Simulation Automation

SQL Server 2008 R2 programming tools were used to develop a software system for creating input and executing the model for 1,320 model runs, storing output data and processing simulation results for all these runs, and making project data available on the Internet. The system was designed to do the following:

- **Execute 1320 model runs in series**. Simulating all general circulation model, greenhouse gas emission scenarios, and 11-year moving window combinations can result in a large number of runs. For this study, 1,320 runs were needed, but the system is designed so that many different combinations of additional greenhouse gas emission scenarios and general circulation models can be used to simulate future hydrologic conditions.

- **Process large numbers of records**. Model runs produce a large amount of output data (for this study, more than 127 million records were processed). An application was developed for processing simulation results and storing them in a relational database.

- **Provide stakeholders with many output variable options for analyzing data**. The model has a large number of output variables. The system was designed so that different combinations of variables could be selected for output without having to alter program code.

- **Provide public Internet access to data and analysis products**. Because one goal of this project was to have stakeholders use simulation data for planning, a Web site was developed to provide access to all data in the project database.

The system uses a database-driven approach for automating model runs and processing simulation results. The term “database-driven” means that the user controls processing by changing data in a database rather than to the input files and programs that access the data. For example, using a database-driven approach, a user can determine the number of greenhouse gas emission scenarios and general circulation models to be run by adding configuration data to the database. A database-driven model automation program was developed to use this configuration data to detect and run all the greenhouse gas emission scenarios in the database without any additional user input. A goal of the database-driven approach is to minimize the need for a user to perform manual operations (such as copying files, cutting and pasting data, or altering computer code) and thereby reduce user error, produce consistent results, and make operations easier.

SQL Server 2008 R2 Management Studio was used to develop a single relational database for storing and managing all project data. All tables in the database were normalized to ensure that the database structure was suitable for efficient querying and to reduce insert, update, and deletion errors that could corrupt data.

Using one relational database had many advantages over using text files and spreadsheets stored in directories. A consistent data storage system made querying and managing data easier than it would be if the data were stored in assorted files with various file formats in multiple directories. Data security was enhanced because the database was protected by selectively setting permissions and passwords to limit access to the data to appropriate users. Data integrity was maintained by setting constraints on

data to ensure unwanted values were not permitted in the data tables (such as unplanned null values or values that exceed user-specified ranges). This reduced the likelihood of erroneous values being used for model input or analysis. The database is scalable, which means the number of records in the data tables can be increased without changing database structure. Therefore, the same database can be used to run simulations at the local, state, or regional scale. Built-in database analysis tools made it possible to perform complex queries and can be used for future studies to perform data mining to explore trends and relationships between model input and output data. Most importantly, the database is adaptable to many types of analysis tasks, which present many options for future work. For example, the database can be linked to other deterministic models or used for statistical or data mining studies. The database also can be linked with other data sources to provide information to the public through the Internet.

The SQL Server Integration Services (SSIS), a development tool within SQL Server 2008 R2, was used to develop applications (also called solutions) for running simulations and processing simulation results. SSIS was selected because of its tools for extraction, transformation, and loading (ELT) of data, and for its capability to read data from, and write data to, a wide variety of data storage formats (such as databases, delimited text files, and spreadsheets), which will facilitate integrating the watershed model with decision support systems. A SSIS solution consists of a set of integrated modular subprograms (called packages) that are programmed to perform specific tasks. SSIS graphical tools were used to develop a program framework. C#.NET and SQL code was then written for performing customized complex data manipulation tasks.

To summarize, 1,320 model runs were completed to simulate all combinations of three greenhouse gas emission scenarios, five general circulation models, and eighty-eight 11-year moving windows. More than 127 million records were processed over a 50-hour period to produce a 1-GB database storing daily, monthly, and central year values for 21 output variables. All output data from the simulations are stored in a relational database that is accessible to the public via the Internet.