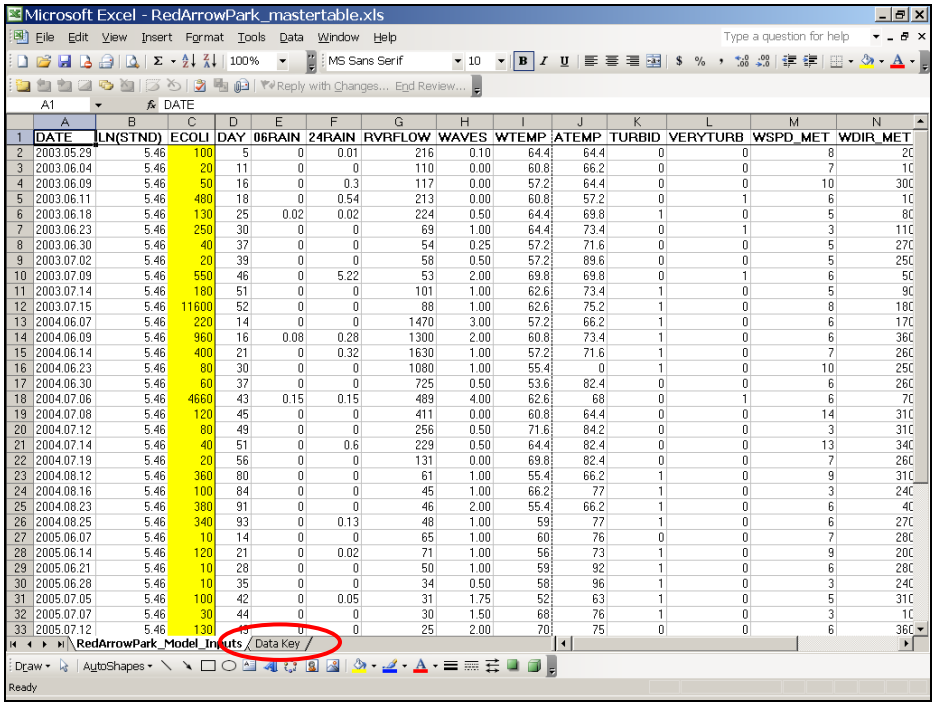
Virtual Beach 1.0
Learning Module I – *Model Building*

In this module you will learn how to:

A. Format and import a data table for use in *Virtual Beach 1.0*
B. Evaluate data using scatter plots
C. Transform variables to better meet Ordinary Least Squares assumptions
D. Exclude unwanted observations and variables
E. Check for multicollinearity (non-independence) among explanatory variables
F. Convert Wind Speed and Direction into "longshore" and "onshore" components
G. Create "interaction terms" (combined variables)

A. Format/Import Data. Data for this module are available for download at http://dnr.wi.gov/org/es/science/contaminants/

A.1. Open the file *RedArrowPark_Mastertable.xls* in MS Excel. Click the "Data_Key" tab (circled below) to view data descriptions, sources, etc.

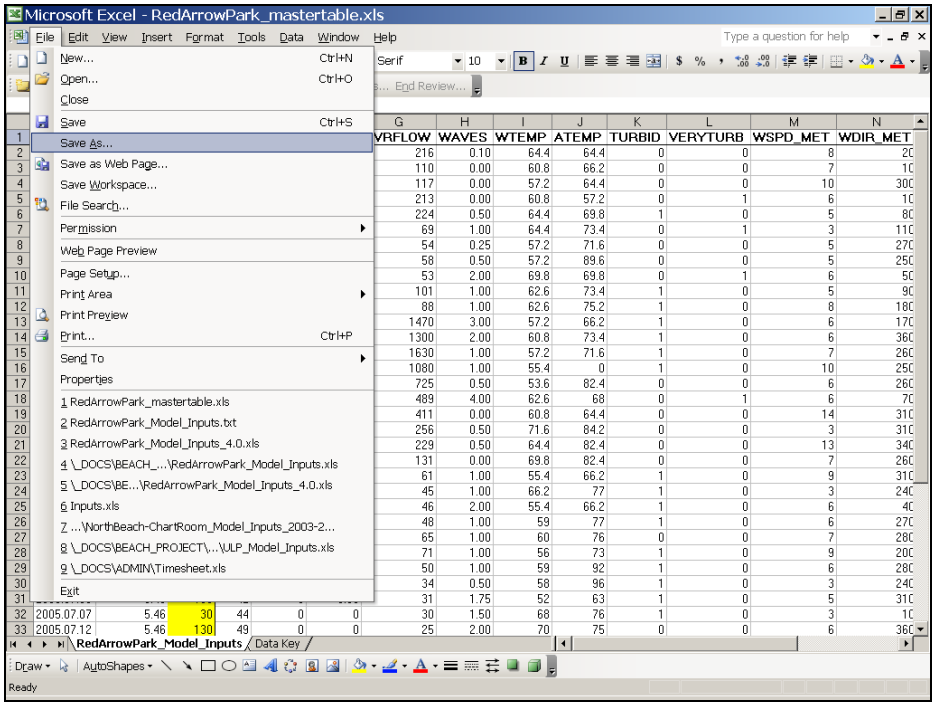


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	DATE	LN(STND)	ECOLI	DAY	06RAIN	24RAIN	RVRFLOW	WAVES	WTEMP	ATEMP	TURBID	VERYTURB	WSPD_MET	WDIR_MET
2	2003.05.29	5.46	100	5	0	0.01	216	0.10	64.4	64.4	0	0	8	20
3	2003.06.04	5.46	20	11	0	0	110	0.00	60.8	66.2	0	0	7	10
4	2003.06.09	5.46	50	16	0	0.3	117	0.00	57.2	64.4	0	0	10	300
5	2003.06.11	5.46	480	18	0	0.54	213	0.00	60.8	57.2	0	1	6	10
6	2003.06.18	5.46	130	25	0.02	0.02	224	0.50	64.4	69.8	1	0	5	80
7	2003.06.23	5.46	250	30	0	0	69	1.00	64.4	73.4	0	1	3	110
8	2003.06.30	5.46	40	37	0	0	54	0.25	57.2	71.6	0	0	5	270
9	2003.07.02	5.46	20	39	0	0	58	0.50	57.2	89.6	0	0	5	250
10	2003.07.09	5.46	550	46	0	5.22	53	2.00	69.8	69.8	0	1	6	50
11	2003.07.14	5.46	180	51	0	0	101	1.00	62.6	73.4	1	0	5	90
12	2003.07.15	5.46	11600	52	0	0	88	1.00	62.6	75.2	1	0	8	180
13	2004.06.07	5.46	220	14	0	0	1470	3.00	57.2	66.2	1	0	6	170
14	2004.06.09	5.46	960	16	0.08	0.28	1300	2.00	60.8	73.4	1	0	6	360
15	2004.06.14	5.46	400	21	0	0.32	1630	1.00	57.2	71.6	1	0	7	260
16	2004.06.23	5.46	80	30	0	0	1080	1.00	55.4	0	1	0	10	250
17	2004.06.30	5.46	60	37	0	0	725	0.50	53.6	82.4	0	0	6	260
18	2004.07.06	5.46	4660	43	0.15	0.15	489	4.00	62.6	68	0	1	6	70
19	2004.07.08	5.46	120	45	0	0	411	0.00	60.8	64.4	0	0	14	310
20	2004.07.12	5.46	80	49	0	0	256	0.50	71.6	84.2	0	0	3	310
21	2004.07.14	5.46	40	51	0	0.6	229	0.50	64.4	82.4	0	0	13	340
22	2004.07.19	5.46	20	56	0	0	131	0.00	69.8	82.4	0	0	7	260
23	2004.08.12	5.46	360	80	0	0	61	1.00	55.4	66.2	1	0	9	310
24	2004.08.16	5.46	100	84	0	0	45	1.00	66.2	77	1	0	3	240
25	2004.08.23	5.46	380	91	0	0	46	2.00	55.4	66.2	1	0	6	40
26	2004.08.25	5.46	340	93	0	0.13	48	1.00	59	77	1	0	6	270
27	2005.06.07	5.46	10	14	0	0	65	1.00	60	76	0	0	7	280
28	2005.06.14	5.46	120	21	0	0.02	71	1.00	56	73	1	0	9	200
29	2005.06.21	5.46	10	28	0	0	50	1.00	59	92	1	0	6	280
30	2005.06.28	5.46	10	35	0	0	34	0.50	58	96	1	0	3	240
31	2005.07.05	5.46	100	42	0	0.05	31	1.75	52	63	1	0	5	310
32	2005.07.07	5.46	30	44	0	0	30	1.50	68	76	1	0	3	10
33	2005.07.12	5.46	130	49	0	0	25	2.00	70	75	0	0	6	360

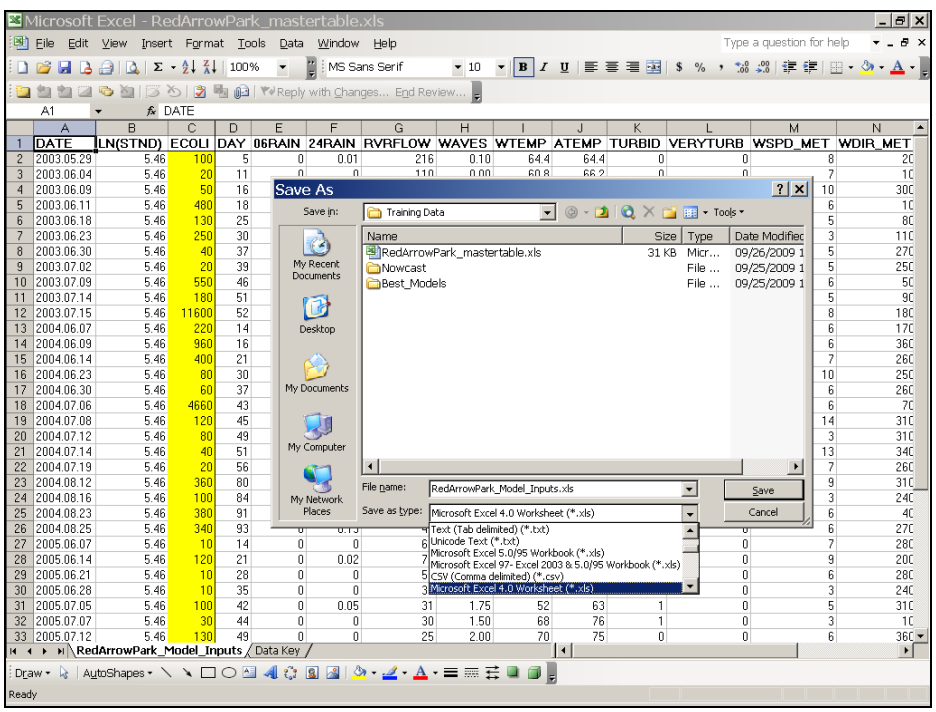
Virtual Beach 1.0 reads only Excel 4.0 files or tab-delimited text (*.txt) files. Newer .xls formats must be converted to Excel 4.0 or tab-delimited text using "Save As…" (see Step A.2 on pg. 2).

Virtual Beach 1.0 automatically reads the third column of an input data table (**column C**) as the eventual model's response variable—make sure this is the variable you are trying to predict.

A.2. Return to the “RedArrorPark_Model_Inputs” worksheet (tab). From the Excel menu in the upper-left corner, click “File” > “Save As…”

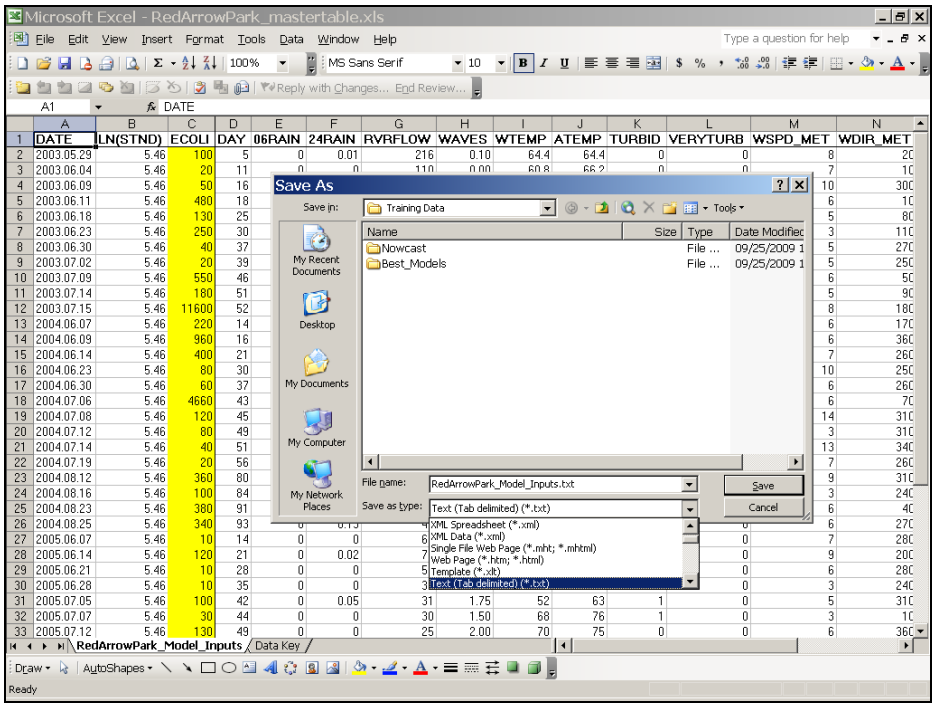


A.3. From the “Save as Type:” pull-down, choose “Microsoft Excel 4.0 Worksheet” and re-name the file “RedArrowPark_Model_Inputs”. **Close Excel**.



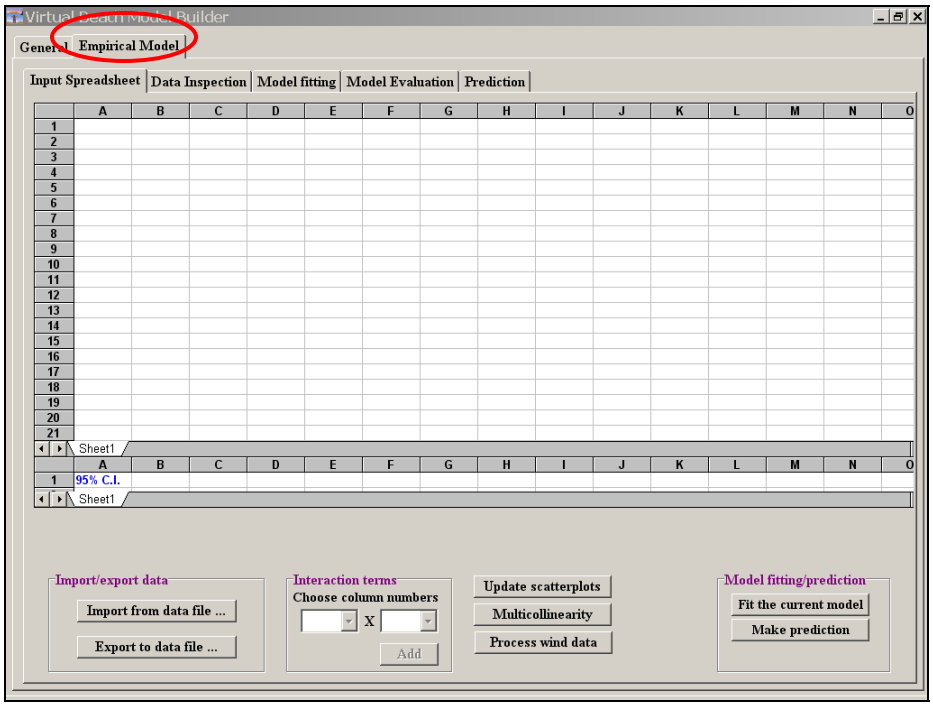
Excel 4.0 files can have only 1 worksheet (tab)—be sure to save the right one.

Note: If you are working with a version of Excel that does not support saving to Excel 4.0, save as a comma-delimited text (.txt) file instead. From the "Save as Type:" pull-down, choose "Text (Tab delimited)." A pop-up window will appear; click "OK" to save only the active worksheet. **Close Excel**.

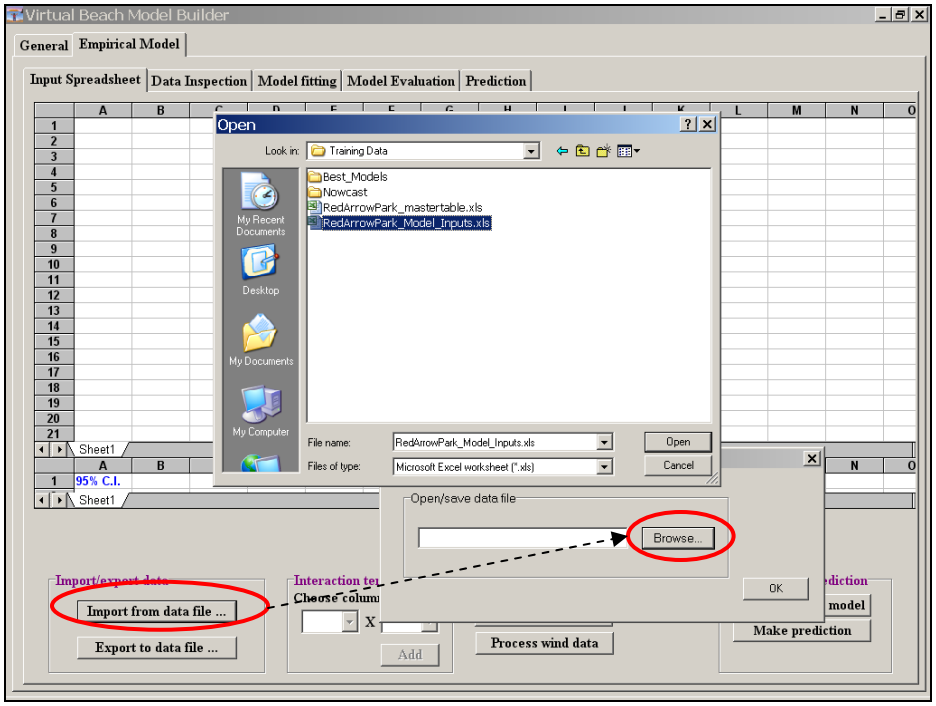


A.4. Open *Virtual Beach*.

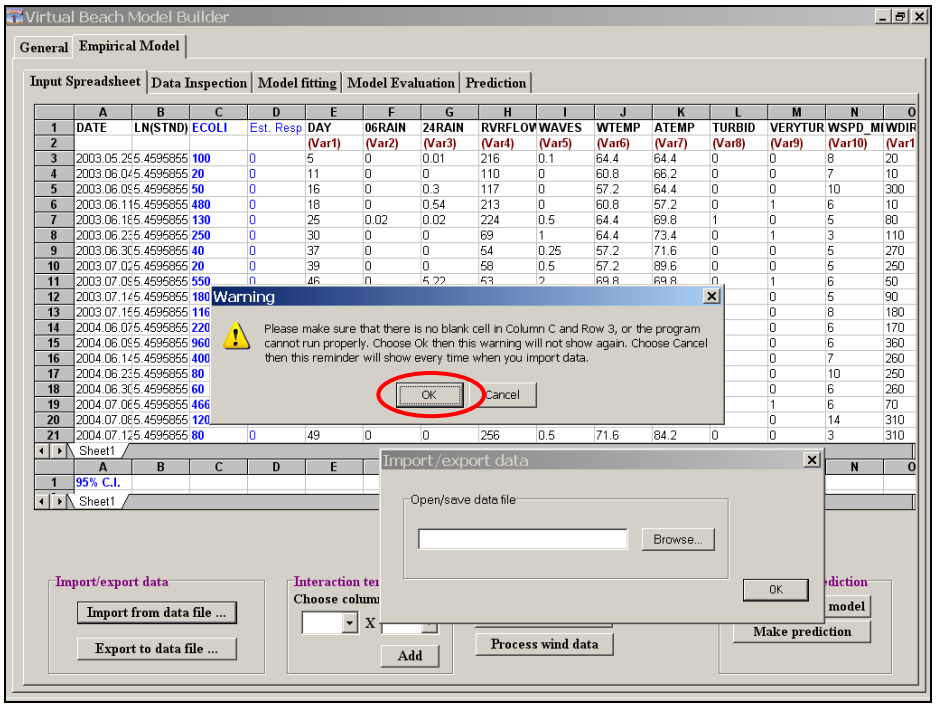
A.5. Click the “Empirical Model” tab, and then “Input Spreadsheet” tab. (This opens two blank spreadsheets: The large spreadsheet on top is for input data. The 1-row spreadsheet below is for making single-case predictions, after a model has been “fit” with the input data.)



A.6. Click the “Import from data file…” button, then “Browse.” Navigate to the input data file you created (*RedArrorPark_Model_Inputs*). Click “Open,” then “OK.”



A.7. A Warning popup will open—click “OK”.
(*VB 1.0* creates a new column (**D**) called “Est. Resp.,” which will later be filled with the predicted values of your response variable. Also adds a new row (2) with ID’s for each explanatory variables: (Var1), (Var2), etc…)



Virtual Beach 1.0 <u>will not import spreadsheets</u> (or tab-delimited text files) <u>if they are simultaneously open in Excel</u>. If the table does not load, first make sure that it is not still open in Excel.

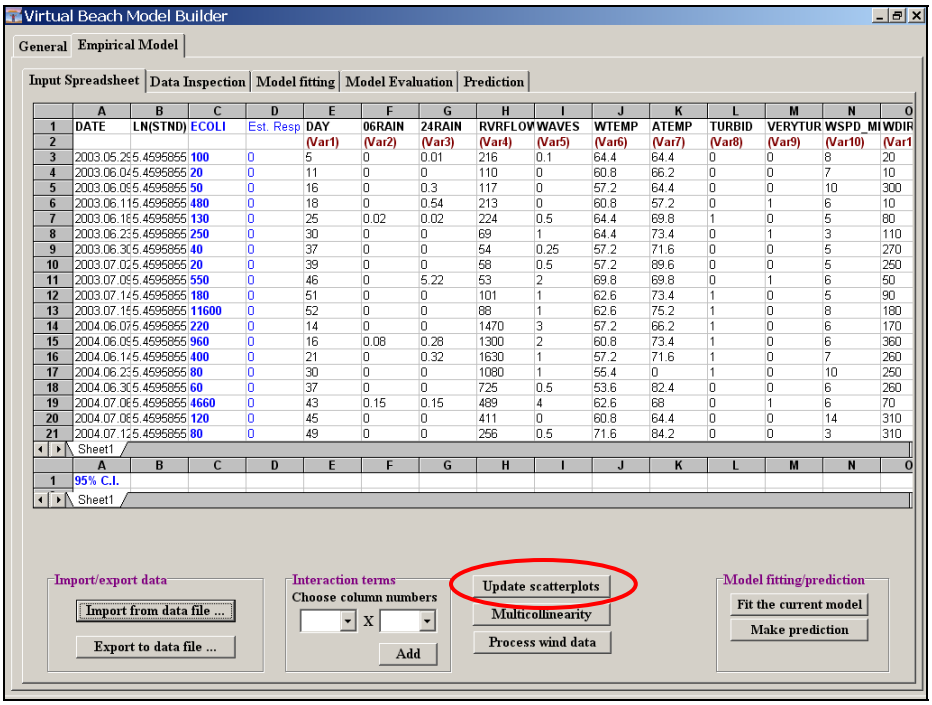
A.8. Practice navigating around the data table.

- Click on any cell and then use the arrow keys to move up, down, or sideways – one cell at a time
- Use the right/left slide-bar on the bottom of the spreadsheet to move sideways.
- Use the up/down slide-bar on the lower right hand side of the spreadsheet to move up or down.

B. Evaluate Data using Scatter Plots

- Plotting values of a potential explanatory variable (e.g., water temp) against the values of the response variable (i.e., *E. coli*) can provide preliminary information on whether or not there is any correlation or relationship between the two. Helps you decide whether or not to exclude or modify the variable in question.
- "Scatter plots" can also help you identify "outliers" in the data set, which may reflect real aberrations or errors in measurement.
- Scatter plots also provide an indication as to whether or not such relationships are linear—and therefore, whether ordinary least-square ("OLS") assumptions will be violated. If the actual relationships are *not* linear, model coefficients and model predictions will less likely be accurate. In many cases, it is possible to transform variables in ways that satisfy OLS requirements.

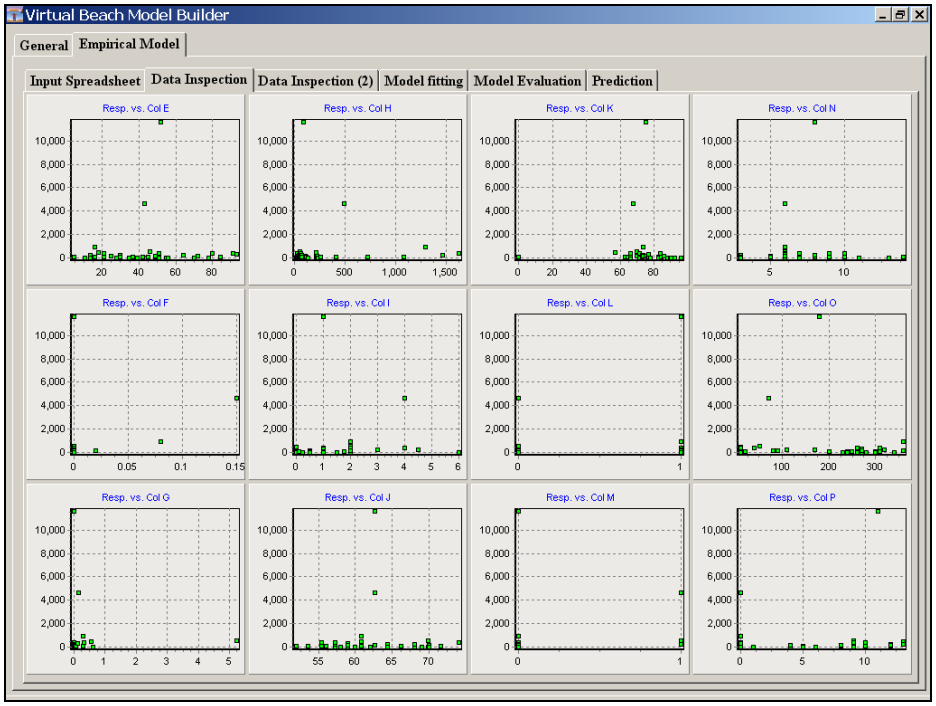
B.1. In the "Input Spreadsheets" tab, click the "Update Scatter Plots" button.



- *Virtual Beach 1.0* generates up to sixteen scatter plots of explanatory variables against the response variable. If you have data to the right of column **AB** in the "Input Spreadsheet" tab these will not be plotted; however, they can still be included in the eventual model.

B.2. The “Data Inspection” tab will open automatically. Evaluate the scatter plots contained in that tab as well as those in the “Data Inspection (2)” tab.

For each plot, the Y axis is the response variable – denoted in the plot title as “Resp.” – in this example *E. coli* in CFU/mL. The X axes correspond to the different explanatory variables – columns in the “Input Spreadsheet” tab – denoted in the plot titles as “Col E,” “Col F,” “Col G,” etc…



B.3. Toggle back and forth between the “Data Inspection” tabs and the “Input Spreadsheet” tab to see which variables correspond to which plots. Note, for example, that “Col L” refers to the categorical variable TURBID, where a value of 1 = “turbid,” and a value of 0 = not “turbid”

B.4. Note the highly-clustered and non-linear pattern of most of the scatter plots.

Generally speaking, you want the scatter plots to resemble “clouds” of data points, suggesting linear relationships (diagonal cylinders), or seemingly random relationships (shapeless). Curved, cone-shaped, or otherwise cyclical patters indicate non-linear relationships between the explanatory variable and the response variable, and thus the potential for violating OLS assumptions. The variable(s) in question may need to be excluded from the model or transformed.

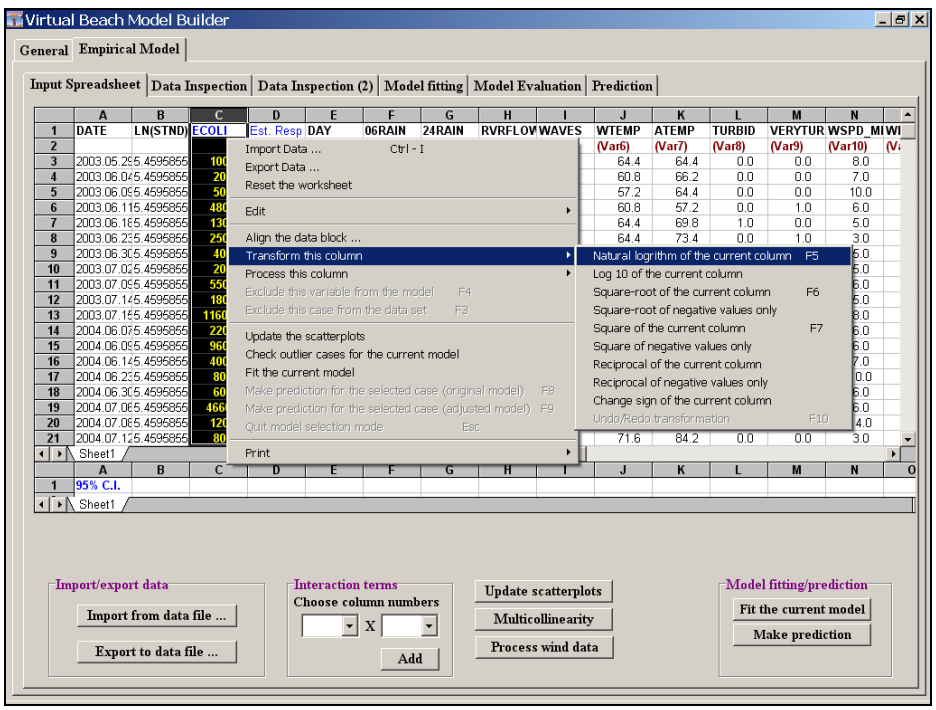
The exception to the above rule of thumb are categorical (0/1) variables, such as TURBID (“Col H”), for which the data points will stack-up on the far ends of the scatter plot; i.e., the X axis values are all 0 or 1.

C. Transform Variables to better meet Ordinary Least Square (OLS) Assumptions. Checking and re-checking scatter plots to make sure relationships are linear is crucial to the success of the model-building effort. Where it appears that OLS assumptions will likely be violated, *Virtual Beach* provides an easy means of transforming / un-transforming variables.

A common transformation used in MLR models for predicting *E. coli* is to take the natural logarithm of the concentration to effectively "straighten" the data, which otherwise will typical exhibit exponential growth under certain environmental conditions.

Another common transformation is to transform antecedent rainfall by taking its square-root. This transformation moderates extreme values associated with large (infrequent) storm events. Taking the square root of streamflow similarly moderates extreme runoff values, particularly in "flashy" watersheds.

C.1. In the "Import Spreadsheet" tab, highlight Column **C** (ECOLI), right click, and select > Transform this column > Natural logarithm of the current column. (Can also use the **F5** key.) Note that the variable name changes to "Ln(ECOLI)."



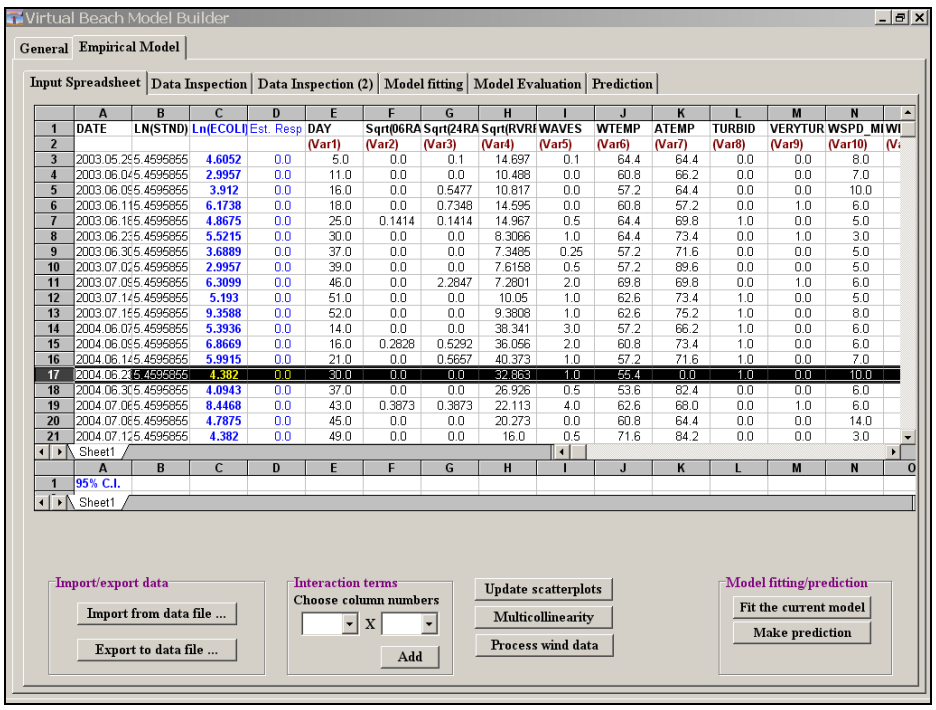
C.2. Transform Columns **F**, **G**, and **H** (06RAIN, 24RAIN, and RVRFLOW) by right clicking and selecting > Transform this column > Square-root of the current column. (Can also use the **F6** key.)

C.3. Return to the “Input Spreadsheet” tab and click the “Update scatterplots” button. Note that the updated scatter plots for the non-categorical variables, such as RVRFLOW (the plot titled “Resp. vs. Col H”), are no longer highly-clustered, and instead resemble diagonal or horizontal clouds of data points.

D. **Exclude Unwanted Observations and Variables.** It is likely that at least some of the potential variables you consider for inclusion in your model will prove to be poor predictors of beach water quality (for any number of reasons). Others may clearly violate OLS assumptions, as described above, even after they are transformed. In addition, individual observations may be "outliers," caused by extreme/atypical conditions, or errors made during data collection or entry. These can reduce the predictive power of the overall model. *Virtual Beach* provides an easy means of excluding / un-excluding both observations and variables.

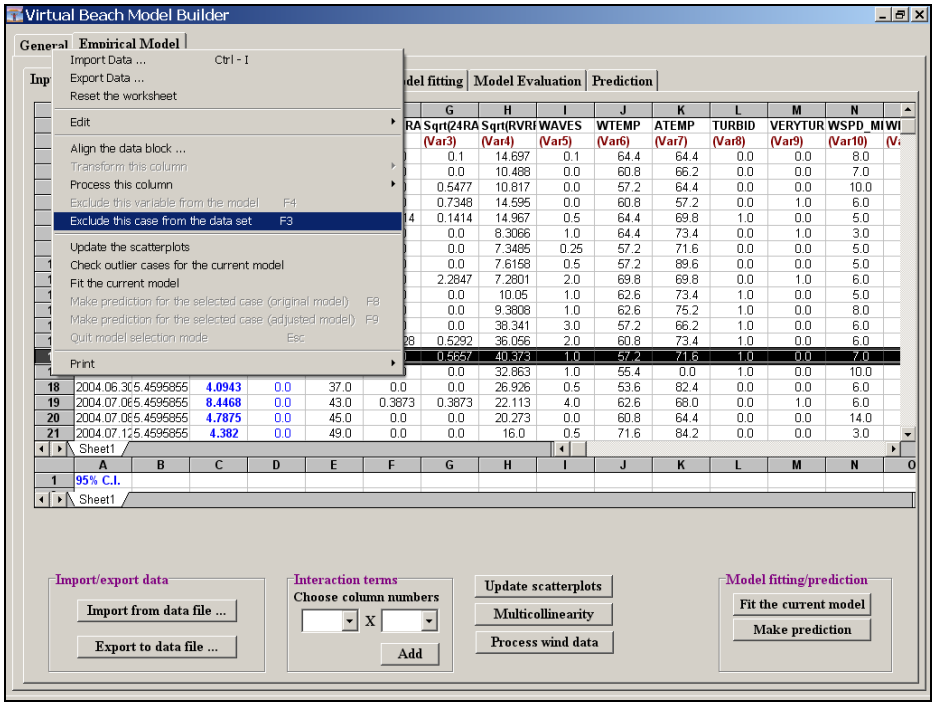
D.1. In the first "Data Inspection" tab (illustrated in Part C.3 above), note the data anomaly indicated in the scatter plot titled "Resp. vs. Column K" (i.e., ATEMP = 0° F).

D.2. Return to the "Input Spreadsheet" tab. Look down the values in Column **K** (ATEMP) until you find the 0 value. Click on the corresponding row number (**17**) to select that observation.

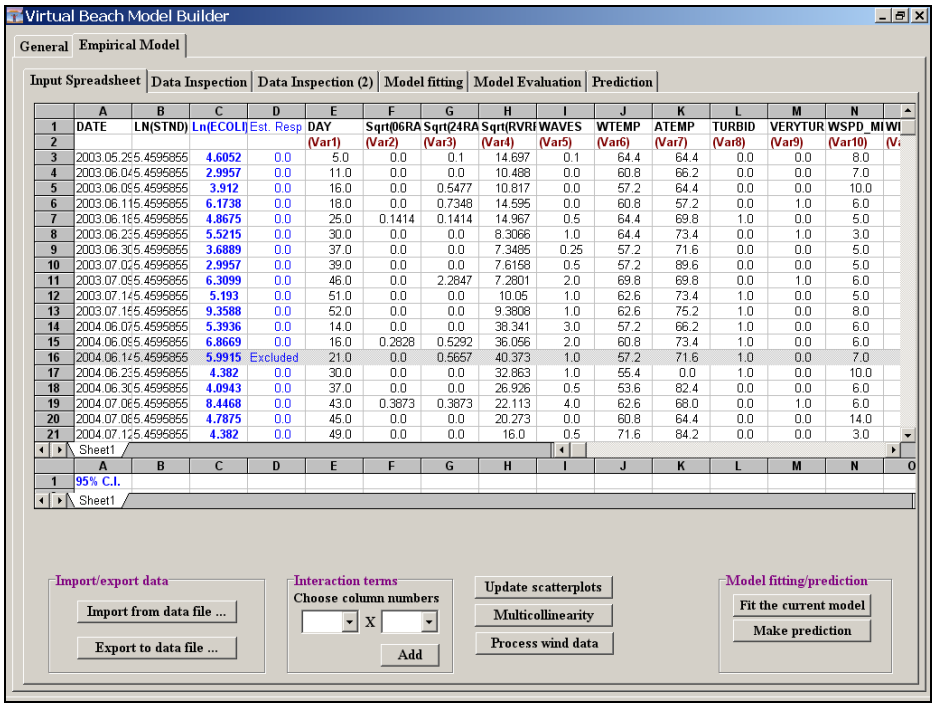


In *VB 1.0* you can select multiple rows the same way you would using MS Excel: Click the first row you want selected. Then (holding the mouse button down) scroll up or down to the last row you want selected and release.

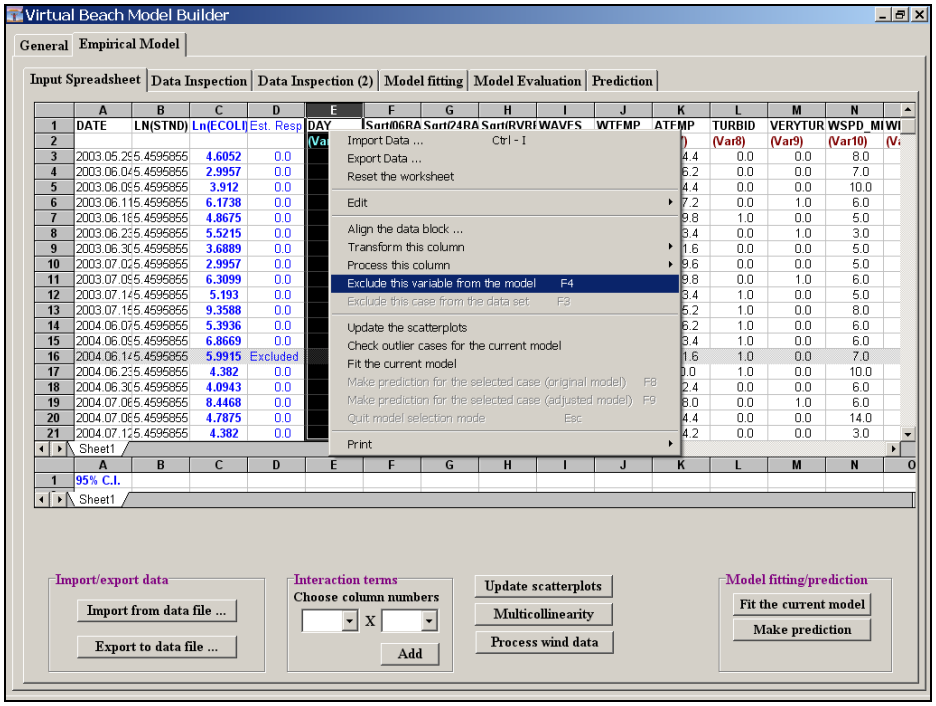
D.3. With row 16 highlighted, right-click with your mouse and select “Exclude the case from the data set” from the menu. (You can also use the **F3** key.)



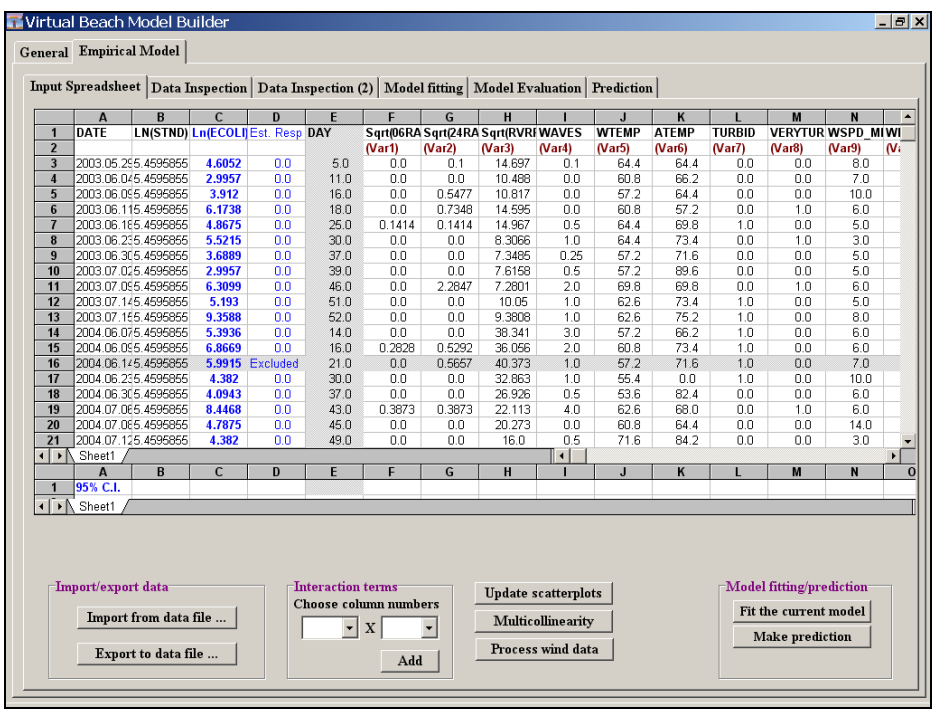
Note that the row turned gray, and the value listed under Est. Resp. (**D**) becomes “Excluded.” <u>Excluded rows will be left out of any model you build</u>. (To re-include them, simply reselect and repeat Step D.3.).



D.4. Exclude the variable DAY. First click on the column letter (**E**), to highlight, then right-click and select "Exclude the variable…" (or hit the **F4** key).



Note that the column turned gray and the identifier in row 2, "(Var1)", disappears. <u>Excluded columns will be left out of any model you build</u>. (To re-include them, simply repeat Step D.4.).

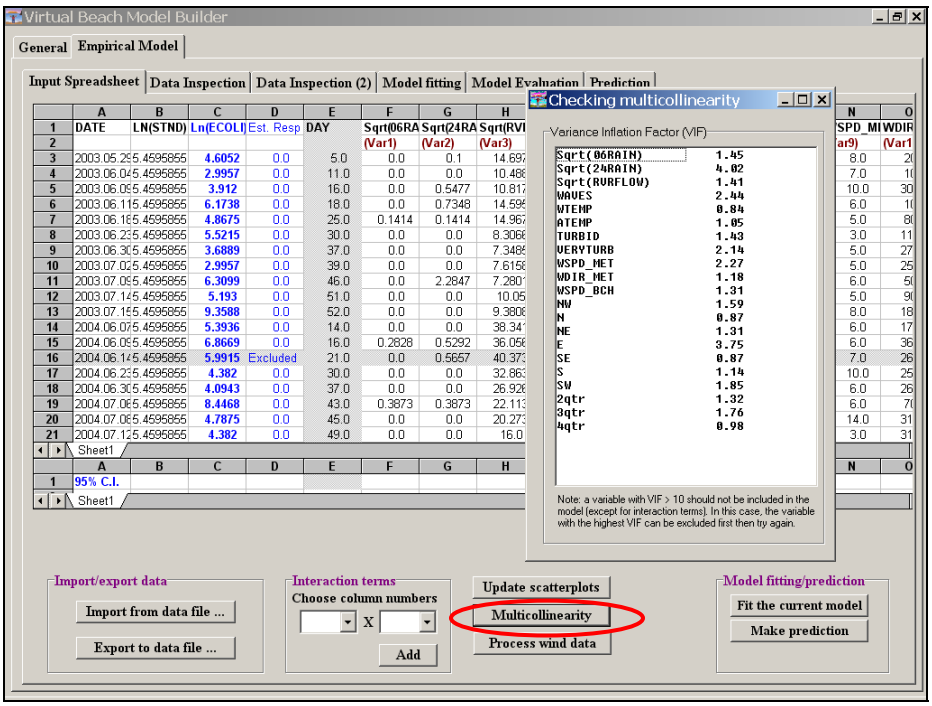


E. Check for Multicollinearity (Non-Independence) among Explanatory Variables

Another assumption of multiple linear regression (MLR) modeling is that individual explanatory variables are independent of one another. That is, they are not correlated with one another, and so do not affect one another's relationship with the response variable. This is obviously not always the case—two or more explanatory variables a model may turn out to positively or negatively correlated; i.e., "collinear."

Severe multicollinearity results in erroneous partial regression coefficients (discussed in Module 2), which are used to estimate the change in *E. coli* expected in response to a one-unit change in a given explanatory variable. As such, the model will not provide information that could otherwise have been used to help identify the most significant factors contributing to contamination events.

E.1. In the "Input Spreadsheet" tab click the "Multicollinearity" button. The window that pops-up will list the Variance Inflation Factors (VIF) for each of the non-excluded explanatory variables. For a given variable, a <u>VIF value > 10 indicates the presence of severe multicollinearity</u>. Note that there are no such values in this dataset.

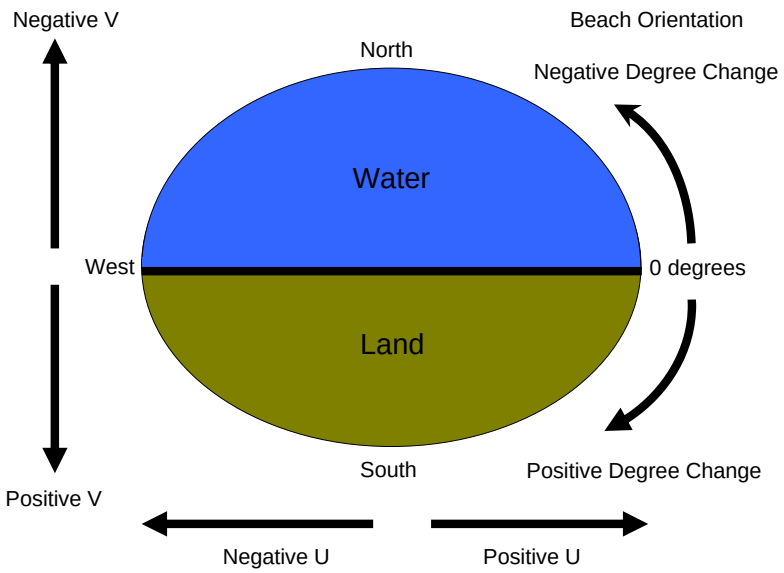


If *Virtual Beach* reports VIF values > 10 you can either exclude the variable(s) in question (see D.4), or create an "Interaction term" (see G.1 below).

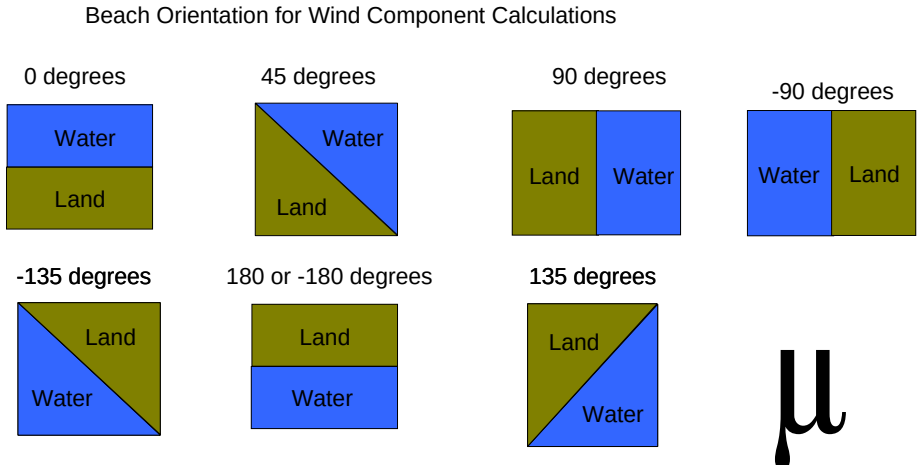
F. Convert Wind Speed and Direction into "Longshore" and "Onshore" components

- While more precise than cardinal directions (e.g., "north," "northeast," etc.), wind direction measured in degrees (0-360) is a problematic variable for regression analysis, since its range of continuous values are not truly sequential. For example, 10° is not less than 350° (nor is it more).

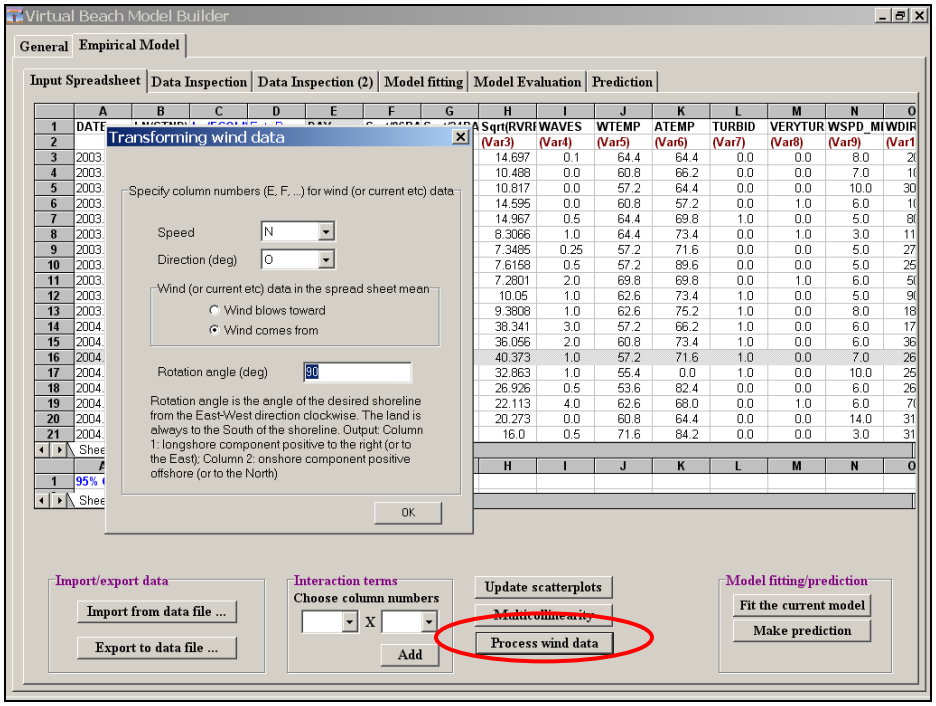
- One way to solve this problem is to combine and re-process wind into separate "Longshore" (U) and "Onshore" (V) components. Facing the water, positive Longshore values occur when wind blows from left to right. Positive Onshore values occur when wind blows toward shore. Higher velocities increase the magnitude of the values (positive or negative) for both components.



- *Virtual Beach* allows you input the orientation of your beach in order to calculate numeric values for these components. For example, many Lake Michigan beaches in Illinois and Wisconsin have orientations around +90°.



F.1. Click the “Process wind data” button. Set the “Speed” column to **N** (WSPD_MET) and the “Direction (deg)” column to **O** (WDIR_MET). Check “Wind comes from” and enter a Rotation angle of 90° (the orientation of Red Arrow Park Beach). Click “OK” and then “OK” again on the Warning window about checking for multicollinearity.



F.2. Scroll down to the newly-created columns (**AA** and **AB**) and rename them by clicking on them and typing, respectively, “CRSSWND” and “SHRWND.”

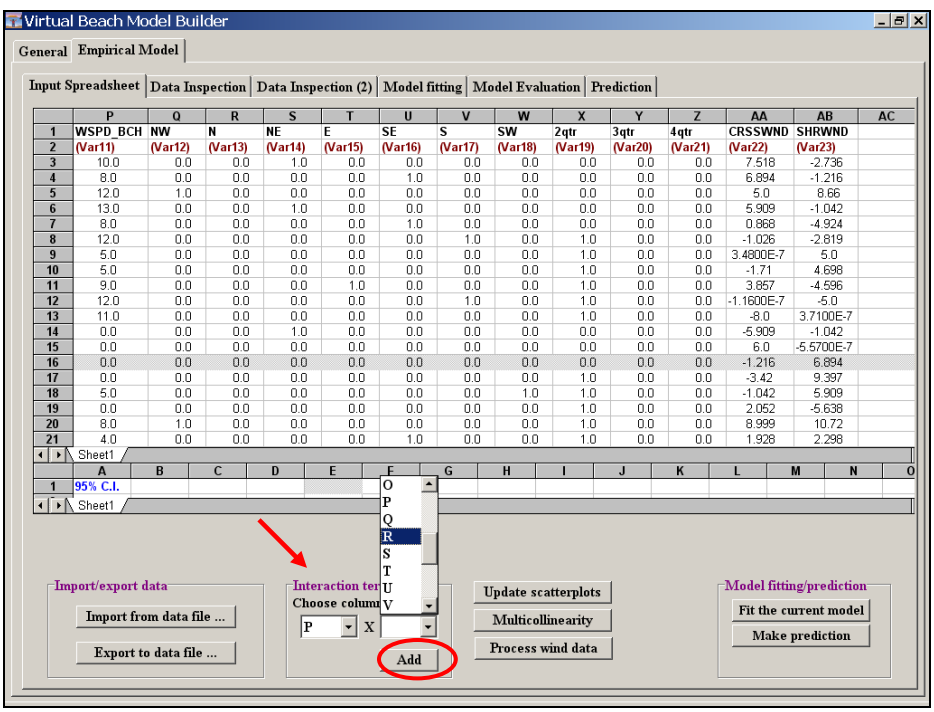
F.3. Select column **N** (WSPD_MET) and exclude it from the model by hitting the **F4** key. Do the same for column **O** (WDIR_MET).

G. Create "Interaction Terms" (combined variables).

A standard approach for dealing with collinear variables is to combine them into a single interaction term, via multiplication. This method may also be used with non-collinear variables as a means to improve model fit.

Interaction terms can be particularly useful for combining categorical variables for different wind vectors (cardinal directions) and wind velocity. In this example, where a major river outlets to the north of the beach, one could hypothesize that winds directly out of the north or out of the northeast (i.e., partially onshore) would correlate with elevated pathogen indicators. If true, increased wind velocity would correlate with increased indicator elevations *only* when winds were out of those directions. Interaction terms for Wind Speed (miles per hour) times North (1 if winds out of the north, 0 if other), and Wind Speed times Northeast, should capture these relationships.

G.1. In the "Input Spreadsheet" tab locate the "Interaction terms" box (see arrow below). In the two pull-down menus choose **P** (WSPD_BCH) and **R** (N), and then click the "Add" button. A new variable will appear in column **AC** titled "P*R". Repeat for columns **P** and **S** (NE) to create a new variable "P*S" (in column **AD**).

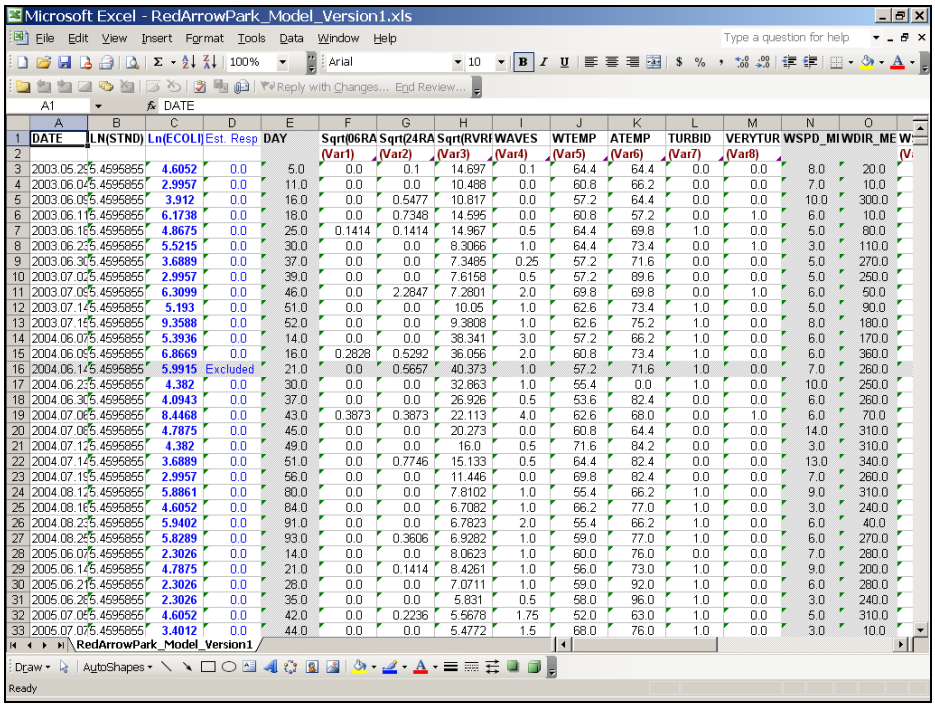


G.2. Click on the first cell in column **AC** ("P*R") and rename it by typing "N_WSPD." Similarly, rename "P*S" "NE_WSPD."

Note: *Virtual Beach 1.0* **does not save your work when you shut down** the program. To save your model, you must export the input spreadsheet to Excel.

In the lower left corner of the "Input Spreadsheet" tab, click on the "Export to data file …" button. In the Browse window that opens, navigate to the desired directory on your hard drive or storage device and name the file something like "RedArrowPark_Model_Version1.xls." Click Save and then OK.

The resulting Excel 4.0 spreadsheet (below) will have similar formatting to *Virtual Beach*, and serves as your backup. **If possible, leave *Virtual Beach* open for Module II**.



DATE	LN(STND)	Ln(ECOLI)	Est. Resp	DAY	Sqrt(06RA	Sqrt(24RA	Sqrt(RVRI	WAVES	WTEMP	ATEMP	TURBID	VERYTUR	WSPD_MI	WDIR_ME
					(Var1)	(Var2)	(Var3)	(Var4)	(Var5)	(Var6)	(Var7)	(Var8)		
2003.05.29	5.4595855	4.6052	0.0	5.0	0.0	0.1	14.697	0.1	64.4	64.4	0.0	0.0	8.0	20.0
2003.06.04	5.4595855	2.9957	0.0	11.0	0.0	0.0	10.488	0.0	60.8	66.2	0.0	0.0	7.0	10.0
2003.06.09	5.4595855	3.912	0.0	16.0	0.0	0.5477	10.817	0.0	57.2	64.4	0.0	0.0	10.0	300.0
2003.06.11	5.4595855	6.1738	0.0	18.0	0.0	0.7348	14.595	0.0	60.8	57.2	0.0	1.0	6.0	10.0
2003.06.18	5.4595855	4.8675	0.0	25.0	0.1414	0.1414	14.967	0.5	64.4	69.8	1.0	0.0	5.0	80.0
2003.06.23	5.4595855	5.5215	0.0	30.0	0.0	0.0	8.3066	1.0	64.4	73.4	0.0	1.0	3.0	110.0
2003.06.30	5.4595855	3.6889	0.0	37.0	0.0	0.0	7.3485	0.25	57.2	71.6	0.0	0.0	5.0	270.0
2003.07.02	5.4595855	2.9957	0.0	39.0	0.0	0.0	7.6158	0.5	57.2	89.6	0.0	0.0	5.0	250.0
2003.07.09	5.4595855	6.3099	0.0	46.0	0.0	2.2847	7.2801	2.0	69.8	69.8	0.0	1.0	6.0	50.0
2003.07.14	5.4595855	5.193	0.0	51.0	0.0	0.0	10.05	1.0	62.6	73.4	1.0	0.0	5.0	90.0
2003.07.18	5.4595855	9.3588	0.0	52.0	0.0	0.0	9.3808	1.0	62.6	75.2	1.0	0.0	8.0	180.0
2004.06.07	5.4595855	5.3936	0.0	14.0	0.0	0.0	38.341	3.0	57.2	66.2	1.0	0.0	6.0	170.0
2004.06.09	5.4595855	6.8669	0.0	16.0	0.2828	0.5292	36.056	2.0	60.8	73.4	1.0	0.0	6.0	360.0
2004.06.14	5.4595855	5.9915	Excluded	21.0	0.0	0.5657	40.373	1.0	57.2	71.6	1.0	0.0	7.0	260.0
2004.06.23	5.4595855	4.382	0.0	30.0	0.0	0.0	32.863	1.0	55.4	0.0	1.0	0.0	10.0	250.0
2004.06.30	5.4595855	4.0943	0.0	37.0	0.0	0.0	26.926	0.5	53.6	82.4	0.0	0.0	6.0	260.0
2004.07.06	5.4595855	8.4468	0.0	43.0	0.3873	0.3873	22.113	4.0	62.6	68.0	0.0	1.0	6.0	70.0
2004.07.08	5.4595855	4.7875	0.0	45.0	0.0	0.0	20.273	0.0	60.8	64.4	0.0	0.0	14.0	310.0
2004.07.12	5.4595855	4.382	0.0	49.0	0.0	0.0	16.0	0.5	71.6	84.2	0.0	0.0	3.0	310.0
2004.07.14	5.4595855	3.6889	0.0	51.0	0.0	0.7746	15.133	0.5	64.4	82.4	0.0	0.0	13.0	340.0
2004.07.19	5.4595855	2.9957	0.0	56.0	0.0	0.0	11.446	0.0	69.8	82.4	0.0	0.0	7.0	260.0
2004.08.12	5.4595855	5.8861	0.0	80.0	0.0	0.0	7.8102	1.0	55.4	66.2	1.0	0.0	9.0	310.0
2004.08.18	5.4595855	4.6052	0.0	84.0	0.0	0.0	6.7082	1.0	66.2	77.0	1.0	0.0	3.0	240.0
2004.08.23	5.4595855	5.9402	0.0	91.0	0.0	0.0	6.7823	2.0	55.4	66.2	1.0	0.0	6.0	40.0
2004.08.25	5.4595855	5.8289	0.0	93.0	0.0	0.3606	6.9282	1.0	59.0	77.0	1.0	0.0	6.0	270.0
2005.06.07	5.4595855	2.3026	0.0	14.0	0.0	0.0	8.0623	1.0	60.0	76.0	0.0	0.0	7.0	280.0
2005.06.14	5.4595855	4.7875	0.0	21.0	0.0	0.1414	8.4261	1.0	56.0	73.0	1.0	0.0	9.0	200.0
2005.06.21	5.4595855	2.3026	0.0	28.0	0.0	0.0	7.0711	1.0	59.0	92.0	1.0	0.0	6.0	280.0
2005.06.28	5.4595855	2.3026	0.0	35.0	0.0	0.0	5.831	0.5	58.0	96.0	1.0	0.0	3.0	240.0
2005.07.05	5.4595855	4.6052	0.0	42.0	0.0	0.2236	5.5678	1.75	52.0	63.0	1.0	0.0	5.0	310.0
2005.07.07	5.4595855	3.4012	0.0	44.0	0.0	0.0	5.4772	1.5	68.0	76.0	1.0	0.0	3.0	10.0