U.S. DEPARTMENT OF ENERGY

PNNL-16781

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

FRAMES Software System: Linking to the Statistical Package R

KJ Castleton
G Whelan
BL Hoopes

December 2006

Pacific Northwest
NATIONAL LABORATORY

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-ACO5-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576 5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@nits.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

PNNL-16781

FRAMES Software System:
Linking to the Statistical Package R

K. J. Castleton
G. Whelan
B. L. Hoopes

December 2006

Prepared for
Engineer Research and Development Center
U.S. Army Corps of Engineers
Vicksburg, MS
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Summary

The Framework for Risk Analysis in Multimedia Environmental Systems (FRAMES) is a Windows-based software platform that provides an interactive user interface and, more importantly, specifications to allow a variety of DOS and Windows-based environmental codes to be integrated within a single framework. The major components of the FRAMES software include modules (module user interface, environmental code and potentially pre- and/or post-processors), the Framework User Interface (FUI), sensitivity/uncertainty module, databases, and data viewers. Although FRAMES is domain independent, it is being used herein to support environmental assessment. Modules represent a particular step in the risk assessment process, such as type of contaminant selection, source release, fate and transport (groundwater, vadose zone, surface water, air, overland), exposure pathway (farm foodchain, ingestion, inhalation, dermal, external), and risk (dose, cancer incidence or fatalities, and hazard quotient). Modules can accept data from the user or other modules and can calculate some portion of the risk assessment. The FUI allows the user to interact with the system. The sensitivity/uncertainty module allows the user to conduct a Monte Carlo analysis, and the viewers allow the user to review results from a particular stage in the process. To achieve realistic scenarios, users often need to combine multiple modules and databases together, where data seamlessly flow from one component to the next.

FRAMES currently has a mechanism to perform Monte Carlo uncertainty analyses, including Latin Hypercube sampling. A need still exits to help the user understand the significance of input parameters in affecting output results in any Monte Carlo analysis. Because FRAMES is specifically designed to allow for direct linkages to disparate legacy software products developed by others, legacy off-the-shelf software products can be linked to the system to meet this need. The purpose of this effort is to relate, through a mathematical (linear or nonlinear) model, a dataset of user-defined predictor (i.e., input or independent) and response (i.e., output or dependent) variables, and to define how to best understand the relative contributions of the predictors. This document discusses the linkage between the statistical package “R” (also known as R-Project) and FRAMES Versions 1.x and 2.0. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques and is highly extensible. R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and Mac OS.

This document provides requirements, design, data-file specifications, test plan, and Quality Assurance/Quality Control (QA/QC) protocol for the linkage between R and FRAMES Versions 1.x and 2.0. The requirements identify the attributes of the system. The design describes how the system has been structured to meet those requirements. The specification presents the specific modifications to FRAMES to meet the requirements and design. The test plan confirms that the basic functionality listed in the requirements (black box testing) actually functions as designed, and QA/QC confirms that the software meets the client’s needs.

Acronyms

ADDAMS	Automated Dredging Decision, Analysis, and Modeling System
ARAMS	Adaptive Risk Assessment Modeling System
BSAF	Bio-Sediment Accumulation Factor (database)
COC	Compound of concern
CSM	Conceptual Site Model, simplified description of the environmental problem to be modeled
DoD	U.S. Department of Defense
DOE	U.S. Department of Energy
DOS	Disk Operation System, basic operation system on the computer
ecdf	empirical cumulative distribution function
EHQ	Ecological hazard quotient
EPA	U.S. Environmental Protection Agency
ERD	Ecosystems Research Division
ERDC	Engineer Research and Development Center
FRAMES	Framework for Risk Analysis in Multimedia Environmental Systems
FUI	Framework User Interface
GMS	Groundwater Modeling System
MEPAS	Multimedia Environmental Pollutant Assessment System
NERL	National Exposure Research Laboratory
NRC	U.S. Nuclear Regulatory Commission
OCRWM	Office of Civilian Radioactive Waste Management
ORD	Office of Research and Development
ORIA	Office of Radiation and Indoor Air
OSW	Office of Solid Waste
PNNL	Pacific Northwest National Laboratory
QA/QC	Quality Assurance and Quality Control, processes that confirm the quality of the product
RAIS	Risk Assessment Information System
TTD	Terrestrial Toxicity Database

Acknowledgments

The authors would like to extend their appreciation to Dr. Mark Dortch and Mr. Jeff Gerald of the U.S. Army Corps of Engineers, Engineer Research and Development Center (ERDC) in Vicksburg, MS, for supporting this work. Thanks are also extended to Dr. James G. Droppo, Jr. for reviewing the material and Mr. Wayne Cosby for the final edit of the document. This work was funded by ERDC under Contract DE-AC05-76RL01830.

Contents

Figures

Tables

1.0 Introduction

The Framework for Risk Analysis in Multimedia Environmental Systems (FRAMES) is a Windows-based software platform that provides an interactive user interface and, more importantly, specifications to allow a variety of disk operations system (DOS) and Windows-based environmental codes to be integrated within a single framework. The major components of the FRAMES software include modules (module user interface, environmental code, and potentially pre- and/or post-processors), the Framework User Interface (FUI), sensitivity/uncertainty module, and data viewers. Modules can accept data from the user or other modules and can calculate some portion of the risk assessment. The FUI allows the user to interact with the system. The sensitivity/uncertainty module allows the user to conduct a Monte Carlo analysis, and the viewers allow results to be reviewed from a particular stage in the process.

FRAMES is a software platform that allows users the ability to select and implement environmental software models for risk assessment and management problems. This program is a flexible and holistic approach to understanding how activities affect humans and the environment. It links models that integrate across scientific disciplines, allowing for tailored solutions to specific activities, and it provides meaningful information to business and technical managers. FRAMES is the key to identifying, analyzing, and managing potential environmental, safety, and health risks. The purpose of FRAMES is to assist users in developing environmental scenarios and to provide options for selecting the most appropriate computer codes to conduct human and environmental risk-management analyses.

1.1 Background

The U.S. Environmental Protection Agency (EPA) is charged with developing, implementing, and enforcing regulations concerning protecting human and ecological health from the myriad of chemical and non-chemical stressors imposed on the environment as a result of human activities. The U.S. Department of Energy (DOE), U.S. Department of Defense (DoD) Army Corps of Engineers, Engineer Research and Development Center (ERDC), and U.S. Nuclear Regulatory Commission (NRC), in response to existing and emerging regulatory requirements for environmental protection, have developed significant programs to assess exposure and risk at their facilities. In pursuing these activities, ERDC, NRC, DOE, and EPA share a common need to understand the environmental processes (physical, biological, and chemical) that collectively release, transform, and transport contaminants and result in exposure and finally a probability of deleterious health effects. At ERDC, NRC, EPA, and DOE, computer models are key tools to organize the knowledge of environmental science and apply it to the decision-making process.

Through the past 13 years, EPA and DOE have jointly pursued common interests related to environmental modeling. For example, in 1995, DOE’s Pacific Northwest National Laboratory (PNNL) and the EPA Office of Air and Radiation in the Office of Radiation and Indoor Air (ORIA) joined efforts to design and develop a prototype multimedia modeling system. The unique aspect of this effort was the incorporation of software modules representing individual steps of a risk assessment (e.g., source release of contaminants, fate and transport in surface water, exposure) within a software framework. A module represents the computer model (i.e., code), pre- and post-processors, and user interface. The software framework was designed using “object-oriented” design and, as such, allowed for the decoupling of individual models. This design greatly improved the ability of model developers (e.g., a modeler developing a new surface-water model) to “plug” the new model into a full multimedia modeling system without the need to develop a complete modeling system. The product of this effort was FRAMES. FRAMES Version 1 allows a user to conduct multimedia simulations of contaminant-based exposure and risk at a single facility.

Concurrent to the development of FRAMES Version 1, DOE and the EPA Office of Research and Development (ORD), Ecosystems Research Division in Athens, Georgia, also initiated a joint effort in 1995 to study existing technology and future needs of EPA and DOE related to multimedia/multi pathway exposure and risk assessment. The latest stage of joint efforts between DOE and EPA (1998 to 2001) was to extend and refine FRAMES to build a modeling system capable of conducting a national assessment of exposure and risk as a result of contaminant releases from hazardous waste sites (NERL 2006). This national assessment modeling approach is called FRAMES-3MRA and was developed and implemented for the EPA Office of Solid Waste (OSW). In the past several years, the EPA National Exposure Research Laboratory (ERD 2006) has responded to these needs by establishing specific research and development tasks to "integrate" all multimedia modeling-based activities, including the FRAMES-based efforts. The goal of this initiative is to design and implement, over the next decade, a multimedia integrated modeling system that will facilitate future environmental assessments and related research. EPA-ORD-ERD is currently supporting the merger of the FRAMES Version 1, which is designed for site-specific assessments, with FRAMES-3MRA, which is designed to perform regional and national assessments. The intent is to capture the best attributes of both into one system for use by others to form FRAMES Version 2.

Concurrent to and in parallel with EPA and DOE, ERDC and NRC have programs that support various aspects associated with the FRAMES approach. ERDC initiated the development of the Adaptive Risk Assessment Modeling System (ARAMS 2006) because the military needed a system that was compatible and consistent with similar systems at other agencies. NRC initiated the linkage of various models contained in the Groundwater Modeling System (GMS 2006) with FRAMES, so users would have direct access to more science-supported groundwater models for their assessment exercises. In each case, FRAMES represents middleware that seamlessly connects these disparate components, and FRAMES is used as the execution manager.

ARAMS is a computer-based, information delivery, modeling, and analysis system that integrates multimedia fate/transport, exposure, uptake, and effects of constituents (i.e., compounds of concern [COCs], including chemicals and radionuclides) to assess human and ecological health impacts and risks (http://el.erdc.usace.army.mil/arams/). ARAMS uses an object-oriented, system framework to construct a computational conceptual site model composed of the environmental pathways and exposure routes linked to various models and databases for exposure and effects assessments.

The Multimedia Environmental Pollutant Assessment System (MEPAS) is one of the modeling systems in ARAMS that helps form the basis for many of the preliminary- or screening-level models/calculations, including releases from contaminant sources; fate/transport in air, streams, vadose zone, and groundwater; multimedia exposure pathways for humans, such as food via uptake through crops and farm animals; human intake; and human health impact/risk calculations. The MEPAS fate/transport models are typically semi-analytical solutions to simplified transport equations for each medium. Other ERDC and related models (e.g., RECOVERY, TPB, TWEM, WEAP, and HELPQ) have been added, and others can be added later as needed. RECOVERY and HELPQ are part of the U.S. Army Corps of Engineers Automated Dredging Decision, Analysis, and Modeling System (ADDAMS).

ARAMS provides additional ecological features that many of the current systems do not. Several web-based ecological databases have been linked to the system, including the Terrestrial Toxicity Database

(TTD), which contains ecological soil screening levels and ecological-effects toxicity reference values for wildlife that can be used to screen soil concentrations and compute ecological hazard quotients (EHQ), respectively. There is also a Bio-Sediment Accumulation Factors (BSAF) database that contains BSAF and percent lipid values for various aquatic organisms. BSAF is used to compute tissue residues (mg/kg) of constituents of concern in aquatic organisms. The system is also linked to the web-based Risk Assessment Information System (RAIS). This allows for the direct download of chemical-specific factors (physicochemical properties, exposure factors, and bioaccumulation factors) and human health toxicity reference values.

The NRC required seamless linkages between numerical groundwater models and models addressing other aspects of the hydrosphere, biosphere, atmosphere, and geosphere. Specifically, when licensed nuclear facilities want to terminate their nuclear operating license, they need to demonstrate that the facility, when vacated, does not create an unacceptable health impact. NRC provides approaches for the license terminees, but those approaches are very conservative and may not show that the site has *de minimus* impacts. As such, NRC wants to provide the license terminees with access to some of the most sophisticated groundwater models in the world, contained in the GMS (http://www.emrl.byu.edu/gms.htm). FRAMES provides the capability to link models to GMS, so more sophisticated and science-based models can be used in the license termination process with other non-groundwater modeling systems.

While the various federal agencies (i.e., DoD, EPA, NRC, and DOE) are sponsoring their own individual projects, they are also working together to confirm that their investments in FRAMES are leveraged to the products that are useful to others. The efforts described herein represent a product of the leveraged investment.

1.2 Purpose

The purpose of this effort is to relate, through a mathematical (linear or nonlinear) model, a dataset of user-defined predictor (i.e., input or independent) and response (i.e., output or dependent) variables and to define how to best understand the relative contributions of the predictors. One could assume that all input variables are independent of each other. This analysis is fairly straightforward, except that many parameters are correlated to each other. One approach to address this is by using partial correlations in the predictor-variable selection process.

Partial correlation is used to understand the extent of variance in a dependent variable, accounted for by the independent variables. A distinction needs to be made between prediction and explanation. To predict a dependent variable from a set of predictor variables, it is less important to know the relative contributions of the independent variables than it is to minimize the prediction error (maximize R2, coefficient of determination); however, it still may be important to have an efficient set of predictor variables (i.e., minimum set required to calculate adequate predictions). On the other hand, if one is trying to understand causal relationships, interactions among variables can be very complex, even if there are only a few predictor variables. If the predictor (independent) variables are not correlated with each other, then the problem is simple; each will be independently added to an explanation of the variance in the dependent variable. If they are correlated (or some of them are correlated), then it is difficult to tease apart the relative contributions of the independent variables. A predictor could be highly correlated with a criterion (or response) variable, and yet add nothing to a given set of predictors because it is redundant. Or a predictor could have zero correlation with a criterion, yet greatly improve the prediction because of how it is correlated with other predictors (a suppressor variable). Consequently, understanding the interactions among predictor variables can involve convoluted reasoning, even when there are as few as three or four predictors, let alone a dozen or more.

When the number of independent variables increases beyond three or four, several approaches have been proposed for both finding the optimal set of regression variables and understanding the relative contributions of the predictors to explain the independent variable. The regression weights for correlated variables depend on the order in which the regression is carried out (e.g., step-wise regression), and different methods are available to determine the order in which variables are introduced. In addition to the order of entry of variables, the data should be explored to determine if the relationships are linear and whether other technical assumptions used to justify the use of linear models apply. Ideally, data should be plotted to determine if linear models are appropriate or whether it is necessary to transform the data before using a linear model.

1.3 Component Description

Some of the options for determining correlations and causality among variables are listed below:

1. Calculate combinations of first-, second-, third-, etc. order partial correlations. These are easily calculated from a correlation matrix, but interpretation can be difficult. First-order partial correlations hold constant (i.e., partials out) one variable; second-order, two variables, etc. Also, the number of predictor variables can vary. With four predictor variables, there are 28 first-order partial correlations, 18 second-order correlations, and 4 third-order correlations.
2. Calculate semi-partial correlations (part correlations). Whereas partial correlations capture the percent of variance accounted for by a set of predictor variables with other variables held constant, part correlation is the percent of variance uniquely accounted for by a predictor variable.
3. Provide scatter plots based on simple and partial correlations. As more and more predictor variables are involved, informative displays become an increasingly important consideration.

Some of the options for determining an efficient set of predictor variables are listed below (Neter et al. 1985):

1. Perform stepwise regression, and calculate the F-statistic to determine if additional variables are statistically significant. Stepwise regression adds and removes variables via an automated search procedure. One criterion for adding or removing variables is based on partial correlations. This is a good method in the case of dozens of predictor variable candidates.
2. Perform an all-subsets regression. This method requires that all possible regression models be examined based on some criterion, such as coefficient of multiple determination (R2), mean-square error, and Mallows Cp.
3. Provide principal-components regression to combine correlated variables. This is an excellent addition to the analysis
   - as a data-screening technique
   - to identify the variables that contribute the most to the system.
4. Implement the simulation approach. Randomly take samples of approximately 5 to 10 variables and fit a model, noting which ones always have high coefficients.

To increase the comfort level for users without a strong statistics background, it would be very valuable and useful to provide documentation within FRAMES to explain to how to interpret and understand partial correlations. The level of effort would depend in part on the extent of the multiple correlation/regression analysis supported in FRAMES. The documentation could also include sections on regression and correlation diagnostics to indicate redundant variables, ill-conditioned covariance matrices, etc.

Software for these statistical procedures has been in existence for a long time (e.g., FORTRAN codes and off-the-shelf packages), and codes suitable for inclusion in FRAMES may well exist in the public domain (e.g., NetLib software repository or some NIST software repositories). The freeware statistical package R (also referred to as R-project because it is a web-searchable phrase) is available (http://www.r-project.org) (The R Project 2006). Because the R Project appears to be nearly equivalent to SASS or S-Plus, it has been linked to FRAMES. It relates a dataset of user-defined predictor (i.e., input or independent) and response (i.e., output or dependent) variables and defines how to best understand the relative contributions of the predictors. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and Mac OS.

Using R, three viewers have been developed: Scatter Plot (SP), Cumulative Distribution Function (CDF), and Histogram. These viewers allow a user to observe the relationships of the inputs to outputs or simply find a value from a distribution for a given probability. The Scatter Plot (SP) viewer constructs a scatter plot that graphs all input variables against all other input variables and all output variables. R's default approach is used to generate a scatter plot. The SP performs an optimized regression analysis using R's linear model and stepwise capability. It performs this analysis once for each output variable that is provided in the output dataset. The development of the CDF plot uses R's capability to plot CDFs of values, which allows a user to find a given input or output value for given a percentile. A 95th percentile line is added for reference into the CDF viewers output. The histogram plot provides the frequency of occurrence versus all independent and dependent variables chosen by the user for the statistical analysis.

When graphics are produced, they are created in Portable Network Graphics (png) format and are then shown to the user through a web browser. The graphics can be used in web pages as well as added to documents. The general approach is to read data from the input or output dataset then write a text file that can be directly read by R. A file that contains the "R commands" is also written at the same time. R is then invoked with the command file. When R's invocation is completed, any graphics that are produced are rendered in a web browser, and any text results are rendered in the user's default text editing program. For most users, this is Windows' Notepad. From there, the text could be copied and pasted into documentation.

2.0 Requirements

Requirements are characteristics and behaviors that a piece of software must possess to function adequately for its intended purpose. The requirements for the Scatter Plot (SP), Cumulative Distribution Function (CDF), and histogram viewers are nearly identical except the fact that they perform different analyses on the same type of data.

1. When sampled input and selected output are generated by the same icon, as in the FRAMES-1.x Sensitivity/Uncertainty icon, then the viewer can be described as a right-click viewer and produced by a right-click on the icon menu. This is currently a FRAMES-1.x requirement.
2. When the sampled input and selected output are generated by different icons, as in the FRAMES-2.0 Sampler and Summarizer icons, respectively, then the viewer is described by a Drag&Drop icon and connected to the icons producing the sampled input and selected output. This is currently a FRAMES-2.0 requirement.
3. The R results produce a scatter plot in png format of all sampled input and selected output.
4. For each selected output, a regression analysis is performed that summarizes the regression in the R standard form, whose results are provided in a standard Text (*.txt) file.
5. For each selected output, a regression analysis is performed that shows the contribution of each sampled input to the Multiple R-squared, as reported in the R standard summary in a standard Text (*.txt) file, capturing the degree that each sampled input has on influencing the variability of each selected output.
6. A viewer is produced that provides a CDF, using R’s empirical cumulative distribution function (ecdf) plotting capability, for each sampled input and selected output.
7. Each CDF plot contains a line at the 95th percentile for the user’s reference.
8. A viewer is produced that provides a histogram, using R’s plotting capability, for each sampled input and selected output.

3.0 Design

Design elements are strategies for meeting requirements. The SP, CDF, and Histogram viewers are designed to meet the requirements identified in Section 2.0. Key to meeting those requirements is the use of R's built-in capabilities to perform statistical analyses. The SP, CDF, and Histogram viewers use R as a scripted interface tool, so these viewers write a script that R then reads and performs. The results of the computation are stored in either textual or graphical files. Graphics are stored in a png format that can be added to documents or web pages, and the text is stored as Text (txt) files.

To better explain the functionality of these tools, a simple input-output dataset is presented for purposes of demonstration. This dataset was used as an example by Borcard (2002). In Borcard's example, y represents the dependent variable (i.e., selected output), and x1 and x2 (i.e., sampled input) represent independent variables. The variables and their values are presented as follows:

y	*x1*	*x2*
4	*1*	*8*
2	*1*	*7*
3	*1*	*7*
4	*1*	*9*
5	*1*	*5*
5	*1*	*4*
7	*2*	*3*
5	*2*	*6*
4	*2*	*7*
9	*2*	*2*
7	*2*	*3*
6	*2*	*2*

The SP performs two operations that show the relationships between inputs and outputs, specifically, relationships between the sampled input and selected output, producing a scatter plot and a text file summarizing the regression analysis summary. Both a scatter plot and a text file are produced for every variable in the system (either sampled or summarized) to show their relationships to one another. Figure 3.1 presents a series of scatter plots showing the relationship between each variable. The left diagonal mirrors the right diagonal.

Figure 3.2 presents a regression analysis summary of the analysis, including the contributions to the Multiple R-Squared for each input variable (i.e., x1 and x2). The output provides the residuals, coefficients, residual standard error, etc. It also identifies the significance that each independent variable has on influencing the variability of the output. This information is translated to provide the fraction of the contribution due to each input variable for each output variable, illustrated as follows:

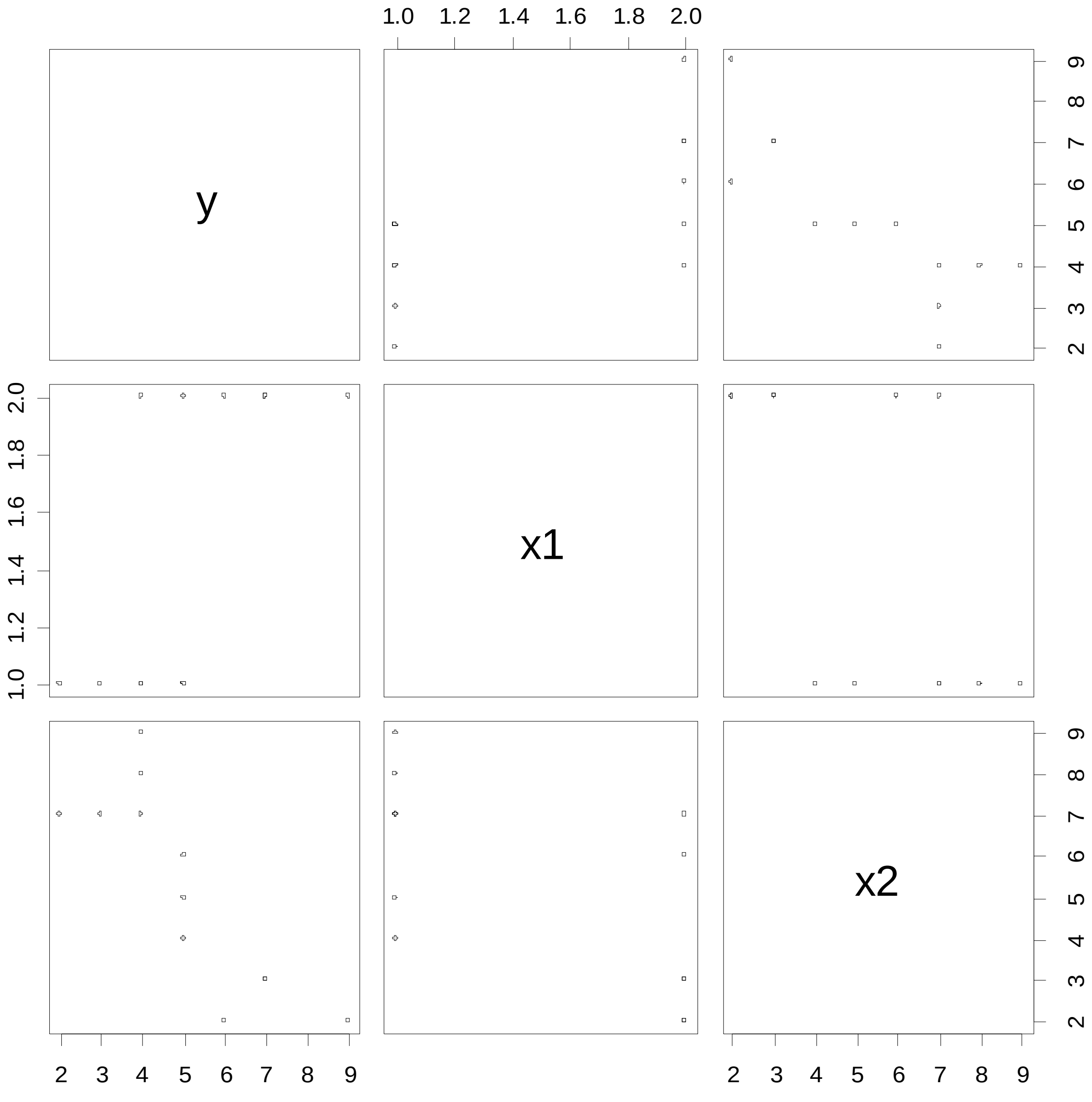
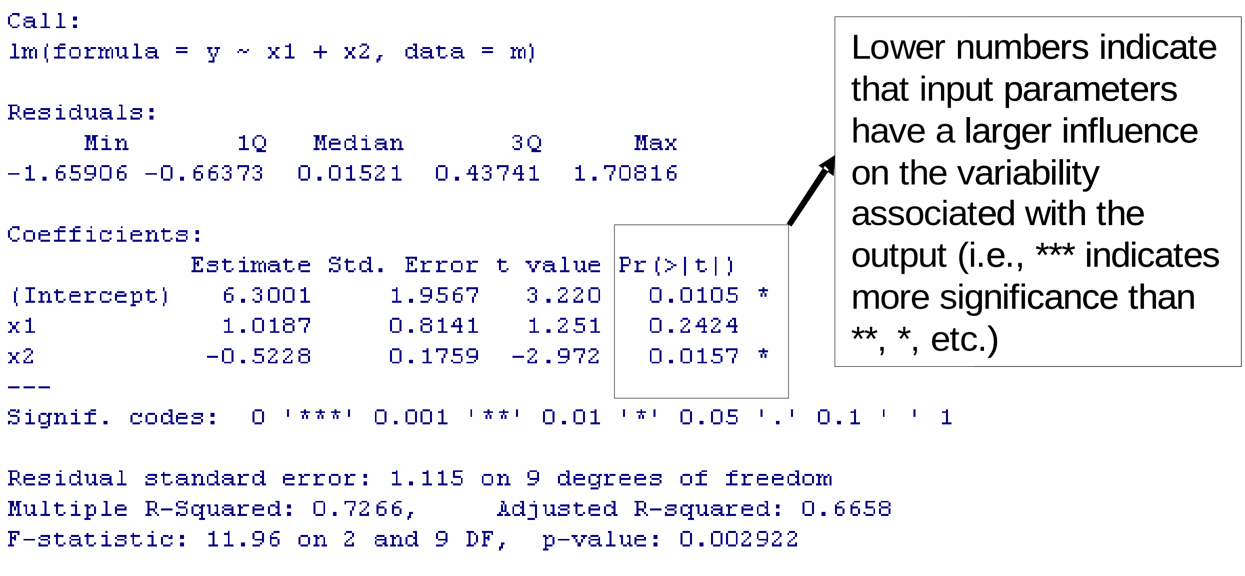


Figure 3.1. Sample Scatter Plot of y, x1, and x2



```
Call:
lm(formula = y ~ x1 + x2, data = m)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65906 -0.66373  0.01521  0.43741  1.70816

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.3001     1.9567   3.220   0.0105 *
x1            1.0187     0.8141   1.251   0.2424
x2           -0.5228     0.1759  -2.972   0.0157 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.115 on 9 degrees of freedom
Multiple R-Squared: 0.7266,     Adjusted R-squared: 0.6658
F-statistic: 11.96 on 2 and 9 DF,  p-value: 0.002922
```

Lower numbers indicate that input parameters have a larger influence on the variability associated with the output (i.e., *** indicates more significance than **, *, etc.)

Figure 3.2. Summary of Simple Regression Analysis

Contributions
x1
0.187
x2
0.540

In Borcard's example, x2's influence is nearly three times that of x1 (0.540/0.187). It is important to note that the summary of the model and the associated contribution are produced for every output. In this example, we only have the output y, but if a y2 were also one of the summarized values, then a complete analysis would be produced for y2 corresponding to the inputs x1 and x2. Summarized values often contain counts, minimums, and maximums. There seems to be little value in analyzing the counts, so they are left out of the scatter plots and regression analysis.

Figure 3.3 presents examples of CDF viewers that would be produced. The CDFs would be produced for each sampled input and selected output. The viewer will generate a CDF with a 95th percentile line for all inputs and outputs. Each output and input variable is sorted and plotted, so the values for each given probability can be found. Because these are discrete distributions, the value is shown as a dot, and then a line extends identical values. Given the small sample size of our example, it would be very difficult to interpret a 95th percentile value from the distributions. The 95th line is drawn as a guide for discussion.

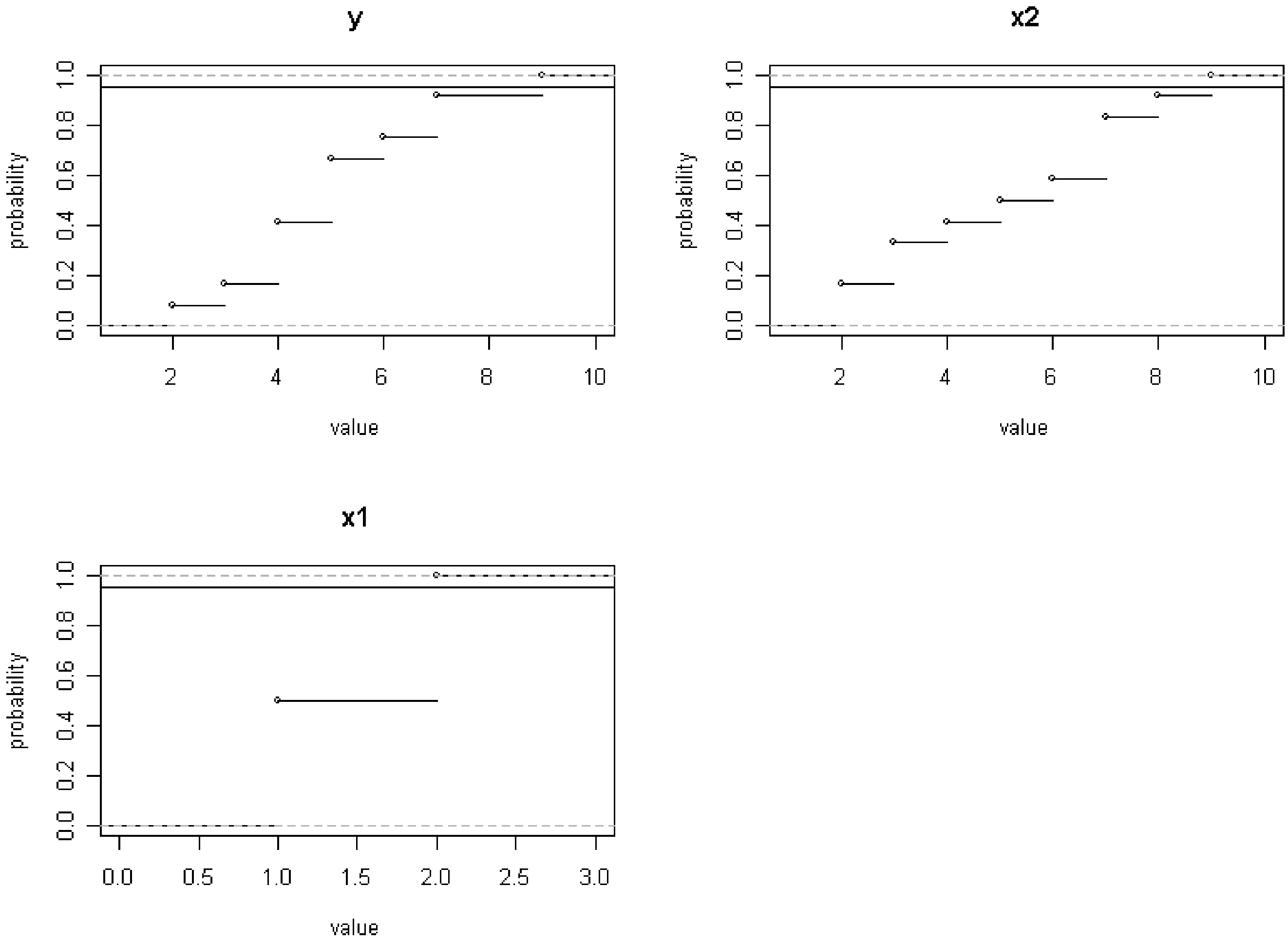


Figure 3.3. CDF Output Graphic

Although a histogram was not developed for this example, the user also has the ability to produce a png of a histogram, as illustrated in Figure 3.4. Figure 3.4 presents histograms for two input variables (Bd and I) and two output variables (Ca_All_Surface_Water_1 and Ca_All_Surface_Water_10). The y axis contains the frequency of occurrence, and the x axis contains the range in values. From the histograms, one can obtain an idea of the distributions chosen for the input parameters and calculated for the output parameters.

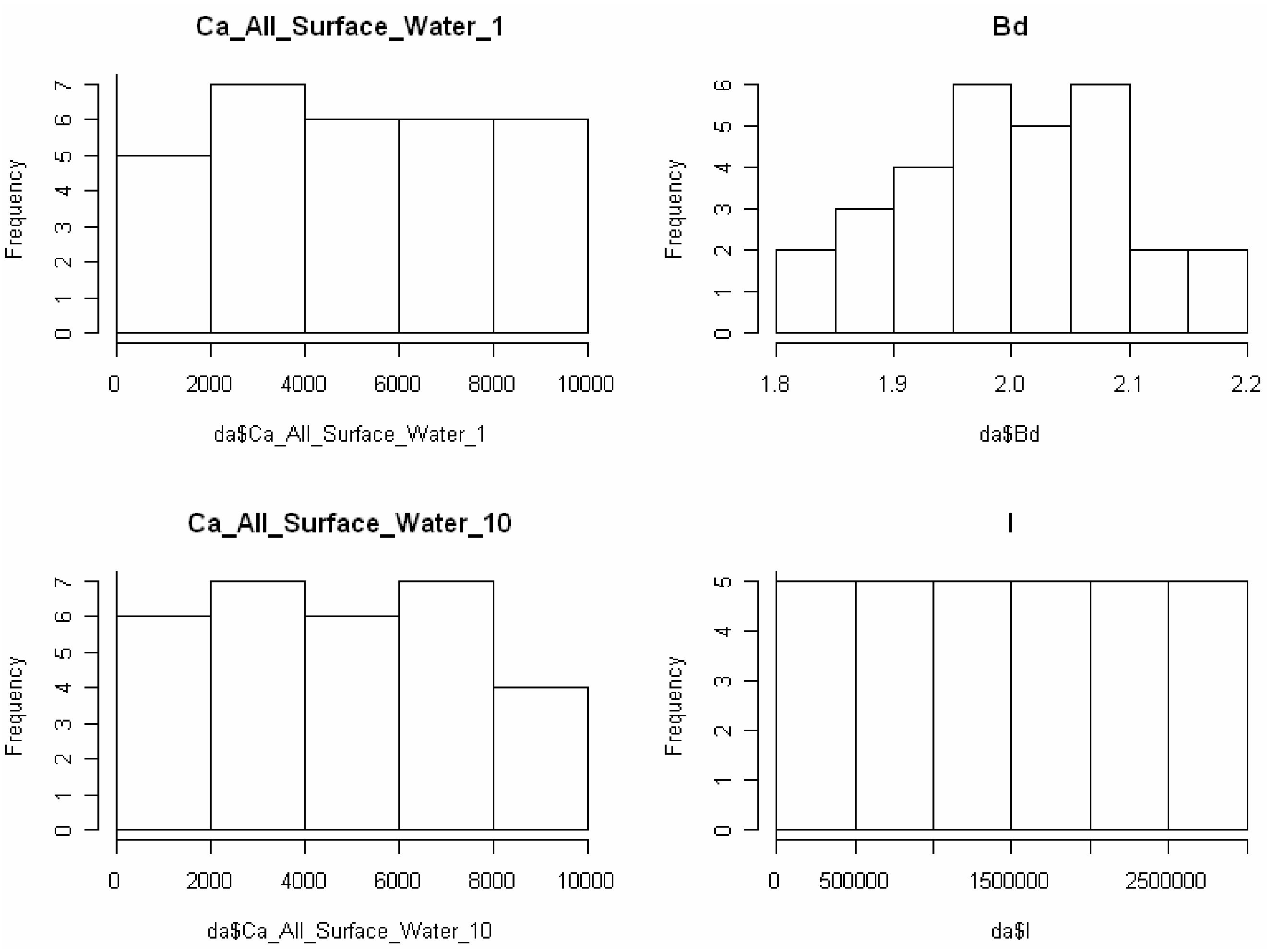


Figure 3.4. Illustrative Example of Input and Output Histograms

3.1 Limitations and Issues

There are no facilities in the SP, CDF, or Histogram viewer to allow the user to interact within R. The data file that was read by R (i.e., sampled input and selected output datasets) can be found in the root directory of the C: drive (the location of FRAMES is ignored in this case). This facilitates R finding the file in the script and running correctly. The line below reads the data prepared for R into a data frame identified by the name “da.” This line can be typed into R to begin more complex analyses of the data by someone familiar with R.

da <-read.csv(\"PartialR.csv\",strip.white=TRUE)

The entire script of the viewer is written to a file name PartialR.R, so the R user could use this as a guide as to how the initial analysis was performed.

4.0 Specifications

Specifications are the descriptions of input and output files that are used during execution of a component. Alongside these input and output descriptions, specifications address how the component is invoked and how the component was compiled. If a virtual environment, like Java, is used for execution, the version associated with that environment also represents a specification. The file format that is being read by the SP, Histogram, and CDF viewer is based on the Sensitivity/Uncertainty File (SUF), which is part of the standard FRAMES-1.x system. The file format is documented at http://mepas.pnl.gov/FRAMESV1/suf.html. This format is not repeated here, but it contains the values of the changed inputs and the values of the resulting outputs. A more modern form of this same information is represented by the Sampled Values and Summary Values datasets in FRAMES-2.0.

4.1 FRAMES-1.x Command Line

The command line for the SP, Histogram, and CDF is the standard FRAMES 1.x module command line. The module, when invoked, is passed to the 1) input file location, 2) output file location, 3) module name, and 4) module ID number. These four pieces of information are used to define the location of the input files that are read by and written by the SP, Histogram, and CDF modules.

4.2 FRAMES-2.0 SampledValues and SummaryValues

The Scatter Plot (SP), Histogram, and Cumulative Distribution Function (CDF) viewers read SampledValues, whose metadata are specified by SampledValues dictionaries, and writes SummaryValues, whose metadata are specified by SummaryValues dictionaries. Tables 4.1 and 4.2 present the SampledValues and SummaryValues dictionaries, respectively.

4.3 FRAMES-2.0 Command Line

The command line for the SP, Histogram, and CDF is the standard FRAMES module command line. The module, when invoked, is passed to the FRAMES 2.0 path, Simulation Name, and Module Name. These three pieces of information are required to connect to the FRAMES system with the SystemIO.dll.

4.4 Compiler Version

The SP, Histogram, and CDF viewers are all Java code that is contained in a BeanShell Script. BeanShell interprets the Java code during runtime. No code is compiled.

4.5 Java Runtime Environment Version

The BeanShell environment requires a Java Runtime environment better than 1.5.

Table 4.1. Sampled Values from Distributions (SampledValues)

Version: 0 Privilege: System Boundary

Name	Description	Unit	Measure	Type	Range	S	D	U	K	Prep	Indices
Indices	Dimension indices for sampled variables.			INTEGER	0-32767	Y	2	N	N		Variables
Iterations	Set of iterations that have been sampled			INTEGER		Y	1	N	N	for	
Set	Data set for sampled variables			STRING		N	1	N	N		Variables
ShortName	Alias name for sampled variables			STRING		N	1	N	N		Variables
Value	Sampled value of variable			FLOAT		N	2	N	N		Variables Iterations
Variables	Variable names that have been sampled			STRING		Y	1	N	N	for	
VarUnits	Units for sampled variables			STRING		N	1	N	N		Variables

Legend:

Column Name	**Meaning**
S	Self-Indexed
D	Dimensional Size
U	Uncertainty can apply (Stochastic)
K	Is the variable a key to others

Table 4.2. Summary Values from Simulation (SummaryValues)

Version: 0 Privilege: System Boundary

Name	Description	Unit	Measure	Type	Range	S	D	U	K	Prep	Indices
Indices	Dimension indices for sampled variables.			INTEGER	0-32767	Y	2	N	N		Variables
Iterations	Set of iterations that have been sampled			INTEGER		Y	1	N	N	for	
Set	Data set name for sampled variables.			STRING		N	1	N	N		Variables
ShortName	Alias name for sampled variables			STRING		N	1	N	N		Variables
Value	Sampled value of variable			FLOAT		N	2	N	N		Variables Iterations
Variables	Variable names that have been sampled			STRING		Y	1	N	N	for	
VarUnits	Units for sampled variables			STRING		N	1	N	N		Variables

Legend:

Column Name	**Meaning**
S	Self-Indexed
D	Dimensional Size
U	Uncertainty can apply (Stochastic)
K	Is the variable a key to others

5.0 Test Sequence

Requirements for linking FRAMES to the Statistical Package R are summarized in the Requirements section of this document and are succinctly presented in Table 5.1. The requirements are written as concise, fundamental, testable requirements.

Table 5.1. Summary of Requirements for Linking FRAMES to the Statistical Package R

Requirement Number	Requirement
1	When sampled input and selected output are generated by the same icon, as in the FRAMES-1.x Sensitivity/Uncertainty icon, then the viewer can be described as a right-click viewer and produced by a right-click on the icon menu. This is currently a FRAMES-1.x requirement.
2	When the sampled input and selected output are generated by different icons, as in the FRAMES-2.0 Sampler and Summarizer icons, respectively, then the viewer is described by a Drag&Drop icon and connected to the icons producing the sampled input and selected output. This is currently a FRAMES-2.0 requirement.
3	The R results produce a scatter plot in png format of all the sampled input and selected output.
4	For each selected output, a regression analysis is performed that summarizes the regression in the R standard form, whose results are provided in a standard Text (*.txt) file.
5	For each selected output, a regression analysis is performed that shows the contribution of each sampled input to the Multiple R-squared, as reported in the R standard summary in a standard Text (*.txt) file, capturing the degree that each sampled input has on influencing the variability of each selected output.
6	A viewer is produced that provides a CDF, using R's ecdf plotting capability, for each sampled input and selected output.
7	Each CDF plot contains a line at the 95th percentile for the user's reference.
8	A viewer is produced that provides a histogram, using R's plotting capability, for each sampled input and selected output.

To confirm that the software meets the requirements listed in Table 5.1, the following test cases were developed to confirm the performance for linking FRAMES to the Statistical Package R. Table 5.2 presents the relationships between these requirements and the test cases, which are described below.

Table 5.2. Relationship Between Test Cases and Fundamental Requirements for Linking FRAMES to the Statistical Package R

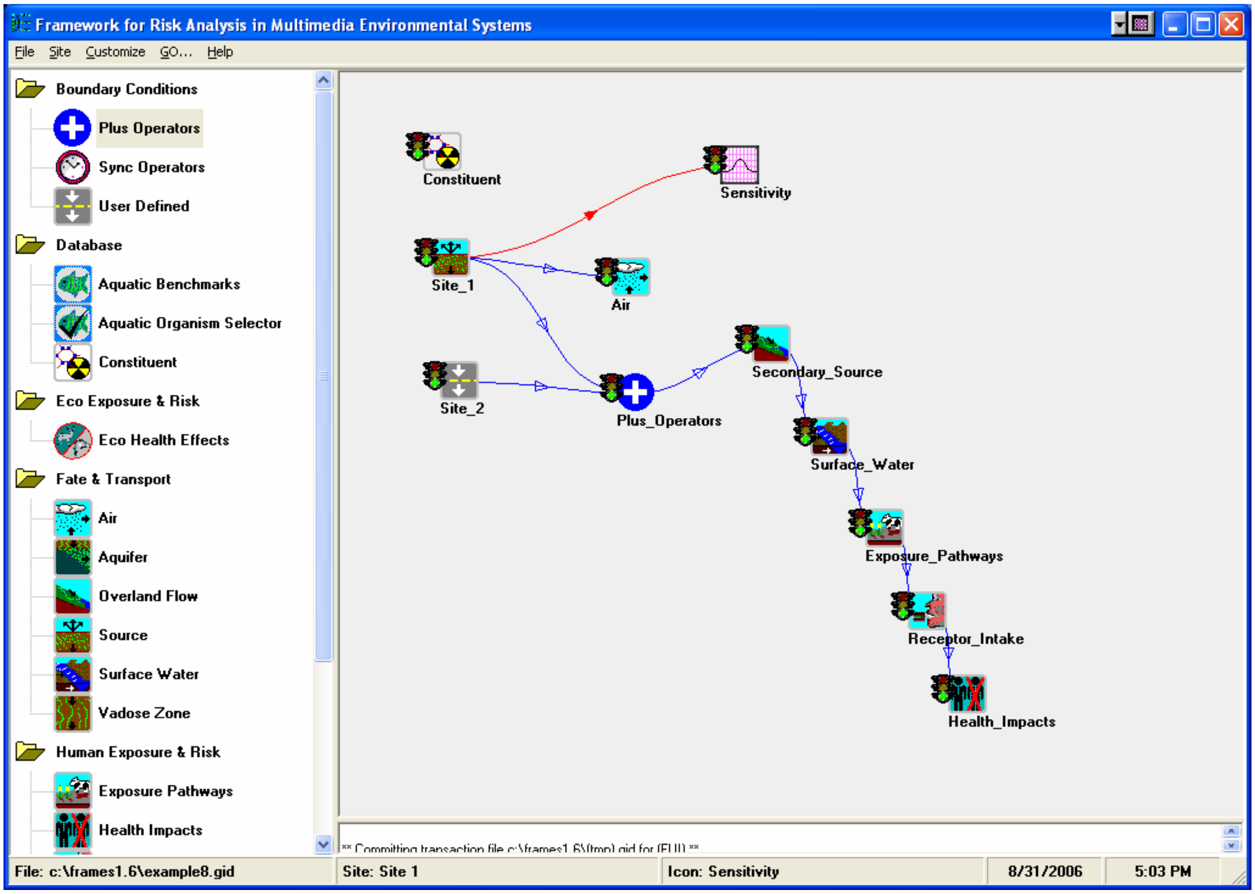
	Test Case Number			
Requirement	**01**	**02**	**03**	**04**
1	X			
2		X		
3	X	X		
4	X	X		
5	X	X		
6	X	X		
7	X	X		
8	X	X		

5.1 FRAMES-1.x: Test Case Example8

5.1.1 Description

The purpose of this test case is to test the general functionality of linking FRAMES-1.x to the Statistical Package R, assuming that the sampled input and selected output are generated by the same icon, where a right-click viewer can be employed to view the output results. Within the FRAMES-1.x system, the Sensitivity/Uncertainty (S/U) module in FRAMES-1.x generates both sampled input and selected output regardless of the number of icons to which it is linked. It generates the sampled inputs by consuming the various FRAMES-1.x file types to which it is linked in the Conceptual Site Model (CSM). The S/U module produces the Sensitivity/Uncertainty File (SUF) type, which generates the selected output. The FRAMES-1.x file types and their specifications can be found at http://mepas.pnl.gov:2080/earth/.

This test case examines and tests Requirements 1 and 3 through 8. Figure 5.1 presents the Conceptual Site Model, associated with Example 8. Both the sampled input and the selected output are generated within the Sensitivity module from which these data are associated with the Site_1 source-term model. The Sensitivity model contains the R statistical package as a series of right-click viewers, corresponding to SP, CDF, and Histogram graphs and a SP statistical summary text file.



```
** Committing transaction file c:\frames1.6\tmp\.gid for (FUI) **
```

File: c:\frames1.6\example8.gid | Site: Site 1 | Icon: Sensitivity | 8/31/2006 | 5:03 PM

Figure 5.1. CSM Testing the Fundamental Requirements of 1 and 3 through 8 for Linking FRAMES to the Statistical Package R

5.1.2 Input Data

The input data for this test case are documented in Example8.gid. Two Site_1 input variables (i.e., Bulk Density and Inventory) and two Site_1 output variables (i.e., adsorbed constituent fluxes at years 1 and 10) are identified in the Sensitivity module. The Random Seed Value equals unity, and the number of iterations equals 30.

5.1.3 Expected Results

The following results are expected:

1. All viewers will be accessed as right-click viewers from the Sensitivity/Uncertainty icon (i.e., “Sensitivity” in Figure 5.1).
2. The R results will produce a scatter plot in png format of all the sampled input and selected output.
3. For each selected output, a regression analysis will summarize the regression in the R standard form, whose results are provided in a standard Text (*.txt) file.

4. For each selected output, a regression analysis will show the contribution of each sampled input to the Multiple R-squared, as reported in the R standard summary in a standard Text (*.txt) file, capturing the degree that each sampled input has on influencing the variability of each selected output.
5. A viewer will be produced that provides a CDF, using R's ecdf plotting capability, for each sampled input and selected output.
6. Each CDF plot will contain a line at the 95th percentile for the user's reference.
7. A viewer will be produced that provides a histogram, using R's plotting capability, for each sampled input and selected output.

5.1.4 Conducting the Test

To proceed with this test, it is assumed that all of the required FRAMES, R Project (2006), and related software (e.g., MEPAS) software are loaded onto the user's computer. The test bed files must also be installed to follow this test plan and execute the tests.

1. Open the test case Global Input Data (gid) file, titled Example8.gid.
2. Re-run each module, using the "Go" button on the menu list.
3. Re-run the "Sensitivity" module.
4. Right click on the "Sensitivity" module, and choose "View/Print Module Output."
5. In three independent and sequential actions, choose the SUF R Scatter Plot Viewer, SUF R CDF Viewer, and SUF R Histogram Viewer.
6. The general approach is to read data from the sampled input or selected output dataset and then write a text file that can be directly read by R. A file that contains the "R commands" is also written at the same time. R is then invoked with the command file. When R's invocation is completed, any graphics that are produced are rendered in a web browser, and any text results are rendered in the user's default text editing program. For most users, this is Windows' Notepad. From there, the text could be copied and pasted into documentation.

5.1.5 Results

Using R, three viewers have been developed: Scatter Plot (SP), Cumulative Distribution Function (CDF), and Histogram. These viewers allow a user to observe the relationships of the inputs to outputs or simply find a value from a distribution for a given probability. The Scatter Plot (SP) viewer constructs a scatter plot that graphs all input variables against all other input variables and all output variables. R's default approach is used to generate a scatter plot. The SP performs an optimized regression analysis using R's linear model and stepwise capability. It performs this analysis once for each output variable that is provided in the SUF dataset. The development of the CDF plot uses R's capability to plot CDFs of values, which allows a user to find a given input or output value for given a percentile. A 95th percentile line is added for reference into the CDF viewer's output. The histogram plot provides the frequency of occurrence versus all independent and dependent variables chosen by the user for the statistical analysis.

When graphics are produced, they are created in png format and are then shown to the user through a web browser. Four sets of results are provided when the user chooses the following three viewers: SUF R Scatter Plot Viewer, SUF R CDF Viewer, and SUF R Histogram Viewer. Two sampled input variables are Bulk Density and Inventory, and the two selected output variables are adsorbed constituent fluxes at years 1 and 10.

1. SUF R Scatter Plot Viewer
    a. Figure 5.2 presents Scatter Plots relating Bulk Density (Bd), Inventory (I), and adsorbed constituent fluxes at years 1 and 10 (Ca_All_Surface_Water_1 and Ca_All_Surface_Water_10, respectively)
    b. Figure 5.3 presents the summarized regression analysis in the R standard form for each selected output, showing the contribution of each sampled input to the multiple R-squared and capturing the degree that each sampled input has on influencing the variability of each selected output. This is presented as a standard Text (*.txt) file. Figure 5.4 illustrates the contributions of each sampled input to the multiple R-squared and captures the degree that each sampled input has on influencing the variability of the selected output Ca_All_Surface_Water_1.
2. SUF R CDF Viewer—Figure 5.5 presents CDF plots of sampled input and selected output, including a 95th percentile line that is added for reference into the CDF viewer’s output.
3. SUF Histogram Viewer—Figure 5.6 presents histograms of sampled input and selected output, including a 95th percentile line that is added for reference into the CDF viewer’s output.

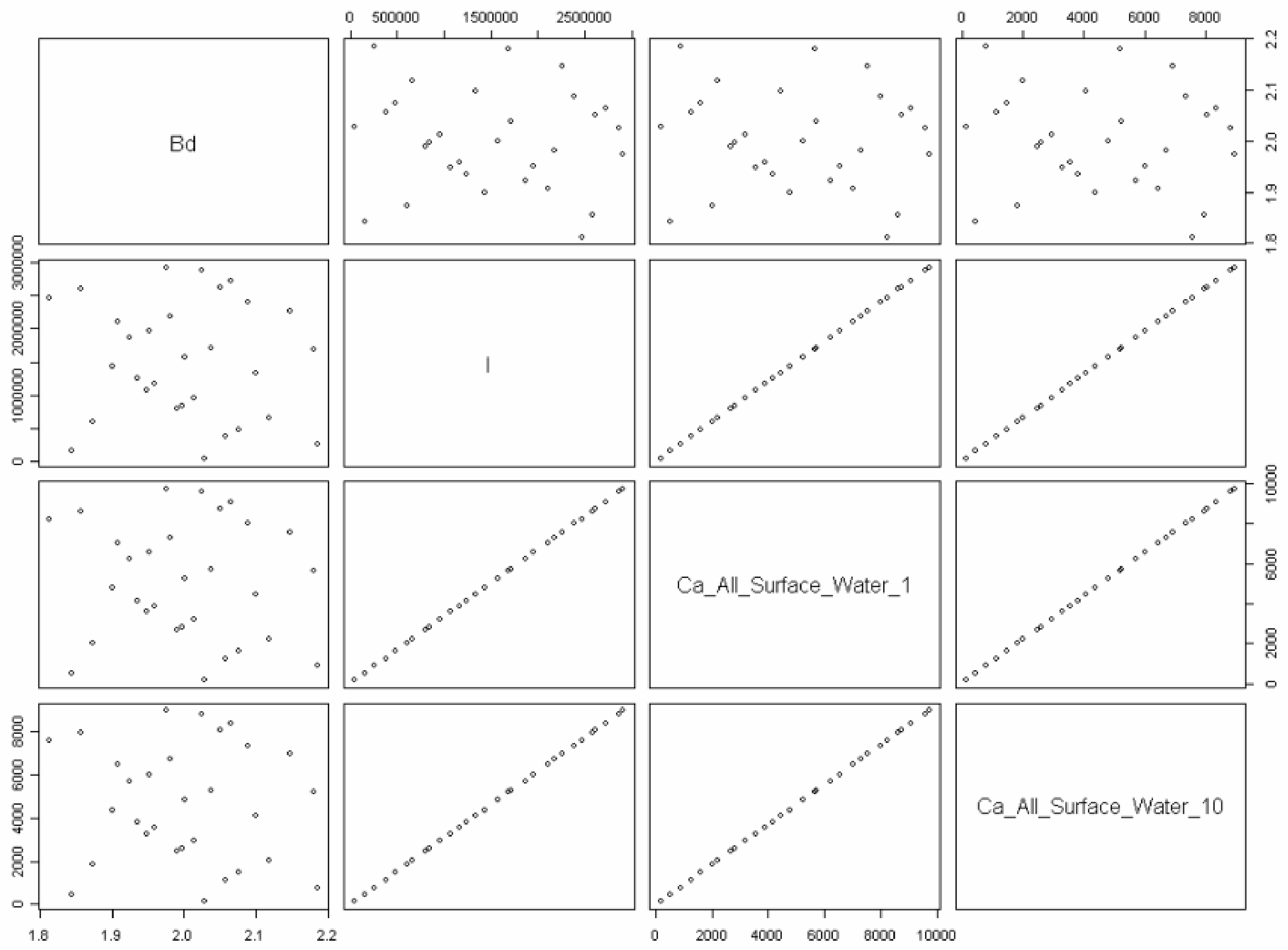


Figure 5.2. Scatter Plots Relating Bulk Density (Bd), Inventory (I), and Adsorbed Constituent Fluxes at Years 1 and 10 (Ca_All_Surface_Water_1 and Ca_All_Surface_Water_10, respectively)

[1] Partial R results

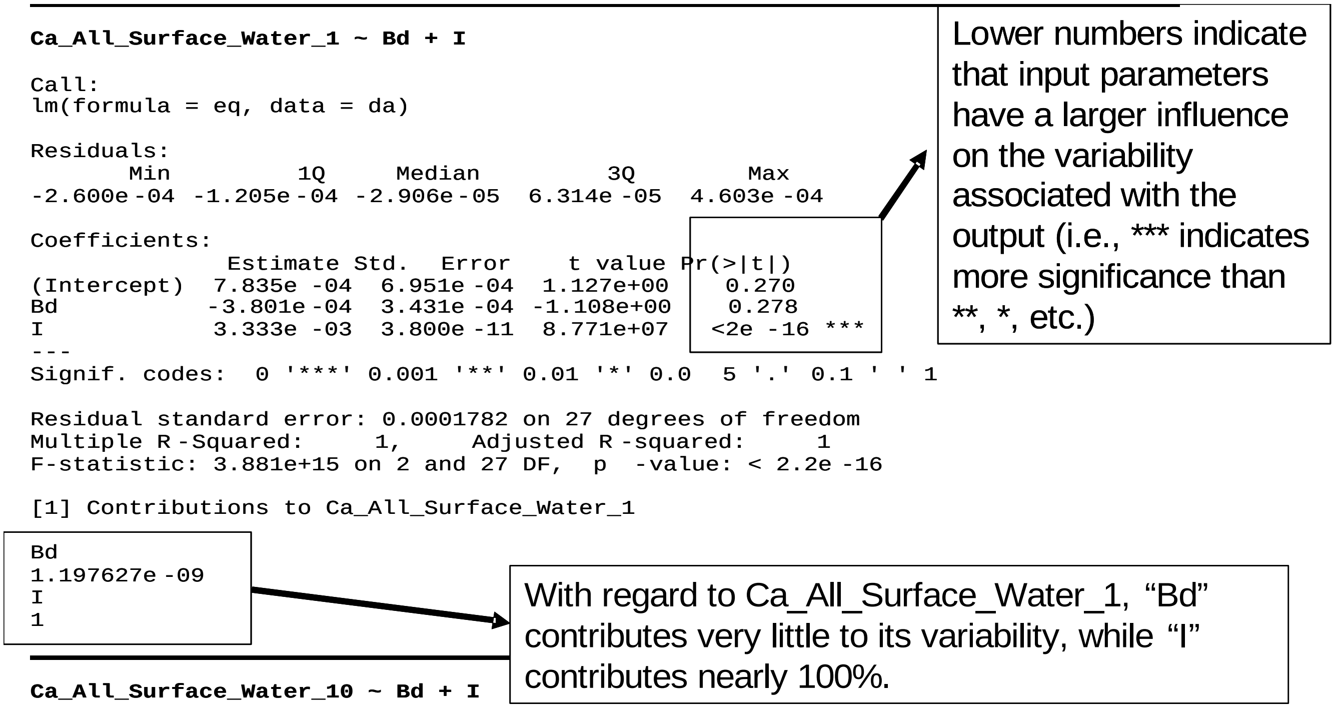
Ca_All_Surface_Water_1 ~ Bd + I

```
Call:
lm(formula = eq, data = da)

Residuals:
       Min         1Q     Median         3Q        Max
-2.600e-04 -1.205e-04 -2.906e-05  6.314e-05  4.603e-04

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  7.835e-04  6.951e-04  1.127e+00    0.270
Bd          -3.801e-04  3.431e-04 -1.108e+00    0.278
I            3.333e-03  3.800e-11  8.771e+07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0001782 on 27 degrees of freedom
Multiple R-Squared:     1,    Adjusted R-squared:     1
F-statistic: 3.881e+15 on 2 and 27 DF,  p-value: < 2.2e-16

[1] Contributions to Ca_All_Surface_Water_1

Bd
1.197627e-09
I
1
```

Ca_All_Surface_Water_10 ~ Bd + I

```
Call:
lm(formula = eq, data = da)

Residuals:
       Min         1Q     Median         3Q        Max
-7.449e-04 -1.538e-04 -4.363e-06  1.275e-04  8.173e-04

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept) -2.949e+01  1.290e-03 -2.286e+04  < 2e-16 ***
Bd           4.863e-03  6.367e-04  7.638e+00 3.25e-08 ***
I            3.082e-03  7.052e-11  4.370e+07  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0003307 on 27 degrees of freedom
Multiple R-Squared:     1,    Adjusted R-squared:     1
F-statistic: 9.636e+14 on 2 and 27 DF,  p-value: < 2.2e-16

[1] Contributions to Ca_All_Surface_Water_10

Bd
-1.657367e-08
I
1
```

Figure 5.3. Summarized Regression Analysis in the R Standard Form for each Selected Output, Showing the Contribution of each Sampled Input to the Multiple R-Squared and Capturing the Degree that each Sampled Input has on Influencing the Variability of each Selected Output



```
Ca_All_Surface_Water_1 ~ Bd + I

Call:
lm(formula = eq, data = da)

Residuals:
       Min         1Q     Median         3Q        Max
-2.600e-04 -1.205e-04 -2.906e-05  6.314e-05  4.603e-04

Coefficients:
              Estimate Std.  Error    t value Pr(>|t|)
(Intercept)  7.835e-04  6.951e-04  1.127e+00    0.270
Bd          -3.801e-04  3.431e-04 -1.108e+00    0.278
I            3.333e-03  3.800e-11  8.771e+07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0  5 '.' 0.1 ' ' 1

Residual standard error: 0.0001782 on 27 degrees of freedom
Multiple R-Squared:     1,     Adjusted R-squared:     1
F-statistic: 3.881e+15 on 2 and 27 DF,  p  -value: < 2.2e-16

[1] Contributions to Ca_All_Surface_Water_1

Bd
1.197627e-09
I
1

Ca_All_Surface_Water_10 ~ Bd + I
```

Lower numbers indicate that input parameters have a larger influence on the variability associated with the output (i.e., *** indicates more significance than **, *, etc.)

With regard to Ca_All_Surface_Water_1, “Bd” contributes very little to its variability, while “I” contributes nearly 100%.

Figure 5.4. Contributions of each Sampled Input to the Multiple R-Squared and Capturing the Degree that each Sampled Input has on Influencing the Variability of the Selected Output Ca_All_Surface_Water_1

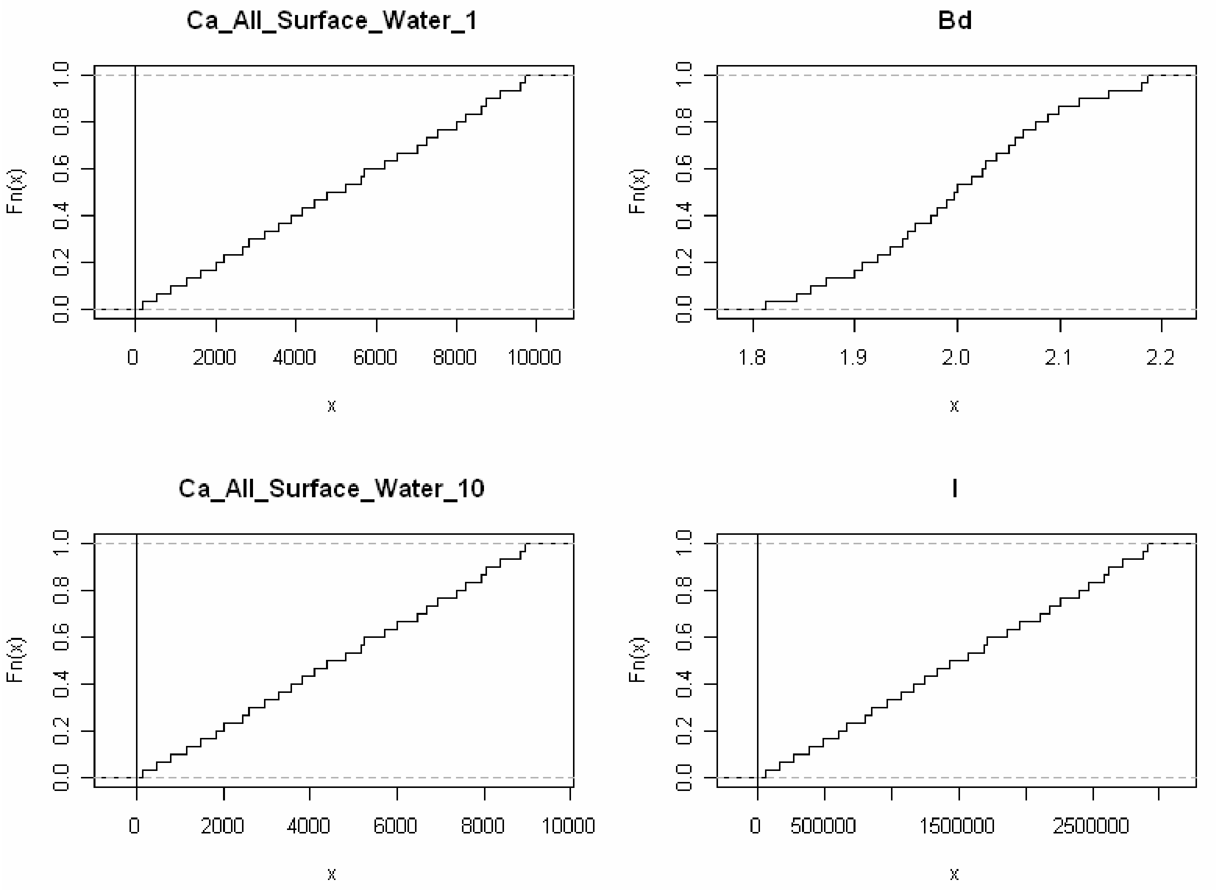


Figure 5.5. CDF Plots of Sampled Input and Selected Output, Including a 95th Percentile Line that is Added for Reference into the CDF Viewer's Output

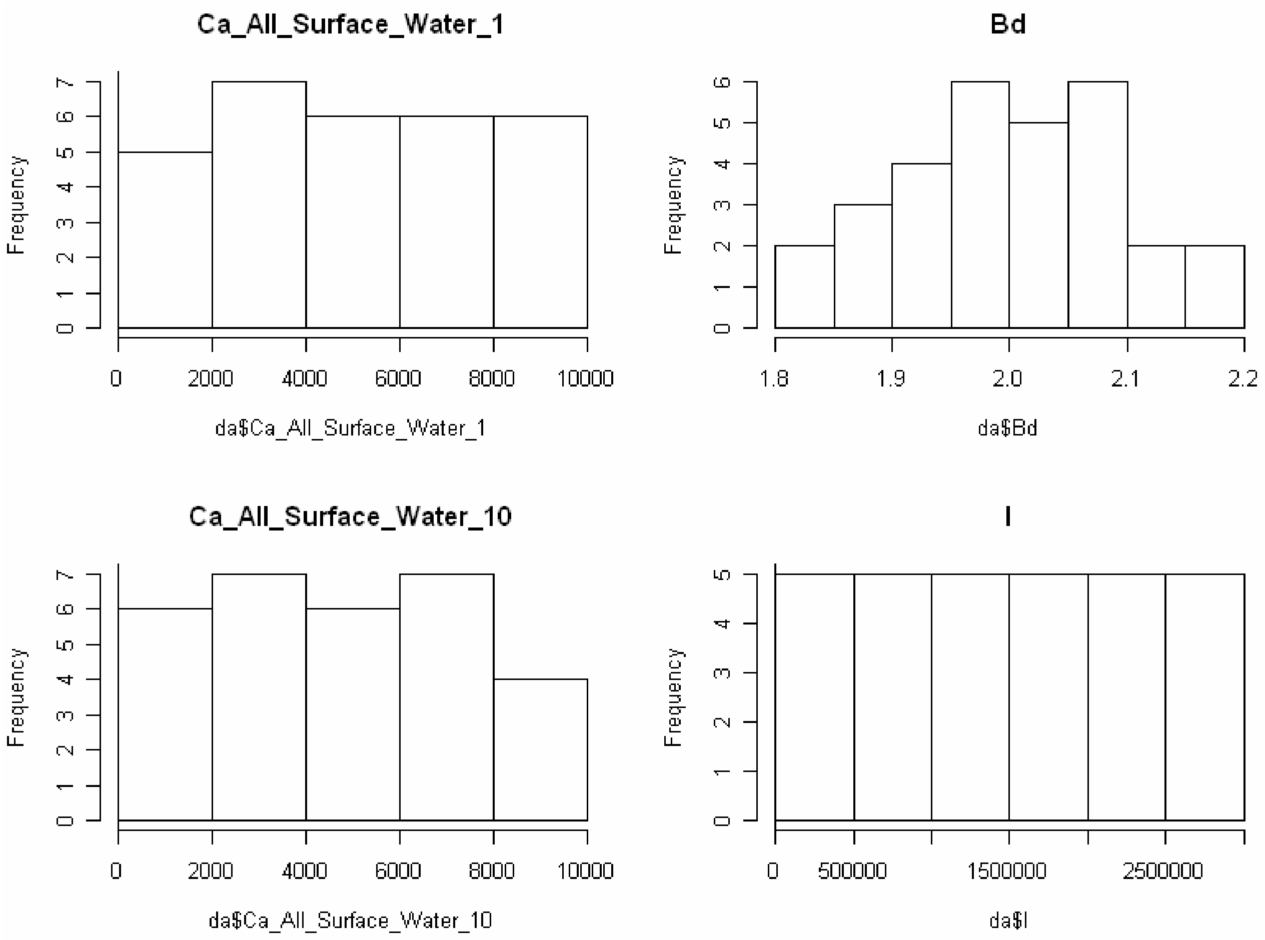


Figure 5.6. Histograms of Sampled Input and Selected Output, Including a 95th Percentile Line that is Added for Reference into the CDF Viewer's Output

5.2 FRAMES-2.0: Test Case SUTEST

5.2.1 Description

The purpose of this test case is to test the general functionality of linking FRAMES to the Statistical Package R, assuming that the SampledValues and SummaryValues are generated by different icons (i.e., models), where a drag-and-drop module can be employed to view the output results. This test case examines and tests Requirements 2 through 8. Figure 5.7 presents the Conceptual Site Model, associated with SUTEST. SampledValues and SummaryValues are associated with Sampler (Mod1) and Output Filter (Mod3), respectively. Viewer (Mod5) contains the R statistical package, providing SP, CDF, and Histogram viewers and a SP statistical summary test file.

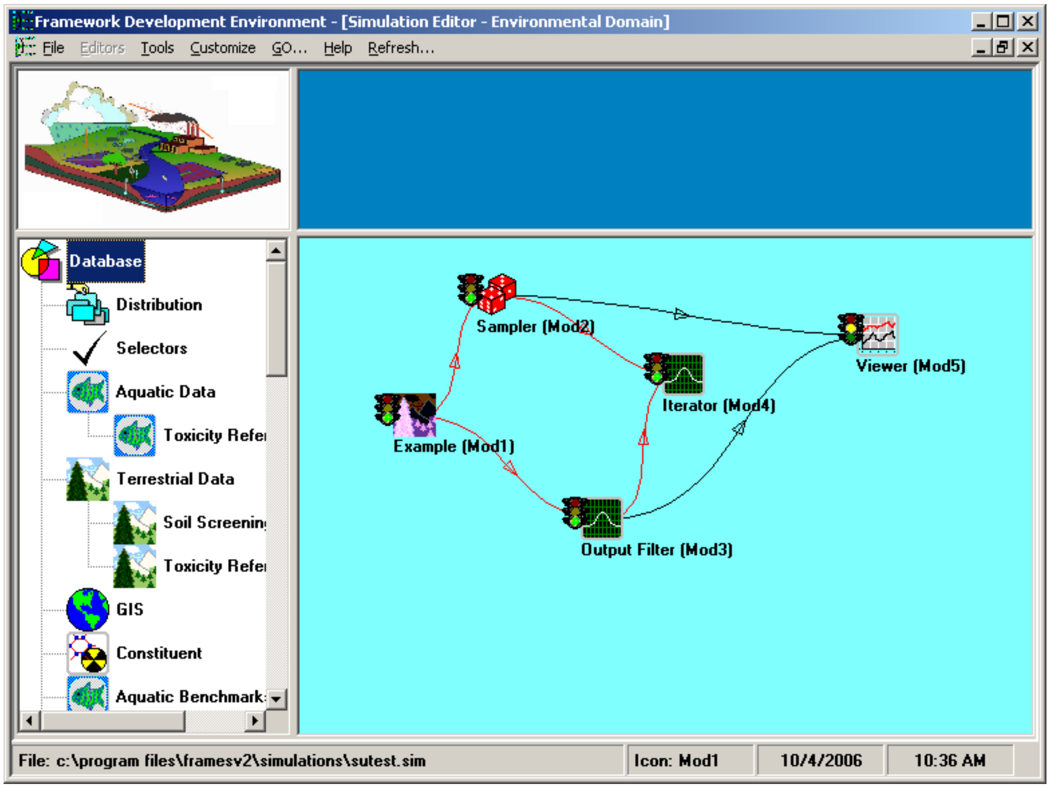


Figure 5.7. CSM Testing the Fundamental Requirements of 2 Through 8 for Linking FRAMES to the Statistical Package R

5.2.2 Input Data

The input data for this test case are documented in SUTEST.gid. Two SampledValues (i.e., “A” and “B”) and two SummaryValues (i.e., “O1” and “O2” for Output 1 and Output 2) are identified. A simple model is being used to test the R project as a drag-and-drop viewer. This is a model that computes O = A/B + C + D. C and D are vectors, so they can contain more than one value. In addition to the initial values for the SampleValues of A and B, the initial values for C and D associated with the module “Example (Mod1)” in Figure 5.7 are as follows:

- A = 1
- B = 2
- C= <1,2>
- D = <1,2>

The module “Sample (Mod2)” in Figure 5.7 randomly samples input parameters from the module “Example (Mod1).” The user-defined input variables with their statistical information are as follows:

- A: Normal distribution, mean of 9, standard deviation of 3, minimum of 0, and maximum of 18.
- B: Uniform distribution, minimum of 1, and maximum of 20.
- The Random Seed Value equals unity.

The module "Output Filter (Mod3)" chooses the variables whose data are collected following each realization, so the user-chosen input variables can be correlated to the user-chosen output variables. The output variables in this example are represented by the SummaryValues of "O1" and "O2."

The module "Iterator (Mod4)" implements the number of iterations defined by the user. The user-defined value for the number of iterations equals 100.

5.2.3 Expected Results

The following results are expected:

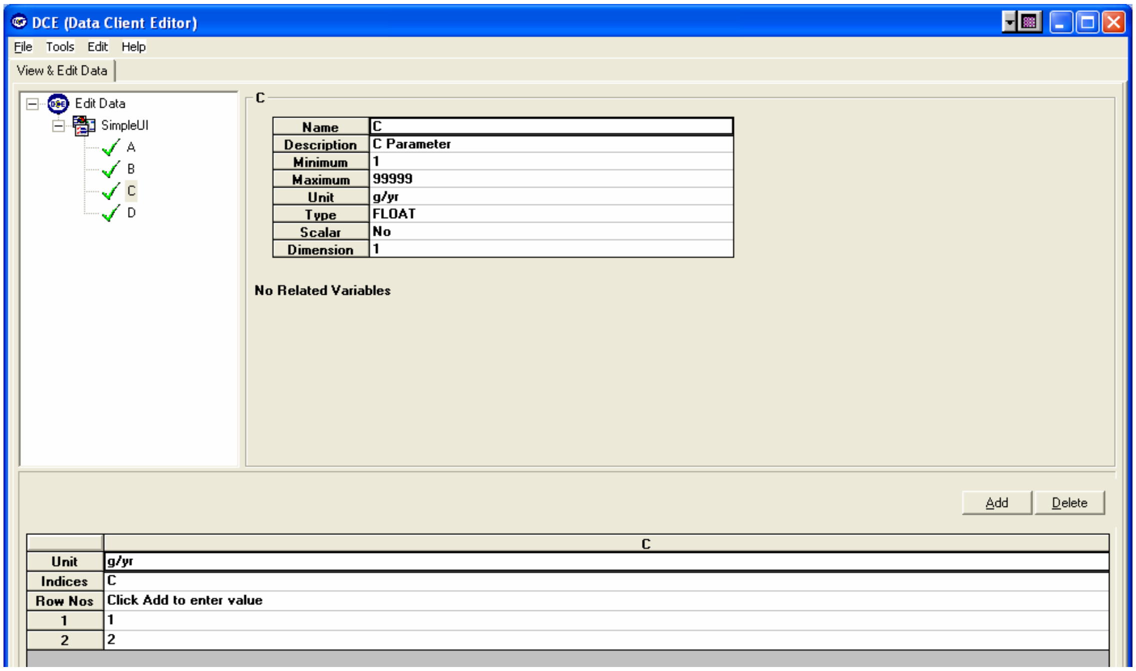
1. All viewers will be accessed as a drag-and-drop module from the Viewer's icon (i.e., "Viewer (Mod5)" in Figure 5.7).
2. The R results will produce a scatter plot in png format of the all SampledValues and SummaryValues.
3. For each SummaryValue, a regression analysis will summarize the regression in the R standard form, whose results are provided in a standard Text (*.txt) file.
4. For each of the SummaryValues, a regression analysis will show the contribution of each of the SampledValues to the Multiple R-squared, as reported in the R standard summary in a standard Text (*.txt) file, capturing the degree that each of the SampledValues has on influencing the variability of each of the SummaryValues.
5. A viewer will be produced that provides a CDF, using R's ecdf plotting capability, for each of the SampledValues and SummaryValues.
6. Each CDF plot will contain a line at the 95th percentile for the user's reference.
7. A viewer will be produced that provides a histogram, using R's plotting capability, for each of the SampledValues and SummaryValues.

5.2.4 Conducting the Test

This test case expects that R-Project 2.3.1 has been installed. R is used as a statistical analysis tool to visualize the output of two of the three components. It is available at http://www.r-project.org.

1. Start with a new simulation.
2. From the environmental domain, add the Example from the Model group.
3. Also add from the System group three sensitivity icons.
4. Connect them together as shown in Figure 5.7.
5. Add a viewer to the diagram with the connection scheme illustrated in Figure 5.7.
6. Name the sensitivity components Sampler, Iterator, and Output Filter.
7. For the Example model, choose the SimpleUI by right clicking and choosing General Information. This is a model that computes O=A/B+C+D. C and D are vectors, so they can contain more than one value.
8. For the sampler choose Sampler. If Sampler is not available, you may need to register it in the Domain editor as being part of the environmental domain.
9. For the Output Filter, choose the Summarizer.
10. For the Iterator, choose Iterator.

11. For the Viewer, choose PartialR.
12. Complete the user input for the Example (Mod1) icon. A=1.0, B=2.0, C={1,2}, D={2,3}. The “Add” button is used to add values to the variable. Figure 5.8 illustrates a completed process for populating variable C.
13. Complete the input for the Sampler. The variables “A” and “B” are aliased as “A” and “B,” respectively. See Figure 5.9.
14. Next, define the distribution for “A” as Normal with a mean of 9, standard deviation of 3, minimum of 0, and maximum of 18, as illustrated in Figure 5.10. Set the seed to the Random Number Generator to “1.”
15. Define the distribution for “B” as Uniform with a minimum of 1 and a maximum of 20.
16. Complete the input for the Output Filter (Mod3). Add an alias for Output named “O1” that has an index of 1, as illustrated in Figure 5.11. This will grab just the first number of the vector Output.
17. Add a second alias for Output named “O2” that has an index of 2.
18. In Output Filter (Mod3), choose every value for summaries of O1 and O2, which capture every value for the statistical analyses using R Project, as illustrated by Figure 5.12.
19. Complete the input for the Iterator (Mod4) module by setting the last iteration to 100.
20. Run Example (Mod1).
21. Run Sampler (Mod2).
22. Run Output Filter (Mod3).
23. Run the Iterator (Mod4). This may take a while as the iterations occur.
24. Press Save in the File menu. Use the simulations directory to store your simulation. Use the name sutest.sim.
25. Open Simulations\sutestsim.Mod2.SampledValues, and the output should be identical to Figure 5.13 when opened in Textpad.
26. SummaryValues are documented in \Simulations\sutest.sim.Mod3.SummaryValues, as illustrated in Figure 5.14.



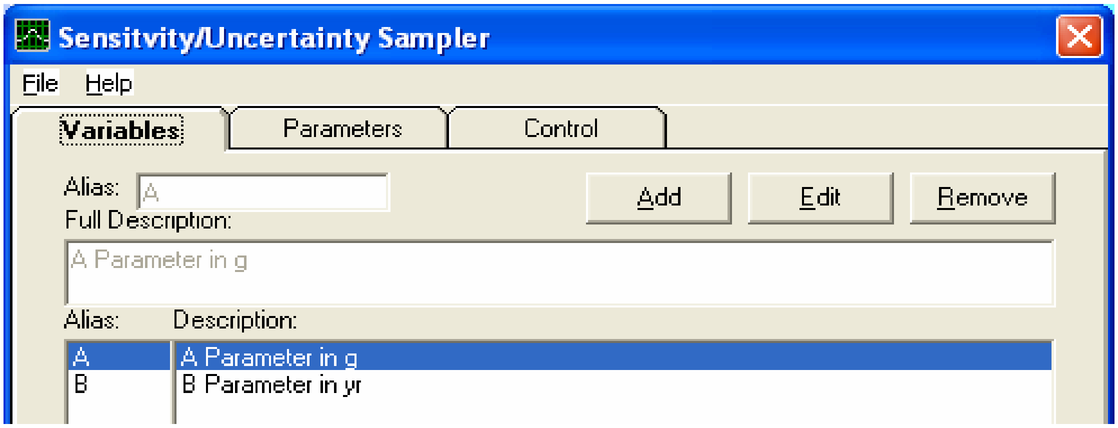
C

Name	C
Description	C Parameter
Minimum	1
Maximum	99999
Unit	g/yr
Type	FLOAT
Scalar	No
Dimension	1

No Related Variables

	C
Unit	g/yr
Indices	C
Row Nos	Click Add to enter value
1	1
2	2

Figure 5.8. The Data Client Editor (i.e., User Interface) Populated Values for Input Variable C in the Example (Mod1) Module



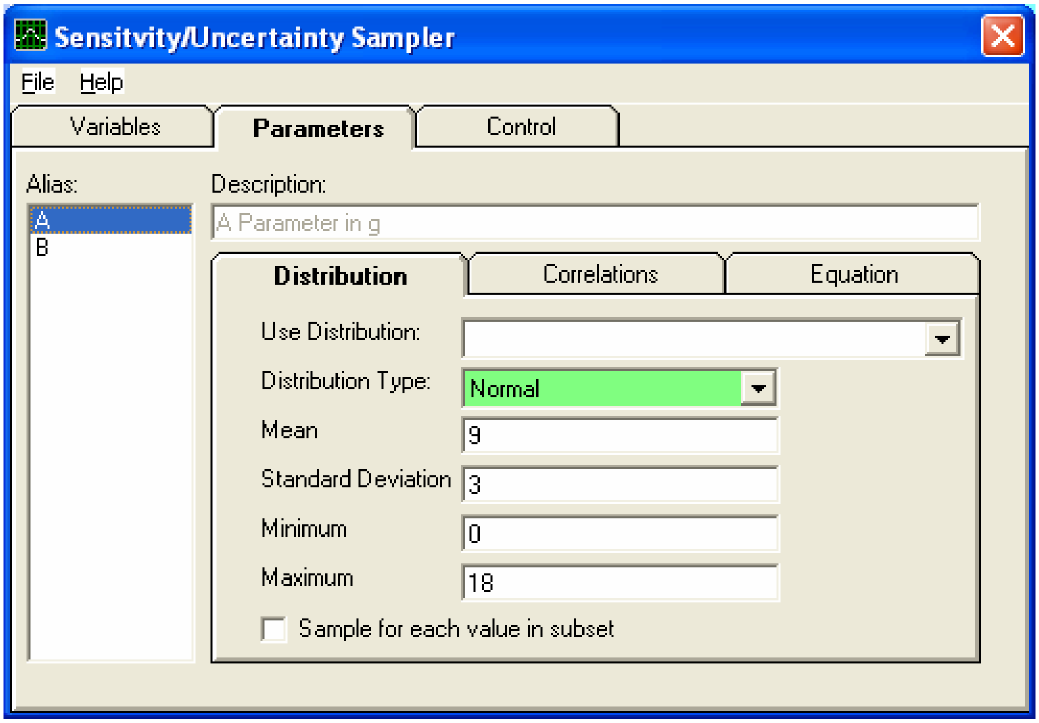
Alias: A

Full Description:

A Parameter in g

Alias:	Description:
A	A Parameter in g
B	B Parameter in yr

Figure 5.9. Sampler (Mod2) User Interface Illustrating the Aliasing of Parameters "A" and "B" as A and B, Respectively, in the Sampler (Mod2) Module



Alias:

A
B

Description: A Parameter in g

Use Distribution:

Distribution Type: Normal

Mean: 9

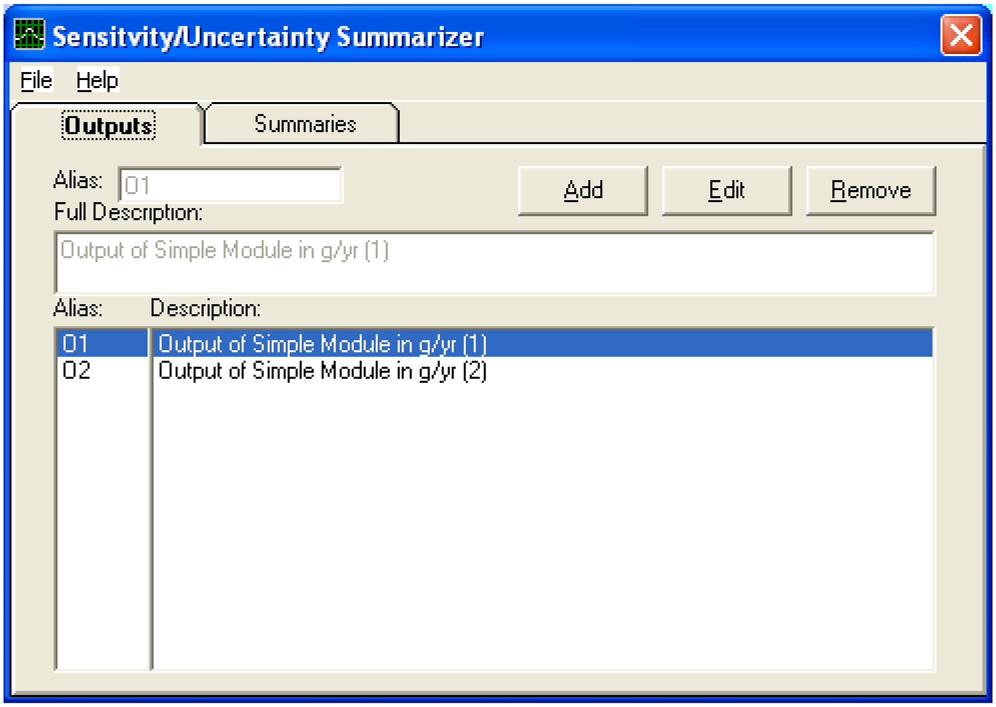
Standard Deviation: 3

Minimum: 0

Maximum: 18

☐ Sample for each value in subset

Figure 5.10. Defining the Statistical Information Associated with Variable “A” in the Sampler (Mod2) Module

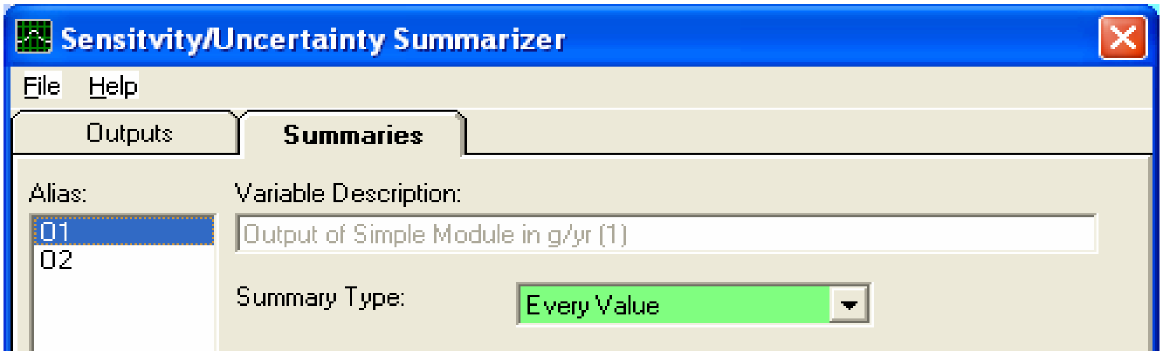


Alias: O1

Full Description: Output of Simple Module in g/yr (1)

Alias:	Description:
O1	Output of Simple Module in g/yr (1)
O2	Output of Simple Module in g/yr (2)

Figure 5.11. Aliasing the Output Parameters “O1” and “O2” in the Output Filter (Mod3) Module

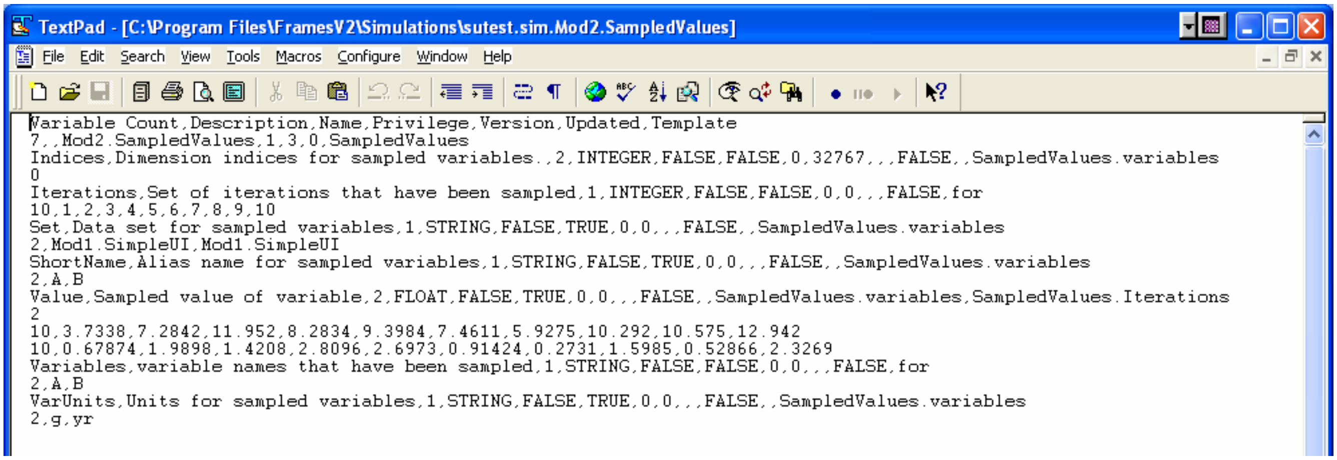


Alias:
O1
O2

Variable Description: Output of Simple Module in g/yr (1)

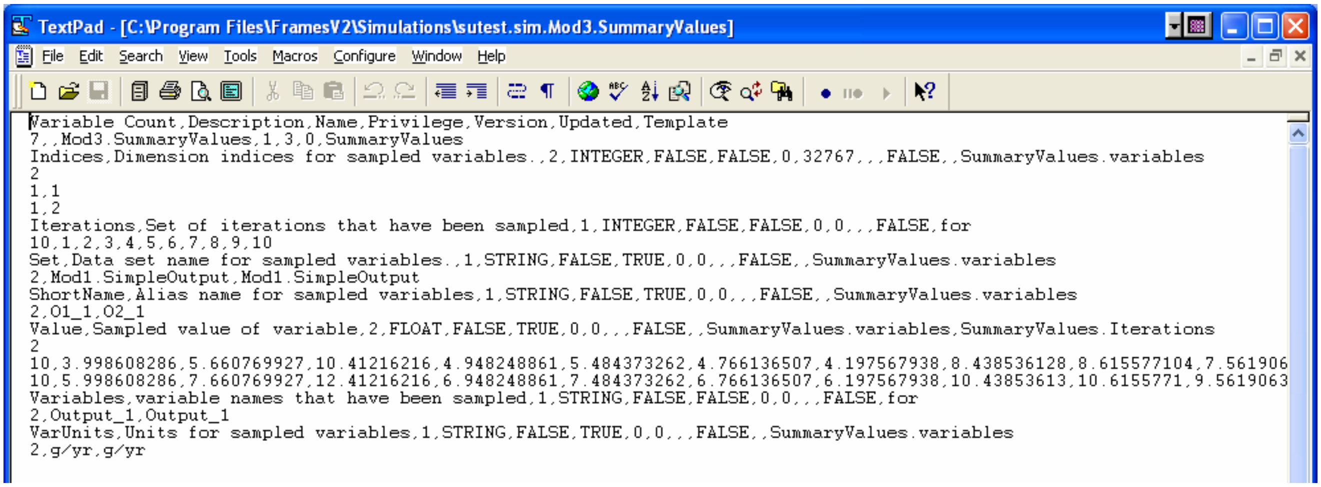
Summary Type: Every Value

Figure 5.12. Choose Every Value for Summaries of O1 and O2, Which Captures Every Value for the Statistical Analyses Using R Project in the Output Filter (Mod3) Module


TextPad - [C:\Program Files\FramesV2\Simulations\sutest.sim.Mod2.SampledValues]
File
Edit
Search
View
Tools
Macros
Configure
Window
Help

```
Variable Count,Description,Name,Privilege,Version,Updated,Template
7,,Mod2.SampledValues,1,3,0,SampledValues
Indices,Dimension indices for sampled variables.,2,INTEGER,FALSE,FALSE,0,32767,,,FALSE,,SampledValues.variables
0
Iterations,Set of iterations that have been sampled,1,INTEGER,FALSE,FALSE,0,0,,,FALSE,for
10,1,2,3,4,5,6,7,8,9,10
Set,Data set for sampled variables,1,STRING,FALSE,TRUE,0,0,,,FALSE,,SampledValues.variables
2,Mod1.SimpleUI,Mod1.SimpleUI
ShortName,Alias name for sampled variables,1,STRING,FALSE,TRUE,0,0,,,FALSE,,SampledValues.variables
2,A,B
Value,Sampled value of variable,2,FLOAT,FALSE,TRUE,0,0,,,FALSE,,SampledValues.variables,SampledValues.Iterations
2
10,3.7338,7.2842,11.952,8.2834,9.3984,7.4611,5.9275,10.292,10.575,12.942
10,0.67874,1.9898,1.4208,2.8096,2.6973,0.91424,0.2731,1.5985,0.52866,2.3269
Variables,variable names that have been sampled,1,STRING,FALSE,FALSE,0,0,,,FALSE,for
2,A,B
VarUnits,Units for sampled variables,1,STRING,FALSE,TRUE,0,0,,,FALSE,,SampledValues.variables
2,g,yr
```

Figure 5.13. SampledValues Summary as Documented in \Simulations\sutestsim.Mod2.SampledValues


TextPad - [C:\Program Files\FramesV2\Simulations\sutest.sim.Mod3.SummaryValues]
File
Edit
Search
View
Tools
Macros
Configure
Window
Help

```
Variable Count,Description,Name,Privilege,Version,Updated,Template
7,,Mod3.SummaryValues,1,3,0,SummaryValues
Indices,Dimension indices for sampled variables.,2,INTEGER,FALSE,FALSE,0,32767,,,FALSE,,SummaryValues.variables
2
1,1
1,2
Iterations,Set of iterations that have been sampled,1,INTEGER,FALSE,FALSE,0,0,,,FALSE,for
10,1,2,3,4,5,6,7,8,9,10
Set,Data set name for sampled variables.,1,STRING,FALSE,TRUE,0,0,,,FALSE,,SummaryValues.variables
2,Mod1.SimpleOutput,Mod1.SimpleOutput
ShortName,Alias name for sampled variables,1,STRING,FALSE,TRUE,0,0,,,FALSE,,SummaryValues.variables
2,O1_1,O2_1
Value,Sampled value of variable,2,FLOAT,FALSE,TRUE,0,0,,,FALSE,,SummaryValues.variables,SummaryValues.Iterations
2
10,3.998608286,5.660769927,10.41216216,4.948248861,5.484373262,4.766136507,4.197567938,8.438536128,8.615577104,7.561906
10,5.998608286,7.660769927,12.41216216,6.948248861,7.484373262,6.766136507,6.197567938,10.43853613,10.6155771,9.5619063
Variables,variable names that have been sampled,1,STRING,FALSE,FALSE,0,0,,,FALSE,for
2,Output_1,Output_1
VarUnits,Units for sampled variables,1,STRING,FALSE,TRUE,0,0,,,FALSE,,SummaryValues.variables
2,g/yr,g/yr
```

Figure 5.14. SummaryValues Summary as Documented in \Simulations\sutest.sim.Mod3.SummaryValues

5.2.5 Results

Using R, three viewers have been developed: Scatter Plot (SP), Cumulative Distribution Function (CDF), and Histogram. These viewers allow a user to observe the relationships of the inputs to outputs or simply find a value from a distribution for a given probability. The Scatter Plot (SP) viewer constructs a scatter plot that graphs all input variables against all other input variables and all output variables. R's default approach is used to generate a scatter plot. The SP performs an optimized regression analysis using R's linear model and stepwise capability. It performs this analysis once for each output variable that is provided in the SummaryValues dataset. The development of the CDF plot uses R's capability to plot CDFs of values, which allows a user to find a given input or output value for a given percentile. A 95th percentile line is added for reference into the CDF viewer's output. The histogram plot provides the frequency of occurrence versus all independent and dependent variables chosen by the user for the statistical analysis. Two SampledValues are "A" and "B", and two SummaryValues are "O1" and "O2."

When graphics are produced, they are created in png format and are then shown to the user through a web browser. Four sets of results are provided when the user chooses the following three viewers:

1. PartialR Scatter Plot Viewer and Partial R Summary Text File
   a. Figure 5.15 presents Scatter Plots, relating inputs "A" and "B" to outputs "O1" and O2 (i.e., O1.1 and O2.1).
   b. Figure 5.16 presents the summarized regression analysis in the R standard form for each of the SummaryValues "O.1" and "O.2," showing the contribution of each of the SampledValues "A" and "B" to the multiple R-squared and capturing the degree that each of the SampledValues has on influencing the variability of each of the SummaryValues. This is presented as a standard Text (*.txt) file. Figure 5.17 illustrates the contributions of each of the SampledValues to the multiple R-squared, capturing the degree that each of the SampledValues has on influencing the variability of the SummaryValues "O.1."
2. CDF: SUF R Cumulative Distribution Function (CDF) Viewer—Figure 5.18 presents CDF plots of SampledValues and SummaryValues, including a 95th percentile line that is added for reference into the CDF viewer's output.
3. Hist R Histogram Viewer—Figure 5.19 presents histograms of SampledValues and SummaryValues, including a 95th percentile line that is added for reference into the CDF viewer's output.

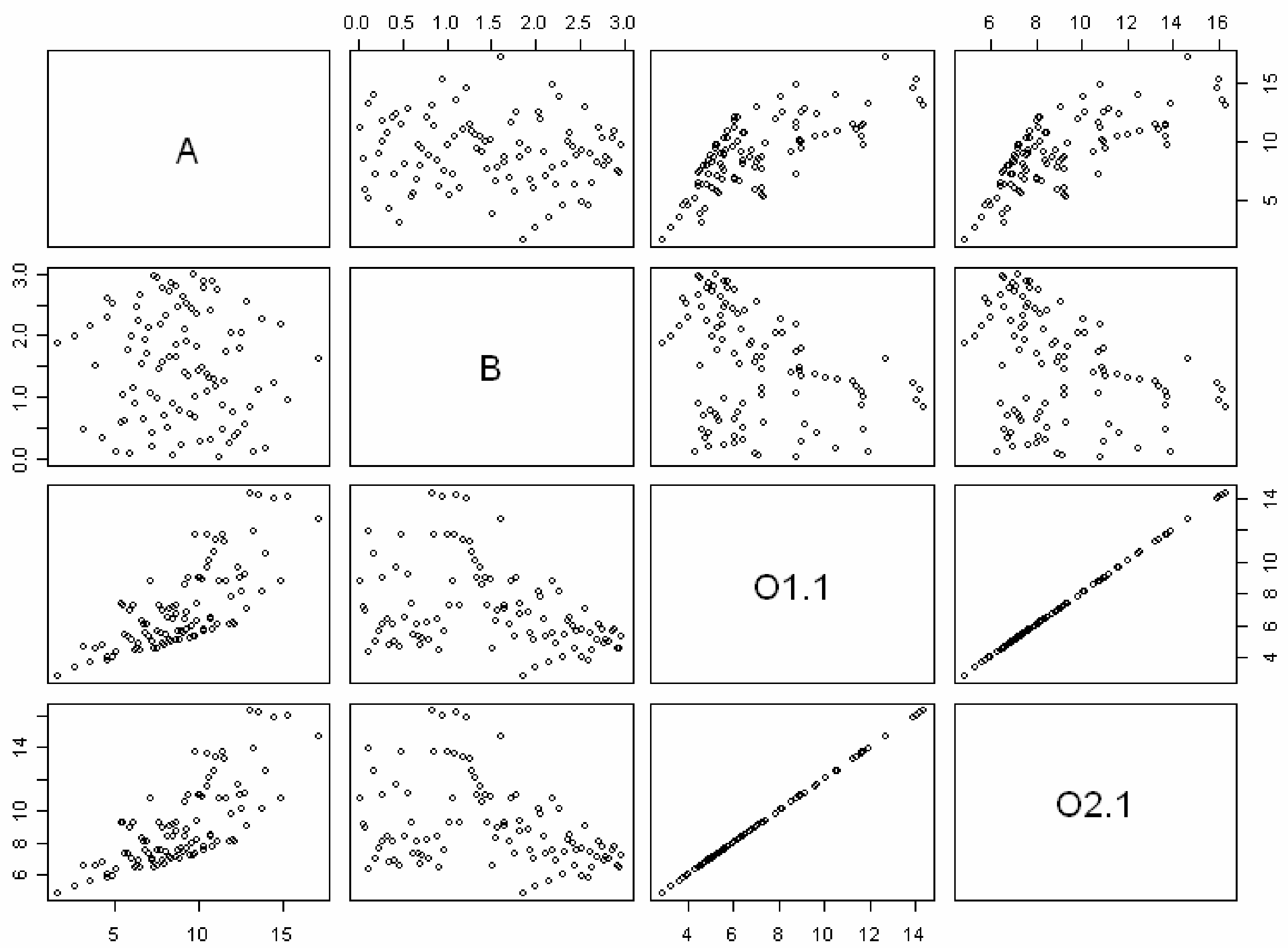
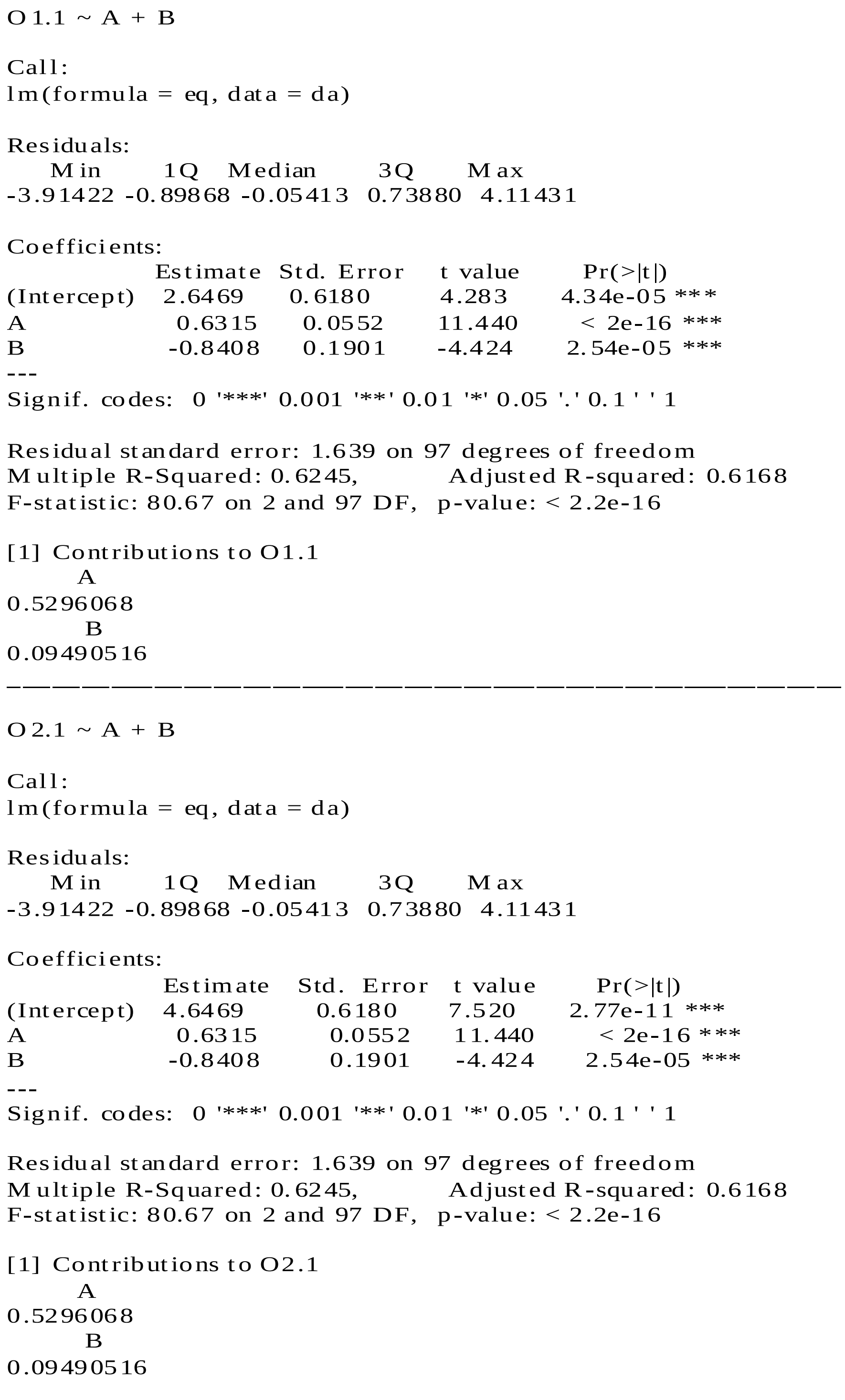
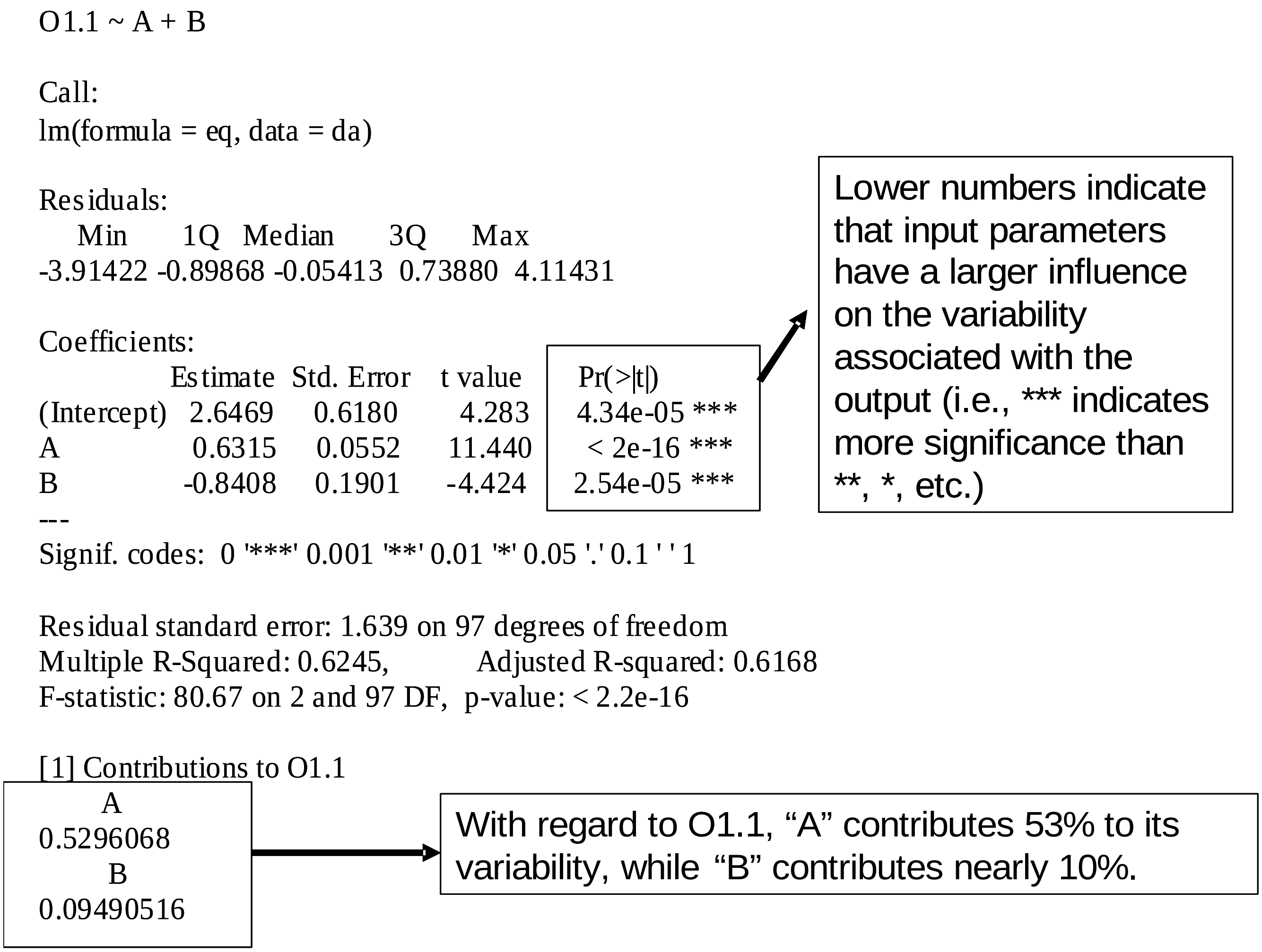


Figure 5.15. Scatter Plots Relating “A” and “B”, and “O1” and “O2” (i.e., O1.1 and O2.1), Respectively



```
O1.1 ~ A + B

Call:
lm(formula = eq, data = da)

Residuals:
    Min       1Q   Median       3Q      Max
-3.91422 -0.89868 -0.05413  0.73880  4.11431

Coefficients:
              Estimate  Std. Error   t value      Pr(>|t|)
(Intercept)   2.6469     0.6180       4.283      4.34e-05 ***
A              0.6315     0.0552      11.440       < 2e-16 ***
B             -0.8408     0.1901      -4.424      2.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.639 on 97 degrees of freedom
Multiple R-Squared: 0.6245,          Adjusted R-squared: 0.6168
F-statistic: 80.67 on 2 and 97 DF,  p-value: < 2.2e-16

[1] Contributions to O1.1
       A
0.5296068
        B
0.09490516
——————————————————————————————————————

O2.1 ~ A + B

Call:
lm(formula = eq, data = da)

Residuals:
    Min       1Q   Median       3Q      Max
-3.91422 -0.89868 -0.05413  0.73880  4.11431

Coefficients:
               Estimate   Std.  Error   t value      Pr(>|t|)
(Intercept)   4.6469       0.6180       7.520      2.77e-11 ***
A              0.6315       0.0552      11.440       < 2e-16 ***
B             -0.8408       0.1901      -4.424      2.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.639 on 97 degrees of freedom
Multiple R-Squared: 0.6245,          Adjusted R-squared: 0.6168
F-statistic: 80.67 on 2 and 97 DF,  p-value: < 2.2e-16

[1] Contributions to O2.1
       A
0.5296068
        B
0.09490516
```

Figure 5.16. Summarized Regression Analysis in the R Standard Form for each of the SummaryValues "O1" and "O2", Showing the Contribution of each of the SampledValues "A" and "B" to the Multiple R-Squared and Capturing the Degree that each of the SampledValues Has on Influencing the Variability of each of the SummaryValues



```
O1.1 ~ A + B

Call:
lm(formula = eq, data = da)

Residuals:
     Min       1Q   Median       3Q      Max
-3.91422 -0.89868 -0.05413  0.73880  4.11431

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   2.6469     0.6180   4.283  4.34e-05 ***
A             0.6315     0.0552  11.440   < 2e-16 ***
B            -0.8408     0.1901  -4.424  2.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.639 on 97 degrees of freedom
Multiple R-Squared: 0.6245,       Adjusted R-squared: 0.6168
F-statistic: 80.67 on 2 and 97 DF,  p-value: < 2.2e-16

[1] Contributions to O1.1
     A
0.5296068
     B
0.09490516
```

Lower numbers indicate that input parameters have a larger influence on the variability associated with the output (i.e., *** indicates more significance than **, *, etc.)

With regard to O1.1, “A” contributes 53% to its variability, while “B” contributes nearly 10%.

Figure 5.17. Contributions of each of the SampledValues “A” and “B” to the Multiple R-Squared and Capturing the Degree that each of the SampledValues Has on Influencing the Variability of the SummaryValues “O1”

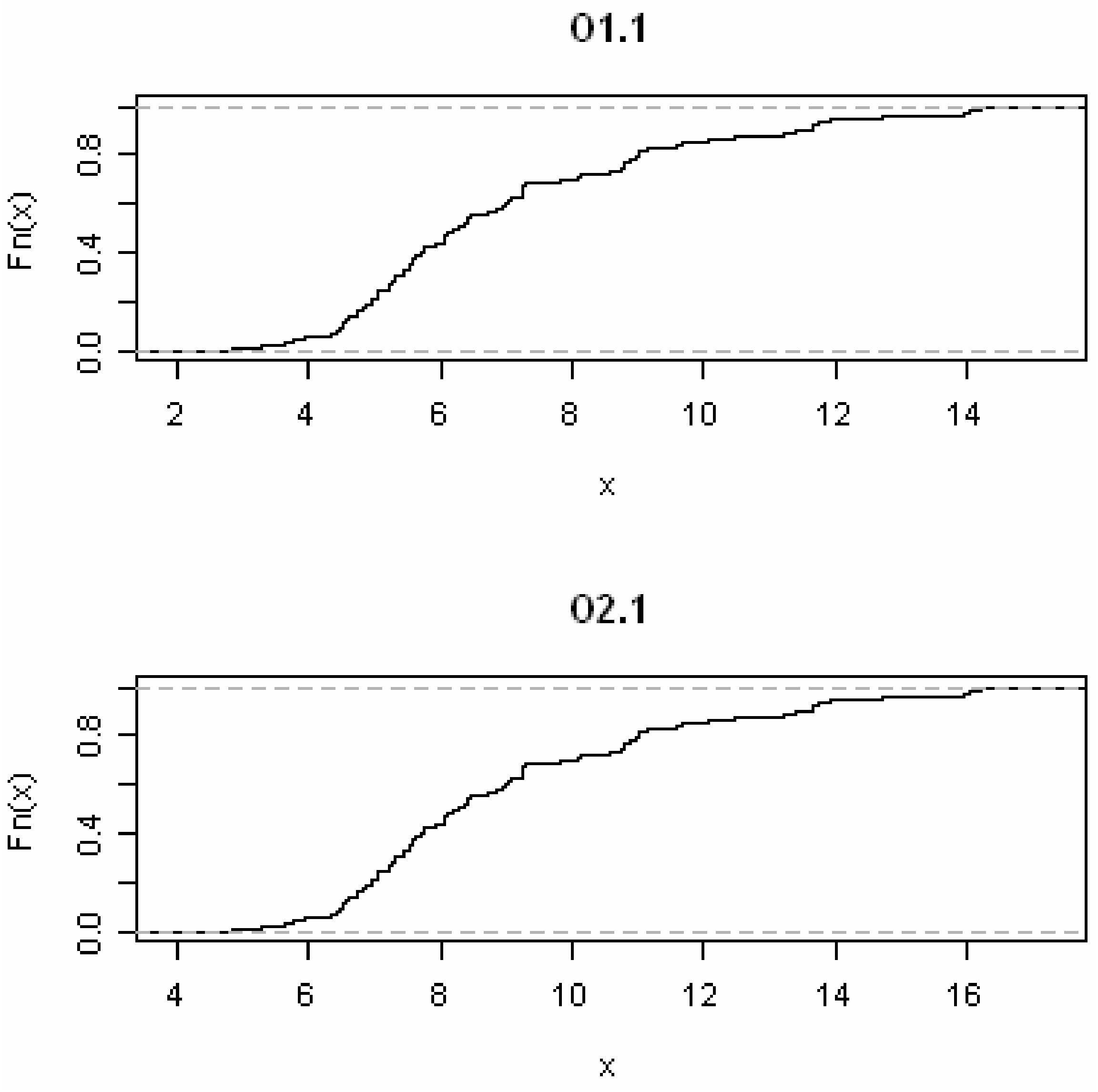


Figure 5.18. CDF Plots of SampledValues and SummaryValues, including a 95th Percentile Line that is Added for Reference into the CDF Viewer's Output

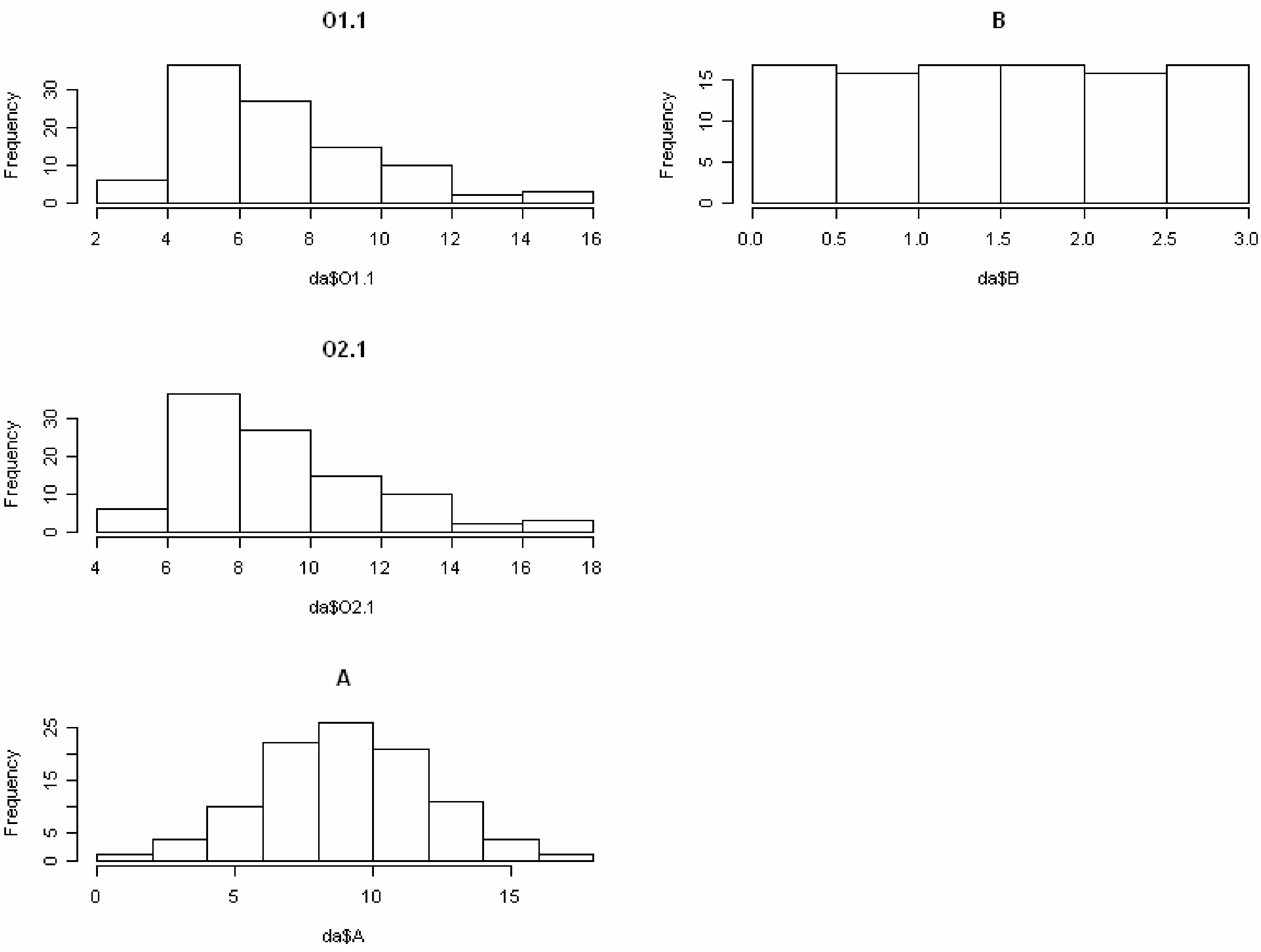


Figure 5.19. Histograms of SampledValues and SummaryValues, Including a 95th Percentile Line that is Added for Reference into the CDF Viewer's Output

6.0 References

Adaptive Risk Assessment Modeling System (ARAMS). Available at: http://el.erdc.usace.army.mil/arams/. Accessed 11-02-06.

Borcard D. 2002. *Multiple and Partial Regression and Correlation Partial r2, Contribution and Fraction [a]*. Université de Montréal, Département de sciences biologiques. Montréal, Québec, Canada. Available at: http://biol10.biol.umontreal.ca/borcardd/partialr2.pdf. Accessed 11-02-06.

Ecosystems Research Division (ERD). *Supercomputer for Model Uncertainty and Sensitivity Evaluation (SuperMUSE)*. U.S. Environmental Protection Agency (EPA). Available at: http://www.epa.gov/athens/research/modeling/supermuse/supermuse.html. Accessed 11-02-06.

Groundwater Modeling System (GMS) (software package). *Environmental Modeling Research Laboratory*. Available at: http://www.emrl.byu.edu/gms.htm. Accessed 11-02-06.

National Exposure Research Laboratory (NERL). *Hazardous Waste Identification Rule Assessment*. U.S. Environmental Protection Agency (EPA). Available at: http://www.epa.gov/nerl/factsheets/2001/goal5_assessment.pdf. Accessed 11-02-06.

Neter J, W Wasserman, and MH Kutner. 1985. *Applied Linear Statistical Models*, 2nd ed., Richard D Irwin, Inc., pp. 288-290, 417-444, 508, Homewood, IL.

The R Project for Statistical Computing. Available at: http://www.r-project.org/. Accessed 11-02-06.

7.0 Bibliography

U.S. Environmental Protection Agency. *Hazardous Waste Identification Rule Assessment*. National Exposure Research Laboratory. Available at (accessed February 28, 2006): http://www.epa.gov/nerl/factsheets/2001/goal5_assessment.pdf

U.S. Environmental Protection Agency. *Supercomputer for Model Uncertainty and Sensitivity Evaluation.* Ecosystems Research Division. Available at (accessed February 28, 2006): http://www.epa.gov/athens/research/modeling/supermuse/supermuse.html

7.1 Documentation for the FRAMES-3MRA Technology Software System

Volume 1: *Overview of the FRAMES-HWIR Technology Software System.* 1998. PNNL-11914, Vol. 1, Pacific Northwest National Laboratory, Richland, Washington.

Volume 2: *System User Interface Documentation.* 1998. PNNL-11914, Vol. 2, Pacific Northwest National Laboratory, Richland, Washington.

Volume 3: *Distribution Statistics Processor Documentation.* 1998. TetraTech, Lafayette, California.

Volume 4: *Site Definition Processor Documentation.* 1998. PNNL-11914, Vol. 4, Pacific Northwest National Laboratory, Richland, Washington.

Volume 5: *Computational Optimization Processor Documentation.* 1998. TetraTech, Lafayette, California.

Volume 6: *Multimedia Multipathway Simulation Processor Documentation.* 1998. PNNL-11914, Vol. 6, Pacific Northwest National Laboratory, Richland, Washington.

Volume 7: *Exit Level Processor Documentation.* 1998. PNNL-11914, Vol. 7, Pacific Northwest National Laboratory, Richland, Washington.

Volume 8: *Specifications.* 1998. PNNL-11914, Vol. 8, Pacific Northwest National Laboratory, Richland, Washington.

Volume 9: *Software Development and Testing Strategies.* 1998. PNNL-11914, Vol. 9, Pacific Northwest National Laboratory, Richland, Washington.

Volume 10: *Facilitating Dynamic Link Libraries.* 1998. PNNL-11914, Vol. 10, Pacific Northwest National Laboratory, Richland, Washington.

Volume 11: *User’s Guidance.* 1998. PNNL-11914, Vol. 11, Pacific Northwest National Laboratory, Richland, Washington.

Volume 12: *Dictionary.* 1998. PNNL-11914, Vol. 12, Pacific Northwest National Laboratory, Richland, Washington.

Volume 13: *Chemical Properties Processor Documentation.* 1998. PNNL-11914, Vol. 13, Pacific Northwest National Laboratory, Richland, Washington.

Volume 14: *Site Layout Processor Documentation.* 1998. PNNL-11914, Vol. 14, Pacific Northwest National Laboratory, Richland, Washington.

Volume 15: *Risk Visualization Tool Documentation.* 1998. PNNL-11914, Vol. 15, Pacific Northwest National Laboratory, Richland, Washington.

7.2 Quality Assurance Program Document

Gelston GM, RE Lundgren, JP McDonald, and BL Hoopes. 1998. *An Approach to Ensuring Quality in Environmental Software*. PNNL-11880, Pacific Northwest National Laboratory, Richland, Washington.

7.3 Additional Sources

Buck JW, BL Hoopes, KJ Castleton, and RY Taira. 1999. *Requirements for the FRAMES User Interface*. PNNL-SA-32277, Pacific Northwest National Laboratory, Richland, Washington (in publication).

Draper N, and H Smith. 1981. *Applied Regression Analysis*, 2nd ed. John Wiley and Sons, New York.

Gelston GM, RE Lundgren, JP McDonald, and BL Hoopes. 1998. *An Approach to Ensuring Quality in Environmental Software*. PNNL-11880, Pacific Northwest National Laboratory, Richland, WA.

Gelston GM, MA Pelton, R Lundgren, KJ Castleton, G Whelan, BL Hoopes, JL Kirk AJ Pospical, M Eslinger, JG Droppo, Jr., and DL Strenge. 2004. *Documentation for the Dictionary Editor of the FRAMEwork System (FRAMES)*. PNWD-3504, Pacific Northwest National Laboratory, Richland, WA.

Hoopes BL, MA Pelton, KJ Castleton, GM Gelston, G Whelan, and RY Taira. 2004. *Documentation for the FRAMEwork Development Environment*. PNWD-3509, Pacific Northwest National Laboratory, Richland, WA

Office of Civilian Radioactive Waste Management (OCRWM). 1995. *Quality Assurance Requirements and Description*, Supplement I, Software. U.S. Department of Energy, Washington, D.C.

U.S. Environmental Protection Agency (EPA). 1997. *System Design and Development Guidance*. EPA Directive Number 2182, Washington, D.C.

Whelan G, KJ Castleton, JW Buck, GM Gelston, BL Hoopes, MA Pelton, DL Strenge, RN Kickert. 1997. *Concepts of a Framework for Risk Analysis In Multimedia Environmental Systems (FRAMES)*. PNNL-11748, Pacific Northwest National Laboratory, Richland, WA.

Pacific Northwest
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)
www.pnl.gov