Investigating Uncertainty and Sensitivity in Integrated, Multimedia Environmental Models: Tools for 3MRA

J.E. Babendreier[a] **and K.J. Castleton**[b]

[a] *Ecosystems Research Division, National Exposure Research Laboratory, Office of Research and Development, United States Environmental Protection Agency, Athens, Georgia, USA (babendreier.justin@epa.gov)*

[b] *Pacific Northwest National Laboratory, Battelle Corporation, Richlands, Washington, USA*

Abstract: Elucidating uncertainty and sensitivity structures in environmental models can be a difficult task, even for low-order, single-medium constructs driven by a unique set of site-specific data. Quantitative assessment of integrated, multimedia models that simulate hundreds of sites, spanning multiple geographical and ecological regions, will ultimately require a comparative approach using several techniques, coupled with sufficient computational power. The Framework for Risk Analysis in Multimedia Environmental Systems - Multimedia, Multipathway, and Multireceptor Risk Assessment (FRAMES-3MRA) is an important software model being developed by the United States Environmental Protection Agency for use in risk assessment of hazardous waste management facilities. The 3MRA modelling system includes a set of 17 science modules that collectively simulate release, fate and transport, exposure, and risk associated with hazardous contaminants disposed of in various land-based waste management units. The 3MRA model encompasses over 700 variables, 185 of which are explicitly stochastic. Design of SuperMUSE, a 125 GHz PC-based, Windows-based Supercomputer for Model Uncertainty and Sensitivity Evaluation is described. Developed for 3MRA and extendable to other computer models, an accompanying platform-independent, Java-based parallel processing software toolset is also discussed. For 3MRA, comparison of stand-alone PC versus SuperMUSE simulation executions showed a distributed computing overhead of only 0.57 seconds/simulation, a relative cost increase of 0.7% over average model runtime. Parallel computing software tools represent a critical aspect of exploiting the capabilities of such systems. Fairly small, easy to write, and well suited for this application, the Java toolset developed here readily handled machine and job management tasks over the distributed system. SuperMUSE can complete over 2.5 million 3MRA model simulations per month.

Keywords: Multimedia Model; Model Evaluation; Parallel Distributed Processing; Supercomputer; Java

1. INTRODUCTION

Elucidating uncertainty and sensitivity structures in environmental models can be a difficult task, even for low-order, single-medium constructs driven by a unique set of site-specific data. The ensuing challenge of examining ever more complex, integrated, higher-order models is a formidable one. This is particularly true in regulatory settings applied on national scales that must ensure the continued protection of humans and ecology, while preserving the economic viability of industry.

Achieving adequate quality assurance in modelling in essence requires a battery of tests designed to establish the model's validity, trustworthiness, and relevance in performing a prospective task of prediction (Chen and Beck, 1999). To this end, model evaluation is seen as an increasingly critical step in the process of establishing confidence in a model's use, and providing a requisite level of safety that decision-makers may rely upon.

Aspects of sensitivity for a given model may be evaluated through a wide array of computational techniques, for example screening methods, local differential-based methods, and global methods (Saltelli *et al.*, 2000). There are equally provocative schemes (Funtowicz and Ravetz, 1990) that can be investigated to more fully characterize elements of uncertainty; reaching well beyond commonly applied Monte Carlo based probabilistic assessments (Cullen and Frey, 1999).

Thoroughly evaluating integrated, multimedia models that simulate hundreds of sites, spanning multiple geographical and ecological regions, will ultimately require a comparative approach using several techniques, coupled with sufficient computational power. This paper provides an overview of the multimedia model FRAMES-3MRA, describes a set of hardware and software supercomputing tools created to facilitate model evaluation, and summarizes 3MRA runtime costs associated with PC-based distributed processing.

2. FRAMES-3MRA MODEL

During the past five years USEPA's Office of Research & Development (ORD) has sponsored, along with other U.S. Federal Agencies, the development of a Windows-based software model Framework for Risk Analysis in Multimedia Environmental Systems (FRAMES). The Multimedia, Multipathway, Multireceptor Risk Assessment (3MRA) modelling system comprises of a unique set of simulation models developed within the FRAMES system (Figure 1).

Designed by researchers at the U.S. Department of Energy's Pacific Northwest National Laboratory in collaboration with ORD, the FRAMES-3MRA Version 1.0 modelling system includes a set of 17 science modules that collectively simulate release, fate and transport, exposure, and risk associated with hazardous contaminants disposed of in various land-based waste management units.

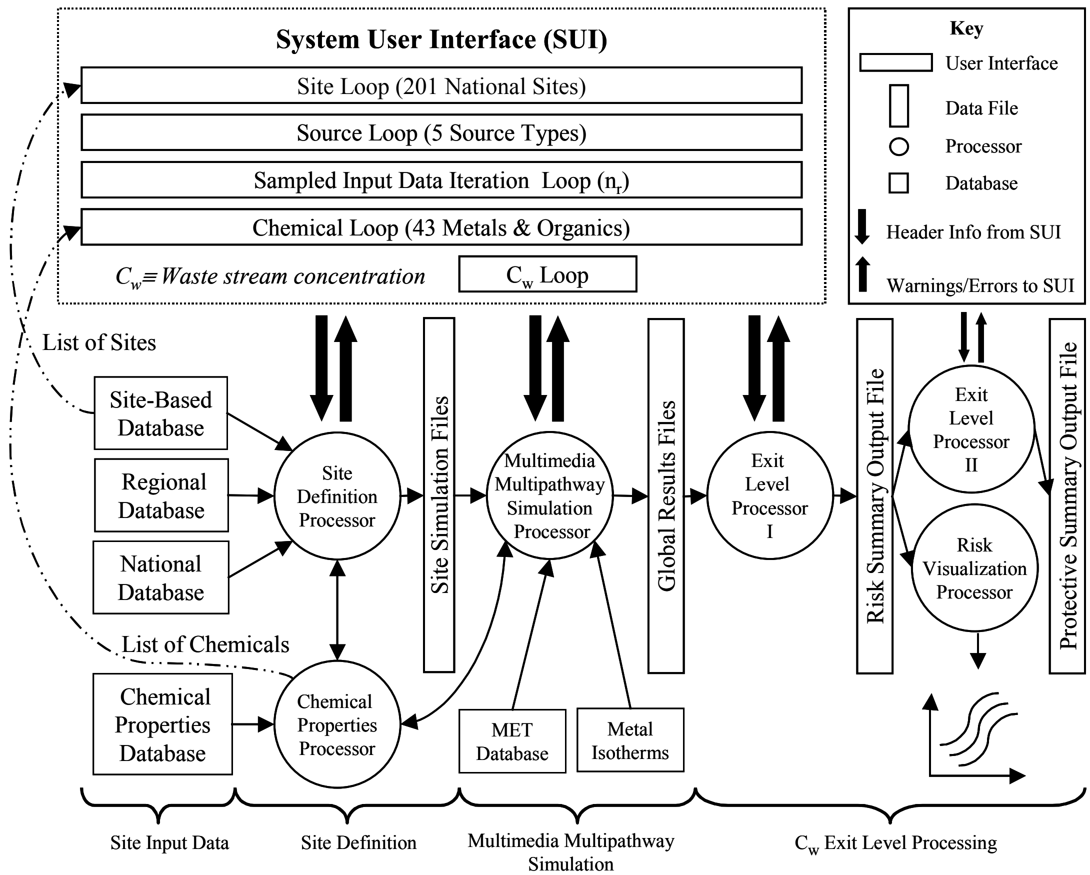

Figure 1. FRAMES 3MRA-HWIR Version 1.0 System Design: Stand-Alone Workstation Application.

2.1 Multimedia Risk Assessment

The 3MRA model encompasses over 700 variables, 185 of which are explicitly stochastic. 3MRA starts with a waste stream concentration in a waste management unit (landfill, waste pile, aerated tank, surface impoundment, or land application unit), estimates the release and transport of the chemical throughout the environment, and predicts associated exposure and risk. 3MRA simulates multimedia (air, water, soil, sediments), fate and transport, multipathway exposure routes (food ingestion, water ingestion, soil ingestion, air inhalation, etc.), multireceptor exposures (resident, gardener, farmer, fisher, ecological habitats and populations; all with various cohort considerations), and resulting risk (human cancer and non-cancer effects, ecological population and community effects). Collating the output for an overall risk analysis, 3MRA offers a probabilistic strategy for regulatory decisions.

Uncertainty and sensitivity analyses are being undertaken to evaluate the 3MRA modelling system. These efforts will emphasize the use of millions of simulations to test effects of small and large changes in model inputs and parameters.

2.2 3MRA Modelling System Overview

For the 3MRA stand-alone PC application, the user begins the simulation process by activating the System User Interface (SUI). The user selects the sites of interest, source types, chemicals, and waste stream concentration ranges (defined as C_w). On launching the simulation, the SUI generates a header file that contains all the information specified by the user. The Site Definition Processor (SDP) next reads in the header file information to determine the first site and scenario to be simulated and creates a complete set of flat-ASCII site simulation files (SSFs). In Figure 1, the innermost looping order for a given scenario set selected follows waste stream concentrations, chemicals, stochastic iteration #, sources, and sites.

The SSFs are created by extracting information from existing databases or randomly generating data using distribution information and statistical subroutines. Various databases represent different levels of data: site-based, regional, and national. These data are constant or stochastic, with cross-correlation data associated with certain variables. The chemical properties database contains needed data related to all chemical-dependent variables. The MET database contains five meteorological datasets describing various national monitoring stations (i.e. representing hourly, daily, monthly, annual, and long-term climatological data).

For each site, using a hierarchical scheme, the site-based database is first accessed to acquire actual site data that are constants, and then available stochastic site data are sampled. Regional data are next accessed to fill in data not in the site-based database, for both constant and then stochastic data. Finally, the national data are used to fill in any remaining missing data. Like the other processors shown in Figure 1, the SDP reports any processor-specific warnings or errors to the SUI.

The Multimedia Multipathway Simulation Processor (MMSP), which executes each site simulation, uses the generated SSF files as input. The SSF file set includes information describing chemical properties, site layout, sources, air data, vadose zone data, aquifer data, watershed data, waterbody network data, farm, terrestrial, and aquatic foodchain data, human and ecological exposure data, and human and ecological risk data.

The global results files (GRFs) contain all the key output data from the MMSP modules that were executed during a given simulation. The Exit Level Processor I (ELPI) output database stores key exposure and risk results for the entire simulation set. The Exit Level Processor II presents this summary of exposure and risk results.

2.3 HWIR National Assessment

The development of 3MRA has been driven by its initial intended application to a nationwide risk assessment underpinning USEPA's proposed Hazardous Waste Identification Rule (HWIR). The HWIR is designed to identify which waste streams can safely be released from existing hazardous waste disposal requirements. Hazardous waste with constituent chemical concentrations less than "exit" levels calculated by 3MRA could be reclassified as nonhazardous solid waste. The assessment is to be based on 419 waste management units located at 201 sites. Intended to extend to hundreds of pollutants, a dataset for 43 metals and organics was initially developed.

3. PARALLEL COMPUTING CLUSTER

A fundamental characteristic of uncertainty and sensitivity analyses is their need for high levels of computational capacity to perform many relatively similar computer simulations, where only model inputs change during each simulation. While this aspect is emerging as a critical area for environmental model evaluation, resources for Windows-based, PC-based modelling have often been limited by an associated lack of supercomputing capacity.

An increasingly common effort, particularly for Linux-based systems, distributed supercomputing achieved through use of PC clusters has expanded rapidly in recent years. Less common are clusters that support dual-boot capabilities for Windows-based models like 3MRA, especially those with dedicated KVM (keyboard, video, mouse) control. To facilitate model evaluation tasks for USEPA's modelling systems, ORD has recently developed a 125-GHz Supercomputer for Model Uncertainty and Sensitivity Evaluation (SuperMUSE).

Major components of the SuperMUSE (Figure 2) include a front-end program server, a back-end data server, and 121 client PCs with a minimum of 256 MB RAM. A variety of Windows operating systems are supported (i.e. Windows 95, 98, NT, 2000). Interconnections were achieved through use of 16-port Raritan KVM switches, and 24-port Linksys (10/100) network switches branching to a master

CISCO 3550-24/2 network switch. The system network protocol is based on TCP/IP. The system design currently provides for gigE channel (1000 megabits/sec) data flow to and from servers, and also allows single-user KVM remote access.

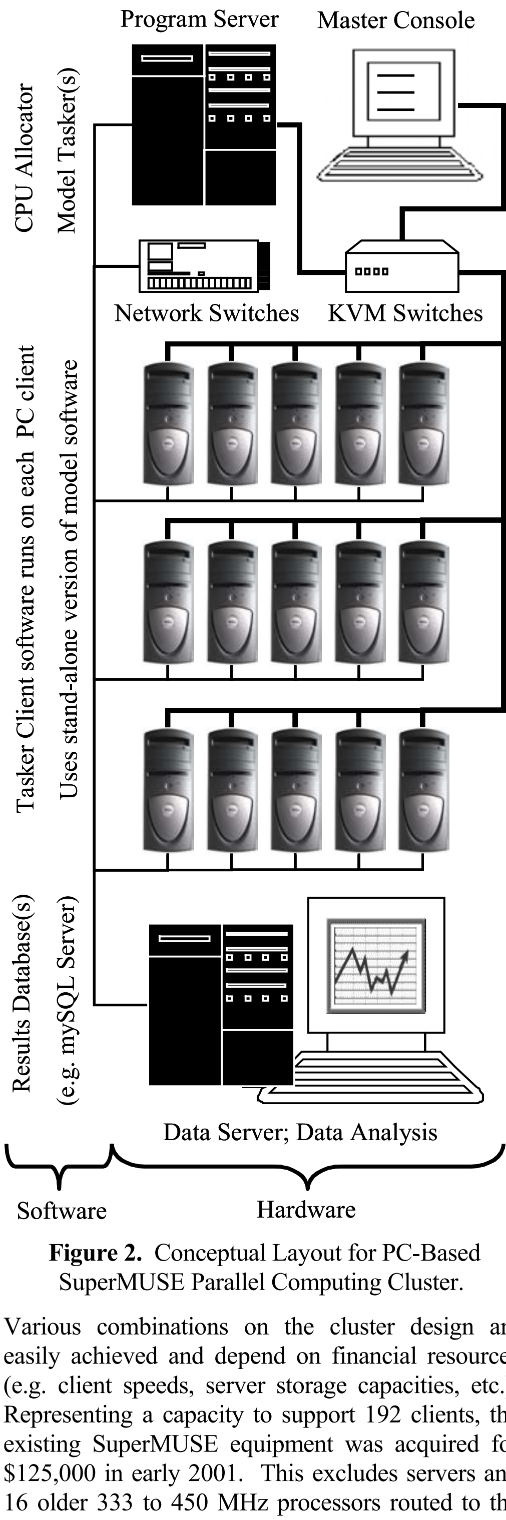

Figure 2. Conceptual Layout for PC-Based SuperMUSE Parallel Computing Cluster.

Various combinations on the cluster design are easily achieved and depend on financial resources (e.g. client speeds, server storage capacities, etc.). Representing a capacity to support 192 clients, the existing SuperMUSE equipment was acquired for $125,000 in early 2001. This excludes servers and 16 older 333 to 450 MHz processors routed to the project. Optimal purchasing based on $/GHz for client PCs will typically identify 3 to 6 month-old CPU technology. Plated wire shelving is best for cabling and heat dissipation. Pre-design considerations include available space, and room heating and cooling capacities. SuperMUSE will be expanded soon to 192 PCs totalling 270 GHz.

4. SUPERMUSE SOFTWARE TOOLSET

With the proliferation of workstation clusters connected by high-speed networks, providing efficient system support has become an important problem (Cruz and Park, 1999). To exploit capabilities of the SuperMUSE parallel computing environment, several software tools were needed. Key functionalities that were identified included 1) managing files across PCs, 2) facilitating the distribution of workloads among PCs, and 3) facilitating data analysis tasks. Java software application programs developed here included:

- Propagate Tool
- Distributed Management Program Toolset
- Site Visualization Tool

Propagate is a file/system management tool employing a dropdown equipment attribute file. For selected machines, the tool executes DOS commands over the network using a batch file script and wildcard designations for various attributes (e.g. machine ID, operating system, etc.).

With no user interface, the Site Visualization Tool extracts data from SSF and GRF file sets for an individual scenario and waste concentration. With integration and statistical processing capabilities, the charts produced via GNUPlot show the major outputs of each MMSP module, starting with the waste management unit source term and ending with human and ecological exposure modules. Collected into an html file, this site tool allows for visual summary of model outputs, spanning the entire MMSP model domain. Risk assessment is handled via computational and visualization tools within the existing ELPI and ELPII processors.

4.1 Distributed Management Program Layout

Shown in Figure 3, with program locations shown in Figure 2, the distributed management program toolset provides an effective, platform-independent parallel tool. Supporting uncertainty and sensitivity analysis evaluation tasks, the distributed processing scheme is capable of managing millions of simulations for 3MRA or other computer models.

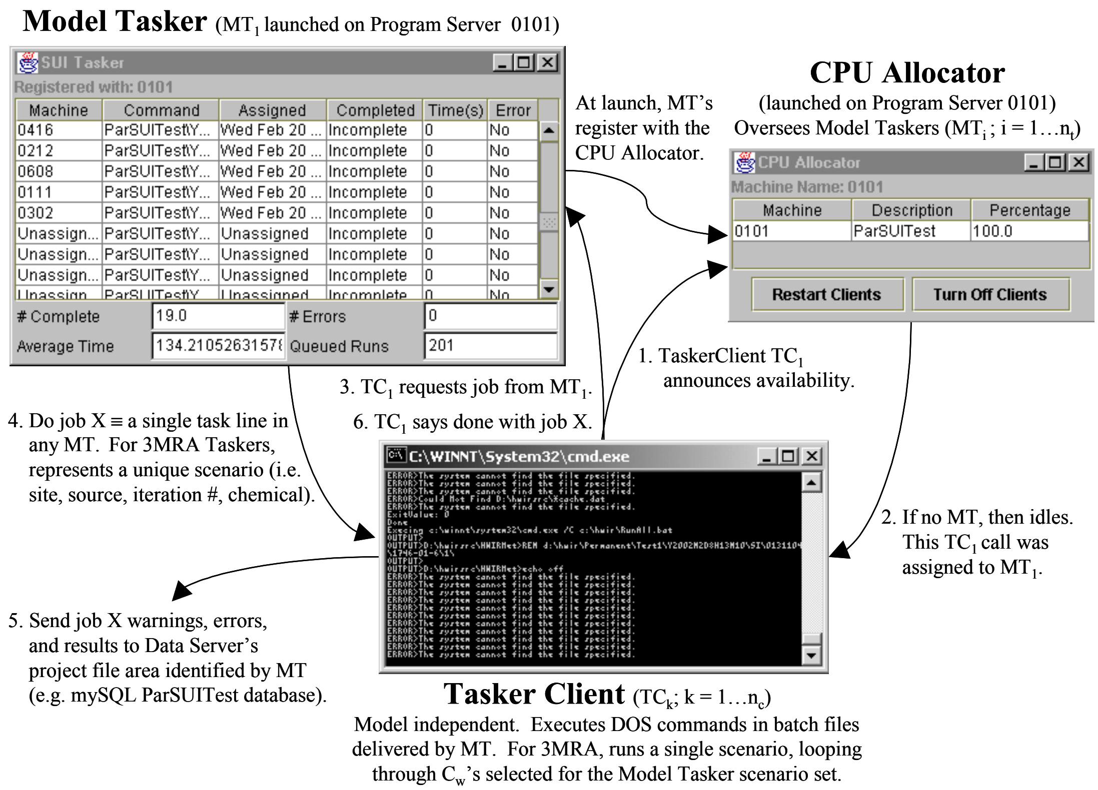

Figure 3. SuperMUSE Distributed Management Program Toolset - Conceptual Layout for 3MRA.

The CPU Allocator and Tasker Client are model independent. A Model Tasker is model dependent, and deconstructs a model's system user interface to generate a set of tasks (e.g. individual model simulations) amenable to distributed processing. For example, a Model Tasker was developed for 3MRA, identified in Figure 3 as the SUITasker. Several Model Taskers can be active, but must currently reside on separate servers. Modifications planned will allow for multiple Tasker Clients to reside on any PC, for example to take advantage of clients with multiple CPUs. The Java toolset is readily extended to Linux by recompiling the 3MRA input/output dll called by the SUITasker.

4.2 CPU Allocator

The CPU Allocator accepts job descriptions from one or several Model Taskers, and provides proportional load balancing across active Taskers. It functions as a TCP/IP server that accepts Model Tasker scenario set descriptions, and randomly assigns Tasker Clients to Model Taskers when clients indicate that they are free to execute tasks.

4.3 Model Tasker

The SUITasker reads a stand-alone 3MRA header file (created in stand-alone mode) to define the overall scenario set to be simulated. It functions as a TCP/IP server that accepts Tasker Clients directed to it by the CPU Allocator. Providing equivalent stand-alone scenario looping (Figure 1), the task list is created, managed, and updated with various statistics to track job performance. It maintains a time-assigned queuing approach and can also handle errant clients that unexpectedly fail to complete requested tasking (e.g. manages client power failures without client UPS backup).

4.4 Tasker Client

The Tasker Client is loaded on each client PC at start-up. Each Tasker Client then periodically calls the CPU Allocator when not tasked. If no Model Tasker is active, it is told to idle. The Tasker Client has no user interface, and functions as a TCP/IP client for the CPU Allocator and active Model Taskers. The client software will connect to the CPU Allocator, receive a Model Tasker machine ID, disconnect, and then connect to the assigned Model Tasker. It then receives a command from the Model Tasker with associated files to execute. In the case of the 3MRA SUITasker, this is a master batch file and a series of 3MRA header files, one for each C_w selected. The file set is first written to the client disk, and the model is then executed in batch sequence.

Providing key connections to the back-end data server, two auxiliary Java applications were also created for client job processing. These also have no user interface. For 3MRA, the tools represent additional calls within the batch file scheme, for each C_w header file executed. The first, a Process Messages Tool, reads normally produced warning and error files and, via JDBC, updates the mySQL server identified by the SUITasker. The second, a Site Summary Tool, extracts results from SSF and GRF files, storing user-selected variables defined by a delimited script file. The Site Summary Tool facilitates sensitivity and uncertainty analysis processing for all available input and output data.

5. 3MRA TIME TRIAL ANALYSIS

A comparison of 3MRA was made between stand-alone execution and parallel execution on SuperMUSE. Using 85 identical PC clients in SuperMUSE for parallel computations, a scenario set describing the 201 national assessment sites was selected representing all 419 site and waste management unit combinations. One chemical was chosen with 5 C_w values, and a single iteration. The Site Summary Tool captured 61 variables.

Shown in Figure 4, the average overhead runtime cost due to paralleling the 3MRA code was 6.0 seconds/simulation for full messaging capabilities, 7.2% of the average stand-alone model runtime for Carbon Disulfide. A more direct comparison of stand-alone PC versus SuperMUSE capabilities, with maximum storage turned off and no message or result processing on SuperMUSE, showed an increase of only 0.57 seconds/simulation, a relative cost increase of 0.7% over average model runtime.

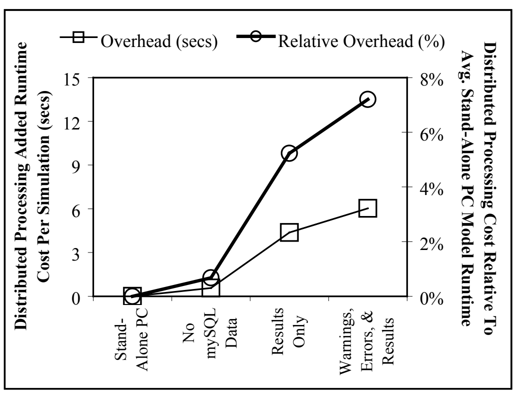

Figure 4. Comparison of Average 3MRA Runtimes Between SuperMUSE and Stand-Alone PC Version: Carbon Disulfide; 2095 Simulations.

For comparison, representing similar parallel execution totalling 90,085 simulations, the average runtime of 15 metal and 28 organic chemicals, with full messaging, was 120 seconds compared to an average of 90 seconds for Carbon Disulfide. On average, SuperMUSE can complete over 2.5 million 3MRA model simulations per month. Such execution in stand-alone mode, using a few PCs, would be prohibited by 1) the actual time expended to execute a given scenario set, 2) the need to optimize job assignments across PCs, and 3) the human capital needed to collect and collate errors, warnings, model input data, and model results.

6. CONCLUSIONS

The SuperMUSE computing cluster is ideal for conducting uncertainty and sensitivity analysis tasking involving relatively redundant model simulations, in both Windows or Linux environs. The supporting Java toolset developed for parallel computing represents a critical aspect of exploiting the capabilities of such systems. Fairly small, easy to write, and well suited for the application, the Java toolset readily handled the tasks of machine and job management over the distributed computing system. For 3MRA, added runtime costs were negligible compared to stand-alone PC execution, while the benefits delivered represented powerful, efficient model evaluation capabilities.

7. ACKNOWLEDGMENTS

The authors wish to thank Gene Whelan and Mitch Pelton at Pacific Northwest National Laboratory for their contributions to design and development of FRAMES-3MRA. Thanks are also given to Robert Ambrose, Kurt Wolfe, Robert Swank, David Brown, and Gerard Laniak of USEPA/ORD/NERL/ERD for their helpful comments during manuscript preparation. This paper has been reviewed by the USEPA. Mention of trade names does not constitute endorsement.

8. REFERENCES

Chen, J., and M.B. Beck, *Quality Assurance of Multi-Media Model For Predictive Screening Tasks*, USEPA, EPA/600/R-98-106, 1999.

Cruz, J., and K. Park, Toward Performance-Driven System Support for Distributed Computing in Clustered Environments, *Parallel Comput. & Distributed Process.,* 59(2), 132-154, 1999.

Cullen, A.C., and H.C. Frey, *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press, 335 pp., New York, 1999.

Funtowicz, S.O., and J.R. Ravetz, *Uncertainty and Quality in Science for Policy*, Kluwer Acad., Dordrecht, The Netherlands, 229 pp., 1990.

Saltelli, A., K. Chan, and E.M. Scott, *Sensitivity Analysis*, John Wiley & Sons, West Sussex, England, 475 pp., 2000.