Integration Of Remote Data Into Water Resources Simulation Software: Now Or Never?

R. M. Argent
Centre for Environmental Applied Hydrology, The University of Melbourne, Parkville, 3052, Australia, R.Argent@civag.unimelb.edu.au

Key words: Remote Data Sources, Water Resource Management

Abstract: Many types of software are used in the management, design, and planning of water resources systems, requiring access to a broad range of data sources. These data are becoming increasingly available from remote sources such as the world-wide web, with access times and levels of reliability that are becoming viable for modelling purposes. The opportunity exists, therefore, to incorporate direct linking and downloading of these data into water resources software development, thereby enabling access to the most recent data for each simulation run. The appropriateness of this approach is determined not only by the reliability and integrity of the data source, but also the way the data are integrated into the software, the sensitivity of the modelling output to the additional data, and the rate of change of the data and the phenomena being measured. Developers and users should examine these issues when deciding where or when to adopt remote data sourcing techniques.

1. INTRODUCTION

Software use in water resources covers an extensive range of activities, and represents possibly one of the longer and broader uses by industry of environmental information systems. Systems software are used by water resources researchers, planners, managers and educators for a wide range of purposes (The Institution of Engineers, Australia, 1996; McDonald and McAleer, 1997; El-Swaify and Yakowitz, 1998).

One of the areas where software is becoming increasingly used is that of management planning and decision making, where simulation modelling is used to provide information to managers over spatial and temporal scales that range from the next storm to alternative future climates. Water

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35503-0_29

R. Denzer et al. (eds.), *Environmental Software Systems*

resources simulation software utilises a range of data from many different scientific disciplines, and includes point and spatial types, monitored over a variety of timescales.

Many of the data used in water resources simulation software are obtained during the software development and testing phases, and subsequently remain relatively unchanged during program life. Other types of data, associated with particular modelling runs, situations or systems, are updated as frequently as the relevant models are used. With ongoing development in the availability of data from remote, or distributed, sources, the opportunity exists to increase the use of these data in water resources simulation software by designing and implementing systems that access these remote data sources on an as-needs basis. The ideal process envisaged is for data-providing agencies and authorities to make available their data through sources such as world-wide web (WWW) sites, CD-ROM, or directly from field instruments. The sources would then be accessed either directly by program software during execution, or else indirectly through prior download and manipulation. This approach utilises sound information science principles by avoiding duplication of data and ensuring use of the most recent data within software systems.

This kind of approach has been *technically* feasible for years, with examples found in real-time flood warning systems (Srikanthan *et al.*, 1994) and environmental radiation measurement and assessment networks (Doberkat *et al.*, 1997). As more data sets become available from remote sources, the temptation for simulation software developers to include access to these sources in their software will also increase. There are, however, issues of access, integration and use of these data and their sources that affect the advisability and practicality of implementing full reliance of software on remote data. This paper explores some of these issues and discusses the question of whether developers should adopt remote data sourcing techniques now, or never, or at some time in the future. A simple case study of a water resources management application is presented, within which the potential for changing from a static, file-based data system to a dynamic, as-needs data access approach, is discussed.

2. DATA CHARACTERISTICS IN WATER RESOURCES SIMULATION

One of the issues to be considered in the adoption of remote data access techniques is that of the period over which the data may be updated or changed, and the subsequent effect on the simulation results of such changes. In this context, updating of data implies the addition of new data from, say, a monitoring network, while changes to data refer to alterations

that may be made in derived data through different calibration or synthesis, or as a result of updates. The most widely used data in water resources are probably times series of climate and hydrological variables, along with spatial variables such as topography, soil type, and land use. Table 1 lists typical spatial and temporal characteristics for data used in water resources management in Australia, and highlights the differing range of scales of the data collected. When used in simulation software, the potential information benefit from having the most recent data can be quite large, depending on data usage and associated information generation. Conversely, for some variables the synthesis process that produces information from data is not strongly affected by the addition of a few data points that may have been obtained since the last time the software was run, and, consequently, incorporation of the most recent data has low value.

Table 1. Typical data characteristics for Australia

Data	**Temporal Scale**	**Spatial Scale**
Rainfall, Flow	1 minute – 1 day	10 – 100 km
Water quality	Daily – monthly	10 – 100 km
Topography; Vegetation cover (remote sensing)	Sub-daily – monthly	5 m – 100 km
Census, Social data	1 – 10 years	Individual – regional
Research data collection	1 – 2 times	Point - regional

The temporal and spatial coverage of various data types are dependent not only on the nature of the variable, but also the equipment available for monitoring, and the social, political and economic context within which they are being monitored. For example, the occurrence of severe drought in south-eastern Australia in 1982-83 and 1997-98 prompted many water authorities to alter or increase their monitoring and assessment of surface and sub-surface water resources.

In essence, the data used in water resources simulation software cover a broad range of spatial and temporal scales, and are changed or updated on periods that vary from sub-daily to numbers of years. When considering the use of remote sources for these data in software development, attention should therefore be given to the likely period over which the data change, particularly in comparison to the design life of the software, and also to the sensitivity of the simulation results to these changes in data.

3. DATA INTEGRATION IN SOFTWARE

The range of data types used in water resources software provide developers with a multitude of data access, storage and software integration problems that must be solved during application development. A variety of methods are available for integrating data into software, and these can be classed loosely into four groups, as follow:

1. Manual input, with the user entering data via a keyboard;
2. Direct integration, where the data are written into the software code;
3. Block loading, where blocks of data are input from files into arrays for later use by the software, and
4. Dynamic loading, where data are obtained by the software from data sources, such as a database, on an as-needs basis.

The selection of an integration technique to meet any given data requirement in software development is determined largely by the data type, the way in which the data are used, although the nature of the developer also has some influence.

Manual data input is not likely to be amenable to remote-source data integration, as manual inputs are often given in response to a particular prompt or are positioned on an application screen. Similarly, software with fully integrated data, such as those in look-up tables, are difficult to integrate with remotely accessed data sources. Although the data are in a defined format that could be matched by remotely sourced data, the requirement in many circumstances for software to be recompiled upon the addition of new data, would make this approach impractical.

The approach that appears to hold the most promise for melding with remotely accessed data sources is block loading, with data files prepared before program execution. Although this approach does not fulfil the ideal of as-needs access during program execution, and has the potential drawback of duplicating locally the remote data, it can be considered a low-risk compromise, with reduced likelihood of programs crashing during execution due to unavailable data sources. Provided remote data sources are nearly always available, the value to users of the locally stored data should be negligible. Consequently, the risk of data duplication should be reduced as users discard data once it is used, in the knowledge that it can be easily obtained from the original source when needed. An issue related to block loading of data is that of transforming data obtained from a remote site into the input format for the model. Water resources management can involve pulling together data from disparate sources, beyond the traditional climate and flow data, with consequential variability in formats and styles of data obtained from remote sources. One approach to this is to use data manipulation toolboxes, or data „wizards", that operate separately of the

main program. An advantage of this is that remote data sites are sometimes subject to change in either address or format, and a fully featured wizard-style system would greatly reduce the propensity for failure during data downloading.

Software that uses dynamic data interactions to read and write data during program execution provide a number of challenges for working with remotely sourced data. The main challenges relate to the incompatibility of styles, with remote data sources generally being read-only, while databases are often designed to facilitate reading and writing of data. A way to meet these challenges is to use the remote source to supply data, such as invariant physical features and dynamic features used in program initialisation, and then to use a local database for storage and exchange of dynamic program output.

The use of remotely sourced data in water resources simulation software seems, therefore, to be a practical alternative for some types of data integration, requiring only that data be readily and reliably available, and that data managing programs can be developed that access and transform data into a format suitable for software use. There are, however, other issues of software development, user interactions, and data interpretation that should be considered.

4. CHARACTERISTICS OF SOFTWARE AND SOFTWARE DESIGN

Beyond the issue of data integration there lie software characteristics that are relevant to the inclusion of remotely sourced data. These include the design life of the software and the general approach taken in software design. Good software design should include a planned redundancy or redevelopment, and include either data redundancy in the design, or specifically design data structures to survive beyond the life of the software. The level to which remote data should be integrated into the software therefore depends on the life spans of both the software and the data, as well as the relative return on effort made in providing remote data integration. For more fundamental data that are subject to significant change over time, and that are likely to be used in redeveloped software, it is sensible to design robust data structures. These obviously would be the most desirable data to be obtained from remote sources, if they were available, to save time and effort when new software was being developed and used. For more model-specific data, the capacity for updating and upgrading is more reliant on the nature of the data, and only those that change significantly over time should be considered.

When considering the issues of software life span and data use, it is worthwhile reflecting that the software that tend to have long lives are either very flexible and generic, and are maintained through popular use, or have a single, specific use, and are retained either by the excellence of their performance or by bureaucratic inertia. Consequently, even the most technically feasible approach to remote data source integration may founder in the face of established procedures. There are, therefore, other issues related to the users of data that must be considered in the process of getting developers and users to integrate remotely sourced data into their systems.

5. HUMAN INPUT AND CONSEQUENCES FOR DATA

Often, much of the data required for a simulation program are obtained during development from knowledgeable staff in the data collecting authorities and agencies, and come with both formal and anecdotal metadata. Metadata provides the developer, and, sometimes, the user, with relevant information on the assumptions behind, and guidance on the appropriate use of, the data. With reduction in human input between data collection and use in software, particularly in the case of remotely sourced data, the potential is much greater for data to be used in an inappropriate way.

To counter this, adoption of good information science principles in software development is essential. These include provision of metadata with the actual data obtained from remote sites, and also interpretation and presentation of the relevant metadata in software. Relevant metadata can include: the quality of the data; the standards used in collection and storage, any history of data processing, manipulation and collation that may have affected the raw data, and changes to metadata.

An aspect where good metadata on processing is required is in the filling of data gaps. Often this is done by users who, through their exposure to the data during collation, have obtained a good "feel" for it, and so are able to fill data gaps in an appropriate manner. Many interesting features of data can be learned when developers and users spend time extracting relevant sub-sets from larger data sets, collating data into formats suitable for user software, and filling data gaps. When remotely sourced data is processed with little or no collation and manipulation, it is essential that metadata relevant to the correct interpretation of output is also obtained and presented.

6. REMOTE SOURCE ACCESS PRACTICALITIES

A final issue in obtaining data from remote sources in that of the accessibility and integrity of the data and the remote data site. Source

integrity is a problem for management of the site, and is treated by good site security, guarantees on the data, and the supply of appropriate metadata. If software is developed that relies entirely on a remote data source, then there is a high cost in having data sites unavailable. Ways to combat this lie in the use of back-ups and mirror sites, and suggest that elegant software design for remote data access will have some of the features listed previously, such as data management tools that can seek out the data if they are in a different place to the last time they were accessed.

The case study of the next section explores a software system that uses a number of different data sets in different ways, and provides some discussion of the practical aspects of integrating data from remote sources.

7. CASE STUDY OF "FILTER"

The city of Melbourne, Australia, with a population of some 3.2 million in the greater urban area, lies at the north end of the semi-enclosed 1950 km^2 Port Phillip Bay (Harris *et al.*, 1996). The Port Phillip catchment covers some 9950 km^2 with a range of urban, agriculture and forest land uses. The Bay is subject to water quality threats, particularly nitrogen, from a range of point and nonpoint sources within the catchment, and government policy has been established with the aim of reducing significantly the nitrogen load delivered to the Bay.

As part of putting this policy into action, key authorities supported the development of a water quality management screening program during 1998. This screening program, named FILTER, was developed to allow managers to obtain estimates of the pollutant loads being generated, transported and reduced in various Bay sub-catchments, and to examine where the potential lay for management intervention to produce cost-effective reductions in pollutants. FILTER combines representative pollutant concentrations with historical flow data to obtain estimates of pollutant loads at various points in the catchment waterways, including flow entrance points to the Bay.

Given these load estimates, and land use data for the Bay sub-catchments, users can compare the loads estimated from the product of flow and concentration with pollutant loads typically generated from the land uses in the catchment. The land-use based load estimates are modified through natural and artificial processes, reflected in the model by "delivery ratio" values. Management interventions variously affect pollutant generation rates, or delivery ratios, depending on the type of intervention, and so alter the modelled value of pollutant load delivered into the Bay.

The management and spatial distribution of data in the program resulted in the use of three forms of spatial descriptor, as follow:

- Basin – one of the eight main drainage basins in the Port Phillip catchment;
- Sub-catchment – the catchment area upstream of a defined point, such as a flow monitoring station, and
- Polygon – the catchment area between that of a defined point and that of the next upstream defined point.

The data sets used in the FILTER software consist of:

- Flow – monthly values over fifteen years, covering a range of wet and dry periods to provide for considerations of the effects of climatic variability on load;
- Water quality – from 53 stations within three different monitoring networks, covering up to 21 years, at frequencies varying from weeks to years, and with a range of numbers of points from three to 512;
- Pollutant generation rates – for the three pollutants total nitrogen (TN), total phosphorus (TP), and total suspended solids (TSS) for 16 land use types, and
- Land use - percentage of each of 16 possible land uses types for 62 polygons making up the Port Phillip catchment.

Problems encountered in data collection included incomplete data sets, missing data, incorrect data, missing or non-existent metadata, and, in the case of the GIS files used for defining and generating basic information on the catchment polygons, the absence of even a basic GIS file delineating the basin boundaries.

All of the data used in the FILTER program were stored as text files, with a small amount of metadata on the file structure and content included in the header. As an example, Figure 1 contains the header information and first few lines from the flow data file.

```
\\Header for Sub-Catchment flow data file named FLOW.CSV
\\Single number refers to number of stations contained in this file
\\Monthly flow data for the years 1983 to 1997 for the 62 sub-
catchments in ML/month - line fixed length - format of lines as
follows:
\\Line 1: Sub-catchment Identifier ;IDNumber
\\Line 2 to 15: 12 monthly flow values Jan to Dec
62
228203,52
182,42,208,198,466,1774,2237,3014,5702,4987,1893,1335
365,185,641,602,326,349,912,2035,9351,2423,623,617
```

Figure 1. Header information in flow data file

In terms of the potential for using remotely sourced data in the FILTER program, the land use and water quality would be appropriate, while the flow and generation rate data would not. The flow data are historical values from a specific fifteen year period, so even though such data could be available from a remote source, they would not be relevant to the current data usage. The generation rates were selected to cover a range of values found in an

extensive review of the literature, and so are not the kind of data that would be monitored and made accessible at a remote source by an agency or authority. The synthesis and interpretation of published values required to provide appropriate generation rates means that these values would be largely unchanged until further study results were published, and the information from these studies taken into consideration. At that stage the simplest way to change these values, if required, would be through manual editing of the generation rate file.

The water quality data used in the program would be suitable for updating with recent data, as the monitoring networks are still operating and producing new data, and some of the water quality sites have low numbers of data points, so would benefit from more data. If such data were available at remote sites, it would be valuable for the block data files to be updated every time the model was run. However, the factors in deciding to provide this capability include:

- A number of different agencies run the FILTER program, so model output from one run would not necessarily be based upon the same data, and same associated assumptions, as other runs. This might lead to some confusion or disagreement when agencies compared results.
- The information developed from examination of the water quality data is not very sensitive to the addition of a few data points, so although the assumptions and model results may change over longer periods as the available data changes significantly, over a short time the effect of having the most recent data would not be great.
- The data are obtained from three separate networks, each of which has a different monitoring regime. Consequently, it is possible that a number of sources would need to be accessed for each run or set of runs, increasing the risk of systemic problems such as unavailable data or changes in the format or address of data sources.
- It is probable that the FILTER program would be used intensively on an occasional basis, so it would be preferable to update the data (either manually or dynamically from a remote site) before a period of intensive use, rather than before or during each model run.
- Finally, the program is relatively new, and there is no history of usage to indicate in what way and how often it is likely to be used. Being a screening program, FILTER may have a very short life, as it is designed to bring users to a common understanding, and then allow them to move on to more detailed studies using different software and techniques.

Overall, these factors indicate that a benefit would accrue from integrating remotely sourced water quality data into FILTER, although this would probably best be done by occasional updating of the input file rather than on an as-needs basis.

The land use data are also open to being accessed from remote sources, as they are maintained by one authority, change over time, and have a relatively significant effect on the model output. In comparison with the water quality data, land use changes significantly over periods on the order of one to five years, but as the model results are sensitive to these values, they should be kept up to date. The possibility of remote access to these data is small, and they are related to the specific polygon areas defined for the FILTER program. These polygons are not used elsewhere outside the program, so a software-specific interpretation would have to be undertaken. Consequently it would be more appropriate for this type of data to be prepared manually.

8. DISCUSSION

The example presented here, and the discussion in previous sections, has shown that it is possible to improve the quality of information provided by water resources software through access to remote data sources. There are many types of software used in management, design, and planning of water resources and other environmental systems, and these have a range of integration approaches for data from various sources. These data are becoming increasingly available from remote sources, with access times and levels of reliability that are becoming viable for modelling purposes. Therefore, there exists the opportunity to move away from the use of static data sets in modelling, whereby the data can be out of date before the modelling is done, to the use of dynamic data through active linking and downloading before or during the modelling task. The appropriateness of this approach is determined by a number of factors that include the sensitivity of the modelling output to additional data, the rate of change of the data and the phenomena being measured, access to the data, and the short and long term reliability and integrity of the data source.

9. CONCLUSIONS

As data supply and monitoring agencies and authorities decrease their personnel involvement in data dissemination, and find further cost cutting methods for handling data by making it available from distributed and remote sources, the risks, challenges and opportunities for water resources software developers and users are large. One of the principles of good environmental software development is to know the data that you are using. It has been suggested that software developers should be drawn from staff with field experience (Lovelock, 1979), as these staff are familiar with the problems and pitfalls associated with the data they are using, and this

principle is valid in many of the fields of modelling and simulation in environmental management (Holling, 1978; Walters, 1986; El-Swaify and Yackowitz, 1988). Developers and users should take heed of this in the headlong rush towards making all things available on "the Net" and be discriminating about the extent to which *instant* data access for modelling is adopted in their software. Whether to provide access to remote data sources now, or never, is a decision that should be made after consideration of a range of issues involving data, software and users, and it is hoped that the discussion provided here has highlighted some of these.

10. ACKNOWLEDGMENTS

The author acknowledges the support of the Land and Water Resources Research and Development Organisation through project UME29. Many of the ideas related here were discussed at the 1998 Workshop on Design Principles for Environmental Information Systems, and the author acknowledges the valuable input obtained from workshop participants.

11. BIBLIOGRAPHY

Doberkat, E.-E., Schmidt, F. and Veltmann, C., 1997: Re-engineering the German Integrated System for measuring and assessing environmental radioactivity. In *Environmental Software Systems, Volume 2.* Denzer, R., Swayne, D. A. and Schimak, G. (Eds), London, Chapman and Hall, 182-189.

El-Swaify, S. A. and Yakowitz, D. S., (Eds), 1998: *Multiple Objective Decision Making for Land, Water, and Environmental Management*, CRC Press, Boca Raton, Fla.

Harris, G., Batley, G., Fox, D., Hall, D., Jernakoff, P., Molloy, R., Murray, A., Newell, B., Parslow, J., Skyring, G. and Walker, S., 1996: *Port Phillip Bay Environmental Study: Final Report*, CSIRO, Canberra, Australia.

Holling, C. S., 1978: *Adaptive Environmental Assessment and Management*, John Wiley and Sons, Chichester.

Lovelock, J. E., 1979: *Gaia: A New Look at Life on Earth,* Oxford University Press, Oxford.

McDonald, A. D. and McAleer, M. (Eds.), 1997: MODSIM 97. *International Congress on Modelling and Simulation Proceedings, Volumes 1-4,* The Modelling and Simulation Society of Australia, Inc., Canberra.

Srikanthan, R., Elliott, J. F. and Adams, G. A., 1994: *A Review of Real-Time Flood Forecasting Methods.* Report 94/2, Cooperative Research Centre for Catchment Hydrology, Melbourne.

The Institution of Engineers, Australia, 1996: *Water and the Environment. Proceedings, Volumes 1-2, 23rd Hydrology and Water Resources Symposium, May, 1996, Hobart, Australia.* I. E. Aust., Canberra.

Walters, C., 1986: Adaptive Management of Renewable Resources, Macmillan, New York.