Biological Stability Analysis Tool (BSAT) User's Guide

Version 1.0

Web Report #4312

Biological Stability Analysis Tool (BSAT) User's Guide

Version 1.0

About the Water Research Foundation

The Water Research Foundation (WRF) is a member-supported, international, 501(c)3 nonprofit organization that sponsors research that enables water utilities, public health agencies, and other professionals to provide safe and affordable drinking water to consumers.

WRF's mission is to advance the science of water to improve the quality of life. To achieve this mission, WRF sponsors studies on all aspects of drinking water, including resources, treatment, and distribution. Nearly 1,000 water utilities, consulting firms, and manufacturers in North America and abroad contribute subscription payments to support WRF's work. Additional funding comes from collaborative partnerships with other national and international organizations and the U.S. federal government, allowing for resources to be leveraged, expertise to be shared, and broad-based knowledge to be developed and disseminated.

From its headquarters in Denver, Colorado, WRF's staff directs and supports the efforts of more than 800 volunteers who serve on the board of trustees and various committees. These volunteers represent many facets of the water industry, and contribute their expertise to select and monitor research studies that benefit the entire drinking water community.

Research results are disseminated through a number of channels, including reports, the Website, Webcasts, workshops, and periodicals.

WRF serves as a cooperative program providing subscribers the opportunity to pool their resources and build upon each other's expertise. By applying WRF research findings, subscribers can save substantial costs and stay on the leading edge of drinking water science and technology. Since its inception, WRF has supplied the water community with more than $460 million in applied research value.

More information about WRF and how to become a subscriber is available at www.WaterRF.org.

Biological Stability Analysis Tool (BSAT) User's Guide

Version 1.0

Prepared by:
Patrick J. Evans and Jennifer L. Hooper
CDM Smith, 14432 Southeast Eastgate Way, Suite 100, Bellevue, WA 98007-6493

Mark W. LeChavallier, Orren D. Schneider, Lauren A. Weinrich, and Patrick K. Jjemba
American Water, 1025 Laurel Oak Road, Voorhees, NJ 08043

and

Rick W. Chappell
Environmental Science Solutions, 8059 Wolf Street, Westminster, CO 80031-7012

DISCLAIMER

CONTENTS

TABLES

FIGURES

FOREWORD

The Water Research Foundation (Foundation) is a nonprofit corporation dedicated to the development and implementation of scientifically sound research designed to help drinking water utilities respond to regulatory requirements and address high-priority concerns. The Foundation's research agenda is developed through a process of consultation with Foundation subscribers and other drinking water professionals. The Foundation's Board of Trustees and other professional volunteers help prioritize and select research projects for funding based upon current and future industry needs, applicability, and past work. The Foundation sponsors research projects through the Focus Area, Emerging Opportunities, and Tailored Collaboration programs, as well as various joint research efforts with organizations such as the U.S. Environmental Protection Agency and the U.S. Bureau of Reclamation.

This publication is a result of a research project fully funded or funded in part by Foundation subscribers. The Foundation's subscription program provides a cost-effective and collaborative method for funding research in the public interest. The research investment that underpins this report will intrinsically increase in value as the findings are applied in communities throughout the world. Foundation research projects are managed closely from their inception to the final report by the staff and a large cadre of volunteers who willingly contribute their time and expertise. The Foundation provides planning, management, and technical oversight and awards contracts to other institutions such as water utilities, universities, and engineering firms to conduct the research.

A broad spectrum of water supply issues is addressed by the Foundation's research agenda, including resources, treatment and operations, distribution and storage, water quality and analysis, toxicology, economics, and management. The ultimate purpose of the coordinated effort is to assist water suppliers to provide a reliable supply of safe and affordable drinking water to consumers. The true benefits of the Foundation's research are realized when the results are implemented at the utility level. The Foundation's staff and Board of Trustees are pleased to offer this publication as a contribution toward that end.

Denise Kruger, Ph.D.   Robert C. Renner, P.E.
Chair, Board of Trustees   Executive Director
Water Research Foundation   Water Research Foundation

ACKNOWLEDGEMENTS

CHAPTER 1
INTRODUCTION

The software tool described in this User's Guide was developed to complement Water Research Foundation Project 4312, "An Operational Definition of Biological Stability in Drinking Water" (LeChevallier 2015a). The purpose of this project was to create a practical definition of water stability and develop an integrated decision-support system that embodies the factors affecting biological stability. The project involved collection of monitoring data from six utility distribution systems for 14 months. These data were used to develop a decision support framework for identifying the relationships between control and response parameters and refine the operational definitions of biostability. The study identified over-arching factors controlling biostability and is generally applicable to multiple facilities with varying water treatment processes and water quality. However, the study was not intended to provide tailored information or site-specific recommendations on system management for specific utilities. This Excel-based biological stability analysis tool (BSAT) was developed to compliment the information provided in the accompanying research report and guidance manual. Utility engineers, scientists, and consultants can use the tool to study individual systems and develop site-specific conclusions and recommendations regarding distribution system biostability.

BSAT is an interactive tool to assess water stability on a site-specific basis. The tool is intended to allow utilities to (1) track historical data, (2) review results from Water Research Foundation Project 4312 as an example for data analysis and interpretation, (3) assess system stability over time, and (4) identify possible controlling factors influencing water stability on a site-specific basis. With this tool, the user can select specific stability metrics to evaluate in detail, such as disinfectant residual stability, biofilm formation rate, and corrosion rate. The user can then determine which factors are most associated with instability, and what threshold values are important. This will allow the utility to identify instability causes and how to develop meaningful operational criteria to attain stability goals. BSAT was created for utility managers, engineers, and scientists for planning purposes to evaluate trends in biological stability using monitoring data collected over the period of seasons or years. The tool is not intended for use by operators on a day-to-day basis. Utilities are ultimately responsible for determining applicability of data based on their own site-specific water quality goals and the data provided in the Report and Guidance Manual from this project.

BSAT consists of four sequential components: 1) Utility Information, 2) Data Entry, 3) Data Analysis, and 4) Data Output. The following is a summary of the general steps for using BSAT (Figure 1.1):

(1) Enter general information including utility and facility name, distribution system disinfectant type (free chlorine or chloramine), flow paths (transmission axes), monitoring locations (data stations), and monitoring parameters (variables).
(2) Enter or import monitoring data for defined transmission axes and data stations.
(3) Conduct data analyses including summary statistics and comparing to the WRF Project 4312 dataset, trends over time, correlations among variables, and instability factor identification.
(4) Export, save and/or print graphical and tabular data analysis results.

Utilities may need to refine their water monitoring program to fill in data gaps to better determine which factors are most important for managing stability.
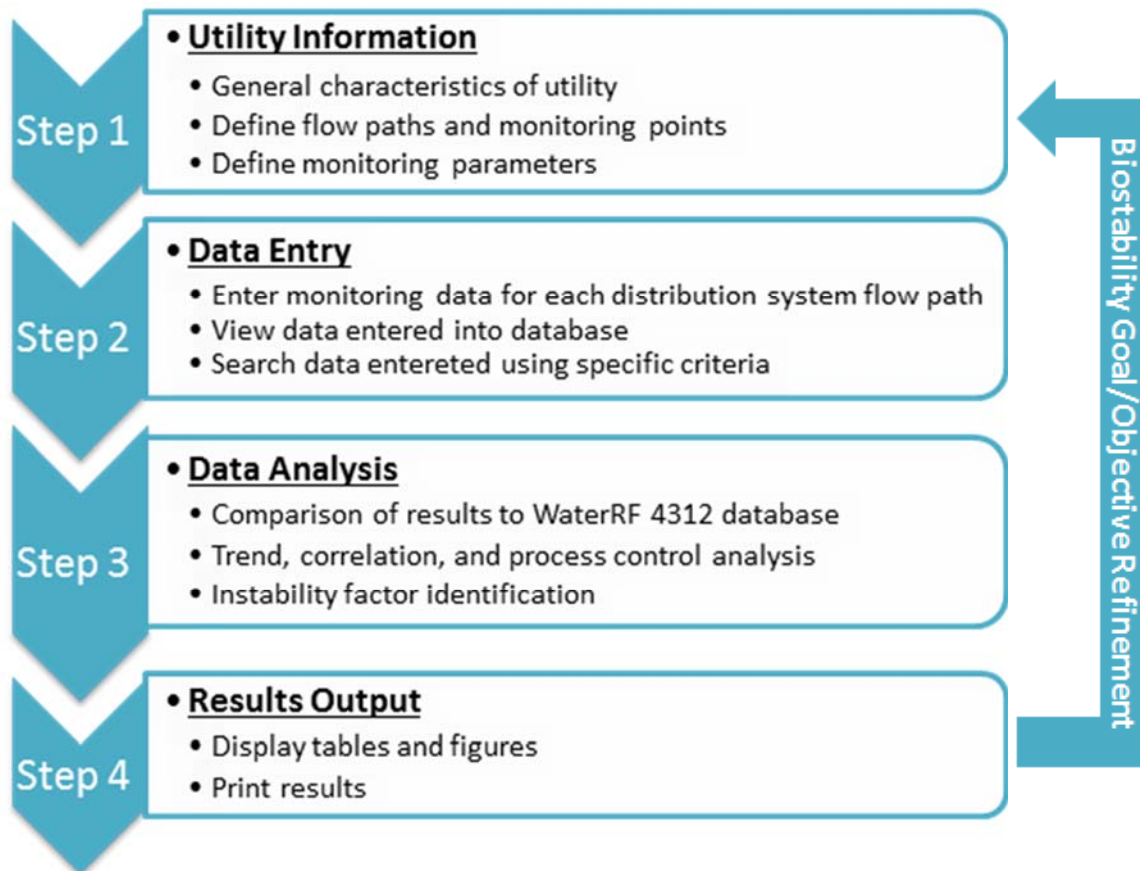
Figure 1.1 BSAT program flow chart

BSAT OVERVIEW

The user interface is accessed through the **BSAT** tab of Excel's ribbon. The **BSAT** tab contains icons separated into three groups: **Instructions**, **Information**, and **Data Analysis**. Selecting an icon opens various tools, or navigates the user to various locations within BSAT. Tools include message boxes and dialog boxes for entering data and modifying selections. Locations pertain to various sheets or pages within BSAT where the user can operate features independently of the ribbon icons via additional controls located on the page. Hovering the mouse over each icon in the ribbon will display more information on what the icon does. Detailed descriptions and operation of the various buttons, tabs, etc. will be shown with accompanying screenshots in the following chapters.

The **Instructions** group on the **BSAT** tab of the ribbon contains various icons designed to provide instructions to the user. Clicking the *Main* icon navigates the user to the main page, which contains a brief description of BSAT and instructions for accessing the **BSAT** tab of the ribbon. Selecting the *User's Guide* or *Guidance Manual* link open the respective portable document file

(pdf). Selecting the *Help* icon displays a drop-down list with four items: (1) *About BSAT*, (2) *Show Variable Codes*, (3) *Page and Button Tips*, and (4) *Basic Operations*. Selecting the Basic Operations item navigates the user to a condensed version of the User's Guide.

The **Information** group on the ***BSAT*** tab of the ribbon contains various icons provided to allow the user to enter data. The *Utility* icon is used to enter general information about their utility, including biological stability issues and/or monitoring variables of concern. The *Data Entry* icon is used to enter data into a database.

The **Data Analysis** group on the ***BSAT*** tab of the ribbon contains various icons provided to allow the user to analyze their entered data, and to generate results output. Three icons pertain directly to data analysis: *Comparison to WRF 4312 Data, Trend and Correlation Analysis, and Instability Factor Identification*. Selecting any of these three icons navigates the user to the associated data analysis page where results of the specific analysis are displayed. Each data analysis page contains a button located in the upper left corner of the page. Clicking the button opens a dialog box where various selections can be made specific to that page or analysis. Finally, the *Results Output* icon is used to select analysis results to save. The results output file is a pdf file saved to the same folder as BSAT.

CHAPTER 2
NAVIGATION

There are three levels of navigation within the BSAT – groups, icons, buttons, and tabs. For the purposes of this User's Guide, **Group** names will be designated by bold font, *icon* names by italic font, button names by underlined font, and ***tabs*** by bold italic font.

SOFTWARE REQUIREMENTS

BSAT is a spreadsheet (workbook) application that requires Microsoft Excel 2007 or later versions. The application contains macros written in Excel's Visual Basic for Applications (VBA) programming language.

RECOMMENDED HARDWARE REQUIREMENTS

- 1 gigahertz (GHz) or faster Processor.
- 1 gigabyte (GB) of memory (RAM).
- 1024 by 768 Pixels or greater color display.

INSTALLATION AND OPERATING INSTRUCTIONS

BSAT does not have any special installation requirements. Simply download the file from the Water Research Foundation website and place it in a designated location or directory (folder) on your computer. If you have already saved this User's Guide and the Guidance Manual, save the BSAT file in the same directory as these documents. Double-click the BSAT file name or open the file to operate BSAT.

USE OF THE RIBBON

Primary operation of BSAT is provided via icons located on the ***BSAT*** tab of the Excel ribbon. If the ***BSAT*** tab is not visible, see page 47 for the security warnings and settings. There are three groups of BSAT icons on the ribbon: **Instructions**, **Information**, and **Data Analysis** (Figure 2.1). The **Instructions** group has a *Main* icon, a *User's Guide* icon, a *Guidance Manual* icon, and a *Help* icon. The **Information** group contains a *Utility* icon and a *Data Entry* icon and the **Data Analysis** group contains three analysis icons, *Comparison to WRF 4312 Data*, *Trend and Correlation Analysis* and *Instability Factor Identification*, along with a Results *Output* icon.
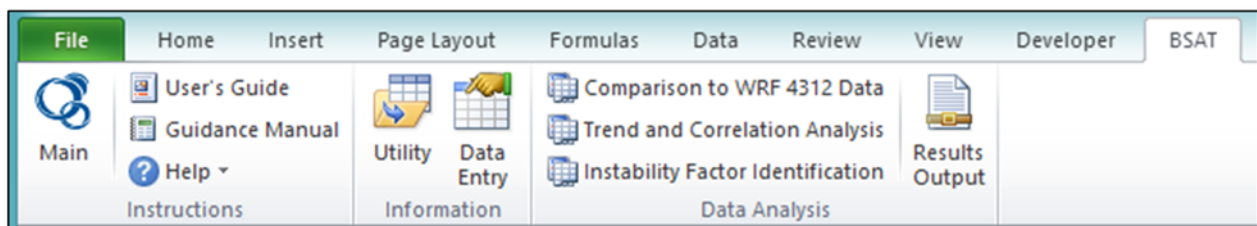
Figure 2.1 BSAT ribbon

PAGE BUTTONS

Page buttons are located at the top of the *Data Entry*, *Comparison to WRF 4312 Data*, *Trend and Correlation Analysis*, and *Instability Factor Identification* pages. Typically, the user clicks on the page button to open a dialog box, which is used for manipulation of the data analysis controls on the dialog box (Figure 2.2).
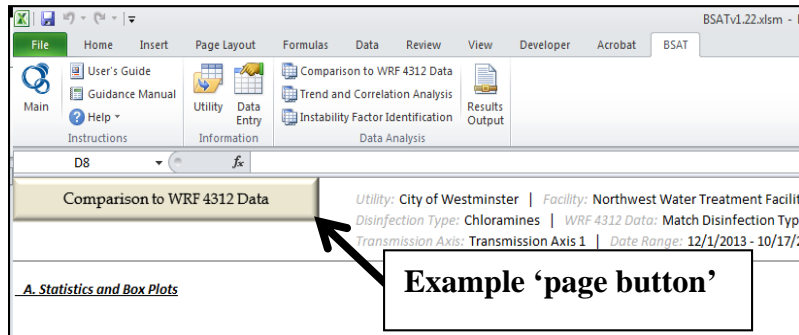
Figure 2.2 Example page button

TIPS

Information and **Data Analysis** groups both have dialog boxes with Tips buttons, which provides additional information for the controls on the dialog box (Figure 2.3). They are typically located in the lower left-hand corner of the dialog box. Clicking the Tips button opens a message box with information intended to help the user with using the dialog box.
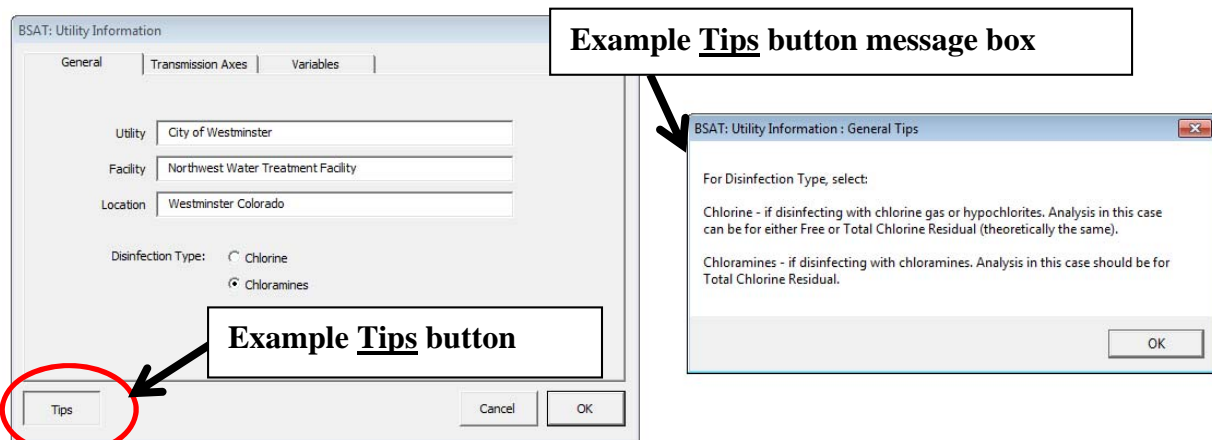
Figure 2.3 Example Tips button

BSAT can be operated entirely by the keyboard (no mouse required). After opening BSAT, press the Alt key to display the ribbon accelerator keys (these appear in little boxes below the ribbon tabs). Press the accelerator key assigned to the *BSAT* tab (typically this will be "Y"). The

BSAT icons will then be displayed with their accelerator keys (typically a letter and number, e.g., "Y5"): press them in sequence to activate an icon. After selecting the *Help* icon, use up/down arrows to make a selection and press Enter. On the *Data Entry* page, press Ctrl-J to activate the Data Entry Form, Ctrl-M to activate the Update Formulas button, and Ctrl-E to activate the Data Options button. On the dialog boxes and message boxes, pressing the Esc key activates the Cancel or Close buttons; if there are no buttons, Esc will close the dialog box. The Enter key defaults to the OK button unless another control has the focus. The Tab button sequences through the controls. Use the Space Bar to toggle selections and the Up/Down or Left/Right keys to change selections for certain controls.

CHAPTER 3
INSTRUCTIONS

Follow the general steps below for data entry and analysis from the **BSAT** tab on the Ribbon (Figure 3.1):

1. Select icons in the **Instructions** group to learn about BSAT and how to operate the program.
2. Open the *Utility* dialog box to enter information about the facility, monitoring locations and monitoring parameters (variables).
3. Navigate to the *Data Entry* page to enter historical monitoring data into the utility database.
4. Navigate to the *Comparison to WRF 4312 Data* page to see summary statistics along with the WRF Project 4312 dataset.
5. Navigate to the *Trend and Correlation Analysis* page to evaluate data over time and relationships among variables.
6. Navigate to the *Instability Factor Identification* page to conduct a regression tree analysis and determine which factors are contributing to instability and what target values are acceptable.
7. Open the *Results Output* dialog box to save selected results as a pdf file.
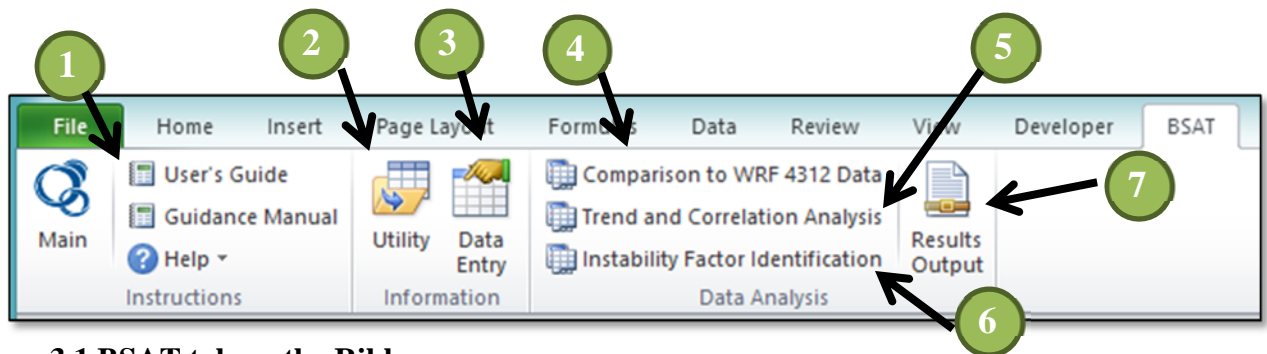
Figure 3.1 BSAT tab on the Ribbon

Figure 3.2 diagrams the various BSAT references available to the user by selecting icons located on the Instructions group of the BSAT tab. These are the references intended to provide guidance on how to use BSAT and how to interpret data analysis results.
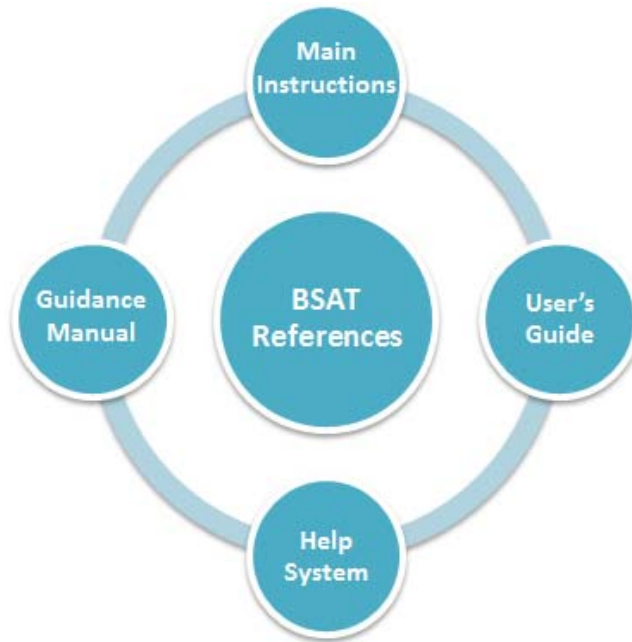
Figure 3.2 BSAT references

MAIN PAGE

The Main Page is accessed through the *Main Page* icon in the **BSAT** tab of the ribbon (Figure 3.3). This page provides a brief description of what the tool was developed for and a quick reference on how to use the tool.
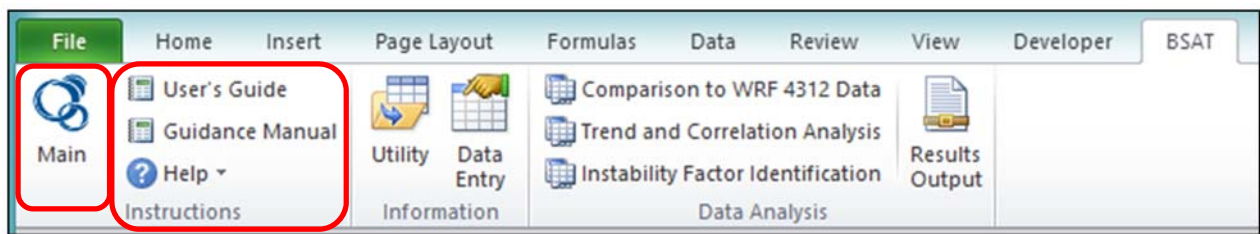
Figure 3.3 Main Page, User's Guide, Guidance Manual and Help icons in the Instructions group.

USER'S GUIDE AND GUIDANCE MANUAL

A link to the User's Guide is located on on the **BSAT** tab of the ribbon. The User's Guide is intended to provide information on how to download, install, and navigate BSAT. The Guidance Manual (LeChevallier 2015b) should be referenced for data interpretation, in addition to the Data Interpretation section included in this User's Guide. The Guidance Manual, in conjunction with the research report for WRF project 4312 (LeChevallier 2015a), provides the framework and

scientific basis for data interpretation. A link to the Guidance Manual is also located on the ***BSAT*** tab of the ribbon.

Opening the User's Guide and Guidance Manual (and also results output pdf files created by BSAT) require that the user have the free program Adobe Reader installed on their computer. To download the latest version of Adobe Reader go to URL: http://get.adobe.com/reader/. Note that these pdf files can also be opened outside of BSAT.

HELP ICON

There are four *Help* icon that can be selected on the ***BSAT*** tab of the ribbon: *About BSAT*, *Show Variable Codes*, *Page and Button Tops*, and *Basic Operations* (Figure 3.5). The *About BSAT* icon displays information on the tool and authorship. The *Show Variable Codes* icon displays the name, units, and code used by the program for analytical variables such as temperature and dissolved organic carbon. For example, the BSAT code for temperature is "TEMP" (all uppercase); thus temperature data must be entered into BSAT as "TEMP", and temperature results are displayed on graphs and tables within BSAT as "TEMP". Additional help features are incorporated in each window interface or page button through the <u>Tips</u> buttons in the *Data Entry* or *Data Analysis* dialog boxes. These Page and button tips can be toggled on or off by selecting the *Page and Button Tips* icon.
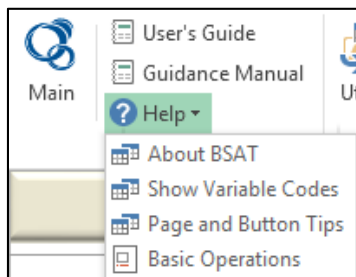
Figure 3.4 Help features include About BSAT, Show Variable Codes, Page and Button Tips, and Basic Operations items

The *Basic Operations* icon navigates the user to the BSAT *Basic Operations* page, which contains a condensed version of the User's Guide intended to be a quick reference for how to operate BSAT controls and features. The *Basic Operations* page can be navigated by scrolling up and down or by clicking the <u>BSAT Basic Operations</u> button to open the Topic Index dialog box (Figure 3.6). The Topic Index dialog box can remain opened while the user navigates around the page. It remains opened until either the <u>Close</u> button is clicked, the "X" is clicked, the Esc key is pressed, or if the user navigates to a different page within BSAT or to a different Excel window. While opened, selecting a topic on the list navigates to the selected topic. The user can also scroll the Topic Index with the scroll bar, or use the keyboard (up arrow, down arrow, Home, End, Page Up, Page Down).
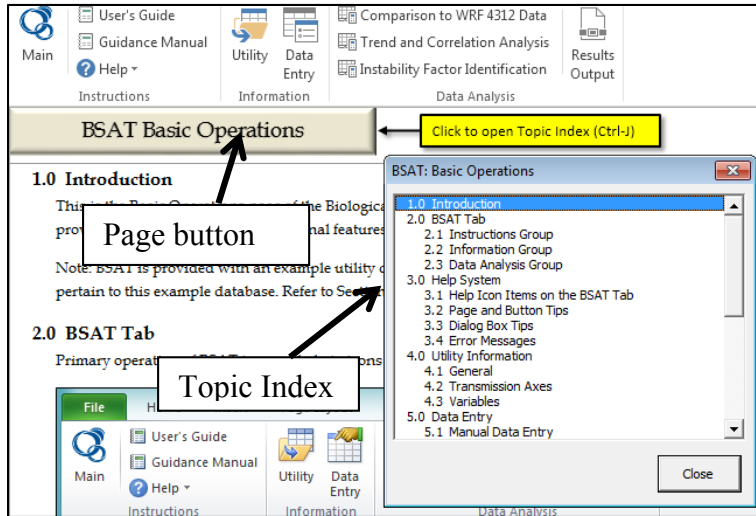
Figure 3.5 Basic Operations page button and Topic Index

CHAPTER 4
UTILITY INFORMATION

The **Information** group on the *BSAT* tab of the ribbon is where the user enters information about the utility, including transmission axes names, sampling location names, and monitoring data (Figure 4.1). The variables included for analysis are those parameters that were found to be associated with biological stability from WRF project 4312, as documented in the Report called *An Operational Definition of Biological Stability* (LeChevallier 2015a) and the *Guidance Manual for Control of Biostability in Drinking Water* (LeChevallier 2015b). The actual data entered will be contingent on the utility's available data, and future sampling from the distribution system may need to be conducted to provide a more robust dataset for evaluation.
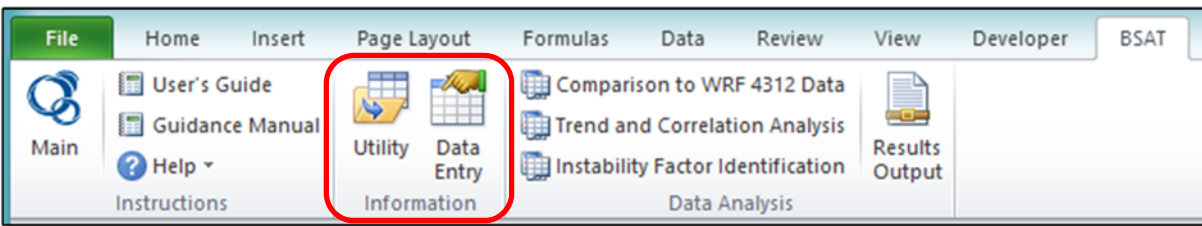
Figure 4.1 Information group on the BSAT tab of the ribbon

UTILITY

This icon is used for entry of basic information on the utility. To open the dialog box, click the *Utility* icon in the **Information** group of BSAT. The *Utility Information* dialog box is a form with three tabs: ***General***, ***Transmission Axes***, and ***Variables***, described in detail below.

General Tab

This tab collects data on the utility name, facility name, water treatment plant location, and the disinfectant type used in the distribution system (Figure 4.2). BSAT is not a web-based program and thus does not share this information with any other program.

The distribution system disinfectant type must be either chlorine or chloramine. This allows the user to compare site-specific data to the WRF Project 4312 utilities that share the same disinfectant type, as systems with chlorines and chloramines have different factors affecting stability. The <u>Tips</u> button on the dialog box provides additional detail on how to select the disinfectant type.
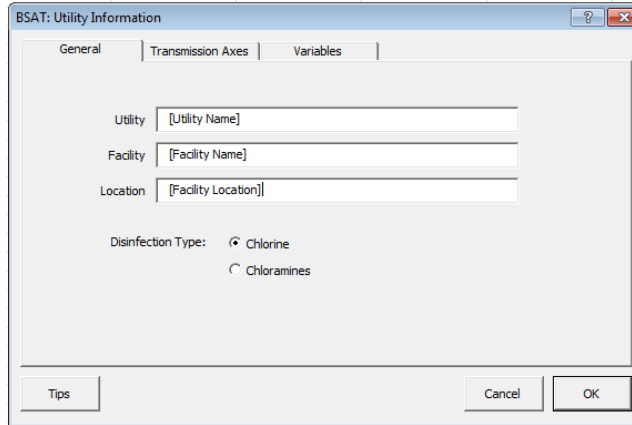
Figure 4.2 General tab of the Utility Information dialog box

Transmission Axes Tab

Each transmission axis [as defined on page 5 of the accompanying Guidance Manual (LeChevallier 2015b)] represents a portion of the distribution system. A transmission axis is not intended to be an exact flow path, but rather to represent general water quality changes as water age increases in the system. There are three locations associated with each transmission axis. The first location (DS1 for distribution system site 1) represents the beginning of the distribution system (e.g., treatment plant finished water), the second location (DS2) represents a point mid-way through the system, and the third location (DS3) represents a point with a longer hydraulic residence time within the system. The actual monitoring point name for DS1, DS2, and DS3 should be entered into this dialog box for tracking and display purposes on the analyses page headers.

If there is more than one facility that provides water for the distribution system, then the area of the distribution system primarily influenced by each treatment plant should have its own Transmission Axis. Additionally, if conditions are significantly different at different locations in the system then consider adding transmission axes with separate monitoring points to capture anticipated variability. Transmission axes can be added to the database by clicking the New button on the right hand side of the *Utility Information* dialog box (Figure 4.3). Transmission axes can also be deleted by clicking the Delete button. The user can also scroll between transmission axes if several are present via the Previous and Next buttons.
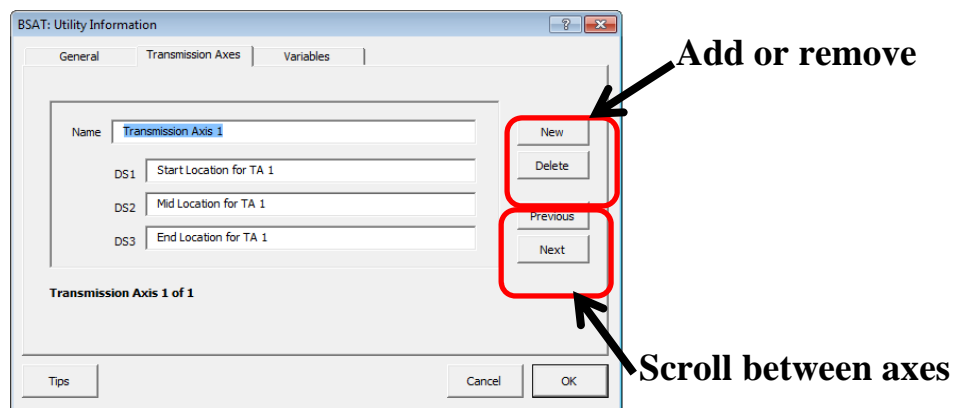
Figure 4.3 Transmission Axes tab of Utility Information dialog box

Variables Tab

The *Variables* tab is used to select the stability issues of concern and monitoring variables (Figure 4.4). Stability issues of concern can be disinfectant residual stability, biofilm formation rate, and/or corrosion rate. Refer to the Guidance Manual and the report *An Operational Definition of Biological Stability* for WRF 4312 for the technical basis for these issues (LeChevallier 2015a and 2015b). The box for disinfectant residual stability is automatically selected and cannot be deselected, as this is the baseline level of information for the program. If data on biofilm formation rate (measured by the accumulation of ATP on a metal coupon over time) or corrosion rate are available, then these stability issues can also be evaluated. Selecting either of these boxes automatically selects the equivalent box listed under Monitoring Variables. Monitoring Variables are the parameters (measured observations) that can be entered or imported for data analysis. The variables selected here will appear in the drop-down list on the Data Entry Form dialog box (see *Data Entry* section). Temperature (TEMP), hydraulic residence time (HRT), dissolved organic carbon (DOC), pH (PH), and chlorine residual (CLR) are automatically selected and cannot be deselected, because these variables are anticipated to be necessary to provide an adequate baseline set of data for analysis; it is recommended (but not required) that sufficient data for these variables be included in the utility database. Additional variables (also recommended but not required) include biofilm formation rate (BFR), corrosion rate (COR), phosphate (PO4), and assimilable organic carbon (AOC); BFR and COR are required if biofilm formation rate and/or corrosion rate, respectively, are biological stability issues of concern.
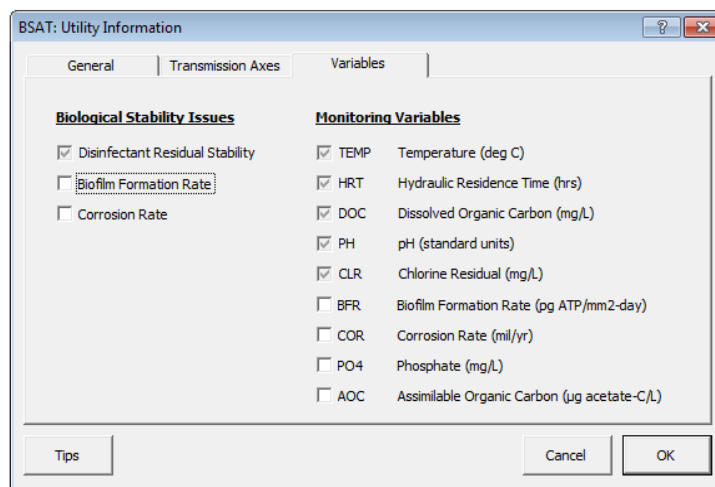
Figure 4.4 Variables tab of the Utility Information dialog box

DATA ENTRY

The Data Entry page is the location for entry of historical monitoring data used by BSAT to conduct analyses. Utilities may need to increase monitoring frequency and/or variables so that sufficient data are available for adequate analyses within BSAT. The Guidance Manual for this project should be referenced for information on which parameters are recommended for monitoring and at what frequency (LeChevallier 2015b). Click the *Data Entry* icon in the **Information** group of BSAT to navigate to the *Data Entry* page. BSAT provides several methods

and flexibility for entry of historical data into the database, including (1) entry using the data entry form, (2) manual entry directly on the page, or (3) entry by importing data from another file. Data must be entered using the variable codes specified in Table 4.1, or else the program will not recognize the data entered; the variables data should also be entered in the measurement units specified in Table 4.1. The dates entered must be in a date format, and the names of the transmission axes entered must match the transmission axes names specified in the *Utility Information* dialog box. The Data Entry form was developed to minimize the potential for entering incorrect data; however, copying/pasting or importing data onto the page is more efficient for entering large amounts of data. The data should be carefully checked by the user prior to performing data analysis to make sure no erroneous data were accidently entered. Data records can be entered in any order (e.g., any order for dates or variable names). BSAT does all necessary sorting for data analysis internally.

Table 4.1

Variable codes, names, and units

Code	Name	Units
TEMP	Temperature	deg. C
HRT	Hydraulic Residence Time	hrs
DOC	Dissolved Organic Carbon	mg/L
PH	pH	standard units
CLR	Chlorine Residual	mg/L
BFR	Biofilm Formation Rate	pg ATP/mm^2-day
COR	Corrosion Rate	mil/yr
PO4	Phosphate	mg/L
AOC	Assimilable Organic Carbon	µg acetate-C/L

Data Entry Form

To open the Data Entry Form, click the Data Entry Form page button at the top of the Data Entry page (Figure 4.5). The Data Entry Form dialog box can then be used to manually enter or modify data for individual data records.
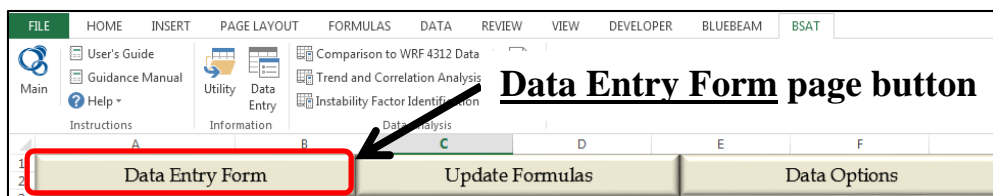
Figure 4.5 Data Entry Form page button

Once the Data Entry Form dialog box opens, a new data record can be entered by clicking the New or Insert buttons in the top right hand corner of the dialog box (Figure 4.6). The New button adds a record at the end of the database and the Insert button inserts a record before the currently displayed record. The first item to enter is the name of the Transmission Axis. The name of the transmission axis is defined in the *Utility Information* dialog box. Multiple defined transmission axes are selectable from the dropdown list. The variable should then be entered or

selected using its drop-down list. The list of variable names and units are shown in Table 4.1. The variable names and units must be entered as shown in the table so that they are comparable with the WRF Project 4312 dataset. This information is also accessible from the *Help* menu (*Help → Show Variable Codes*).

Dates can be entered in any acceptable date format; however, BSAT will internally reformat dates to display in the MM/DD/YYYY format. For example, if a date is entered in MM-DD-YY format, it will be reformatted to display as MM/DD/YYYY. Double-clicking the Date textbox on the Data Entry Form will display the popup calendar date picker; the date picker can be used to select the date. The DS1, DS2, and DS3 textboxes are where measured values are entered for the sequential data stations: DS1 (e.g., plant finished water), DS2 (e.g., a midpoint along the transmission axis), and DS3 (e.g., an endpoint along the transmission axis). Only one transmission axis/variable record should be entered for each date. If multiple data points are collected from a sampling site on the same day, the results should be averaged and then entered into the database.

After the data are entered, click the Previous or Next buttons to register the newly entered data in the database, or click the New or Insert buttons to register the newly entered data and continue to the next record to be entered. The most recent record entered can be removed by clicking the Undo Entry button; this will appear to the left of the Close button. The Undo button will appear as Undo New, Undo Insert or Undo Delete after these operations are performed. To scroll between records, use the Previous or Next buttons on the right hand side of the dialog box, or use the scroll bar on the bottom of the dialog box. The current record number is displayed at the bottom of the Data Entry Form dialog box.

The Mean (MEAN), Coefficient of Variation (CV), and Maximum (MAX) values are automatically calculated for each record based on the data from DS1, DS2 and DS3. These values are used for subsequent data analysis and cannot be entered or edited.
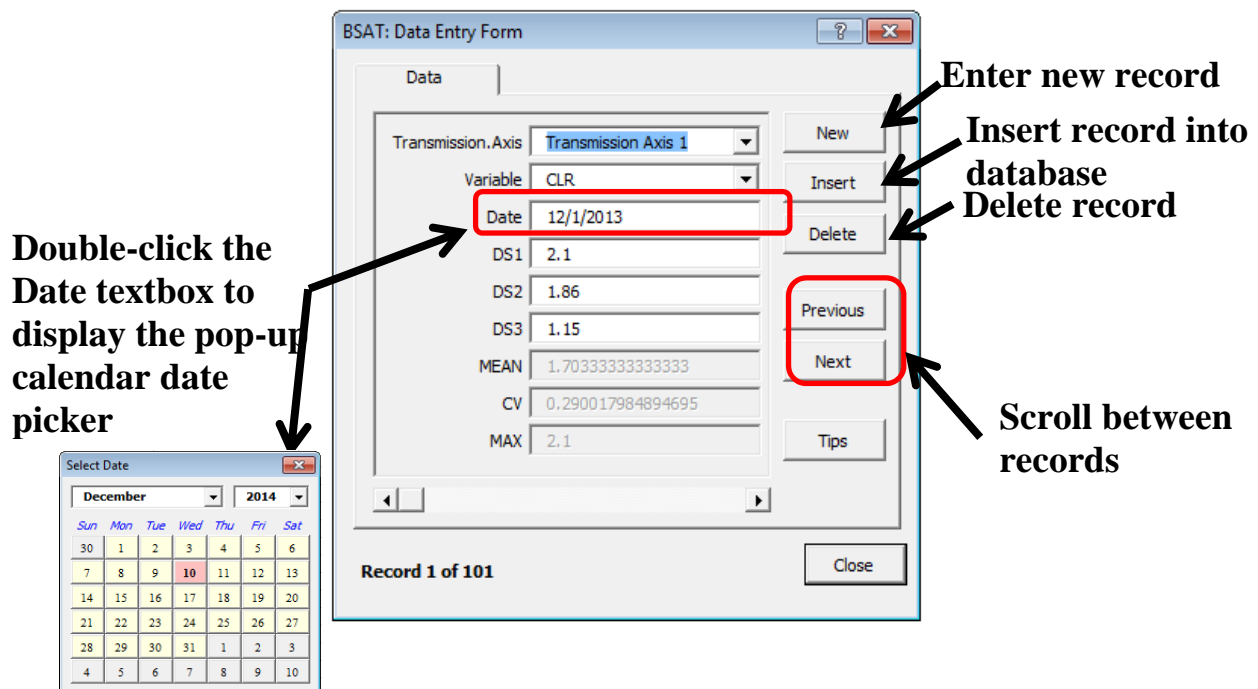
Figure 4.6 Data Entry Form dialog box

Direct Data Entry

Data may also be entered directly onto the Data Entry page (Figure 4.7) either by manually entering values into each cell, or by copying data from another Excel file and pasting it into the database cells under the blue headers. Copying/pasting may be advantageous if there are multiple records to be entered. The transmission axes and variable names must be entered exactly as defined in the *Utility Information* dialog box and as specified in Table 4.1. The Mean (MEAN), CV, and Maximum (MAX) values under the red headers are automatically calculated from the values entered in DS1, DS2, and DS3. Only one record should be entered for each date and transmission axis/variable. If multiple data values for the same transmission axis/variable are collected on the same day, the results should be averaged and then entered into the database. If there are any blank rows in the database they should be deleted before conducting subsequent data analysis. Double clicking a cell under the Date field will open the popup calendar date picker. The user can then select the date to be entered from the date picker. If a date is already present in the cell, the date picker will initially display that date; if the cell is empty or contains text that is not in a proper date format, the date picker will initially display the current (today's) date. Data can also be imported directly from an external Excel file, which is discussed further under the Data Options section below.

Double-click a cell in the Date column to display the pop up calendar date picker
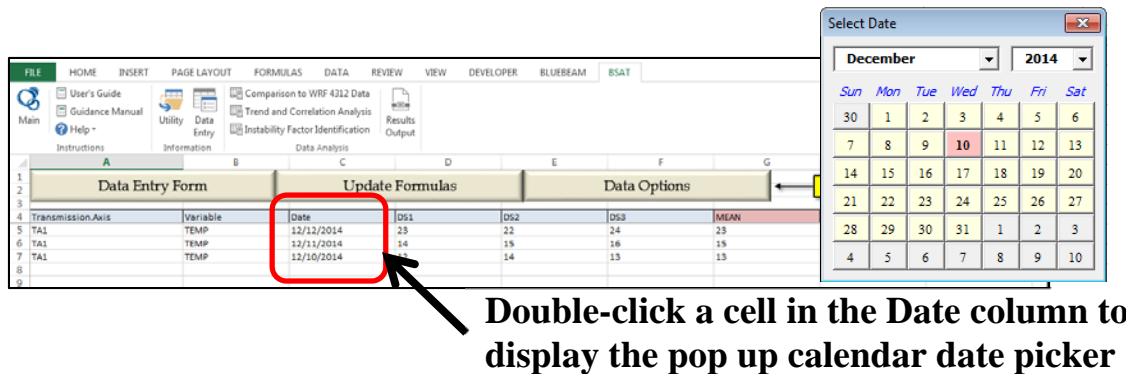
Figure 4.7 Example Data Entry page

Check that the data format is correct prior to data analysis including the variable name, date format, and units for monitoring data (see Table 4.1). If the data are not entered correctly, or if there are gaps in the database, the data may be ignored in the analyses or an error may occur.

Check that the data entered do not contain errors. Check that the values were not accidently entered incorrectly or contain negative numbers.

Update Formulas

Once data are imported, use the Update Formulas page button to insert the statistical calculations (MEAN, CV, and MAX) in the pink colored columns on the right (Figure 4.8). This page button can be clicked at any time to update and recalculate the formulas if changes in the dataset are made. The Update Formulas page button also contains data checking features to check

that values are entered in the correct format. The user should also review the dataset and check for negative or erroneous values that may be inadvertently been imported.

Figure 4.8 Update Formulas page button

Data Options

Clicking the <u>Data Options</u> page button opens a dialog box that can be used to load the example data, unload (e.g., delete) all data, import data from an external file, and run the date checker utility (Figure 4.9). These options are described in further detail below.
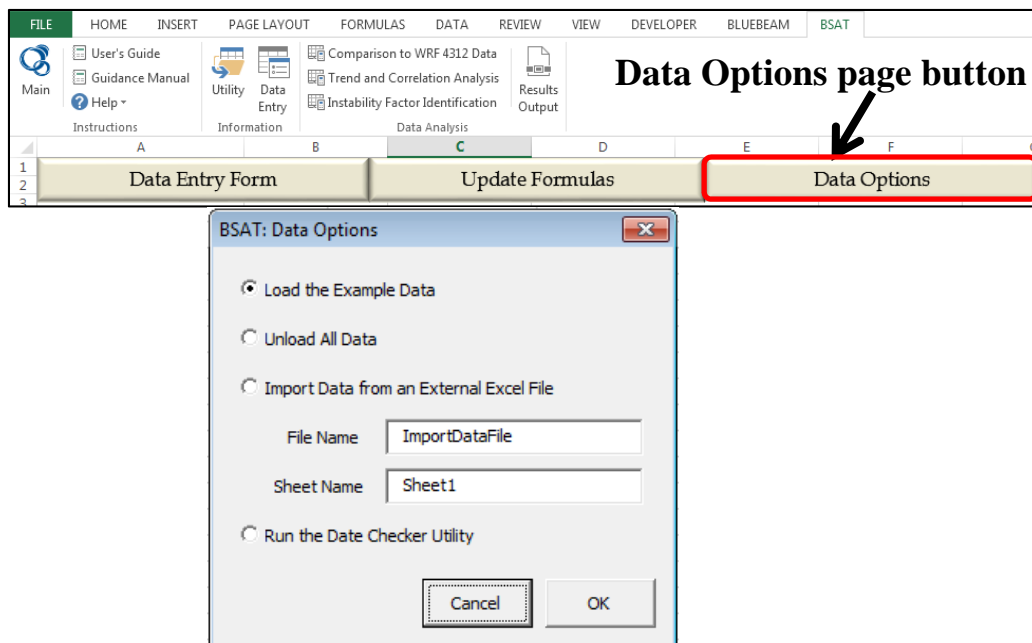
Figure 4.9 Data Options page button and dialog box

Load Example Data

Selecting Load the Example Data will load the example dataset onto the *Data Entry* page. The example dataset is an excerpt from one of the six utilities that participated in WRF 4312 and collected monitoring data over the duration of one year. This dataset is included to serve as an example of what the data input and analysis would look like. A discussion of this dataset and interpretations of the results are included in the Data Interpretation chapter of this User's Guide.

Unload All Data

Selecting Unload All Data will delete all data on the *Data Entry* page and create an empty database. Essentially, this clears the database and thus provides an empty database state for subsequent entry of new or different data.

Import Data from an External Excel File

Selecting Import Data from and External Excel File allows the user to automatically import data from an external Excel file. Enter the file name and sheet name for the external file. Note that the formatting rules discussed earlier must be followed, and the imported data must have the same format as the Data Entry page (e.g., columns for transmission axis, variable, date, DS1, DS2, and DS3). Selecting Import Data from an External Excel File on the Data Options dialog box allows the user to automatically import data stored in an Excel file to the *Data Entry* page. The external file must be located in the same directory (folder) as BSAT, and can be any of the following types: pre-Excel 2007 (".xls" extension), post-Excel 2007 (".xlsx") or a comma-delimited value (".csv") file. Enter the File Name (note: it is not necessary to include the extension in the file name) and the Sheet Name in the designated text boxes and click the <u>OK</u> button to initiate the import. If the file is a ".csv" file, the Sheet Name should be the same as the File Name. The format of the sheet in the external file should be as shown in Figure 4.10.

Figure 4.10 Example spreadsheet for importing data

Columns A:F beginning in Row 1 (i.e., cell A1, no headers) must contain the data that will be transferred to the *Data Entry* page. If cell A1 is empty then no data will be transferred. If a blank row exists in the dataset, only the data above the blank row will be included. Columns H:K beginning in Row 1 may (optionally) contain the list of Transmission Axes and data station (DS1, DS2 and DS3) names that are to be used to populate the **Transmission Axes** tab of the *Utility Information* dialog box. If cell H1 is empty then this information will not be included and the user will then need to manually define the **Transmission Axes** on the *Utility Information* dialog box. Column M (cells M1:M4) may (optionally) contain the utility/facility information used to populate the **General** tab of the *Utility Information* dialog box. If cell M1 is empty then this information will not be included and the user will then need to manually enter it on the *Utility Information* dialog box.

Run the Date Checker Utility

Selecting Run the Date Checker Utility will check the database for errors in date formatting. Use this application prior to performing data analysis. Selecting Run the Date Checker

Utility on the Data Options dialog box initiates a procedure designed to check the validity and formatting of each Date value in the Utility database. The procedure performs the following:

- Ensures that each Date value is a serial integer representing a single day (i.e., removes time decimals if present). This is important because BSAT analyses pair variables by Date. If two variables have the same date (day) but different times, they would not be paired in the analyses.
- Ensures that Date values entered as text are properly converted to Date values. This is also important because BSAT may ignore data for Dates entered as text.
- Identifies Date cells that cannot be modified to a correct date format and therefore need to be updated by the user.

BSAT will ignore data in cases where the Date value is not a correct date format. For example, if a Date cell contained a value such as "Yr 12 Qtr 2 Mo 1 Day 5" (or any text that cannot be evaluated as a date) the associated data record would be ignored in all subsequent analyses.

When the procedure identifies such a case, the cell is selected, and the user is provided with the option of either continuing to the next data record, or exiting to examine the selected Date value (Figure 4.11a). When all Date values have been checked, a confirmation message will be displayed (Figure 4.11b).
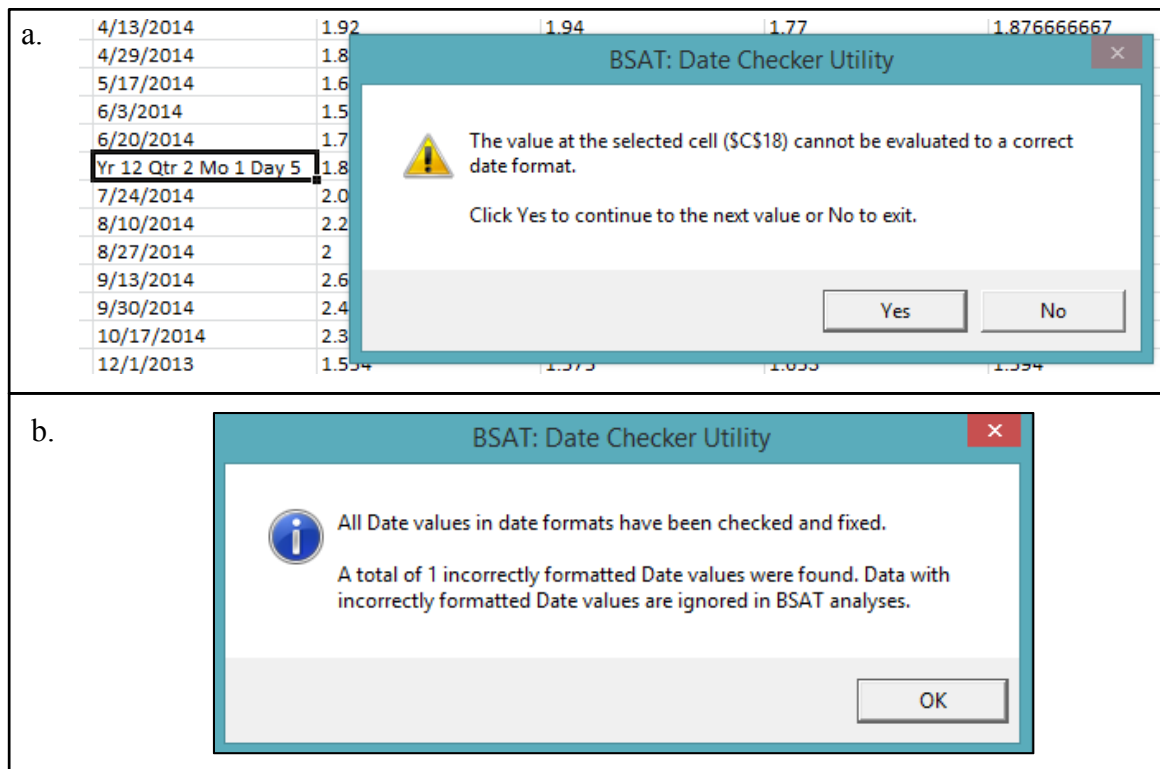
Figure 4.11 Data Checker Utility messages: invalid date identification (a) and confirmation and summary message (b).

CHAPTER 5
DATA ANALYSIS

The **Data Analysis** group on the *BSAT* tab of the ribbon is where the user can select various icons to initiate data analysis on their utility database, including comparing utility data to data collected in the WRF 4312 project, conducting trend and correlation analyses, and identifying factors that impact instability in their distribution system (Figure 5.1). This group also includes the *Results Output* icon for saving the analyses results to an external file.
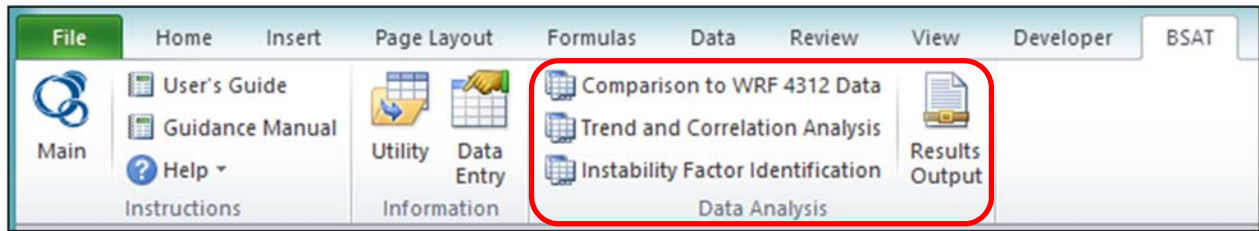
Figure 5.1 Data Analysis group on the BSAT tab of the ribbon

COMPARISON TO WRF 4312 DATABASE

This icon is selected for evaluating historical water quality data. The WRF Project 4312 data was collected from six distribution systems over a one-year duration. This is included to serve as an example for data analysis and interpretation. The overall goal of BSAT is to evaluate the stability of one's own distribution system. The included data are not intended to be used for comparison of what is "good" or "bad" but rather as an example. To access the interface, click the *Comparison to WRF 4312 Data* icon in the **Data Analysis** group on the *BSAT* tab of the ribbon. This will navigate the user to the analysis page. Click the Comparison to WRF 4312 Data button in the upper left hand corner of the page to open the dialog box for this analysis.

Data Tab

The *Data* tab on the *Comparison to WRF 4312 Data* dialog box (Figure 5.2) is used to select data for analysis. The transmission axis may be selected along with the date range. The date range can be entered directly or can be selected using the calendar date picker buttons. The default date range includes all of the dates in the dataset for the selected transmission axis. To select or restore the date range to the default settings click the Default button. Select the "Match Disinfection Type" option to only include data from WRF Project 4312 with the same disinfectant type as the user's utility. If the "Use All Data" option is selected, then data from all WRF Project 4312 systems will be used. Note that only those Transmission Axes defined on the *Utility Information* dialog box will appear on the list. If a Transmission Axis in your utility database does not appear on the list, it is because it has not been defined on the *Utility Information* dialog box. Conversely, if a Transmission Axis appears on the list for which you have no data, attempting to conduct an analysis on that Transmission Axis will either provide no results or will generate an

error message. Note also that the selections on the **Data** tab will be shown in lines two and three of the Analysis page headers (Figure 5.3).
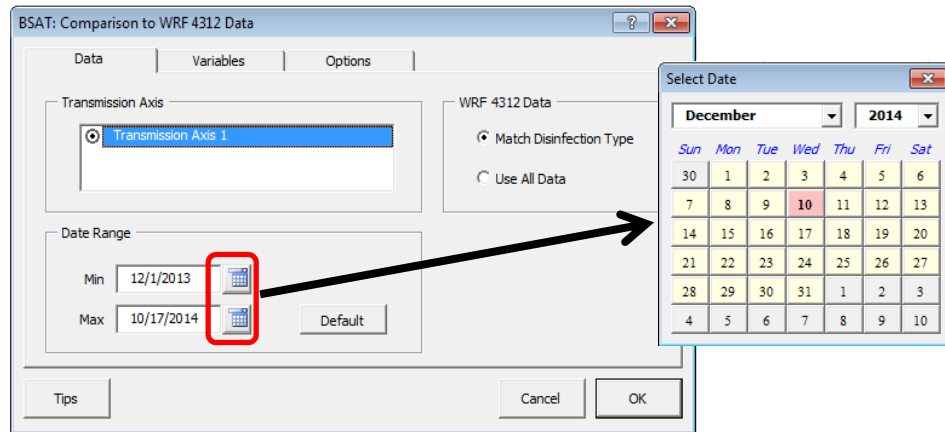
Figure 5.2 Data tab of the Comparison to WRF 4312 Data dialog box
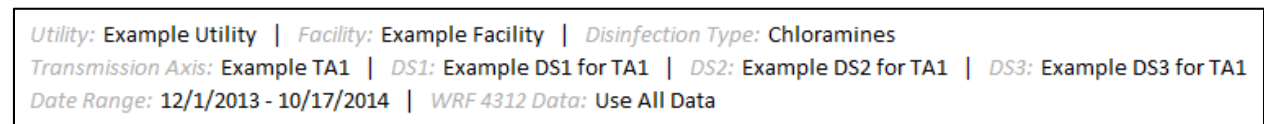
Figure 5.3 Analysis page headers

Variables Tab

The **Variables** tab of the *Comparison to WRF 4312 Data* dialog box (Figure 5.4) allows the user to select the variables to use for comparison to the WRF Project 4312 dataset. The user may select as many variables as desired. However, if no data are available from the user's utility dataset, then only the results from the WRF Project 4312 database will be displayed.
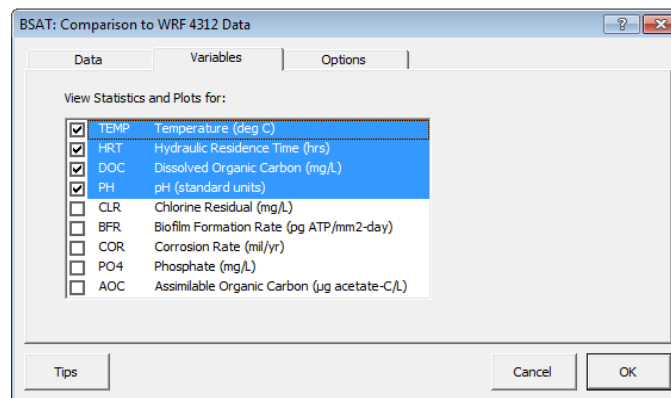
Figure 5.4 Variables tab of the Comparison to WRF 4312 Data dialog box

Options Tab

The summary statistics including Mean (MEAN), CV, and Maximum (MAX) values are calculated for the three distribution system monitoring locations (DS1, DS2 and DS3). The user may also view summary statistics and box plots for various locations (DS1, DS2 or DS3, or any combination thereof). These are specified under Locations on the ***Options tab*** (Figure 5.5). The Combinations option determines the variables to include for statistics and box plots in Part B, and scatter plots in Part C. For a description of Part B and Part C, see the Output Display section below.
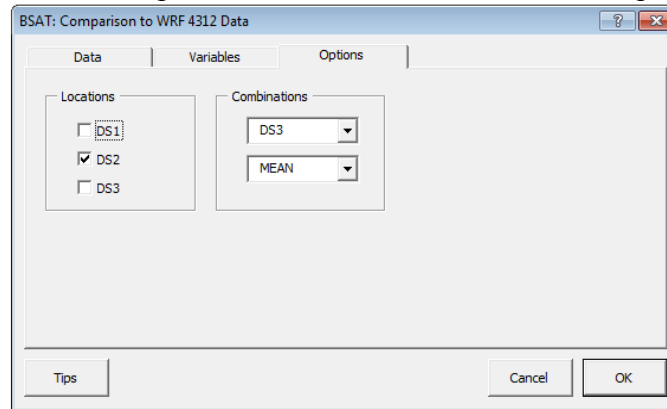
Figure 5.5 Options tab of the Comparison to WRF 4312 Data dialog box

Output Display

A. *Statistics and Box Plots*

Summary statistics and box plots are displayed for the variables that were selected in the *Comparison to WRF 4312 Data* dialog box. Summary statistics are provided in both tabular and graphical form as box and whisker plots. The WRF Project 4312 data are shown in the blue columns and blue box plots and the utility data are shown in the red columns and red box plots. An example box plot is shown in Figure 5.6. The upper whisker indicates the maximum value measured (MAX). The top of the box is the 3rd quartile (Q3) or 75th percentile (75% of the data are less than this value and 25% of the data are greater than this value). The dot in the box plot is the mean or average value (MEAN). The horizontal line through the box plot is the median (MEDIAN or Q2) or the mid-point of the dataset, meaning 50% of the data are greater than this value and 50% of the data are less than this value. The bottom of the box plot is the 1st quartile (Q1) or 75th percentile (25% of the data are below this value and 75% of the data are greater than this value). The lower whisker indicates the minimum value measured (MIN).
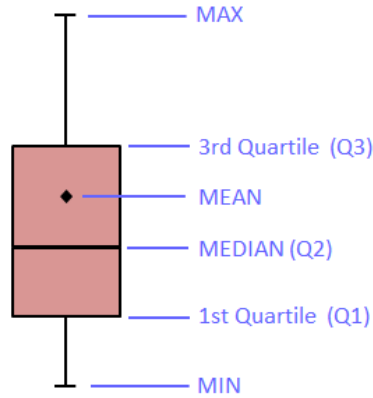
Figure 5.6 Example box and whisker plot

Summary statistics for the location(s) specified on the ***Options*** tab are shown on the far left (Figure 5.7). The next set of results to the right show the distribution of mean values of DS1, DS2, and DS3; followed to the right by the distribution of CV values of DS1, DS2, and DS3; and lastly by the distribution of maximum values of DS1, DS2, and DS3.

A. Statistics and Box Plots

							Temperature (deg C)				
	TEMP.DS2		TEMP.MEAN		TEMP.CV		TEMP.MAX	
	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility
N	49	21	49	21	49	21	49	21
MIN	5.4	7.6	5.6	9.467	3.636E-3	0.0E+0	9.1	12.2
MAX	31.1	28.8	30.633	28.9	0.608	0.3	32.2	31.9
MEDIAN (Q2)	18.5	25.5	20.067	26.	5.806E-2	4.334E-2	23.6	26.4
MEAN	20.569	22.167	20.704	22.849	0.123	6.537E-2	22.573	24.029
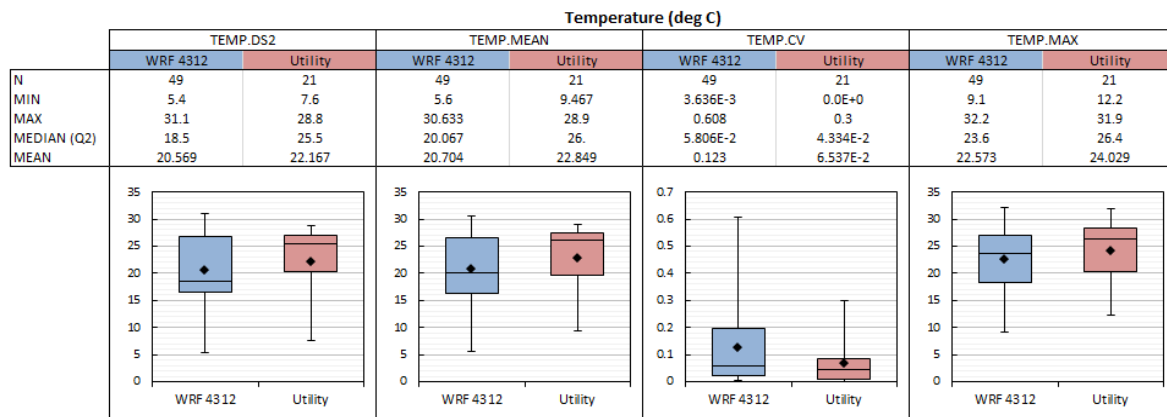
Figure 5.7 Example statistics and box plots output display

Note in the example above that the first set of results are specified as TEMP.DS2 because DS2 (only) was selected on the ***Options*** tab. The specification will vary depending on the combination of Locations selected. For example, if all three locations are selected, the specification would be TEMP.DS1/DS2/DS3 and the statistics would be calculated for all three locations combined; similarly, if two locations are selected, the specification would be, for example, TEMP.DS2/DS3 and the statistics would be calculated for these two locations combined.

B. *Statistics and Box Plots - Partitioned*

The partitioned statistics show the same data as in the Statistics and Box Plots, but partitioned by temperature and hydraulic residence time and specific to DS3. These values (temperature of 20ºC and hydraulic residence time of 80 hours) were found to be key partitions in

the WRF 4312 Project data, where biostability metrics changed significantly above or below these partitions (Figure 5.8).

Disinfectant Residual Stability - CLR.CV

	TEMP.DS3 >= 20		TEMP.DS3 < 20		HRT.DS3 <= 80		HRT.DS3 > 80		ALL DATA	
	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility
N	80	0	40	0	45	0	75	0	120	0
MIN	5.347E-2		4.019E-2		4.951E-2		4.019E-2		4.019E-2	
MAX	1.616		0.674		1.616		1.512		1.616	
MEDIAN (Q2)	0.365		0.308		0.214		0.358		0.323	
MEAN	0.457		0.281		0.273		0.474		0.399	

Figure 5.8 Example partitioned statistics and box plots output display

C. Scatter Plots

Scatter plots are included for disinfectant residual stability (chlorine residual CV versus temperature) and biofilm formation rate (chlorine residual CV versus biofilm formation rate, and biofilm formation rate versus chlorine residual). The location (DS1, DS2, or DS3) and the statistical value shown (MEAN or MAX) are selected on the **Options** tab of the *Comparison to WRF4312 Data* dialog box. An example scatter plot is shown in Figure 5.9. The red vertical line in Figure 5.9 is shown at a temperature of 20ºC. When temperature was greater than 20ºC, higher chlorine residual CV was observed in the WRF 4312 dataset.
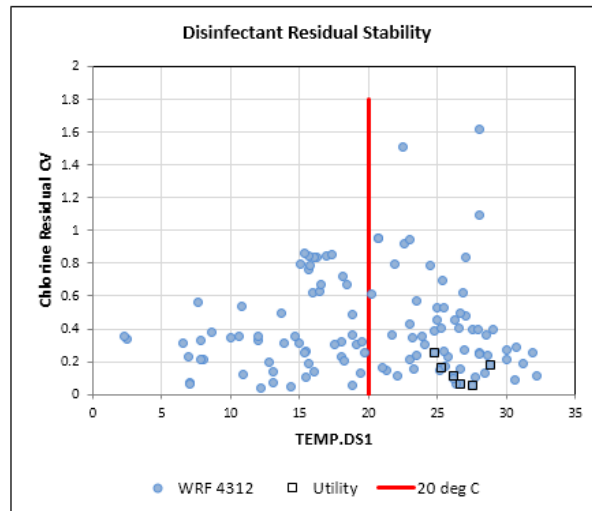
Figure 5.9 Example scatter plot

TREND AND CORRELATION ANALYSIS

This icon is selected for trend analysis and to evaluate correlations between monitoring parameters (variables) in the utility (only) database. The WRF 4312 dataset are not included in this analysis. To navigate to this analysis page, select the *Trend and Correlation Analysis* icon in the **Data Analysis** group on the ***BSAT*** tab of the ribbon. To open the dialog box for this analysis, click the <u>Trend and Correlation Analysis</u> page button in the upper left hand corner of the page.

Data Tab

The ***Data*** tab on the *Trend and Correlation Analysis* dialog box (Figure 5.10) is used to specify which data will be used for analysis. The transmission axis may be selected along with the date range. The date range can be entered directly or can be selected using the calendar date picker buttons. The default date range includes all of the dates in the dataset for the selected transmission axis.
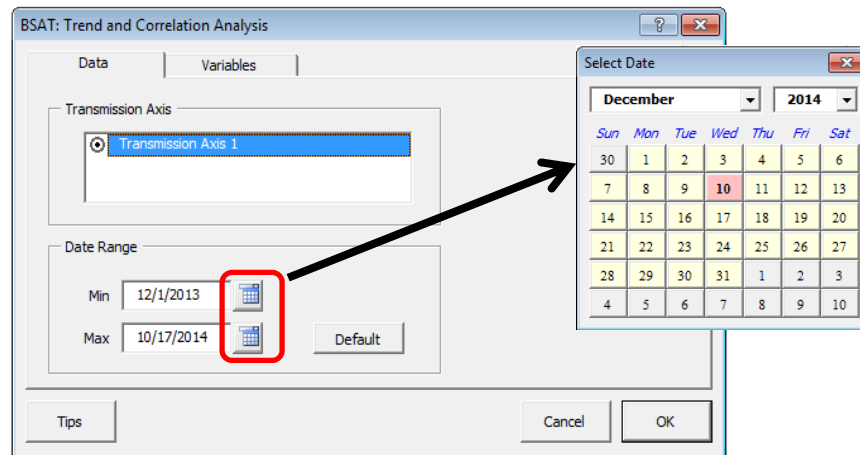
Figure 5.10 Data tab of the Trend and Correlation Analysis dialog box

Variables Tab

The ***Variables*** tab of the *Trend and Correlation Analysis* dialog box (Figure 5.11) is used to select the variables to use for the left and right axes on the time-series plot and for the y-axis and x-axis on the scatter plot (Y-axis/left and X-axis/right). The data to display may either be the original data from DS1, DS2, DS3, the average of these three locations at each time point (MEAN), the maximum values observed per time point (MAX), or the coefficient of variation for the three locations at each time point (CV). For example, if you wish to display the coefficient of variation (CV) for chlorine residual (CLR) as the Y-axis variable on the scatter plot and as the left-axis variable on the time series plot, select "CLR.CV" on the Y/L list (i.e., CLR with extension ".CV"). Variables such as the biofilm formation rate should be viewed on a log scale by the user to evaluate changes over time on an order of magnitude basis. This may be selected below the (Y-axis/left and X-axis/right) variable lists. The correlations of all X-variables with the Y-variable will be calculated when the Update Correlation List box is checked. This box is automatically checked whenever the Y/L variable is changed. Correlations include the Pearson correlation (r) and the

Rank correlation (correlation of the ranked data). The coefficient of determination (r-squared), slope, intercept, and count are also calculated for each variable pair. Further discussion on these statistics is provided in the Scatter Plots and Correlations section below.
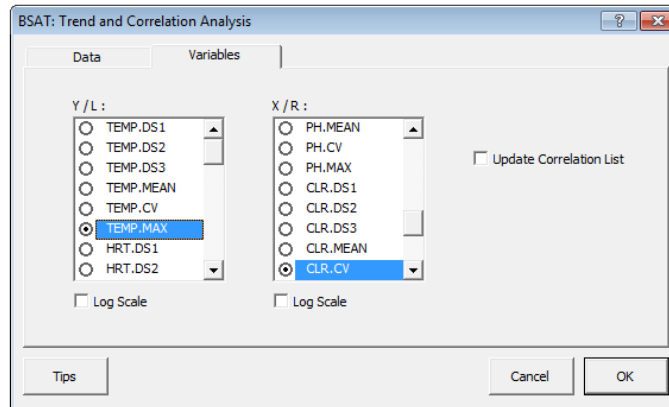
Figure 5.11 Variables tab of the Trend and Correlation Analysis dialog box

Output Display

A. Time Series Plot

This plot shows variations over time for the two selected variables (Figure 5.12). The variable on the left-axis is the selected "Y/L" variable on the *Variables* tab of the *Trend and Correlations* dialog box. The variable on the right-axis is the selected "X/R" variable on the *Variables* tab. These variables may be quickly changed by clicking the Trend and Correlations page button, navigating to the *Variables* tab (if necessary) and changing the variables to display. The date range for the display may also be modified via the *Data* tab.
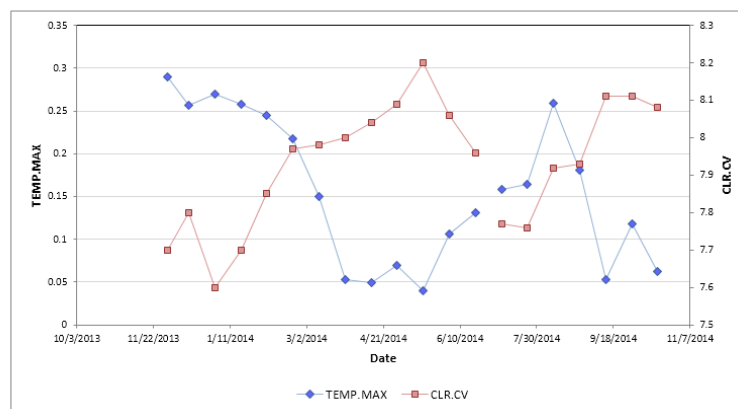
Figure 5.12 Example time series plot

B. Scatter Plot and Correlations

The scatter plot (Figure 5.13) shows the relationship between the two selected variables. These variables may be quickly changed by clicking the Trend and Correlations page button, navigating to the **Variables** tab (if necessary) and changing the variables to display ("Y/L" changes the y-axis variable, "X/R" changes the x-axis variable). A summary table of statistics for the x-y pair in the scatter plot is shown to the right. Statistics provided include the following:

Count. The count is the number of data points shown on the scatter plot (and on the time series plot for data paired by date).

Pearson r Correlation. The Pearson r Correlation is the correlation coefficient calculated for the x-y data pairs as plotted on the scatter plot. The Pearson r Correlation is a measure of the linear association between the two variables. Values of the Pearson r Correlation may range between -1 and 1, where negative values indicate a negative or inverse correlation (the best-fit line slopes downward from left to right) and positive values indicate a positive or direct correlation (the best-fit line slopes upward). The closer the values are to -1 or 1, the higher or more significant is the correlation (negative or positive) and the closer are the data points to the best-fit line. If the Pearson r Correlation value is exactly -1 or 1, the data points would plot exactly on the best-fit line with no scatter. If the Pearson r Correlation value is exactly zero, there is no association or relationship between the two variables at all and the best-fit line would be exactly horizontal with no slope.

Rank Correlation. The Rank Correlation is calculated by ranking the data from lowest to highest value (1, 2, 3, etc.) and then calculating the Pearson r correlation (see description above) on the ranks instead of on the actual values. The Rank correlation is provided solely for evaluation purposes and is NOT used for any other calculation purposes in the workbook. Typically, the Rank correlation and the Pearson r correlation values will be similar. However, if the plotted data are not evenly distributed along the best-fit regression line, or the residuals (differences between the data and the best-fit line) are not normally distributed, then the two correlation values may differ significantly. Hence comparison between the Rank correlation and Pearson r correlation provides a quick and simple way of assessing whether the assumptions (evenly distributed data and normality of residuals) of the Pearson r correlation and thus the best-fit regression line are met.

r-squared (Coefficient of Determination). The r-squared value is simply the Pearson r Correlation coefficient squared. Thus the r-squared value may range between 0 and 1. The higher the r-squared value (closer to 1), the higher the percentage of the total variability (or variance) of the plotted data is explained or accounted for by the linear association or best-fit regression line. For example, an r-squared of 0.748 would mean that approximately 75% of the total variation is accounted for by the linear association, with the remaining 25% due to random scatter. Hence an r-squared value of exactly 1 indicates that all of the variation is accounted for by the linear association, and thus the data points would plot exactly on the best-fit linear regression line with no scatter.

Intercept. The intercept is the lowest possible value that the response variable can be based on the dataset provided. The intercept is the y-value corresponding to an x-value of zero for the best-fit line. In other words, it is the y-value where the best-fit line intersects the y-axis at an x-value of zero. In the equation for the best-fit linear line, $y = a + bx$, parameter a is the intercept.

Slope. The slope of the line provides information on how strongly the explanatory variable impacts the response variable, and whether they are directly or inversely related. The slope is the change in y with x for the best-fit linear line. A positive value indicates the line is sloping upward

from left to right, whereas a negative value indicates the line is sloping downward. In the equation for the best-fit linear line, $y = a + bx$, parameter b is the slope.

The lower table shows the count, Pearson correlation and Rank correlation for the y-variable versus all other x-variables in the dataset (Figure 5.13). The data can be sorted to show the strongest Pearson or Rank correlation (closer to 1 or -1). The Pearson and Rank correlations can be sorted as desired using the filter function; this feature is strictly for ease of evaluating all correlations at a glance. The correlation values are automatically color-coded: Green represents positive or direct correlations and Red represents negative or inverse correlations; the darker the color (green or red), the higher or stronger the correlation (positive or negative).
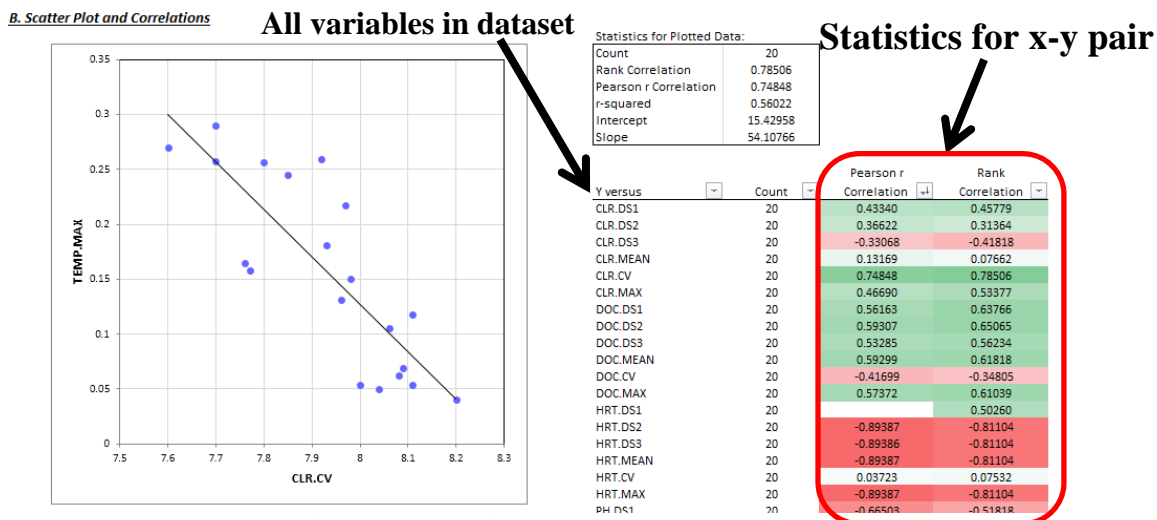
Figure 5.13 Example correlations output display

INSTABILITY FACTOR IDENTIFICATION

This icon is selected to identify factors affecting biostability using regression tree statistical analysis. Regression tree analysis involves selection of an explanatory variable value that splits the response variable data set into two subsets. The explanatory variable value is selected that results in the greatest separation between the two data subsets. For example, an explanatory variable could be temperature and a response variable could be biofilm formation rate. As shown in the accompanying Report and Guidance Manual (LeChevallier 2015a and 2015b), the project demonstrated that higher temperatures increase biological activity and increase biofilm formation rates. The tree analysis finds the value for temperature that results in the greatest separation between high and low values. This is the split point that causes the biggest change in the response variable value (in this case biofilm formation rate). To begin an analysis, select the *Instability Factor Identification* icon in the **Data Analysis** group on the *BSAT* tab of the ribbon to navigate to the page. Click the Instability Factor Identification page button in the upper left corner of the page to open the dialog box for this analysis.

Data Tab

The **Data** tab on the *Instability Factor Identification* dialog box (Figure 5.14) is used to specify which data will be used for analysis. The transmission axis may be selected along with the

date range. The date range can be entered directly or can be selected using the calendar date picker buttons. The default date range includes all of the dates in the dataset for the selected transmission axis.

.
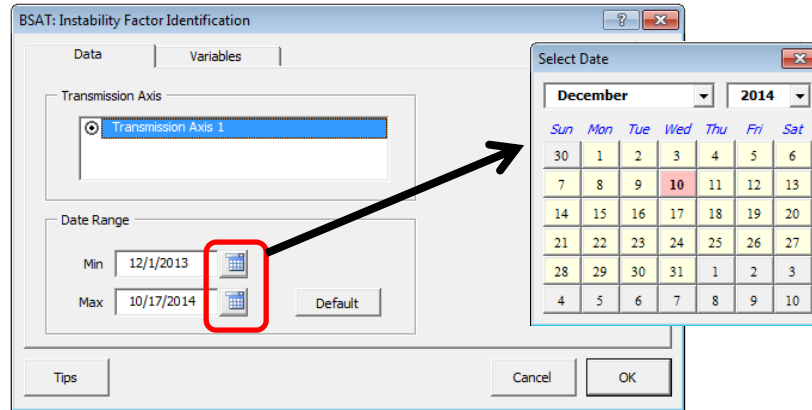
Figure 5.14 Data tab of Instability Factor Identification dialog box

Variables Tab

The ***Variables*** tab on the *Instability Factor Identification* dialog box is used to select the variables to be included in the instability factor identification analysis. The user must select one Biological Stability Issue (the response variable) and one or more explanatory variables (Table 5.1). The default Biological Stability Issue is chlorine residual stability (CLR.CV is the response variable). Other Biological Stability Issues may be analyzed if the data are available for the corresponding response variables (Figure 5.15). Additional explanatory variables may also be selected for inclusion in the analysis, again if the data are available. The explanatory variable extension can be changed by clicking the variable name and selecting the extension from the drop-down list that appears to the right. Here the user can select to analyze data for extensions DS1, DS2, or DS3, or the MEAN, CV, or MAX values for DS1, DS2, and DS3. Each record in the utility database has a DS1, DS2, DS3, MEAN, CV and MAX value. Records corresponding to Response and Explanatory variables are paired by date (same as for Trend and Correlation Analysis). The analysis is then conducted on the date-paired data. Clicking the Default button resets the explanatory variables to their default extensions as shown in Table 5.1.

Table 5.1

Default variable settings (extensions) for Instability Factor Identification analysis

Variable.Extension	Description
TEMP.MEAN	Average temperature for DS1, DS2, and DS3
HRT.MAX	Maximum HRT for DS1, DS2, and DS3
DOC.MEAN	Average DOC for DS1, DS2, and DS3
PH.MEAN	Average pH for DS1, DS2, and DS3
CLR.DS1	Chlorine Residual for DS1
BFR.MEAN	Average Biofilm Formation Rate for DS1, DS2, and DS3
COR.MEAN	Average corrosion Rate for DS1, DS2, and DS3
PO4.MEAN	Average phosphate for DS1, DS2, and DS3
AOC.MEAN	Average AOC for DS1, DS2, and DS3
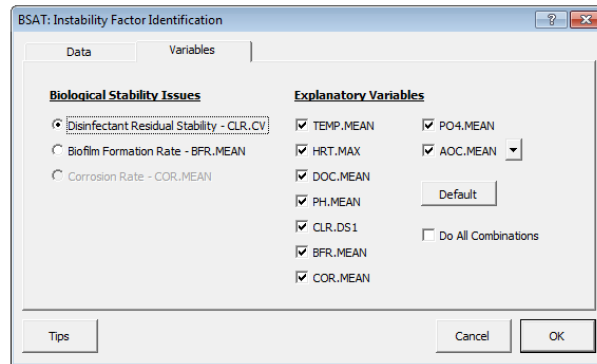
Figure 5.15 Variables tab of the Trend and Correlation Analysis dialog box

If the "Do All Combinations" box is checked, results will be displayed for all selected explanatory variables and all extensions (DS1, DS2, DS3, MEAN, CV, and MAX). However, only those selected explanatory variables with corresponding data in the utility database will be shown. Selected explanatory variables (and extensions) are listed in the page table. The top nine explanatory variables in the table list are displayed, in order, on the regression tree diagram. Sorting or filtering the table (or otherwise changing the order of the table list) will therefore change the corresponding order on the regression tree plot. Ideally, one would endeavor to explore the explanatory variables with the goal of identifying and selecting a sorted subset for display in the table and on the regression tree diagram, such that ultimately the subset contains, lists, and displays in order the most imported explanatory variables for the selected stability issue (Response Variable) of concern.

Click the "Show Table Tips" box to show additional useful tips directly on the output page. The number of factors that should be included will initially be all parameters for which there are historical data. Subsequent data analysis may indicate that only a subset of these parameters are key to a particular distribution system.

Output Display

A. Single-Split Tree Analysis

   The output displays the regression tree analysis results for a series of single split, or single node, trees (Figure 5.16). The variables in blue on the left are the explanatory variables and the variable at the top in red is the response variable. The split points for each explanatory variable are shown above and below the tree branch split. The average values for the response variable above and below the split point are shown in the boxes to the right. The order of these variables can be changed by sorting the values in the table, for example, based on the absolute difference of average values above and below the split point. The display can be sorted by clicking the dropdown arrows on the table headers. This allows the user to determine the most applicable metric and sort the tree diagram as desired. The absolute difference between the average values above and below the split point is particularly useful for determining which explanatory variable split points have the greatest effect on the response variable (the stability issue of concern). Some variables may have a large impact and others may result in only a minor change. The variables to include in subsequent analyses should include those variables with the greatest absolute difference. Note that the values in the last three columns of the table to the right are color-coded for ease in evaluating their significance at a glance. The tree diagram is linked directly to this tabulation – the ordering of the explanatory variables in the tree diagram (top to bottom) will change depending on how the table is sorted; the tree diagram is limited to displaying a maximum of nine explanatory variables. Note that a minimum number of data points are required on each branch of the tree in order for the data to be analyzed: 5 if the number of matched Response/Explanatory Variable data pairs is greater than or equal to 17; 4 if less than 17 and greater than or equal to 13; 3 if less than 13 and greater than or equal to 9; and 2 if less than 9. If an insufficient number of matched Response/Explanatory Variable pairs are available in the utility database to meet any of these criteria, then results will not be displayed for that Explanatory Variable. For further details regarding the tree analysis method, please refer to the accompanying research report (Project 4312).
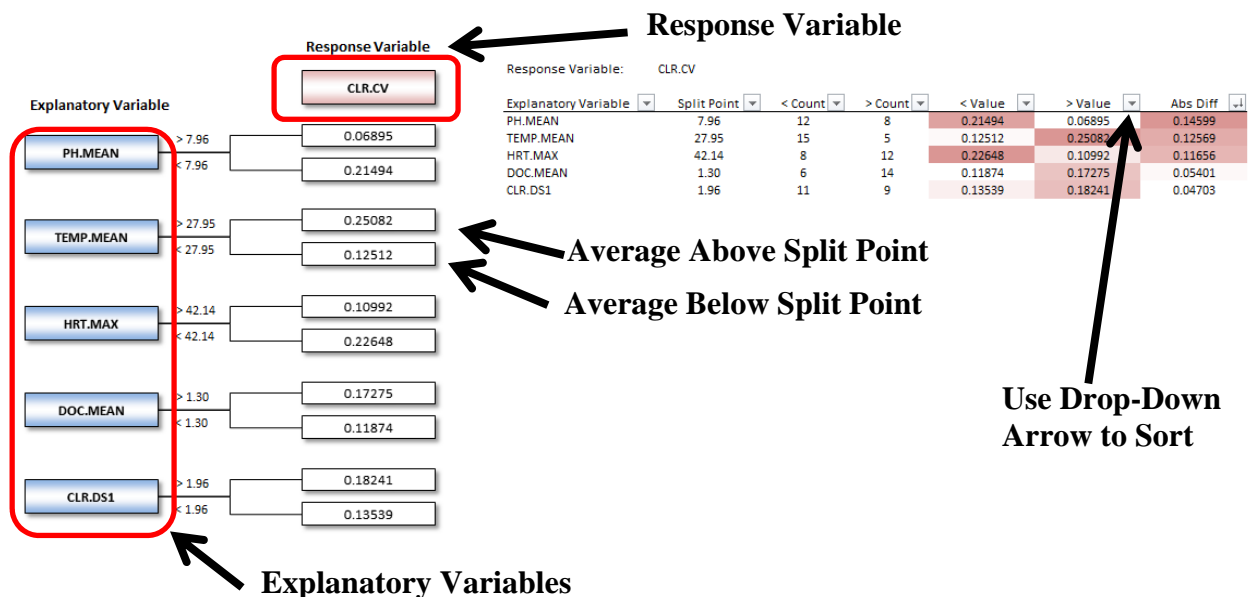
Figure 5.16 Example Instability Factor Analysis output

RESULTS OUTPUT

The data analysis results can be exported (saved) to a separate file. To do so, select the *Results Output* icon in the **Data Analysis** group on the ***BSAT*** tab of the ribbon. The *Results Output* dialog box will open and the user may select the specific analyses to save (Figure 5.17). Selecting the <u>Tips</u> button will display tips on how to save the file. A file name must be specified and at least one data analysis result page must be selected. The results output file is saved as a portable document file (pdf) to the same folder where the BSAT program file is located.
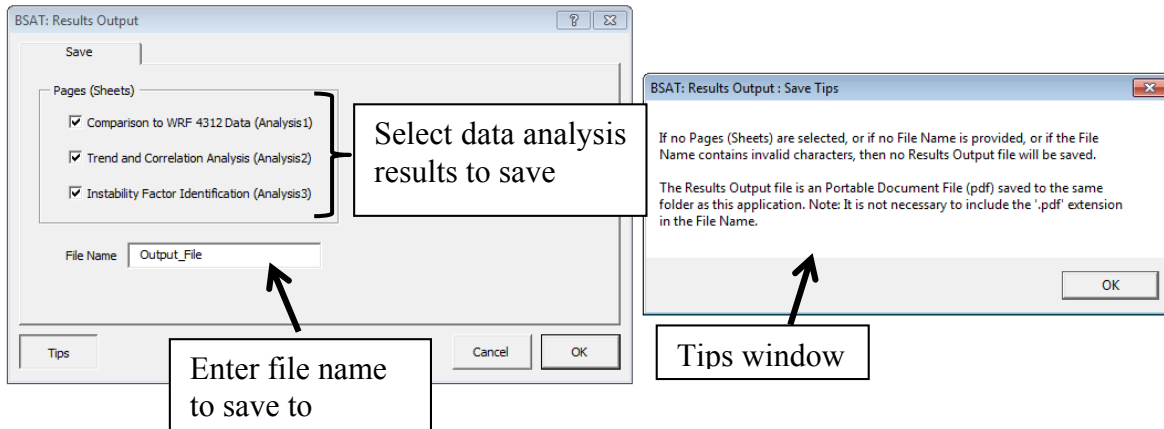
Figure 5.17 Results Output dialog box

CHAPTER 6
DATA ANALYSIS AND RESULTS INTERPRETATION

The BSAT is intended for use by utility managers, engineers, and scientists for planning purposes using data collected over the period of seasons to years. The tool is not intended for use by operators on a day-to-day basis. Utilities are ultimately responsible for determining applicability of data and BSAT output based their own site-specific water quality goals.

This section is intended to provide guidance on how the data should be interpreted and can be integrated into a utility's overall distribution system management program. BSAT is intended for managers, scientists, and engineers who have first read the accompanying WRF Project 4312 report *An Operational Definition of Biological Stability* and the *Guidance Manual for Control of Biostability in Drinking Water* (LeChevallier 2015a and 2015b). The report provides the scientific basis for the recommendations outlined in the Guidance Manual. In addition, utilities should collect and evaluate their own data to best take advantage of the BSAT data analysis capabilities. This excel-based macro tool is intended to complement these reports and allow utilities to evaluate their data on a site-specific basis. Many variables can impact distribution system water stability and there is no "one-size-fits-all" parameter that will be applicable to all distribution systems. However, the methodology outlined in the research report and guidance manual and included in BSAT will enable utilities to have more insight on factors controlling stability in their system. The BSAT also provides utilities with a systematic approach to evaluate distribution system stability that can be continuously updated as new data are collected. In this way, the utility can clearly evaluate whether specific actions taken to improve stability have the desired positive effects.

EXAMPLE UTILITY'S DATA SET

An example dataset is included in BSAT and may be loaded by clicking the Data Options button on the *Data Entry* page and selecting Load the Example Data (see Figure 4.9 and associated discussion). These data were included to use as an example of how the data can be analyzed and displayed in BSAT. The data for the example transmission axis 1 (TA1) were selected for the analysis and interpretation herein.

Comparison to WRF 4312 Dataset

The example dataset includes chlorine residual, DOC, hydraulic residence time (HRT), pH, and temperature. Statistical results for temperature and chlorine residual are shown in Figure 6.1. The number of data points (N), minimum (MIN), maximum (MAX), median [MEDIAN (Q2)], and mean (MEAN) values are summarized for all of the chlorine residual (CLR) data (DS1/DS2/DS3); the mean between DS1, DS2, and DS3 (CLR.MEAN); the coefficient of variation (CV) between DS1, DS2, and DS3 (CLR.CV); and the maximum values between DS1, DS2, and DS3 (CLR.MAX). The minimum and maximum values will bracket the range of data observed. In this case, the chlorine residual measured from DS1, DS2, and DS3 (CLR.DS1/DS2/DS3) ranged from 1.09 to 2.67 mg/L. This immediately identifies that the chlorine residual is relatively stable (compared with the much wider range in the WRF 4312 database). The MIN and MAX values can be a useful indicator of whether chlorine residuals are meeting target concentrations. The chlorine residual CV (CLR.CV) data provides information on the stability

through the distribution system (from DS1 at the finished water to DS3, which has a longer HRT). The lower the value for CV, the more stable the system. The CV is calculated by taking the average of three disinfectant residual measurements on a given day in the distribution system and then dividing the average by the standard deviation of the data. By mathematical definition, the values for CV can range from a minimum of 0 to a maximum of the square root of the number of sample locations ($\sqrt{n} = \sqrt{3} = 1.73$). In this case, the highest value for chlorine residual CV was 0.29 and the mean was 0.157. Both the maximum measured CV and the mean CV were very low relative to the highest possible value for CV of 1.73. A more important statistic for CLR.CV, however, is its range or variation. Here we see that the variation is relatively small, ranging between a minimum of 0.04 and a maximum of 0.29, relative to the WRF 4312 dataset which exhibits a much greater variation, ranging between 0.04 and 1.62, thus indicating that chlorine residual is relatively stable for this transmission axis overall. Acceptable values for CV will vary on a site-specific basis as well as the number of data points collected, but in general, lower relative values indicate greater stability. Ultimately, the utility needs to determine what value of CV is acceptable.

|	Chlorine Residual (mg/L)							
	CLR.DS1/DS2/DS3		CLR.MEAN		CLR.CV		CLR.MAX	
	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility
N	360	60	120	20	120	20	120	20
MIN	1.0E-2	1.09	0.44	1.503	4.019E-2	4.019E-2	0.6	1.73
MAX	4.1	2.67	3.45	2.53	1.616	0.29	4.1	2.67
MEDIAN (Q2)	1.505	1.865	1.603	1.772	0.323	0.154	1.98	1.965
MEAN	1.62	1.825	1.62	1.825	0.399	0.157	2.137	2.039
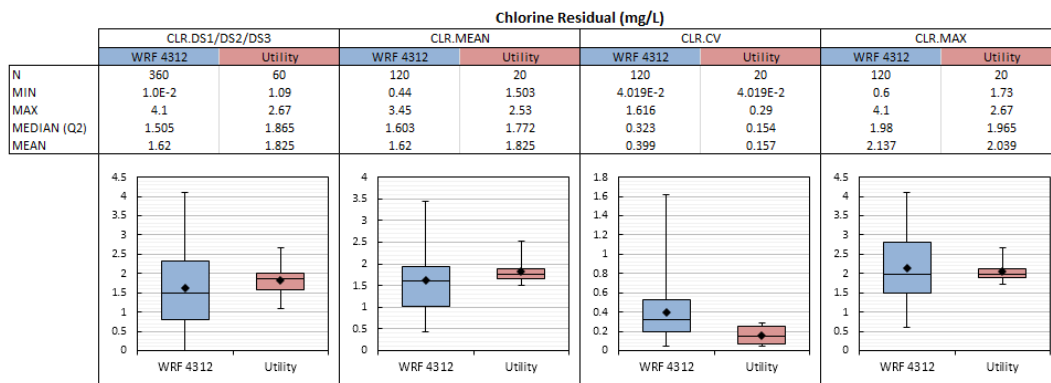
Figure 6.1 Summary statistics and box plots for chlorine residual from example dataset

WRF Project 4312 identified that temperature and HRT were key parameters that affected chlorine residual stability. When temperatures were above 20 degrees Celsius (ºC) chlorine residual was statistically less stable compared to when temperatures were below 20ºC. A similar benchmark was identified for HRT, where HRTs greater than 80 hours exhibited significantly less stable chlorine residual. When chlorine residual is less stable, chlorine residual concentrations may drop below the minimum values needed to control biofilm growth. The *Guidance Manual for Control of Biostability in Drinking Water* recommends a minimum of 0.8 to 1.5 mg/L of free chlorine or 1 to 2.1 mg/L of chloramine to attain a 2-log reduction in the biofilm formation rate. Figure 6.2 shows summary statistics for disinfectant residual stability CV (CLR.CV) where the data has been partitioned by temperature (greater and less than 20ºC) and HRT (above and below 80 hours). This data summary is intended to be a first cut evaluation of whether temperature and HRT are important parameters based on the site-specific data provided by the utility. The WRF 4312 project results are included to aid in interpretation of the data. Based on the example utility dataset, the mean CLR.CV when temperature was greater than 20ºC was slightly higher than when it was less than 20ºC (0.185 vs. 0.103), indicating this system is relatively stable as a function of the 20 ºC temperature split point. As an example, let's say the example utility independently determined that a CLR.CV of 0.15 was acceptable. The data just presented could be used to conclude that instability problems would be observed only when the temperature was greater than

20 ºC because that is when CLR.CV > 0.15. While the utility cannot control the temperature, knowledge of the temperature above which problem can be expected to occur is useful and facilitate planning. The maximum values for CLR.CV are only slightly lower below 20 ºC compared to above 20 ºC (0.269 vs 0.290). There were no data points with an HRT greater than 80 hours, so the effect of HRT on chlorine residual stability is not possible here for the 80 hour split point; however, HRT at a different split point could still prove to be an important stability-differentiating factor. These questions can be explored and evaluated further using *Trend and Correlation Analysis* and *Instability Factor Identification* analysis (as discussed further below).

Disinfectant Residual Stability - CLR.CV

	TEMP.DS1 >= 20		TEMP.DS1 < 20		HRT.DS1 <= 80		HRT.DS1 > 80		ALL DATA	
	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility	WRF 4312	Utility
N	64	13	56	7	120	20	0	0	120	20
MIN	5.347E-2	5.347E-2	4.019E-2	4.019E-2	4.019E-2	4.019E-2			4.019E-2	4.019E-2
MAX	1.616	0.29	0.858	0.269	1.616	0.29			1.616	0.29
MEDIAN (Q2)	0.299	0.181	0.326	6.939E-2	0.323	0.154			0.323	0.154
MEAN	0.407	0.185	0.389	0.103	0.399	0.157			0.399	0.157
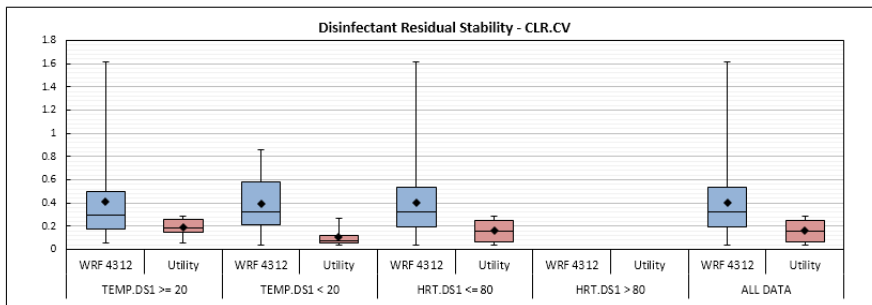
Figure 6.2 Summary statistics and box plots for data partitioned for temperature and hydraulic residence time (HRT) for example dataset with key partitions identified in the WRF Project 4312

Scatter plots for disinfectant residual stability and biofilm formation rate are included because finished water temperature (TEMP.DS1) and chlorine residual (CLR.DS1) were found to correlate with the disinfectant residual stability (CLR.CV) and biofilm formation rate (BFR.MEAN) response variables in the WRF 4312 study. Data from the WRF 4312 database are plotted as blue-filled circle symbols and example utility data are plotted as black square outlined symbols (Figure 6.3). Figure 6.3 indicates that while the chlorine residual stability (CLR.CV) from the example dataset increases with increasing finished water temperature (TEMP.DS1), it nevertheless remains relatively stable (low overall CLR.CV values) even at higher temperatures. There are two other plots for BFR.MEAN vs CLR.CV and the chlorine residual at the longest water age (CLR.DS3) vs BFR.MEAN. The example utility dataset does not include BFR.MEAN data so these are not discussed in this section. However, if biofilm formation rate data are collected, the BSAT yield more a more thorough assessment of the distribution system and information on how disinfectant residual stability and concentration impact the biofilm formation rate. As an example, let's again assume the utility independently determined that CLR.CV of 0.15 was acceptable. If the utility desires to decrease CLR.CV to less than 0.15 when temperatures are greater than 20 ºC (see Figure 6.3), then measurement of the biofilm formation rate at specific locations in the distribution system will help the utility to identify where biofilm growth is leading to instability problems. The utility is then in a better position to develop potential mitigation strategies. Similarly, if corrosion is thought to be the cause of elevated CLR.CV, then direct measurement of corrosion rate can be used to develop potential mitigation strategies.
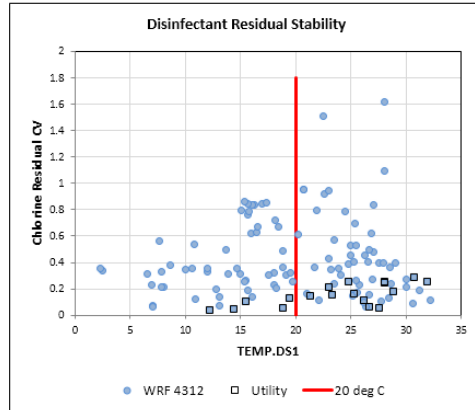
Figure 6.3 Disinfectant residual stability (CLR.CV) as a function of temperature of the finished water (DS1)

Trend and Correlation Analysis

Time series plots of two variables can be displayed to look for seasonal variations and relationships. Process changes or upsets that occur at particular time points may assist in explaining variations in the data. In the example database, finished water temperature (TEMP.DS1) and chlorine residual variation (CLR.CV) appear to be positively correlated and are influenced by seasonal variations (Figure 6.4).
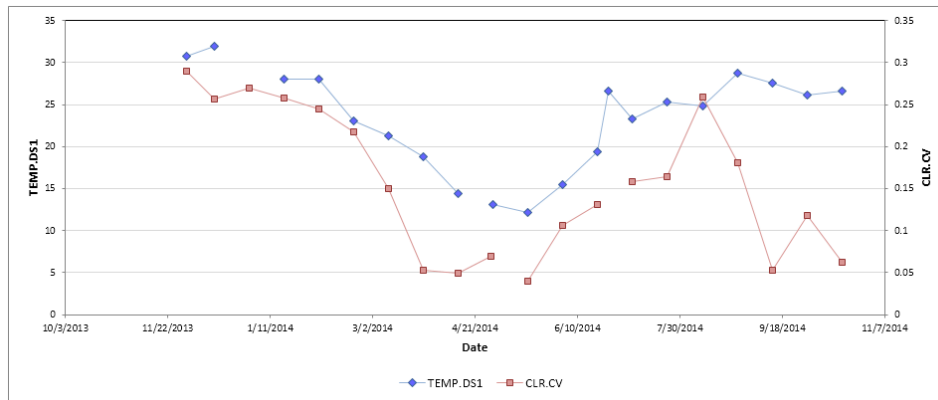
Figure 6.4 Time-series plot of chlorine residual CV (CLR.CV) and finished water temperature (TEMP.DS1)

The best-fit correlation line between CLR.CV and finished water temperature (TEMP.DS1) are provided in the scatter plot below the time series plot. The statistics for this correlation are shown in the upper right hand box of Figure 6.5 titled Statistics for Plotted Data. In this case there are 18 paired data points and the Pearson correlation (r) is 0.66. Generally, data with an r value of greater than 0.5 or less than -0.5 are considered to be moderately correlated, and data with an r value greater than 0.8 or less than -0.8 are considered to be strongly correlated. The closer the r value is to 1, the more directly correlated are the two variables. The closer the r value is to -1, the more inversely correlated are the two variables. The summary table below the Statistics

for Plotted Data show all of the Pearson r and Rank correlations for the y value (in this case CLR.CV) and all of the possible x values (all other variable combinations). Since CLR.CV is the response variable we are interested in evaluating, it should be listed on the y-axis (select from the Y/L dropdown list on the *Variables* tab of the <u>Trend and Correlations</u> page button). The Pearson and Rank correlations are color coded to identify the strength of the correlation, with darker red values indicating stronger inverse correlations, and darker green values indicating stronger direct correlations. These values can be sorted from highest to lowest to quickly identify which variables are most closely related. In this example, CLR.CV exhibits a moderate direct correlation with TEMP.DS1 (Pearson r = 0.66), i.e., higher temperature is generally associated with higher chlorine residual CV) and TEMP.MAX (the greatest temperature in the distribution system at any given time point) exhibits a stronger correlation (r = 0.73). These values allow for fast screening of multiple combinations of variables without having to view each plot individually. When strongly correlated variables are identified, the data may then be plotted for further evaluation of time series trends and correlations. In our example, the utility can use the scatter plot to more precisely identify the temperature associated with a specific CLR.CV (e.g., 0.15). The data in Figure 6.5 indicate the associated temperature is 23 °C which is slightly greater than the split point of 20 °C. A few degrees difference may have a significant cost impact when a utility must determine when to implement a specific instability mitigation strategy (e.g., additional chlorine dosing).
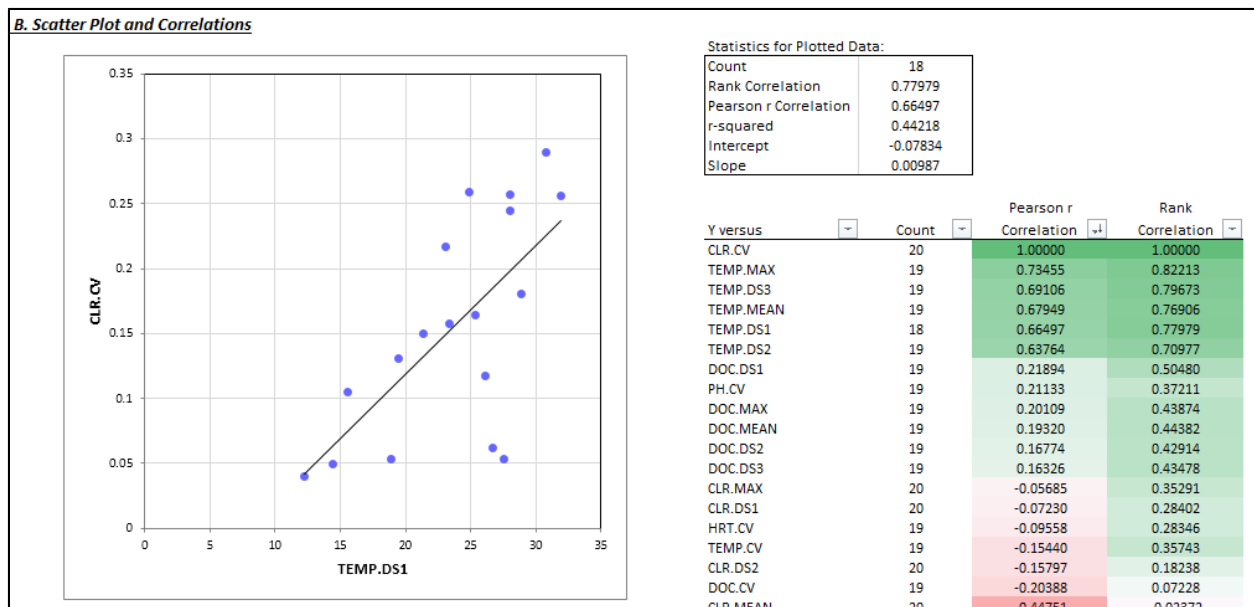

Figure 6.5 Correlations between chlorine residual CV (CLR.CV) and finished water temperature (TEMP.DS1); all possible correlations with TEMP.DS1 displayed and sorted by Pearson r correlation

Instability Factor Identification

Instability Factor Identification allows the utility to further explore the data and determine which explanatory variables have the greatest impact on the values of the response variable. In this case, Figure 6.6 shows an analysis of the chlorine residual stability (CLR.CV) response variable. The table on the right hand side is sorted by the absolute difference between values above and below the split point to show the greatest difference at the top. PH.MEAN exhibits the greatest difference in chlorine residual stability (CLR.CV) at the pH split point of 7.96 SU. The average chlorine residual CV when pH is above 7.96 SU is 0.069, compared to an average of 0.215 when pH is below 7.96 SU. Both chlorine residual CVs (either above or below the split point) are relatively stable in this case because they are much lower than the maximum possible CV of 1.73. If chlorine residual stability was a concern, then a recommendation from this analysis would be to increase the pH to above 7.96 SU if possible. For example, if the example utility again uses a CLR.CV < 0.15 as their acceptability criterion, then they may increase the pH when the CLR.CV > 0.15. The remaining variables can also be evaluated to determine which create the greatest change in CLR.CV. The table on the right in Figure 6.6 can also be sorted by values greater or less than the split point. In this case, the greatest average chlorine CV (under the >Value column in Figure 6.6) was when temperature was above 27.95°C – the average CV was 0.251. Temperature cannot be modified, but knowing what temperatures most impact stability can be important for managing other factors that work synergistically to dissipate chlorine residual. For example, the utility may want to increase the pH to over 7.97 when the temperature is above 27.95 °C because both factors likely affect chlorine residual stability based on the example data. It is recommended that all explanatory variables with data be used in this analysis, with particular attention given to those variables identified as important (highest absolute correlation) from the trend and correlation analysis.
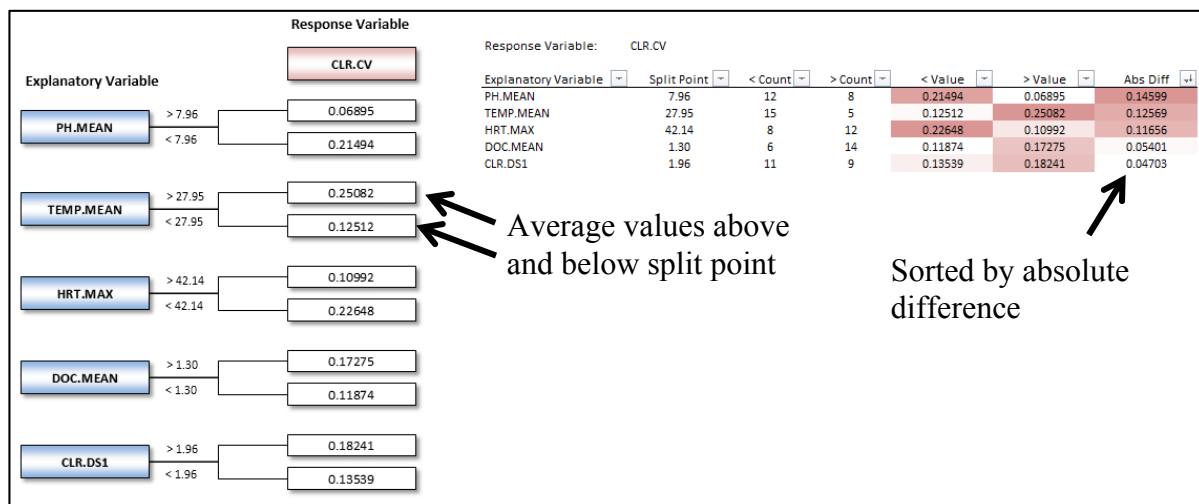
Figure 6.6 Instability factor analysis for chlorine residual CV

Once the type and degree of instability is determined, mitigation strategies can be identified and implemented. For example, WRF project 4312 found that for biofilm accumulation, the most important control variable is chlorine residual – increasing the residual has a powerful impact on reducing the biofilm formation rate. In general, to achieve a goal of 2-log (99%) reduction in biofilm formation rate, a chloramine residual of 1.0-2.1 mg/L would be required; for a 3-log

reduction, a chloramine dose of 2.0-3.0 mg/L would be required. For utilities that use free chlorine, a residual of 0.8-1.5 mg/L would diminish the biofilm formation rate by 2-log; for a 3-log reduction, a residual of 1.4-2.1 mg/L would be required. If for some reason the utility cannot increase (and maintain) the residual (due to trihalomethane formation, for instance), other control measures include reducing the mean corrosion rate through the improvement of corrosion control programs, or improving the coagulation/filtration process to improve removal of DOC or AOC. If loss in disinfectant residual is the primary concern, control measures could include reducing biofilms (through flushing for example), reducing the corrosion rate, or improving the removal of DOC and/or AOC. If corrosion is the primary concern, removing biofilms and increasing the phosphate and/or pH should be considered. The exact values and parameters may vary depending on site-specific characteristics and the functions of BSAT can be used to help identify site-specific parameters have the most impact on biostability. The *Guidance Manual for Control of Biostability in Drinking Water* (LeChevellier 2015b) should be consulted for additional information on how to interpret the data collected and design a mitigation strategy.

In summary, the BSAT is a data analysis tool that facilitates systematic evaluation of distribution system instability. It allows comparison to a database developed during by the WRF 4312 Project. However this comparison is just the beginning of a thorough analysis. Collection of distribution-specific data over a period of seasons and years is critical. Only through collection of these data will a utility be able to identify and analyze trends and correlations between causes and effects. This knowledge can be used to develop and implement instability mitigation strategies. The quantitative effects of these mitigation strategies can be observed using the BSAT provided that relevant data are collected before and after implementation. Ultimately, this dynamic process of distribution system analysis and improvement will lead to improved water quality and decreased costs.

CHAPTER 7
TROUBLESHOOTING

This section provides guidance for basic troubleshooting of BSAT error messages. If an error is encountered, first do the following:

(1) Check that the transmission axes names defined in the *Utility Information* dialog box match the transmission axes names as entered on the *Data Entry* page,
(2) Click the Update Formulas button on the *Data Entry* page, and
(3) Check that the date range is representative of your dataset. Click the Default button if you would like to use the complete utility dataset.

If an error message is encountered after resolving the above three steps then review Table 7.1 to determine what resolutions are required for each error message. If an error regarding security settings is displayed, then review the Security Settings section below.

Table 7.1 Common BSAT error messages

Error Message	Function	Resolution
Difficulty was encountered during the opening of BSAT. This may be due to incompatible or unanticipated security settings. To ensure that BSAT operates correctly, it is recommended that you modify your security settings. Then close and re-open BSAT.	Opening Workbook	See Security Warnings and Settings discussion after Table 7.1 below.
The value at the selected cell cannot be evaluated to a correct date format.	Data Entry	The date format is either blank or not in an acceptable format to perform data analysis. Reformat the date to be MM/DD/YYYY format.
You must select at least 1 Explanatory Variable	Instability Factor Identification	This message is displayed if the *Instability Factor Identification Variables* tab did not have any explanatory variables selected. Select at least one variable for inclusion in the analysis.

(continued)

Table 7.1 (Continued)

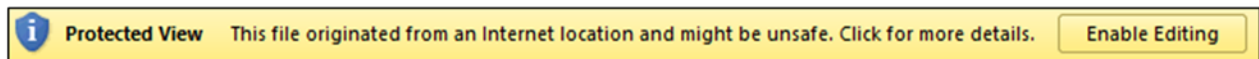| There are insufficient data to conduct an analysis. Try selecting a different Transmission Axis or expanding the Date Range. You may need to enter more data.	Instability Factor Identification	If the database is deleted or is empty then analyses cannot be completed until data are entered. Results will appear blank and a warning message will appear indicating that there are insufficient data. This may also occur if the transmission axis is defined on the *Utility Information* dialog box but there are no data for that transmission axis in the database.
No matches were found between your selected Explanatory Variables and those in your dataset. Try making different selections.	Instability Factor Identification	Select at least one variable that is included in the utility dataset.
There are an insufficient number of data pairs to conduct an analysis. A minimum of 4 matched Explanatory-Response variable data pairs are required. Try selecting a different set of Explanatory Variables. Or you may need to collect more data.	Instability Factor Identification	A minimum of 4 data points must be included to perform the analysis. More data must be added to utilize this function.
No data were found for the selected Response Variable. Try making different selections.	Instability Factor Identification	To perform this analysis, a minimum of 4 data points must be provided for the Biological Stability Issues shown on the *Variables* tab of the *Instability Factor Identification* dialog box (e.g., disinfectant residual stability, biofilm formation rate, or corrosion rate).
A file cannot be saved because no Pages (Sheets) are selected.	Results Output	On the *Save* tab of the *Results Output* dialog box, select at least one of the analyses to save.

(continued)

Table 7.1 (Continued)

A file cannot be saved because the File Name is blank.	Results Output	Add a file name to save as a pdf file; note that the file name extension (.pdf) does not need to be included.
A file cannot be saved because the File Name contains a reserved (not allowed) character.	Results Output	Revise file name to use only allowed characters for Windows file names.
BSAT encountered a problem in trying to save the results output file. There are several possible reasons, including: the file path does not exist, has been deleted, or is inaccessible; the file is being used by another program; or the file you are trying to save has the same name as a currently opened file.	Results Output	If a file with the same File Name already exists, the warning message appears, prompting whether or not the user wants to overwrite the file. Also, an error message will appear if the file is already opened or is being used somehow by another opened application.
An error occurred during this process. Try making different selections. Also, check to be sure sufficient data have been entered.	General processing error	There may not be sufficient utility data entered to perform the selected analysis. Confirm that the date and transmission axes were entered correctly in the *Utility Information* dialog box and on the *Data Entry* page.

Security Warnings and Settings

If the ***BSAT*** tab is not visible, then the workbook file has likely opened in Protected View mode, which typically occurs after the file has been downloaded from an internet location. Normally in this case you will see a notification message similar to the following:

Protected View   This file originated from an Internet location and might be unsafe. Click for more details.   Enable Editing

Click Enable Editing to exit Protected View mode. If this message does not appear, and Protected View mode is still active, then you will need to change your Trust Center settings; after which you may then need to close and re-open this file.
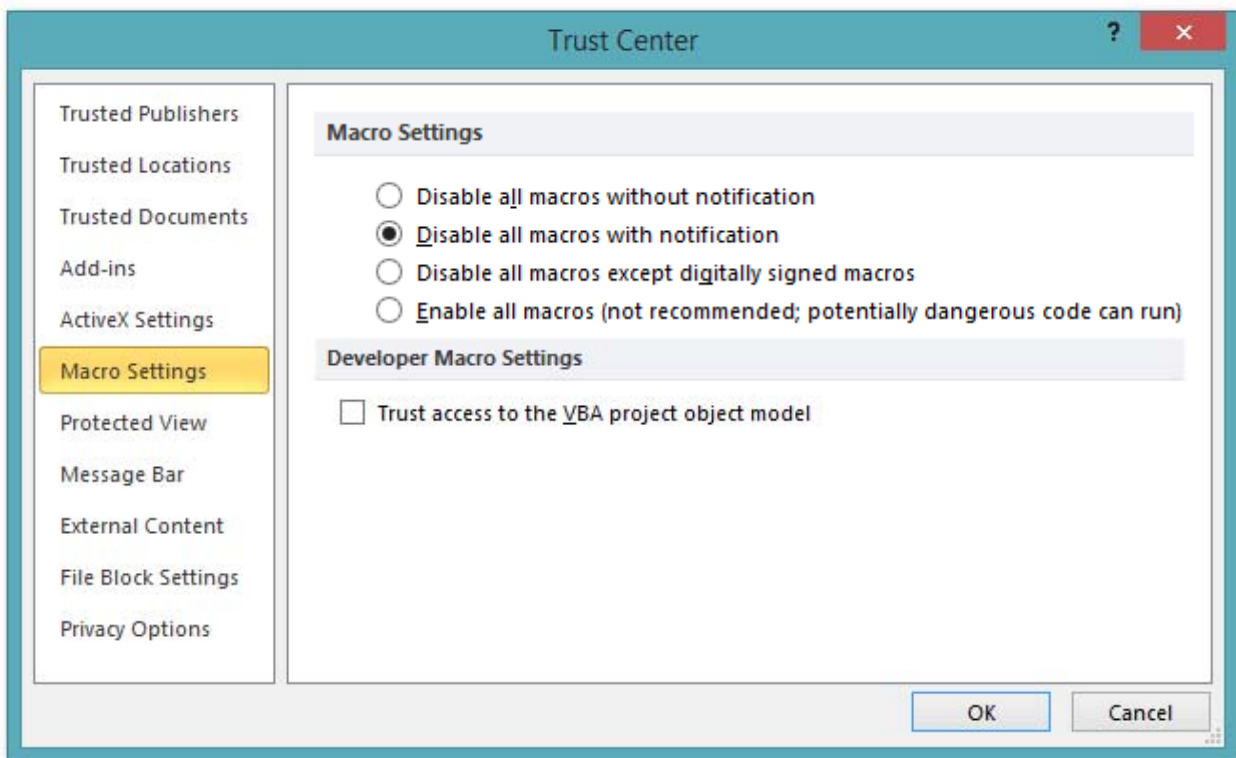
Macros

This workbook file contains macros, and therefore to operate the application features of this workbook it must be opened with the Macros Enabled. If the ***BSAT*** tab is visible but you experience no response when selecting icons on the ***BSAT*** tab, then the workbook file has likely

opened with the macros disabled. Normally in this case you will see a notification message similar to the following:
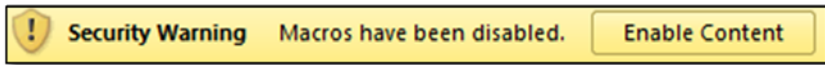
Click Enable Content to enable the macros. In Excel 2007, the warning message is slightly different: Instead of Enable Content, you will need to click Options and then select the Enable this Content option from the list that appears. In either case, if the "Macros have been disabled" notification does not appear, and there is still no response when selecting the **BSAT** icons, then you will need to change your Trust Center settings; after which you may then need to close and re-open this file. The recommended (default) Trust Center setting is:

After opening BSAT the first time and clicking Enable Macros or selecting Enable this content, Excel thereafter opens the workbook with the macros automatically enabled (i.e., the yellow Message Bar notification does not appear again).

REFERENCES

LeChevallier, M.W., O.D. Schneider, L.A. Weinrich, P.K. Jjemba, P.J. Evans, J.L. Hooper, and R.W. Chappell. 2015a. *An Operational Definition of Biostability*. Denver, Colo.: Water Research Foundation.

LeChevallier, M.W., O.D. Schneider, Lauren A. Weinrich, Patrick J. Jjemba, Patrick J. Evans, Jennifer L. Hooper, and Rick W. Chappell. 2015b. *Guidance Manual for Control of Biostablity in Drinking Water*. Denver, Colo.: Water Research Foundation.