Hydrosanity: a starting point for hydrological analysis

Andrews, F.T. [1, 2]

[1] Integrated Catchment Assessment and Management Centre (iCAM), Fenner School of Environment and Society, The Australian National University, Canberra
[2] Department of Mathematics, The Australian National University, Canberra
Email: felix.andrews@anu.edu.au

Keywords: *Exploratory data analysis, visualisation, sanity checks, rainfall, streamflow*

EXTENDED ABSTRACT

This paper describes the software *Hydrosanity*, a starting point for hydrological analysis. It was developed to allow exploratory data analysis of catchment hydrology data, to check model assumptions and identify inconsistencies in the data. The methods are based on interactive graphical displays where possible. It is argued that rapid visualisation and arbitrary extensibility (through integration with a full statistical language) are necessary for effective exploratory data analysis with hydrological data.

Methodological features available in the Hydrosanity software include: iterative refinement of study scope; data management and transformation; dataset overview; basic visualisation; checks for transient data errors; checks for systematic errors and trends; and water balance checks. Major technical features used to support these are: *imputing* (modelling rainfall based on neighbouring sites, to fill gaps); and estimation of *areal rainfall*. In both cases, the intent is to provide several alternative methods.

The features are demonstrated with a brief case study application in Tuross River catchment in coastal NSW, Australia. Daily rainfall and stream flow data were analysed.

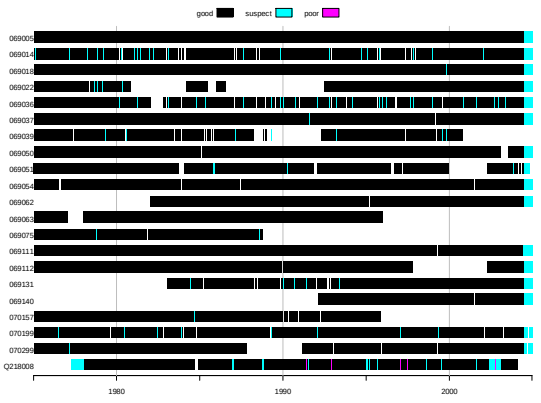

Figure 1. Data timeline for 20 rain gauge sites in and around Tuross catchment, and a stream gauge.

Spatial and temporal data summary views are available. The data timeline (Figure 1) is useful in showing periods of missing data and varying data quality. The spatial mosaic of rain gauge data (Figure 2) gives context to multiple time series, which helps to identify inconsistencies in the data.

As an error-checking test, rainfall values were imputed from nearby sites and plotted against the actual values, and outliers were identified. These outliers point to inconsistencies in the data between otherwise comparable sites, which may be data errors or localised events.

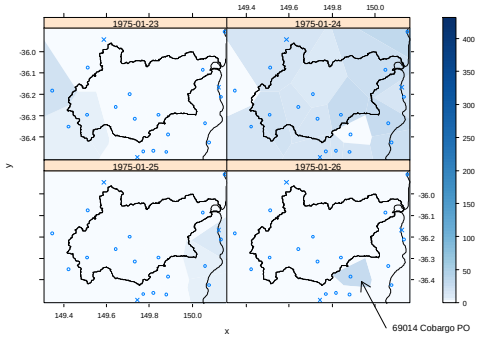

Figure 2. Spatial mosaic of rain gauge data, showing a sequence of four time steps (days). A likely timing error presents itself, at site 69014.

A typical model assumption is the form of the *rainfall-runoff relationship*. This can be visualised by plotting rises in stream flow as a function of areal rainfall. The relationship can be further explored by *conditioning* on other variables to produce multiple panels: here it can be conditioned on *season*, *time*, and/or *antecedent flow*.

Hydrosanity is *free software*, made available under the *GNU General Public License*. You therefore have permission to inspect, improve and reuse the source code (see the licence for details). It is implemented as a package for R (R Development Core Team, 2007), which is also free software. The package is online at http://hydrosanity.googlecode.com/.

1. INTRODUCTION

Many problems in hydrology centre on recorded measurements of components of the water cycle, most obviously rainfall, surface water flow and evaporation. The processes involved are complex and depend strongly on local conditions. Thus the models that hydrologists construct to explain and predict these processes necessarily make many simplifying assumptions. Furthermore, quantitative models generally take their parameter values from observational data, which is subject to various kinds of error. However, published modelling studies usually do not report convincing tests of the model assumptions and data integrity. Such tests are sometimes called *sanity checks*.

No matter how many sanity checks are performed, there will still be possible faults with the data and model assumptions. It is therefore important that the data and analysis procedure in a published study be specified clearly enough that others can test it. Indeed, this is arguably fundamental to the scientific method. Specifically, the *replication standard* holds that "sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." (King, 1995).

The best way to understand a data set is to explore it without a pre-stated method: Exploratory Data Analysis (EDA; Tukey, 1977). This can be done efficiently with modern interactive visualisation. Also, a graphical user interface helps to move rapidly between different displays and different methods, without the cost of learning syntax. However, it is sometimes necessary to break out of the limits imposed by an interface, to experiment with methods using the full power of a statistical language (Fox, 2005).

Catchment hydrology data consists of multiple time series at point locations, and also involves working with continuous spatial fields. EDA with such complex data structures would be difficult without specially designed software support.

There is a great deal of existing software for working with hydrological data, although research software tends to be distributed only informally if at all. One commonly used package is the Hydstra Time-Series Data Management Suite, previously known as HYDSYS (http://www.kisters.com.au/). This provides a large set of tools for working with hydrological time series, including many types of summaries and plots. However, it is proprietary software and thus does not allow arbitrary modifications to the implemented methods, and has limited capabilities for data analysis compared to a full statistical environment.

There is also plenty of general-purpose software for data analysis, including time series analysis and spatial data analysis. Suffice it to say that the author has not yet found a tool that meets the twin aims of rapid visualisation and arbitrary extensibility that are necessary for effective EDA with hydrological data.

This paper describes the software *Hydrosanity*, a starting point for hydrological analysis. It was developed to allow exploratory data analysis of catchment hydrology data, to check model assumptions and identify inconsistencies in the data. The methods are based on interactive graphical displays where possible.

As this is a tool for scientific investigations, all actions from the interface are recorded in a log, in the underlying statistical language. This log helps the practitioner to understand exactly what is happening, to modify it, and also to learn to use the more powerful statistical language. The log is kept in a project file along with the data, which can be distributed to support replication. The software also has potential for teaching purposes.

Hydrosanity is *free software*, made available under the *GNU General Public License*. You therefore have permission to inspect, improve and reuse the source code (see the licence for details). It is implemented as a package for R (R Development Core Team, 2007), which is also free software. The package is online at http://hydrosanity.googlecode.com/.

2. FEATURES

Methodological features available in the Hydrosanity software are listed here. They are demonstrated with a brief case study in the next section.

2.1. Iterative refinement of study scope

Users of this software may not have a well-defined problem scope. The time period and spatial extents chosen may well depend on what data turns out to be available, and change as one begins to understand the data. To support this, Hydrosanity can *interactively* import time-series from a database (such as measurements from rainfall stations), as follows: (i) the stations are displayed on a map and the user selects a region; (ii) the available data from each station in that region is shown on a timeline, and the user selects a time period; (iii) a minimum number of years of data

can be set, so that stations are only imported if they have sufficient data within the chosen time period. This process can be iterated to refine the study scope.

This procedure is currently implemented for the Australian Daily Rainfall Data archive (Australian Bureau of Meteorology, 2006), but could be linked to other databases.

2.2. Data management and transformation

Whether time series are imported from a database or from files, they need to be managed and perhaps transformed. Metadata for each item includes the identification number, name, units, start and end times, time step, location, elevation, data role (e.g. rain or flow), as well as data quality codes for each time step. Ultimately this should also include uncertainty estimates, such as the resolution limit of the data (Croke, 2007); support for this may be included in future.

Hydrosanity uses three-level quality codes based on the idea of rough sets: *good*, *suspect* and *poor*, as well as codes for *imputed* and *disaccumulated* values. Existing quality codes can be mapped into this scheme.

Rainfall time-series often have a peculiar type of partially-missing data: if rainfall was not recorded it accumulates until the next time step. It is common for Australian rainfall to be accumulated over weekends. This needs to be accounted for, and the accumulated values can be *disaccumulated* prior to analysis. This is trivial for analysis at an aggregated temporal resolution (it is enough to uniformly smooth out the accumulated value). For analysis at the data resolution a method of *imputing* is used to fill in the gaps (described in section 2.5) and the imputed values are then scaled to match the accumulated total.

Time series can be transformed by aggregation to a longer time step, and by taking the ratio of series. Arbitrary transformations can be done using the R syntax. And since Hydrosanity is just a starting point for analysis, care was taken to ensure the data can be exported flexibly.

2.3. Dataset overview

One important summary is the data timeline. It shows the times for which data is available at each site. Data quality codes are shown if available. This display reveals how much simultaneous data is available, and highlights gaps and malfunctions.

A literal "overview" of the dataset is provided by a map of sites. This is enhanced with other spatial layers for context: notably catchment boundaries (which can be imported) and coastlines. Elevation can be interpolated from the point locations as a rough guide, or imported from a file.

There is also a numerical summary of the data available within the chosen time period, consisting of: overall data coverage, the number of sites that are typically active at any one time, and "five-number" distribution summaries (Tukey, 1977).

2.4. Basic visualisation

The most basic visualisation for this data is a time-series plot. In this context a notable enhancement is to highlight data quality codes. These plots can show multiple series superposed or in parallel, and may be augmented by aggregating or smoothing the data. As with all other plots described here, interactivity is a key feature. The user can zoom in to investigate possible problems in detail.

The empirical cumulative distribution of data values is a useful graphical summary, particularly for comparing sites and identifying extreme values. Disaggregating into seasonal or monthly distributions is also useful, to find the strength and period of seasonality, as well as to assess the consistency of this pattern between sites. Box-and-whisker plots are suitable here (Tukey, 1977).

To help understand the spatial dimension, maps showing rain gauge data as a mosaic are available. Such maps can be constructed to show a series of single time-steps, so the user can flick through the data to look for anomalies. Also available are annual time series (annual anomalies), the map of overall mean rainfall, and maps of mean rainfall in each season. Sophisticated spatial interpolation techniques such as Kriging and thin plate smoothing splines (e.g. Hutchinson, 1998; Jeffrey et al., 2001) are not used in Hydrosanity, although such functions do exist in other R packages.

2.5. Checks for transient data errors

There are many possible types of transient errors in hydrological data. One is timing error: the data are accurate but recorded for the wrong time (e.g. day). These can be detected by estimating the lag between pairs of time-series, and plotting it over time. Typically many rainfall sites will be plotted against one reference site, which could be either rainfall or stream flow.

Once timing errors have been eliminated as much as possible, the data values at each site can also be

plotted against those of a reference site. This gives a visual indication of the (perhaps non-linear) relationship between time-series. Further, the data can be aggregated, to produce scatterplots with lower temporal resolution. One important use of this is to reveal which rainfall stations have a good direct relationship with stream flow. These can then be further analysed as described in section 2.7.

A complementary test to the scatter-plot just mentioned, is a plot of correlation over time (calculated in a moving window). This can indicate whether the relationship between two series breaks down at some point, which is likely to indicate transient data errors.

Hydrosanity implements a model-based approach to detecting inconsistencies in rainfall data. A model is constructed to estimate data at one site from data at other sites. Then the model-estimated values are plotted against the original observed values. This can help to identify errors, such as gauge calibration tests, transcription errors or false zeros, which are confined to one site. It will also pick up actual localised events if the set of rain gauges is too sparse.

These model-based predictions can also be used to fill gaps in time-series, a procedure known as *imputing*. There are two kinds of models available: temporal prediction based on correlations, and areal prediction based on distances. Areal prediction can be biased where topographic effects on rainfall are significant. On the other hand, the temporal method (as it is currently implemented) may be inefficient as it only uses the one site with best correlation to predict where possible, the site with second-best correlation to predict in cases where the first site had no data, and so on.

2.6. Checks for systematic errors and trends

Systematic errors are usually more important than random errors in their effect on model results. One kind of systematic error may lie in the distribution of missing values. For instance, the measuring equipment may be more likely to fail during storms. This is known as *non-ignorable missingness* (Honaker and King, 2006). Models often assume that values are *missing completely at random* (MCAR), where the occurrence of missing values is unrelated to the actual values. Hydrosanity includes a test of this, which is to compare the distribution of imputed values for times when data is missing with the distribution of imputed values for times when data is not missing.

Another kind of model assumption is *stationarity*: that the underlying distribution remains constant over time. As well as simple checks for “trends” (in fact any systematic change over time) with non-parametric smoothing, changes in the ratio of series can be tested. This may indicate shifts in broad weather patterns over the study region.

2.7. Water balance checks

Hydrological models are generally based on accounting for water movement through the water cycle, and the mass of these stocks and flows must balance. Models often account for rainfall, changes in soil moisture, surface water runoff, groundwater gains and losses, and evapo-transpiration. But in catchment hydrology, the most basic quantity is rainfall. This requires extrapolating from point measurements to areal (total rainfall over the catchment area), which is inherently difficult due to its spatio-temporal variability (Jeffrey et al., 2001).

Hydrosanity includes some facilities for constructing areal rainfall time series. The simplest method is to extrapolate each rainfall gauge to the surrounding area for which it is the closest (the Thiessen polygon method). But this assumes that each gauge is representative of the total rainfall over a large area, ignoring local variations: a big assumption. A better estimate of areal rainfall can be produced by sophisticated spatial models, which take account of elevation and climatic patterns. If spatial grids of rainfall (output from such models) are available, their areal average value over the catchment can be extracted as a time series. For instance, if spatial grids are available for each year in sequence, an annual time series can be extracted. Finally, areal rainfall from spatial grids can be *downscaled* using the rain gauge data. The resulting time series combines the temporal pattern of each gauge (say, at daily scale) with the aggregated totals of the grids (say, at annual scale).

Once areal rainfall has been estimated, several other analyses become possible. The ratio of stream flow at the catchment outlet to areal rainfall is known as the *runoff coefficient*. This should be calculated over a time scale long enough for residual stream flow to be negligible. It represents the proportion of rain that flows through the catchment outlet, the remainder being accounted for by evapo-transpiration, and the change in groundwater and soil storage. The runoff coefficient should be less than 1 (unless there is significant subsurface flow into the catchment), and its value should be consistent with physical knowledge of the aridity of the region,

imperviousness of the catchment surfaces, and so on. Any systematic changes in its value over time demand explanation.

Areal rainfall can also be plotted against stream flow to show the *rainfall-runoff relationship*. Actually the *increase* in flow should be plotted, to show the response of flow to each input of rainfall. Furthermore, at small time steps compared with the time taken for water to reach the outlet, the rainfall needs to be lagged to correspond to flow peaks.

The rainfall-runoff relationship plot generally reveals two things: a (fuzzy) threshold – the level of rainfall required to produce runoff – and the response above the threshold (Boughton, 2004). These are typical parameters of hydrological models.

The rainfall-runoff relationship can be further explored by *conditioning* the relationship on other variables to produce multiple panels. This is known as Trellis graphics (Cleveland, 1993). Here it can be conditioned on *season*, producing four plots if canonical seasons are used. The seasons may have different climatic patterns, such as more or less intense storm events. The relationship can also be conditioned on *time*, dividing up the study period into several sub-periods. This might be used to check for changes in catchment response over time. Finally, the relationship can be conditioned on *antecedent flow*, which is an indicator of catchment wetness. The form of this relationship is a typical model assumption (e.g. Croke and Jakeman, 2004).

3. CASE STUDY

Tuross River catchment is a coastal catchment in NSW, Australia. The data considered here are daily rainfall from the Bureau of Meteorology Australian Daily Rainfall Data (2006 edition), and daily stream flow from Pinneena 8 (NSW Department of Natural Resources, 2004). The replication project file is available online at http://hydrosanity.googlecode.com/files/modsim07.hydrosanity.

Rainfall stations were selected using the iterative procedure described in Section 2.1. A snapshot of the process is shown in Figure 3. The time period 1970-01-01 to 2006-01-01 was chosen, and a criterion was set for sites to have at least 10 years of data in that period. The spatial region chosen was from 149.3° to 150.2° longitude and from -36.5° to -35.9° latitude. Finally 20 sites met the criteria and were imported: see Figure 4.

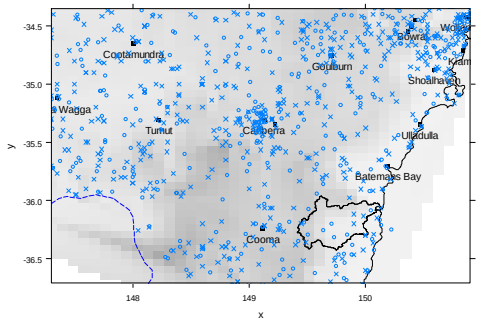

Figure 3. Navigating the Australian Daily Rainfall Database. Sites were considered acceptable if they had 10 years of data in the period 1970-2006, and are shown as circles. The Tuross catchment boundary is overlaid. Shading shows elevation interpolated from sites.

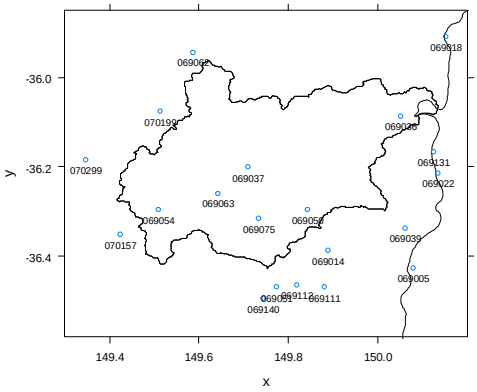

Figure 4. A map of the 20 sites selected from the database, with Tuross catchment boundary.

The data timeline shown in Figure 1 confirms that all selected sites have at least 10 years of data. Most of the data has a "good" quality code, with data in the most recent months marked "suspect" as it had not yet been checked by the Bureau of Meteorology. Many short gaps are visible, most of which are due to accumulations over several days.

A text summary corresponding to the data timeline is available, as follows:

- Overall, 18% of data is missing. There are 20 time series, of which 0 are complete. 7 are > 95% complete and 13 are > 75% complete.

- No time steps have data from all series. The median number of active sites is 16 (80%). Half the time, the number of active sites is between 15 and 17.

Such information is important context for modelling studies based on this data.

One stream flow time series was imported (simply from a file): site 218008 “Tuross River at Eurobodalla”, which has a catchment area of 1586 km^2, the largest in the study region. Also, a subset of six rain gauges was chosen for further analysis: sites 69036, 69037 69050, 69054, 69075 and 70199. These collectively represent most of the catchment. The remaining sites act as spatial context, and as inputs for imputing missing values.

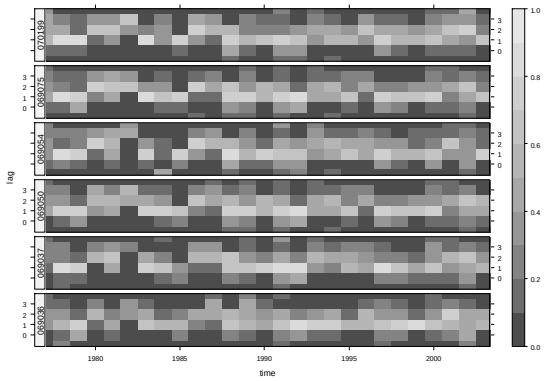

Figure 5. Time series of cross-correlation between each of six rainfall series and stream flow rises, aggregated into years. The lag (in days) is marked white where correlation with flow rises is strong.

To test for systematic timing errors, cross-correlations were calculated between rain gauges and stream flow for each year of record. The result is shown in Figure 5. The lag between rain and flow rises is about 1 or 2 days. Variation in lag seems to be consistent across rain gauges, suggesting there are no severe timing errors at individual gauges. It does vary over time, probably reflecting the lag time of one or two events each year, which may vary in such a large catchment.

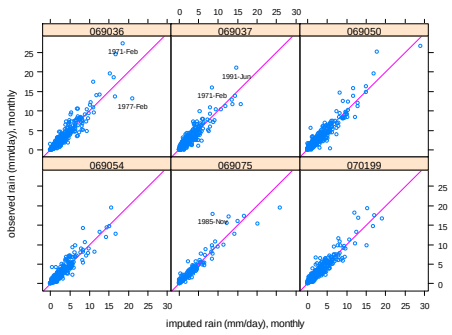

Figure 6. Imputed vs. actual values of daily rainfall, aggregated to monthly, at each of six selected rainfall sites. Some outliers are marked.

As an error-checking test, rainfall values at each of the six selected sites were imputed from other nearby sites. The correlation method was used. Imputed values were then plotted against the actual values, and outliers identified (Figure 6). These outliers point to inconsistencies in the data between otherwise comparable sites, which may be localised events or data errors. At this aggregated time scale they are most likely not timing errors. Inconsistencies in the raw data can be followed up with a spatial time series plot, to put it in spatial and temporal context: see Figure 2.

Areal rainfall was estimated with the downscaling method described in section 2.7. The spatial grids of rainfall used were generated by ANUSPLIN (Hutchinson, 1998), consisting of one grid per month from 1968 to 2006. Monthly areal values from the grids (over each Thiessen polygon) were used to scale the daily data from each gauge; these scaled time series were then combined with area weighting as in the simple Thiessen polygon method. Prior to estimating areal rainfall, gaps in all rain gauge time series were filled by imputing, as described above.

The runoff coefficient was calculated by dividing the flow time series by the areal rainfall time series in a 1-year moving window. The result is shown in Figure 7. Decadal-scale variations can be seen.

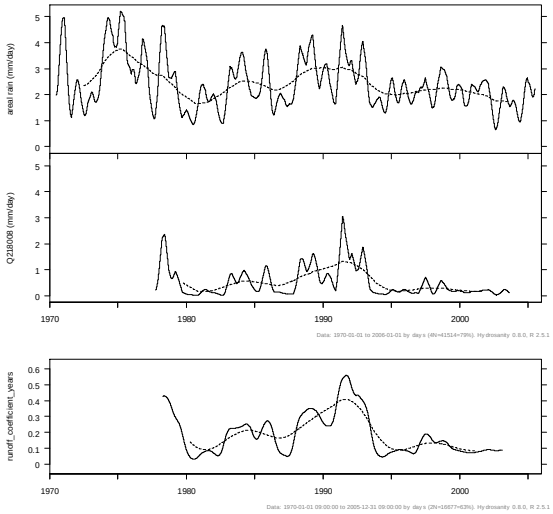

Figure 7. Time series of catchment areal rainfall (top), stream flow (middle) and their ratio over a one-year window, the runoff coefficient (bottom). Each series is smoothed with (triangular) kernels of width 1 year and 5 years.

Finally, the rainfall-runoff relationship was plotted (Figure 8), conditioned on two levels of antecedent flow. This reveals an apparent threshold level of rainfall as expected, but there is not an obvious effect from antecedent flow. At this daily scale, the relationship is affected by the inconsistent lag as shown in Figure 5. There are further complexities in the data that are not pursued here.

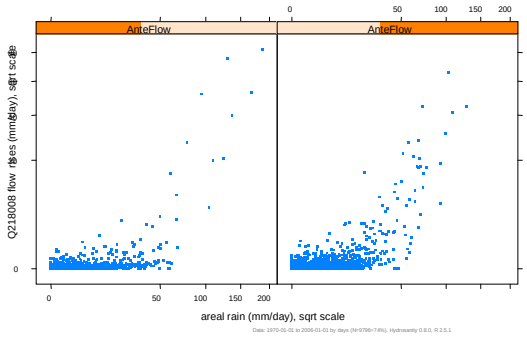

Figure 8. Rainfall-runoff relationship showing flow rises at site 218008 as a function of areal rainfall over its catchment, lagged by 1 day. The panels show subsets of the data based on antecedent flow (left panel has days with lowest half of antecedent flows, right panel has the rest).

4. FURTHER DEVELOPMENT

A priority for all the facilities in Hydrosanity is to allow variations in methodological choices, and to easily see their effect. Thus different methods will be implemented for imputing rainfall, estimating areal rainfall and calculating stream flow statistics. In particular, techniques of *multiple imputing* (Honaker and King, 2006) should help to characterise the uncertainty due to missing values.

Arguably the most powerful sanity tests come from applying models with explicit assumptions: *model-based tests*. One approach is to fit a simple model but allow one or more parameters to vary over time to fit the data, as in the Linear Time-Varying (LTV) method described by Norton and Chanat (2005). Time-varying parameter estimates can reveal un-modelled behaviour, and thus suggest other factors that are relevant. This approach may be trialled.

As Hydrosanity is *free software*, developments may come from the wider hydrology community.

5. CONCLUSION

Rapid visualisation and arbitrary extensibility (through seamless integration with a full statistical language) are necessary for effective EDA with hydrological data. This includes checks for transient and systematic errors, and other tests of model assumptions. This paper described free software developed to support such analysis.

ACKNOWLEDGMENTS

I am grateful for many helpful discussions with Tony Jakeman and Barry Croke. The Hydrosanity Graphical User Interface was based on *Rattle* by Graham Williams (http://rattle.togaware.com).

6. REFERENCES

Boughton, W. (2004), The Australian water balance model, *Environmental Modelling & Software*, 19(10), 943-956.

Cleveland, W.S. (1993), *Visualizing Data*, Hobart Press, Summit, New Jersey, U.S.A.

Croke, B.F.W. (2007), The role of uncertainty in design of objective functions, this issue.

Croke, B.F.W. and A.J. Jakeman (2004), A Catchment Moisture Deficit module for the IHACRES rainfall-runoff model, *Environmental Modelling and Software*, 19.

Fox, J. (2005), The R Commander: A Basic-Statistics Graphical User Interface to R, *Journal of Statistical Software*, 14(9).

Honaker, J. and G. King (2006), What to do about Missing Values in Time Series Cross-Section Data, *(preprint)* URL: http://gking.harvard.edu/preprints.shtml

Hutchinson, M.F. (1998), Interpolation of Rainfall Data with Thin Plate Smoothing Splines - Part I: Two Dimensional Smoothing of Data with Short Range Correlation, J*ournal of Geographic Information and Decision Analysis*, 2(2), 139-151.

Jeffrey, S.J., J.O. Carter, K.B. Moodie and A.R. Beswick (2001), Using spatial interpolation to construct a comprehensive archive of Australian climate data, *Environmental Modelling & Software*, 16, 309-330.

King, G. (1995), Replication, replication, *PS: Political Science and Politics*, XXVIII(3) 443-499.

Norton, J.P. and J.G. Chanat (2005), Linear time-varying models to investigate complex distributed dynamics: A rainfall-runoff example, *Mathematics and Computers in Simulation*, 69, 123-134.

R Development Core Team (2007), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org

Tukey, J.W. (1977), Exploratory Data Analysis.