Why collect data when we can model it?

R.P. Silberstein

CSIRO Land and Water, Floreat, WA 6913.
Richard.Silberstein@csiro.au

Abstract: The "art of modelling" natural systems has progressed at an enormous rate over the last 10 to 20 years. In particular, during the last decade as computational power increased to the stage where we can now have a super-computer on our desk, the detail and fine scale processes that can be included in models is fantastic. This has opened doors our forebears could only have dreamed of. However, as modelling power has increased, there has been an accompanying reduction in "datapower" in some areas – in particular in the collection of hydrological data. While we undoubtedly have access to huge datasets of extraordinary technological finesse such as the remotely sensed data from satellites, our collection of more basic and traditional datasets suffers woefully. We can read car number plates from outer space, but we still, in the main, measure rainfall by putting a bucket out in a paddock. The argument often mounted by those with the purse is that with current modelling power, data needs are reduced. This is an extraordinarily dangerous and arrogant statement. Our current generation of models are powerful, and do give insights we may not have previously had, but they are only models. Real insight into natural systems comes from observation of them, and the true role of models is to assist this process, not to replace it.

Keywords: *modelling, hydrology, data collection*

1. INTRODUCTION

This paper was stimulated by comments received when the author was attempting to reinvigorate hydrological monitoring in Australia. While senior bureaucrats in data collection agencies and their policy clients generally appreciate the need to collect data on the systems we are trying to manage, two comments made were that "models have reached a level of sophistication that renders data collection less important" and "data collection these days is principally to calibrate models". These comments strike the author as presumptuous and arrogant, and this paper is an attempt to place the dual activities of modelling and field data collection in some relative context.

Hydrological modelling has progressed to the stage where virtually any catchment process can be modelled. Indeed, modelling is at least as old as Darcy's equation (1856). This if course was significantly improved by Richards (1931) who added conservation of mass and unsaturated dispersion processes. By 1871 (Saint Venant, 1871) we had river hydraulics, and thus we had infiltration, groundwater flow and river routing. Along with runoff generation (e.g. Horton, 1933), either by simple partitioning or with kinematic surface flow we have the basic representation of water pathways once the rain strikes the ground.

A model is by definition, of course, a simpler representation of the real thing, and it is not a unique opinion that modelling is fine as long as it isn't confused with the real thing. I conjecture that in many cases models take on a life, especially in the eyes of their creators, that equals or exceeds the real world in importance. We have had whole conferences devoted to single models (such as the TOPMODEL workshop in 2000), and major experiments investigating not the real world, but inter-model comparisons, as if they were different races of a new species (e.g., PILPS, Henderson-Sellers et al., 1993).

As models progressed we have added vegetation interactions: rainfall interception (Rutter et al., 1971; Gash et al., 1980) and transpiration. The representation of vegetation can now include dynamic response to environmental conditions, with carbon assimilation, partitioning above and below ground in response to stresses in both 1-D (Dawes et al., 1998) and full catchment models (Silberstein et al., 1999b; Vertessy et al., 1996).

The single biggest advance for modelling the reverse flux – evaporation – came with Penman's model (1948) based on data collected during World War II, elegantly added to by Monteith in 1965 (Monteith, 1965), although he later contended that he received far more than his share of recognition for this. Thus we had methods to simulate runoff generation, subsurface unsaturated and saturated flow, river storage and routing and evaporation and transpiration.

2. MODEL SPECIES

Hydrological models fall into three main categories (Wheater et al., 1993; Grayson and Chiew, 1994) being conceptual, physics based and empirical or statistical. The vast majority of models in the literature are conceptual, in which a catchment is represented as a series of moisture stores, with fluxes between them and out of the catchment represented by parametric equations. The stores and fluxes are usually reminiscent of real identifiable attributes and processes, but are not generally independently measurable. The parameters must be estimated from input and output data (Ye et al., 1997). Empirical models are generally the simplest models and are utterly dependent on data as they represent the relationship between input and output series, generally as "transfer functions" between these series. Ye et al. (1997) discussed the performance of conceptual and empirical models in simulating the streamflow of semi-arid catchments in Australia. This study was one of a number of inter-model comparisons that are dependent on data not only for their tests, but the models themselves could not be even started without calibration data. Indeed, there is no reason to suppose that things should be otherwise.

In principle, physics based models may be operated without streamflow data for calibration, as they purport to represent the important physical processes with parameters that can be measured independently and assigned a priori to the relevant model characteristics (e.g. *SHE*, Abbott et al., 1986; *IHDM*, Beven et al., 1987; *TOPOG*, Vertessy et al., 1996). In the quest for truth and beauty these models have become so sophisticated their successors would seek to include everything. In practice, however, this approach is defeated by the lack of sufficient data to adequately parameterise the model, and by the fact that no model adequately represents the internal heterogeneity of the catchment unit. Thus the most sophisticated of catchment models still rely on input and output data for some level of calibration because there is never enough characterisation to avoid it. Grayson et al. (1992) argue that this is really because the models do not yet include all processes, and do not adequately deal with the internal heterogeneity. All models have spatial and temporal limits to their discretisation and description, which is another way of saying that the "scale problem" remains unsolved. While the physically based models cannot be seriously run without data, the empirical and conceptual models cannot really be run *at all* without such data.

In the last five years or so personal computers have indeed developed enough power that these models can be used to run on moderately sized catchments. In fact, there is a reasonable argument that we would not have bothered to develop these sophisticated models without the power to exploit it. Over the last 10 to 15 years, at the same time as this explosion in computer power and the associated advance in model sophistication, we have seen the development of much simpler models. TOPMODEL, for example (Beven and Kirkby, 1979) uses a relatively simple terrain based attribute (the "topographic wetness index") to represent catchment flow and water storage processes. In doing so it retains some of the internal complexity but is much simpler than the more complete physical representations.

At the simplest level of complexity are the simple "bucket" models, such as those presented by Boughton (1995). In these models, the whole catchment is represented by a single or a small number of buckets. The underlying principle being that a catchment has two major properties that control most of its response to drivers – the ability to store water and the release of that water. Depending on the storage size, relative to the timescale for forcing, and intensity of forcing, streamflow is generated. This idea was explored in some detail by Farmer et al. (2003) who showed that most catchments could be well represented with a small number of buckets, but the exact number depended on relatively few characteristics.

The next simpler level is possibly a model like LASCAM (Sivapalan et al., 1996a,b) in which a large catchment is broken in to a manageable number of sub-catchments, each represented as a single lumped entity. Most of the model parameters are determined by calibration on streamflow response to rainfall and potential evaporation drivers.

We have also had the development of what might be termed hybrid models, (such as Silberstein et al., 2003) in which relatively simple conceptual catchment models are coupled to more complex energy balance models to determine evaporation, and surface temperatures. The surface temperatures can then be linked to satellite data, either for model testing (Silberstein et al., 1999a) or as inputs (McVicar and Jupp, 1999).

The ultimate in model sophistication are probably the global climate models, that are run continuously on the world's most powerful computers, and are linked to so-called meso-scale meteorological forecasting models. We can now simulate the weather over the last 100's of years, and the next. These capabilities are truly awesome, and genuinely raise the question: Given we can now simulate how the world works so well, can we cut down the cost of collecting all this data that no one needs?

3. THE USE OF MODELS

Models serve three main purposes, firstly, they give us a framework to assemble our process understanding and to explore the implied system behaviours that come from that understanding. We can examine the model results, and consider whether they concur with our overall system analysis or not. If not, we have a structured framework to analyse whether it is our model or our overall understanding, or both, that is in error. The most dangerous thing in hydrology may be a model that fits with expectations (Dooge, 1988), because then we are not learning if we accept that our encapsulated understanding is adequate.

The second main use of a model is as a mechanism for testing data, to check for inconsistencies and errors, and to fill in missing information. It also gives us a method to explore the implications of our measurements. In fact, this may be the most useful function of models, because they help structure scientific enquiry that can elucidate further details behind observations.

The third use of models, and probably the most widely publicised and "commercial" use, is to explore scenario options, rather like the agroforestry example cited later. These may be options for management of a system, or exploring possible outcomes under a range of different input conditions, perhaps depending on future climate, political or economic scenarios. However, it is my contention that these activities should be confined largely to stimulate discussion, and always be tempered by some healthy scepticism and recognition of whole of system understanding.

3.1. Scenario Modelling

Using models like IQQM we could simulate water quality at any point along the Murray River (Viney et al., 2003, studied a small section), and with simple tools we can get fantastic pictures of different vegetation and salinity scenarios across the whole basin by using large GIS based models. The results of these modelling exercises are used to inform political and environmental decision making. The models are used because they are much cheaper and faster than doing real experiments. They also have the ability to predict things that we may not be able to do in the real world, or perhaps that would not really happen. The models are also used because they can give a nice set of pictures to represent possible outcomes.

An example of the use of models as a substitute for data was the application and analysis of 9000 simulations of trials (Silberstein et al., 2001) to distil out the over-riding principles for design guidelines for agroforestry in Australia. There is no way we would get resources to perform such a set of trials in a real catchment, and of course very little chance we would be given the 10 years required to make the measurements and study the outcomes. These simulations relied on a sophisticated biophysical hydrological model and intensive analysis to extract the emergent properties. This proved a huge task, but nothing compared to the job of carrying out real experiments.

4. MODELS DEMAND DATA

As models gain complexity, or expand the processes represented, the demand for data to calibrate and validate them increases. At the same time, as our technology improves and we have the ability to measure more attributes with greater precision, models expand to make use of these opportunities, for example the host of models being driven by remotely sensed data, especially surface temperature and vegetation cover data. Data availability now is truly impressive. In southern Western Australia we have a true 2 m digital elevation model to generate catchment networks, and over the whole of Australia the SILO data resource (Bureau of Meteorology and Queensland Department of Natural Resources and Mining http://www1.ho.bom.gov.au/silo/) has over 100 years of daily rainfall and 45 years of daily climate data from thousands of stations across the country, and anywhere in between. However, these examples are of data used to drive models, not to test them.

The literature is full of modelling and data studies, and while there are many studies exploring data without models, there are far fewer exploring models without data – at least in the natural sciences. Studies comparing models to data (Moore and Mein, 1975; Wheater et al., 1993; Grayson and Chiew, 1994; Ye et al., 1997; and so on) seek to assess model performance by comparing to data. The aim is to find a model that will work in any catchment, with any climate, and any vegetation. This may never happen, or certainly not for a long time, and until it does we know data are required. The hydrological community has been divided into "modellers" and "experimentalists", and despite the obvious benefits to both camps calls to combine forces (Dunne, 1983; Klemes, 1986) are rarely observed. Data-model comparisons invariably result in differences, and these differences are the source of insight into how natural systems work. Seibert and McDonnell (2002) explore a mechanism for getting the understanding of the experimentalist into the mathematics of the modeller, through use of "soft data", or experimentalists' understanding of "how the catchment works", in addition to the "hard data" of streamflow, climate and soil parameters modellers are used to. They conclude that it is better to be "less right for the right reason" than "right for the wrong reason".

4.1. Modelling difficult environments

Much of Australia is arid or semi-arid, conditions typically much harder to model than humid temperate catchments like those in most of Europe and North America. The hydrology of these regions is the subject of increasing interest (Pilgrim et al., 1988; Karnieli and Ben-Asher, 1993; Smakhtin, 2001), particularly as much of the worlds population lives in such regions. Many of these catchments (and most in Australia) also have low relief, with for example, the average slope in the Murray-Darling Basin being around 1:10,000 (and the Basin having an area 10^6km^2). These catchments typically have rainfall to potential evaporation ratios of much less than 1 and under natural vegetation have average annual streamflow less than 10% of rainfall. Such flow statistics are very difficult to get, because the flow may be zero much of the time. Catchments in south-west Western Australia, at 1200 mm rainfall typically have streamflow of about 10% rainfall, and at 700mm have streamflow of 2% rainfall. Such catchments are difficult to model because the primary drivers of streamflow are the residual of evaporation losses and soil moisture storage, and soil characteristics that control moisture redistribution. With fewer flow statistics parameter estimation is, of course, much more difficult. Because of the high evaporation, the catchment spends a large proportion of the time with soil moisture either too low for streamflow generation, or distributed such that discharge does not occur. Either way, a zero hydrograph gives no information about the water storage, or the redistributions that may be taking place. With evaporation being 90% or more of rainfall, then a 10% error in its estimation leads to 100% error in streamflow.

5. THE NEED FOR DATA

In Australia, as in the rest of the world, we in the hydrological community are beholden to the data collectors. Models are our attempt to encapsulate our understanding of the real world. They are not the real world, and without data they are simply imagination and computer games. Imagination, as we know thanks to Einstein's poster, is more important than knowledge, but it is not a substitute for it. Einstein made his special relativity insights by considering the data. Imagination is at its most useful when it can make use of knowledge or extend knowledge. We can fly to the moon and back on our computer, but it would not be half as interesting as the real trip. Hydrological modelling, as with any other, is not half as interesting as the real water world.

In many parts of the world, not the least Australia, data collection is shrinking by the day. The attitude of those in positions to make these decisions seems to be that we have enough data, and that not enough use is made of it. The view is that we do not need to collect because "we can model" and in any case "the main use of data is to calibrate models". These statements were made to the author by senior people charged with collecting hydrological data in Australia, and both of them indicate a significant lack of understanding of how science works, and how we should be using it to improve the management of our environment. Data are the real world. We don't need data if we don't care about the real outcome - if we are happy with computer games.

5.1. Data are needed in a changing environment

The thesis here is that data are essential. Modelling the impacts of land clearing on stream changes, flows and salinity would be impossible without data. The change to flow regimes, such as those in the Collie research catchments (Silberstein et al., 2003) that have converted ephemeral streams to perennial streams are completely beyond the data boundaries prior to these impacts. No model can demand great confidence when it goes beyond the boundaries of its data and calibration.

Australian taxpayers should demand that, having embarked on (another) $1,400M experiment that is the National Action Plan on Salinity and Water Quality, significant data are collected to actually confirm whether the experiment is a success. Government after government harangue their predecessors with assertions that throwing money at a problem is not the answer, and we are about to see the biggest single slab of money splurged at abating land degradation in history, with no guarantee of success, and not even a guarantee that we will measure whether it was a success. The "spin" from the government is that they have committed the funds, but not that the funds will be spent wisely or that they will even undertake an audit to see whether they have been. If a bank or a big energy company was to spend this kind of money there would be (or should be) checks and balances to audit the process, to stage the process to ensure that it was spent wisely, and to learn as the process was undertaken with mechanisms in place to improve outcomes as the programme proceeded. Australia currently spends about 1% of this amount on water monitoring. I suggest that a small additional investment would ensure that we could account for how well the other 99% was spent. We can use models to explore the possibilities but it is only data that will tell us which possibility has become a reality.

6. CONCLUDING REMARKS

Models are enormously useful as test beds for ideas, and for exploring the implications of our understanding of natural systems. They are also extremely valuable as data processing and analysis aids, often showing up data errors and inconsistencies that might otherwise have gone unnoticed. Models are also useful for exploring scenarios that cannot be tested in the real world. However, while this last use is a rapidly expanding one, it is also their most dangerous, as high level managers appreciate the nice graphics and simplistic sets of options it can be easy to lose sight of the limitations of the process that generated them. It is in this mode that models are often run outside their tested bounds, and by definition little or no data are available to constrain the scenario results. If we are to continue to learn about our environment and to continue to improve our management of it, we must continue to observe it – that means collect data. Modelling is a nice accompaniment to it, but is no substitute for it.

7. REFERENCES

Abbott, M.B., J.C. Bathurst, J.A. Cunge, P.E. O'Connell, and J. Rasmussen, An introduction to the European Hydrological System - Systeme Hydrologique Europeen, "SHE", 2: Structure of a physically-based, distributed modelling system, *Journal of Hydrology,* 87, 61-77, 1986.

Beven, K.J., A. Calver, and E.M. Morris, The Institute of Hydrology Distributed Model, Wallingford, England, Institute of Hydrology, 1987.

Beven, K.J. and M.J. Kirkby, A physically based, variable contributing area model of basin hydrology, *Hydrological Sciences,* 24(1), 43--69, 1979.

Boughton, W.C., An Australian water balance model for semiarid watersheds., in *Water research and management in semiarid environment, Proceedings of an international symposium, Tucson, Arizona, USA, 1-3 November, 1994,* 1995.

Darcy, H., *Les Fontaines Publique de la Ville de Dijon*, Paris, Dalmont, 1856.

Dawes, W.R., L. Zhang, and P.A. Dyce, WAVES V3.5 User Manual, Canberra, ACT, 2601, CSIRO Land and Water, 1998.

Dooge, J.C.I., Hydrology in perspective, *Hydrological Sciences Journal,* 33(1), 61-85, 1988.

Dunne, T., Relation of field studies and modeling in the prediction of storm runoff, *Journal of Hydrology,* 65, 25-48, 1983.

Elhames, A.S. and K.S. Richards, Progress in arid-lands rainfall-runoff modelling, *Progress in Physical Geography,* 18(3), 343-365, 1994.

Farmer, D., M. Sivapalan, and C. Jothityangkoon, Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis, *Water Resources Research,* 39(2),1035,10.1029/2001WR000328,2003.

Gash, J.H.C., I.R. Wright, and C.R. Lloyd, Comparative estimates of interception loss from three coniferous forests in Great Britain, *Journal of Hydrology,* 48, 89-105, 1980.

Grayson, R.B. and F.H.S. Chiew, An approach to model selection, in *Water Down Under 94 Symposium,* Institution of Engineers Australia, 507-512, 1994.

Grayson, R.B., I.D. Moore, and T.A. McMahon, Physically based hydrologic modeling. 2. Is the concept realistic? *Water Resources Research,* 28, 2659-2666, 1992.

Henderson-Sellers, A., Z.L. Yang, and R.E. Dickinson, The Project for Intercomparison of Land-Surface Parameterization Schemes, *Bulletin of the American Meteorological Society,* 74(7), 1335-1349, 1993.

Horton, R.E., The role of infiltration in the hydrologic cycle, *Transactions of the American Geophysical Union,* 14, 446-460, 1933.

Hughes, D.A., Monthly rainfall-runoff models applied to arid and semiarid catchments for water resource estimation purposes., *Hydrological Sciences Journal,* 40, 751-769, 1995.

Karnieli, A. and J. Ben-Asher, A daily runoff simulation in semi-arid watersheds based on soil water deficit calculations., *Journal of Hydrology,* 149, 9-25, 1993.

Klemes, V., Dilettantism in hydrology; transition or destiny? *Water Resources Research,* 22(9), 177S-188S, 1986.

McVicar, T.R. and D.L.B. Jupp, Estimating one-time-of-day meteorological data from standard daily data as inputs to thermal remote sensing based energy balance models, *Agricultural and Forest Meteorology,* 96, 219-238, 1999.

Monteith, J.L., Evaporation and environment, in *The State and Movement of Water in Living Organisms, Symposium of the Society of Experimental Biology* , Fogg, G.E.(Ed.), New York, Academic Press, 205-234, 1965.

Moore, I.D. and R.G. Mein, Evaluating rainfall/run-off model performance, *Journal of Hydraulics Division, ASCE,* 102, 1390-1395, 1976.

Penman, H.L., Natural evaporation from open water, bare soil and grass, *Proceedings of Royal Society,* A 193, 120-45, 1948.

Pilgrim, D.H., T.G. Chapman, and D.G. Doran, Problems of rainfall-runoff modelling in arid and semiarid regions., *Hydrological Sciences Journal,* 33, 379-400, 1988.

Richards, L.A., Capillary conduction of liquids in porous mediums, *Physics,* 1, 318-331, 1931.

Rutter, A.J., K.A. Kershaw, P.C. Robins, and A.J. Morton, A predictive model of rainfall interception in forests. I: Derivation of the model from observations in a plantation of Corsican Pine, *Agricultural Meteorology,* 9, 367-384, 1971.

Saint-Venant, Barre de, Theory of unsteady water flow, with application to river floods and to propagation of tides in river channels, *French Academy of Science,* 73, 148-154, 237-240, 1871.

Seibert, J. and J.J. McDonnell, On the dialog between experimentalist and modeller in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resources Research,* 38(11), 2301-2314, 2002.

Silberstein, R.P., M. Sivapalan, and A. Wyllie, On the validation of a coupled water and energy balance model at small catchment scales, *Journal of Hydrology,* 220, 149-168, 1999a.

Silberstein, R.P., R.A. Vertessy, J.D. Morris, and P.M. Feikema, Modelling the effects of soil moisture and solute conditions on long-term tree growth and water use: a case study from the Shepparton irrigation area, Australia, *Agricultural Water Management,* 39(2-3), 283-315, 1999b.

Silberstein, R.P., R.A. Vertessy, and D. Mc Jannet, A parameter space and simulation response surface for agroforestry design, in *MODSIM 2001*, The Modelling and Simulation Society of Australia and New Zealand Inc., 1889-1894, 2001.

Silberstein, R.P., M. Sivapalan, N.R. Viney, A. Held, and T.J. Hatton, Modelling the energy balance of a natural jarrah (*Eucalyptus marginata*) forest in Western Australia, *Agricultural and Forest Meteorology,* 115(3-4), 202-232, 2003.

Sivapalan, M., J.K. Ruprecht, and N.R. Viney , Water and salt balance modelling to predict the effects of land-use changes in forested catchments. 1. Small catchment water balance model, *Hydrological Processes,* 10, 393-411, 1996a.

Sivapalan, M., N.R. Viney, and C.G. Jeevaraj , Water and salt balance modelling to predict the effects of land-use changes in forested catchments. 3. The large catchment model, *Hydrological Processes,* 10, 429-446, 1996b.

Smakhtin, V.Y., Low flow hydrology: a review, *Journal of Hydrology,* 240, 147-186, 2001.

Vertessy, R.A., W.R. Dawes, L. Zhang, T.J. Hatton, and J. Walker, Catchment-Scale Hydrologic Modelling to Assess the Water and Salt Balance Behaviour of Eucalypt Plantations, Technical Memorandum No. 96/2, CSIRO Div. Water Resources, 1996.

Viney, N.R., B.C. Bates, S.P. Charles, I.T. Webster, M. Bormans, and S.K. Aryal, Impacts of climate variability on riverine algal blooms, in *MODSIM 2003* Int. Congress on Modelling and Simulation, 2003.

Wheater, H.S., A.J. Jakeman, K.J. Beven, M.B. Beck, and M.J. McAleer, Progress and Directions in Rainfall-Runoff Modelling., Modelling Change in Environmental Systems., Chichester, UK, John Wiley Sons Ltd, 1993.

Ye, W., B.C. Bates, N.R. Viney, M. Sivapalan, and A.J. Jakeman, Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments., *Water Resources Research,* 33, 153-166, 1997.