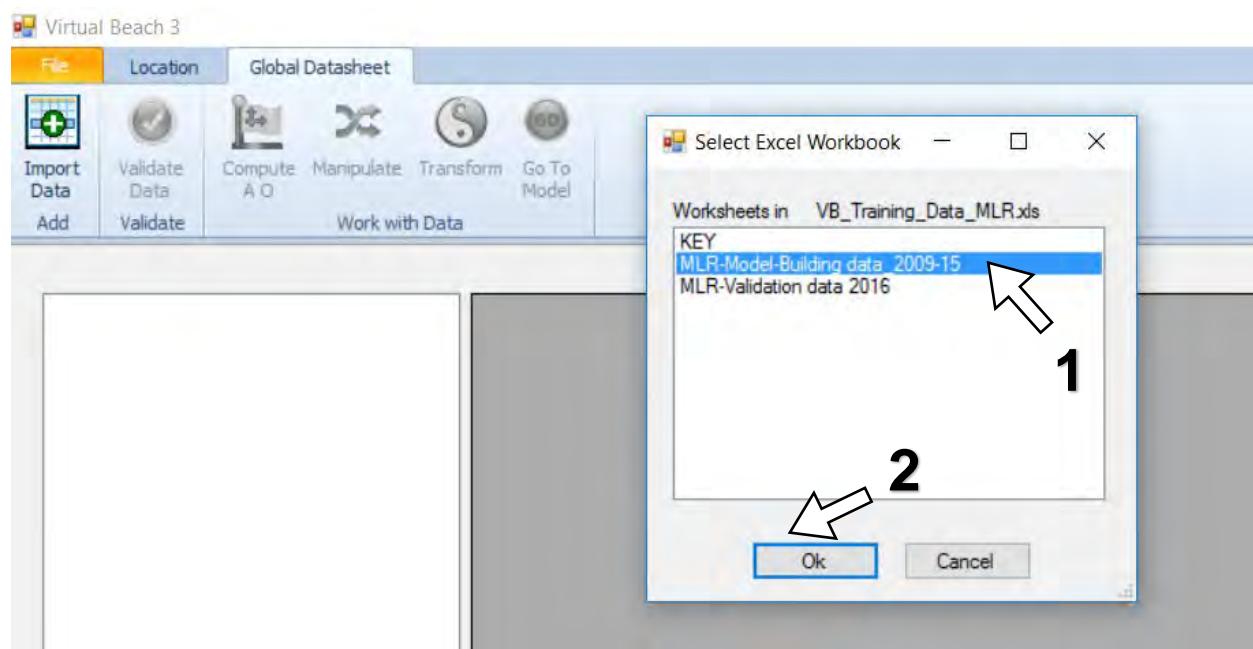
	A	B	C	D	E	F	G	H	I
1	DATETIME	ECOLI	QTRSEASON	PRE_JUNE21	JUNE21_JULY15	JULY16_AUG10	POST_AUG10	WATERTEMP_F	DOY
2	5/21/2009 12:05	2		1	1	0	0	0	62
3	5/28/2009 12:20	5		1	1	0	0	0	55
4	6/4/2009 11:55	1		1	1	0	0	0	65
5	6/11/2009 12:35	345		1	1	0	0	0	56
6	6/12/2009 14:15	18		1	1	0	0	0	66
7	6/15/2009 11:25	29		1	1	0	0	0	62
8	6/16/2009 10:30	8		1	1	0	0	0	68
9	6/17/2009 14:05	120		1	1	0	0	0	64
10	6/18/2009 14:05	17		1	1	0	0	0	68
11	6/22/2009 10:40	4		2	0	1	0	0	70
12	6/23/2009 11:45	76		2	0	1	0	0	75
13	6/24/2009 12:55	15		2	0	1	0	0	72
									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_MLR.xls”. 3. Click **Open**.

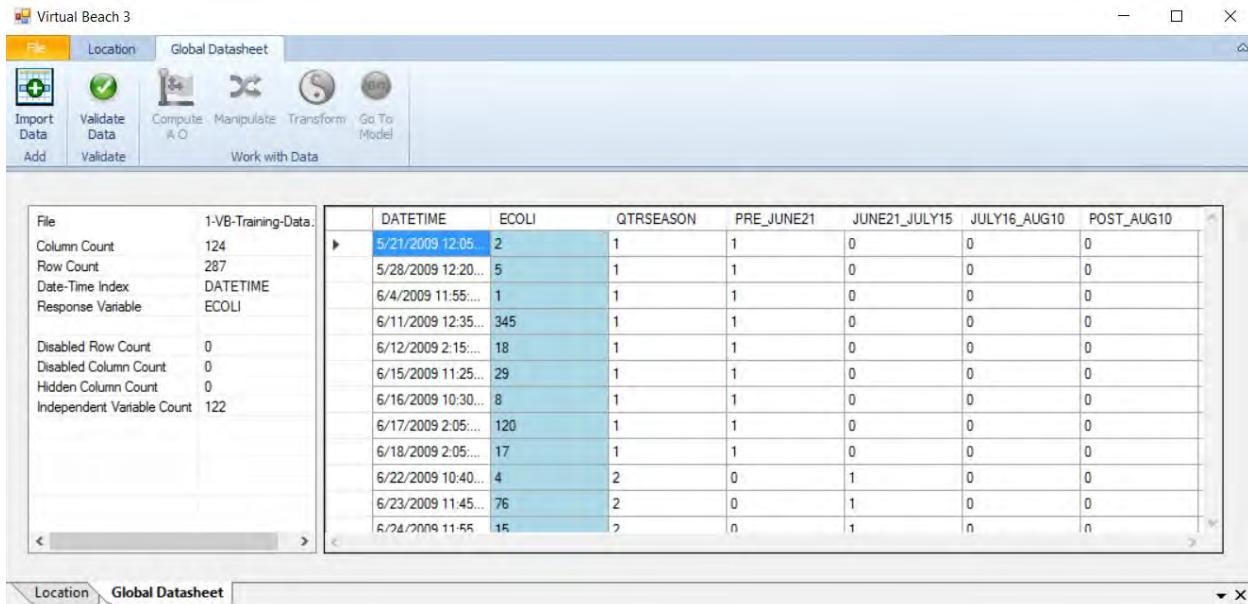


A.3. In this example there is more than one worksheet in the Excel file, so you must choose which one to import. 1. Select the worksheet **MLR-Model-Building data_2009-15**. 2. Click **OK**.



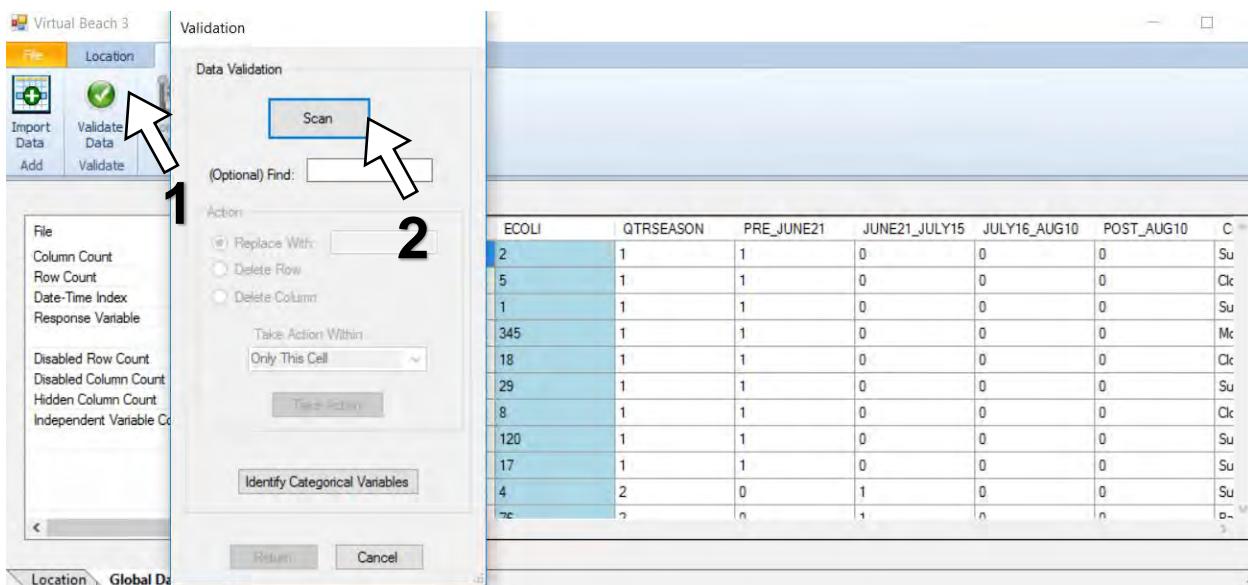
A.4. The data table will open in **Virtual Beach 3**.

 **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.



The screenshot shows the Virtual Beach 3 interface. The left panel displays dataset statistics: Column Count (124), Row Count (287), Date-Time Index (DATETIME), Response Variable (ECOLI), and Independent Variable Count (122). The right panel shows a "Global Datasheet" with columns: DATETIME, ECOLI, QTRSEASON, PRE_JUNE21, JUNE21_JULY15, JULY16_AUG10, and POST_AUG10. The data includes rows for dates from 5/21/2009 to 8/24/2009, with ECOLI values ranging from 1 to 345. The "ECOLI" column is highlighted in blue, indicating it is the response variable.

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**.



The screenshot shows the Virtual Beach 3 interface with the "Validation" dialog box open. Step 1 highlights the "Validate Data" icon in the toolbar. Step 2 highlights the "Scan" button in the validation dialog. The dialog also includes fields for "Find" and "Action" (Replace With, Delete Row, Delete Column), "Take Action Within" (Only This Cell), and "Identify Categorical Variables". The background shows a "Global Datasheet" with the same E. coli data as the previous screenshot.

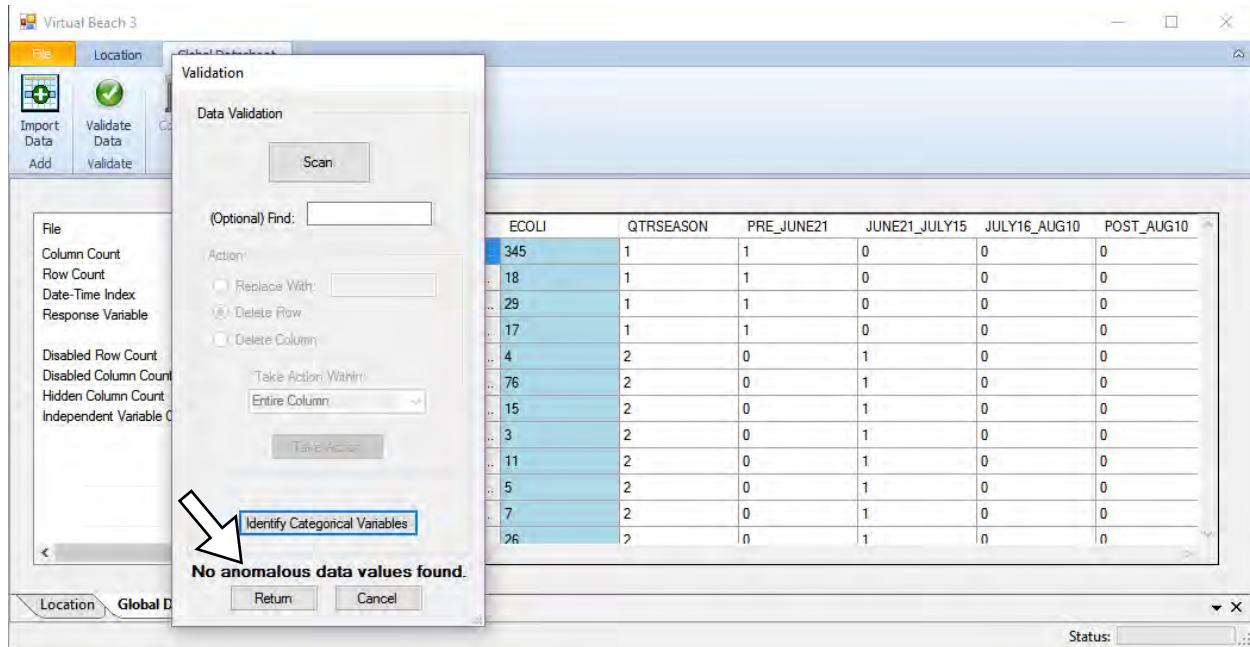
A.6. In this example, the “CLOUDCOV_qual” column is flagged because the values are text, or non-numeric. 1. Click the radio button next to **Delete Column**. Under Take Action Within make sure **Only This Column** is selected. 2. Click **Take Action**.

The screenshot shows the SPSS Data Editor with a validation dialog box open. The validation dialog has several options: 'Replace With:' (radio button), 'Delete Row' (radio button), and 'Delete Column' (radio button, which is selected). Below these is a dropdown menu for 'Take Action Within' with 'Only This Column' selected. A large callout arrow points from the 'Delete Column' radio button to the 'CLOUDCOV_Qual' column in the main data table. Step 1 is labeled '1' near the 'Delete Column' radio button. Step 2 is labeled '2' near the 'Take Action' button. The data table shows columns for QTRSEASON and various dates, followed by the CLOUDCOV_Qual column which contains text values like 'Sunny' and 'Cloudy'.

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**.

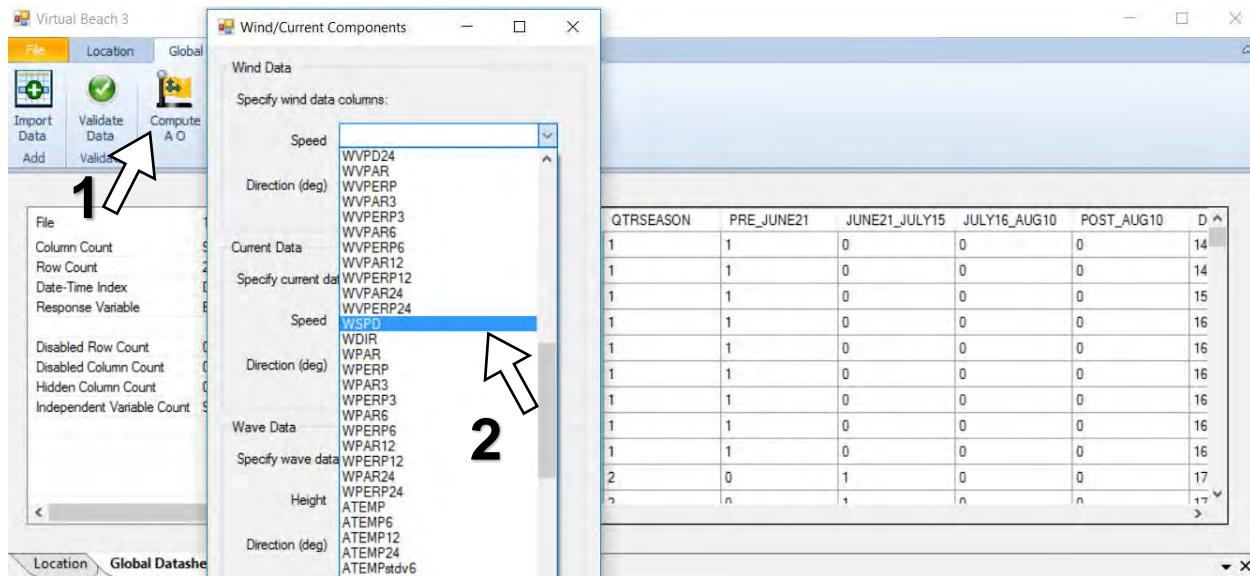
The screenshot shows the SPSS Data Editor with a validation dialog box open. The validation dialog has three radio buttons: 'Replace With:', 'Delete Row' (which is selected), and 'Delete Column'. Below these is a dropdown menu for 'Take Action Within' with four options: 'Only This Row', 'Only This Row' (selected), 'Entire Column' (highlighted with a callout), and 'Entire Sheet'. A large callout arrow points from the 'Entire Column' option to the TRIB6 column in the main data table. Step 1 is labeled '1' near the 'Delete Row' radio button. Step 2 is labeled '2' near the 'Take Action' button. The data table includes columns for RRRAIN24 through RRRAIN144, followed by the TRIB6 column which contains numeric values.

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.

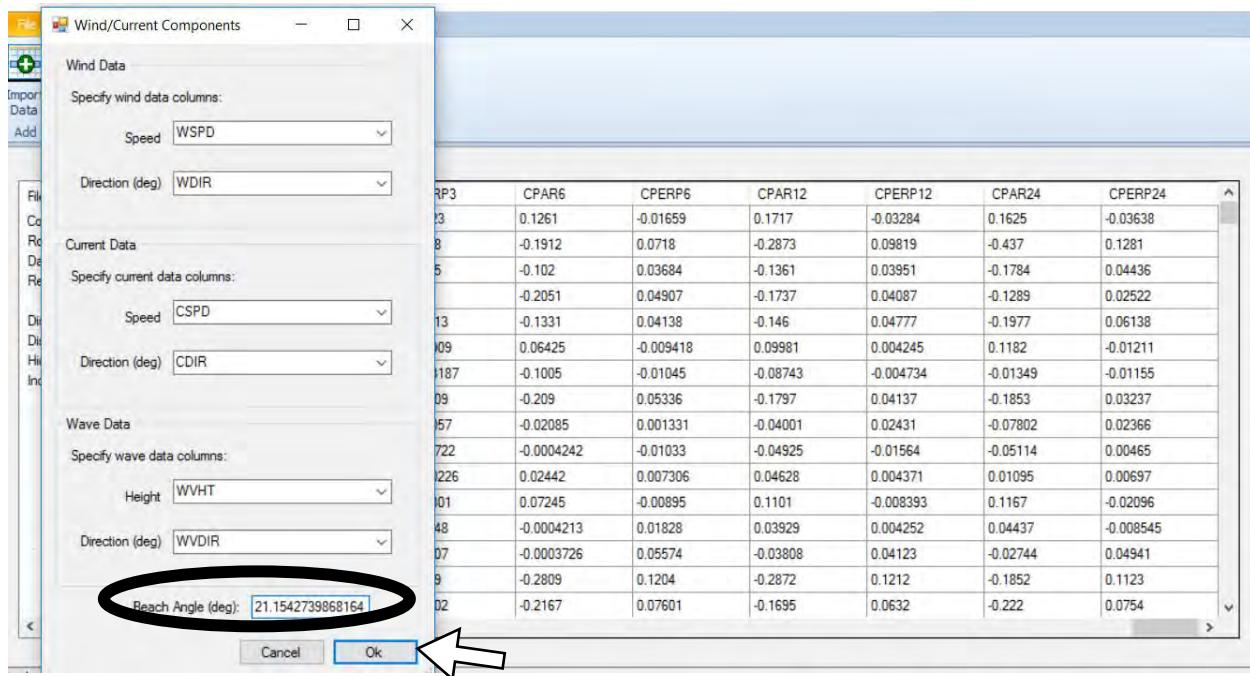


B. Process wind and current data

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_O_comp: on-shore wave height

The screenshot shows the 'Global Datasheet' interface. On the left, there's a sidebar with 'File', 'Location', and 'Global Datasheet' tabs, and icons for Import Data, Validate Data, Compute A/O, Manipulate, Transform, and Go To Model. Below the sidebar, a table has the following information:

File	1-VB-Training-Data.	CPERP24	WindA_comp[WDII]	WindO_comp[WDI]	CurrentA_comp[CD]	CurrentO_comp[CD]	WaveA_comp[WVI]	WaveO_comp[WV]
Column Count	126	0.12638	5.224	1.006	0.09794	0.01167	0.2734	0.2603
Row Count	281	0.1281	-0.2055	-0.009418	0.09819	-0.437	0.1281	-0.1104
Date-Time Index	DATETIME	0.04436	-0.8499	-2.613	-0.07477	-0.03144	-0.07845	-0.06703
Response Variable	ECOLI	0.02522	-8.08	-3.52	-0.2327	-0.06925	-0.6112	0.01645
Disabled Row Count	0	0.06138	-4.677	-1.274	-0.1248	-0.04007	-0.3093	0.0552
Disabled Column Count	6	-0.01211	-1.941	0.4918	0.0493	0.02722	-0.01497	0.01758
Hidden Column Count	0	-0.01155	0.04079	-2.718	-0.0708	-0.01735	-0.1998	-0.002227
Independent Variable Count	118	0.03237	-1.62	-2.762	-0.2096	-0.04655	-0.3067	0.06824

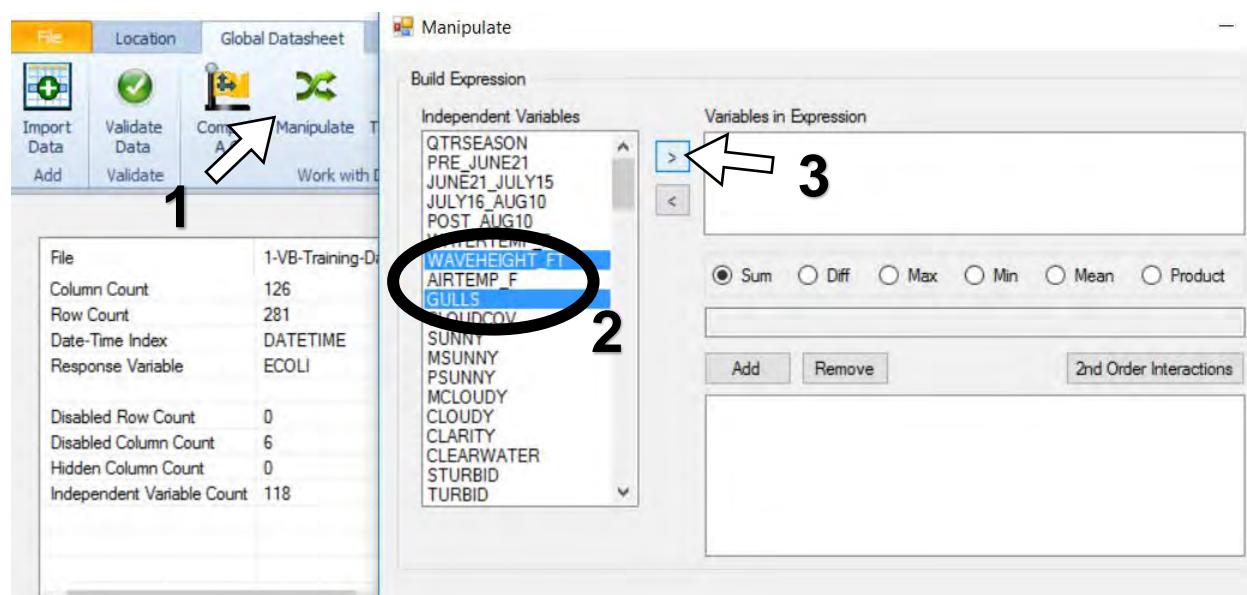
C. Combine two or more predictive variables

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.

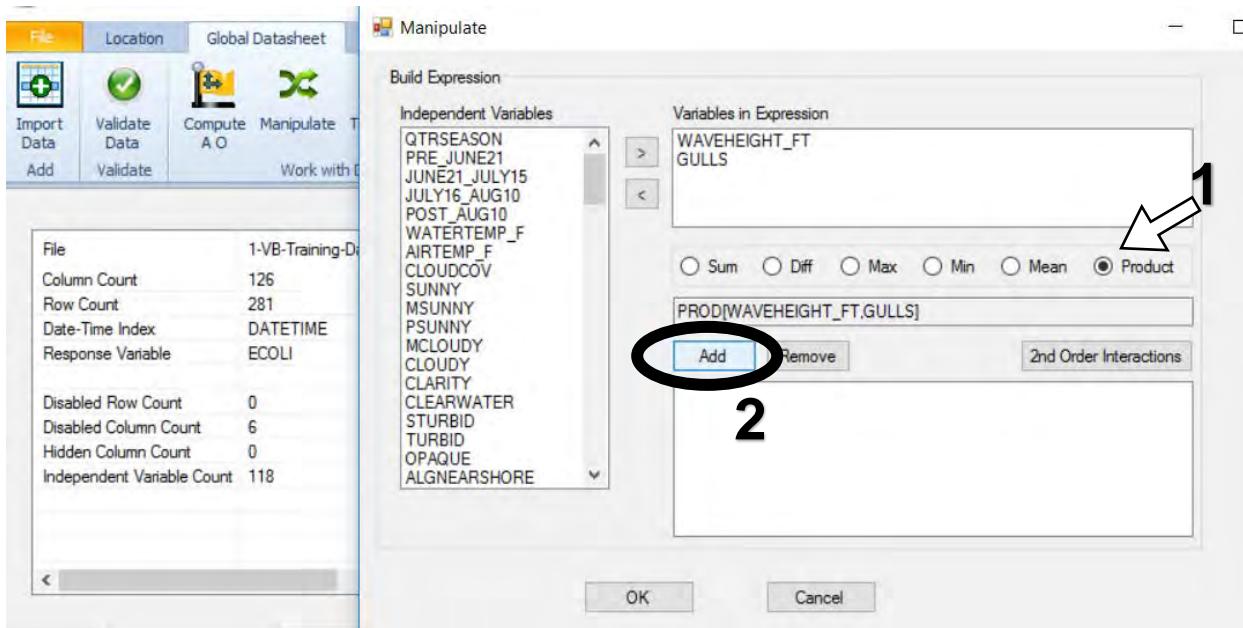
Combined Categories: Some variables are either yes or no. The 0 is “no” and the 1 is “yes”. In situations where binary variables represent successive categories of some qualitative variable, like visually-observed water clarity, it may be useful to combine them into a single binary variable by **summing** 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.

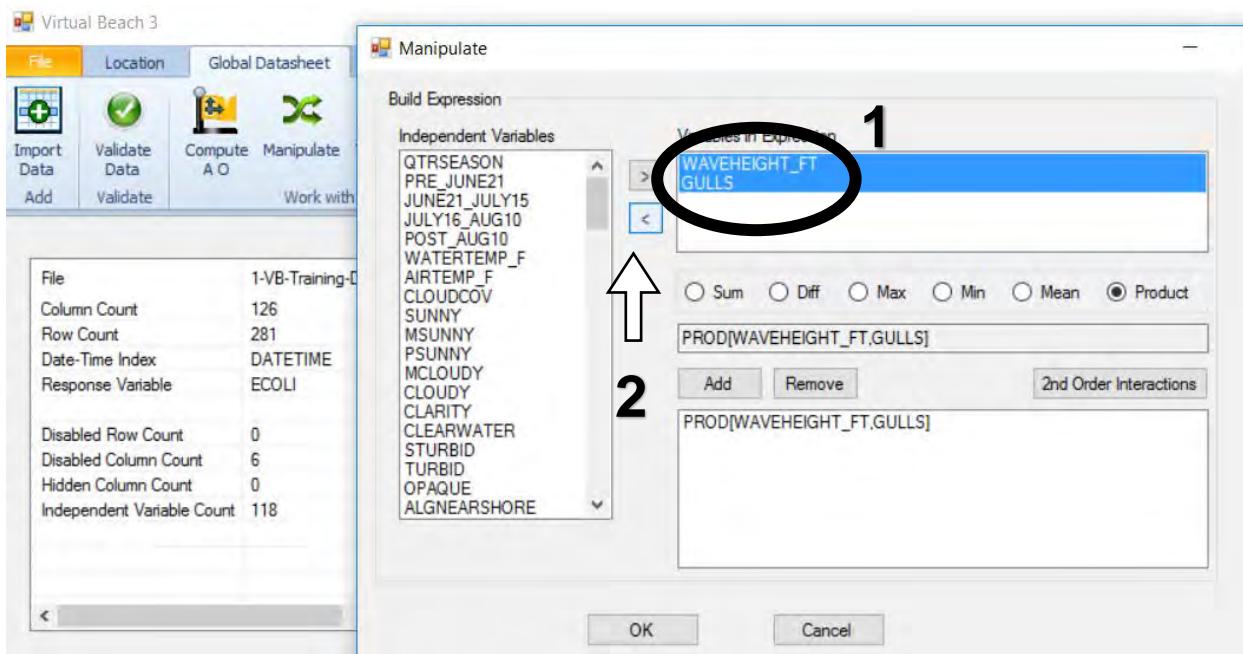
- C.1. First create an interaction term by multiplying two variables together. **1.** Click the **Manipulate** icon. **2.** In the pop-up window, ctrl-select WAVEHEIGHT_FT and GULLS. **3.** Click the right-arrow “>” button.



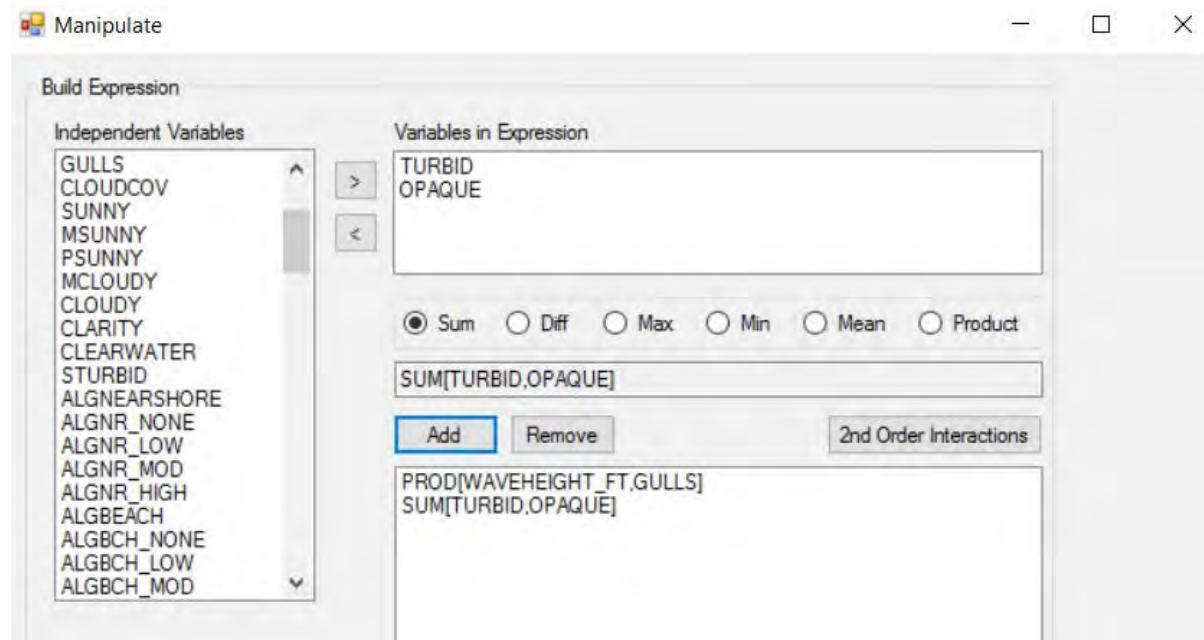
C.2. 1. Click the radio button next to **Product**. 2. Click the **Add** button.



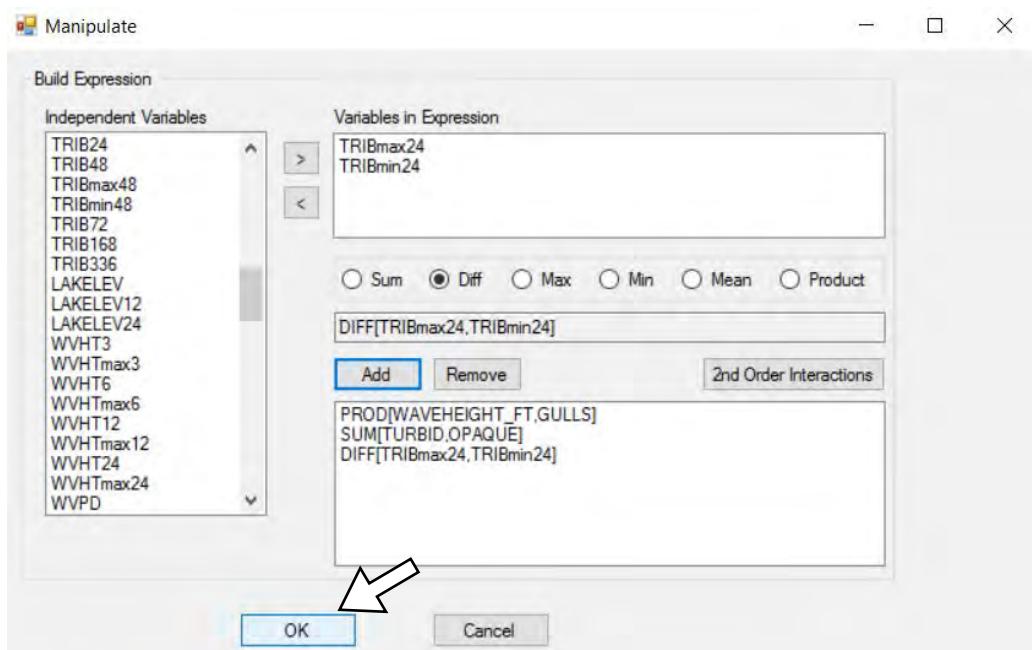
C.3. This creates an interaction term that may describe more accurately how wave height influences the number of gulls on the beach. 1. Shift-select WAVEHEIGHT_FT and GULLS. 2. Click the left-arrow “<” button to move them back to the main list.



C.4. Repeat the steps in C.2 and C.3 to create the expression of combined categories TURBID and OPAQUE using the **Sum** radio-button. This combined expression creates a variable where a visual observation of either TURBID or OPAQUE water receives a value of 1.



C.5. Repeat the steps in C.2 and C.3 to create the change in flow variable with TRIBmax24 and TRIBmin24 using the **Diff** radio-button. This approximates whether and to what extent the previous 24 hours of tributary discharge has been constant or varied a lot. Other manipulations can be added as needed. Click "OK" when complete.



C.6. Scroll to the far-right end of the table to see any new columns added through this process.

File	1-VB-Training-Data..	01	CurrentA_comp[CD]	CurrentO_comp[CD]	WaveA_comp[WVI]	WaveO_comp[V]	PROD[WAVEHEIG]	SUM[TURBID,OPA]	DIFF[TRIBmax24,T]	V
Column Count	129		0.09794	0.01167	0.2734	0.2603	600	0	36	
Row Count	281		-0.1678	-0.0771	-0.3524	-0.1104	2700	1	56	
Date-Time Index	DATETIME		-0.07477	-0.03144	-0.07845	-0.06703	375	0	23	
Response Variable	ECOLI		-0.2327	-0.06925	-0.6112	0.01645	45	1	156	
Disabled Row Count	0		-0.1248	-0.04007	-0.3093	0.0552	125	0	142	
Disabled Column Count	6		0.0493	0.02722	-0.01497	0.01758	137.5	0	11	
Hidden Column Count	0		-0.0708	-0.01735	-0.1998	-0.002227	200	0	21	
Independent Variable Count	121		-0.2096	-0.04655	-0.3067	0.06824	400	1	50	
			0.04153	0.01672	0.07274	0.07508	50	0	37	
			-0.004248	-0.02261	-0.09014	0.02469	8	0	20	

D. Transform variables

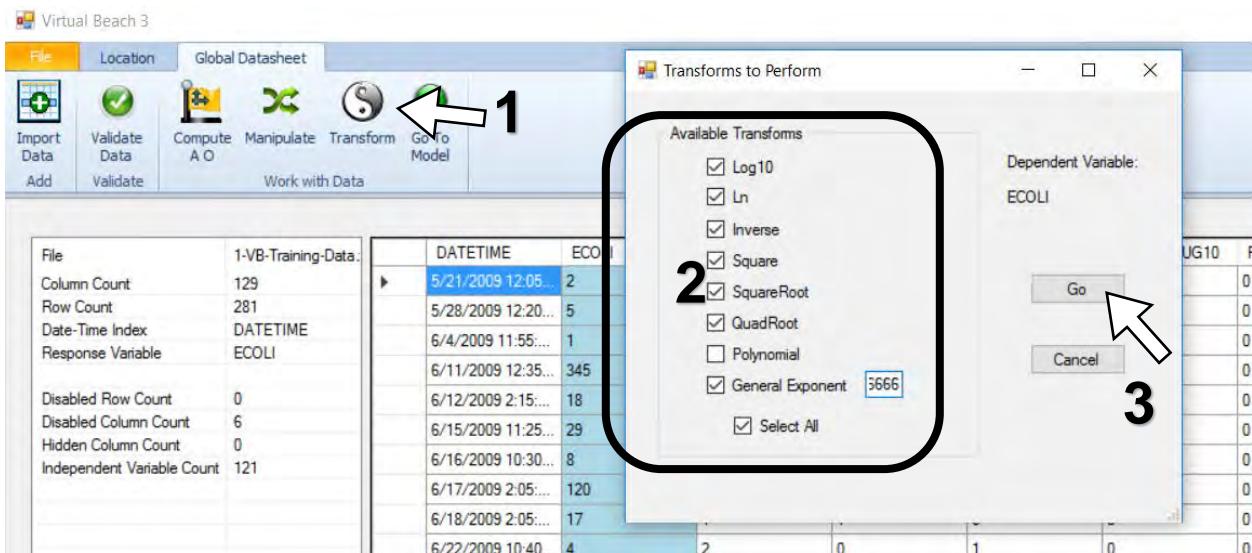
D.1. Right-click on the “ECOLI” column header and select Transform > Log10.

To build a usable nowcast model, bacteria counts must be transformed. Log10 is a common transformation for microbial concentrations.

File	1-VB-Training-Data..	DATETIME	ECOLI	Transform	Log10	Y15	JULY16_AUG10	POST_AUG10	V
Column Count	129	5/21/2009 12:05...	2			0	0	6	
Row Count	281	5/28/2009 12:20...	5			0	0	5	
Date-Time Index	DATETIME	6/4/2009 11:55:...	1			0	0	6	
Response Variable	ECOLI	6/11/2009 12:35...	345			0	0	5	
Disabled Row Count	0	6/12/2009 2:15:...	18			0	0	6	
Disabled Column Count	6	6/15/2009 11:25...	29			0	0	6	
Hidden Column Count	0	6/16/2009 10:30...	8			0	0	6	
Independent Variable Count	121	6/17/2009 2:05:...	120			0	0	6	
		6/18/2009 2:05:...	17			0	0	6	
		6/22/2009 10:40...	4			0	0	7	

D.2. In addition to transforming the response variable, transforming explanatory variables can significantly improve model fit. 1. Click the **Transform** icon. 2. Check all options, EXCEPT Polynomial, and type 0.6666 next to **General Exponent**. 3. Click **Go**.

 As of July 2017, the polynomial transformation causes problems in the Virtual Beach program. Future updates will address this issue.



D.3. A pop-up will open listing all of the optional transformations for each explanatory variable. Those in black represent the transformation with the best correlation (Pearson's coefficient) with the response variable LOG(ECOLI).

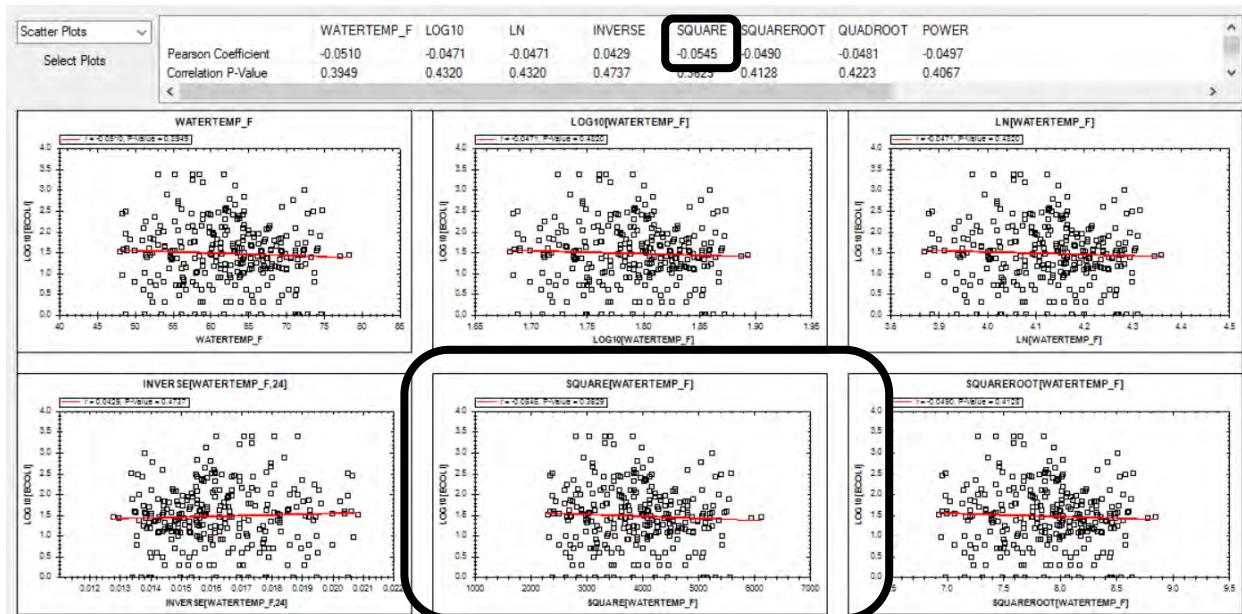
Dependent Variable: LOG10[ECOLI]		
	Variable	Transform
▶	QTRSEASON	none
	QTRSEASON	LOG10[QTRSEASON]
	QTRSEASON	LN[QTRSEASON]
	QTRSEASON	INVERSE[QTRSEASON,0.5]
	QTRSEASON	SQUARE[QTRSEASON]
	QTRSEASON	SQUAREROOT[QTRSEASON]
	QTRSEASON	QUADROOT[QTRSEASON]
	QTRSEASON	POWER[QTRSEASON,0.6666]
	WATERTEMP_F	none
	WATERTEMP_F	LOG10[WATERTEMP_F]
	WATERTEMP_F	LN[WATERTEMP_F]
	WATERTEMP_F	INVERSE[WATERTEMP_F,24]
	WATERTEMP_F	SQUARE[WATERTEMP_F]
	WATERTEMP_F	SQUAREROOT[WATERTEMP_F]

D.4. Right-clicking in any cell to the left of a variable will enable you to view scatter plots of all the transformations of that variable versus LOG(ECOLI). For example, right-click to the left of the variable WAVEHEIGHT_FT.

The screenshot shows the 'Data Transformation' dialog box. The 'Dependent Variable' is set to 'LOG10[ECOLI]'. On the left, there are sections for 'Auto-Select' and 'Threshold Select'. In the main area, a table lists variables and their transformations. The row for 'WATERTEMP_F' has a red arrow pointing to the 'View Plots' button in the 'Transform' column.

Variable	Transform
QTRSEASON	none
QTRSEASON	LOG10[QTRSEASON]
QTRSEASON	LN[QTRSEASON]
QTRSEASON	INVERSE[QTRSEASON,0.5]
QTRSEASON	SQUARE[QTRSEASON]
QTRSEASON	SQUAREROOT[QTRSEASON]
QTRSEASON	QUADROOT[QTRSEASON]
QTRSEASON	POWER[QTRSEASON,0.6666]
WATERTEMP_F	none
WATERTEMP_F	LOG10[WATERTEMP_F]
WATERTEMP_F	LN[WATERTEMP_F]
WATERTEMP_F	SQUARE[WATERTEMP_F]

D.5. Note that in this case the best transformation, in terms of Person's r, is square transformation. The scatter plot confirms this selection. Close the window to return to the list of transformation options.



If you decide to select an alternative best transformation for a given variable, simply click on that row. When you are finished, click **OK**.

D.6. New columns are added if the newly-transformed variable had a better fit than untransformed original variable. The new columns here are **SQUARE(WATERTEMP_F)**, **QUADROOT(WAVEHEIGHT_FT)**, and **QUADROOT(AIRTEMP_F)**. The original columns are now disabled as indicated by red text. Disabled columns will NOT be used in the model. Save your project file. You can now move onto the next module, “Building an MLR Model”.

File	1-VB-Training-Data.
Column Count	226
Row Count	281
Date-Time Index	DATETIME
Response Variable	LOG10(ECOLI)
Disabled Row Count	0
Disabled Column Count	97
Hidden Column Count	1
Independent Variable Count	127

WATERTEMP_F	SQUARE[WATERTEMP_F]	WAVEHEIGHT_FT	QUADROOT[WAVEHEIGHT_FT]	AIRTEMP_F	QUADROOT[AIRTEMP_F]
62.1	3965	12	1.861	7.1	3.055
55.4	3069	12	1.861	60.1	2.784
65.1	4238	1.5	1.107	62.8	2.815
56.1	3147	1.5	1.107	61	2.795
66.1	4369	0.5	0.8409	57.5	2.754
61.5	3782	0.5	0.8409	65.3	2.843
68.1	4638	1	1	71.7	2.91
64.3	4134	2	1.189	71.3	2.906
68.2	4651	1	1	64.3	2.832