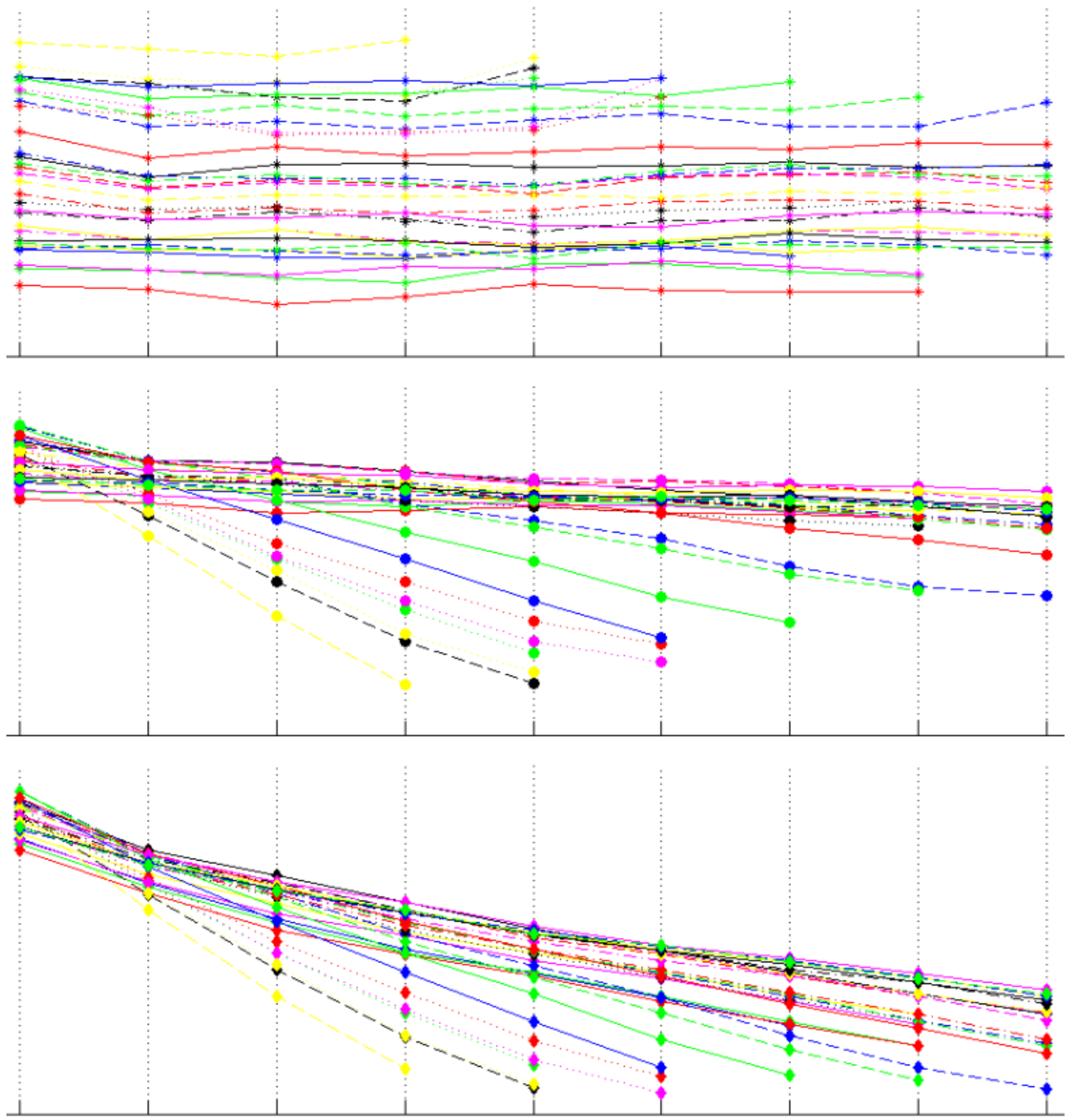
Surrogate Analysis and Index Developer (SAID) Tool



Manual Version

20140528

Software Version

20140528

Please send your questions, comments, and suggestions to

Marian at 217-328-9758 or mdomanski@usgs.gov

and Tim Straub at 217-621-9587 or tdstraub@usgs.gov

Contents

Overview

The Surrogate Analysis and Index Developer (SAID) tool is a MATLAB based tool to assist in the creation of ordinary least squares (OLS) regression models that relate response and predictor variables by providing visual and quantitative diagnostics to the user (Figure 1). The tool also processes acoustic parameters to be used as predictor variables for suspended-sediment concentrations using a constant spatial suspended sediment concentration method. The method utilizes acoustic data from fixed-mount stationary acoustic Doppler velocity meters (ADVM). Within the program, you can set ADVM configuration and processing options, transform a loaded variable, build linear regression models, view linear model diagnostic statistics and plots, export the model information, and generate a predicted time series.

SAID allows you to quickly adjust ADVM processing options, change the variables used in the regression, and evaluate the created model.

The state of the model can be saved into a MATLAB MAT-file to be loaded at a later time within SAID or to perform an analysis of the dataset and linear model using MATLAB.

There is no limit on the number of variables used in the linear model, and there are no restrictions to which variables must be used.

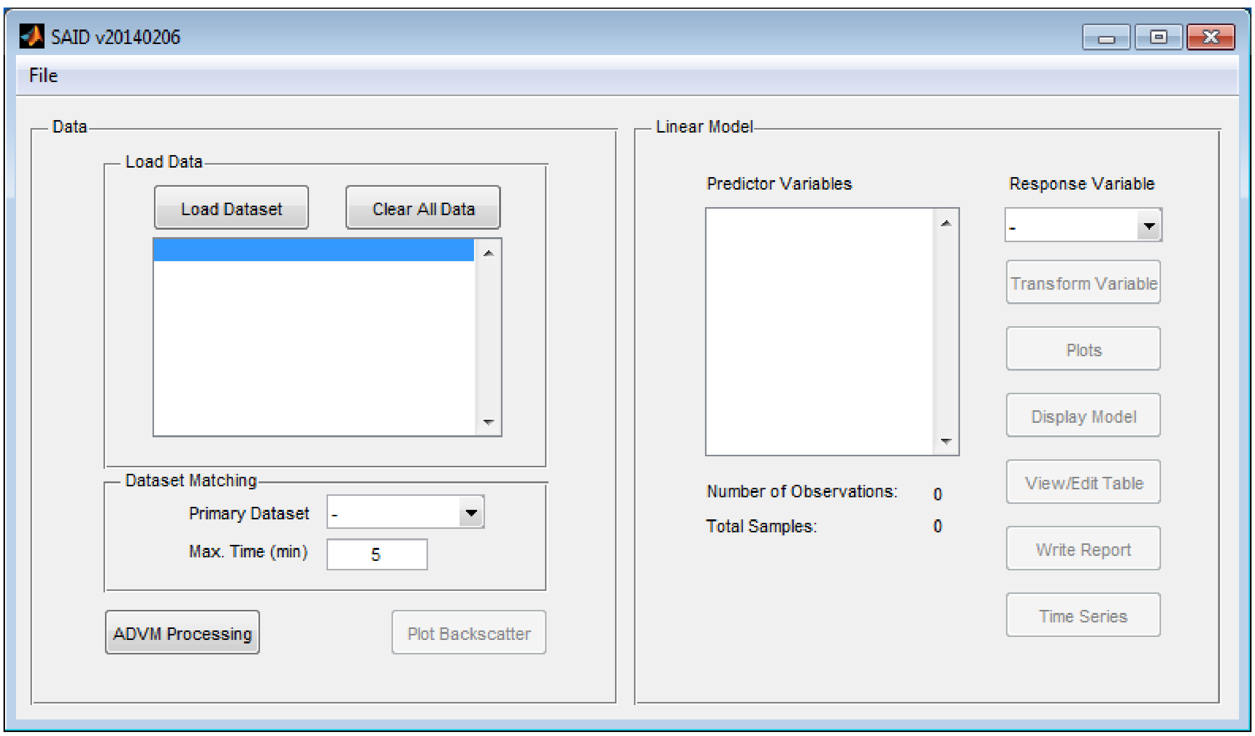


Figure 1 - Main SAID window

An overview of the dataset workflow is briefly described below with a diagram shown in Figure 2.

- **Load datasets**
  - Data that is stored on disk in ASCII files is loaded into memory by SAID.
- **Choose primary datasets**
  - You choose a dataset that serves as the primary time series to synchronize observations to. The chosen dataset is known as the primary dataset and other loaded datasets become the secondary datasets.
  - You can adjust the maximum time difference value (Max. Time) in order to change the upper limit of the time difference that observations are synchronized.
- **Choose linear model variables**
  - Available variables will be displayed in the Predictor Variables list box and the Response Variable drop down list. SAID creates a linear model when a valid set of predictor and response variables are selected.
- **Evaluate linear model**
  - SAID provides several diagnostic plots to determine if the linear model you created fits the assumptions of the OLS method.

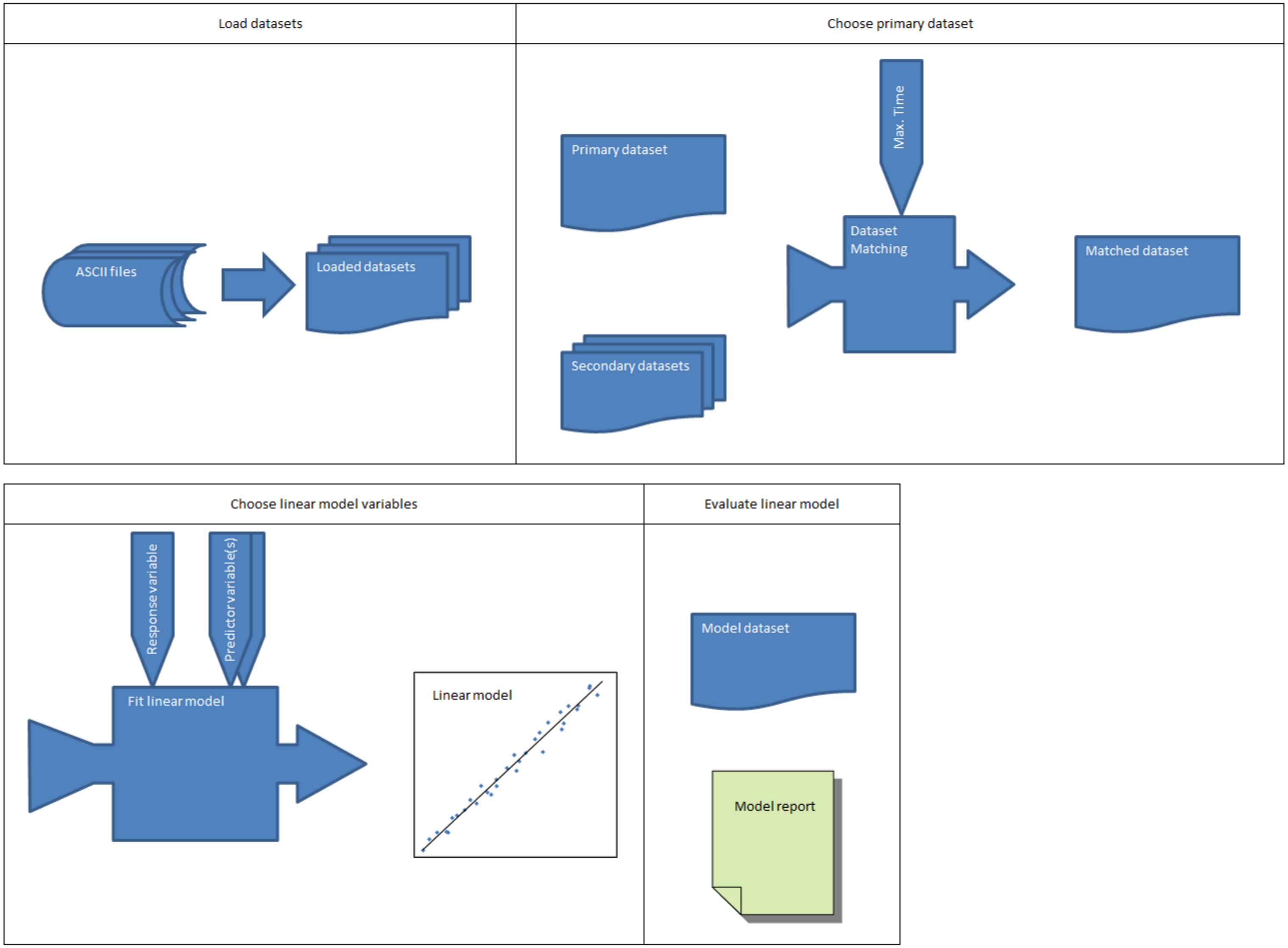


Figure 2 - Diagram of dataset workflow in SAID

Data

Datasets

An observation is a set of coinciding observed values and a dataset is a collection of observations. In SAID, each observation contains a variable with the date and time of the observation as well as observed values for other associated variables measured at that time.

Data Sources

SAID is capable of loading two types of data. The term loaded datasets refers to datasets that contain data that was loaded from disk. The data can be stored within tab delimited ASCII files or a collection of Argonaut ASCII files. Loaded datasets aren't necessarily stored in separate files because of a constituent/surrogate relationship. Having dissimilar time steps is a typical reason for storing and loading datasets separately.

In the format of the tab delimited ASCII file, observations are represented by rows and variables by columns. Variable names are taken from the header row of the text file that contains the dataset information. All time and date information must be present and labeled by headers shown in Table 1.

Header label	Data type
y	Year
m	Month
d	Day
H	Hour
M	Minute
S	Second

Table 1 – Date and time variables that are necessary in a tab delimited ASCII file

The program converts the date and time information into a MATLAB date and time serial number. The resulting variable is named 'DateTime[DatasetName]', where [DatasetName] is the root name of the file loaded. Characters that are invalid in the use of dataset variable names are removed (for example, "." will be change to "_"). All loaded variable values must be numerical.

SAID also has the ability to load ADVM data directly from ASCII files exported by Sontek's ViewArgonaut software. In order for data to be loaded this way, a collection of ASCII files that have the same root file name with .ctl, .snr, and .dat file extensions must be present. The program takes all variables needed to compute the acoustic parameters as well as the ADVM setup information from these files.

Loading data

To navigate to a new directory, click Select New Directory (Figure 3). A dialog box will prompt you to navigate to a new directory to search for dataset files.

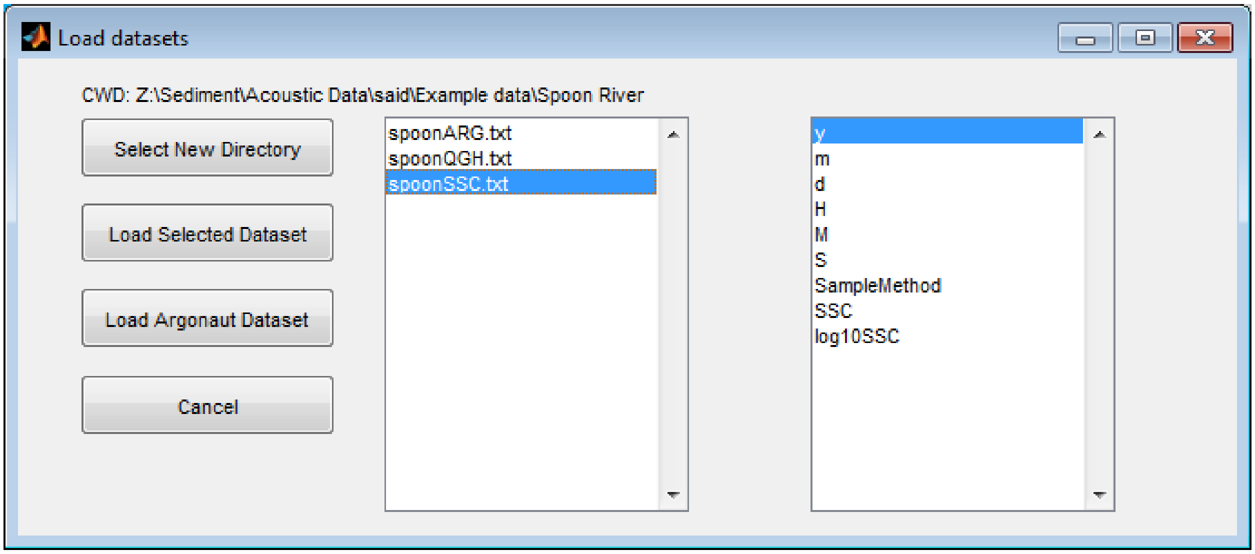


Figure 3 - Load datasets window

Loading tab delimited ASCII files

After clicking the Load Dataset button, a window displaying files with a .txt and .rdb extension is shown. When you select one of these files, the list to the right of the file name list box displays the variable names in the dataset file. The variable names are taken from the first line of the tab-delimited ASCII file. Unless the date and time variables are detected, the program will not load the file.

Loading ViewArgonaut ASCII files

To load an Argonaut ADVM dataset, click on the Load Argonaut Data button within the Load datasets window and you will be prompted for a directory to search for Argonaut datasets. After selecting OK, if datasets are detected in the directory selected, you will be able to select a dataset. If datasets are not detected, you will be notified by a message dialog box and again prompted for a new directory to search.

Dedicated variable names

Variables with names that match the patterns CellXXAmpY and CellXXSNRY (where XX is the cell number, from 00 to 99, and Y the beam number, either 1 or 2) are dedicated variables for backscatter counts and signal-to-noise ratio and are used in the computation of the sediment attenuation coefficient and mean sediment corrected backscatter. These variables are not available for use in the creation of a linear model but are necessary in the computation of the ADVM acoustic surrogate metrics.

Variables named ADVMTemp and Vbeam are also dedicated variables used for the temperature and water depth. The temperature must be in units of degrees Celsius and is directly used in computing the ADVM parameters and is therefore necessary. The water depth is used to determine if the cell is out of water when the vertical orientation is selected in the ADVM Processing dialog box. A minimum Vbeam value is also set by the user in order to exclude samples taken when the water is below a certain depth. (Note, using fixed vertical beam ADVMs to obtain acoustic surrogates may violate an underlying assumption in this method. The method assumes that there are no significant variations in SSC concentration in the acoustic beam for a given measurement).

While a dataset is being loaded, the program checks for variable names that are already loaded. If a variable that is in the dataset that is to be loaded exists in an already loaded dataset, then you will not be able to load the selected dataset.

The names of the successfully loaded datasets will be displayed in the list within the Load Data section of the main SAID window Figure 4).

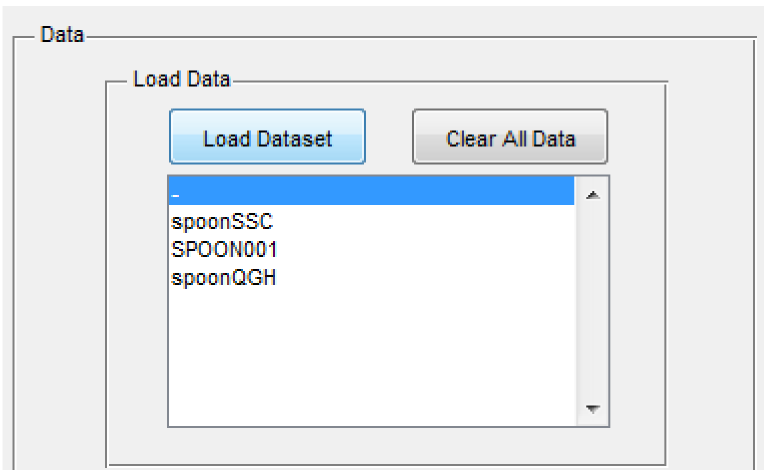


Figure 4 - Load Data interface on the main SAID window

Once the datasets are loaded, the names of the variables that are available for the creation of a linear regression model are shown in the Predictor Variables list under the Linear Model box.

Clicking the Clear All Data button removes the currently loaded datasets from memory and resets the program.

Matching observations

In order to build a linear model, it is necessary that a dataset with observations of predictor (surrogate) and response (water-quality) variables exist. Matching occurs in order to synchronize observations from the loaded datasets. The result of matching is a matched dataset, which is passed to a linear model.

The primary dataset is the loaded dataset that contains the observations whose date-time data will form the basis for matching in the secondary datasets. In a typical application for SAID, the primary dataset is the dataset that contains the suspended sediment concentration observations. You choose which dataset is the primary dataset by selecting it from the Primary Dataset drop-down list (Figure 5). You will not be able to create a linear model until you select a primary dataset.

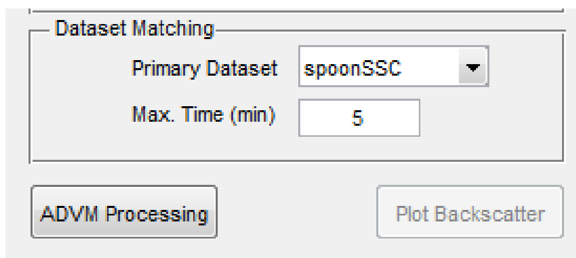


Figure 5 - Dataset matching interface on the main SAID window

Secondary datasets are loaded datasets that contain observations that are matched to primary dataset observations. The term secondary dataset refers to all of the datasets that are not the primary dataset. Observations from the secondary datasets are only copied to the synchronized dataset if they have a date-time that matches a primary dataset observation within the user specified time interval allowance (Max. Time).

Selecting a primary dataset initiates the matching algorithm. For each observation in the primary dataset, SAID calculates the minimum absolute time difference between the observation's date and time variable and the date and time variable of the secondary dataset being compared. If the minimum absolute time difference is less than or equal to the user entered value for the maximum time difference, the observations from the secondary dataset being compared are matched with the observations of the primary dataset and the values are copied to the matched dataset (Figure 6 and Table 2). If the minimum absolute time difference is greater than the user allowed maximum time difference (Max. Time), then the corresponding variables in the observation in the matched dataset are set to an invalid value. In other words, the observation in the primary dataset will not be matched to an observation in the dataset being compared. The decision at the core of the matching process is illustrated in Figure 6. An example showing the matching process and the creation of the matched dataset is presented in Table 2.

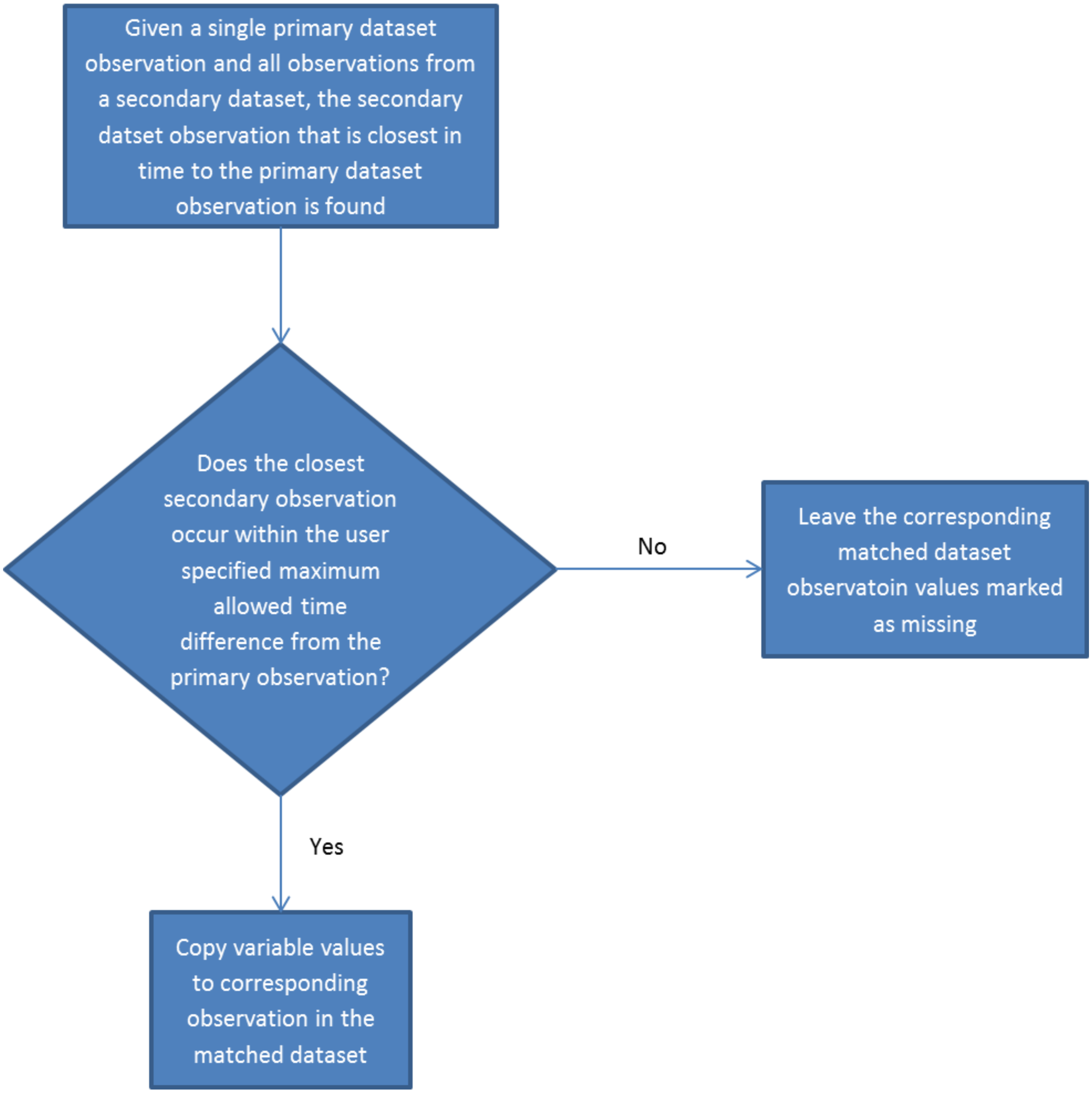


Figure 6 - Flowchart showing the process to match a secondary observation a primary observatoin

Primary dataset			Secondary dataset				Matched dataset			
SSC			ADVM							
Date	Time	Value	Date	Time	Value	If MaxTime = 5 min	Date	Time	SSC Value	ADVM Value
			4/4/2013	0:00	55					
4/4/2013	0:12	30	4/4/2013	0:15	56	Match	4/4/2013	0:12	30	56
			4/4/2013	0:30	64					
			4/4/2013	0:45	71					
4/4/2013	0:53	275	4/4/2013	1:00	75	No Match	4/4/2013	0:51	275	NaN
			4/4/2013	1:15	72					
			4/4/2013	1:30	79					
			4/4/2013	1:45	82					
4/4/2013	1:56	550	4/4/2013	2:00	85	Match	4/4/2013	1:56	550	85

Table 2 - Example of the second step in the matching process (diamond shape of in the flow diagram in Figure 6). The example shows a primary and secondary dataset with a maximum matching time (MaxTime) set to 5 minutes. The resulting matched dataset includes three samples with two observations available for linear model building.

A result of this synchronizing method is that the primary dataset observations occur only once in the matched dataset while the observations in the secondary datasets can occur multiple times. For instance, this can happen when an observation in the secondary dataset is found to be the observation with the minimum absolute difference in time and within the maximum allowed time value for two or more primary observations.

After you select a primary dataset, the program will indicate that it is matching datasets and will remain unresponsive until the matching is complete. The time it takes for the program to create a matched dataset depends on the number of loaded datasets and the number of observations in each dataset.

When the program has completed the matching algorithm, the variables available for use in the linear model are shown in the Predictor Variables and Response Variable lists.

Parameters required for processing backscatter data

The following parameters are required before the backscatter data is processed and the ADVM parameters are computed.

- Configuration
  - Frequency
  - Effective Transducer Diameter
  - Slant Angle
  - Blanking Distance
  - Cell Size
  - Number of Cells
- Processing
  - Intensity Scale Factor (if Amp is selected for Backscatter Values)
  - Minimum Mid-Point Cell Distance
  - Maximum Mid-Point Cell Distance
  - Minimum Vbeam

Once the required parameters in the ADVM Processing window have valid values, the ADVM parameters with at least one valid observation will be available in the Predictor Variables list.

Setting ADVM processing options

By clicking on the ADVM Processing button, you can change the ADVM configuration and processing options used in the calculation of the ADVM parameters (Figure 7).

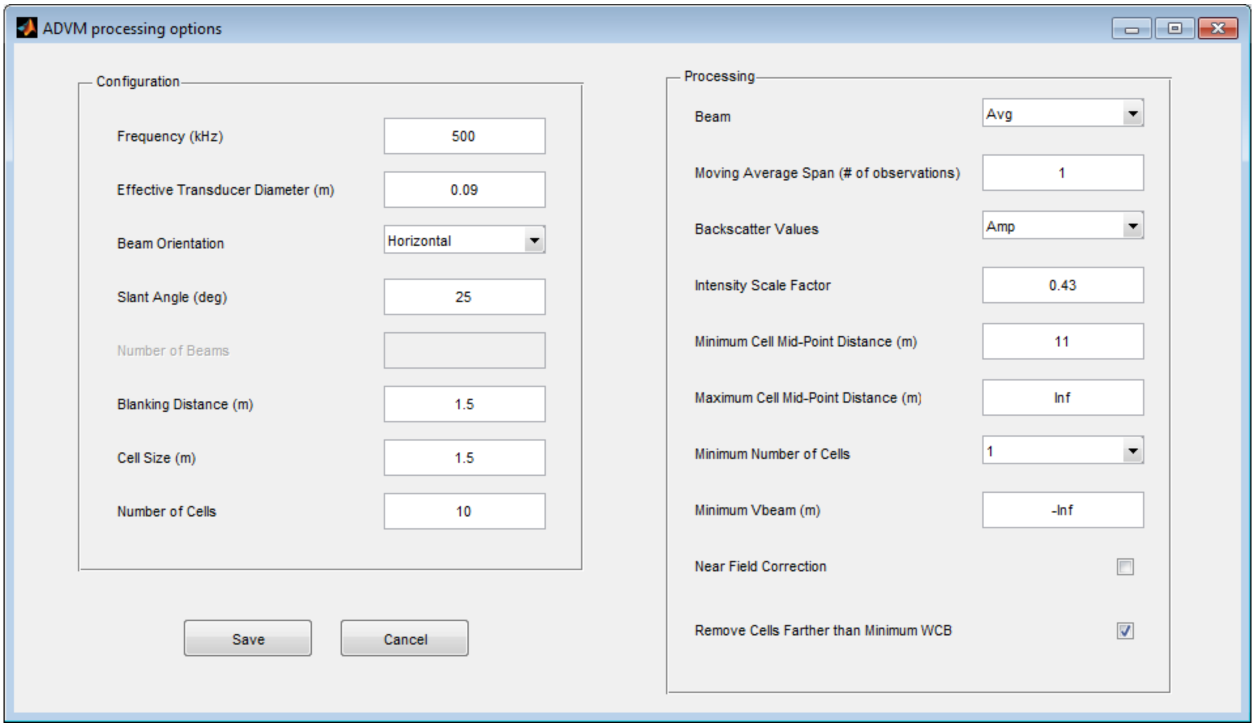


Figure 7 - ADVM deployment configuration and acoustic parameter options window

Configuration

These parameters define the configuration of the ADVM (Figure 8). If the ADVM data is read from an Argonaut dataset, these values are automatically loaded from the .ctl file.

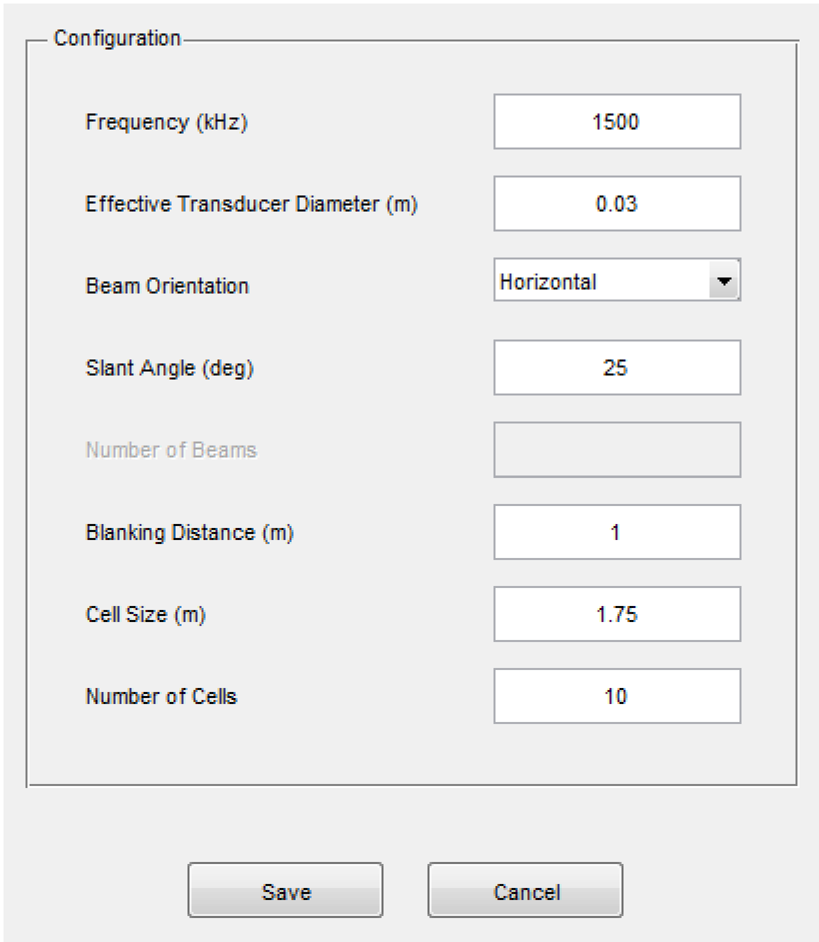


Figure 8 - ADVM deployment configuration settings within the ADVM processing options window

- **Frequency** – The frequency of the ADVM's acoustic signal.

- **Effective Transducer Diameter** – The effective diameter in meters of the ADVM transducer. With Argonauts, the effective diameter corresponds to a given frequency. For Sontek Argonaut ADVMs, these values are as follows (personal communication from SonTek March 14, 2012) (note that these values are not the same as the physical diameter that is measured on the instrument; and that they could change with new hardware versions for these instruments):

    - 3000 kHz – 0.015 m
    - 1500 kHz – 0.030 m
    - 500 kHz – 0.090 m

If any of the above frequency values are inserted into the Frequency field, or the frequency is read from an Argonaut .ctl file, the corresponding default value for the effective diameter is automatically inserted into the Effective Transducer Diameter field. If one of the three frequencies shown above is being used, but the instrument is not an Argonaut, then the manufacturer of the instrument should be contacted to get the effective transducer diameter. The effective transducer diameter is only used when the Near Field Correction option is selected in the Processing section.

- **Beam Orientation** – The orientation of the acoustic beams of the ADVM. If 'Vertical' is selected for this field, then the Vbeam for each observation is compared to the cell edges, and each cell that is out of water is marked as invalid.

- **Slant Angle** – The angle of the acoustic beam with respect to the vector that represents the cell distance from the instrument. This angle, along with the blanking distance, cell size, and number of cells, is used to find the mid-point distance of each cell along the acoustic beam.

- **Number of Beams** – The number of acoustic beams on the instrument. This value is not used. SAID assumes that the instrument has two beams.

- **Blanking Distance** – The distance in meters from the instrument to the beginning of the first cell. This value is used in the computation of the mid-point distance of each cell along the acoustic beam.

- **Cell Size** – The length of each cell in meters. This value is used in the computation of the mid-point distance along the acoustic beam of each cell.

- **Number of Cells** – The number of cells in the configuration of the ADVM under analysis. The number of cells directly affects the values displayed in the Minimum Number of Cells drop down list.

Processing

These parameters control how ADVM backscatter data is screened and processed (Figure 9).

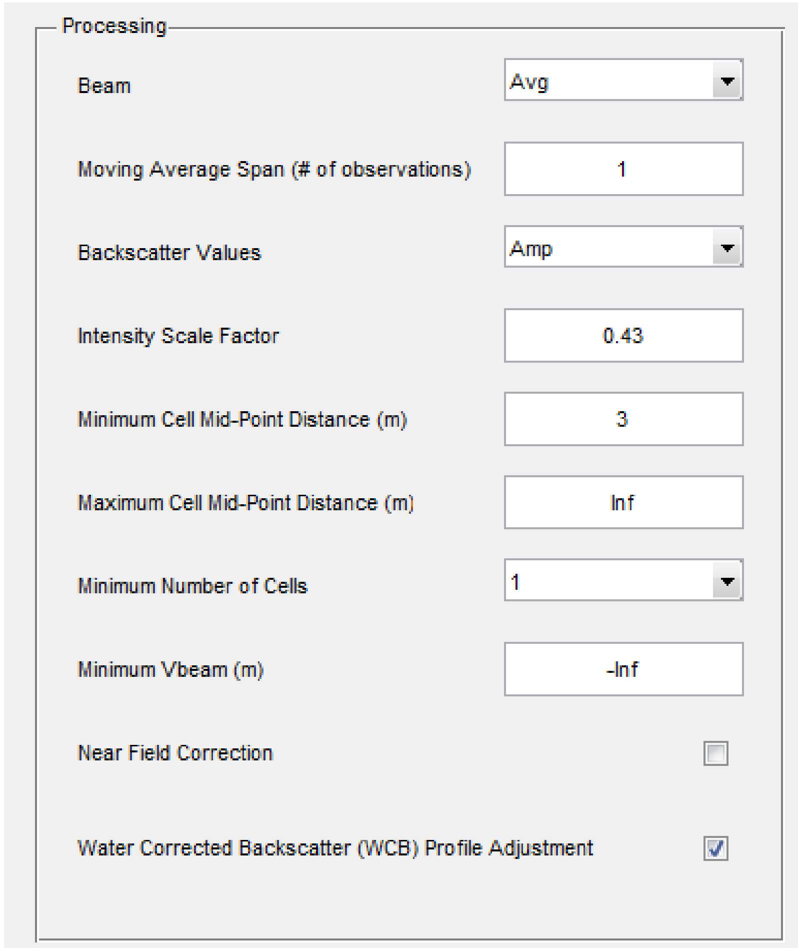


Figure 9 - Acoustic parameter processing options within the ADVM processing options window

- **Beam** – The beam number from which the backscatter values are taken. When 'Avg' is selected for this field, the average cell backscatter values are used.

- **Moving Average Span** – The span, in number of observations, used in a centered moving averaging of the backscatter time series. The span must be an odd positive integer.

  The center averaging method used in SAID is demonstrated in Table 3 and

  Table 4 where the moving average span is **3** observations. Note that the samples taken at 13:00 and 14:00 are time series endpoints and are not averaged.

Table 3 – Example of measured backscatter time series

	Backscatter				
Time	Cell01Amp2	Cell02Amp2	Cell03Amp2	Cell04Amp2	Cell05Amp2
13:00	65.4	53.8	43.9	35.7	28.4
13:15	64.5	54.2	44.3	36.9	30.1
13:30	64.5	53.8	44.3	36.1	28.8
13:45	64.9	53.8	43.4	35.0	29.7
14:00	65.4	53.3	43.9	36.1	30.2

Table 4 – Example of center averaged backscatter time series from Table 3

	Backscatter for Moving Average Span of 3				
Time	Cell01Amp2	Cell02Amp2	Cell03Amp2	Cell04Amp2	Cell05Amp2
13:00	65.4	53.8	43.9	35.7	28.4
13:15	64.8	53.9	44.2	36.2	29.1
13:30	64.6	53.9	44.0	36.0	29.5
13:45	64.9	53.6	43.9	35.7	29.6
14:00	65.4	53.3	43.9	36.1	30.2

- **Backscatter Values** – The backscatter values used in the computation of the ADVM parameters. When 'Amp' is selected, the backscatter values are multiplied by the value in the Intensity Scale Factor field. The Intensity Scale Factor field is made available only when 'Amp' is selected. (Caution: the model developed will be specifically for SNR or AMP units and cannot be switched without building a new model. An empirical (testing for best model using SNR or Amp) and theoretical considerations should be evaluated. The Sontek Argonaut-SL System Manual, 2009 says "If you are trying to relate the acoustic return to the amount of suspended sediment in the water, this data should be assessed as signal strength and not as SNR." In other words, use Amp unless you have no choice but to use SNR)

- **Intensity Scale Factor** – The scaling factor to convert backscatter counts to decibels. This field is only available when 'Amp' is selected in the Backscatter Values drop down list. The factor defaults to 0.43 (typical for Sonteks); but should be taken from manufacturer literature for specific ADVMs.

- **Minimum Cell Mid-Point Distance** – The minimum distance in meters from the transducer that the mid-point of a cell has to be in order for it to be used in the computation of the ADVM parameters.

- **Maximum Cell Mid-Point Distance** – The maximum distance in meters from the transducer that the mid-point of a cell can be in order for it to be used in the computation of the ADVM parameters.

- **Minimum Number of Cells** – The required minimum number of valid cells that an ADVM sample has to have in order for its computed parameter to be included as an observation in the linear model.

- **Minimum Vbeam** – The minimum value for Vbeam that a sample must have in order for it to be used as an observation. Vbeam is the water height in meters that the ADVM reports.

- **Near Field Correction** – When the box is checked, a near field correction to the backscatter values is made (Downing and others, 1995). When the box is not checked, no near field correction is applied. In general, data from the near field should be avoided by setting the blanking distance and/or Minimum Cell Point distance greater than the near field for a given instrument.

- **Water Corrected Backscatter (WCB) Profile Adjustment**– When this box is checked, the range of cells that include and are beyond the cell with the minimum water corrected backscatter (minWCB) are not included in the calculation, unless the cell with the minWCB is the last or first cell in the range considered.
    - If the cell with the minWCB is the last cell, the value is retained and all cells are used to calculate the sediment corrected backscatter and attenuation coefficient.
    - If the cell with the minWCB is the first cell, all other cells are not considered, and the water corrected backscatter value in the first cell is used as the sediment corrected backscatter value for the observation, and no attenuation coefficient is calculated.

The following two figures (Figure 10 and Figure 11) demonstrate this feature. The first figure shows a backscatter profile of a single sample without Remove Cells Farther than Minimum WCB feature enabled. As shown, the water corrected backscatter of the seventh cell is the minimum value. In Figure 11, with the feature enabled, the backscatter values including and beyond the seventh cell are set to invalid values and are not plotted or used further in the calculations.

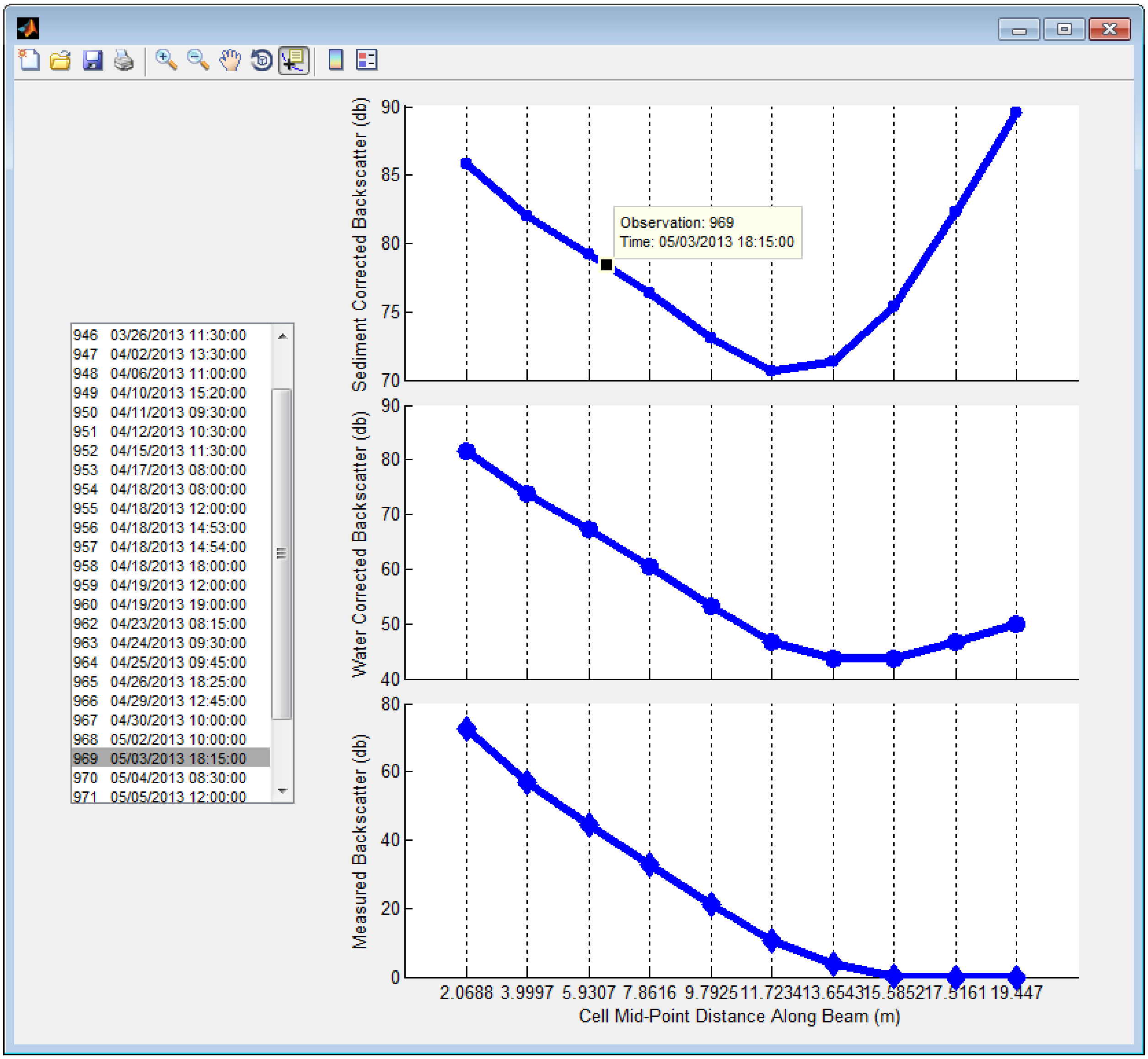


Figure 10 - Backscatter profile of a single ADVM sample without the Remove Cells Further than Minimum WCB option enabled

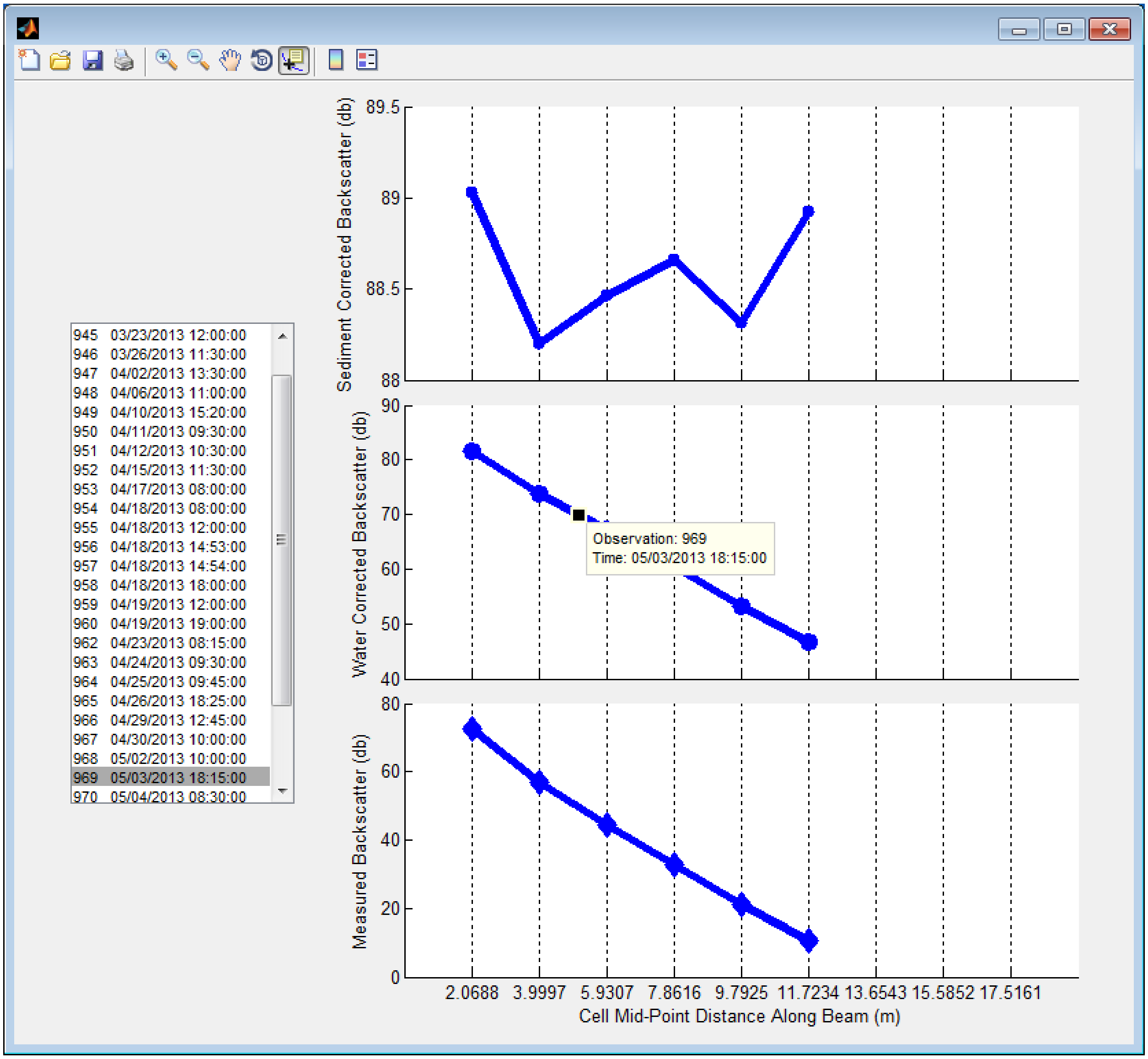


Figure 11 - Backscatter profile of a single ADVM sample with the Remove Cells Further than Minimum WCB option enabled

Viewing backscatter profiles

When a valid Response Variable is matched with valid Predictor Variables, the Plot Backscatter button will be made available. When this button is clicked, a window with three sets of axes is displayed. From the top, the axes show Sediment Corrected Backscatter (SCB), Water Corrected Backscatter (WCB), and Measured Backscatter (MB), all in decibels, versus the cell mid-point distance along the acoustic beam (Figure 12). Also shown in the window is a list of observation numbers and times from the model. The observation times are taken from the primary dataset. Only the backscatter samples that correspond to observations in the linear model are shown. Selecting sample times in the list displays the plots of the backscatter values on the axes. Multiple observations can be selected and plotted.

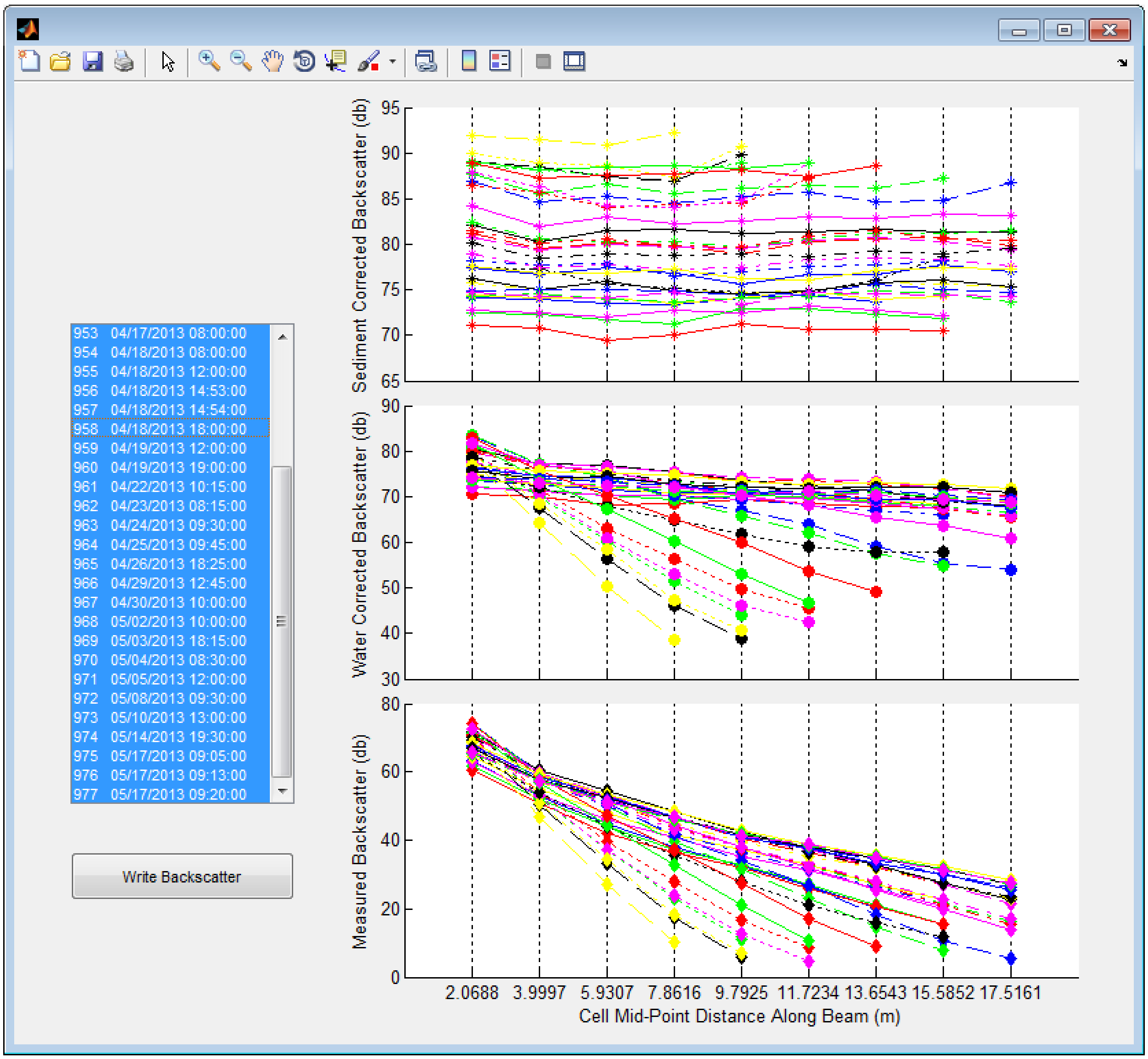


Figure 12 - Backscatter profile plotting window

Clicking on the data cursor icon enables data cursor mode. When this feature is enabled, you can click on a backscatter profile line to select it. In the axes that you select a line, the observation number and date and time of the sample is shown in a label box near the place you clicked it. The observation number corresponds to the same observation number that is shown in the observation table that pops up when you click on the View/Edit Table button, as well as the observation numbers shown in the model diagnostic plots.

The Write Backscatter button will allow you to write the measured backscatter, water corrected backscatter, and sediment corrected backscatter observations shown in the window to a csv format file.

A note on invalid cells

Invalid cells are those that have invalid values. Throughout the process, cells are assigned invalid values when they

- have erroneous or incomplete input
- are missing (e.g., cells 1 and 4 are loaded with valid data and cells 2 and 3 are not, so cells 2 and 3 are loaded as invalid values)
- are outside of the Minimum and Maximum Cell Mid-Point Distance range
- are equal to and are farther than the cell with the minimum water corrected backscatter when the Remove Cells Farther than Minimum WCB box is checked in the ADVM processing options window
- have a portion greater than the sample Vbeam value when ‘Vertical’ is selected for Beam Orientation in the ADVM processing window (in other words, the cell is out of water)

Transforming variables

Once the Match Variable is selected, and the available variables for use in the model appear in the Predictor Variable list and Response Variable drop down list, the Transform Variable button is made available. This feature provides the option to transform a loaded variable using a transform function (Figure 13).

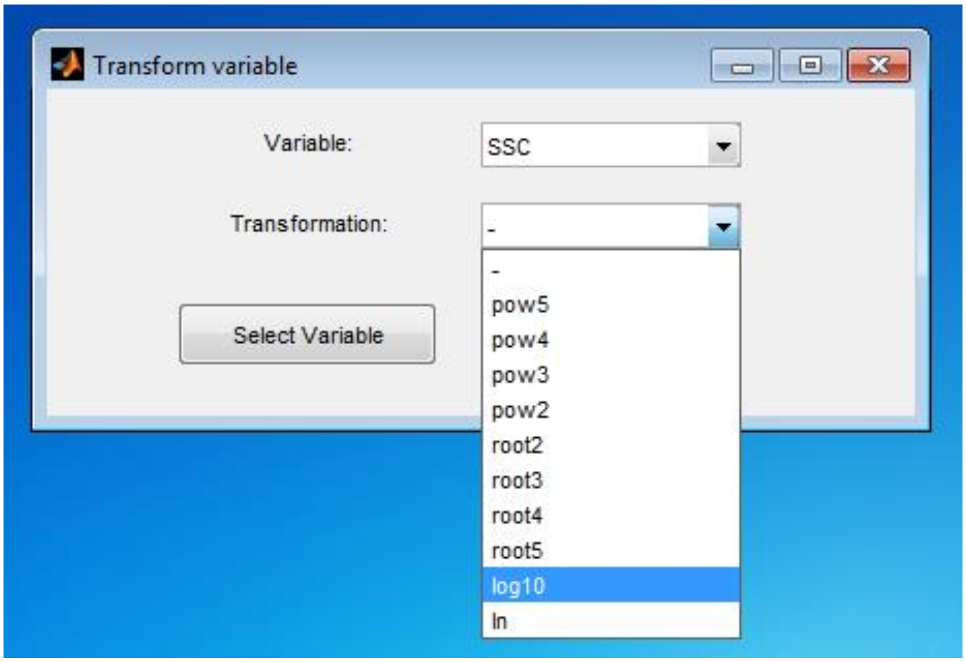


Figure 13 - Transform variable window

After clicking on the Transform Variable button, a small window pops up. To transform a variable, select a variable from the Variable drop down list and a transformation function from the Transformation drop down list. In the current release of SAID, the available transformations are power from 2 to 5, base 10 and the natural logarithms, and the 2nd to the 5th root.

Click on the Select Variable button after the desired selections have been made. The transformed variable will be available as a selection in the Predictor Variable list and the Response Variable drop down list. The new variable will have a prefix that indicates the transformation. For example, a variable named SSC that has been transformed with the base-10 logarithm transformation function will have the name log10SSC.

Linear Model

There is no limit on the number of variables used in the creation of an OLS (Ordinary Least Squares) linear regression model, and there are no restrictions to which variables must be used (Figure 14).

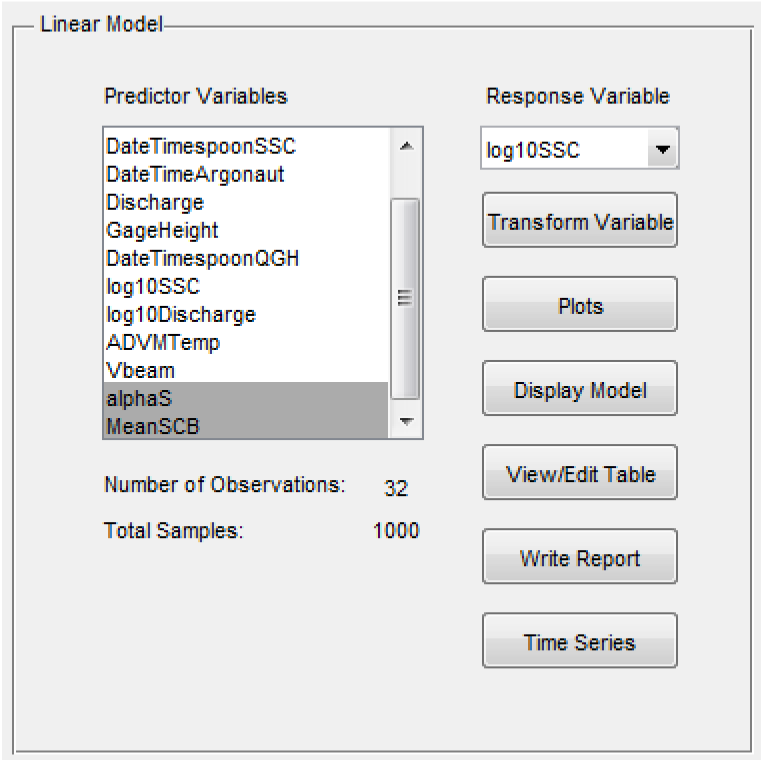


Figure 14 - Linear Model options on the main SAID window

Choosing variables

As datasets are loaded, and if the Match Variable is a valid selection, the variables that are available for use in creation of the linear model are shown in the Predictor Variables list and the Response Variable drop down list. Selecting a variable in the Predictor Variables list, then one from the Response Variable drop down list, creates a model. Selecting the variable that is used for the Response Variable in the Predictor Variables list deselects the predictor variables and resets the response variable selection to the first in the list.

After a model is successfully created by selecting variables, a user can begin to evaluate the model results. This program includes tools to assist in model evaluation, available using the Plot Backscatter, View/Edit Table, Display Model, Write Report, Plots, and Time Series buttons. Also, the number of observations used in the model is shown next to the Number of Observations label.

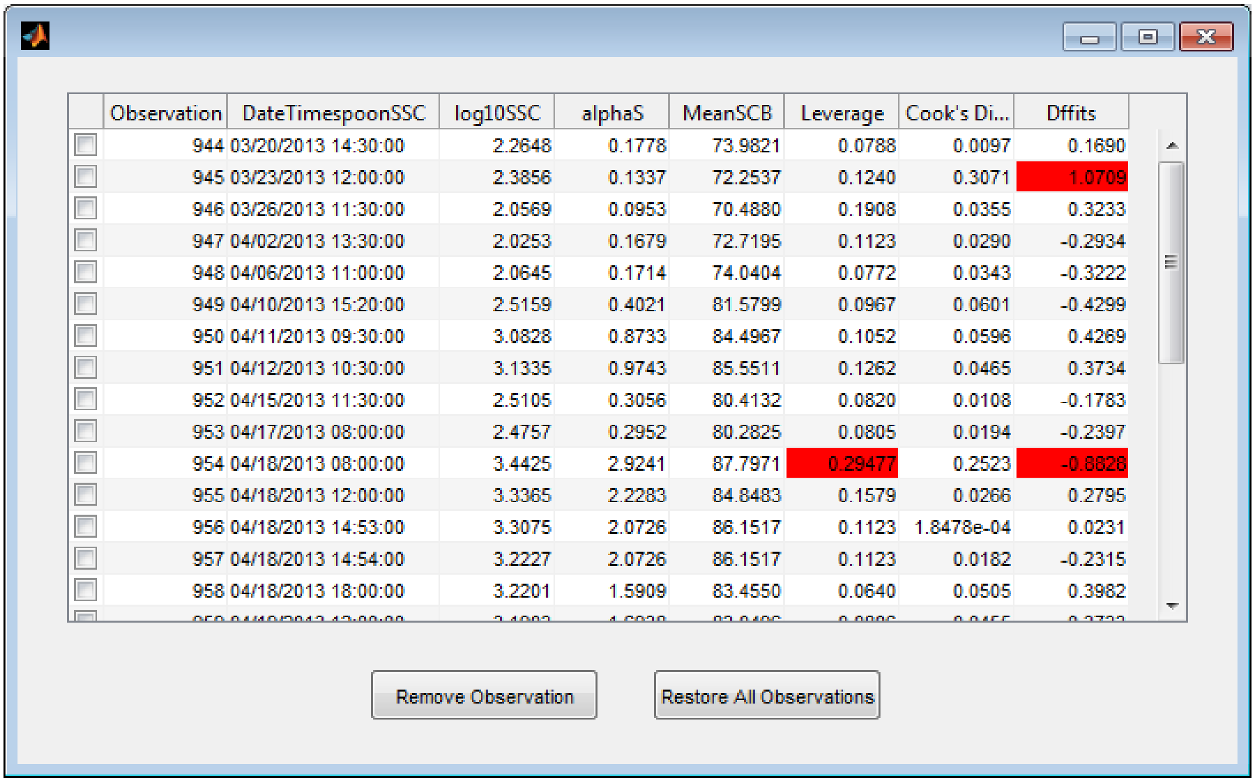
Number of Observations

If a valid linear model exists within the program, the Number of Observations field will show the number of samples used in the creation of the linear model. This corresponds to the number of valid observation values for the selected variables within the primary dataset.

Total Samples

The Total Samples field shows the total number of samples in the loaded dataset. This number corresponds to the number of samples in the dataset that is selected in the Match Variable drop down list.

Viewing and removing observations from the model dataset



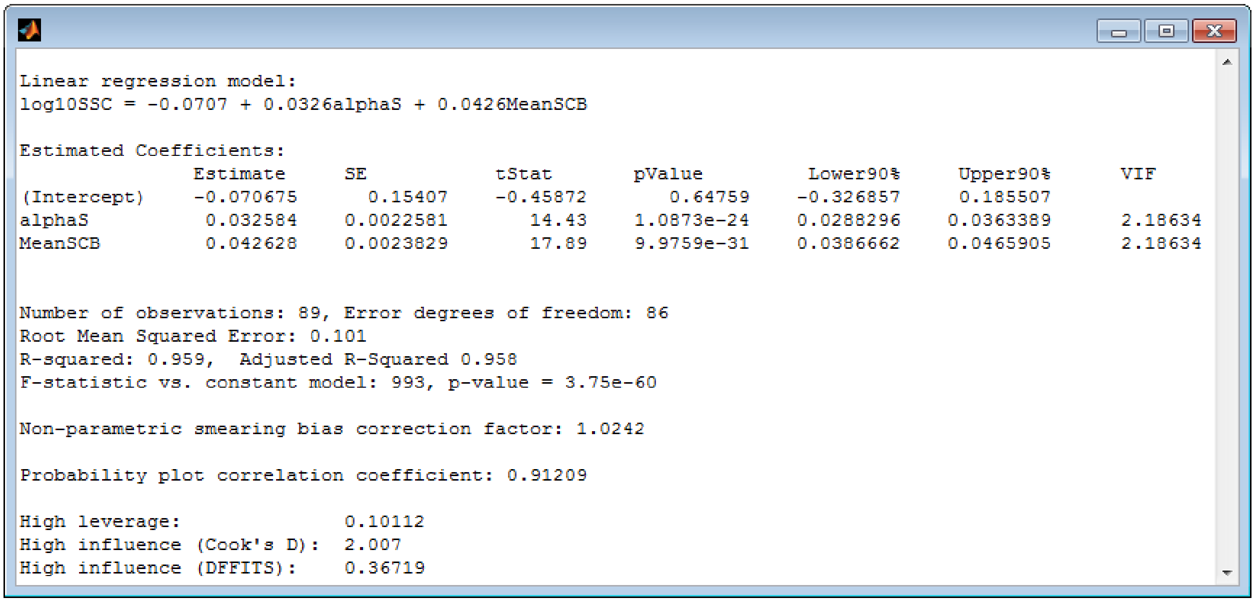
	Observation	DateTimespoonSSC	log10SSC	alphaS	MeanSCB	Leverage	Cook's Di...	Dffits
☐	944	03/20/2013 14:30:00	2.2648	0.1778	73.9821	0.0788	0.0097	0.1690
☐	945	03/23/2013 12:00:00	2.3856	0.1337	72.2537	0.1240	0.3071	1.0709
☐	946	03/26/2013 11:30:00	2.0569	0.0953	70.4880	0.1908	0.0355	0.3233
☐	947	04/02/2013 13:30:00	2.0253	0.1679	72.7195	0.1123	0.0290	-0.2934
☐	948	04/06/2013 11:00:00	2.0645	0.1714	74.0404	0.0772	0.0343	-0.3222
☐	949	04/10/2013 15:20:00	2.5159	0.4021	81.5799	0.0967	0.0601	-0.4299
☐	950	04/11/2013 09:30:00	3.0828	0.8733	84.4967	0.1052	0.0596	0.4269
☐	951	04/12/2013 10:30:00	3.1335	0.9743	85.5511	0.1262	0.0465	0.3734
☐	952	04/15/2013 11:30:00	2.5105	0.3056	80.4132	0.0820	0.0108	-0.1783
☐	953	04/17/2013 08:00:00	2.4757	0.2952	80.2825	0.0805	0.0194	-0.2397
☐	954	04/18/2013 08:00:00	3.4425	2.9241	87.7971	0.29477	0.2523	-0.8828
☐	955	04/18/2013 12:00:00	3.3365	2.2283	84.8483	0.1579	0.0266	0.2795
☐	956	04/18/2013 14:53:00	3.3075	2.0726	86.1517	0.1123	1.8478e-04	0.0231
☐	957	04/18/2013 14:54:00	3.2227	2.0726	86.1517	0.1123	0.0182	-0.2315
☐	958	04/18/2013 18:00:00	3.2201	1.5909	83.4550	0.0640	0.0505	0.3982

Figure 15 - Observation table window

Clicking on the View/Edit Table button will display a window that contains the observation information used in the model. The information shown is the observation number, the corresponding primary date and time variable, the response variable, and the predictor variables. Also shown are values diagnostic statistics including leverage, Cook's distance, and DFFITS for each observation for outlier detection. Calculated indicator values that exceed the corresponding critical values for the model are highlighted in red.

Observations can be removed by checking the boxes in the far left column and clicking the Remove Observation button. This action flags the date and time within the program and sets the variables that correspond to the date and time to an invalid value. Once a date and time is flagged as removed, any future variable that is used in the model will have the corresponding values set to invalid. This will continue until the Restore All Observations button is clicked, which clears the date and times flagged.

Viewing model statistics



```
Linear regression model:
log10SSC = -0.0707 + 0.0326alphaS + 0.0426MeanSCB

Estimated Coefficients:
                   Estimate     SE           tStat       pValue        Lower90%      Upper90%      VIF
(Intercept)        -0.070675    0.15407      -0.45872    0.64759       -0.326857     0.185507
alphaS              0.032584    0.0022581    14.43       1.0873e-24    0.0288296     0.0363389     2.18634
MeanSCB             0.042628    0.0023829    17.89       9.9759e-31    0.0386662     0.0465905     2.18634


Number of observations: 89, Error degrees of freedom: 86
Root Mean Squared Error: 0.101
R-squared: 0.959,  Adjusted R-Squared 0.958
F-statistic vs. constant model: 993, p-value = 3.75e-60

Non-parametric smearing bias correction factor: 1.0242

Probability plot correlation coefficient: 0.91209

High leverage:              0.10112
High influence (Cook's D):  2.007
High influence (DFFITS):    0.36719
```

Figure 16 - Linear model regression statistics window

The Display Model button will show a window that displays the model results and statistics. The information includes the linear equation, coefficient estimates, estimate confidence intervals, R^2 values for the model, and root mean squared error. This information is also written to a report with the Write Report button on the main SAID window.

Writing a model report

To write a summary report for the linear model, click on the Write Report button within the main SAID window. You will be prompted for a location and name of a comma separated value file to write the report to. Selecting and entering a valid location and file name will write the report.

The contents of the report include

- ADVM configuration and processing options
- Data set file names and locations
- Linear model summary and statistics
- Critical outlier indicator values
- The dataset observations that were used in the creation of the model along with
    - Observation number
    - Fitted response variable values
    - Raw residuals
    - If the response variable is transformed, an estimate of the non-transformed variable with bias correction applied
    - Calculated outlier indicator values
- The observations that were removed from the model dataset

Viewing diagnostic plots of the model

SAID provides several ways to graphically evaluate the linear model. Clicking on the Plots button within the main SAID window will display another window that provides several plotting options.

In any plot figure, if Data Cursor Mode is enabled, then you can select any observation data point and observation number will be shown along with the values plotted.

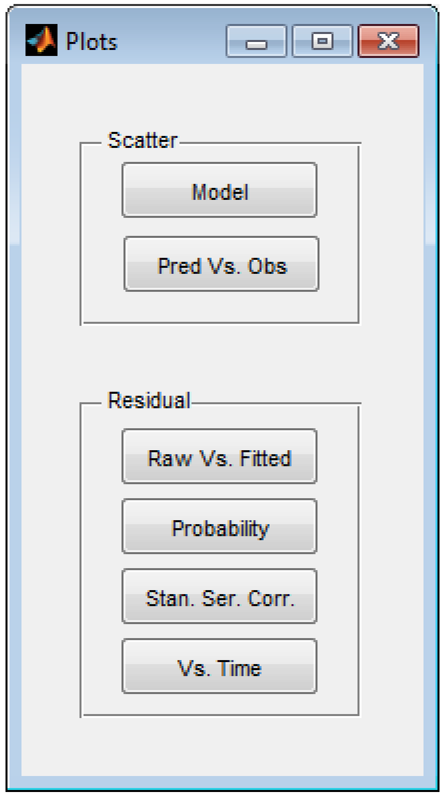


Figure 17 – Linear model plotting window

Scatter plots

Model

The Model button within the Scatter button group will show different things depending on if the linear model is a simple linear regression or a multiple linear regression or if the response variable is transformed.

Simple linear regression model

If the existing linear model is an SLR model, then a figure with the response observations plotted against the explanatory observed values will be shown. Also shown are a fit line and 95% confidence bounds.

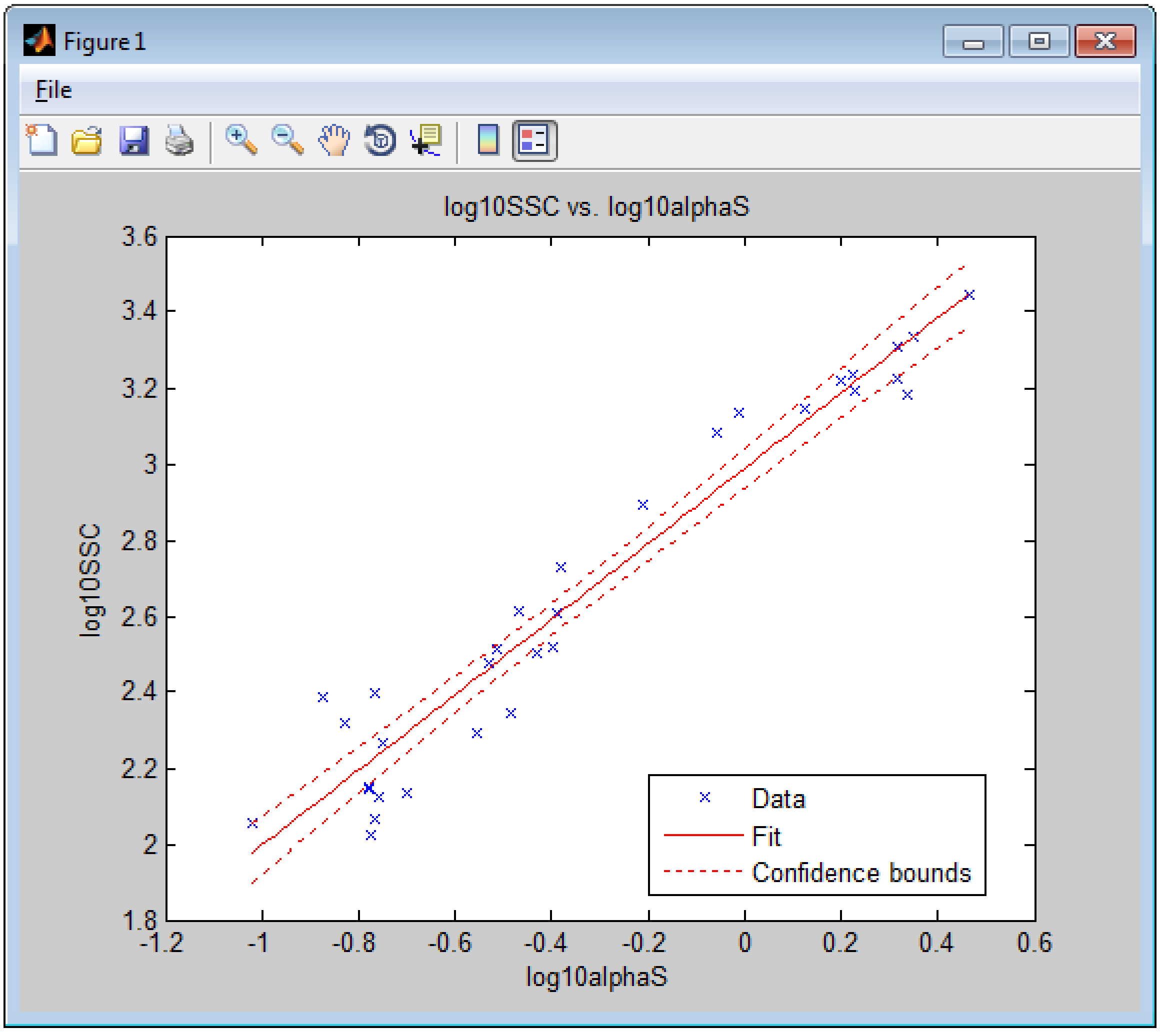


Figure 18 - Linear model scatter plot for a SLR model

If the response variable is transformed, then a linear-space plot will be shown with a smeared estimate fit line and confidence bounds.

Multiple linear regression model

When the existing model is an MLR, a partial residual plot for each variable in the model will be shown. Partial residual plots are described in Helsel and Hirsch, section 11.5.1, page 301.

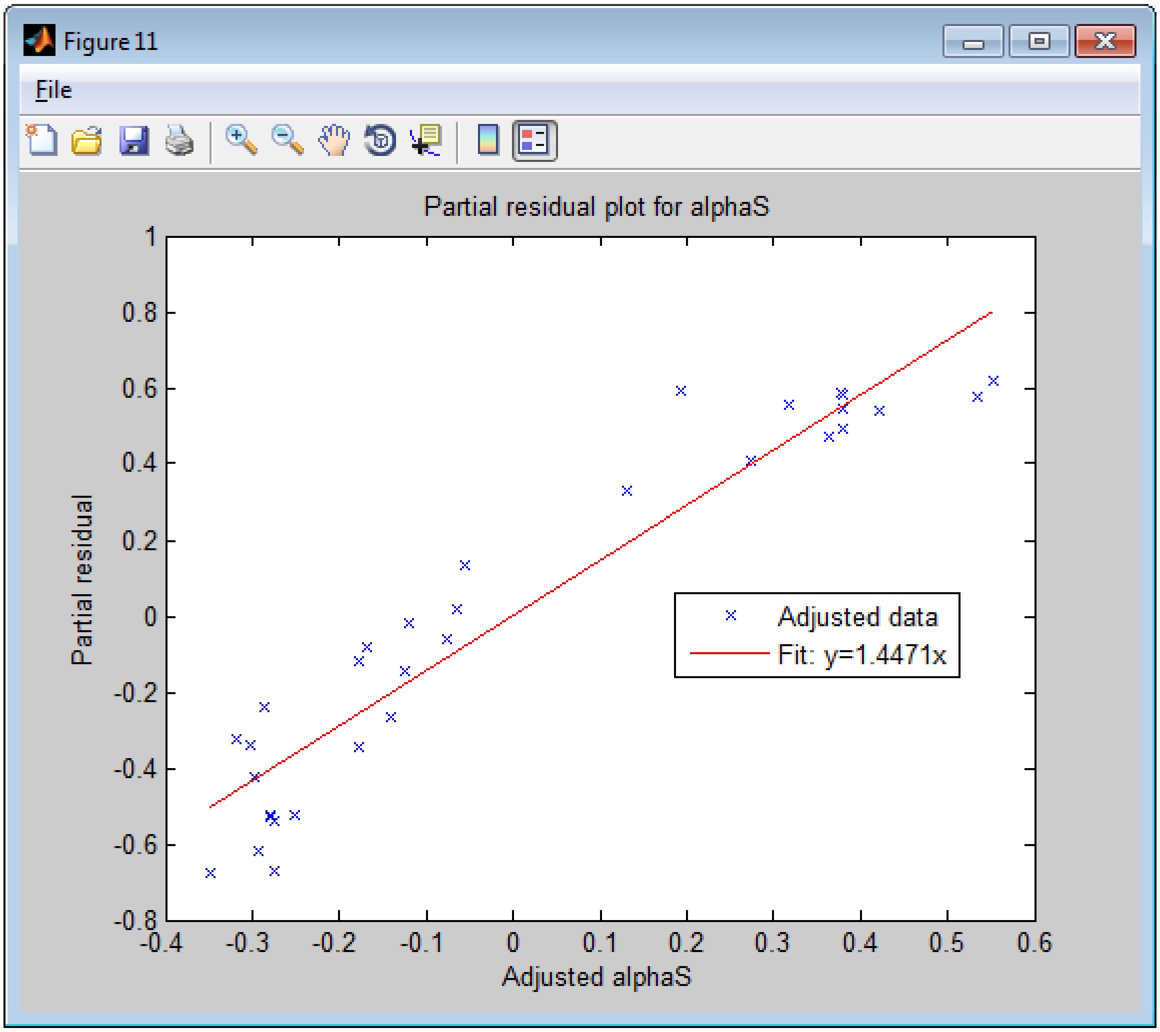


Figure 19 - Partial residual plot for a single variable from an MLR model

Predicted versus observed

Clicking this button will show a plot of the predicted response variable against the observed response variable. Also plotted is a one-to-one data line for comparison. If the response variable is transformed then an additional figure will show the predicted versus observed values in linear space.

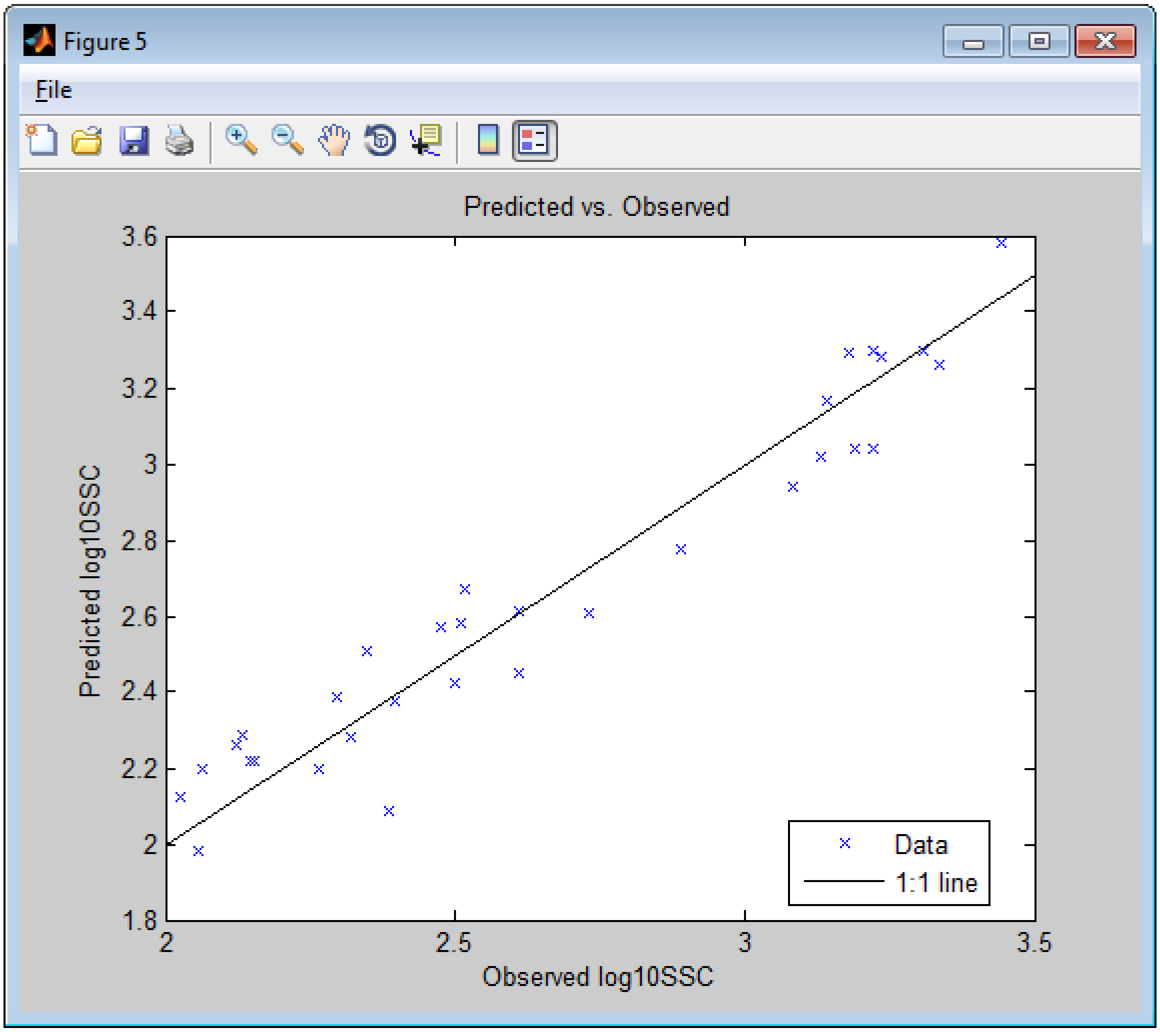


Figure 20 – Predicted response variable plotted against observed response variable

Residual plots

- **Raw Vs. Fitted**—Shows a plot of the raw residuals against the fitted response values

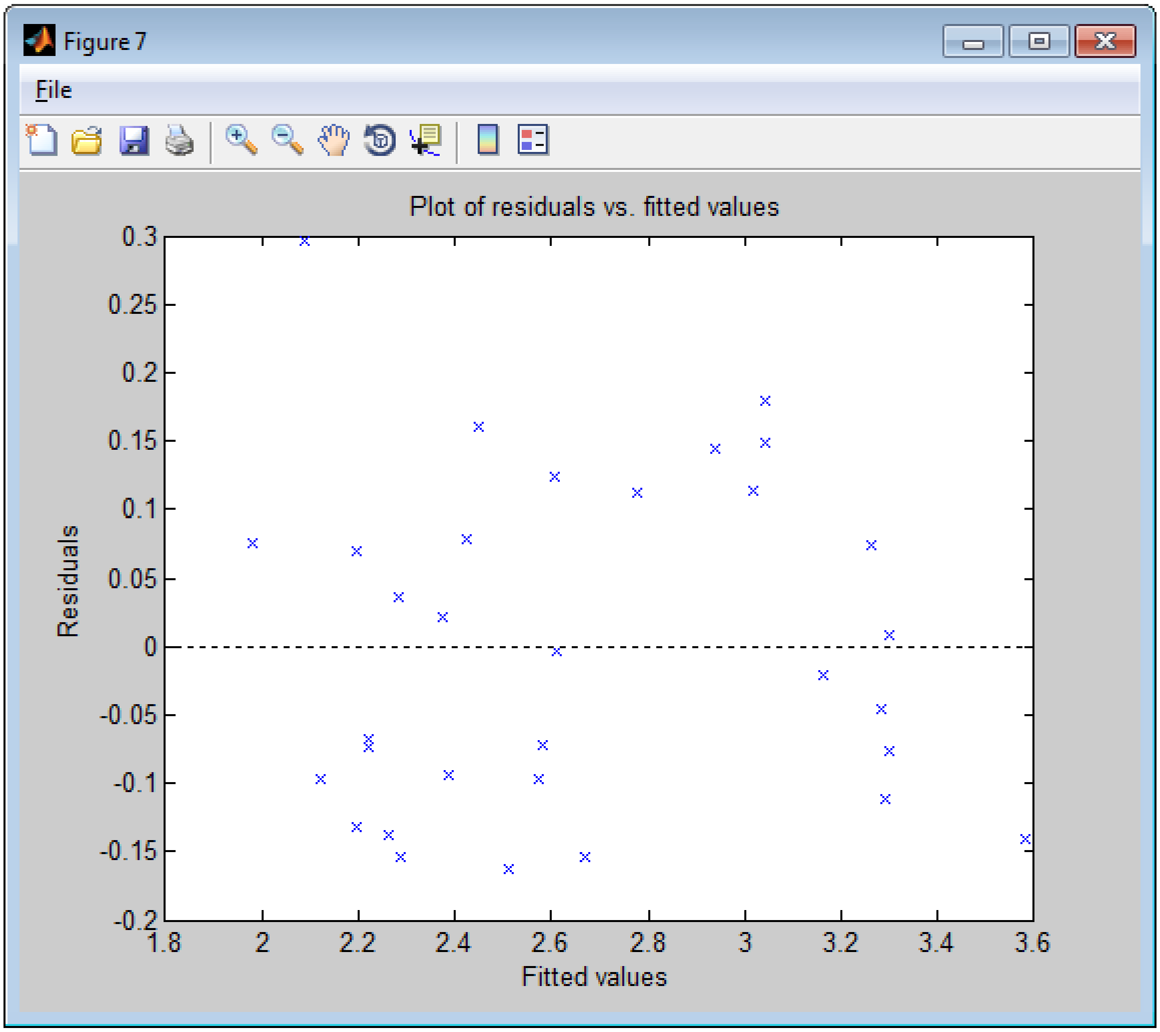


Figure 21 - Raw residuals plotted against the fitted response variable

- **Probability**—Normal probability plot of raw residuals

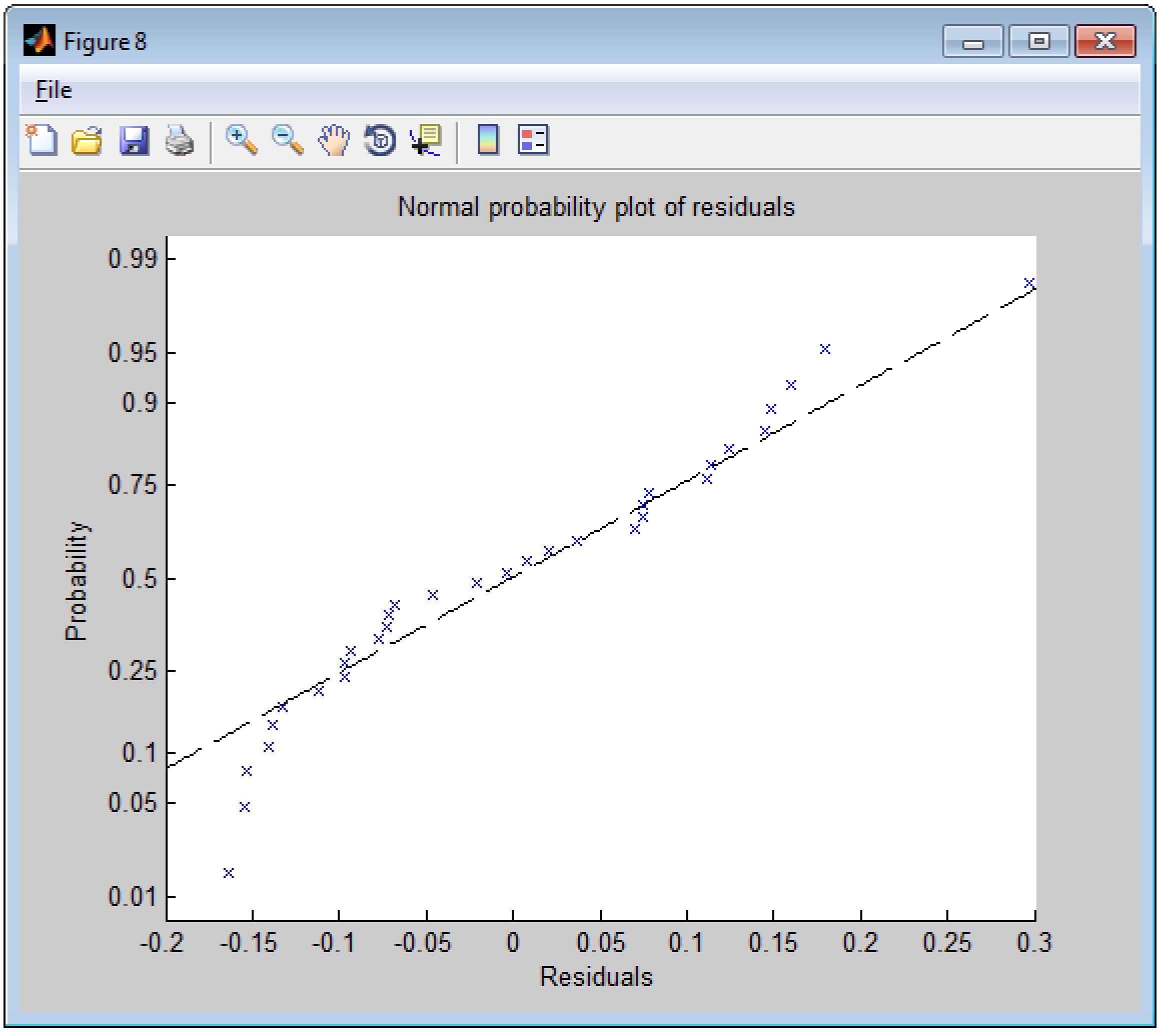


Figure 22 - Normal probability plot of the raw residuals from a linear regression

- **Stan. Ser. Corr.**—Standard serial correlation plot of the residuals shown with a LOWESS fit line to detect autocorrelation. If the LOWESS fit line shows a trend that deviates far from 0, serial correlation may be present.

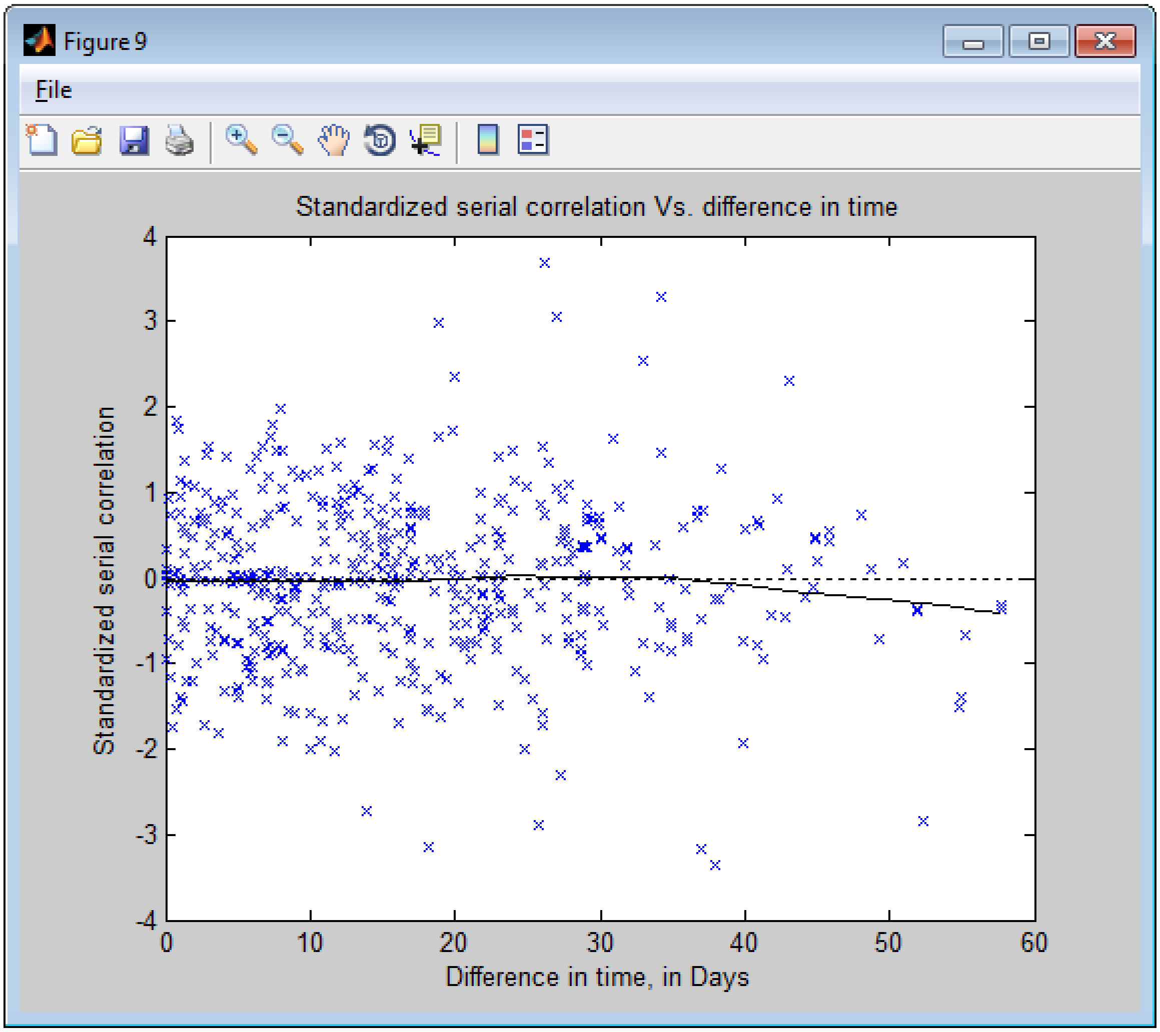


Figure 23 - Standard serial correlation plot

- **Vs. Time**—Raw residuals plotted against time to see if a time dependent trend exists with the residuals

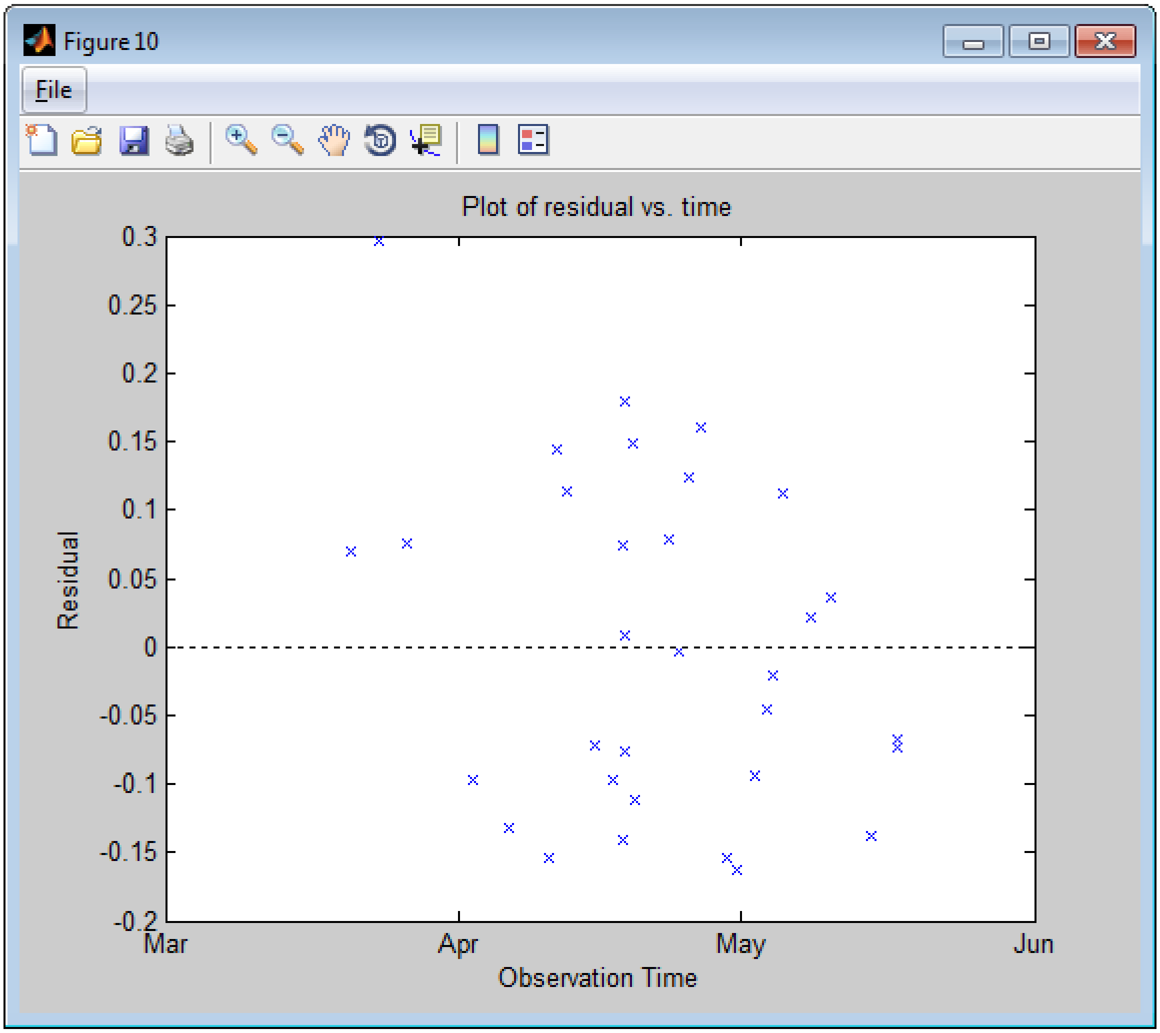


Figure 24 - Raw residuals plotted against time

Exporting a time series

The Time Series button, when clicked, will first prompt you for a dataset file to load. This dataset must contain either the predictor variables used in the model or enough information to compute ADVM parameters if they are used as predictor variables. If ADVM variables are used, the same processing and configuration options are applied to the loaded dataset in computing the ADVM parameters.

After selecting a dataset file, a figure showing the predicted time series will be shown with a 90% prediction interval (Figure 25). You will also be prompted for a name and location of a tab delimited text file to write the predicted time series to. If you do not want to write a text file, click on the Cancel button.

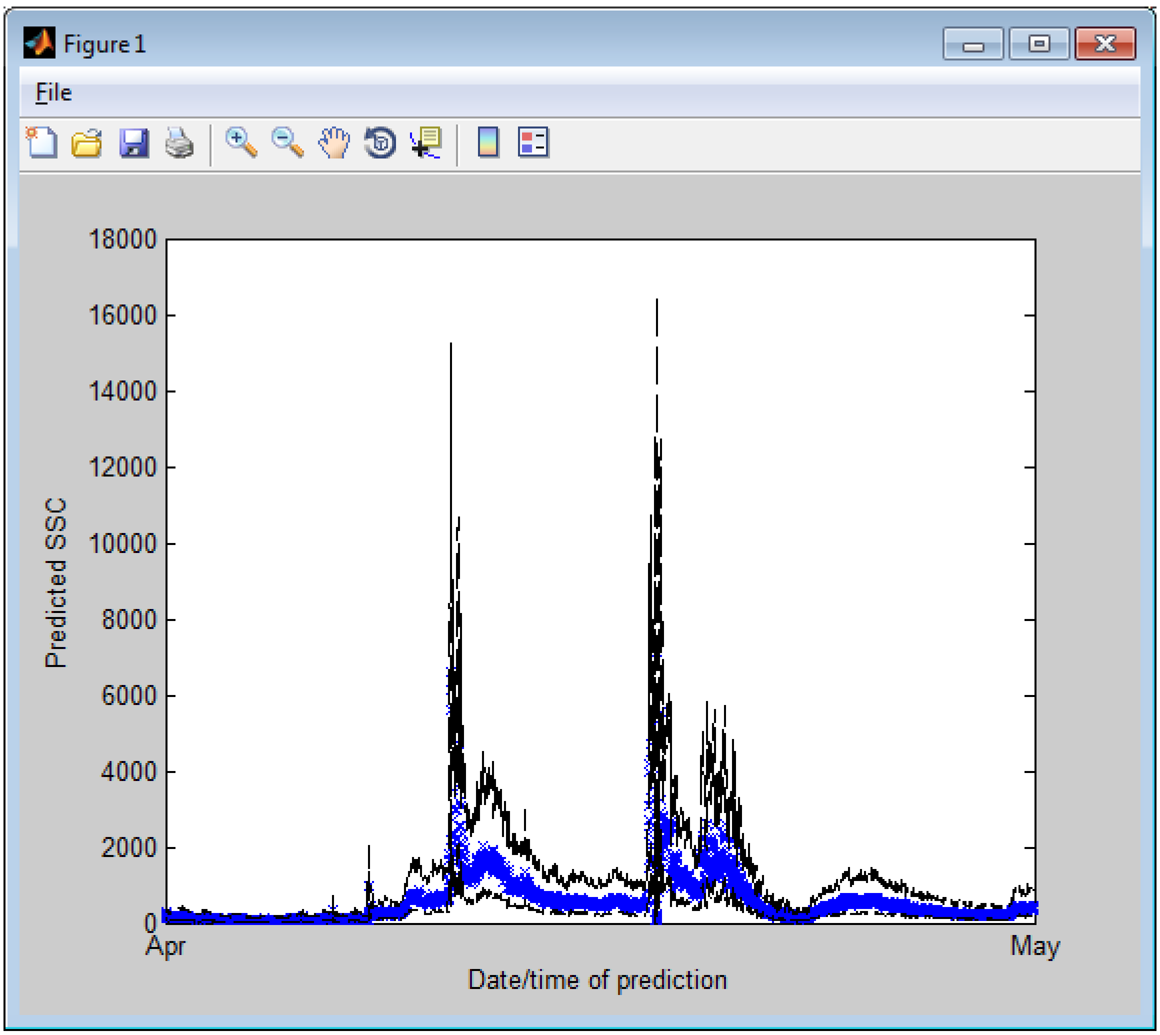


Figure 25 - Predicted time series with prediction interval plotted against time

Saving and loading a model state

At any time, you may save a model state in order to load it at a different time. To do so, select either Save or Load in from the File menu. The saved MAT-file contains information that can be loaded into a MATLAB session for analysis. This saves all loaded datasets, matched datasets, configuration data, and user-specified model characteristics.

To clear the data stored in SAID, select New from the File menu.

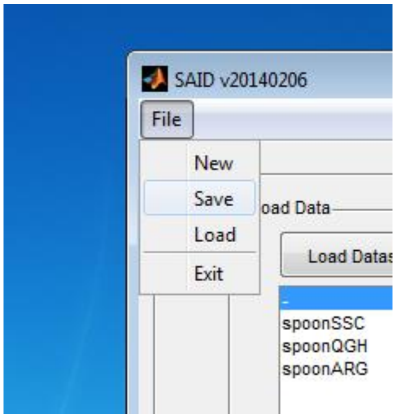


Figure 26 - File menu on the main SAID window