Virtual Beach 1.0
Learning Module II – *Model Evaluation and Nowcasting*

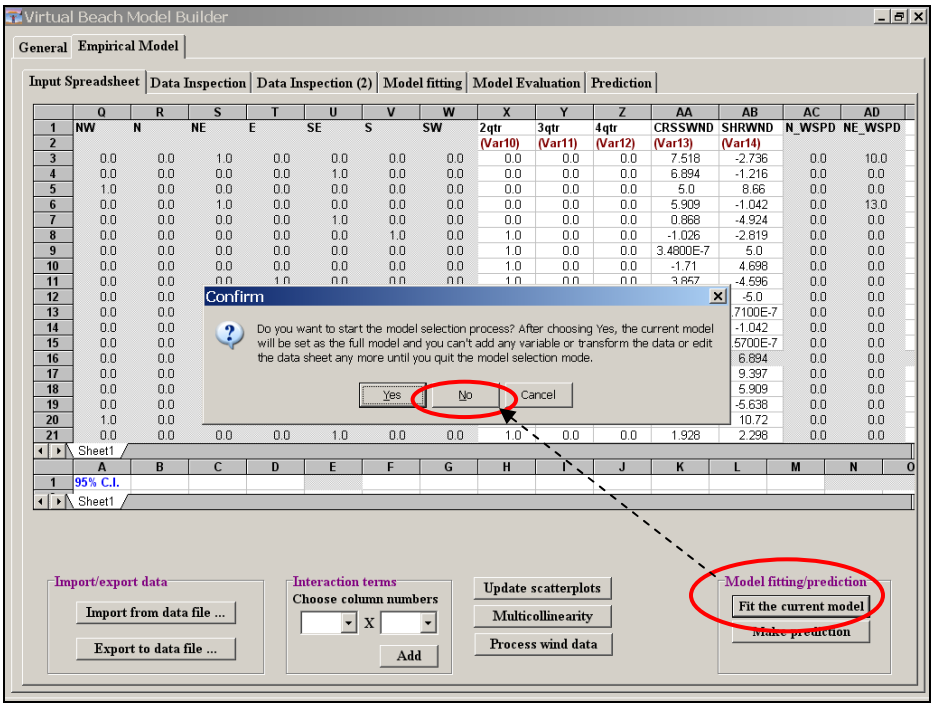
In this module you will learn how to:

A. "Fit" your model
B. Identify influential outliers
C. Identify "best, unbiased" models based on *Mallow's Cp*
D. Make single-day predictions with 95% confidence intervals
E. Save your model
F. Make real-time predictions

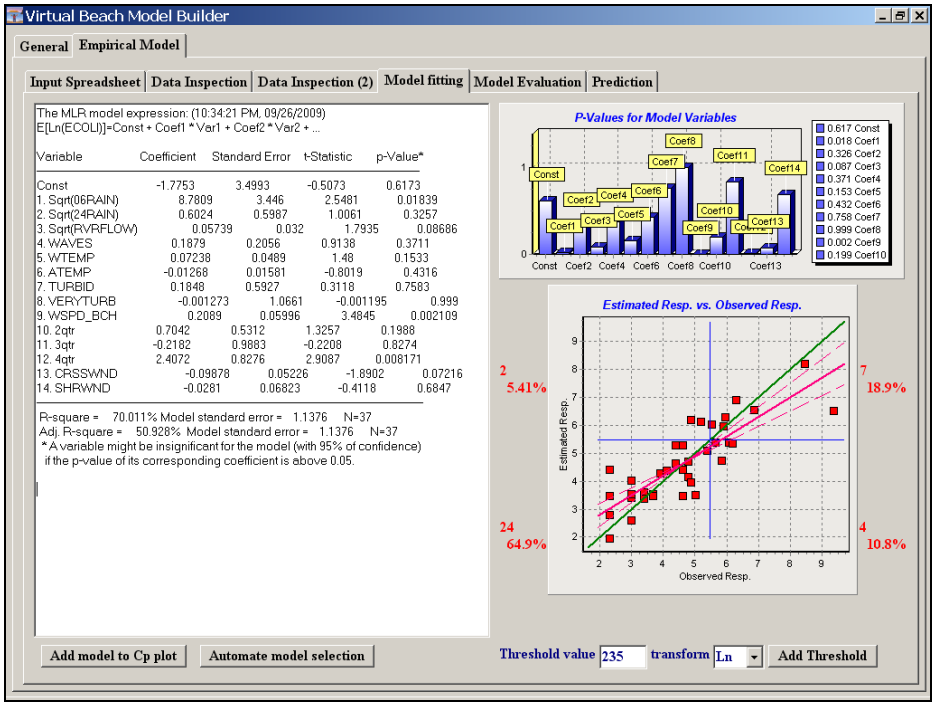
A. "Fitting" Your Model. Start in the "Input Spreadsheet" tab with the data set used in Module I.

A.1. In addition to the previously excluded columns: **E** (DAY), **N** (WSPD_MET), and **O** (WDIR_MET), use the F4 key to exclude the following variables (Note: You can only exclude one variable at a time): Columns **O** through **W** (the original wind speed and categorical wind directions as measured at the beach), and columns **AC** and **AD** (the wind vector times velocity interaction terms).

A.2 In the lower right-hand corner, click the "Fit the current model" button. (A pop-up window will ask whether you want to start the "model selection" process. Clicking yes will freeze the ability to conduct certain operations, such as transforming variables, although these functions can be restored later by canceling-out of the model selection process.) **For now, click "No."**



A.3. The “Model fitting” tab will automatically open, and you will see three main figures. To the left is a text/table summary of the resulting model. In the upper right corner is a bar chart showing *P-values* (measures of statistical significance – **lower is better**) associated with the different explanatory variables. In the lower right is a plot of predicted ln(ECOLI) versus observed. See below for details.



Information presented in the “Model fitting” tab is interpreted as follows:

<u>The MLR results table</u> (left side of the window)

- *Variable*: Name of variables included in the fit model
- *Coefficient*: The partial regression coefficient for the constant and each of the explanatory variables.
  - For the constant (Const.), the coefficient is the hypothetical value of ln(ECOLI) if all explanatory variables had a measure of zero (in this case, that value is negative: -1.7753).
  - For the explanatory variables (everything from “1. Sqrt(06RAIN)” down), the coefficients represent the expected increase or decrease in ln(ECOLI) in response to a one unit (e.g., 1 inch of rainfall) change in the variable in question—controlling for variation in all of the other explanatory variables
- *t-Statistic*: Used to calculate the *P-value* (described below). Higher values indicate higher levels of confidence in the accuracy of the estimated coefficients.

Continued

- *P-value*: Commonly-used measure of statistical significance. **Lower values** indicate **higher levels of confidence** in the accuracy of the estimated coefficients. A *P-value* of .05 or less indicates a 95% confidence that the estimated relationship (coefficient) between the explanatory variable in question and ln(ECOLI) is statistically significant.
- *R-square*: Known as the coefficient of determination, this is a measure of the model's "goodness of fit," interpreted as the percent of the variation in the response variable that is explained by the combined variation in the explanatory variables. In this example approximately 70% of the variation in ln(ECOLI) is explained by the variation in the explanatory variables.
- *Adj. R-square*: An adjusted version of R-square, discounted according to the number of explanatory variables in the model (additional variables typically inflate un-adjusted R-square). Unlike R-square, adjusted R-square is not interpretable by itself; however, it provides a relative basis for comparing the efficiency of models with different numbers of variables.

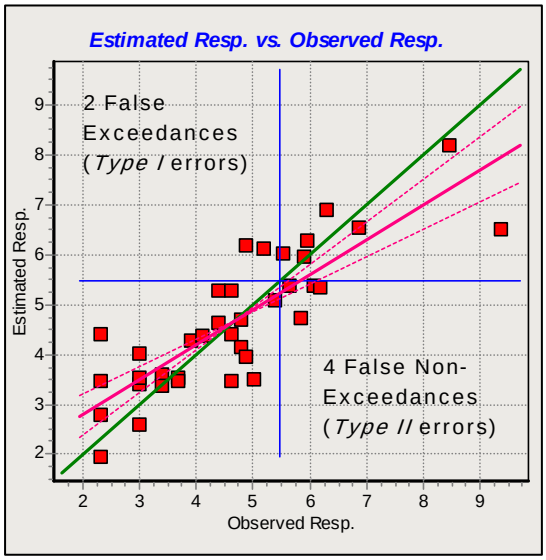
P-Values for Modeled Variables bar chart

Graphically displays the *p-Values* (statistical significance) of the constant ("Const.") and the different explanatory variables, listed from left to right in the order they are listed in the MLR results table on the left of the screen. Low bars represent explanatory variables for which we have higher levels of confidence in our estimated model coefficients.

Estimated Resp. Vs. Observed Resp. scatter plot

Graphically displays how well model-estimated values of the response variable match observed values for the various data points (days) included the model.

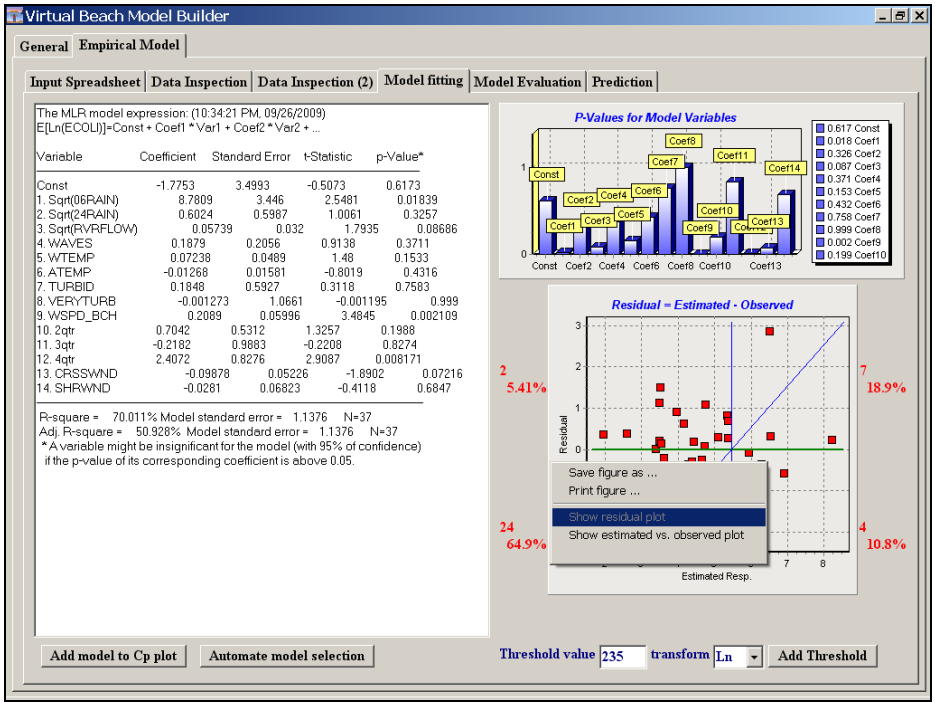
- Red dots are data points of observed values (X axis), plotted against the corresponding model estimates (Y axis).
- The perpendicular and horizontal blue lines represent the 235 CFU/100mL standard in both dimensions, enabling the visual and quantitative evaluation of "false exceedances" (*Type I Errors*) and "false non-exceedances" (*Type II Errors*).
- The purple line shows the model trend (bias) in relation to the perfect fit (1-to1) green line.



A.4. Look at the *Estimate Resp. vs. Observed Resp.* scatter plot and note that 4 of the 12 observed exceedances (the number of points to the right of the blue horizontal line) are "false non-exceedances" (*Type II Errors*).

<u>Note</u>: One of the primary goals of model fitting and evaluation is to ***minimize the number of false exceedances and non-exceedances***. Altering and re-fitting the model will hopefully produce more accurate estimates.

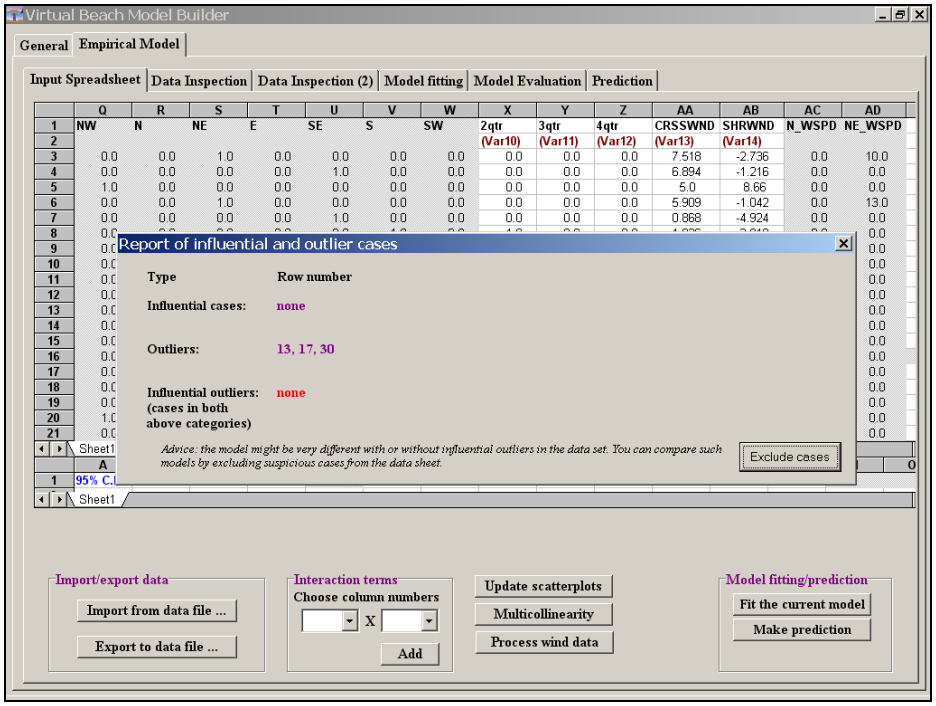
A.5. Move your mouse to inside the *Estimate Resp. vs. Observed Resp.* scatter plot, right click and select "Show residual plot." The graph will change to a plot of observed ln(ECOLI) against model residuals (error = estimated-observed values). The blue horizontal blue line (235 CFU threshold) rotates and the number of "false positives" and "false negatives" remain unchanged.



<u>Residual plots</u> are an important diagnostic for ensuring that OLS assumptions – especially the assumption that errors are random – are not violated. Tapering of cone-shaped clouds of data points indicate *heteroschedasticty* (one type of non-randomness) which would indicate that the model's estimates of ln(ECOLI), while on average unbiased, should be viewed as having a lower degree of confidence in their accuracy than the our tests for statistical significance suggest.

B. **Identify influential outliers.** In addition to our own visual inspection of the scatter plots, *Virtual Beach* can automatically detect data outliers, including "influential outliers (those likely to skew the model as fitted).

B.1 In the "Input Spreadsheet" tab, right click anywhere within the table and select "Check outlier cases for the current model." (Click "OK" in the warning window that pops up—we haven't made any changes to the model yet.)



B.2. Note that there are no "influential outliers;" i.e., those observations likely to skew the model as a whole. (Had there been any, clicking on the "Exclude cases" button would automatically exclude them from the model.)

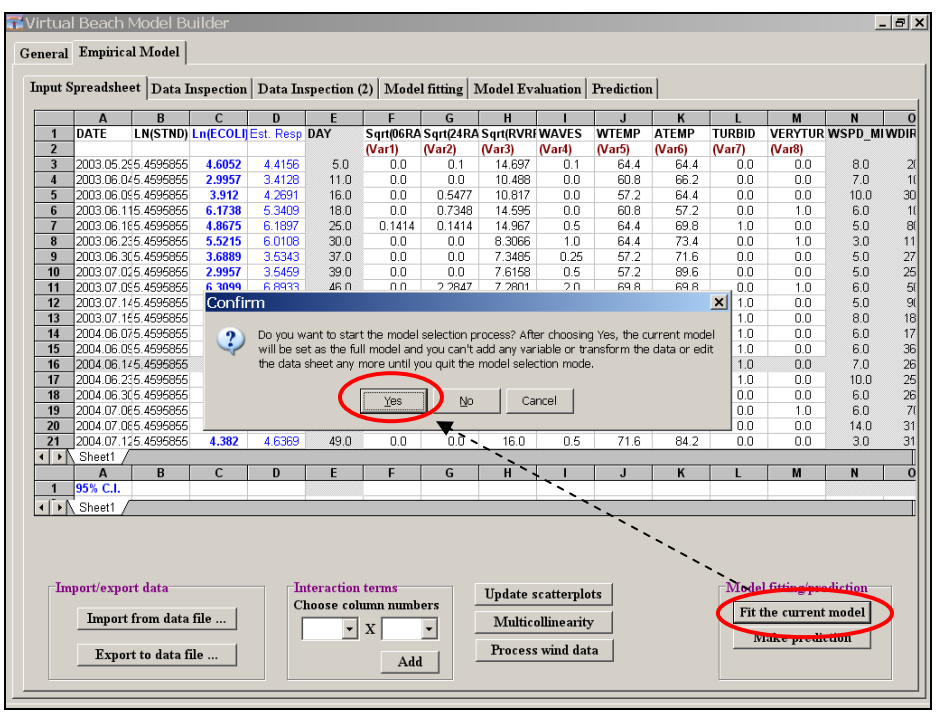
C. Identify "Best, Unbiased" Models based on *Mallow's Cp*

Building an MLR model for your beach should be an *intuitive process*. Start with a *conceptual model* – that is, a list of the factors or variables that you believe affect water quality at your beach, based on your experience. Use *Virtual Beach* to translate your conceptual model into an MLR model, and refine that model as you go – by adding, excluding, transforming, or combining variables; checking scatter plots; and evaluating the results in the "Model fitting" tab.

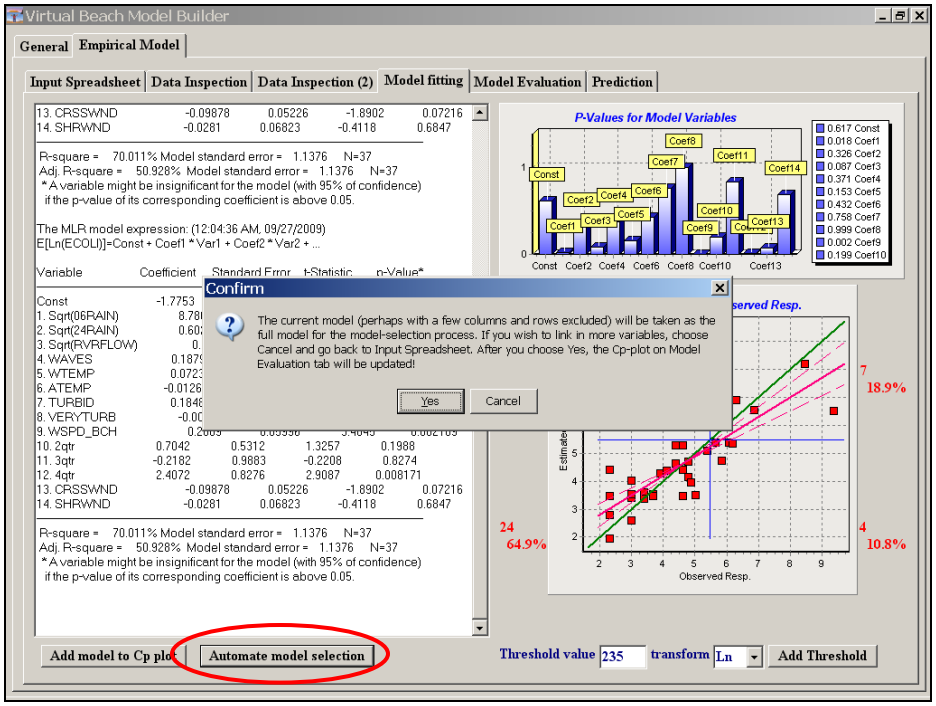
Selecting a "best" model is a difficult process if you are manually re-fitting the model with each change you make to the data and variables. Therefore, *Virtual Beach* provides an automated process selecting a "best, unbiased" model using a test statistic known as *Mallow's Cp*. The program conducts a *backwards regression*, whereby the "full" model (all of the non-excluded variables) is fit, and then variables are removed one at a time according to their level of statistical significance. At each step, the model is re-fit until there is only one variable left—that with the highest level of statistical significance.

The model with the lowest *Mallow's Cp* is considered the best, unbiased model, in that it has the best fit for the smallest possible number of explanatory variables. Too many variables in a model can lead to reduced statistical significance and can also be inefficient in terms of future data collection.

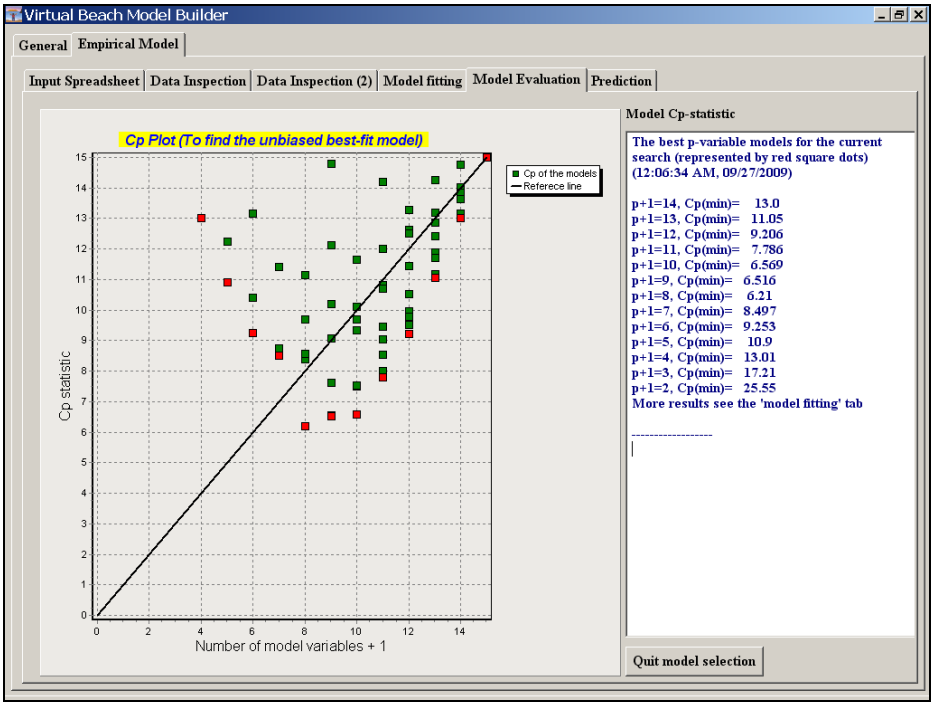
C.1 In the "Input Spreadsheet" tab, click the "Fit the current model" button in the lower right-hand corner. This time click "Yes" when the program asks if you would like to begin the "model selection" process.



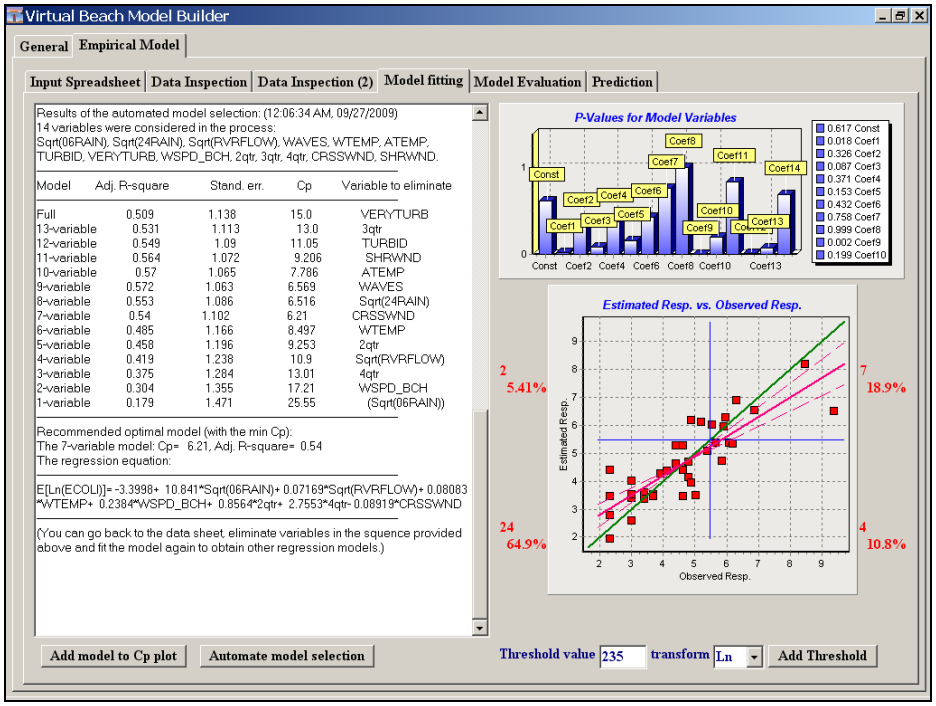
C.2. The “Model fitting” tab will automatically re-open. Towards the bottom-left of the window, click the “Automate model selection” button. Click “Yes” on the Confirm window that pop’s up.



C.3. The “Model Evaluation” tab will automatically open and you will see a *Cp Plot* on the left side and list of *Model Cp-statistics* on the right. The different models mare denoted by the number of explanatory variables (p) +1.



C.4 In the *Cp Plot* and *statistics* in the "Model Evaluation" tab, notice that the model with the lowest Cp value is the 7 variable (+1) model. Click on the model fitting tab to view more detailed information.

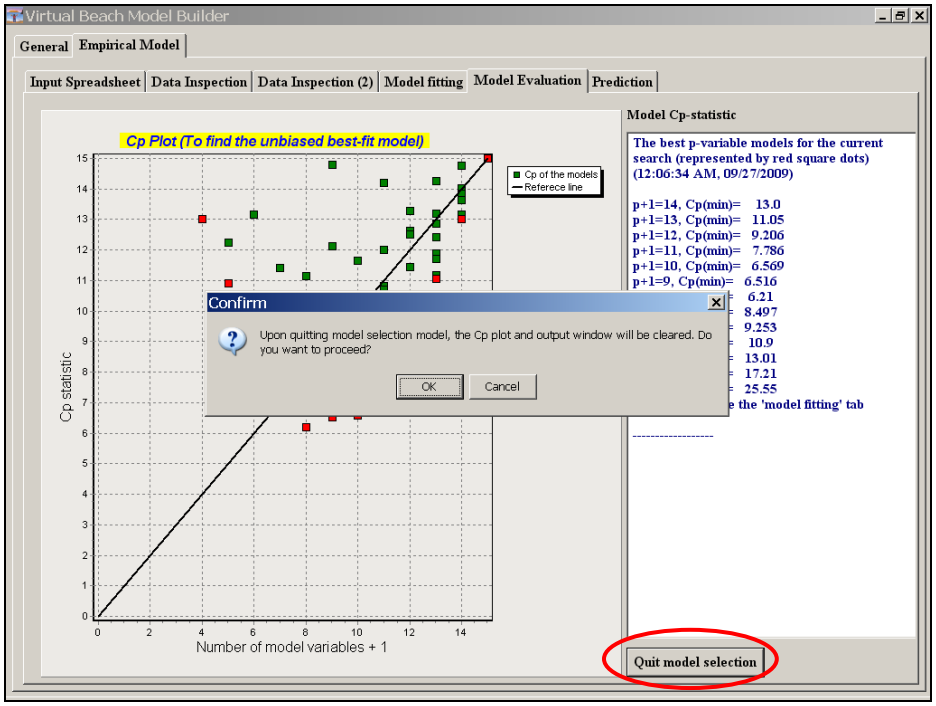


C.5. Copy the backwards regression results from the text box on the left-hand side of the screen (highlight > Ctrl-C) and then paste (Ctrl-V) into a text editor such as MS Word. There you can highlight the variables from the recommended "best, unbiased" model. (Those variables that are not highlighted below are recommended for exclusion.)

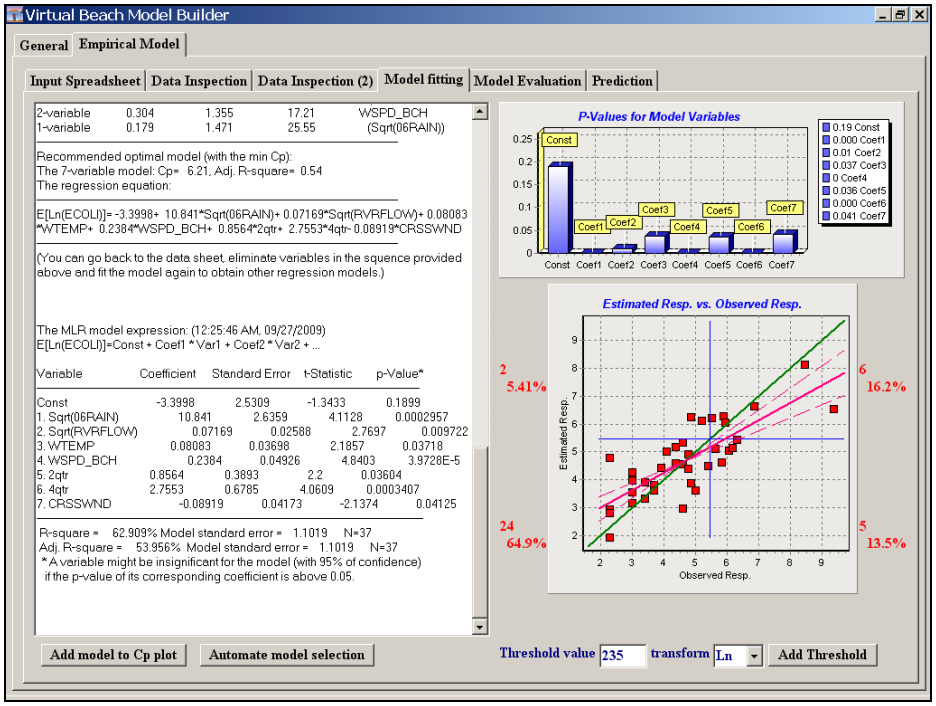
Model	Adj. R-square	Stand. err.	Cp	Variable to eliminate
Full	0.509	1.138	15.0	VERYTURB
13-variable	0.531	1.113	13.0	3qtr
12-variable	0.549	1.09	11.05	TURBID
11-variable	0.564	1.072	9.206	SHRWND
10-variable	0.57	1.065	7.786	ATEMP
9-variable	0.572	1.063	6.569	WAVES
8-variable	0.553	1.086	6.516	Sqrt(24RAIN)
7-variable	0.54	1.102	6.21	CRSSWND
6-variable	0.485	1.166	8.497	WTEMP
5-variable	0.458	1.196	9.253	2qtr
4-variable	0.419	1.238	10.9	Sqrt(RVRFLOW)
3-variable	0.375	1.284	13.01	4qtr
2-variable	0.304	1.355	17.21	WSPD_BCH
1-variable	0.179	1.471	25.55	(Sqrt(06RAIN))

Recommended optimal model (with the min Cp):
The 7-variable model: Cp= 6.21, Adj. R-square= 0.54

C.6. In the “Model Evaluation” tab, click the “Quit model selection” button in the lower right-hand corner. Click OK in the Confirm window that pops-up. (This will free you to add new variables, make new transformations, etc.)

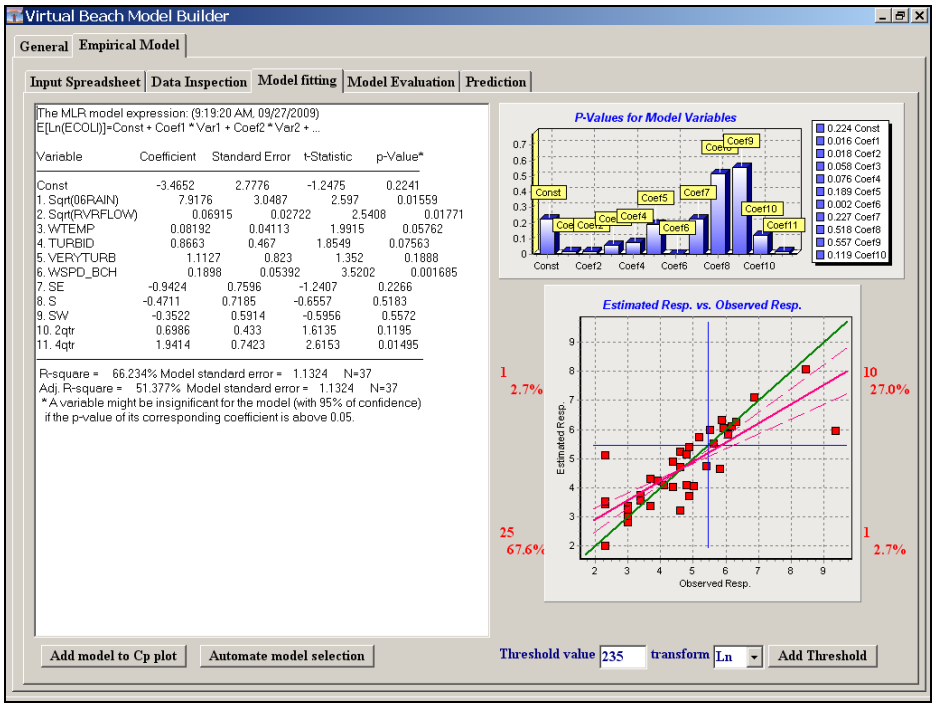


C.7. Go back to the “Input Spreadsheet” tab and fit the current model (see Step A.2). Note that while adjusted R-square improves from 51% to 54%, and *P-values* generally decreased, false exceedances (*Type I Errors*) increased from 4 to 5.



C.8. Revisit the list of variables copied in Step C.5 and use as a guide for re-fitting your model (in the "Input Spreadsheet" tab). Experiment with excluding and un-excluding different variables, and re-fitting the models under different combinations. Be sure to check for multicollinearity.

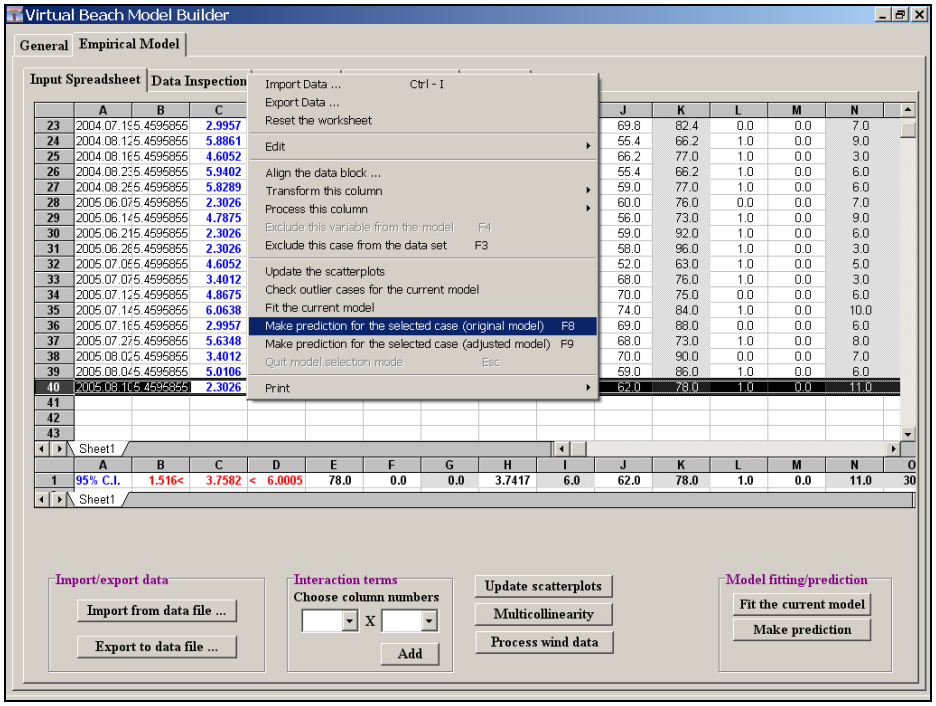
<u>How few "false exceedances" and "false non-exceedances" can you get?</u>



D. Make Single-day Predictions with 95% Confidence Intervals.

When you fit a model in the "Input Spreadsheet" tab, the values that populate column **D** ("Est. Resp.") are "point estimates." The likelihood that these point estimates will precisely match their corresponding observed values is extremely small. It is possible, however, to estimate ranges of likely values within a high level of confidence; e.g., 95%.

D.1. In the "Input Spreadsheet" tab, scroll down to last observation (row 40) and select by clicking on the row number to the left of the date column. Right click and select "Make prediction for the selected case (original mode)." (You can also hit the **F8** key.)



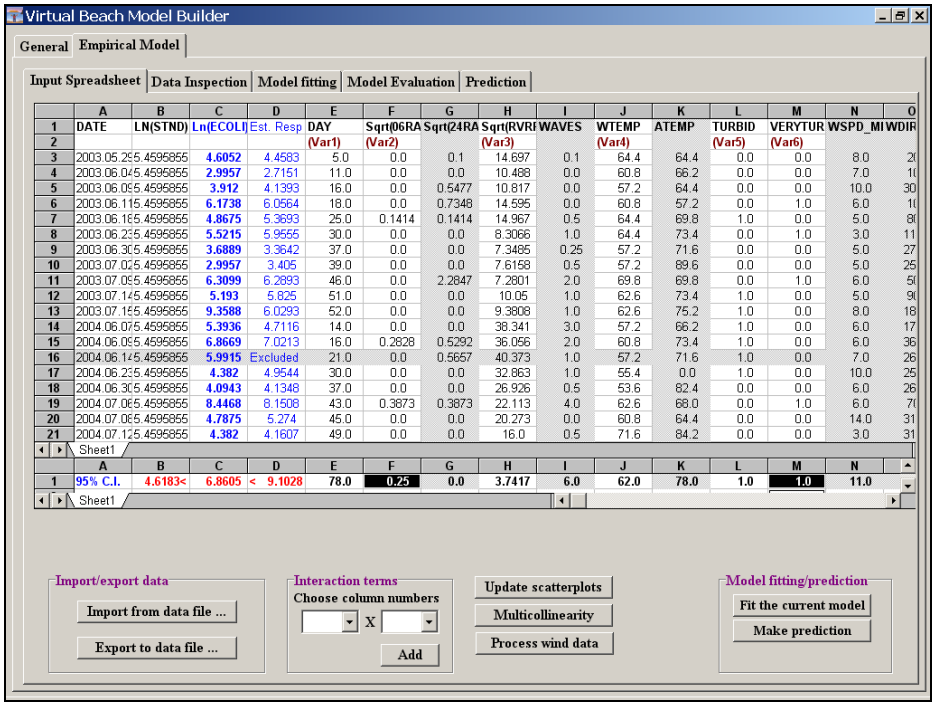
D.2. Note that the single-row spreadsheet below the larger input spreadsheet is now populated with data. **Red values** represent the point estimate of ln(ECOLI) flanked by the low value of a 95% confidence interval (to the left) and the high value of the interval to the right.

Models that violate OLS assumptions (regarding the distribution of residuals) may produce erroneous confidence intervals.

(Continued)

You can also *manually* enter and/or edit values in the small prediction spreadsheet at the bottom of the “Input Spreadsheet” tab and from these predict the response in your pathogen indicator variable.

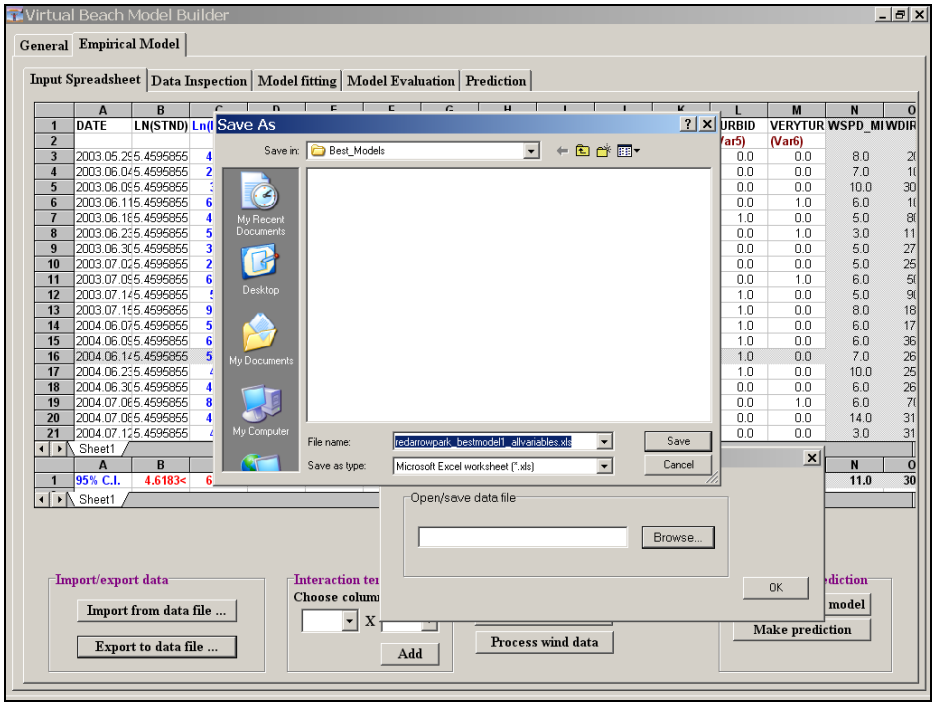
D.3. Change some of the values in the small prediction spreadsheet; for example, increase the value under column **F** (Sqrt(06RAIN)) or under column **M** (VERYTURB). Click the “Make Prediction” button in the bottom right corner. Note the change in the point estimate and confidence interval values.



E. Saving your Model

Once you have determined the final set of predictive (explanatory) variables in your model, you can save your model as an Excel spreadsheet and build a separate spreadsheet containing current data on those variables. This spreadsheet can be updated regularly with real-time data, imported into *Virtual Beach*, and used to make real-time predictions of *E. coli* concentrations at your beach.

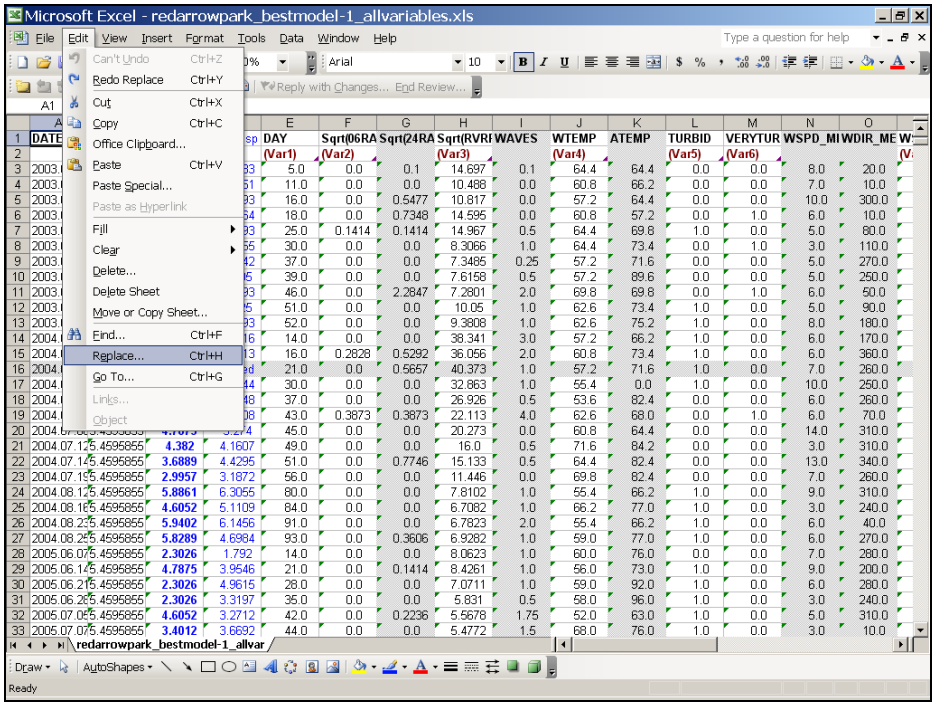
E.1. In the "Input Spreadsheet" tab, click the "Export to data file…" button. An "Import/export data" pop-up will open. Click the "Browse…" button, navigate to the folder where you wish to save the output data, name the file something like "redarrowpark_**bestmodel_allvariables**.xls", and click "Save" then "OK."



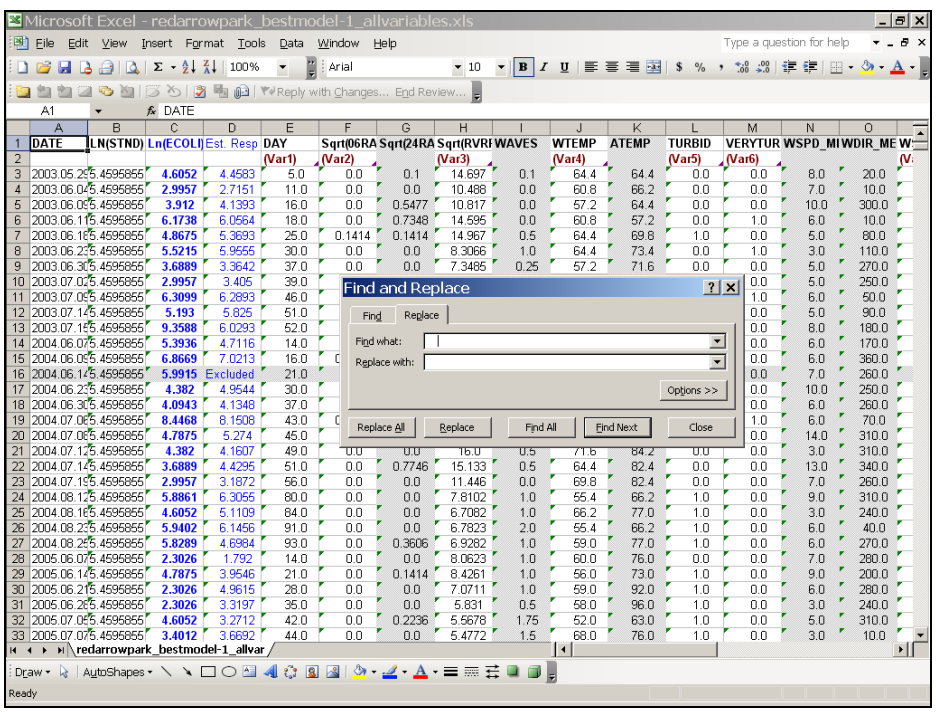
When exporting tables to Excel, *Virtual Beach* saves the formatting of the spreadsheet, including the gray highlighting of excluded variables (columns) and cases (rows). <u>Use these markings as visual guides within Excel</u> to edit the exported table.

(Continued)

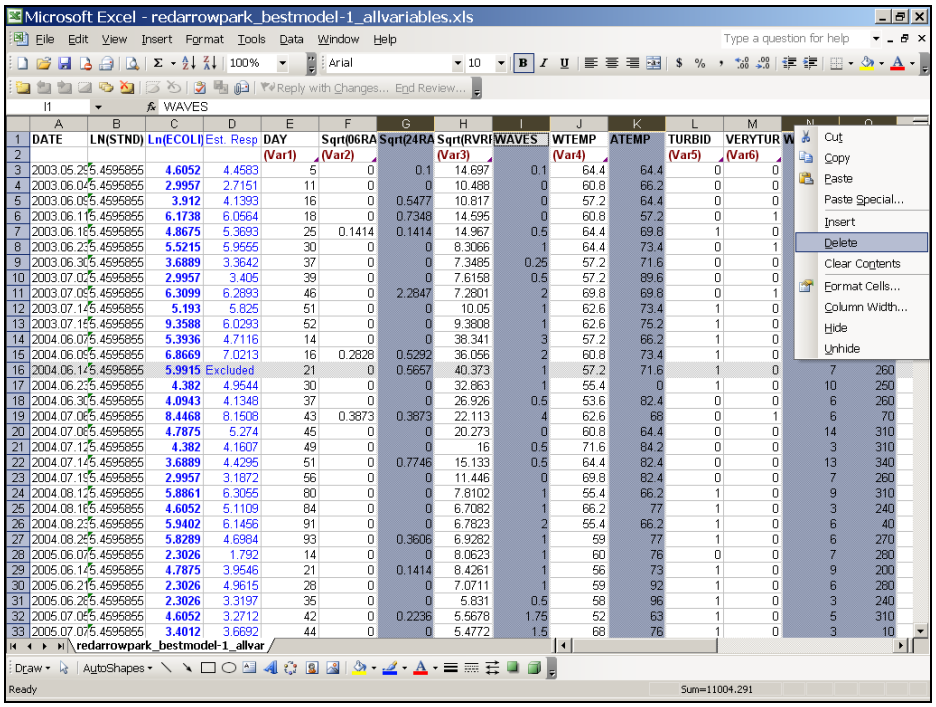
E.2. ⚠ <u>Note</u>: Excel files exported from *Virtual Beach* contain sequences of 4 blank spaces at the beginning of data cells. To remove these, first open the exported table in Excel and from the pull-down menu select “Edit” > “Replace.”



E.3. A “Find and Replace” pop-up window will open. Next to “Find what” type <u>4 blank spaces</u>, then hit the “Replace All” button. (This will eliminate all blank spaces and correctly format the data cells.)



E.4. While still in Excel, select all of the grayed-out columns (excluded variables) by clicking on one column header then, holding the Ctrl key, click on each of the other columns. When all of the grayed-out columns are selected, right-click and select “Delete” to remove them from the table.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	DATE	LN(STND)	Ln(ECOLI)	Est. Resp	DAY	Sqrt(06RA	Sqrt(24RA	Sqrt(RVRI	WAVES	WTEMP	ATEMP	TURBID	VERYTUR	W	
2					(Var1)	(Var2)		(Var3)		(Var4)		(Var5)	(Var6)		
3	2003.05.25	5.4595855	4.6052	4.4583	5	0	0.1	14.697	0.1	64.4	64.4	0	0		
4	2003.06.04	5.4595855	2.9957	2.7151	11	0	0	10.488	0	60.8	66.2	0	0		
5	2003.06.09	5.4595855	3.912	4.1393	16	0	0.5477	10.817	0	57.2	64.4	0	0		
6	2003.06.11	5.4595855	6.1738	6.0564	18	0	0.7348	14.595	0	60.8	57.2	0	1		
7	2003.06.18	5.4595855	4.8675	5.3693	25	0.1414	0.1414	14.967	0.5	64.4	69.8	1	0		
8	2003.06.23	5.4595855	5.5215	5.9555	30	0	0	8.3066	1	64.4	73.4	0	1		
9	2003.06.30	5.4595855	3.6889	3.3642	37	0	0	7.3485	0.25	57.2	71.6	0	0		
10	2003.07.02	5.4595855	2.9957	3.405	39	0	0	7.6158	0.5	57.2	89.6	0	0		
11	2003.07.09	5.4595855	6.3099	6.2893	46	0	2.2847	7.2801	2	69.8	69.8	0	1		
12	2003.07.14	5.4595855	5.193	5.825	51	0	0	10.05	1	62.6	73.4	1	0		
13	2003.07.15	5.4595855	9.3588	6.0293	52	0	0	9.3808	1	62.6	75.2	1	0		
14	2004.06.07	5.4595855	5.3936	4.7116	14	0	0	38.341	3	57.2	66.2	1	0		
15	2004.06.09	5.4595855	6.8669	7.0213	16	0.2828	0.5292	36.056	2	60.8	73.4	1	0		
16	2004.06.14	5.4595855	5.9915	Excluded	21	0	0.5657	40.373	1	57.2	71.6	1	0	7	260
17	2004.06.23	5.4595855	4.382	4.9544	30	0	0	32.863	1	55.4	0	1	0	10	250
18	2004.06.30	5.4595855	4.0943	4.1348	37	0	0	26.926	0.5	53.6	82.4	0	0	6	260
19	2004.07.06	5.4595855	8.4468	8.1508	43	0.3873	0.3873	22.113	4	62.6	68	0	1	6	70
20	2004.07.08	5.4595855	4.7875	5.274	45	0	0	20.273	0	60.8	64.4	0	0	14	310
21	2004.07.12	5.4595855	4.382	4.1607	49	0	0	16	0.5	71.6	84.2	0	0	3	310
22	2004.07.14	5.4595855	3.6889	4.4295	51	0	0.7746	15.133	0.5	64.4	82.4	0	0	13	340
23	2004.07.19	5.4595855	2.9957	3.1872	56	0	0	11.446	0	69.8	82.4	0	0	7	260
24	2004.08.12	5.4595855	5.8861	6.3055	80	0	0	7.8102	1	55.4	66.2	1	0	9	310
25	2004.08.18	5.4595855	4.6052	5.1109	84	0	0	6.7082	1	66.2	77	1	0	3	240
26	2004.08.23	5.4595855	5.9402	6.1456	91	0	0	6.7823	2	55.4	66.2	1	0	6	40
27	2004.08.25	5.4595855	5.8289	4.6984	93	0	0.3606	6.9282	1	59	77	1	0	6	270
28	2005.06.07	5.4595855	2.3026	1.792	14	0	0	8.0623	1	60	76	0	0	7	280
29	2005.06.14	5.4595855	4.7875	3.9546	21	0	0.1414	8.4261	1	56	73	1	0	9	200
30	2005.06.21	5.4595855	2.3026	4.9615	28	0	0	7.0711	1	59	92	1	0	6	280
31	2005.06.28	5.4595855	2.3026	3.3197	35	0	0	5.831	0.5	58	96	1	0	3	240
32	2005.07.05	5.4595855	4.6052	3.2712	42	0	0.2236	5.5678	1.75	52	63	1	0	5	310
33	2005.07.07	5.4595855	3.4012	3.6692	44	0	0	5.4772	1.5	68	76	1	0	3	10

Note: Following Step E.4, there should be no columns *to the right of column D* that are blank in row 2. Each of the remaining columns (starting at column D) should have a model variable number (e.g., **Var1**). Delete any columns that do not have a variable number.

E.5. Delete the grayed-out row (16) by right-clicking and selecting delete, similar to E.4 above.

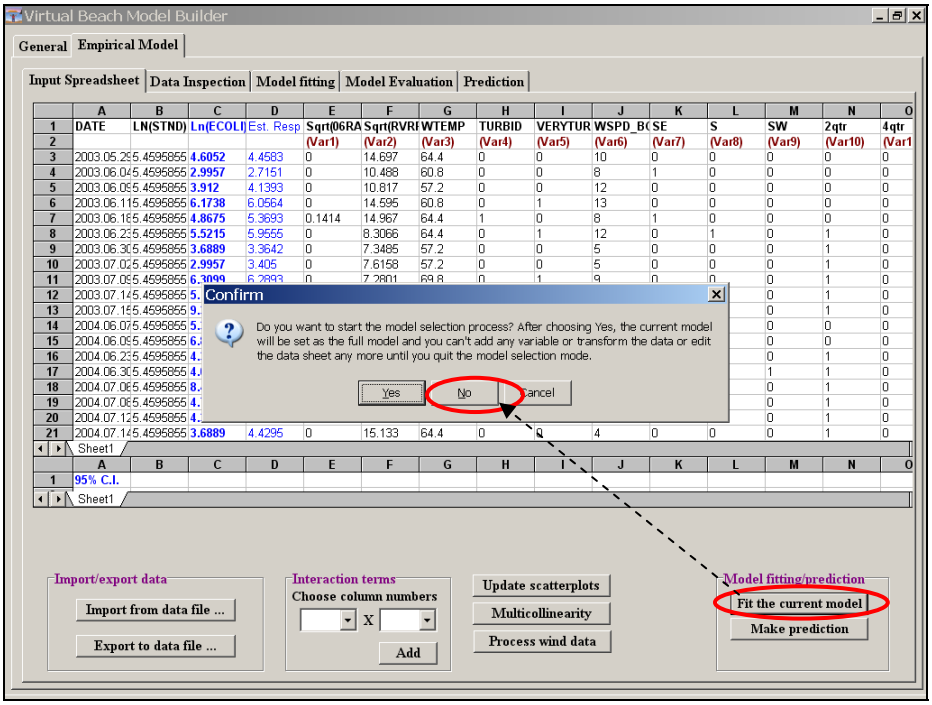
E.6. Re-save the table as an Excel 4.0 spreadsheet (or tab-delimited text file) named something like “redarrowpark_**bestmodel_inputs**.xls”.

F. **Making Real-Time Predictions**. Once you have built and saved your model, you can predict values of your response variable (e.g., *E. coli*) using separate data on your chosen explanatory variables. These data can be updated and imported into *Virtual Beach* on a daily basis, in order to make real-time predictions of water quality at your beach.

F.1. Open a new session of *Virtual Beach*.

F.2. In the "Import Spreadsheet" tab, click on the "Import from data file…" button, and open the file *RedArrowPark_Bestmodel-1_Inputs.xls* (available for download at http://dnr.wi.gov/org/es/science/contaminants/_____).

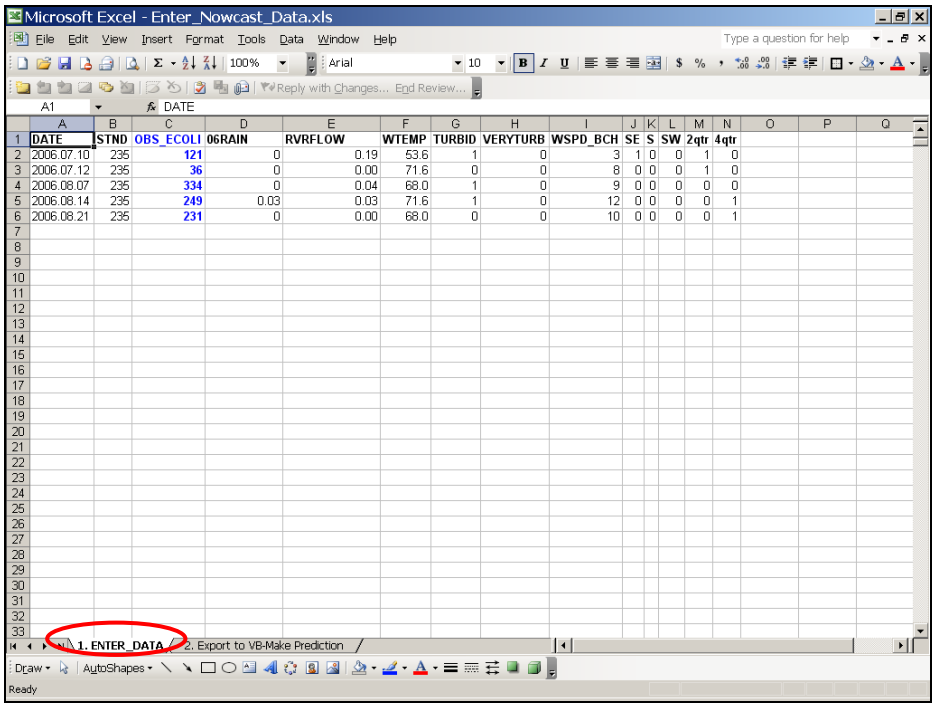
F.3. Click the "Fit the current model" button and then "No" (on the Confirm window).



F.4. The "Model fitting" window will open automatically, as in A.3 and C.4 above. View the text summary of the resulting model (R-square = 66.234%, Adj. R-square = 51.377%), the bar chart of *P-values*, and the plot of estimated versus observed values (1 "false exceedance" and 1 false "non-exceedance").

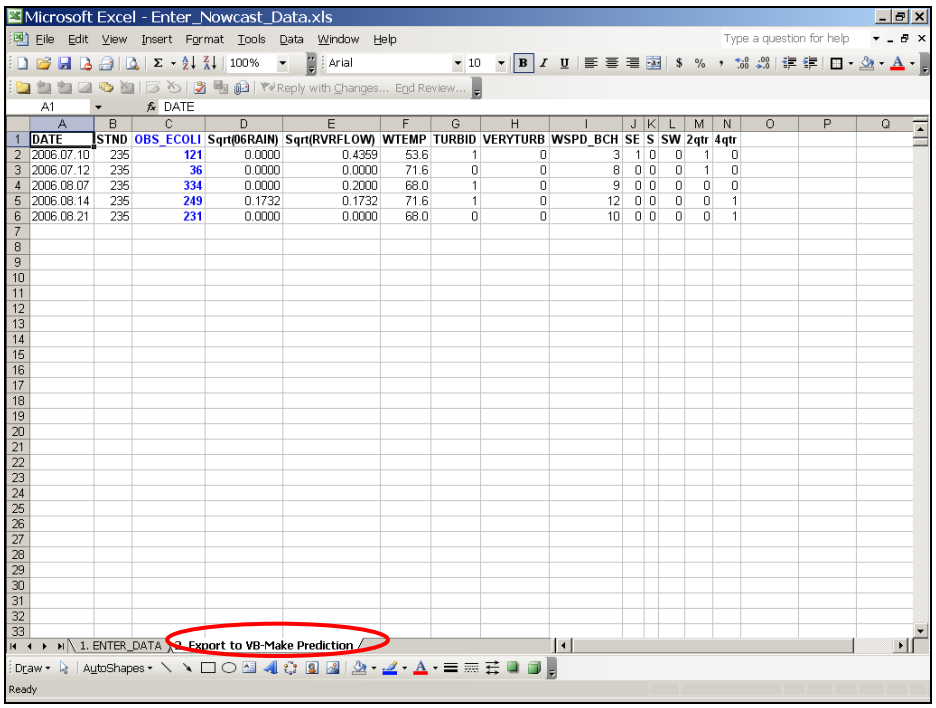
Note: *Virtual Beach 1.0* does not have a function for directly entering daily values for making real-time predictions "on the fly" (future versions will). You can use MS Excel to enter and automatically transform/format daily data for use in the "Make Predictions" tab of *Virtual Beach*.

F.5. Minimize *Virtual Beach* and open the file *Enter_Nowcast_Data.xls* (available at http://dnr.wi.gov/org/es/science/contaminants/_____.)



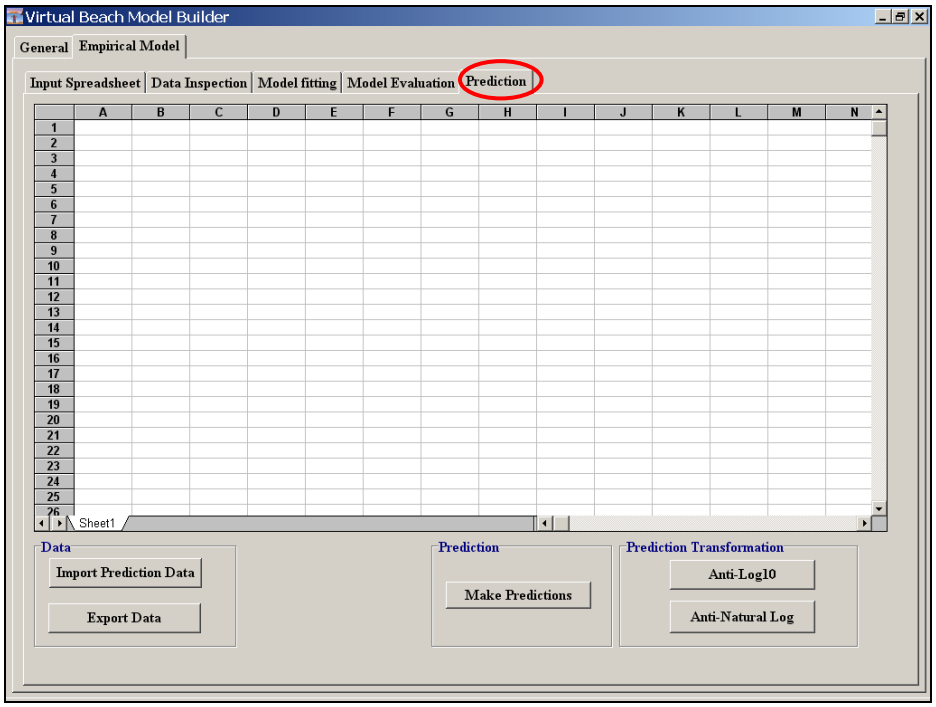
DATE	STND	OBS_ECOLI	06RAIN	RVRFLOW	WTEMP	TURBID	VERYTURB	WSPD_BCH	SE	S	SW	2qtr	4qtr
2006.07.10	235	121	0	0.19	53.6	1	0	3	1	0	0	1	0
2006.07.12	235	36	0	0.00	71.6	0	0	8	0	0	0	1	0
2006.08.07	235	334	0	0.04	68.0	1	0	9	0	0	0	0	0
2006.08.14	235	249	0.03	0.03	71.6	1	0	12	0	0	0	0	1
2006.08.21	235	231	0	0.00	68.0	0	0	10	0	0	0	0	1

F.6 Click on the two tabs: 1. “ENTER_DATA” and “2. Export to VB-Make Predictions” to see how daily data entry works.

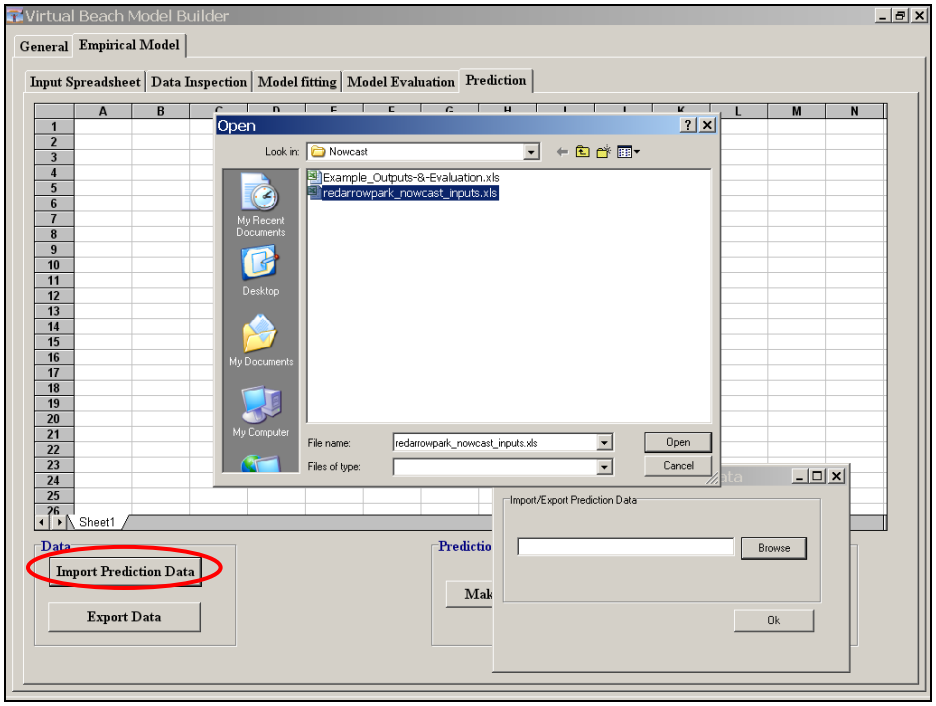


DATE	STND	OBS_ECOLI	Sqrt(06RAIN)	Sqrt(RVRFLOW)	WTEMP	TURBID	VERYTURB	WSPD_BCH	SE	S	SW	2qtr	4qtr
2006.07.10	235	121	0.0000	0.4359	53.6	1	0	3	1	0	0	1	0
2006.07.12	235	36	0.0000	0.0000	71.6	0	0	8	0	0	0	1	0
2006.08.07	235	334	0.0000	0.2000	68.0	1	0	9	0	0	0	0	0
2006.08.14	235	249	0.1732	0.1732	71.6	1	0	12	0	0	0	0	1
2006.08.21	235	231	0.0000	0.0000	68.0	0	0	10	0	0	0	0	1

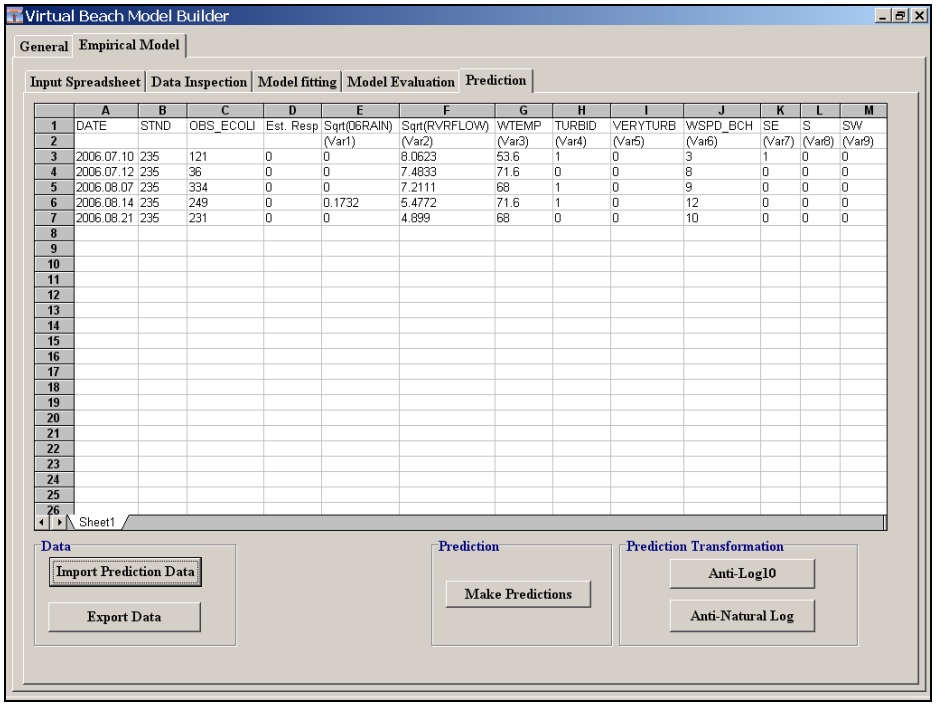
F.7. Return to *Virtual Beach* and click on the “Prediction” tab. This will open a blank spreadsheet view.



F.8. Click on the "Import Prediction Data" button and open the Excel file titled: *Redarrowpark_Nowcast_Inputs.xls* (available for download at http://dnr.wi.gov/org/es/science/contaminants/)

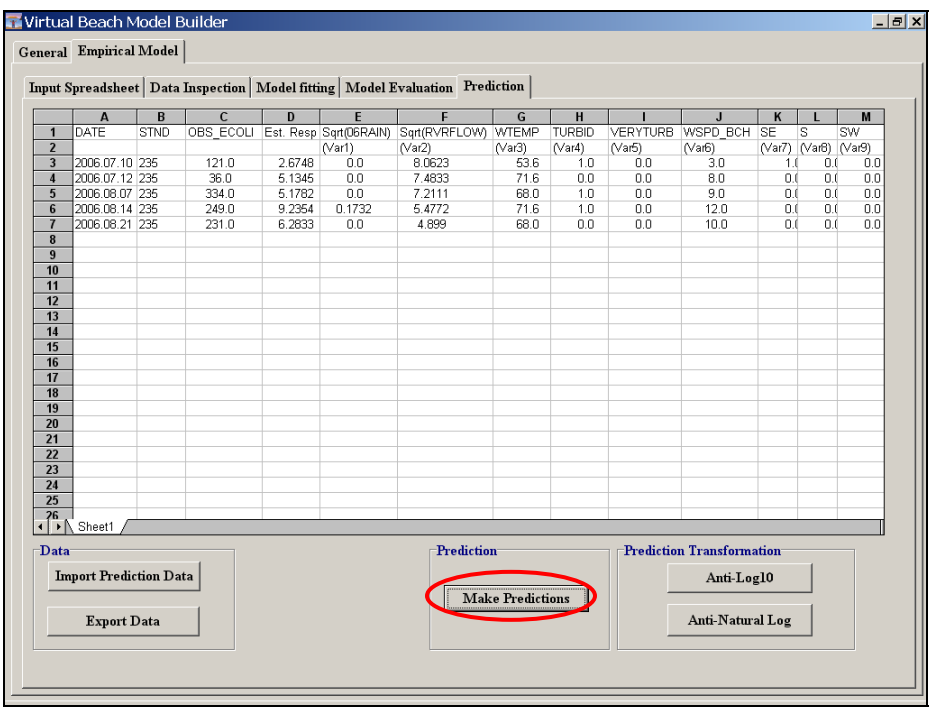


F.9. The table will open formatted with a new second row listing variable numbers ("Var1," Var2," etc.) and an added column, **D**, "Est. Resp". This will be populated with model-predicted values of the response variable.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	DATE	STND	OBS_ECOLI	Est. Resp	Sqrt(06RAIN)	Sqrt(RVRFLOW)	WTEMP	TURBID	VERYTURB	WSPD_BCH	SE	S	SW
2					(Var1)	(Var2)	(Var3)	(Var4)	(Var5)	(Var6)	(Var7)	(Var8)	(Var9)
3	2006.07.10	235	121	0	0	8.0623	53.6	1	0	3	1	0	0
4	2006.07.12	235	36	0	0	7.4833	71.6	0	0	8	0	0	0
5	2006.08.07	235	334	0	0	7.2111	68	1	0	9	0	0	0
6	2006.08.14	235	249	0	0.1732	5.4772	71.6	1	0	12	0	0	0
7	2006.08.21	235	231	0	0	4.899	68	0	0	10	0	0	0

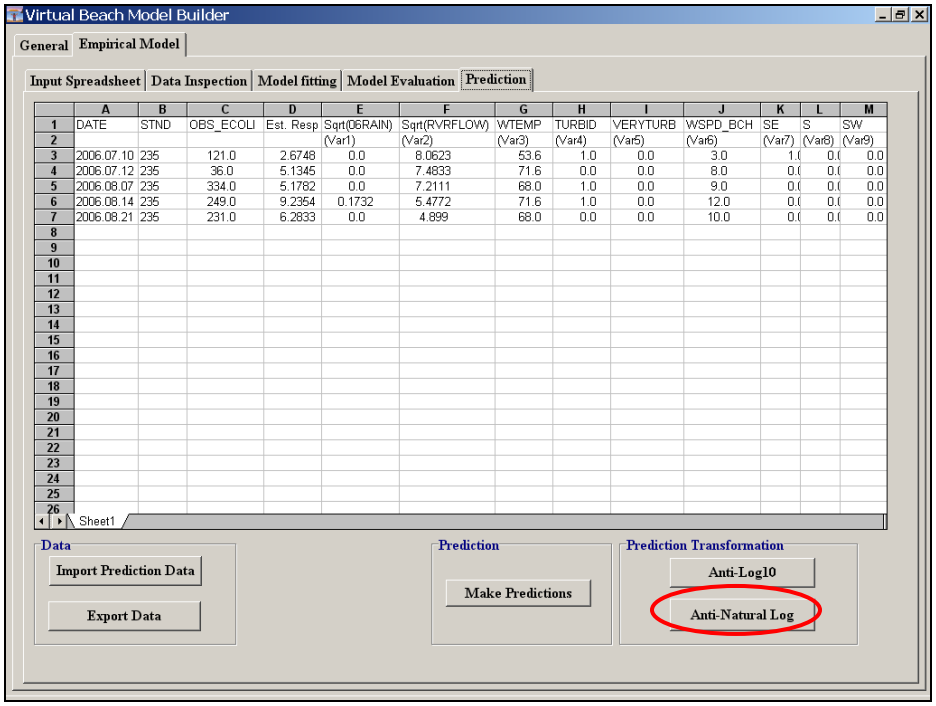
F.10. Click the “Make Predictions” button to generate model-based predictions of your response variable (the natural log of *E. coli*) in **column D**.



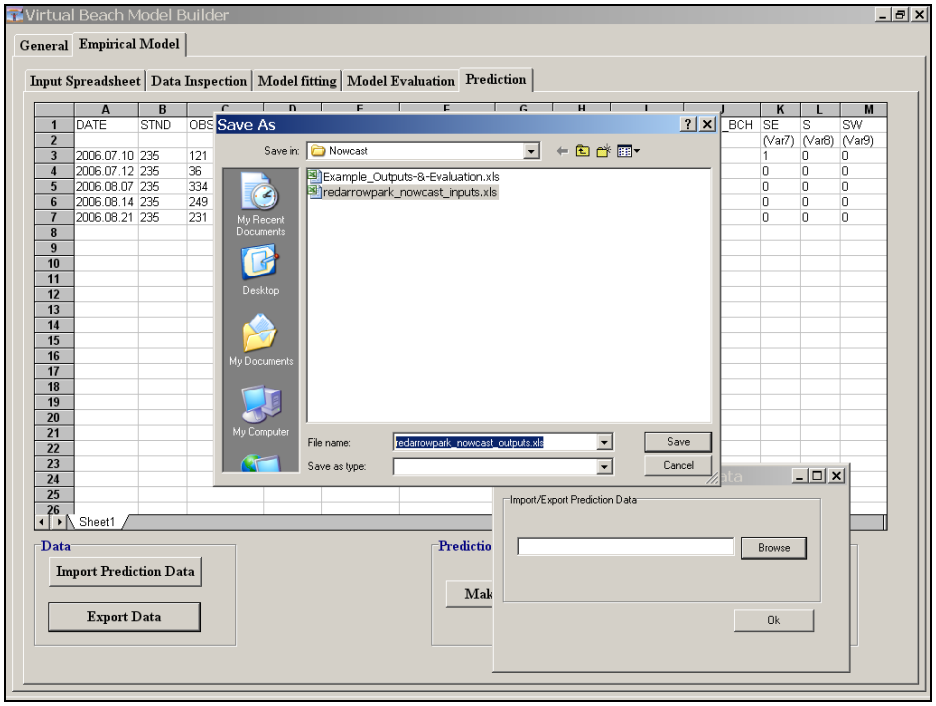
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	DATE	STND	OBS_ECOLI	Est. Resp	Sqrt(06RAIN)	Sqrt(RVRFLOW)	WTEMP	TURBID	VERYTURB	WSPD_BCH	SE	S	SW
2					(Var1)	(Var2)	(Var3)	(Var4)	(Var5)	(Var6)	(Var7)	(Var8)	(Var9)
3	2006.07.10	235	121.0	2.6748	0.0	8.0623	53.6	1.0	0.0	3.0	1.(	0.(	0.0
4	2006.07.12	235	36.0	5.1345	0.0	7.4833	71.6	0.0	0.0	8.0	0.(	0.(	0.0
5	2006.08.07	235	334.0	5.1782	0.0	7.2111	68.0	1.0	0.0	9.0	0.(	0.(	0.0
6	2006.08.14	235	249.0	9.2354	0.1732	5.4772	71.6	1.0	0.0	12.0	0.(	0.(	0.0
7	2006.08.21	235	231.0	6.2833	0.0	4.899	68.0	0.0	0.0	10.0	0.(	0.(	0.0

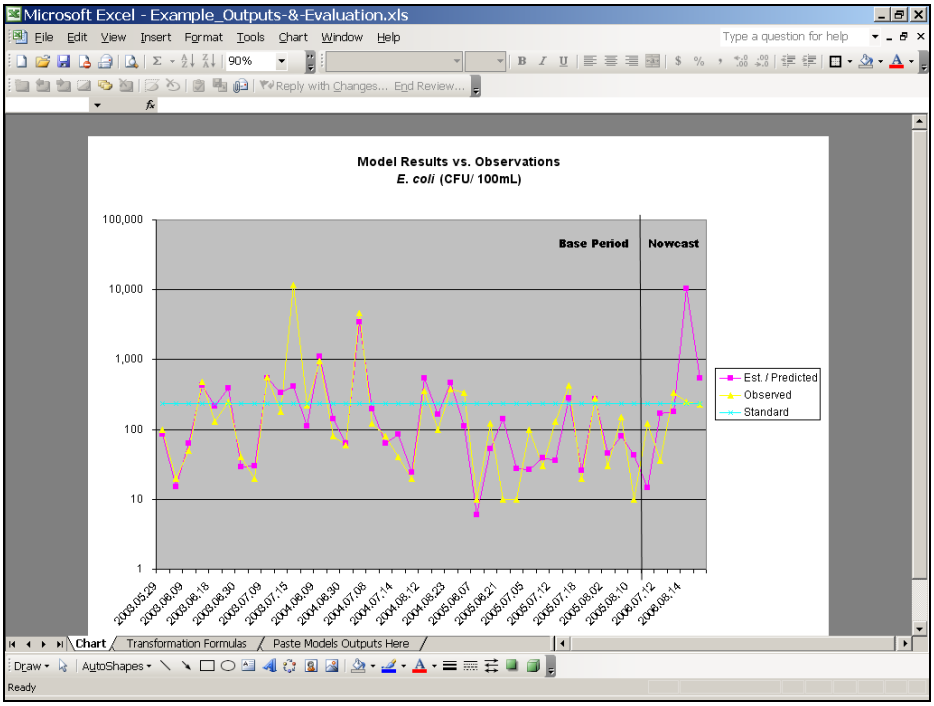
F.11. In the lower right-hand corner, under “Prediction Transformation,” are two buttons for un-transforming estimated response values from either Log10 or

Natural Log. Click the “Anti-Natural Log” button to convert the model predictions from ln(ECOLI) to concentrations of *E. coli* in CFU/100 mL.



F.12. Click the “Export Data” button, then browse (navigate) to a directory where you wish to save the output data, name the file, and click OK.



- Once you have exported model outputs to MS Excel you can conduct further evaluations – such as plotting observed data and model estimates/predictions over time.

- The file titled *Example_Outputs-&-Evaluation.xls* (available for download at http://dnr.wi.gov/org/es/science/contaminants/) contains the Excel commands and an example chart, shown below.