Sediment Spreadsheet Manual

7/21/2010

U.S. Geological Survey

Contents

Introduction

Getting Started

In order to use this program, Macros must be enabled. For Excel 2003, go to the Tools menu. Under the Macros option select Security. Set the Security level to either low or medium. You must then close and reopen excel and this program. If medium is selected, indicate that you wish to enable macros. For Excel 2007, above the formula far, there is a Security warning. Click options and choose the "enable this content" option. Click OK, and macros will be enabled.

Note: Program was created on Microsoft Windows XP using Microsoft Office Excel 2007 with 32 bit color. All properties and formats are described as they appear using this software. Variations may occur based on operating programs and settings.

Data Entry

First select the "Regression and Analysis 1" tab located at the bottom of Excel. Enter the Site Number in cell B3 and the Site Name in cell F3. If this information is not entered prior to running a macro the program will require you to enter this information before proceeding. Insert date and time beginning in cells C7 and D7 respectively. Insert Turbidity, Streamflow, Suspended Sediment Concentration, and Percent Sand data beginning in cells E7, F7, G7, and H7 respectively. All data that is entered by the user is shaded in a darker shade of green. The Date+Time and Julian Day columns will be automatically filled by the macro once the "Data Analysis" button is clicked. The program will not work if there are empty cells in streamflow, SSC, or turbidity. Empty cells are permissible for percent sand. The program can handle data up to row 2500, as more data is entered more time will be required for the program to process. For a sample of the data entry see Image 1 in Appendix.

Process

After the data has been entered, the buttons should be clicked in increasing numerical order. The only exception is button #3 which does not need to be clicked to proceed to button #4. Button #3 should only be pressed when a Candidate Model is wished to be generated for the current data on the current Regression and Analysis sheet. For a sample of the buttons see Image 2 in Appendix.

Regression and Analysis

For all buttons, simply follow the prompts. The following buttons can be found on the "Regression and Analysis" sheets. Plots of the data are automatically generated above the data as soon as values are entered.

Use the drop down boxes in Cells J5 through L5 to choose the transformations for each variable (user.selected.transform). For a sample of drop down boxes see Image 3 in Appendix.

i. Log (Logarithm to the base 10)
ii. Ln (logarithm to the base e)
iii. Sqrt (square root)

NOTE: Currently, MSPE and Duan BCF cannot be computed using the sqrt transformation for the response variable.

1. Data Analysis

This button does a number of calculations designed to help the user choose the appropriate model for the data submitted. Residual plots are automatically generated as a result of these calculations. This should be done for each Regression and Analysis sheet.

a. Transforms the Turb, Q, and SSC data based on the user selected transformations
b. Computes six different regression models
    a. (user.selected.transform)SSC vs (user.selected.transform)Turb
    b. SSC vs Turb
    c. (user.selected.transform)SSC vs (user.selected.transform)Q
    d. SSC vs Q
    e. (user.selected.transform)SSC vs (user.selected.transform)Turb + (user.selected.transform)Q
    f. SSC vs Turb + Q
c. Computes a number of regression statistics for each of the six models
    a. Slope(s)
    b. Intercept
    c. R^2
    d. Adjusted R^2
    e. Standard Error
    f. T-Statistics
    g. Duan Bias Correction Factor
    h. PRESS
    i. Mallow's C_p
    j. MSPE
    k. VIF (for multiple explanatory variables)
    l. P-value for the explanatory variable(s)
d. Calculates estimated SSC values and residuals
e. Gives each data point a data reference number

2. Normal Quantiles

This button calculates the normal quantiles for the residuals of each model. Normal quantiles plots are automatically generated. Possible outliers are identified using quartiles. Upper and lower outliers are defined by:

3^{rd} Quartile + 1.5*Inter-Quartile Range < Residual ≤ 3^{rd} Quartile + 3*Inter-Quartile Range

Or 1^{st} Quartile – 1.5*Inter-Quartile Range > Residual ≥ 1^{st} Quartile – 3*Inter-Quartile Range

The extreme upper and extreme lower outliers are defined by:

Residual > 3rd Quartile + 3*Inter-Quartile Range

Or Residual < 1st Quartile – 3*Inter-Quartile Range

The user is asked to review the outliers, models, and plots to determine whether or not some possible outliers should be removed. Points can be removed by analyzing columns BK through BP, and typing the work “yes” into column BQ in the row of the corresponding data reference number. Extreme outliers are marked in red. If the user determines that no points should be removed, then button #4 is unnecessary. This should be done for each Regression and Analysis sheet.

3. Choose Candidate Model

This button allows the user to identify up to six models that fit the data well. A number of Userforms will guide the user through selecting the candidate model. The user first selects whether a simple linear regression (slr) or multiple linear regression (mlr) model should be used. For a sample of the Userform see Image 4 in Appendix. The user then selects the transformed or untransformed model to be transferred. If there is a possibility that another model is a more accurate fit the program will show a new Userform explaining that another model may be more accurate while displaying the two models’ graphs along with the statistics that support it. For the comparison of slr to mlr the lowest MSPE should be used. The program will only suggest a change if the MSPE of another model is less than 90% of the chosen model’s MSPE. For the comparison of transformed and untransformed the PPCC closer to one should be used. The user will then have the option of switching models or continuing with the originally chosen model. The program will then start the process over until all models that might be more accurate are proposed. Once a candidate model is chosen the corresponding information will be transferred to a candidate model sheet along with the appropriate graphs and data sets used in that model. For samples of Userforms see Images 5 and 6 in Appendix. There are a total of six candidate model sheets. The program will loop back to the first model after the sixth model has been filled and overwrite any data on that sheet.

This does not have to be done for each Regression and Analysis sheet, only when the user wishes to choose a model.

4. Delete Points

If the user has indicated that a point is to be deleted, this button deletes the point(s) and carries the rest of the data over to the next Regression and Analysis sheet. This entire process (buttons 1-4) can be repeated on all Regression and Analysis sheets except for sheet R & A 4, which will not delete points.

Clear Computed Results

This button is on each of the Regression and Analysis sheets and clears all of the results from buttons 1 and 2. A similar button is on each candidate model sheet and it will delete all data that has been transferred to that sheet. For each Regression and Analysis sheet, excluding Regression and Analysis 1, once the computed results are cleared the sheet is hidden from the user to simplify the tabs located at the bottom of the screen.

SSC/SSL Model Switch

This button in on each of the Regression and Analysis sheets, by clicking on this button the user switches between analyzing the data relative to SSC and SSL. The button label will change to display which model is active. When the button is clicked a message box will pop us asking the user if he/she would like to switch models. The "yes" button switches the model and "no" cancels the macro. The SSC model should be used the majority of the time, the top two rows on the candidate models only apply if the SSC model is in use. The SSL should only be used if there is a bias for the SSC model.

For samples of buttons see Image 7 in Appendix.

Candidate Model

There are a total of six candidate model sheets that can be created from any of the four Regression and Analysis sheets. Each candidate model's layout goes as follows:

The top two rows hold all statistical and numerical information for easy transition to the USGS *National Real-Time Water Quality* website. The top row consists of the labels, while the second row holds the data which can be copy and pasted for website use. Next the layout matches the *National Real-Time Water Quality* model info tab for computed suspended sediment. It displays the model equation, model calibration, explanatory variables, and covariance matrix. See Image 8 in Appendix. The candidate model sheet also shows the measured data from its specific Regression and Analysis sheet, including the regression computed data, residuals and normal quantiles. The four corresponding graphs for that model are also included on the sheet.

If a simple linear regression model is chosen an additional graph will be created to compare the accuracy of the transformed vs. the accuracy of the untransformed models. This graph will demonstrate whether the chosen model is bias at lower or higher values.

Cell B5 contains a reference from where the candidate model was created; if this cell is double clicked it will take the user to that Regression and Analysis page.

The candidate model sheet also contains two buttons, see Image 9 in Appendix. The two buttons are:

Clear Results

This button clears all data and graphs from the candidate model sheet. Once the candidate model is cleared it is hidden from the user to simplify the tabs located at the bottom of the screen.

Transfer to Time-Series Sheet

This button transfers the model's equation and measured data to one of two time-series sheets which then computes an estimated SSC or SSL along with upper and lower 90% confidence intervals.

The Regression and Analysis sheet from which the candidate model came from is reported in cell B5 of each candidate model sheet. When double-clicked on it will take the user to that sheet.

Time-Series

There are a total of two time-series sheets, which compute estimated SSC or SSL for a chosen candidate model along with upper and lower 90% confidence intervals.

Once a model is decided on the user should click the "Transfer to Time-Series Sheet" button on that candidate model sheet. The equation and measured data will then be moved to an empty time-series sheet where the estimated SSC and SSL will be calculated along with upper and lower 90% confidence intervals for both SSC and SSL. The loads table located to the right of the time-series table will compute the total loads for each month and then sum the loads for the annual year and for the water year.

The samples, estimated SSC or SSL, upper and lower confidence intervals will then be graphed together so the user can visually compare the estimated data to the measured data. A drop down bar lets the user see either SSC or SSL and a check box lets the user see the data on a log or linear scale.

Each time-series sheet will be filled in increasing order. If both sheets are full a Userform will allow the user to choose which model they would like to replace. The equation used for the time-series will be given for both time-series sheets. If the user does not wish to replace either of the time-series models clicking continue without selecting a model to replace ends the transfer. See Image 10 in Appendix.

The user must input data into columns "J-P" before transferring a candidate model to the time series sheet. The entered time series data will be graphed with the sample data so the user can see graphically the accuracy of the model. If there is no data in the first row of the required columns a message box will be displayed asking the user to input the data before proceeding and no action will be made. Excel graphs are only capable of graphing up to 32,000 data points, if more are entered in the time series columns, anything over 32,000 will not be displayed. For a sample of the Time-Series sheet layout see Image 11 in Appendix.

The candidate model sheet form which the time-series model was generated from is reported in cell D2 of each time-series sheet. When double clicked on it will take the user to that sheet. If the candidate model sheet has been cleared and is hidden, no action will be made. If the candidate model sheet has been replaced the new candidate model will have no relevance to the time series.

Show More Model Estimates

This button brings up a Userform for the user to choose which other model or models, from the same Regression and Analysis page, to show along with the chosen Candidate Model Time-Series. The Userform will display the generic equation of the model, the specific equation of the model, the MSPE, and the PPCC, to help the user decide which model to choose. If no model is chosen from the list then all models except the original will be removed. For a sample of the Userform see Image 12 in Appendix. If the user selects only one model to add to the time series then that model along with its 90% prediction intervals will be calculated and graphed with the chosen Candidate Model and Samples. If the user selects two or more models to add to the time series then all prediction intervals will be hidden to simplify the graph.

Note: The additional models added to the time series graph will not have their specific equation displayed; only the generic equation form will be displayed.

Note: If the Regression and Analysis sheet or Candidate Model sheet in which the time series model was generated has been deleted then there is no way for the program to know where to obtain the equations for additional models, and none will be added.

Interactive User Buttons on Graph

The drop down bar on the time series graph allows the user to choose between displaying the SSC data or the SSL data. When the time series is generated both sets of data will be calculated, the drop down button allows the user to switch between which is viewed on the graph.

The "Log Scale" check box allows the user to switch between a log base 10 scale and linear scale. It is suggested that a log scale is used for log transformed equations and a linear scale is used for an untransformed equation.

The top scroll bar allows the user to scroll through starting dates on the graph. The minimum will start on the first day of the earliest month of the earliest year in the time series dates. Each click of the forward and back arrows changes the start date forward and back by one month. If the empty space to the left or right of the indicator moves the start date back or forward one year, respectively. The starting date is displayed just above the scroll bar.

The bottom scroll bar allows the user to change the period of time displayed on the graph. The shortest period is 15 days and each click of the forward and back arrow increases and decreases the period by 15 days. After the 90 day period, the last position is for a display of one year. A scale is placed under the scroll bar to assist the user. The period length is displayed just above the scroll bar.

For a sample of the interactive graph buttons see Image 13 in Appendix.

X-Section

2.33 Cross-Section

This button will create the same number of cross-section sheets as you have samples. From here, cross-section data can be entered. This step is not required, but may be helpful. To use button 2.67, the date, time and turbidity values are required.

2.67 Cross-Section Summary

This button calculates the average time, average turbidity, median turbidity, maximum turbidity, and minimum turbidity for each cross section entered. These statistics are then pasted in the X-Section Summary sheet and compared with the dates and turbidity values entered by the user. These values are plotted in the X-Section Summary sheet as well. As before, this step is not required, but may be helpful.

Other

There are notes throughout the first frame detailing how some processes are done. Just hold the mouse over the red "flags" to see the notes.

Methods

LINEST

Linest reports regression statistics as follows:

$Slope_m$	$Slope_{m-1}$
Standard $error_m$	Standard $error_{m-1}$
R^2	Standard error y (RMSE)
F-statistic	Degrees of freedom
Regression sum of squares	Residual sum of squares

T-Statistics

A hypothesis test that determines whether each slope coefficient is useful. The higher the T-stat is, the more useful it is.

$(slope_m)/(standard\ error_m)$

Adjusted R-Squared

The Adjusted R-squared in an R-squared adjusted for the number of explanatory variables in the regression model. The regression model with the highest adjusted R-squared is identical to the regression model with the lowest standard error.

$Adj\ R^2 = 1-(1-R^2)*(n-1)/(n-p)$ where p is the number of explanatory variables plus one and n is the number of observations

Estimated SSC

$Y_i(estimated) = (slope_1)*(x_{1,i})+b$ SLR

$Y_i(estimated) = (slope_1)*(x_{1,i})+(slope_2)*(x_{2,i})+b$ MLR

Residuals

$e_i = (Actual\ y)-y_i(estimated)$

Duan Bias Correction Factor

A smearing factor for the transform. This smearing estimator is based on each of the residuals being equally likely, and smears their magnitude in the original units across the range of x.

$BCF = sum(f^1(e_i))/n$ for i = 1 to n

f^1 is the inverse of the transform function used, and n is the number of observations

Variance Inflation Factor

A measure of co linearity used in multiple linear regression (MLR). Multi co linearity occurs when two of more explanatory variables are related to each other. There are serious problems when VIF > 10 and one should be cautious when VIF > 5.

$VIF = 1/(1-R_j^2)$ where R_j^2 is the R-squared from the regression of the jth explanatory variables with all other explanatory variables

Prediction Error Sum of Squares

A validation-type estimator of error. PRESS uses n-1 observations to develop the regression model, then estimates the value left out. The prediction error are then squared and summed. Minimizing PRESS means the model the least error when making new predictions.

$PRESS = sum(e_i^2)$ where $e_i = y_i$-(Actual y)

Mallow's C_p

Mallow's C_p is designed to minimize bias and to minimize the standard error by keeping the number of coefficients small. The best model is the one with the lowest C_p.

$C_p = p+(n-p)*(sp^2-\sigma^2)/\sigma^2$ where n is the number of observations, p is the number of explanatory variables plus one, sp is the standard error of this p coefficient model, and σ is the standard error of the model with the highest p

Model standard Percentage Error

MSPE can be used to compare any regression model. The lowest MSPE corresponds to the model with the least uncertainty associated with regression computed values. Far an SLR model with an MSPE greater than 20, consider and MLR model.

For log units

$\text{Upper MSPE} = 100*(10^{RMSE}-1)$

$\text{Lower MSPE} = -100*(1-10^{-RMSE})$

For regular units

$MSPE = \pm(RMSE/y_{avg})*100$

References

Heisel, D.R., and R.M. Hirsch, 2002, Statistical Methods in Water Resources, U.S. Geological Survey Report

Hurvich, C.M., and Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika, Vol 76.pp.297-307

NOTE: These statistics are not appropriate for comparing models with different units of y.

Appendix

Site No:	12345678			Site Name:	No Name		
						Calibration Dataset	
1				Measured data			
Date + Time	Date	Time	Turbidity	Streamflow, ft^3/s	SSC, mg/L	% Sand	SSL, lb/s
1/19/2000 10:45 AM	1/19/2000	10:45 AM	33	2720	47		7.977216
2/23/2000 10:15 AM	2/23/2000	10:15 AM	50	2430	38		5.762016
3/7/2000 10:40 AM	3/7/2000	10:40 AM	100	4960	114		35.283456
4/6/2000 8:50 AM	4/6/2000	8:50 AM	250	5680	304		107.747328
5/10/2000 9:20 AM	5/10/2000	9:20 AM	88	3020	199		37.501152
6/8/2000 11:30 AM	6/8/2000	11:30 AM	110	2190	126		17.218656
7/6/2000 10:00 AM	7/6/2000	10:00 AM	1600	12300	1610		1235.7072

Image 1: Data Entry Sample

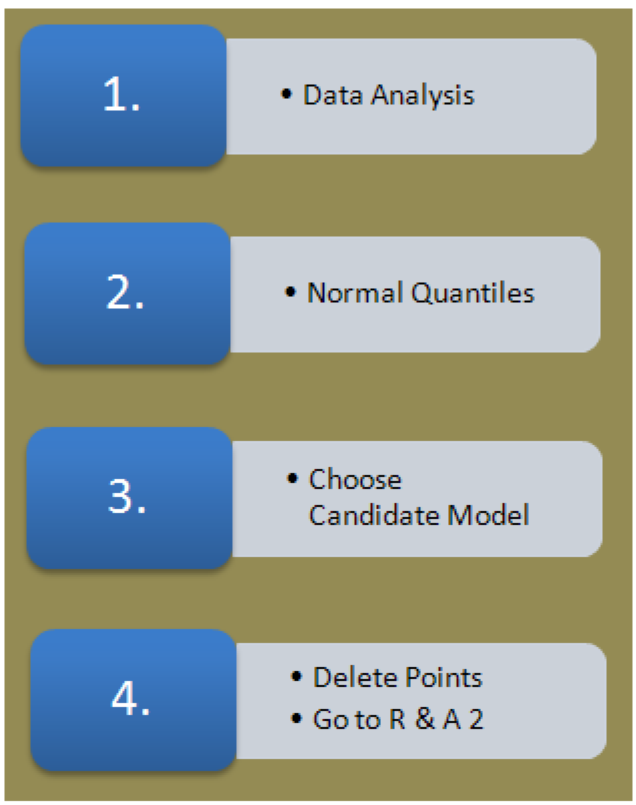


Image 2: Regression and Analysis Buttons

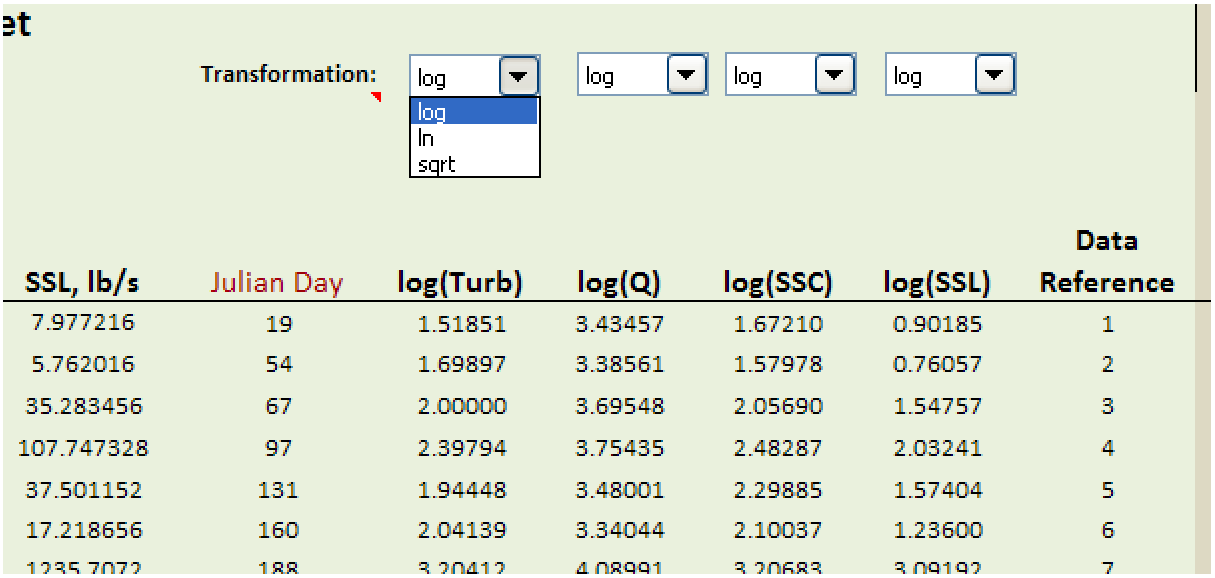


Image 3: Drop Down Box Sample

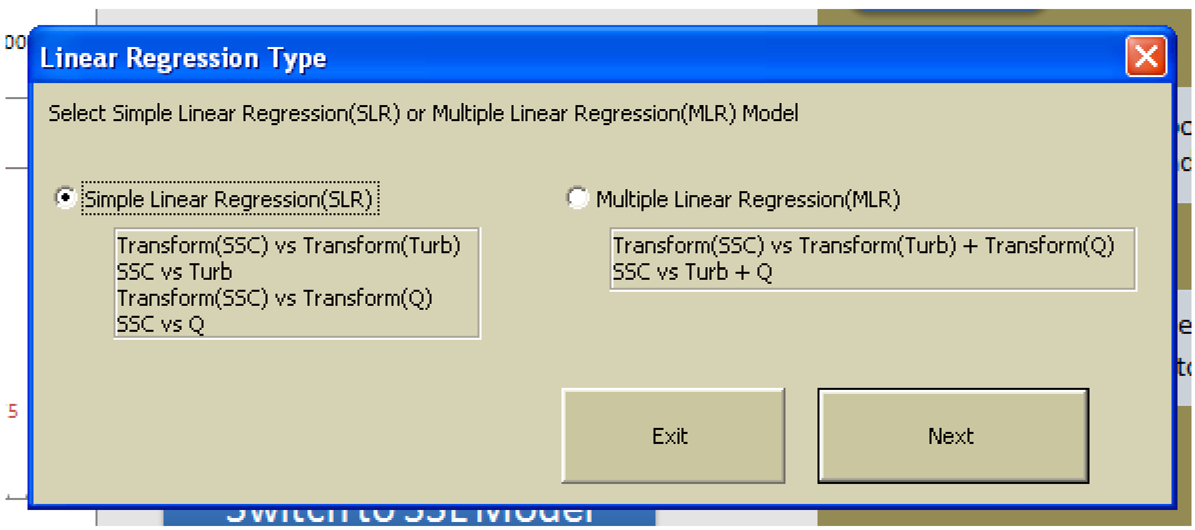


Image 4: Linear Regression Type Userform Sample

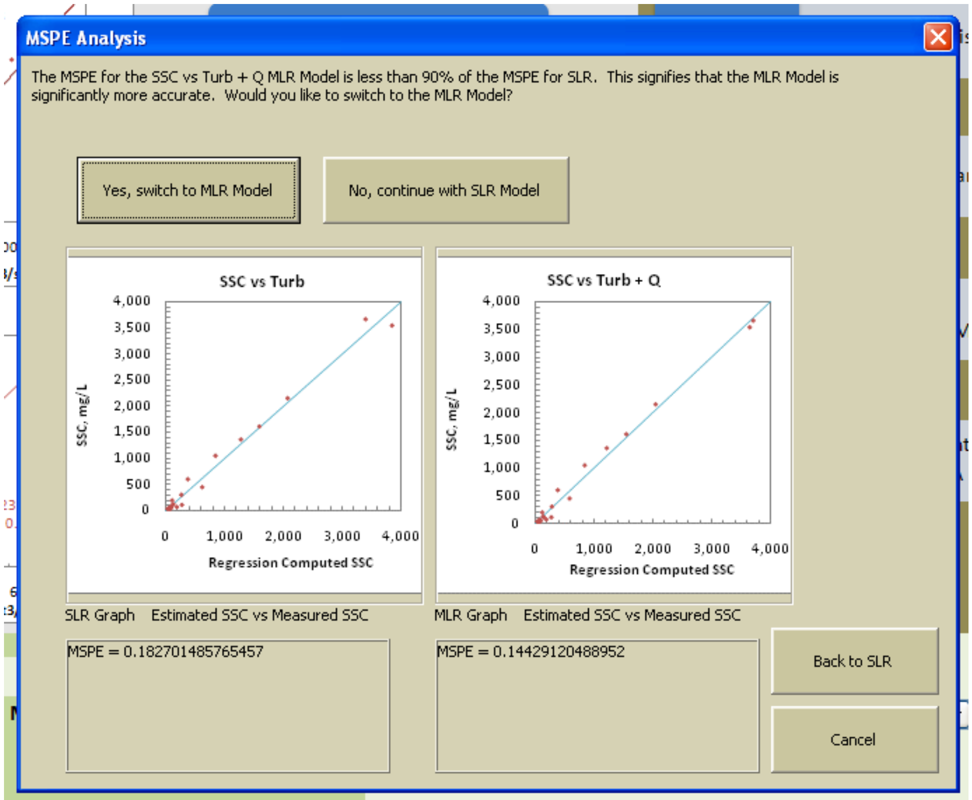


Image 5: MSPE Analysis Sample

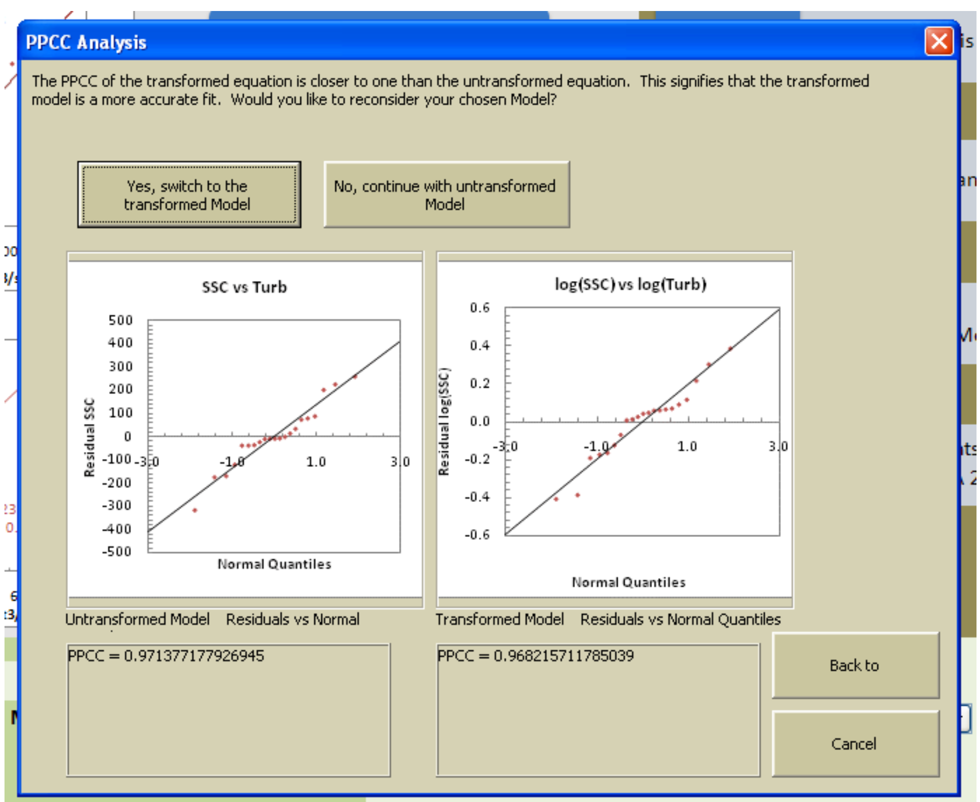


Image 6: PPCC Analysis Sample

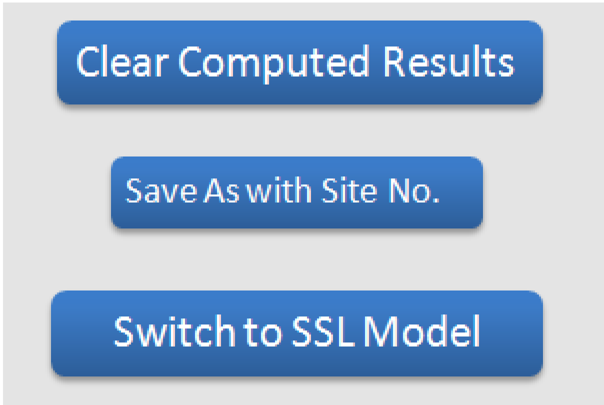


Image 7: Regression and Analysis Buttons Sample

	A	B	C	D	E	
3	**Model Form**					
4	log(SSC) =	0.194	+	0.927	* log(Turb)	
5	Page:	Regression and Analysis 1				
6	**Model Calibration**					
7	Basic Data					
8	Number of Measurements:	21				
9	Standard Error:	0.1988				
10	R^2	0.92				
11	Adj R^2	0.92				
12	Duan BCF:	1.1				
13						
14						
15						
16	Explanatory Variables					
17	Variable	Value	Standard Error			
18	Intercept	0.194	1.299			
19	log(Turb)	0.927	15.16			
20						
21						
22	Covariance Matrix					
23		Intercept	log(Turb)			
24	Intercept	1	-0.957			
25	log(Turb)	-0.957	1			
26						
27						
28						
29	Date and Time	Turb	Q	SSC	SSL	lo
30	1/19/2000 10:45	33	2720	47	7.977	
31	2/23/2000 10:15	50	2430	38	5.762	
32	3/7/2000 10:40	100	4960	114	35.28	
33	4/6/2000 8:50	250	5680	304	107.7	

Image 8: Candidate Model Page

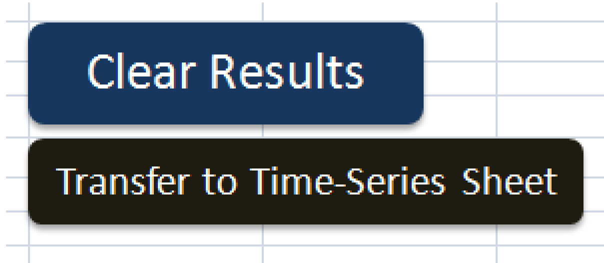


Image 9: Candidate Model Buttons

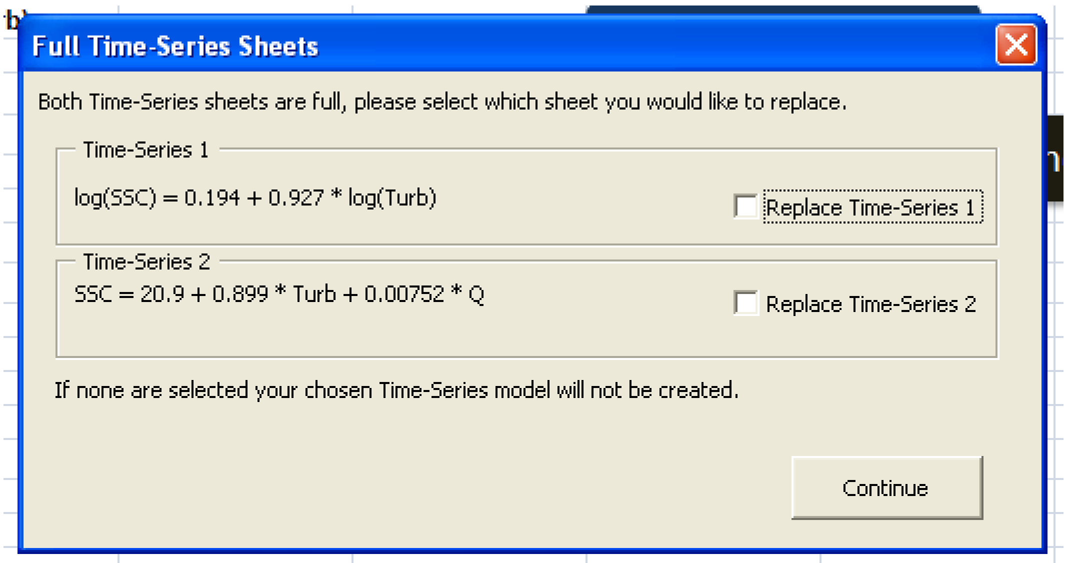
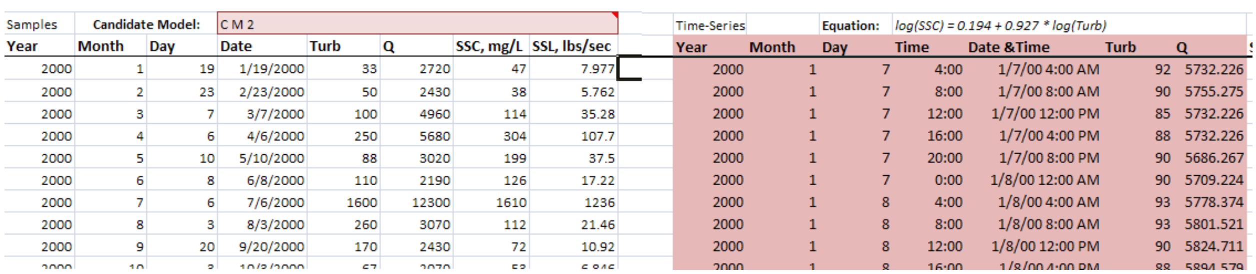


Image 10: Full Time Series Userform



Samples	Candidate Model:		C M 2				
Year	Month	Day	Date	Turb	Q	SSC, mg/L	SSL, lbs/sec
2000	1	19	1/19/2000	33	2720	47	7.977
2000	2	23	2/23/2000	50	2430	38	5.762
2000	3	7	3/7/2000	100	4960	114	35.28
2000	4	6	4/6/2000	250	5680	304	107.7
2000	5	10	5/10/2000	88	3020	199	37.5
2000	6	8	6/8/2000	110	2190	126	17.22
2000	7	6	7/6/2000	1600	12300	1610	1236
2000	8	3	8/3/2000	260	3070	112	21.46
2000	9	20	9/20/2000	170	2430	72	10.92

Time-Series		Equation:	log(SSC) = 0.194 + 0.927 * log(Turb)			
Year	Month	Day	Time	Date &Time	Turb	Q
2000	1	7	4:00	1/7/00 4:00 AM	92	5732.226
2000	1	7	8:00	1/7/00 8:00 AM	90	5755.275
2000	1	7	12:00	1/7/00 12:00 PM	85	5732.226
2000	1	7	16:00	1/7/00 4:00 PM	88	5732.226
2000	1	7	20:00	1/7/00 8:00 PM	90	5686.267
2000	1	7	0:00	1/8/00 12:00 AM	90	5709.224
2000	1	8	4:00	1/8/00 4:00 AM	93	5778.374
2000	1	8	8:00	1/8/00 8:00 AM	93	5801.521
2000	1	8	12:00	1/8/00 12:00 PM	90	5824.711

Image 11: Time-Series Sheet Layout

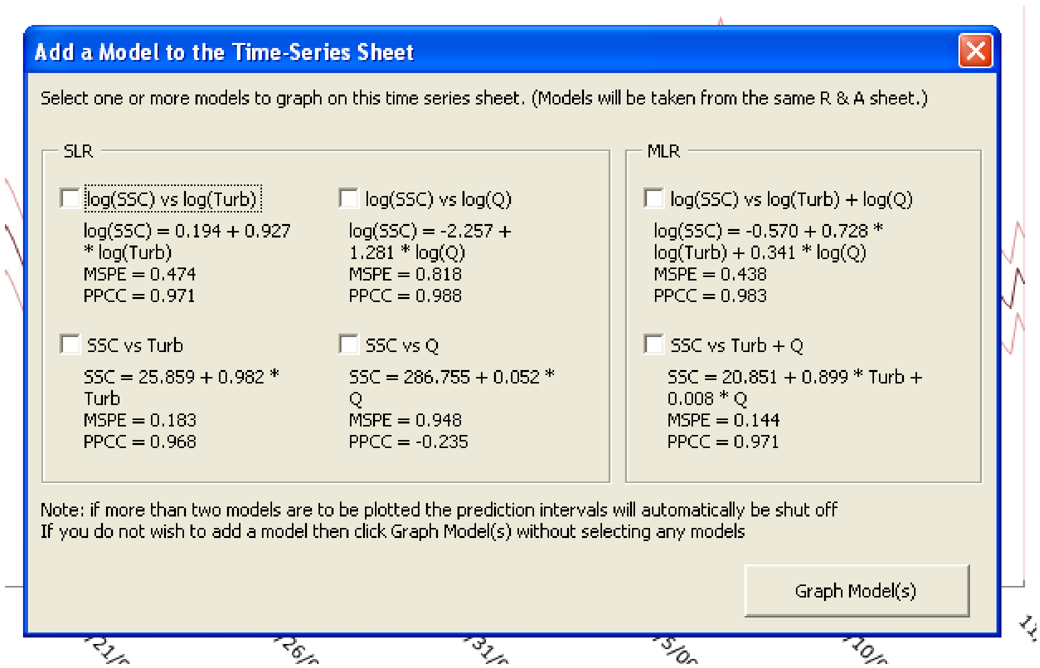


Image 12: Time-Series Add a Model Userform

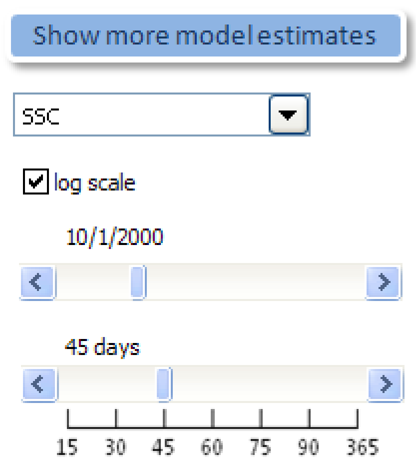


Image 13: Time-Series Interactive Graph Buttons