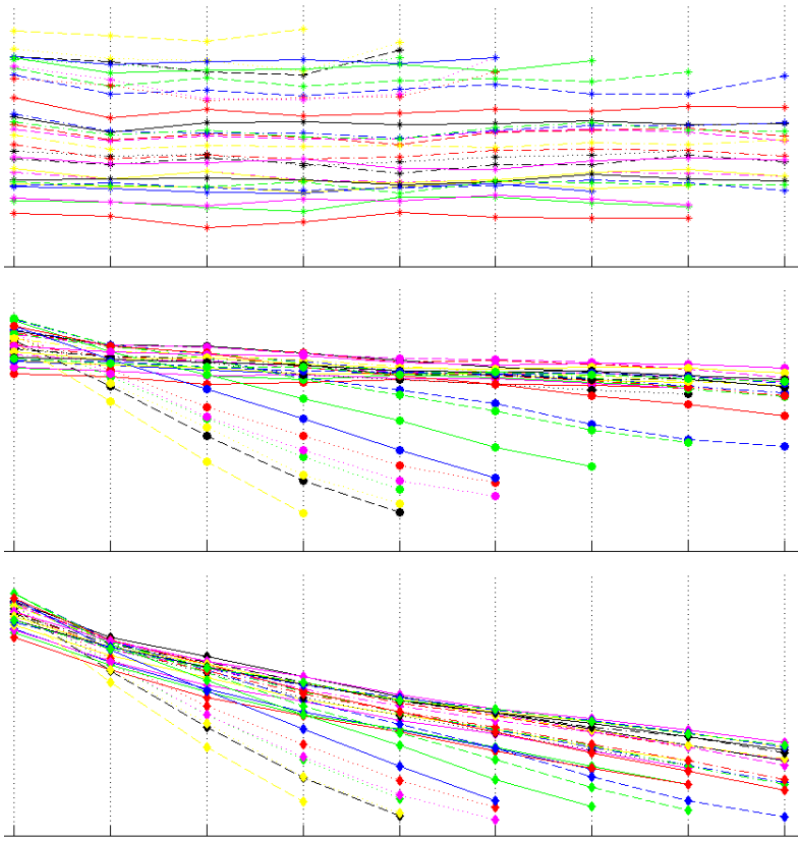
USGS
science for a changing world

Federal Interagency Sedimentation Project and Midwest Region River Sediments and Nutrients Investigations Initiative

Surrogate Analysis and Index Developer (SAID) Tool



Open-File Report 2015–1177

U.S. Department of the Interior
U.S. Geological Survey

Federal Interagency Sedimentation Project and Midwest Region River Sediments and Nutrients Investigations Initiative

Surrogate Analysis and Index Developer (SAID) Tool

By Marian M. Domanski, Timothy D. Straub, and Mark N. Landers

Open-File Report 2015–1177

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2015

Suggested citation:
Domanski, M.M., Straub, T.D., and Landers, M.N., 2015, Surrogate Analysis and Index Developer (SAID) tool (version 1.0, September 2015): U.S. Geological Survey Open-File Report 2015–1177, 38 p., http://dx.doi.org/10.3133/20151177.

ISSN 2331-1258 (online)

Contents

Figures

Tables

Conversion Factors

Inch/Pound to International System of Units

Multiply	By	To obtain
	Length	
foot (ft)	0.3048	meter (m)

Temperature in degrees Fahrenheit (°F) may be converted to degrees Celsius (°C) as °C = (°F – 32) / 1.8.

Abbreviations

ADVM	acoustic Doppler velocity meter
MB	measured backscatter
MLR	multiple linear regression
OLS	ordinary least squares
SAID	Surrogate Analysis and Index Developer
SCB	sediment corrected backscatter
SLR	simple linear regression
SNR	signal-to-noise ratio
SSC	suspended-sediment concentration
USGS	U.S. Geological Survey
WCB	water corrected backscatter

Surrogate Analysis and Index Developer (SAID) Tool

Abstract

The use of acoustic and other parameters as surrogates for suspended-sediment concentrations (SSC) in rivers has been successful in multiple applications across the Nation. Tools to process and evaluate the data are critical to advancing the operational use of surrogates along with the subsequent development of regression models from which real-time sediment concentrations can be made available to the public. Recent developments in both areas are having an immediate impact on surrogate research and on surrogate monitoring sites currently (2015) in operation.

The Surrogate Analysis and Index Developer (SAID) standalone tool, developed by the U.S. Geological Survey (USGS), assists in the creation of linear regression models that relate constituent and surrogate parameters by providing visual and quantitative diagnostics to the user. SAID also processes acoustic parameters to be used as explanatory variables for SSC. The sediment acoustic method utilizes acoustic parameters from fixed-mount stationary equipment. The theory and method used by the SAID tool have been described in recent publications. The tool also serves to support sediment-acoustic-index methods and other surrogate guidelines such as turbidity and SSC (Rasmussen and others, 2009).

The regression models created in SAID can be used in utilities that have been developed to work with the USGS National Water Information System (NWIS) and for the USGS National Real-Time Water Quality (NRTWQ) Web site. The real-time dissemination of predicted SSC and prediction intervals for each time step has substantial potential to improve understanding of sediment-related water quality and associated engineering and ecological management decisions.

Introduction

Streamflow, sediment, and water-quality data are needed to establish baseline information for water-resource managers to evaluate historical and current conditions and plan management alternatives. Real-time, continuous suspended-sediment concentration (SSC) data can be useful for monitoring river response downstream of areas affected by recent wildfires, construction or remediation activities, levee failures, or changing land uses. Additionally, real-time data can provide an early warning for operators of municipal water supply and hydroelectric facilities concerned with avoiding damage to infrastructure from sediment. Surrogates are becoming widely used to better understand physical and chemical processes in natural systems (Rasmussen and others, 2009). Acoustic technology is becoming increasingly used for velocity measurements and also is being used as a surrogate for sediment concentrations.

The Surrogate Analysis and Index Developer (SAID) tool is a standalone tool to assist in the development of ordinary least squares (OLS) linear regression models that relate constituent and surrogate parameters (Helsel and Hirsch, 2002) by providing visual and quantitative diagnostics to the user (fig. 1). The tool is written in the MATLAB® programming language. There is no limit on the number of explanatory variables to be used in a linear regression model and no requirement of which explanatory variables to use.

This manual provides an overview on processing and loading data into SAID for developing regression models among surrogate data and measured constituents. In addition, this manual discusses the acoustic Doppler velocity meter (ADVM) configuration parameters that are used in the calculation of acoustic surrogate parameters.

Overview

SAID has applications for relating surrogate-technology parameters such as turbidity, acoustics, and others. SAID can be used for processing acoustic parameters to be used as explanatory variables for SSC. The sediment-acoustic method, which assumes a spatially constant acoustic attenuation due to the presence of suspended particles, utilizes acoustic data from fixed-mount stationary ADVMs. Some of the earliest U.S. Geological Survey (USGS) application and research was done by Topping and others (2004, 2006, 2007), Wright and others (2010), Landers (2012), and Wood and Teasdale (2013). The sediment-acoustic method, as described in these references, is used in SAID to compute the sediment attenuation coefficient and sediment corrected backscatter from ADVM acoustic parameters. SAID allows for quick adjustment of complex ADVM data-processing options, changes in the variables used in the regression, and evaluation of the created model. The SAID tool also enables the user to transform loaded variables, build linear regression models, view linear model diagnostic statistics and plots, export the model information, and generate a predicted time series. An overview of the dataset workflow is briefly described in the next section; the following sections describe the process to develop regression models in more detail.

- **Data**—In the development of linear regression models, explanatory and response variables must be selected. These variables are contained in time series or discrete dataset files stored on disk and can be loaded into SAID.
    - ***Load data***—Datasets that are stored on disk in ASCII files are loaded into memory by SAID.
        - Constituent—A single constituent dataset is loaded. The response variables used for the linear regression model are chosen from the variables in the constituent dataset. The constituent dataset observation times serve as the primary key for matching surrogate observations.
        - Surrogate—Multiple surrogate datasets can be loaded. At least one dataset with valid variables must be loaded in order to choose explanatory variables for the linear regression.
        - Matching—Surrogate dataset observations are matched to the closest in time constituent observations. If the closest time difference is greater than the selected maximum time difference, then no valid observation match is made. The resulting matched dataset is used in the creation of the linear regression model.
    - ***Process ADVM data (optional)***—Acoustic variables are calculated from data in ADVM observations using configuration and processing parameters entered by the user. After a valid linear model is created, the acoustic backscatter profiles used to calculate the acoustic variables can be viewed using the backscatter profile plotting utility.
- **Linear Model**—After datasets are loaded and valid explanatory and response variables exist within the workspace, variables are chosen and a linear regression model is created and evaluated. 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, which are available using the Plot Backscatter, View/Edit Table, Display Model, Write Report, Plots, and Time Series buttons.

- *Choosing variables*—Available variables will be displayed in the Explanatory Variables list box and the Response Variable dropdown list.
  - Response variable—The list of response variables is taken from the variables in the constituent dataset. Only one response variable can be chosen.
  - Explanatory variable(s)—The list of explanatory variables is taken from the variables in the surrogate datasets. If a variable name exists as a response variable, it will not be available for selection as an explanatory variable.
- *Transform variables*—Transformation of variables can be done within SAID, and the transformed variable will be available in the respective list of variables.
- *Evaluate model and write report*—SAID provides several diagnostic plots, tables, and reports to determine if the linear model created fits the assumptions of the OLS method. After a model is decided upon, a report can be generated that contains details of the ADVM calculation parameters, model statistics and diagnostics, and observation information.
- *Generate time series*—A time series of predicted response values can be calculated using the linear model created within SAID by loading a dataset that contains the necessary explanatory variables.

- **SAID workspace**—At any time, the state of the SAID workspace can be saved in order to be loaded later. The file generated after saving can be reopened in SAID for later use or can be opened directly in MATLAB.

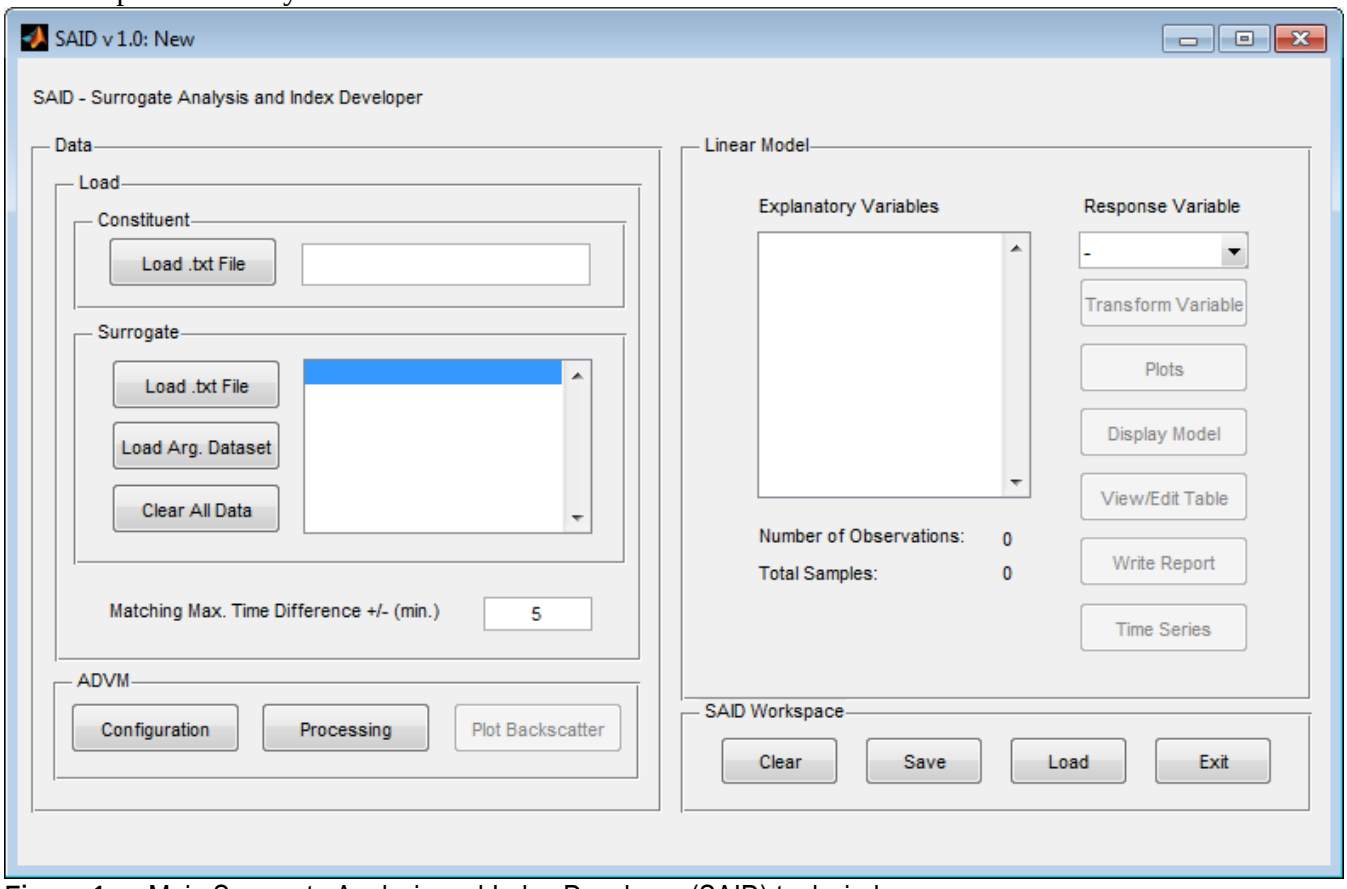


Figure 1. Main Surrogate Analysis and Index Developer (SAID) tool window.

Data

The following definitions are useful as the data workflow is described in more detail.

- Variable—A quantification of a particular phenomenon being observed.
    - Observation—A set of coinciding observed values for one to several variables.
    - Dataset—A collection of observations.

It is typical for surrogate and constituent time series to be stored in separate dataset files because surrogate observations are usually continuous and constituent observations generally occur at irregular time intervals. Additional reasons for keeping surrogate and constituent datasets separate are as follows:

- In SAID, constituent and surrogate data must be stored and loaded separately.
- Groups of variables with observations that are not exactly coincident in time should be stored in separate dataset files.
- Variables that have observations at intermittent time intervals, as typical with constituent data, should be stored in separate dataset files.

Formats

SAID is capable of loading two types of datasets. Datasets can be stored on disk as tab delimited ASCII files or as a collection of SonTek® Argonaut ASCII files.

Tab Delimited ASCII File

In the tab delimited ASCII file format, 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. The names of variables must be valid MATLAB identifier names. A valid MATLAB identifier name begins with a letter and contains letters, numbers, or underscores. Upon loading, characters that are invalid in the use of dataset variable names are removed and replaced (for example, "." will be changed to "_").

Each dataset must contain date and time information. All date and time information must be present and labeled by headers; the header labels are case sensitive. Date and time information can be stored three different ways (bold indicates header label):

1. In the same column under the header label **DateTime**
2. In separate columns under header labels **Date** and **Time**
3. In separate columns for each value: **y** (year), **m** (month), **d** (day), **H** (hour), **M** (minute), **S** (second)

For items 1 and 2, the following table (table 1) shows examples of the different date and time formats that are acceptable (from http://www.mathworks.com/help/matlab/ref/datenum.html#inputarg_DateString). The program converts the date and time information into a MATLAB date and time serial number. For converting columns Date, Time, and DateTime, SAID uses the MATLAB function datenum and passes the columns as a DateString argument.

Table 1. Examples of acceptable formats of date and time variables.

Date string format	Example
'dd-mmm-yyyy HH:MM:SS'	01-Mar-2000 15:45:17
'dd-mmm-yyyy'	01-Mar-2000
'mm/dd/yyyy'	03/01/2000
'mm/dd/yy'	03/01/00
'mm/dd'	03/01
'mmm.dd,yyyy HH:MM:SS'	Mar.01,2000 15:45:17
'mmm.dd,yyyy'	Mar.01,2000
'yyyy-mm-dd HH:MM:SS'	2000-03-01 15:45:17
'yyyy-mm-dd'	2000-03-01
'yyyy/mm/dd'	2000/03/01
'HH:MM:SS'	15:45:17
'HH:MM:SS PM'	3:45:17 PM
'HH:MM'	15:45
'HH:MM PM'	3:45 PM

ADVM variables that are loaded from tab delimited ASCII files can be used in SAID. In order for the program to recognize the variables, specific variable names must be used. For backscatter data, 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 (Amp) and signal-to-noise ratio (SNR), 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 also are dedicated variables used for the temperature and water depth. ADVMTemp is the temperature recorded by the ADVM, and Vbeam is the water height in meters that the ADVM reports. The temperature must be in units of degrees Celsius and is directly used in computing the ADVM parameters. 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 also is set by the user in order to exclude samples taken when the water is below a certain depth. Unlike the backscatter variables, ADVMTemp and Vbeam are available for use in the creation of a linear regression model.

SonTek® Argonaut Dataset

The SAID tool also has the capacity to load ADVM data directly from ASCII files exported by SonTek's ViewArgonaut® software. SAID requires the header lines to be exported with the data in ViewArgonaut. 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 extracts all variables needed to compute the acoustic parameters as well as the ADVM configuration information from these files.

When loading Argonaut datasets, SAID checks for incompatibility. A dataset is considered incompatible if the frequency, slant angle, blanking distance, cell size, or number of cells is different than what is currently loaded. If an Argonaut dataset has already been loaded, the subsequently loaded datasets are checked against the previously loaded dataset. If no Argonaut datasets have been loaded,

datasets are compared to the dataset first in the list of datasets to be loaded. If you leave the frequency, slant angle, blank distance, cell size, or number of cells blank (see fig. 2), the program fills the configuration values from the first dataset in the set of datasets to be loaded.

Loading Data

Data are loaded into the SAID workspace through dataset files. The method of loading a dataset depends on the type of variable (constituent or surrogate) and the type of dataset being loaded (tab delimited ASCII text or SonTek® Argonaut exported). The "Load .txt File" buttons load tab delimited ASCII files. The "Load Arg. Dataset" button loads SonTek® Argonaut datasets exported from ViewArgonaut. Refer to figure 1 for the location of the buttons.

Constituent

Only one dataset can be loaded as the constituent dataset. The constituent dataset must be in the tab delimited ASCII text file format. Variables loaded in a constituent dataset are used only as response variables and are not available for use as explanatory variables in the linear regression model.

To load constituent data, click the "Load .txt File" button (fig. 1) within the Constituent box, and follow the prompts. After a constituent dataset is loaded, the constituent dataset name will appear in the box to the right. The variables loaded from the constituent dataset will appear in the Response Variable dropdown list.

Surrogate

Multiple surrogate datasets can be loaded. A single surrogate variable can be stored among several datasets. Surrogate variables are used only as explanatory variables in the linear regression model.

To load surrogate data, click on either the "Load .txt File" or "Load Arg. Dataset" buttons (fig. 1), and follow the prompts. If observations of variables being loaded already exist within the loaded variables, you will be prompted with three choices: overwrite the currently loaded observations, keep the current observations, or cancel loading the dataset. Once loaded, surrogate variables are stored individually within the SAID workspace. The time of observation is taken from the dataset the variable was loaded from, as described in more detail in the next section (Matching Observations). All surrogate data can be cleared by clicking on the "Clear All Data" button.

The names of the loaded surrogate datasets appear in the list box within the Surrogate section, with Argonaut datasets indicated by an (Arg) next to the root name. Available surrogate variables will appear in the Explanatory Variable list box.

In order to develop a linear regression model, observations from the surrogate and constituent time series must be matched. Matching takes place after variables have been loaded.

Matching Observations

In order to build a linear model, it is necessary that a dataset exists with observations of explanatory and response variables. Matching occurs in order to synchronize observations from the surrogate datasets to observations from the constituent dataset. The observation date and time of the constituent dataset serve as the date and time to which observations of other variables are matched. The result of matching is the creation of a single dataset (referred to as the matched dataset) containing the matched observations. The matched dataset is then used to develop the linear model. The matched dataset observation times are taken from the constituent observations. When surrogate observations are put into the matched dataset, the surrogate observation time is lost. 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.

The process of creating a matched dataset begins with making a copy of the constituent dataset. For each surrogate variable, a variable containing null observations is created in the matched dataset. For each observation in the matched dataset, the minimum absolute difference between the times of observation of the constituent and surrogate variables is found. If the minimum absolute time difference is less than the entered value for Matching Max Time Difference, then the variable observation is matched with the constituent observation. Otherwise, the observation for the variable is left as null. An example of the process is shown in table 2, in which the resulting matched dataset includes three samples with two observations available for linear model building. When the program has completed the matching algorithm, the variables available for use in the linear model are shown in the Explanatory Variables and Response Variable lists.

Table 2. Example of the matching process showing a constituent and surrogate dataset with a maximum Matching Max Time Difference set to 5 minutes.

[SSC, suspended-sediment concentration; ADVM, acoustic Doppler velocity meter; min, minutes]

Constituent dataset			Surrogate dataset				Matched dataset			
SSC			ADVM						SSC	ADVM
Date	Time	Value	Date	Time	Value	If MaxTime = 5 min	Date	Time	Value	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	2:02	550	4/4/2013	2:00	85	Match	4/4/2013	2:02	550	85

ADVM Parameter Processing and Plotting (Optional in SAID)

The following ADVM-related parameters are required by SAID before the acoustic backscatter data are processed and acoustic surrogate parameters are computed:

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

By clicking on the ADVM Configuration and Processing buttons (fig. 1), the ADVM configuration and processing options used in the calculation of the ADVM parameters can be changed (fig. 2). ADVM configuration parameters needed for input to SAID can be found in the setup parameters section of the ADVM software. When loading SonTek® Argonaut datasets, the configuration parameters are taken automatically from a configuration record file that is saved automatically with each data file by the ViewArgonaut software.

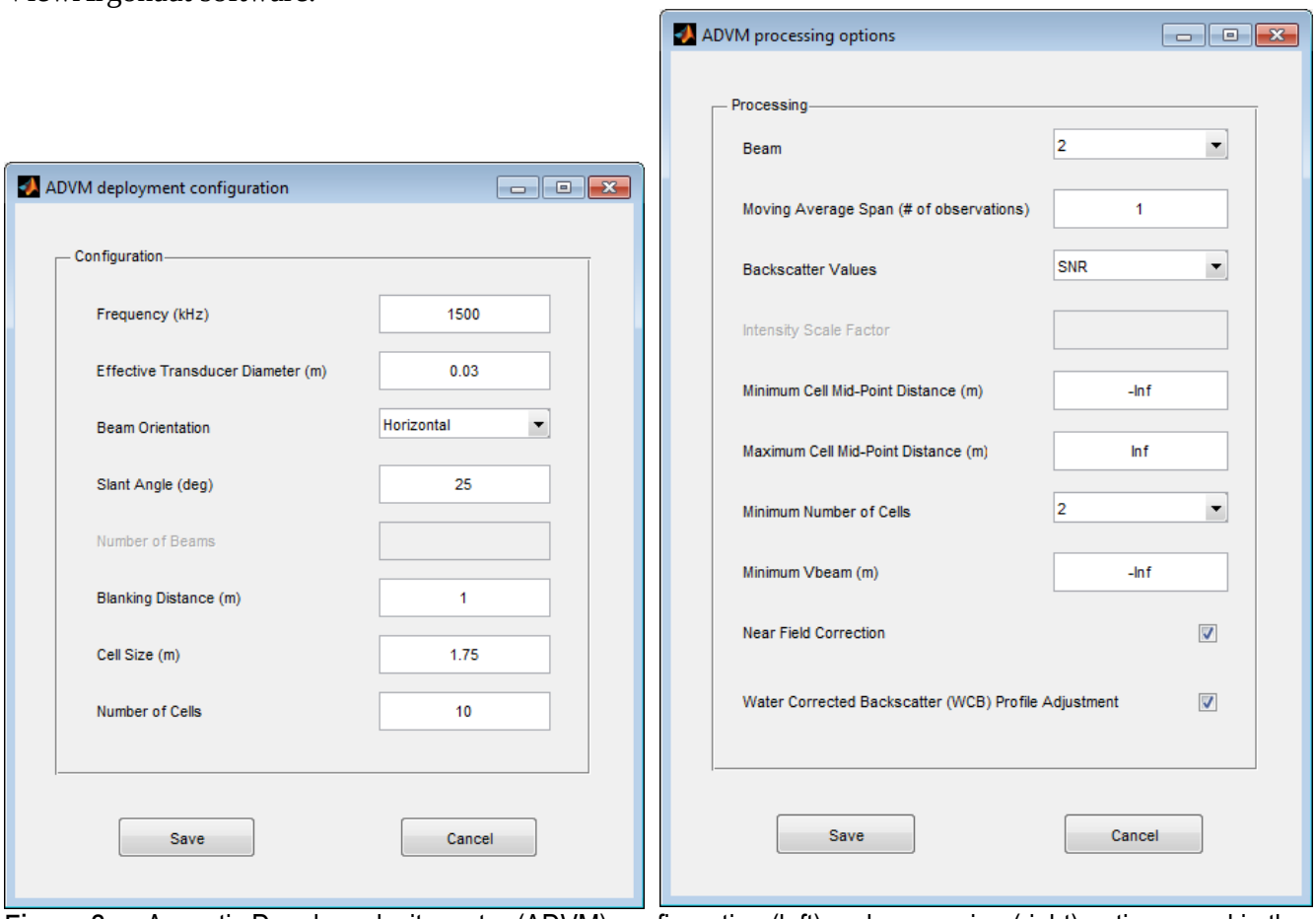


Figure 2. Acoustic Doppler velocity meter (ADVM) configuration (left) and processing (right) options used in the calculation of the ADVM parameters.

Configuration Parameters

ADVM configurations needed for input to SAID can generally be found in the setup parameters section of the ADVM software. When loading SonTek® Argonaut files, as described in the Loading Data section, the configuration information is loaded automatically. The following parameters indicate the ADVM type and setup and are necessary to compute the acoustic surrogate parameters.

- Frequency—The frequency of the ADVM acoustic signal.
- Effective Transducer Diameter—The effective diameter (in meters) of the ADVM transducer. The effective transducer diameter is only used when the Near Field Correction option is selected in the Processing section. For SonTek® Argonaut ADVMs, these values are as follows: 3,000 kilohertz (kHz) – 0.015 meter (m); 1,500 kHz – 0.030 m; 500 kHz – 0.090 m (SonTek, written commun., March 14, 2012). Note that these values are not the same as the physical diameter that is measured on the instrument, and they could change with new hardware versions for these instruments.
- 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 the backscatter for 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—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 dropdown list.

Processing Parameters

By clicking on the Processing button, the processing options used in the calculation of the ADVM parameters can be changed. The following parameters control how ADVM backscatter data are screened and processed (fig. 2).

- Beam—The beam number from which 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. Note that the span only indicates the number of observations to be averaged and not a time period. Care must be taken when ADVM time series with different time steps are loaded. For example, 5-minute data will be averaged with 15 minute data, etc.
- 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. The model developed will be specifically for either SNR or Amp units and cannot be switched

without building a new model. All empirical testing for best model using SNR or Amp should be evaluated.

- 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 dropdown list. The factor is typically 0.43 for SonTek® Argonaut instruments, 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 observation 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.
- 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 Mid-Point distance greater than the near field for a given instrument.
- Water Corrected Backscatter (WCB) Profile Adjustment—When this box is checked (example plots shown in the View Backscatter Profiles section), 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; no attenuation coefficient is calculated.

Once the required ADVM Configuration and Processing parameters have acceptable values, the ADVM variables with at least one observation will be available in the Explanatory Variables list.

An unexpectedly low number of matched acoustic variable observations could be due to a high number of invalid cells. Invalid cells are those that have invalid values. Throughout the process, cells are assigned invalid values because they either

- 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 filled with 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, or
- 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).

Viewing Backscatter Profiles

When a valid response variable is matched with valid explanatory variables, the Plot Backscatter button will be made available (fig. 1). 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 (fig. 3). The calculation of these parameters is described elsewhere (Topping and others, 2006; Topping and others, 2007; Landers, 2012; Wood and Teasdale, 2013). Also shown in the window is a list of observation numbers and times from the model. The observation times are taken from the constituent 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. Plots without and with the WCB Profile Adjustment box checked (see Processing Parameters section) are shown in figures 4 and 5, respectively. The Write Backscatter button allows you to write the MB, WCB, and SCB observations shown in the window to a comma-separated value (csv) format file.

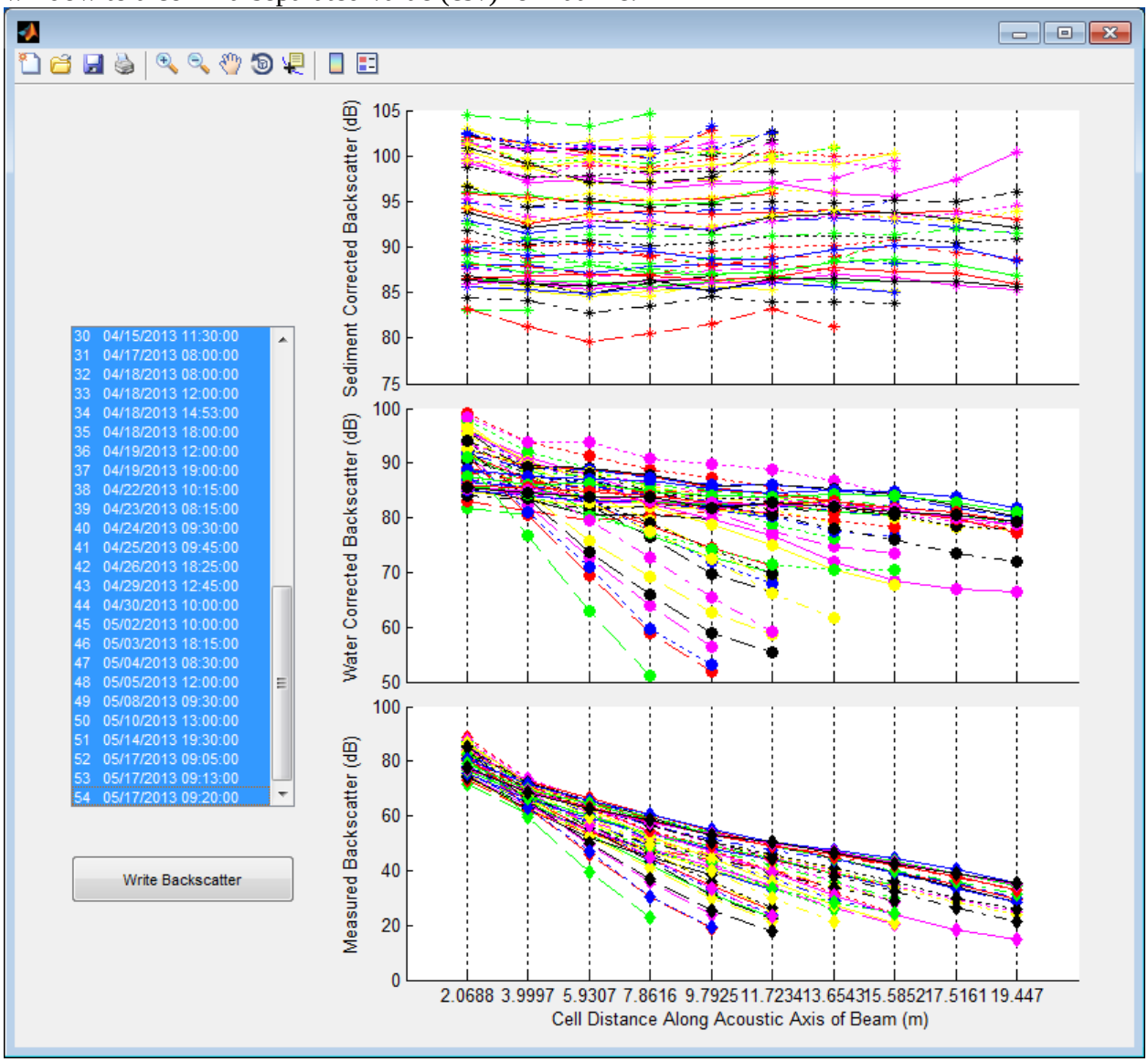


Figure 3. Backscatter profile plotting window.

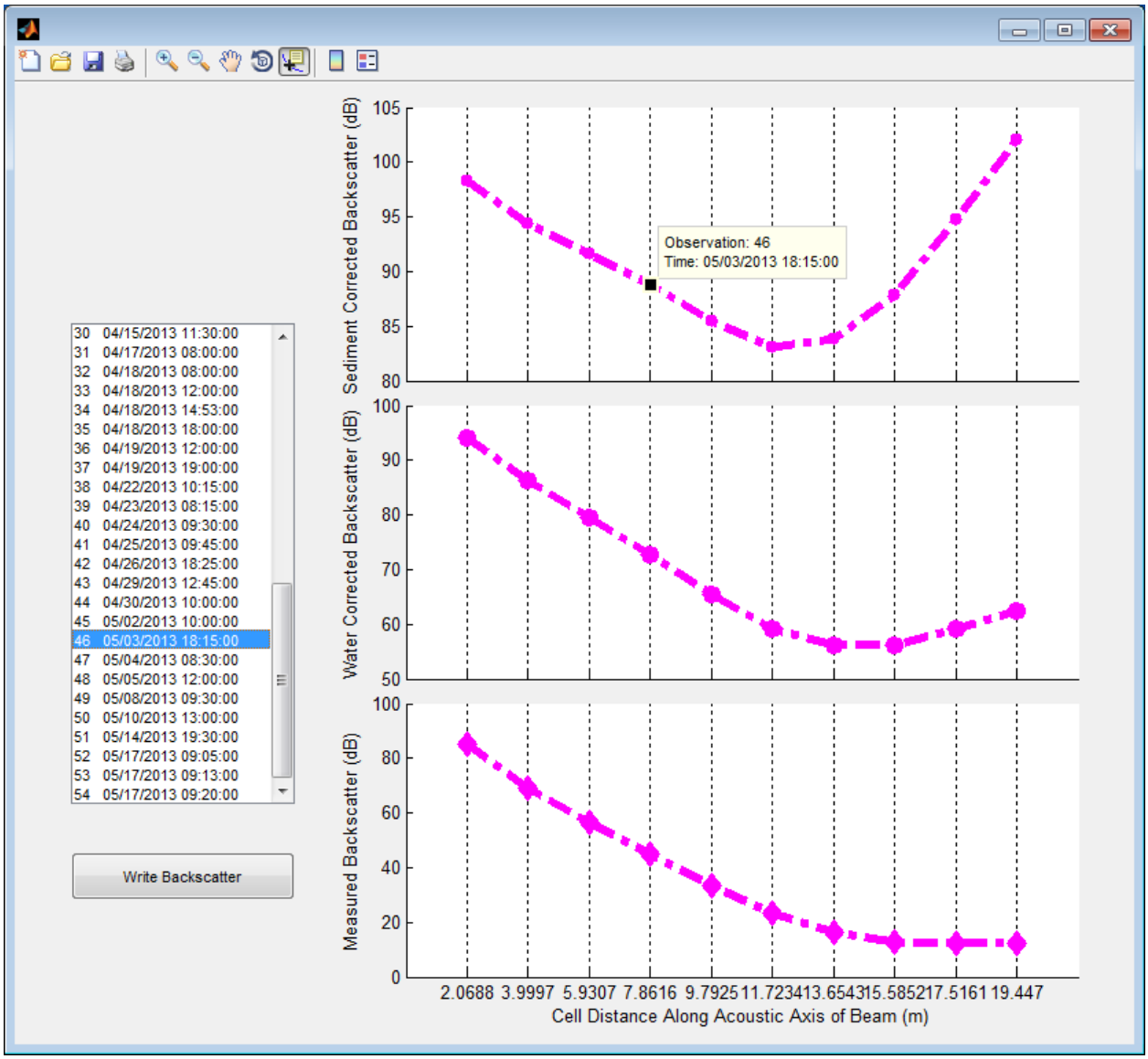


Figure 4. Backscatter profile of a single Acoustic Doppler velocity meter (ADVM) sample without the Water Corrected Backscatter (WCB) Profile Adjustment option enabled.

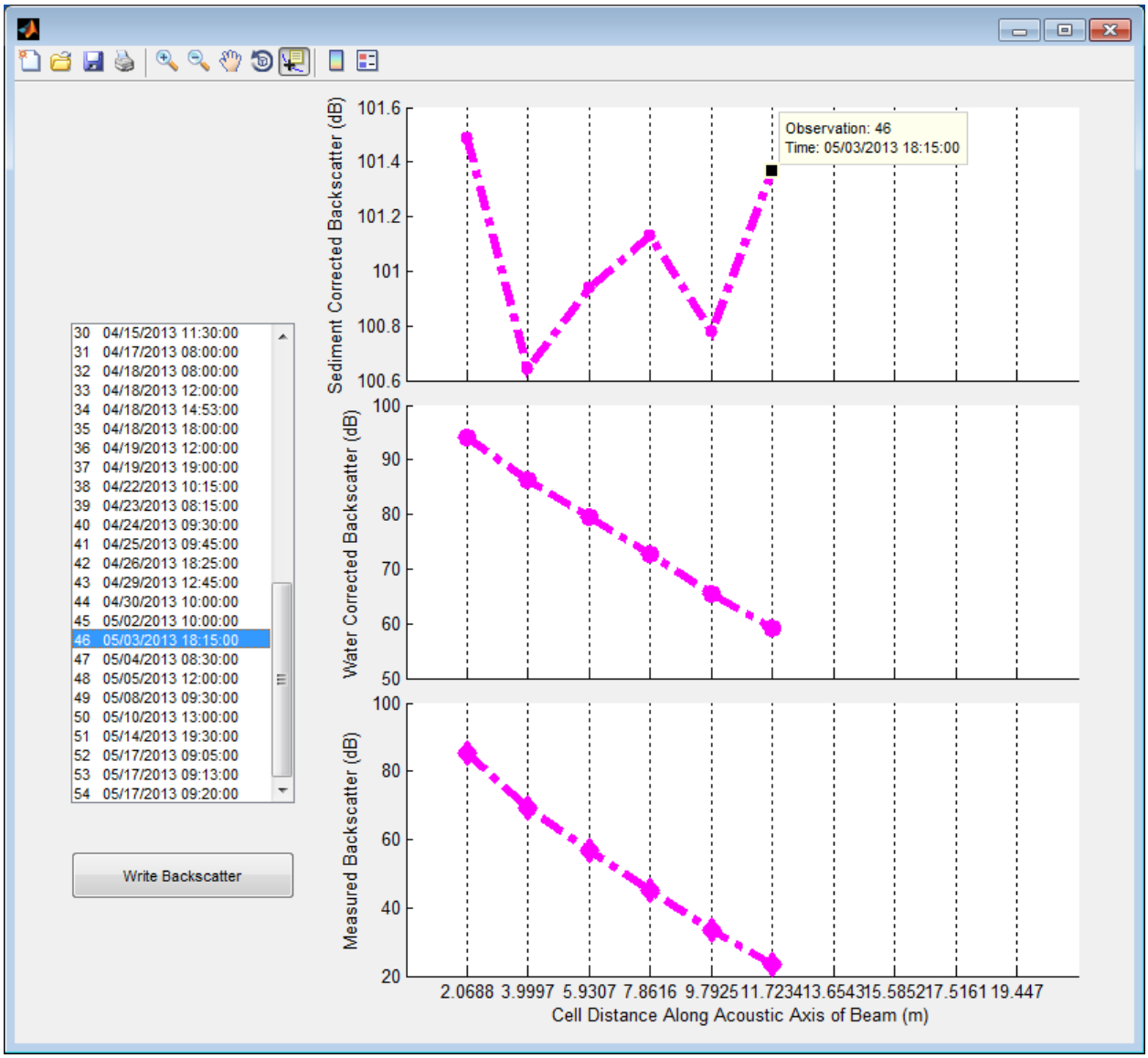


Figure 5. Backscatter profile of a single Acoustic Doppler velocity meter (ADVM) sample with the Water Corrected Backscatter (WCB) Profile Adjustment option enabled.

Linear Model

After constituent and surrogate datasets are loaded and response and explanatory variables are made available, an OLS linear regression model can be created. Selecting a variable in the Explanatory Variables list and one from the Response Variable dropdown list will result in the generation of a model (fig. 6). The number of observations used in the model is shown next to the Number of Observations label. Details of observations within the model can be viewed and observations can be removed.

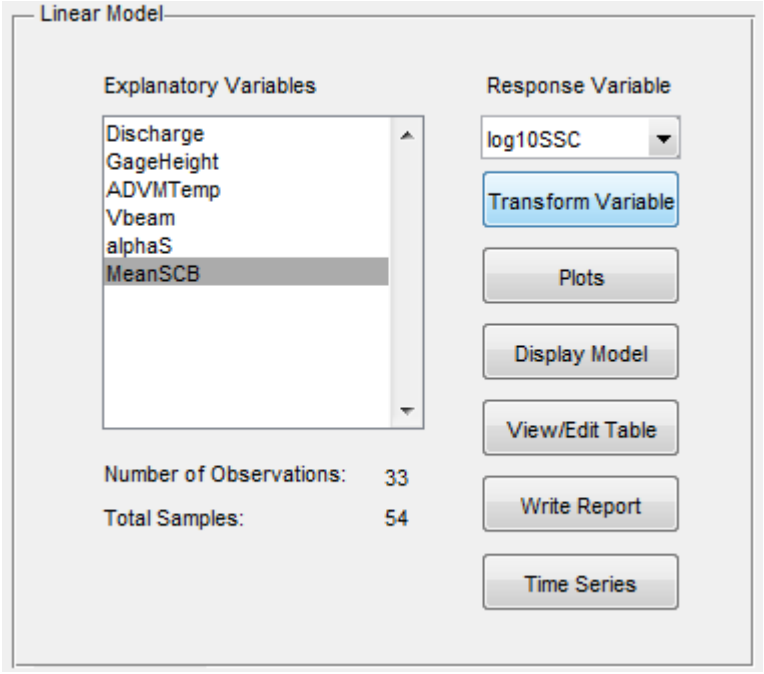


Figure 6. Linear Model options on the main Surrogate Analysis and Index Developer (SAID) window.

Choosing Variables

The available variables for use in the development of linear regression models in SAID appear in the Explanatory Variables list box and Response Variable dropdown list in the Linear Model section. Response variables are taken from the constituent dataset and explanatory variables are taken from the surrogate datasets. Variables loaded in a constituent dataset are not available for use as explanatory variables in the linear regression model. If a variable is loaded in the constituent dataset and it also appears in a surrogate dataset, it will not be available as an explanatory variable. There is no limit on the number of variables used in the creation of an OLS linear regression model, and there are no restrictions regarding which variables must be used.

Model Observations

If a valid linear model exists within the program, the Number of Observations field will show the number of observations used in the development of the linear model. The number of observations is the number of observation sets with valid values for all variables used within the model. Observations removed by the user, as described later in the View/Edit Table section, are not counted within the number of observations. The Total Samples field shows the total number of samples in the loaded constituent dataset.

Transforming Variables

The Transform Variable button provides the option to transform a loaded variable using a transform function in order to correct for heteroscedasticity or non-linearity. When transformed, the variable will be available as a selection in the Explanatory Variables list and the Response Variable dropdown list.

Model Evaluation

SAID provides graphical, tabular, and report formats to evaluate linear models. This section will describe the functions of the following buttons: Plots, Display Model, View/Edit Table, and Write Report.

Plots

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 (fig. 7). In any plot figure, if Data Cursor Mode is enabled, any observation data point can be selected and the corresponding observation number will be shown along with the values plotted. A legend is automatically generated for all scatter plots. For residual plots, the user has the option of including a legend. All legend entries are modifiable by the user.

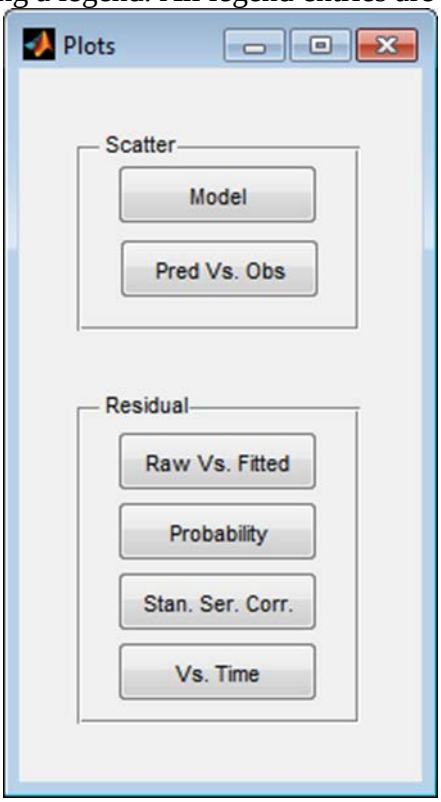


Figure 7. Linear model plotting window.

Scatter Plots

Depending on the type of model and whether or not the response variable is transformed, there are several varieties of scatter plots that can be shown.

- Model—The Model button will show different figures depending on whether the linear model is a simple linear regression (SLR) or a multiple linear regression (MLR), or if the response variable is transformed. If the response variable is transformed, then a linear-space plot will be shown with a smeared estimate fit line and confidence bounds.
    - SLR—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 (fig. 8). If the response variable is transformed, then a linear-space plot will be shown with a smeared estimate fit line and confidence bounds.
    - MLR—When the existing model is an MLR, a partial residual plot for each variable in the model will be shown (fig. 9).
- Predicted versus observed—Predicted versus observed plots can be selected to display the predicted response variable with the observed response variable (fig. 10). 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.

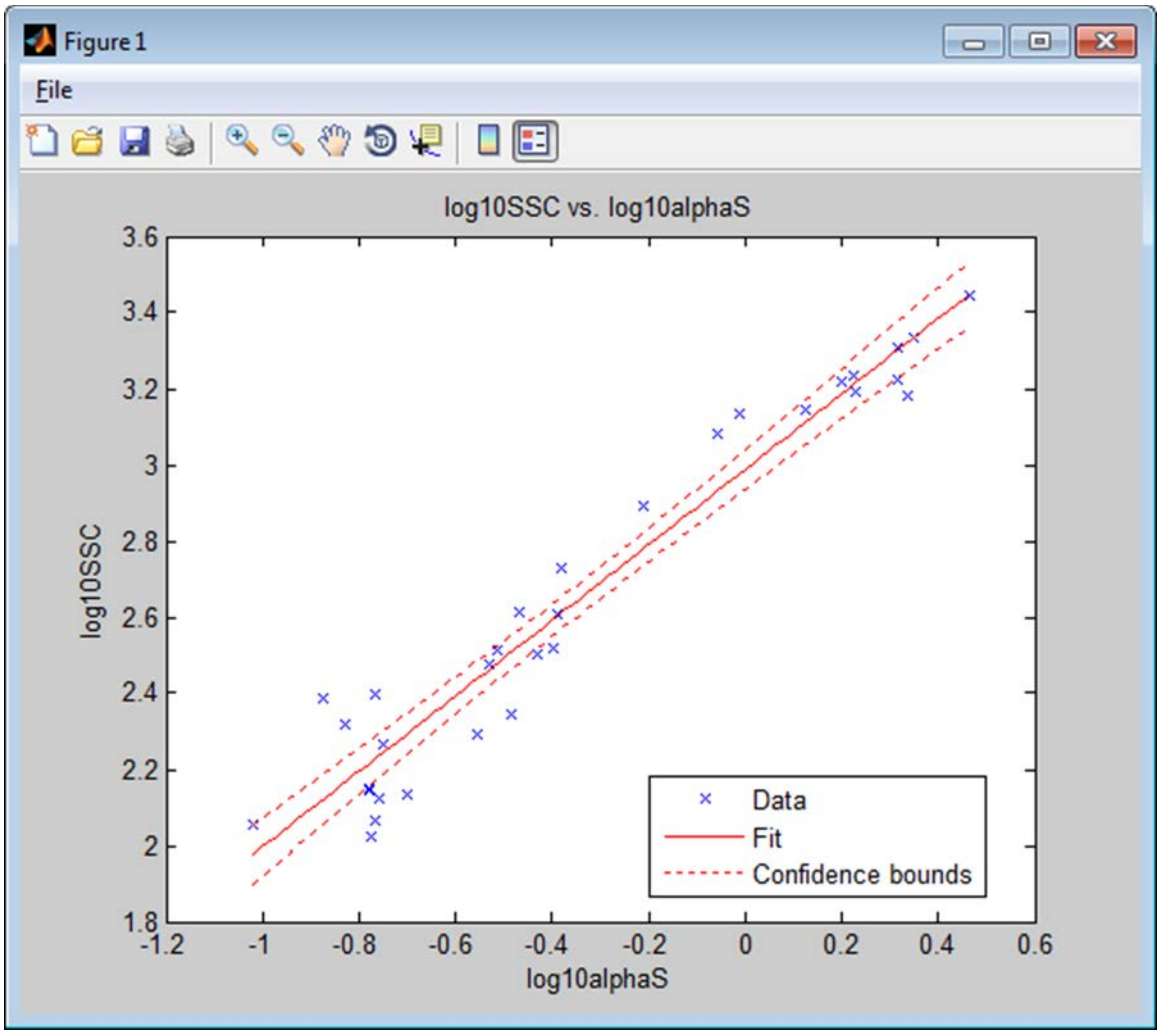


Figure 8. Linear model scatter plot for a simple linear regression (SLR) model.

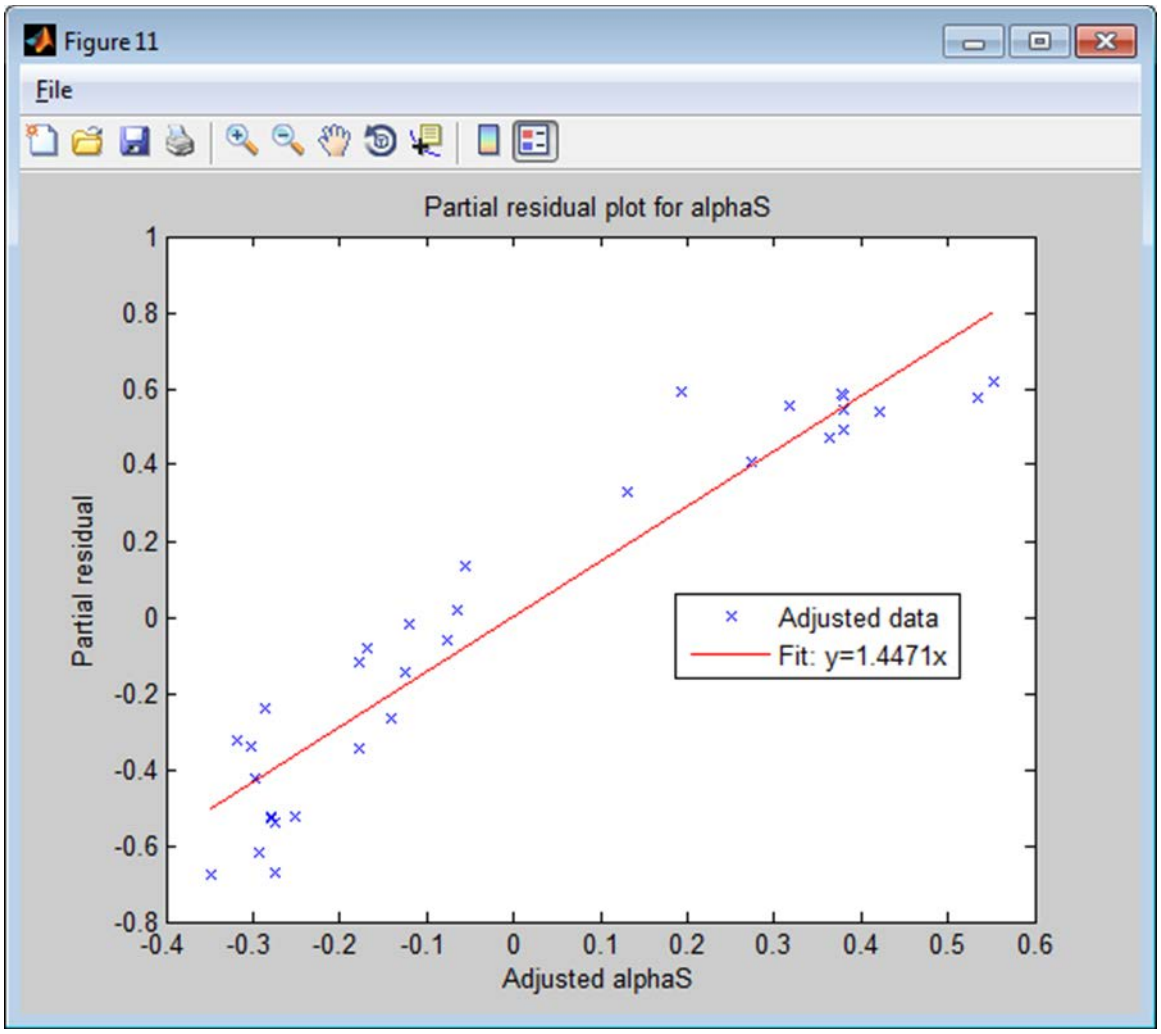


Figure 9. Partial residual plot for a single variable from a multiple linear regression (MLR) model.

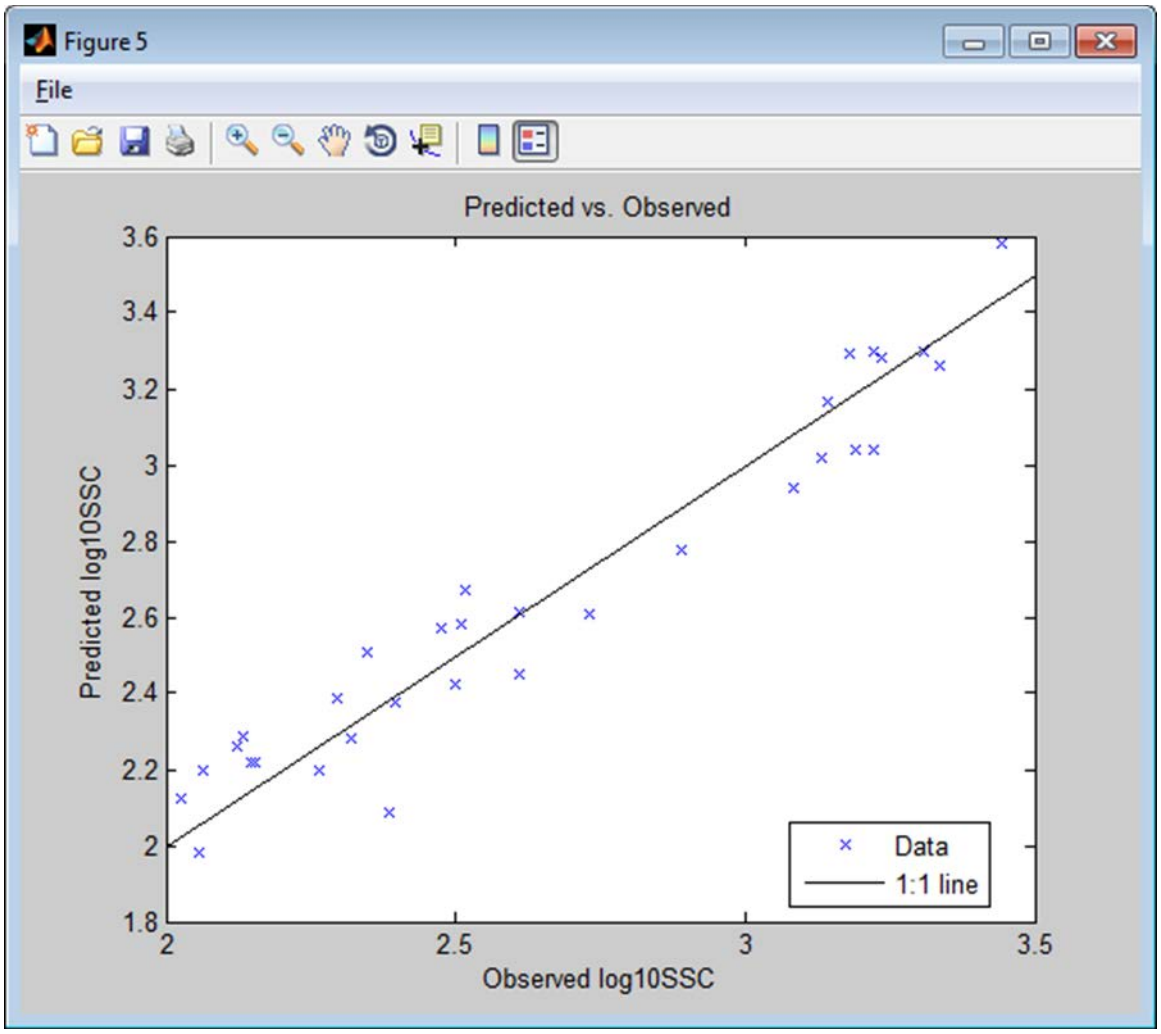


Figure 10. Predicted response variable plotted against observed response variable.

Residual Plots

In order to assess whether or not the assumptions required of the residuals have been met when creating an OLS regression (Helsel and Hirsch, 2002, p. 231), several plots are available.

- Raw residual versus fitted response variable—Illustrates a plot of the raw residuals against the fitted response values and is useful for demonstrating homoscedasticity (fig. 11).
- Normal probability plot of raw residuals—Useful for demonstrating normality of residuals. The probability plot correlation coefficient (PPCC) calculated in the model report is a measure of the linearity of residuals on the normal probability plot (fig. 12).
- Standardized serial correlation—Stan Ser. Corr. plot of the residuals shown with a locally weighted least squares (LOWESS) fit line to detect autocorrelation (fig. 13). If the LOWESS fit line shows a trend that deviates far from 0, serial correlation may be present.
- Raw residuals versus time—Raw residuals plotted against time to determine whether a time dependent trend exists with the residuals (fig. 14).

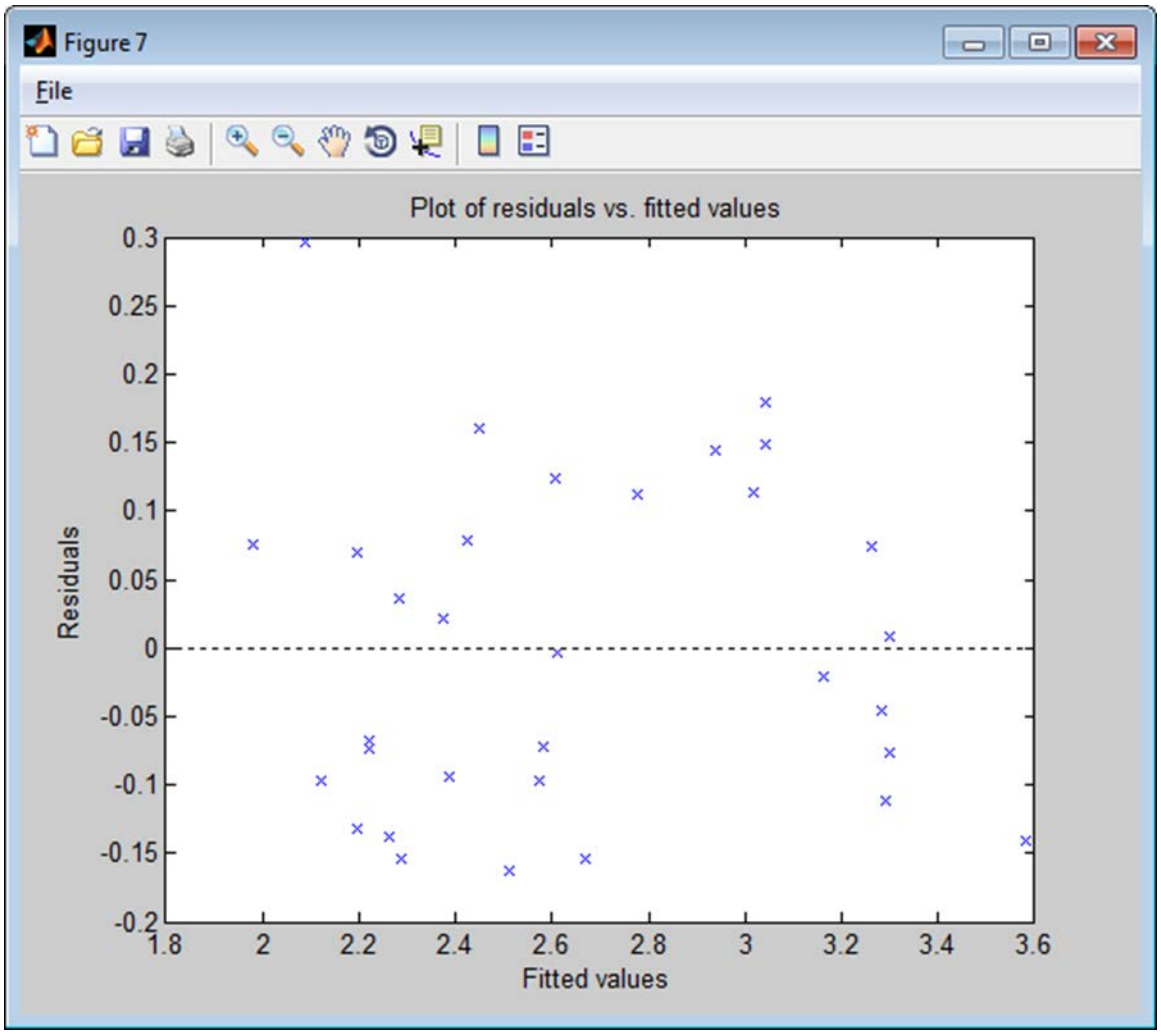


Figure 11. Raw residuals plotted against the fitted response variable.

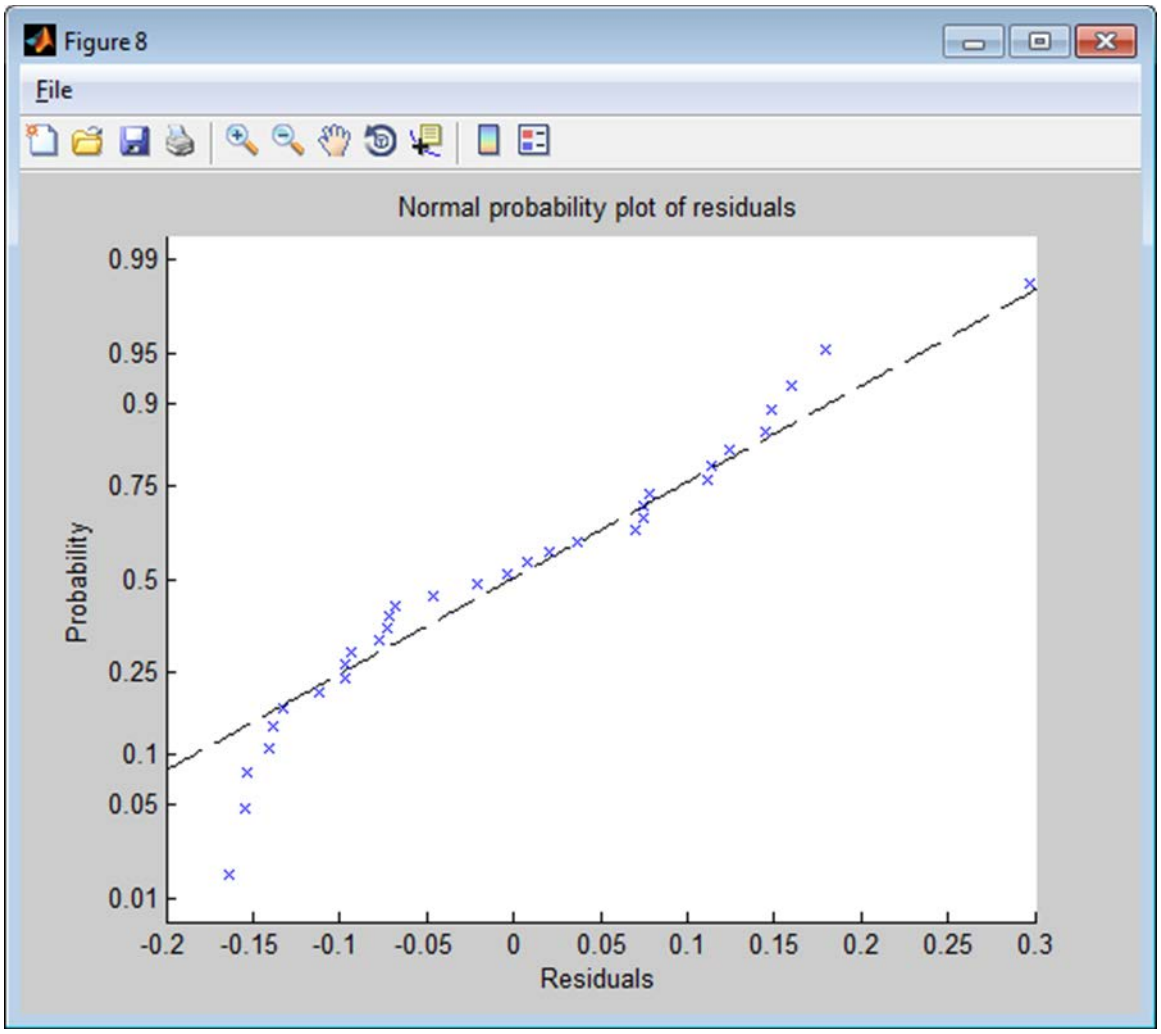


Figure 12. Normal probability plot of the raw residuals from a linear regression.

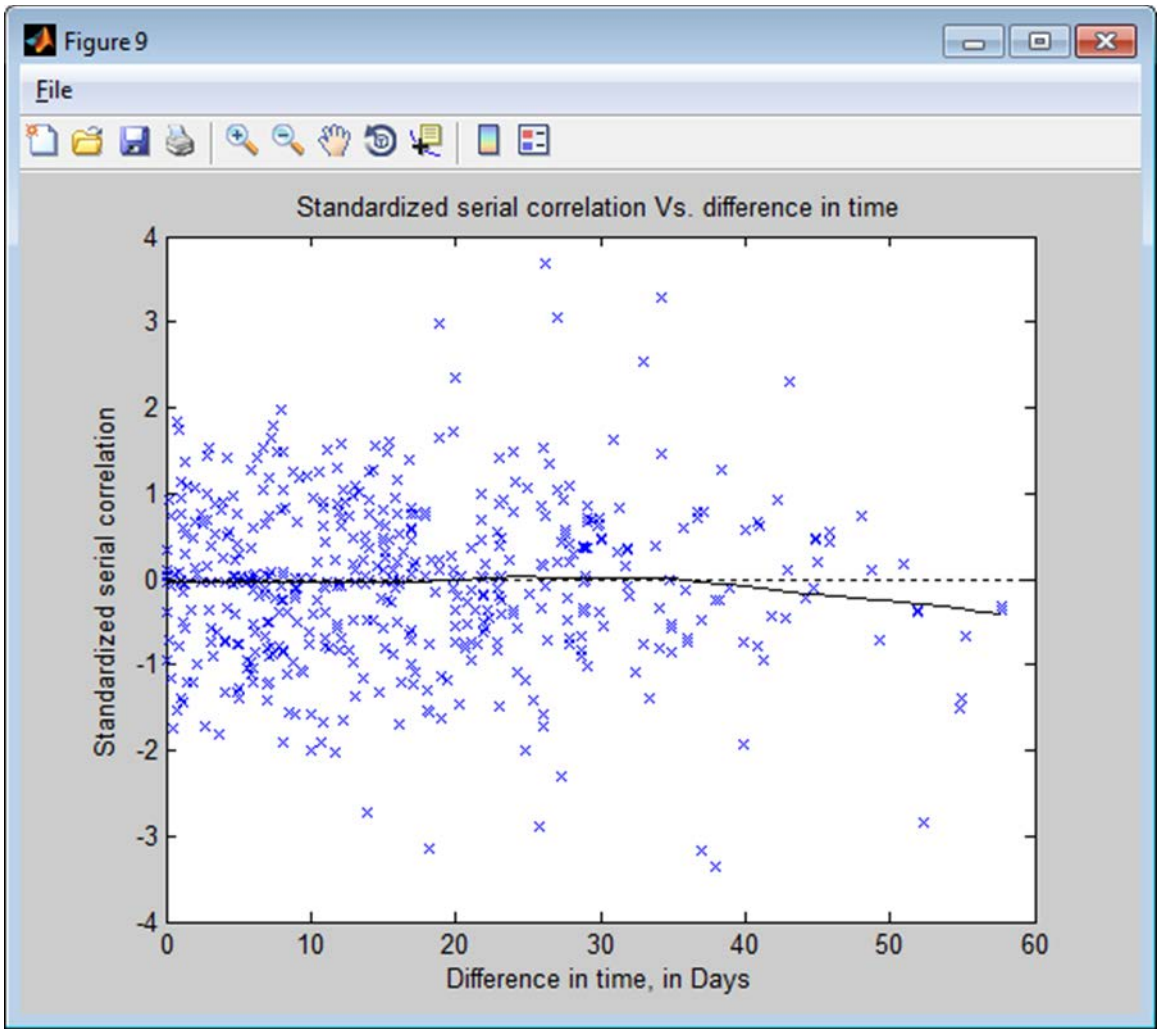


Figure 13. Standard serial correlation plot.

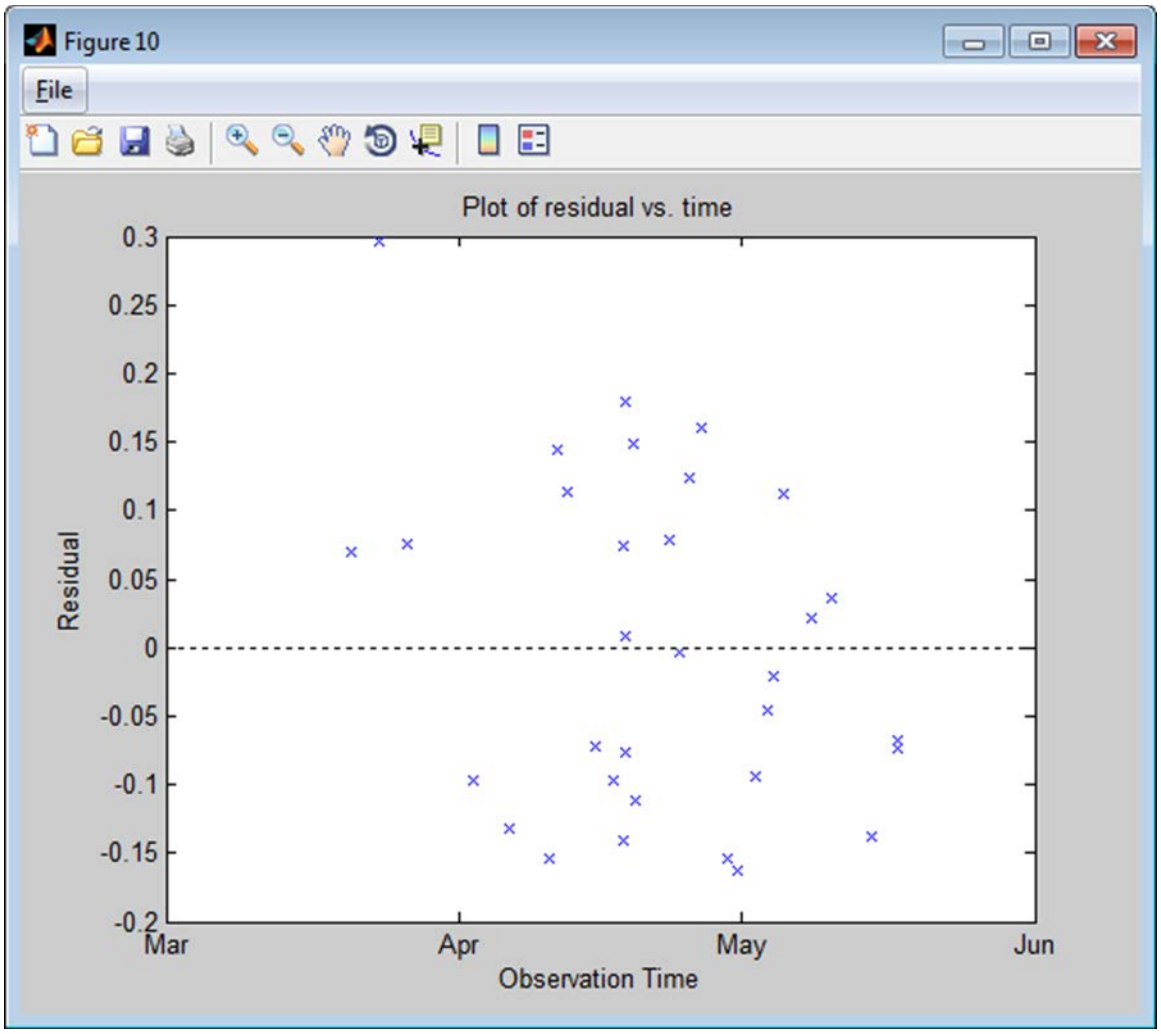


Figure 14. Raw residuals plotted against time.

Display Model

The Display Model button will provide a window that displays the model results and statistics (fig. 15). This information also can be written to a report with the Write Report button, along with other information as described later in this report. The following information can be viewed by clicking the Display Model button:

- The linear equation
- Coefficient estimates
- Estimated confidence intervals
- Coefficient of determination (R-squared) values for the model
- Root mean squared error

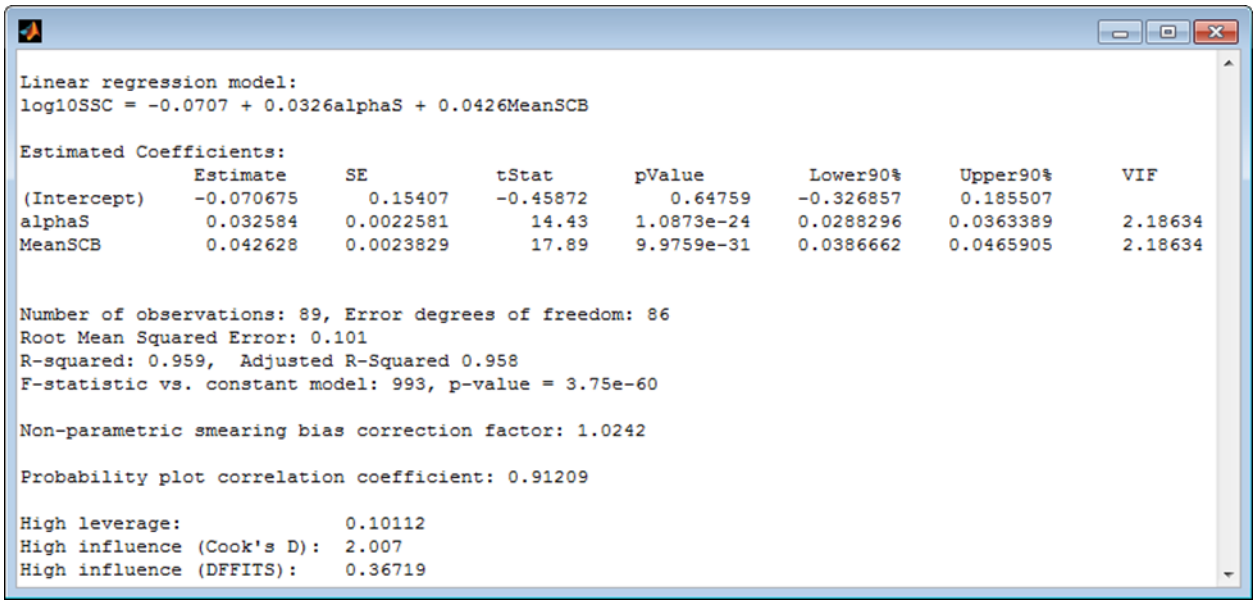


Figure 15. Linear model regression statistics window.

View/Edit Table

Clicking on the View/Edit Table button will display a window that contains the observation information used in the model (fig. 16). The information shown is the observation number, the corresponding constituent date and time variable, the response variable, and the explanatory variables. Also shown are diagnostic statistics for each observation for outlier detection (Helsel and Hirsch, 2002). 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 excludes the corresponding observations from the model. Once a date and time is flagged, the corresponding observation will be excluded (regardless of which variables are chosen) until the Restore All Observations button is clicked. The Restore All Observations button clears the flags from all date and time entries.

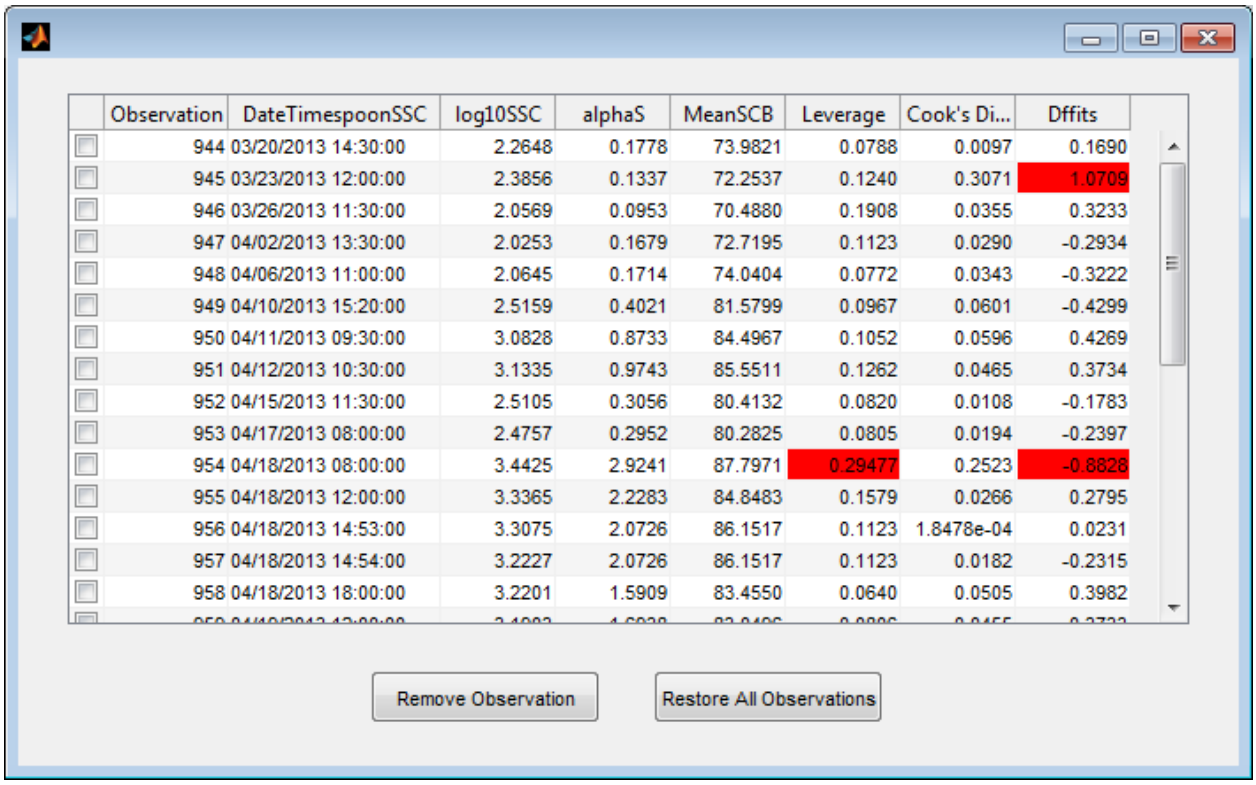


Figure 16. View/Edit Table window.

Write Report

To write a full 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 which you can write the report. Selecting and entering a valid location and file name will write the report. The contents of the report include the following:

- ADVM configuration and processing options
- Dataset 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
    - An estimate of the non-transformed variable with bias correction applied (if the response variable is transformed)
    - Calculated outlier indicator values
- The observations that were removed from the model dataset

Time Series

It also is possible to calculate a predicted time series of the constituent variable using the current linear model in the SAID workspace. The 90-percent prediction interval also is calculated. The constituent variable is predicted from a single dataset that contains all of the necessary surrogate

variables. The reported time of the predicted constituent is taken from this surrogate dataset. If ADVM variables are used as explanatory variables in the linear model, the dataset must contain the variables necessary to compute them. The same processing and configuration options are applied to the loaded dataset in computing the ADVM parameters. Only one dataset can be loaded. After SAID calculates the predicted variables, a plot of the time series is shown. You also have the option to write the predicted time series to a text file.

The Time Series button, when clicked, will first prompt you for a dataset file to load. After selecting a dataset file, a figure showing the predicted time series will be shown with the 90-percent prediction interval (fig. 17). You also will be prompted for a name and location of a tab delimited text file to which you can write the predicted time series. If you do not want to write a text file, click on the Cancel button.

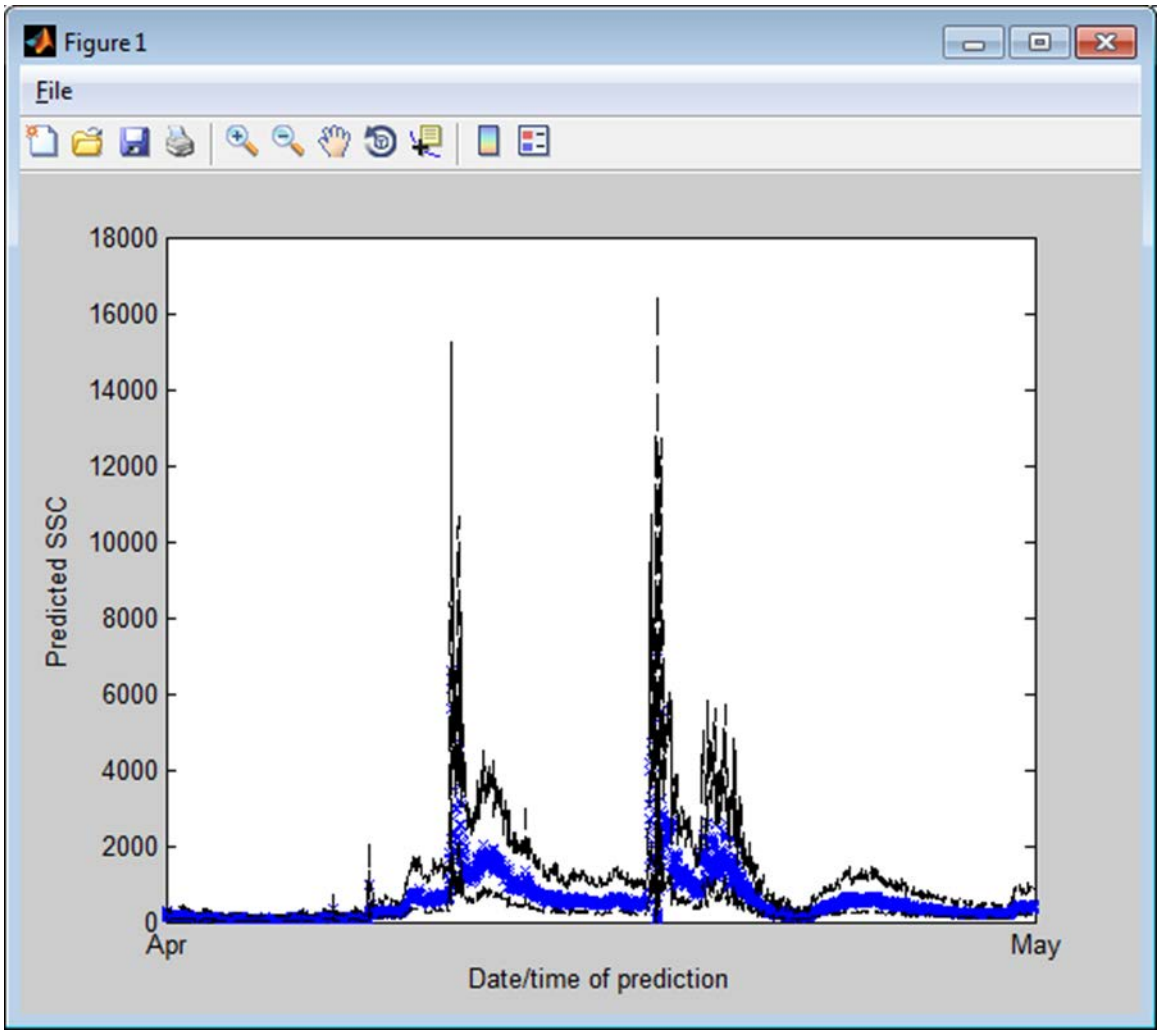


Figure 17. Predicted time series with prediction interval plotted against time.

SAID Workspace

At any time, a SAID workspace can be cleared, saved, loaded, or exited. Clicking the Clear button clears all loaded datasets and models. Clicking the Save button allows a user to save all the loaded data and existing models so that they can be loaded at a later time if needed. Clicking the Exit button will close the SAID program completely. More details on the file structure and items saved are presented in Appendix 1.

Acknowledgments

The USGS Office of Surface Water, the Federal Interagency Sedimentation Project, and the USGS Midwest Region (MWR) supported the development of this tool. The MWR support was through the River Sediments and Nutrients Investigations Initiative. The authors thank the many beta testers for their thorough testing and comments.

References Cited

Downing, Andrew, Thorne, P.D., and Vincent, C.E., 1995, Backscattering from a suspension in the near field of a piston transducer: Journal of The Acoustical Society of America, v. 97, no. 3, p. 1614–1620.

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources—Hydrologic analysis and interpretation: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 510 p.

Landers, M.N., 2012, Fluvial suspended sediment characteristics by high-resolution, surrogate metrics of turbidity, laser-diffraction, acoustic backscatter, and acoustic attenuation: Atlanta, Ga., Georgia Institute of Technology, Ph.D. dissertation, 236 p., accessed July 28, 2015, at http:hdl.handle.net/1853/43747.

Rasmussen, P.P., Gray, J.R., Glysson, G.D., and Ziegler, A.C., 2009, Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity-sensor and streamflow data: U.S. Geological Survey Techniques and Methods book 3, chap. C4, 53 p.

Topping, D.J., Melis, T.S., Rubin, D.M., and Wright, S.A., 2004, High-resolution monitoring of suspended-sediment concentration and grain size in the Colorado River in Grand Canyon using a laser acoustic system, *in* Proceedings of the 9th International Symposium on River Sedimentation, Yichang, China, October 18–21, 2004: Tsinghua University Press, p. 2507–2514.

Topping, D.J., Wright, S.A., Melis, T.S., and Rubin, D.M., 2006, High-resolution monitoring of suspended-sediment concentration and grain size in the Colorado River using laser-diffraction instruments and a three-frequency acoustic system, *in* Proceedings of the 8th Federal Inter-Agency Sedimentation Conference, Reno, Nev., April 2–6, 2006: p. 555–559.

Topping, D.J., Wright, S.A., Melis, T.S., and Rubin, D.M., 2007, High-resolution measurement of suspended-sediment concentrations and grain size in the Colorado River in Grand Canyon using a multi-frequency acoustic system, *in* Proceedings of the 10th International Symposium on River Sedimentation, Moscow, Russia, August 1–4, 2007: p. 330–339.

Wood, M.S., and Teasdale, G.N., 2013, Use of surrogate technologies to estimate suspended sediment in the Clearwater River, Idaho, and Snake River, Washington, 2008–10: U.S. Geological Survey Scientific Investigations Report 2013–5052, 30 p.

Wright, S.A., Topping, D.J., and Williams, C.A., 2010, Discriminating silt-and-clay from suspended-sand in rivers using side-looking acoustic profilers, *in* Proceedings of the 2nd Joint Federal Interagency Conference, Las Vegas, Nev., June 27–July 1, 2010: 12 p.

Appendix 1. The Surrogate Analysis and Index Developer (SAID) Tool Workspace Structure

In order to run SAID from source code or work with the SAID MAT file in MATLAB, you must have the Statistics and Curve Fitting Toolboxes installed.

Data Organization

Within SAID, data are organized within instances and collections of instances of the MATLAB dataset class. The linear model regression is done by the MATLAB LinearModel class. The dataset and LinearModel classes are defined within the Statistics Toolbox.

- Constituent dataset—Dataset that contains the constituent observations. The constituent dataset is loaded from a tab delimited text file and left unaltered.
- Surrogate variable structure—Structure that holds the explanatory variable datasets. Each explanatory variable, along with the backscatter variables, is stored in a separate dataset. As surrogate datasets are loaded into SAID, surrogate variables are separated into single variable datasets. The single variable datasets are then added to the structure.
- Matched dataset—Dataset that contains observations of constituent and surrogate variables that have been matched. The matched dataset is created by making a copy of the constituent dataset and filling in surrogate variable observations during matching.
- Linear model class instance—An instance of the MATLAB LinearModel class. The matched dataset is used in the creation of the linear regression model.

Update Process

- Acoustic Doppler velocity meter (ADVM) variable calculation—The ADVM variables are calculated.
- Variable transformation—Any variables that require transformation are transformed.
- Matching—Surrogate variable observations are matched to constituent dataset observations.
- Model creation—A linear regression model is created.
- Graphic user interface (GUI) refresh—The GUI is updated.

Clearing, Saving, and Loading

- Clearing—All loaded data can be cleared and settings reset to default.
- Saving—At any time, you can save a model state in order to load it at a different time. To save, click the Save button under the SAID workspace. This saves all loaded datasets, matched datasets, configuration data, and user-specified model characteristics. The saved MAT-file contains information that can be loaded into a MATLAB session for analysis.
- Loading—To load a previously saved workspace, click the Load button under the SAID workspace.

MAT File Contents

- advmParamStruct—MATLAB structure containing configuration and processing parameters used to calculate the ADVM variables. The fields of the structure are as follows: Frequency, EffectiveDiameter, BeamOrientation, SlantAngle, Nbeams, BlankDistance, CellSize,

NumberOfCells, BeamNumber, MovingAverageSpan, BSValues, IntenScale, RMin, RMax, MinCells, MinVbeam, NearField, and RemoveMinWCB.

- loaded_var_struct—Structure that holds the datasets of surrogate variables.
- CWD—Current working directory of the SAID workspace.
- surr_full_file—Cell array that holds the full file paths to the loaded surrogate datasets.
- const_full_file—Cell array that holds the full file path to the constituent dataset.
- trans_vars—Cell array that holds the names and transformations of variables within the workspace.
- bsPlotsFigNum—The figure number of the backscatter plots figure. This is needed for the program to keep track of the figure for opening and closing.
- max_time_min—This is the maximum absolute difference in time used for matching observations.
- const_ds—Dataset class instance that holds the constituent observations.
- Update flags—(UpdateMatch, UpdateProc, UpdateTrans, UpdateGUI, and UpdateMDL).
- matched_ds—Dataset class instance. Holds the observations of variables that have been matched. Created after the matching process. Gets passed to the linear model.
- mdl—LinearModel class instance. Performs linear regression. Provides statistics.
- ExcludeDates—Array that holds the date serial numbers of the observations that are excluded from the linear regression.
- session_name—The name of the current SAID session. Taken from the root name of the saved MAT file.

ISSN 2331-1258 (online)
http://dx.doi.org/10.3133/20151177