Prediction intervals for rainfall–runoff models: raw error method and split-sample validation

John Ewen and Greg O'Donnell

ABSTRACT

A method (the ghost method) is developed here that calculates prediction intervals for the discharge hydrograph for a river catchment. It uses a calibrated rainfall–runoff model and a dataset containing raw errors such as residuals between observation and simulation. When calculating prediction intervals, raw errors are selected from the dataset and applied to the simulated hydrograph. The selection method is based on matching the simulated hydrological conditions to the hydrological conditions associated with the raw errors. To test the method, the split-sample calibration-validation approach advocated by Klemeš and used widely in hydrology is extended so that the data available for calibrating and testing are divided into three parts rather than two, called periods A, B and C. The rainfall–runoff model is calibrated for period A. For period B, the method by which prediction intervals are calculated is calibrated to give a specified high level of containment (e.g. 99% of observations lie within the prediction interval). Period C is used for testing, carried out in a way that shows the performance expected under operational conditions for real-world problems. Prediction intervals are calculated for the Hodder catchment, northwest England.

Key words | calibration, prediction interval, rainfall–runoff modelling, uncertainty

John Ewen (corresponding author)
Greg O'Donnell
School of Civil Engineering and Geosciences,
Newcastle University,
Newcastle upon Tyne,
NE1 7RU,
UK
E-mail: *john.ewen@tiscali.co.uk*

INTRODUCTION

The problem studied here is the prediction of the discharge at the catchment outfall for a period of time when the discharge is not known, based on data available for a period when the discharge is known. This is relevant to practical problems such as reconstructing historical discharge hydrographs (Gyau-Boakye & Schultz 1994) and the use of synthetic time series datasets for rainfall and evaporation when estimating the impacts of variability and change in climate and land use management (Reynard *et al.* 2001; Diaz-Nieto & Wilby 2005; Sivapalan *et al.* 2005; Fowler & Kilsby 2007). The aim is to create accurate prediction intervals that comprise an upper and lower bounding time series which show the likely range within which observations would be expected to lie.

The source of information on prediction intervals is the simulated and observed discharges for the period of time where the discharge is known (the calibration period). Information must therefore be carried forward from the calibration period to the period of time when the discharge is not known (the prediction period). Traditionally, the method for carrying forward information is calibration using split-sample methods (Klemeš 1986). In the simplest case, the calibration period is divided into two periods, A and B. The model is calibrated for period A to maximise some performance measures (i.e. measures of the quality of the simulation such as the Nash–Sutcliffe Efficiency (NSE); Nash & Sutcliffe 1970), then the calibrated model is tested for period B. If the testing is successful, the calibrated model is then applied to the prediction period. In this traditional approach, information is carried forward in the form of: (1) a rainfall–runoff model; (2) a calibrated parameter set; and (3) a set of performance measures.

In the past decade or so, catchment rainfall–runoff modellers have concentrated on the direct calculation of prediction intervals, with associated probabilities. For example, an interval may be calculated in such a way that

doi: 10.2166/nh.2012.038

it might be expected to contain 95% of the available observations (or potential observations if no observations are available). The obvious way to calculate a prediction interval is to use statistical methods, including Monte Carlo approaches and Bayesian statistical techniques (e.g. Beven & Binley 1992; Ewen & Parkin 1996; Thiemann *et al.* 2001; Krzysztofowicz 2002; Georgakakos *et al.* 2004; Kavetski *et al.* 2006b; Mantovan & Todini 2006; Beven *et al.* 2008; Blasone *et al.* 2008; Xiong & O'Connor 2008; Liu *et al.* 2009; Jin *et al.* 2010).

When using these approaches, information is carried forward in the form of: (1) rainfall–runoff models; (2) error models, such as models for the statistical properties of the residuals; (3) parameter sets; and (4) likelihood values associated with parameter sets. Reading the literature on this topic, it is very clear that there is no consensus about how best to estimate prediction intervals, especially when working with the very messy datasets and practical constraints faced in real-world problems (Montanari *et al.* 2009). However, work is progressing on several fronts, including on handling the effects of heteroscedastic errors, non-normal errors, auto-correlated errors, errors in forcing datasets, non-stationary error characteristics and the use of machine learning techniques (Kavetski *et al.* 2006a; Yang *et al.* 2007; Vrugt *et al.* 2008; Solomatine & Shrestha 2009; Thyer *et al.* 2009; Renard *et al.* 2010; Schoups & Vrugt 2010).

Simulations use and can produce a huge set of raw information, such as time series for: (1) rainfall; (2) evaporation; (3) residuals (i.e. difference between simulated and observed discharge); (4) simulated hydrographs; (5) time derivatives of simulated hydrographs; and (6) sensitivities of simulated hydrographs to the parameters of the rainfall–runoff model. Ewen (2010) recently suggested that it is realistic and reasonable to carry forward a huge amount of raw information, and thus perhaps eliminate or bypass some of the problems faced when creating prediction intervals for real-world problems. This is all quite speculative, so this paper describes some very basic testing of the approach. A dynamic programming method is described by Ewen (2011) that, in theory, allows raw information on magnitude and timing errors to be carried forward. However, the only information on error that will be carried forward here is for magnitude errors (i.e. residuals).

Two rainfall–runoff models are used so that the effect of differences in model quality can be studied: results for an 8-parameter version of the Probability Distributed Model (PDM; Moore 2007) are compared against results for a simple two-parameter Two Bucket Model (TBM). PDM and TBM are described in the Appendix. The data used are for the Hodder catchment, northwest England (Ewen *et al.* 2010). There are some errors and limitations in the dataset which could, with further work, be partly eliminated (these are discussed later). However, the aim is to have a robust method for estimating prediction intervals that works for real-world problems; it is normal, rather than exceptional, to have to work with datasets that have errors and limitations. The approach taken to testing extends the split-sample validation approach advocated by Klemeš. The testing is carried out under operational conditions. Basically, this means that the data used in testing were not made available until all the calibration work had been completed. This is in the spirit of the 'blind' validation approach of Ewen & Parkin (1996), in which measures are taken to ensure that the modeller is not contaminated by knowing the likely final outcome of testing while still involved in calibration (because for real-world problems, by their very nature, there are usually no discharge data for the time period for which the predictions are to be made).

METHOD

Three time periods, A, B and C, are used in a split-sample approach. These are used for: (a) calibrating the rainfall–runoff model; (b) calibrating the method used to calculate the prediction interval; and (c) testing the quality of the prediction interval when the approach is used under operational conditions. For period A, the rainfall–runoff model is calibrated to optimise some measure or measures of performance for the comparison between the simulated discharge q_{sim} and the observed discharge q_{obs}. Once the optimum parameter set has been found for the rainfall–runoff model, the residual errors ($r=q_{obs}-q_{sim}$) are saved along with q_{sim} and any other raw data thought relevant to defining the hydrological state to which the residuals correspond. For period B, a method (the ghost method) is used that calculates upper and lower prediction bounds using

the calibrated rainfall–runoff model and the saved raw data. The ghost method, so-called because it gives plots that show ghosts of previous storm responses, is calibrated to minimise the width of the prediction interval for period B for a specified level of containment (95 and 99% are used here). For period C, the calibrated rainfall–runoff model and ghost method are tested to see if, under operational conditions, the containment and width of the prediction intervals are adequately close to the values expected.

When the model is calibrated for period A, the final task is to calculate and save the values associated with the residuals and hydrological conditions. The saved raw data includes a set of residuals, and for each residual there is an associated set of values for hydrological conditions. Here, the data for the hydrological conditions include only: (1) the simulated discharge at the time where the residual applies; and (2) the rate of change in simulated discharge, as given by its average over a period of 6 hours prior to the time where the residual applies (this helps distinguish, for example, between residuals associated with the simulation of rising limbs of storm responses and residuals associated with falling limbs). In theory, a large amount of raw data of this type could have been saved, including data on rainfall, evaporation and the observed discharge, and data associated with the mathematical structure of the model such as the sensitivity of discharge to the parameters of the model.

The process of calculating the prediction interval involves selecting residuals, based on the level of match between the hydrological conditions for the simulation and the hydrological conditions recorded for the residuals. Once a set of residuals has been selected, the prediction bounds can be found. If at a given time in period B the simulated discharge is q and the number of residuals selected is N, such that the residuals are $r_1, r_2, \ldots, r_N$, then on a plot of the simulated hydrograph it is possible to plot points (ghost points) at $q+r_1$, $q+r_2, \ldots q+r_N$. These ghost points give a scatter around the simulated value, so define a prediction envelope which can be used to calculate the bounds. At any time, the upper and lower bounding values simply correspond to the maximum and minimum of the residuals selected for that time.

The ideal outcome is a prediction interval with high containment but narrow width, so the calibration of the ghost method for period B involves specifying some high level of containment (e.g. 99%) and minimising the width of the prediction interval that achieves that level of containment. 'Width' is the difference between the upper and lower bounding values and the measure minimised is the root-mean-square width (RMSW). The parameters to be optimised are: (1) the number N of residuals selected from the saved set; and (2) weighting parameter for the relative importance of the various data on hydrological conditions. The hydrological conditions are here defined by only two variables, so there is only one weighting parameter: α. The ghost method works by matching the simulated hydrological conditions to the saved raw data for the conditions in period A. For any given time in period B, the hydrological conditions are represented by the current values for the discharge and rate of change in discharge, $q_{\mathrm{sim,B}}$ and $q'_{\mathrm{sim,B}}$, respectively, and for any given residual in the saved set the hydrological conditions are represented by $q_{\mathrm{sim,A}}$ and $q'_{\mathrm{sim,A}}$. A measure of distance (closeness) D is used as a basis for matching:

$$D = \left(q_{\mathrm{sim,B}} - q_{\mathrm{sim,A}}\right)^2 + \alpha\left(q'_{\mathrm{sim,B}} - q'_{\mathrm{sim,A}}\right)^2 \quad (1)$$

At every time in period B where ghost points are to be generated, the distance D is calculated for every saved residual from period A and the saved residuals are ranked depending on their associated distances (residuals with the lowest values of D are at the top of the ranked set). If N residuals are selected, this involves taking the top N residuals from the ranked set. For any given value of α, the value of N is found by trial and error such that the target containment percentage is achieved exactly (within the resolution possible with discrete values). A standard algorithm for one-dimensional (1D) optimisation can therefore be used to optimise α to minimise RMSW, and this will automatically give the corresponding value for N.

The information carried forward for use when a prediction is to be made operationally, or in testing for period C, is: (1) the rainfall–runoff model; (2) the calibrated parameter set for the rainfall–runoff model; (3) the saved dataset of residuals and associated hydrological conditions; (4) the ghost algorithm for calculating prediction bounds, including

the definition for the distance measure D; and (5) the ghost parameter values N and α.

HODDER CATCHMENT

The area of the River Hodder catchment, northwest England, is 260 km^2 and its altitude ranges from 40 m at the flow gauge at Hodder Place to 544 m (Figure 1). The median annual flood at Hodder Place is 225 m^3 s^{-1} and the highest recorded flood since recording began in 1969 is 488 m^3 s^{-1} (Oct 1980; hi flows on http://www.environment-agency.gov.uk). Around 1,500 mm of rainfall falls annually in the uplands, where rich organic soils support grassland and moorland used for rough grazing and game rearing. Around 1,200 mm falls in the lower areas of the catchment, where mineral soils support substantial areas of improved grassland. Water is abstracted for domestic supply from Stocks Reservoir. This reservoir collects runoff from an area of 35 km^2. There are also water abstraction sites in the Dunsop and the Langden catchments. The daily operation of the abstraction system depends on various factors including water colour, the storage volume available and the prevailing weather conditions. At Slaidburn, near the outlet of Stocks Reservoir, the mean annual temperature is 8 °C and typically there are 18 snow days per year.

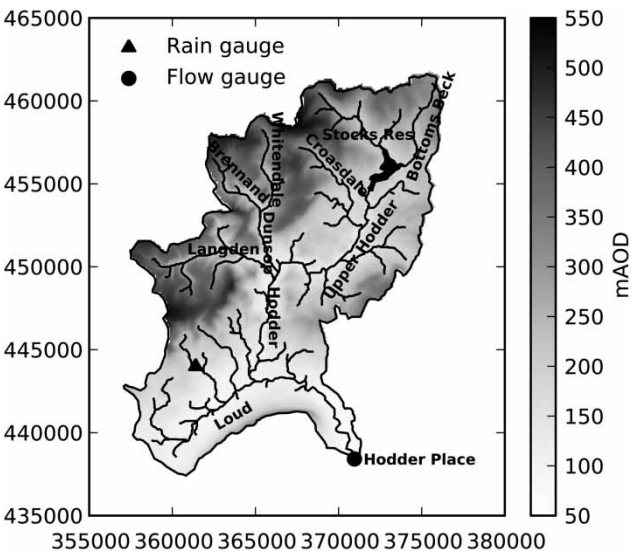

Figure 1 | Hodder catchment (using standard national UK grid References, metres).

The data used here are for the 19-year period 1991–2009 and comprise: (1) 15-minute rainfall measured at a tipping bucket raingauge located in the Loud catchment; (2) 15-minute discharge measured at Hodder Place by the UK Environment Agency; and (3) 15-minute data for potential evaporation, derived from data from an Automatic Weather Station (AWS) located in the Dunsop catchment, following the US Food & Agricultural Organization (FAO) method for calculating the potential evaporation for a hypothetical grass reference crop (Allen *et al.* 1998). As noted in the Introduction, this dataset has errors and limitations that could in part be eliminated:

1. The datasets have quality flags that show when values are suspect. A few periods of up to 6 weeks are marked as suspect.
2. The raingauge is located at a relatively low altitude (166 m) and the observations at a single gauge cannot be expected to always give good estimates for the rainfall over a catchment with a land area of 260 km^2. For the 19-year period, the mean annual rainfall at the gauge was 1,390 mm.
3. Evaporation data are available only for 2008–2010. The data for the 2009 calendar year were simply assumed to apply every year.
4. No explicit allowances were made for the water abstractions in the rainfall–runoff modelling.

RESULTS

For convenience, all the discharges are quoted as area-averaged values in millimetres per hour. A discharge of 1 mm h^{-1} corresponds to 72.22 m^3 s^{-1}, so the median annual flood (225 m^3 s^{-1}) is 3.12 mm h^{-1} and the highest recorded flood since recording began (488 m^3 s^{-1}) is 6.76 mm h^{-1}.

Each of the split-sample time periods covers four or more full calendar years: (A) 1997–2005; (B) 2006–2009; and (C) 1991–1996. The data for period C was not seen by the modeller until the work for periods A and B had been completed. The two models used in the work, PDM and TBM, were calibrated for time period A. This involved maximising the NSE using a shuffled complex algorithm (Duan *et al.* 1992). The resulting NSE values are 0.81 for PDM and

0.70 for TBM. The calibrated parameter values are given in the Appendix, where the models are described.

The models were calibrated for the full extent of period A but excluding: (1) times where the data suppliers had flagged suspect rainfall or flow data and for 120 hours following flagged suspect rainfall data; and (2) the time from 1st January to 1st October in 1997, to allow the effects of model initial conditions to decay. The purpose is to predict storm responses, rather than low flows, so raw data were saved only when the observed discharge exceeded 0.3 mm h^{-1}.

Figure 2 shows a summary of the residuals for period A, in the form of an exceedance plot. NSE and all the other measures used here are based on mean-square or root-mean-square values, so implicitly the relative contribution made by a residual is proportional to its square. For this reason, the scale for the x axis in the exceedance plot is the product of the residual and its absolute value. To allow the tails of the exceedance distribution to be seen, the scale for the y axis is a normal distribution with zero mean and unit variance. One of the reasons that the NSE is higher for PDM than for TBM is that PDM gives peakier storm responses resulting in a better match for the peaks in the observed hydrograph, hence the exceedance curve for PDM is lower than for TBM for large positive residuals. The relationship between residuals and discharge is not straightforward however, as a model that gives peaky simulations can also generate large residuals as a result of timing errors. This is the reason why the exceedance curve for PDM is lower than for TBM for large negative residuals.

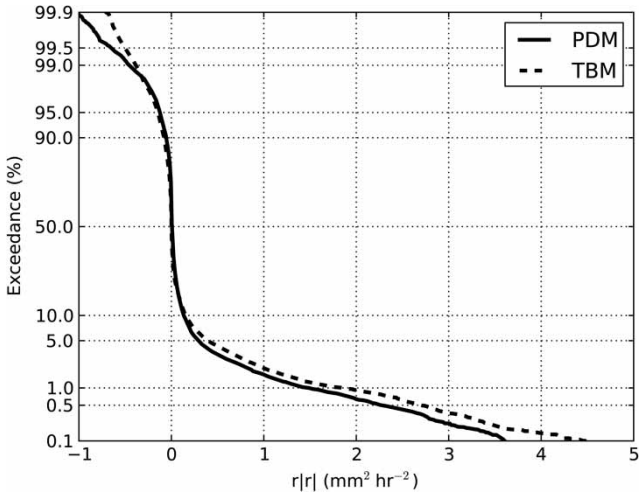

Figure 2 | Exceedance for the residuals for time period A.

When working with time periods B and C and creating prediction intervals, it is useful to have reference values against which the widths (i.e. RMSW values) of prediction intervals can be compared. The exceedance plot is the obvious source for reference values, but it would be too crude an approach simply to take the RMS of the residuals. Instead, the reference values (quoted against time period A in Table 1) were derived using a trivial method of calculating prediction bounds. In this trivial method (which is independent of the ghost method), the prediction interval is uniform such that, relative to the simulated discharge, the upper bound is always at U and the lower bound at L (or at a discharge of zero if the lower bound would be negative). Values for U and L were extracted from the exceedance data such that the required fraction of the population of residuals is contained (e.g. for a containment of 99%, 0.5% of the residuals lie above U and 0.5% lie below L).

The full set of results is given in Table 1 and the change in performance between calibration (period B) and testing (period C) is depicted in Figure 3. It can be seen that: (1) the widths of the prediction intervals were significantly lower than the reference values; (2) for the test period C the containment achieved (CP, the containment percentage)

Table 1 | Width, containment and calibrated parameter values for prediction intervals

Time period	Target CP	Measure/parameter	PDM	TBM
A	95	RMSW	1.203	1.349
	99	RMSW	1.991	2.154
B	95	RMSW	0.985	1.068
		CP	95.0	95.0
		N	57	45
		α	24.8	3.98
	99	RMSW	1.214	1.291
		CP	99.0	99.0
		N	243	191
		α	21.45	5.86
C	95	RMSW	1.043	1.097
		CP	91.9	93.3
	99	RMSW	1.294	1.338
		CP	97.2	97.9

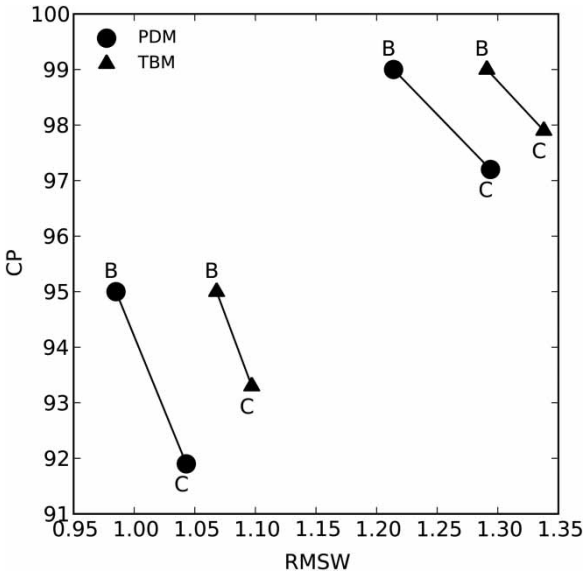

Figure 3 | Containment versus width for prediction intervals for periods B and C; the lines show the change in performance between calibration and testing.

fell only two or three percentage points below the target values; and (3) the widths of the prediction intervals do not change substantially between calibration and testing.

The prediction interval for two storms in the test period, period C, are shown in Figures 4 and 5. These figures are for target containments of 95 and 99%, respectively. The ghost method is under stress for these storms because the storms have large return periods (so are rare), with the result that there is a shortage of appropriate matching raw data for period A. By studying the behaviour of the ghost method under stress, its strengths and weaknesses can be gauged. In both figures, the storm shown in the left-hand plots has a peak discharge approximately equal to the median annual flood and the storm in the right-hand plots are for the highest flood recorded during the 19-year period covered by the dataset (at 5.55 mm h^{-1} it is the second highest since recording began in 1969). At high levels of containment it is rare for the observed hydrograph (thick line) to stray outside the prediction interval (grey envelope).

The individual ghost points are plotted as small dots so, depending on the resolution of the visual medium on which you are viewing this plot, you may be able to see the fine detail of the distribution of the ghost points. The time step is 15 minutes, so there are ghost points every 15 minutes. Twice per hour the median ghost point is shown as a dark dot, giving a broken-line effect. The solid thin line in the plots is the simulated hydrograph.

It can be seen that the median points vary more smoothly for PDM than TBM, partly because rainfall can give rise instantly to discharge when the deficit bucket in TBM is full, with the result that the simulated hydrograph for TBM is less smooth than for PDM. The difference in smoothness between the models also explains why the prediction bounds (i.e. the top and bottom of the grey envelope) are more ragged for TBM than for PDM. On comparing these bounds with the exceedance plot, it can be seen that in calculating the discharge for the ghost points some residuals must have been selected from the extremes of their distribution. Note that the simulated hydrograph sometimes lies above or below the prediction interval, as for example can be seen for a short period halfway down the recession in Figure 4(d). The reason for this is that, for some sets of hydrological conditions, there can be systematic under- or over-prediction in the calibrated simulation for period A.

PDM is the better model (as demonstrated by the results for NSE) but the results in Table 1 and Figure 3 for the prediction interval for period C, calculated under operational conditions, show that the percentage containment for TBM is slightly better than for PDM. The reason for this can be seen clearly in the left-hand plots in Figure 4. For the second peak in these plots the simulation for TBM is worse than for PDM but, in terms of containment, TBM performs better than PDM because for PDM the observed hydrograph strays outside the prediction interval just prior to reaching the peak and then again part-way down the following recession. For both models, there is clearly a difference in timing between the peak in the upper bound and the observed peak. However, the peak in the upper bound for PDM is narrower and it is for this reason that it fails to contain the observations.

As mentioned above, at high levels of containment it is rare for the observed hydrograph to stray outside the prediction interval. This means that for both Figures 4 and 5 the plots on the right-hand side, which are for the largest storm during the 19-year period, show rare behaviour. During period A, the only comparable peak is

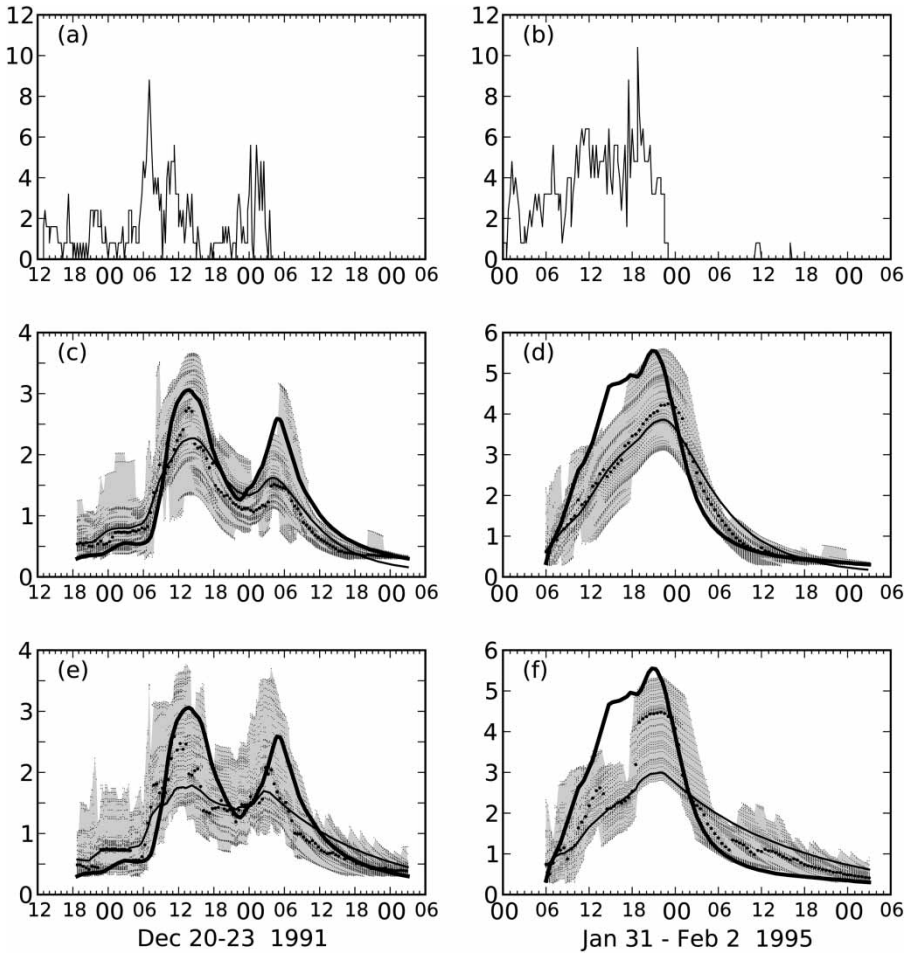

Figure 4 | (a, b) Rainfall and discharge for (c, d) PDM and (e, f) TBM for 95% containment for two storms during period C. Thick lines are observed discharge; thin lines simulated discharge; grey envelopes prediction intervals; dark dots median ghost points (plotted only twice per hour); light dots other ghost points; units are mm h^{-1}.

5.29 mm h^{-1} on 31 October 2000 (the next highest is 3.68 mm h^{-1} on 11 February 2002). It is therefore to be expected that the quality of the prediction interval might be poorer for this peak than for others because very few of the raw data from period A can be an adequate match. The poor quality of matching helps to explain why the patterns of ghost points are so regular and less fragmented in the right-hand plots.

Figure 6 shows some details of the outcome of matching. It corresponds to Figure 4(c), so is for PDM at 95% containment for the smaller of the two storms. The fine details cannot be seen, but the blocks of various shades of grey show that raw data have been selected from across time period A. As a result of the problem described above of finding an adequate match, a similar plot for the larger storm would show less variation (especially around the peak) as a result of repeatedly selecting from the same small subset of raw data. For storms with smaller return periods than the two storms considered here, plots similar to Figure 6 will show greater variation.

The ghost method takes its name from the fact that 'ghosts' of storm hydrographs from time period A can be seen when the method is applied. Ghosts are clearly visible in Figure 6, but can also be seen in a more subtle form where ghost points cluster together in bands in Figures 4 and 5. These bands should not be confused with the patterns and banding seen during the peaks. The ghost bands most clearly visible in the figures are those running down from the saw-tooth upper bounds for the tail of the recession in Figure 4(f).

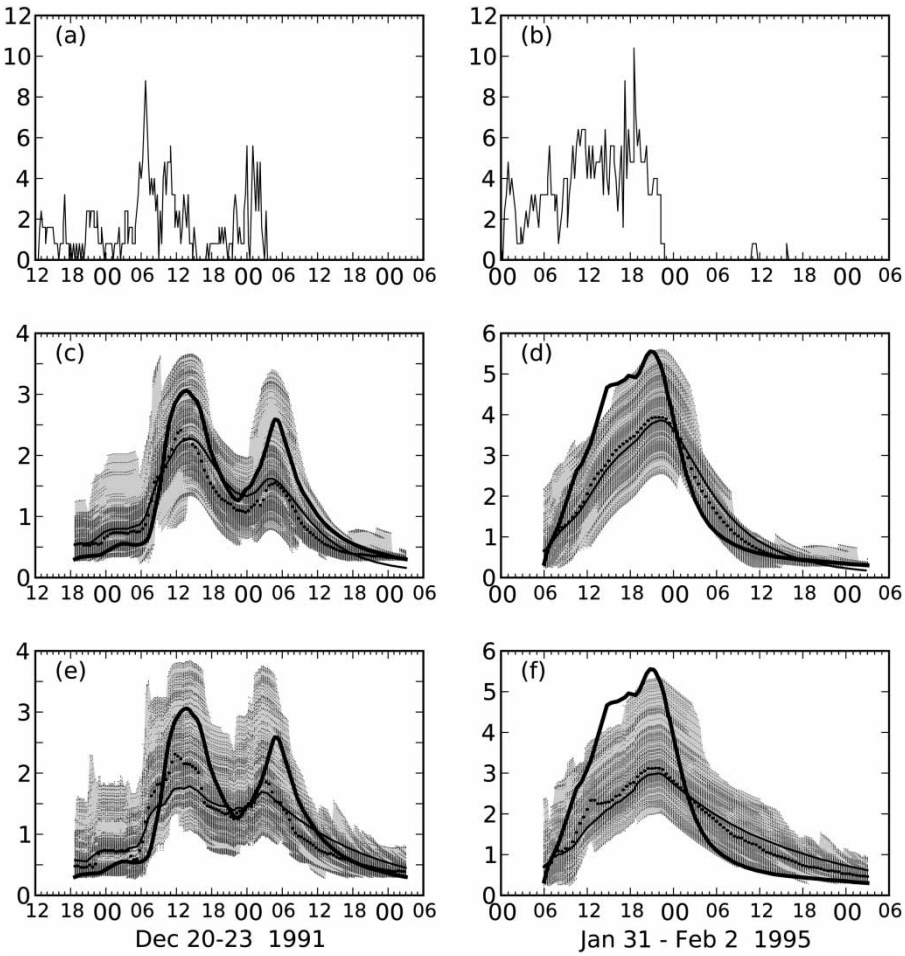

Figure 5 | (a, b) Rainfall and discharge for (c, d) PDM and (e, f) TBM for 99% containment for two storms during period C. Thick lines are observed discharge; thin lines simulated discharge; grey envelopes prediction intervals; dark dots median ghost points (plotted only twice per hour); light dots other ghost points; units are mm h^{-1}.

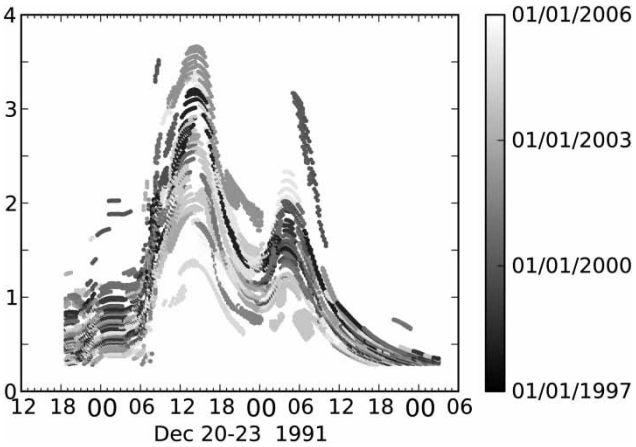

Figure 6 | Matching plot for Figure 4(c) showing the times in period A at which the raw data are selected for period C for PDM at 95% containment.

DISCUSSION AND CONCLUSIONS

A method (the ghost method) for use with calibrated rainfall–runoff models has been developed that calculates prediction intervals for the discharge at a river catchment outfall. The ghost method has some unusual features, including that: (1) in the calculation of prediction intervals it makes direct use of a dataset that contains raw errors (such as model residuals) and data on the hydrological conditions associated with those errors; and (2) the calculation method is calibrated to give a specified level of containment, such as 99% containment of observations.

To test the method, the two-period split-sample calibration-validation approach advocated by Klemeš was

extended to three periods, where the middle period (which is labelled period B) is used to calibrate the prediction interval. The ultimate goal is to have a practical method by which narrow prediction intervals can be calculated that have a specified level of containment, so the calibration for period B involves minimising the width of the prediction interval. To be most useful in hydrological analysis and operational work, the containment of the prediction interval must be very high and so 95 and 99% were used here. The prediction intervals are designed to contain observations or potential observations because the aim is to account for the overall effect of the errors arising from all sources (i.e. total error). Total error includes the large errors that can arise, for example, from the difficulties inherent in gauging very high flows and from limitations in the spatial distribution, spatial coverage and performance of rain gauges, as well as errors arising from the limitations of the model and modeller. We have not attempted to calculate prediction intervals for the actual discharge (i.e. the real discharge) because this would require making gross assumptions about the systematic and/or random nature of observation errors.

The method was applied using data for the Hodder Catchment, northwest England. The link between the quality of the prediction interval and the quality of the rainfall–runoff model was investigated. The models used are described in the Appendix. They are an eight-parameter version of PDM and the simple two-parameter model TBM.

The three time periods in the split-sample validation are labelled A, B and C. For period A, the rainfall–runoff model is calibrated. For period B, the ghost calculations are calibrated. For period C, prediction intervals made under operational conditions are checked to see if they are appropriately narrow yet contain the required percentage of observations. There is huge scope for working with a wide range of different types of raw data, but only three sets of raw data are used here: (1) residuals (observed discharge minus simulated discharge, for period A); (2) simulated discharge for period A; and (3) simulated rate of change in discharge for period A.

The measure used for the width of the prediction interval is the RMSW. The aim is to predict storm responses, so the raw data and the data used in the calculation of RMSW were extracted from the results for period A only for times when the observed discharge exceeded 0.3 mm h^{-1}. Reference values for RMSW were calculated using a trivial method for calculating prediction intervals based on the population of residuals for time period A. For 99% containment, this gave values around 2.0 mm h^{-1}. In both calibration and testing, the ghost method gave narrower intervals with widths around 1.3 mm h^{-1}. When the containment specified in time period B was 99% the values in testing for period C were over 97% for both models, which is quite acceptable given that some deterioration in performance can be expected between calibration and testing.

As expected, because of its pedigree as a well-established and widely used model, PDM was the better model at predicting the hydrograph for period A (NSE 0.81 compared to 0.70 for TBM). However, the performance of the prediction intervals for the two models was approximately the same.

For those unfamiliar with high-containment prediction intervals, the large widths of the intervals might come as a surprise. This is simply a consequence of the fact that there are large residuals in period A. From a hydrological perspective, the way to narrow the prediction intervals is to better understand the hydrology by finding out more about the discharge hydrograph, rainfall and evaporation, and to carry forward as much hydrologically relevant information as possible. (The similarity of the performance of the prediction intervals for PDM and TBM suggest that improving the mathematical structure of the rainfall–runoff model will not necessarily help.)

A study of the prediction intervals for period C showed that there are specific problems that deserve attention. These include the handling of timing errors for storm peaks and handling situations where the simulated hydrological conditions are quite different to the conditions for which the raw data apply. However, one of the main reasons for using raw data and the ghost method is that it gives direct control over the handling of specific problems. This control is exercised by specifying which raw data are carried forward and/or by redesigning the 'distance' measure D (a measure of closeness of match) so that the relevant hydrological interpretation is taken into account when comparing the simulated hydrological conditions against the raw data. For example, the dynamic programming algorithm in Ewen (2011) can calculate residuals and timing errors

simultaneously, and the data for timing errors could be carried forward to help with timing problems.

When faced with hydrological conditions not met during calibration, the distance measure D will be poor and will thus give an indication that the prediction interval is expected to be poor. In fact, using the distance measure D it might be possible to calculate confidence time series for the prediction interval (e.g. the confidence with which 99% containment is expected). Narrower prediction intervals might be obtained if the length of period A is increased so that it includes a fuller representation of the hydrological conditions experienced at the catchment (we have not tested this). However, the relationship between the interval width and the amount of available information is far from simple.

In the Introduction, it was stated that there is no consensus about how best to calculate prediction bounds; the literature shows signs of three distinct branches of approach: (1) informal hydrologically based approaches; (2) formal statistical approaches; and (3) approaches using artificial intelligence. The ghost approach, as it stands, lies somewhere close to branch (1) but part-way along towards branch (3). The aim is to have a method that works under operational conditions for messy real-world problems, rather than to test hypotheses about theories of uncertainty. A very pragmatic approach has therefore been taken here in which, if anything, rather more emphasis has been placed on how the prediction intervals were tested than on how they were calculated. In much of the literature, the emphasis is quite markedly the other way around.

As set out in the Introduction, the purpose was to carry out some very basic testing of the idea of using raw errors to calculate prediction intervals. A simple method (the ghost method) was developed based on that idea. The basic testing was quite successful, but detailed extensive testing for a wide variety of different catchments and types of storm response would be required before the ghost method could be proposed for use in operational work.

ACKNOWLEDGEMENTS

Work on the Hodder Catchment was supported under Environment Agency Project SC060092, Natural Environment Research Council programme 'Flood Risk from Extreme Events' (FREE; NE/F001134/1) and the Engineering and Physical Sciences Research Council programme 'Flood Risk Management Research Consortium' (FRMRC Phase 2).

REFERENCES

Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. 1998 *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56, FAO - Food and Agriculture Organization of the United Nations, Rome, Italy.

Beven, K. J. 2001 *Rainfall-Runoff Modelling: The Primer*. Wiley, Chichester, UK.

Beven, K. & Binley, A. 1992 The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* **6**, 279–298.

Beven, K. J., Smith, P. J. & Freer, J. E. 2008 So just why would a modeller choose to be incoherent? *J. Hydrol.* **354**, 15–32.

Blasone, R.-S., Madsen, H. & Rosbjerg, D. 2008 Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling. *J. Hydrol.* **353**, 18–32.

Diaz-Nieto, J. & Wilby, R. L. 2005 A comparison of statistical downscaling and climate change factor methods: impacts on low flows in the River Thames, United Kingdom. *Climatic Change* **69**, 245–268.

Duan, Q., Sorooshian, S. & Gupta, V. 1992 Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **28**, 1015–1031.

Ewen, J. 2010 Building a 'Virtual Hydrologist' into hydrological models to assess performance and confidence. In: *British Hydrological Society 3rd International Symposium: Role of Hydrology in Managing Consequences of a Changing Global Environment*, (C. Kirby, ed). British Hydrological Society, Newcastle, UK, pp. 507–511.

Ewen, J. 2011 Hydrograph matching method for measuring model performance. *J. Hydrol.* **408**, 178–187.

Ewen, J. & Parkin, G. 1996 Validation of catchment models for predicting land-use and climate change impacts. 1. Method. *J. Hydrol.* **175**, 583–594.

Ewen, J., Geris, J., O'Donnell, G., Mayes, W. & O'Connell, P. E. 2010 Multiscale experimentation, monitoring and analysis of long-term land use changes and flood risk - SC060092: Final Science Report, Newcastle University, Newcastle upon Tyne, UK.

Fowler, H. & Kilsby, C. 2007 Using regional climate model data to simulate historical and future river flows in northwest England. *Climatic Change* **80**, 337–367.

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J. & Butts, M. B. 2004 Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.* **298**, 222–241.

Gyau-Boakye, P. & Schultz, G. A. 1994 Filling gaps in runoff time series in West Africa. *Hydrolog. Sci. J.* **39**, 621–636.

Jin, X., Xu, C.-Y., Zhang, Q. & Singh, V. P. 2010 Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. *J. Hydrol.* **383**, 147–155.

Kavetski, D., Kuczera, G. & Franks, S. W. 2006a Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resour. Res.* **42**, W03407

Kavetski, D., Kuczera, G. & Franks, S. W. 2006b Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resour. Res.* **42** W03408

Klemeš, V. 1986 Operational testing of hydrological simulation models. *Hydrolog. Sci. J.* **31**, 13–24.

Krzysztofowicz, R. 2002 Probabilistic flood forecast: bounds and approximations. *J. Hydrol.* **268**, 41–55.

Liu, Y., Freer, J., Beven, K. & Matgen, P. 2009 Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error. *J. Hydrol.* **367**, 93–103.

Mantovan, P. & Todini, E. 2006 Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. *J. Hydrol.* **330**, 368–381.

Montanari, A., Shoemaker, C. A. & van de Giesen, N. 2009 Introduction to special section on uncertainty assessment in surface and subsurface hydrology: an overview of issues and challenges. *Water Resour. Res.* **45**, W00B00.

Moore, R. J. 1985 The probability-distributed principle and runoff production at point and basin scales. *Hydrolog. Sci. J.* **30**, 273–297.

Moore, R. J. 2007 The PDM rainfall-runoff model. *Hydrol. Earth Syst. Sci.* **11**, 483–499.

Nash, J. E. & Sutcliffe, J.V. 1970 River flow forecasting through conceptual models Part I – a discussion of principles. *J. Hydrol.* **10**, 282–290.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. & Franks, S. W. 2010 Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resour. Res.* **46**, W05521.

Reynard, N. S., Prudhomme, C. & Crooks, S. M. 2001 The flood characteristics of large UK rivers: potential effects of climate and land use. *Climate Change* **48**, 343–359.

Schoups, G. & Vrugt, J. A. 2010 A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resour. Res.* **46**, W10531.

Sivapalan, M., Blöschl, G., Merz, R. & Gutknecht, D. 2005 Linking flood frequency to long-term water balance: Incorporating effects of seasonality. *Water Resour. Res.* **41**, W06012.

Solomatine, D. P. & Shrestha, D. L. 2009 A novel method to estimate model uncertainty using machine learning techniques. *Water Resour. Res.* **45**, W00B11.

Thiemann, M., Trosset, M., Gupta, H. & Sorooshian, S. 2001 Bayesian recursive parameter estimation for hydrologic models. *Water Resour. Res.* **37**, 2521–2535.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W. & Srikanthan, S. 2009 Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. *Water Resour. Res.* **45**, W00B14.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M. & Robinson, B. A. 2008 Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.* **44**, W00B09.

Xiong, L. & O'Connor, K. M. 2008 An empirical method to improve the prediction limits of the GLUE methodology in rainfall-runoff modeling. *J. Hydrol.* **349**, 115–124.

Yang, J., Reichert, P., Abbaspour, K. C. & Yang, H. 2007 Hydrological modelling of the Chaohe Basin in China: statistical model formulation and Bayesian inference. *J. Hydrol.* **340**, 167–182.

First received 9 February 2011; accepted in revised form 20 December 2011. Available online 3 May 2012

APPENDIX: RAINFALL–RUNOFF MODELS

Two rainfall models are used in this work: the Probability Distributed Model (PDM) and a simple custom-designed model called the Two Bucket Model (TBM). These are described below and depicted in Figure A.1. The main 'building brick' for modelling used in both models is a reservoir that drains at rate S^m/k where k is a time constant, S the depth of water stored, and m is unity if the reservoir response is linear. Both models were run with a fixed time step of 15 minutes.

PDM

PDM is well established and is used widely in the UK (Moore 1985, 2007). The central concept in PDM is that runoff is generated by saturation excess runoff. It has a lumped equation for mass balance, derived assuming that each point in the catchment can store water and that the population of storage capacities within the catchment can be described by a known probability distribution. To adapt PDM to different catchments and applications, many different components and

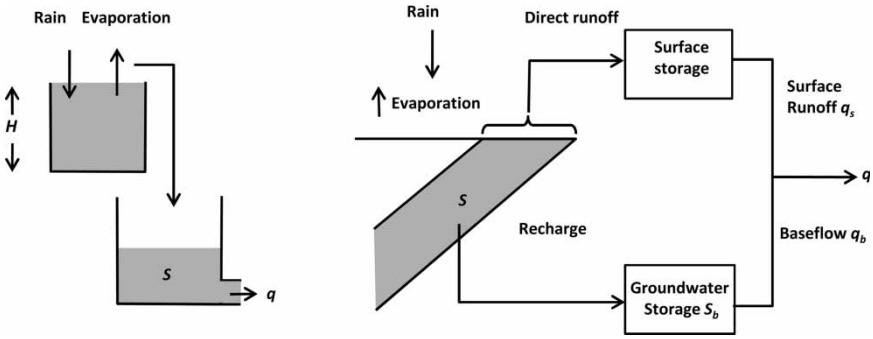

Figure A.1 | Schematic diagrams for TBM (left-hand side) and PDM (right-hand side).

formulations can be used. The components and formulations selected for modelling the Hodder catchment have been used widely in previous work on catchments in the UK. An outline of the formulation used is given below. In total, there are 8 parameters.

A constant multiplier F_c (calibrated value 0.9120) is applied to the observed rainfall. The fraction F of area that has a storage capacity less than c is given by a Pareto function:

$$F(c) = 1 - \left(1 - \frac{c}{c_{\max}}\right)^b \quad \text{(A.1)}$$

where $c_{\max}$ (40.233 mm) is the maximum storage capacity and b (0.2221) controls the variability of storage over the catchment. The area-average depth of storage S calculated by mass balance is used in conjunction with Equation (A.1) in the calculation of the areal extent which is saturated and thus capable of generating saturation excess runoff. Saturation excess runoff is routed via two linear reservoirs connected in series (time constants 6.9522 h and 2.3359 h). The storage S is used as an indicator for wetness in the calculation of evaporation; the ratio of actual to potential evaporation is assumed equal to the ratio of storage S to the maximum possible storage $S_{\max}$. The store S acts as a linear reservoir discharging to the groundwater store (time constant for drainage is 426.54 h). The groundwater store acts as a non-linear linear reservoir (time constant 20.215 $\text{mm}^{2.116}$ h and the dimensionless constant m is 3.116).

TBM

TBM is a simple model that captures the essence of the response of the Hodder Catchment. It is based on two ideas that have been applied within many lumped rainfall–runoff models (e.g. Beven 2001): (1) that the rate of runoff is strongly related to the volume of water stored in the catchment; and (2) the rates of evaporation and surface runoff are related to a moisture 'deficit' in near-surface storage. TBM uses these two ideas in their simplest form. There are two storage buckets. The upper bucket is for the calculation of deficit and produces overflow, and the lower bucket is a linear bucket that receives the overflow. The parameters are the capacity of the upper bucket, H (86.1 mm) and the time constant k (17.92 h) for the lower bucket.

Water evaporates freely from the upper bucket at the potential rate, limited only by supply. When the upper bucket is full (i.e. when the deficit is zero), any excess water overflows to the lower bucket. The mass balance equations for the buckets are solved exactly, using analytic solutions.