

Error analysis for the evaluation of model performance: rainfall–runoff event summary variables

Edzer J. Pebesma,^{1*} Paul Switzer² and Keith Loague³

¹ Department of Physical Geography, Utrecht University, PO Box 80.115, 3508 TC Utrecht, The Netherlands

² Department of Statistics and Department of Geological and Environmental Sciences, Stanford University, Stanford, CA, USA

³ Department of Geological and Environmental Sciences, Stanford University, Stanford, CA, USA

Abstract:

This paper provides a procedure for the evaluation of model performance for rainfall–runoff event summary variables, such as total discharge or peak runoff. The procedure is based on the analysis of model errors, defined as the differences between observed values and values predicted by a simulation model. Model errors can (i) indicate whether and where the model can be improved, (ii) be used to measure the performance of a model, and (iii) be used to compare model simulations. In this paper, both statistical and graphical methods are used to characterize model errors. We explore model recalibration by relating model errors to the model predictions, and to external, independent variables. The R-5 catchment data sets that we used in this study include summary variables for 72 rainfall–runoff events. The simulations used in this study were previously conducted with the quasi-physically based rainfall–runoff model QPBRRM for 11 different characterizations of the R-5 catchment, each with increasing information or a refined spatial discretization of the overland flow planes. This paper is about proposing model diagnostics and not about procedures for using diagnostics for model modification. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS model comparison; recalibration; linear regression; model errors

Received 12 April 2005; Accepted 9 May 2006

INTRODUCTION

Simulation models are often used in the Earth sciences to describe systems and to forecast the future behaviour of these systems. One example of such a system is the hydrologic response of a catchment after a rainfall event, in terms of groundwater recharge, surface runoff, and channel discharge (e.g. Güntner *et al.*, 1999). Simulation models have proven to be an indispensable tool for (i) improving our understanding of Earth science systems and (ii) forecasting a system's state where or when such a state, for practical reasons, cannot be measured. For example, the response of a system to change or a spatial–temporal prediction can often only be assessed with the aid of simulation. Whether a model's predictions are adequate to answer the questions being posed depends on how well the model performs. Currently, there is very little guidance as to how model performance should be assessed. There is a need for an organized way to conduct model performance evaluation, ultimately leading to a model evaluation protocol based upon performance standards. This paper attempts to set a foundation for such a protocol.

Model performance is usually assessed by numerical and graphical analysis of model errors, or residuals,

where

$$(\text{residual}) = (\text{observed value}) - (\text{predicted value})$$

Ideally, residuals should be small, and should contain no predictive information. Often, a model evaluation concentrates on a single summary statistic that measures the overall size of the residuals, such as the root-mean-square error (RMSE) used by Loague and Kyriakidis (1997) and the model efficiency used by Nash and Sutcliffe (1970), Loague and Freeze (1985), and Donnelly-Makowecki and Moore (1999). Plots of errors are often limited to scatter plots showing predicted versus observed values. Some workers (e.g. Loague and Green, 1991) have presented multiple summary statistics in conjunction with several different graphs, whereas others evaluated multiple measured response variables (Mroczkowski *et al.*, 1997; Kuczera and Mroczkowski, 1998) or evaluated multiple statistical measures describing the characteristics of residuals (e.g. Young, 2001).

Although measures of the size of residuals are certainly valuable, they usually do not reveal whether model residuals are random and unpredictable, or whether they behave in some sense in a predictable way. Predictable errors occur when underprediction is clearly distinct from overprediction. For example, (i) if peak channel flow is overpredicted following large rainfall events but underpredicted when following small rainfall events, or (ii) if the magnitude of the error relates to an external variable, such as a model input variable. The degree to which

* Correspondence to: Edzer J. Pebesma, Department of Physical Geography, Utrecht University, PO Box 80.115, 3508 TC Utrecht, The Netherlands. E-mail: e.pebesma@geo.uu.nl

errors are predictable may indicate how much room there is for improvement in the model. Predictable errors may give rise to new hypotheses related to the process being simulated and point us to areas where the model can be improved. In this paper, we use predictable errors to recalibrate model output and show how recalibration statistics may help to diagnose model deficiencies. In an operational setting, recalibration can be useful to end users of a model who do not have the possibility to modify or adjust the model or its parameters.

Pebesma *et al.* (2005) diagnosed errors for the catchment discharge time-series data of two large events. In contrast, this paper deals with event summary variables, such as peak discharge and total discharge, and analyses these variables for a large set of events. As in Pebesma *et al.* (2005), we present a three-part framework for model performance evaluation, based on an increasingly detailed examination of model errors. In the first part, the distributions of model errors are examined and properties of selected summary statistics are discussed. In the second part, we relate errors to model predictions to determine how much a simple adjustment improves the model output prediction. In the third part, the errors are related to external variables, such as rainfall characteristics or initial conditions, to explore whether part of the error variation can be further explained by available information. For selected model error summary statistics, bootstrap confidence intervals will be used to reflect their sampling variability.

To illustrate the model performance evaluation methods, we employed the R-5 catchment data set (observed and predicted values), because it comprises (i) a large number of rainfall–runoff events and (ii) a large amount of near-surface soil hydraulic property data that facilitated 11 increasingly complex and detailed characterizations of the catchment for simulation with the quasi physically-based rainfall-runoff model QPBRRM. These QPBRRM simulations were conducted and published over the past 20 years (Loague *et al.*, 2000).

Model recalibration in this paper is used only to identify error patterns that are predictable in a relatively simple way. To the extent that its errors are predictable, a model output could in principle be adjusted by simple recalibration. Thus, allowing for such recalibration serves to sharpen our model comparison statistics. If the goal were model building, i.e. model structure identification or model parameter calibration, rather than our goal of model intercomparison, then we refer, for example, to the GLUE framework of Beven and Binley (1992; Beven, 2001), to the Bayesian frame work of Thiemann *et al.* (2001), or to the data-based mechanistic modelling approach of Young (2001). In this paper we do not address real-time recalibration of models.

R-5 RAINFALL–RUNOFF SIMULATIONS

Catchment

The 0.1 km² R-5 catchment (see Figure 1), located within the Washita River Experimental Watershed, is near

Chickasha, Oklahoma, in rolling prairie grass-land that was subjected to continuous, well-managed grazing for decades (see Loague *et al.* (2000)). The R-5 catchment has hosted several field studies and simulation efforts designed to characterize the near-surface hydrologic response of the catchment. There has been a concerted effort to simulate observed R-5 rainfall–runoff events with Horton-type models of overland flow. An overview of the work at R-5 was recently provided by Loague *et al.* (2000). The R-5 data set is summarized in Figure 1; also see Sharma and Luxmoore (1979) and Loague and co-workers (Loague and Freeze, 1985; Loague and Gander, 1990; Loague, 1992c,d).

Events

A rainfall–runoff event can be characterized by a response in a channel's discharge caused by a rainstorm on the surrounding catchment. The events that we consider in this paper are the 72 events (and a subset of these) that were used by Loague and Freeze (1985). The 72 selected R-5 events each met the following four criteria: (i) they showed an obvious cause-and-effect rainfall–runoff relationship; (ii) there was no snowmelt component; (iii) the rainfall duration was less than 24 h; (iv) the storm flow depth (amount of run off for the event) was at least 0.0025 mm. For the R-5 catchment, no base flow separation was necessary because the channel is ephemeral.

At R-5, the rainfall data were measured with a tipping-bucket gauge in break-point form at the downstream end of the catchment (Figure 1). For rainfall, three event summary variables were defined by Loague and Freeze (1985): (1) rainfall depth P_D , which is the total amount of rainfall of an event; (2) peak rainfall rate P_{MX} , which is the maximum 2 min average rainfall rate; (3) the time of the maximum intensity t_{MX} , which is related to the start of the event (Figure 2, Table I). The start of an event is always the moment at which rainfall starts.

At R-5, the runoff data was measured at a V-notch weir at the outlet of the catchment (Figure 1). The discharge was measured continuously at the weir. The raw data were recorded in a breakpoint form. For the R-5 runoff events, three summary variables were defined by Loague and Freeze (1985): (1) the runoff depth Q_D , which is the total discharge during an event; (2) the peak discharge rate Q_{PK} , which is the maximum two-minute average discharge over the whole event; and (3) the time of peak discharge t_{PK} , which is related to the start of the event (Figure 2, Table I).

For the R-5 events, soil-water content values at the start of the event were estimated by linear interpolation (in time) between the two nearest soil-water measurements. The assignment of soil-water content and saturated hydraulic conductivity values to spatial locations across the R-5 catchment depends on the particular model simulation, and is discussed below. Other available event-specific variables used in this paper are the time between events Δt_{bet} and the event duration Δt_{dur} . The measured event-specific data are summarized in Table I.

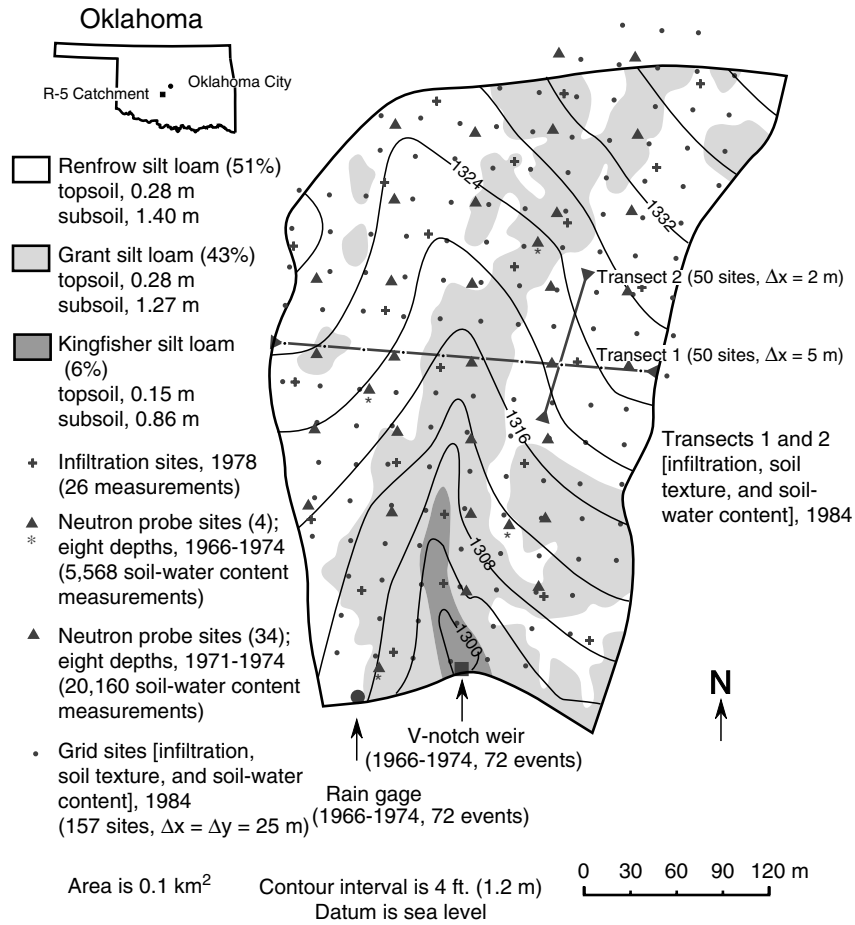


Figure 1. Map of the R-5 catchment showing topography, soil types, and measurement locations

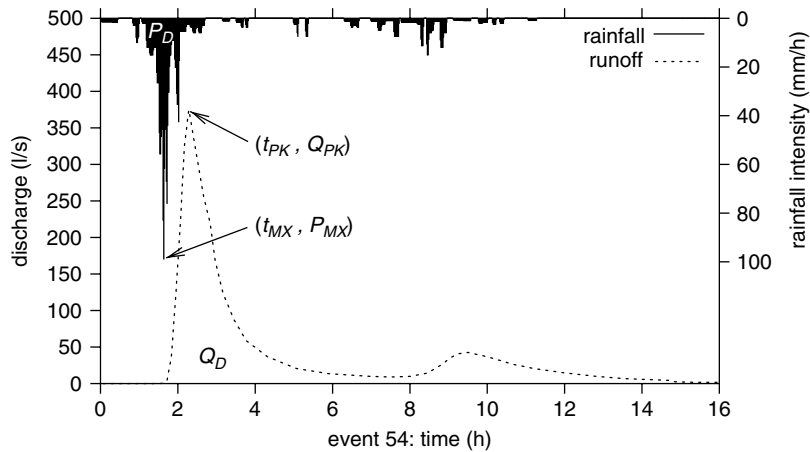


Figure 2. The six summary variables (see Table I) for event 54. P_D and Q_D refer to the area under the hyetograph and hydrograph respectively

QPBRM

The event-based rainfall–runoff model results used in this study are from 11 simulation scenarios with the quasi-physically based rainfall–runoff model QPBRM of Horton overland flow. QPBRM was originally presented by Engman and Rogowski (1974), and further described by Loague and Freeze (1985). The operating algorithms for QPBRM are based on solutions and/or simplifications to the full set of coupled partial differential equations which describe near-surface hydrologic

response. The model combines three major components: (i) an infiltration algorithm that allows the calculation of excess precipitation; (ii) a routing algorithm that translates partial area rainfall excess, generated on overland flow planes, into lateral inflow hydrographs at the stream channel; (iii) a routing algorithm for tracking the stream-flow hydrograph through the channel system.

The three major components of QPBRM are represented by one-dimensional equations that, when coupled together, result in quasi-three-dimensional representation

Table I. List of the available event-based summary data. The Q_D , Q_{PK} , t_{PK} , P_D , P_{MX} and t_{MX} estimates were obtained from 2 min aggregated break-point data (Loague and Freeze, 1985)

Variable	Symbol	Units
Rainfall depth	P_D	mm
Maximum rainfall intensity	P_{MX}	mm h ⁻¹
Time to P_{MX}	t_{MX}	h
Initial soil-water content	θ_{g^*}	—
Time between events	Δt_{bet}	days
Event duration	ΔL_{dur}	h
Runoff depth	Q_D	mm
Peak discharge	Q_{PK}	s ⁻¹
Time of peak discharge	t_{PK}	h

of rainfall–runoff via the Horton mechanism. The infiltration equation used is similar to Philip’s two-parameter equation (Philip, 1969). The equations for both overland and channel flow routing are each based upon numerical solution to the kinematic wave equations for shallow water flow. QPBRRM allows partial source areas to expand and contract during a storm, and allows for reinfiltration of runoff water. Two different flow-plane characterizations from the R-5 catchment are shown in Figure 3. The time step for the 11 simulation scenarios for R-5 events with QPBRRM was 2 min.

Simulation scenarios

The 11 QPBRRM simulation scenarios for the R-5 events that are compared in this study are summarized in Table II. The simulation scenarios listed in Table II reflect an increase in effort. They differ with respect to the quantity of data used to estimate the distribution of saturated hydraulic conductivity K_s (i.e. the number of measurements, 26 or 247), the amount of soil-water content data θ (i.e. 4 or 34 measurements), the method of

assessing initial soil-water content, and the characterization of overland flow planes (Figure 3). There was no parameter calibration involved for any of the QPBRRM simulation scenarios used in this study.

The differences between the results of the 11 simulation scenarios are mostly related to the amount of water that runs off Q_D and, to a much lesser extent, the time of peak runoff t_{PK} . Time of peak runoff t_{PK} is mostly determined by surface roughness coefficients for the overland flow planes and stream channel, which were held constant in all 11 simulation scenarios. For this reason, the variable t_{PK} will not be addressed in this paper. It should be pointed out that some of the simulation scenarios have fewer than 72 R-5 events. For example, some scenarios depended on the larger set of water content measurements that was only available for the last 39 events. For certain summary variables, the information that was available for all 11 simulations was used.

MODEL ERRORS

To examine how well individual events are predicted by a given simulation scenario, we will look at model residuals e_i :

$$e_i = o_i - p_i \quad i = 1, \dots, n \quad (1)$$

where n is the number of events considered, o_i is the reported value of a summary variable for event i (such as the runoff depth), and p_i is the corresponding model predicted value for that summary variable. Defined this way, negative errors indicate overprediction by the model, and positive errors indicate underprediction.

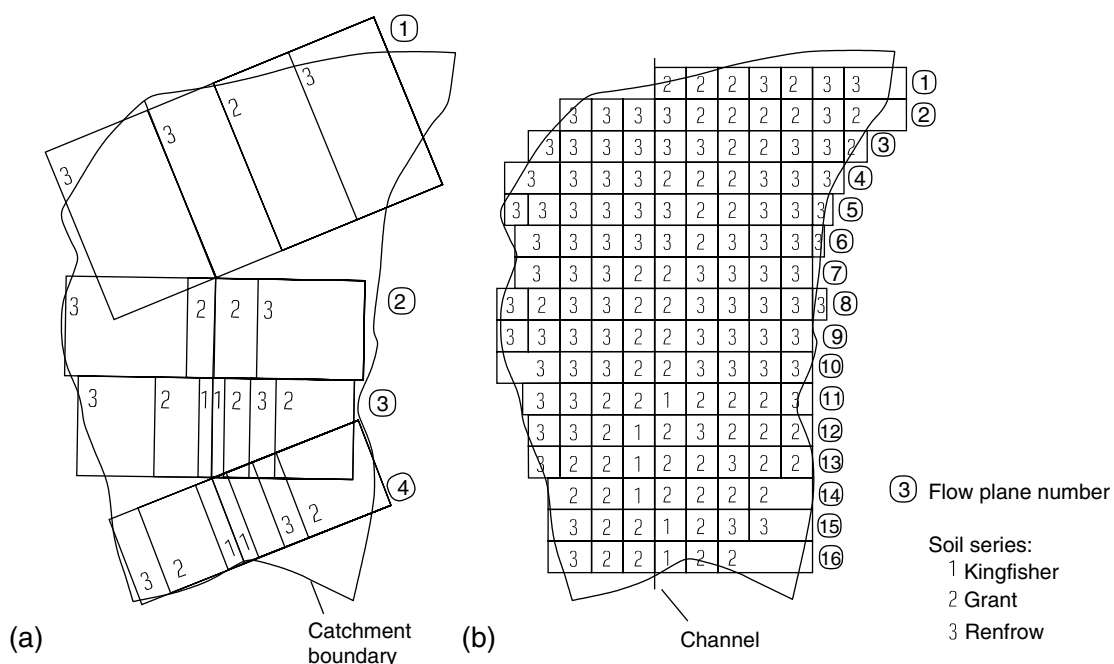


Figure 3. Different discretizations of the overland flow planes: (a) 22 planes (Loague and Freeze, 1985); (b) 145 planes (Loague, 1992a)

Table II. Summary of the 11 simulation scenarios with the QPBRRM. The locations of the K_s and θ measurements are shown in Figure 1; the overland flow planes are shown in Figure 3

Simulation scenario	No. of rainfall-runoff events	No. of overland flow planes	K_s estimate ^a	No. of K_s measurements	θ_{θ^*} estimate ^b for topsoil and subsoil	No. of θ measurement locations ^c	Reference
1 SiltLm	72	22	Silt loam value	—	Average	4/34	Loague (1992b)
2 Loam	72	22	Loam value	—	Average	4/34	Loague (1992b)
3 LF85	72	22	Global average	26	Average	4/34	Loague and Freeze (1985)
4 L90	72	22	Average per soil type	247	Average	4/34	Loague (1990)
5 Max	39	22	Average per soil type	247	Maximum	34	Loague (1992d)
6 Min	39	22	Average per soil type	247	Minimum	34	Loague (1992d)
7 Dist	39	22	Average per soil type	247	Distributed	34	Loague (1992d)
8 Grid-L90	68	145	Average per soil type	247	Average	4/34	Loague (1992a)
9 Grid-Max	37	145	Average per soil type	247	Maximum	34	Loague (1992a)
10 Grid-Min	37	145	Average per soil type	247	Minimum	34	Loague (1992a)
11 Grid-Dist	37	145	Average per soil type	247	Distributed	34	Loague (1992a)

^a K_s is the saturated hydraulic conductivity.

^b θ_{θ^*} is the initial soil-water content.

^c 4/34 refers to all measurements available (four locations for the first 33 events, 34 locations for the remaining 39 events).

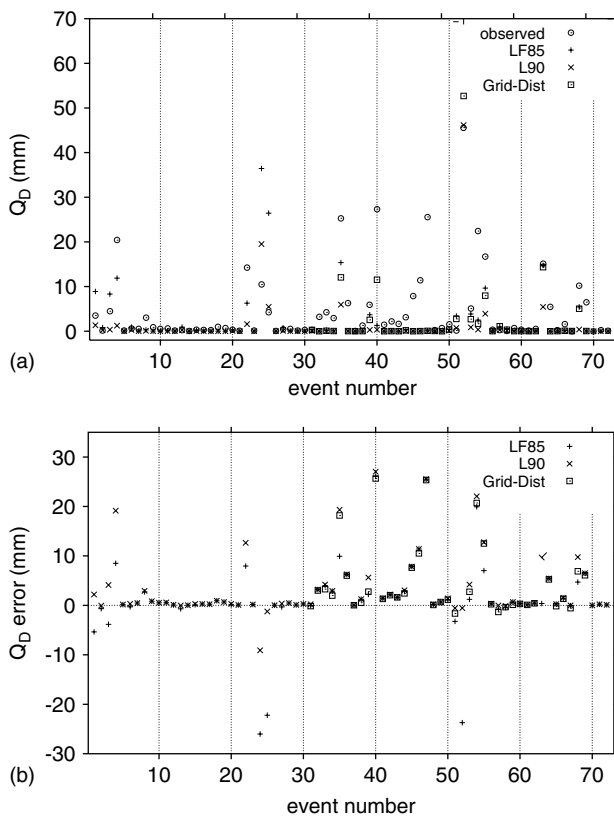


Figure 4. (a) Observed Q_D values for 72 events, with predicted values for three simulations with QPBRRM; (b) errors in the Q_D values for the simulation scenarios shown in (a). Note that for Grid-Dist, only 37 events are available

Graphical methods

Figure 4a shows the observed values, for all 72 events, of runoff depth Q_D and the corresponding predicted values for three different simulation scenarios. Figure 4b shows the magnitude of the errors, for the same events, in the same three simulations shown in Figure 4a. In Figure 4, the event numbers for extreme (large positive

or negative) residuals are shown. The complete set of graphs with results for all three summary variables and all 11 simulations are available as separate electronic documents (<http://www.geog.uu.nl/~pebesma/R5/> or contact corresponding author).

Figure 5 shows the cumulative distributions of the errors for the three simulations shown in Figure 4. In Figure 5, for example, it is shown that the fraction of Q_D errors larger than 10 mm is approximately 6% for LF85 and 16% for the Grid-Dist simulation. The error distribution in Figure 5 confirms what could be gleaned from Figure 4: many errors have a value close to zero, and large underpredictions occur much more often than large overpredictions. Figure 5 has the advantage that it corrects for the different number of events available for the three simulations, at the cost of their connection; it does not, for example, show whether for different simulations the large errors relate to the same set of events, as is shown in Figure 4.

The box-and-whisker plots shown in Figure 6 summarize the residual distributions and simplify comparison

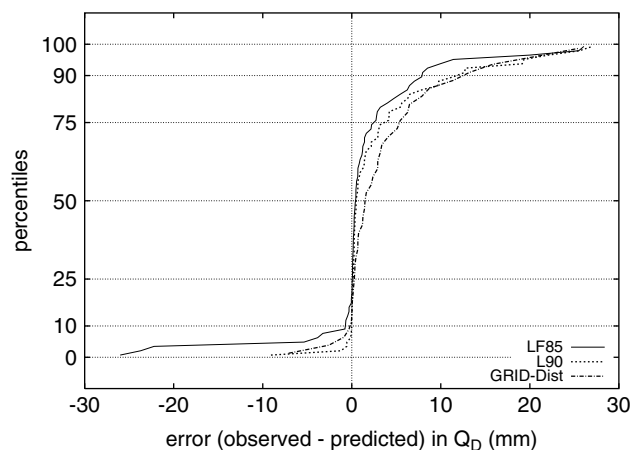


Figure 5. Cumulative density plots for the errors from three different R-5 simulations with the QPBRRM

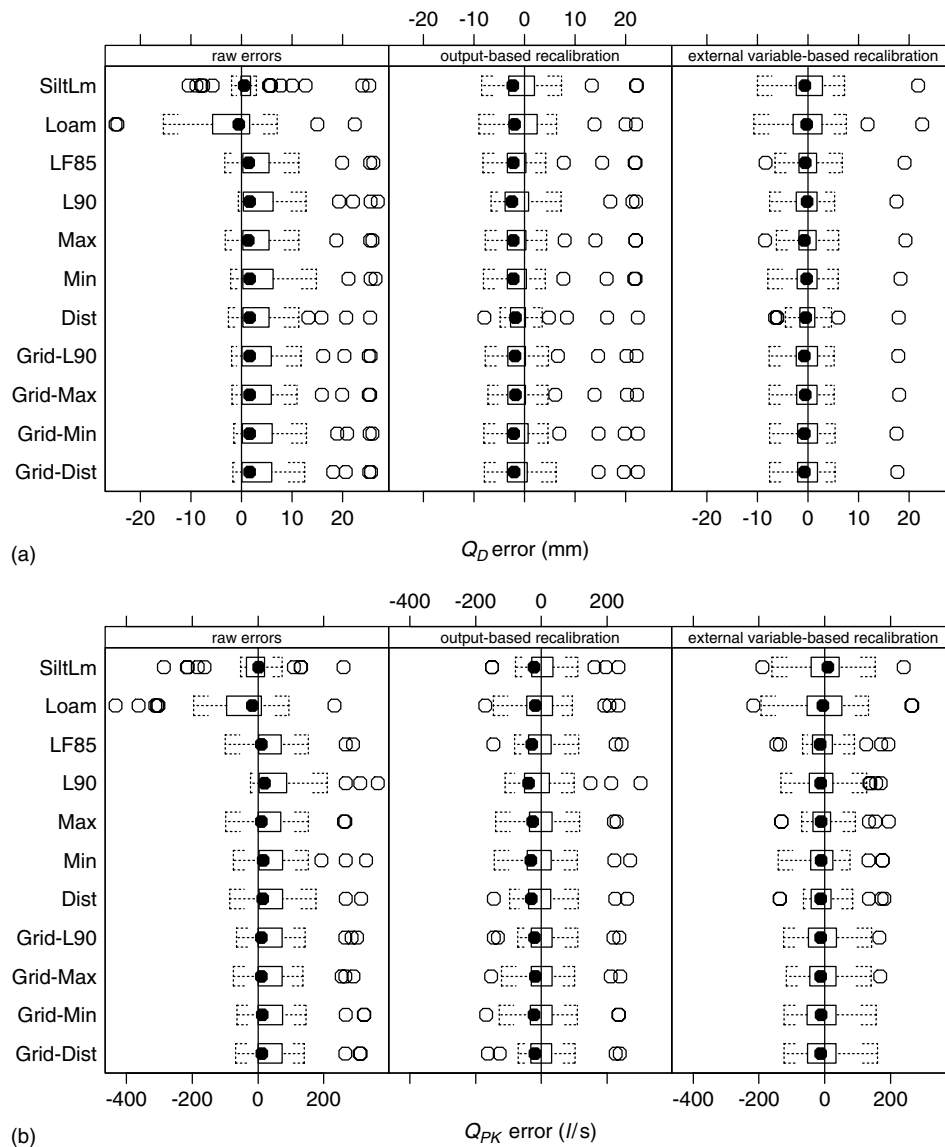


Figure 6. Box-and-whisker plot for (a) Q_D errors and (b) Q_{PK} errors of the 11 QPBRRM simulations of R-5 events, before and after recalibration. Only data from the common subset (including outliers) are shown. The filled bullet denotes the median, solid boxes range from the lower to the upper quartile, whiskers show the data range. Data that are further than 1.5 times the interquartile range from the nearest quartile are shown as open bullets

between different simulations. The left-hand panels of Figure 6 show box plots for the errors from the 11 simulations, using the common subset of simulated events. The box-and-whisker plots show the median, the upper and lower quartile (box), the data ‘range’ (whiskers), and outliers relative to the interquartile range. Figure 6 shows for the raw errors that many simulations have similar distributions, with many positive and few large positive errors, except for SiltLm and Loam, which have some negative and a few large negative errors.

Ideally, the model error is unrelated to other variables, such as the predicted values, or any known, external variable (Anscombe and Tukey, 1963; Garrick *et al.*, 1978), such as listed in Table I. To check whether such relations exist, scatter plots can be used. Figure 7 shows a scatter plot of model prediction errors for simulation scenario LF85 plotted against the model predicted values. Figure 7 shows that the larger predictions are mainly

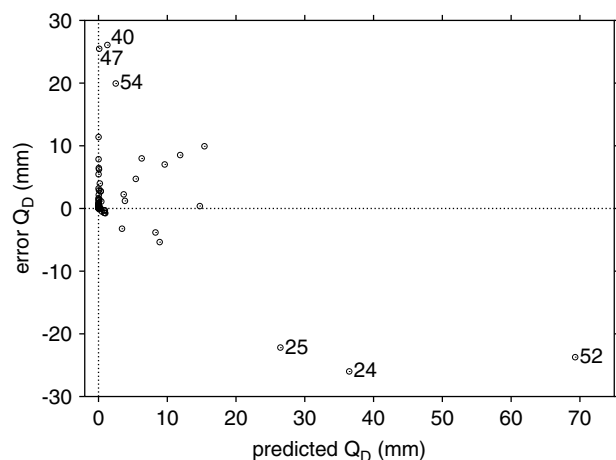


Figure 7. Scatter plot of Q_D errors against predicted values from simulation LF85. The six labelled points are the outliers

overpredictions. If we relate model errors to *pairs* of external variables, then we see the *combined* effect of two external variables on the error. Figure 8 shows for all Q_D simulations the error magnitude (symbol) plotted simultaneously as a function of maximum rainfall intensity P_{MX} (x-axis) and event duration Δt_{dur} (y-axis). Figure 8 shows that the large overpredictions or underpredictions occur as clusters, indicating that specific combinations of these two external variables are associated with large model errors. In contrast, Figure 9 demonstrates that the negative Q_{PK} errors, plotted as a function of θ_{θ^*} and Δt_{bet} , show much less clustering, whereas the overpredictions (positive errors) are clustered: they occur typically at wet initial conditions and at short times

between events. Figures 8 and 9 have a high potential for revealing specific circumstances under which large over- or under-predictions typically occur. The full set of graphs with Q_D and Q_{PK} errors as functions of pairs of external variables is available as a separate electronic document (<http://www.geog.uu.nl/pebesma/R5/> or contact corresponding author).

Numerical methods

The mean error \bar{e} tells how much *on average* the model over- or under-predicts and is called the bias. A common measure of variability for the n errors is the standard deviation $\sigma(e)$. A measure that combines the bias and

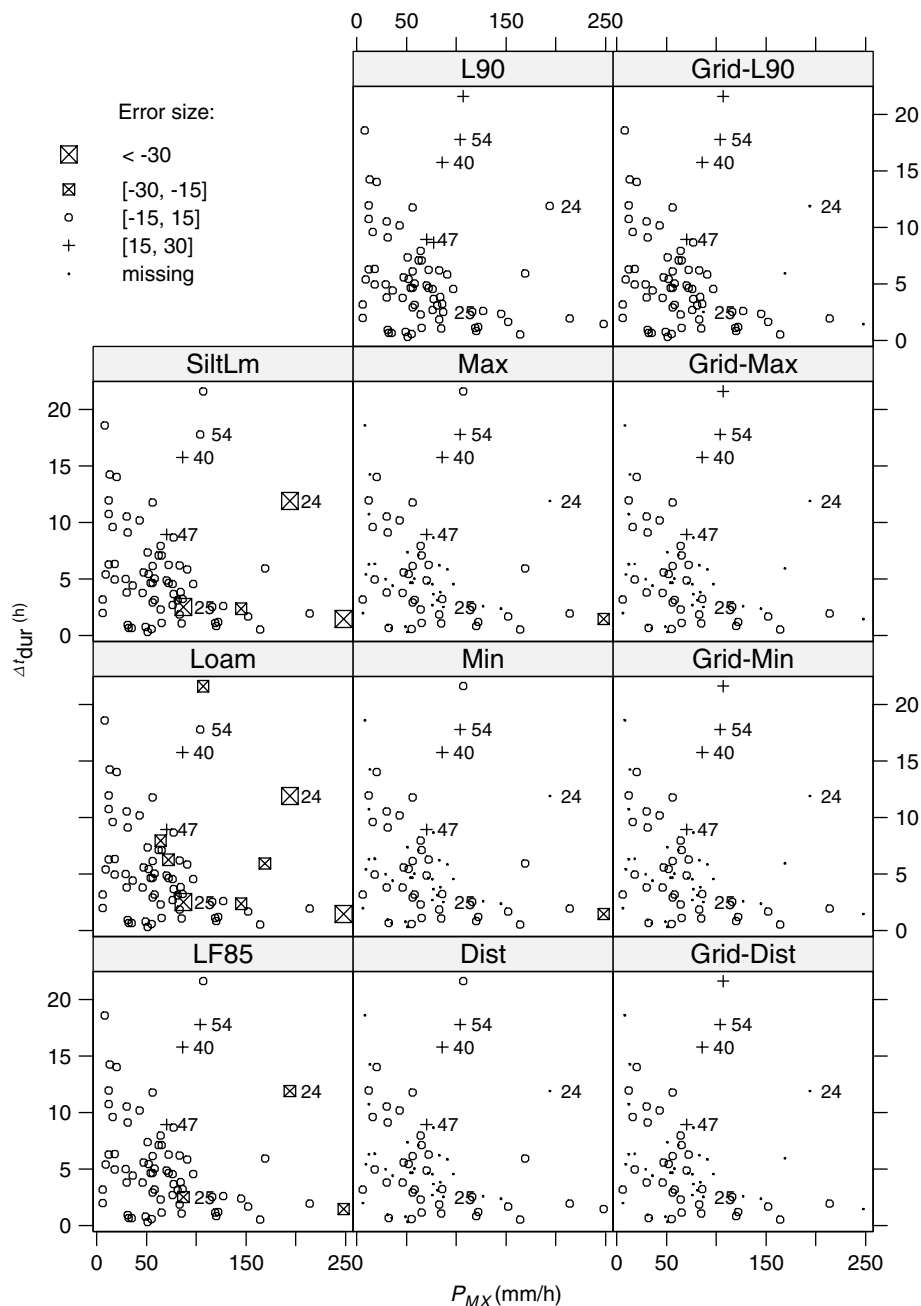


Figure 8. Errors for Q_D for the 11 QPBRRM simulations of R-5 events (see Table II), as a function of maximum rainfall intensity P_{MX} and event duration Δt_{dur}

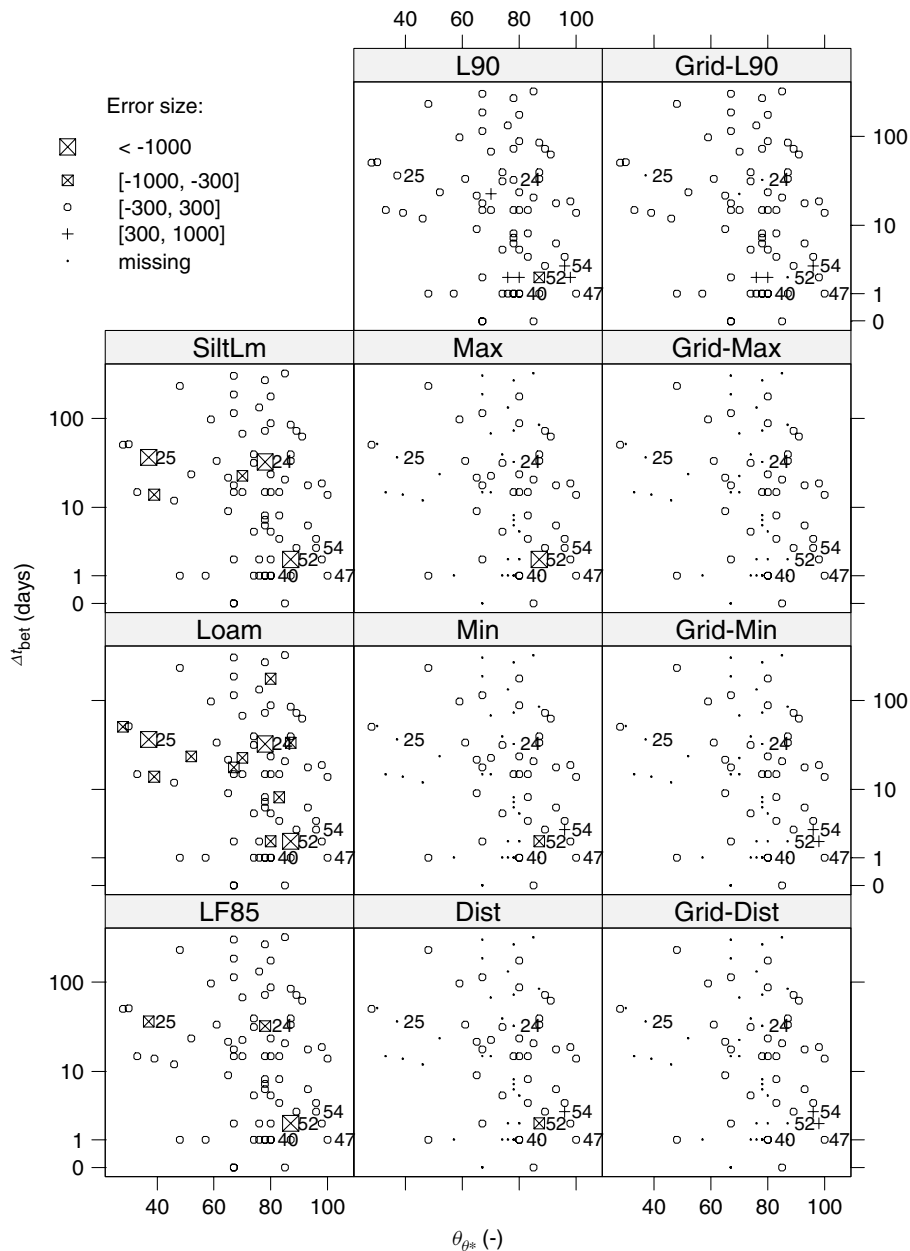


Figure 9. Errors for Q_{PK} for the 11 QPBRRM simulations of R-5 events (see Table II), as a function of initial soil-water content θ_{0^*} and time between events Δt_{bet}

variability is the RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} = \sqrt{\bar{e}^2 + \sigma^2(e)} \quad (2)$$

To compare the RMSE between different variables and different simulation scenarios, for non-negative variables it can be normalized by the mean observed value \bar{o} , and the relative RMSE, expressed as a percentage, is

$$RMSE_r(\%) = \frac{RMSE \times 100}{\bar{o}} \quad (3)$$

Another measure that normalizes the RMSE is the model efficiency E_f , proposed by Nash and Sutcliffe (1970) and used by Loague and Freeze (1985) and many others.

It is defined as

$$E_f = \frac{\sum_{i=1}^n (o_i - \bar{o})^2 - \sum_{i=1}^n e_i^2}{\sum_{i=1}^n (o_i - \bar{o})^2} \quad (4)$$

It should be pointed out that E_f does not have a lower limit of zero when the model predictions do not involve a fitted (i.e. calibrated) mean value (Anderson-Sprecher, 1994); it does, however, have an upper limit of one.

Neither RMSE or E_f separate the bias (mean error \bar{e}) and variance. Figure 10 shows five artificial examples, four of which (a, b, c and e) have a model efficiency very close to zero (Table III), so that none of them would be preferred judging by E_f alone. Still, we are

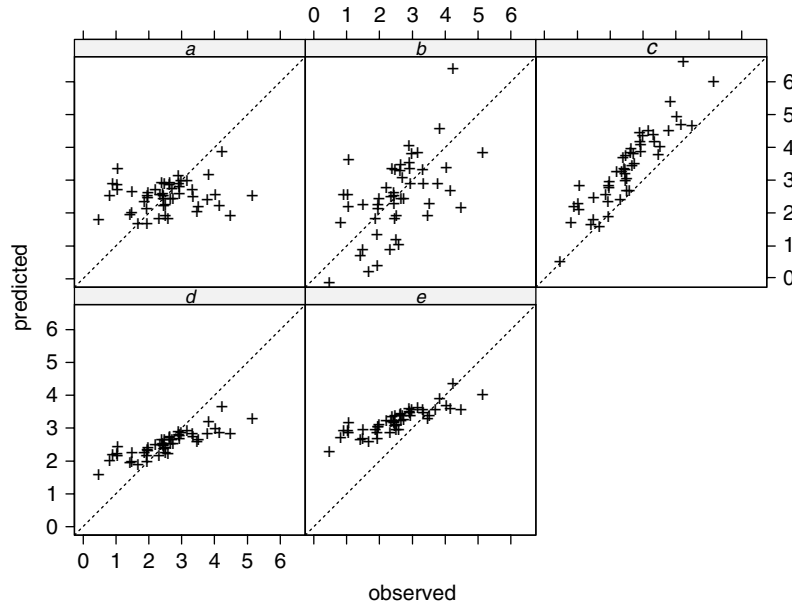


Figure 10. Scatter plots for five hypothetical cases of observed (x-axis) versus predicted (y-axis). The diagonal line indicates where observed equals predicted

Table III. Summary statistics for the errors of the five hypothetical cases of Figure 10

Case	Bias	σ^2	E_f	r
a	0	1.1	-0.05	0.15
b	0	1.1	-0.08	0.55
c	-0.87	0.24	0	0.91
d	0	0.48	0.52	0.86
e	-0.71	0.48	0.02	0.86

Bias is the mean error, σ_2 is the variance of the error, E_f is the model efficiency, defined in Equation (4); r is the correlation between predicted and observed.

inclined to prefer c and e over a and b, and even prefer b over a. This preference is based on a clear linear relation between observed and predicted, associated with a large linear correlation (Addiscot *et al.*, 1995). In the case of a low E_f but a high correlation coefficient, a simple linear adjustment of the predicted value could improve the model prediction dramatically, as will be shown below.

The different R-5 simulation scenarios did not always involve the same catalogue of rainfall events. Apparent performance differences among simulation scenarios might be attributed to differences between the simulation performance or to differences between the event subsets. Therefore, summary statistics are also presented for the subset of events that is common to all simulations.

Outliers

In contrast to percentiles or fractions above or below a threshold, the mean, variance, RMSE and related summary statistics are highly sensitive to outliers. Outliers are evident in Figures 4 and 5. In this study, an event was labelled as an outlier when both Q_D and Q_{PK} predictions for the event qualitatively showed atypically large prediction errors, compared with the bulk of the errors. The six outlier events in this study are events 24, 25, 40, 47, 52 and 54. They are specifically labelled in the scatter plots of Figures 7–9. For summary statistics, we will consider both the full set of events and the set of events without outliers. The largest prediction errors are

Table IV. Listing of Q_D and Q_{PK} errors for the six outlier events. Empty cells indicate missing data

	Q_D						Q_{PK}					
	24	25	40	47	52	54	24	25	40	47	52	54
Event simulation	24	25	40	47	52	54	24	25	40	47	52	54
SiltLm	-45.3	-39.5	23.9	25.3	-38.3	12.7	-1191.9	-1243.8	72.8	260.5	-1533.6	109.4
Loam	-55.3	-51.5	15	22.5	-40.9	7.1	-1328.1	-1557.3	43.3	232.7	-1904.2	-12.5
LF85	-26	-22.2	26.1	25.5	-23.7	19.9	-716.1	-696	90	266.3	-1317.7	289.1
L90	-9.1	-1.2	27	25.5	-0.6	22.1	-269.3	-65.3	120.4	267.2	-453.7	364.8
Max			25.9	25.5	-25.1	18.8			89.1	266.2	-1300.3	261.2
Min			26.5	25.5	-19.5	21.2			104.3	267.2	-863.9	328.4
Dist			15.8	25.5	-7.1	20.8			97.9	266.8	-905.4	313.8
Grid-L90			25.5	25.2		20.4			107.1	265.1		300.8
Grid-Max			25.5	25.2		19.9			106.6	265.3		290.7
Grid-Min			25.9	25.4		20.9			113.1	266.5		321.6
Grid-Dist			25.7	25.3		20.7			107.4	266.2		312.1

always related to large events: either a large stormflow is heavily underpredicted or a big rainstorm causes a significant overprediction of runoff.

Pebesma *et al.* (2005) investigated discharge error time-series for two of the larger events (54 and 68) in detail, one of them having large prediction residuals. For the six outlier events, the Q_D and Q_{PK} errors are shown in Table IV.

MODEL BIAS INVESTIGATIONS USING RECALIBRATION

We investigate model bias using optimized linear recalibration of model outputs, and optimized linear recalibration using external variable adjustments. For this, we used the event subsets without outliers to prevent that a very small number of extreme residuals completely dominate the recalibration statistics. Alternative approaches are given in the ‘Discussion’ section.

Model output-based recalibration

A simple way to recalibrate model outputs, such as those shown in Figure 10b–e, is to adjust model predictions by using a linear function of the predicted value. The errors can be related to the predicted values through

$$e_i = a_0 + a_1 p_i + \varepsilon_i \tag{5}$$

with ε_i a noise term representing deviation of errors from a straight line, and where a_0 and a_1 are unknown parameters that can be estimated by \hat{a}_0 and \hat{a}_1 from the set of events used for recalibration using ordinary least squares. The ‘fitted’ errors

$$\hat{e}_{i,a} = \hat{a}_0 + \hat{a}_1 p_i \tag{6}$$

are added to the original model predictions to obtain ‘output-based’ recalibrated predictions. We will call this procedure output-based recalibration.

The errors that remain *after* this type of recalibration are $e_i - \hat{e}_{i,a}$, and we can calculate the corresponding RMSE by using these modified errors in Equation (3). We now divide by $n - 2$ instead of n to obtain an unbiased estimate (because two parameters, a_0 and a_1 , were fit to the e_i).

Table V shows the estimated regression coefficients and their standard errors. In Table V, the predicted values p_i were centred by subtracting their mean value before the regression line of Equation (5) was calculated. This does not affect the slope a_1 , but it makes the interpretation of a_0 easier: a_0 now equals the bias \bar{e} . A positive slope a_1 indicates that larger predicted values are associated with the larger positive (or least negative) residuals. It should be pointed out that no centring was used to recalibrate model predictions.

The standard errors of these calibration regression coefficients indicate how well the regression coefficients are determined from the available recalibration event

Table V. Regression coefficient estimates (\hat{a}_0 and \hat{a}_1), standard errors (SE), and fraction of the error variation explained by the regression line (R^2) for output-based recalibration of 11 simulation scenarios with the QPBRRM for Q_D and Q_{PK} . Only data from the common subset without outliers were used ($n = 34$). Parameter a_0 is the mean error (bias)

Summary variable	Simulation	\hat{a}_0	SE	\hat{a}_1	SE	R^2
Q_D	SiltLm	0.1	0.57	-0.40	0.08	0.43
	Loam	-3.8	0.62	-0.62	0.06	0.78
	LF85	2.4	0.51	0.41	0.16	0.17
	L90	3.4	0.53	2.76	0.45	0.55
	Max	2.3	0.52	0.33	0.16	0.12
	Min	2.9	0.51	1.06	0.23	0.39
	Dist	2.7	0.52	0.76	0.20	0.31
	Grid-L90	2.6	0.52	1.47	0.29	0.45
	Grid-Max	2.5	0.51	1.36	0.26	0.45
	Grid-Min	2.9	0.54	2.32	0.40	0.51
	Grid-Dist	2.8	0.53	2.01	0.36	0.49
Q_{PK}	SiltLm	-23	11	-0.45	0.07	0.56
	Loam	-75	12	-0.59	0.06	0.73
	LF85	27	9	0.03	0.11	0.00
	L90	49	10	0.89	0.23	0.32
	Max	24	9	0.00	0.11	0.00
	Min	36	9	0.29	0.13	0.13
	Dist	32	9	0.19	0.12	0.07
	Grid-L90	33	9	0.73	0.19	0.32
	Grid-Max	30	9	0.63	0.17	0.31
	Grid-Min	39	10	1.06	0.23	0.40
	Grid-Dist	36	10	0.91	0.21	0.37

data. In general, only estimated parameters that are at least twice as large as their standard error are significantly different from zero. When the slope parameter differs significantly from zero, a steeper slope indicates a larger change in fitted errors when the model predicted value changes one unit. The R^2 represents the fraction of error variation explained by the regression line.

Table V shows that for both Q_D and Q_{PK} most simulation scenarios have a significant bias. For the SiltLm and Loam scenarios, the bias is negative (or approximately zero for Q_D at SiltLm), indicating overprediction on average in the series of recalibration events. The bias is positive for the remaining simulation scenarios, indicating underprediction on average. Most of the regression slopes are significantly different from zero. The negative slopes of the SiltLm and Loam scenarios indicate that the larger predictions are overpredicted. The positive slopes of the other nine simulations indicate greater underpredictions in the case of larger predictions, which suggests that better models would yield larger predicted values for the larger events. This could, for instance, be accomplished by modifying the infiltration component of the model, or by adding other runoff-generating mechanisms.

For the R^2 values given in Table V it can be concluded that output-based recalibration is more successful for Q_D than it is for Q_{PK} . The $RMSE_r$ before and after output-based recalibration is found by comparing columns r and a in Table VI for the three different subsets. For the common event subset without outliers, all simulations show improved $RMSE_r$. In addition, the $RMSE_r$ values

Table VI. RMSE_r of mean observed value \bar{o} of Q_D and Q_{PK} for the 11 simulation scenarios with the QPBRRM. RMSE_r is given for model errors r , errors remaining after output-based recalibration a , and errors remaining after external variable-based recalibration b . Evaluations are for all data, for the non-outlier data, or for the common subset without outliers, and n is the number of events used. Note that for the whole table, \bar{o} for Q_D (3.76) and Q_{PK} (63.0) were obtained from the common subset without outliers

Summary variable	Simulation	All data				Non-outliers				Common non-outliers			
		n	RMSE _r (%)			n	RMSE _r (%)			n	RMSE _r (%)		
			r	a	b		r	a	b		r	a	b
Q_D	SiltLm	72	275	165	217	66	115	83	82	34	114	89	100
	Loam	72	347	163	211	66	208	85	103	34	221	96	130
	LF85	72	203	166	183	66	89	78	62	34	107	80	67
	L90	72	190	170	144	66	135	87	81	34	148	82	77
	Max	39	226	183	175	35	102	85	69	34	104	81	66
	Min	39	231	190	168	35	123	92	71	34	125	79	69
	Dist	39	193	167	133	35	116	89	70	34	118	80	68
	Grid-L90	68	171	139	109	65	110	73	74	34	127	81	75
	Grid-Max	37	215	172	125	34	123	79	72	34	123	79	72
	Grid-Min	37	227	175	125	34	139	83	77	34	139	83	77
	Grid-Dist	37	223	174	125	34	134	82	76	34	134	82	76
Q_{PK}	SiltLm	72	470	185	328	66	186	110	114	34	148	98	118
	Loam	72	591	188	370	66	284	116	151	34	237	110	163
	LF85	72	331	168	291	66	96	93	80	34	92	84	81
	L90	72	217	186	208	66	179	136	137	34	135	94	98
	Max	39	355	138	312	35	92	86	85	34	91	85	83
	Min	39	265	144	244	35	104	88	84	34	105	84	84
	Dist	39	270	140	246	35	97	87	85	34	98	84	84
	Grid-L90	68	149	116	119	65	128	95	108	34	115	88	95
	Grid-Max	37	149	120	112	34	108	83	90	34	108	83	90
	Grid-Min	37	166	127	119	34	128	89	99	34	128	89	99
	Grid-Dist	37	161	125	118	34	122	88	97	34	122	88	97

for the 11 simulation scenarios are much more alike after output-based recalibration than before.

External variable-based recalibration

External variable-based recalibration was done in away similar to model output-based recalibration. However, here, a linear model was used to relate prediction residuals to the six ‘external’ variables given in Table I (P_D , P_{MX} , t_{MX} , θ_{θ^*} , Δt_{bet} , and Δt_{dur}) as follows:

$$e_i = b_0 + \sum_{j=1}^6 b_j X_{ij} + \varepsilon'_i \tag{7}$$

where ε'_i is a noise term and where X_{ij} is the value of external variable j for event i . Parameters were estimated by ordinary least squares over a set of recalibration events. The ‘fitted’ errors

$$\hat{e}_{i,b} = \hat{b}_0 + \sum_{j=1}^6 \hat{b}_j X_{ij} \tag{8}$$

are added to the original model predictions to obtain ‘external variable-based’ recalibrated predictions. The errors that remain after this type of recalibration are $e_i - \hat{e}_{i,b}$, and we can calculate the corresponding RMSE by using these modified errors in Equation (3). We now divide by $n - 7$ instead of n to obtain an unbiased estimate (because seven parameters were fit to the e_i).

Before calculating the linear regression coefficients of Table VII, the errors e_i , as well as all the external variables, were standardized by subtracting their respective means and then dividing by their respective standard deviations. This removes the bias and makes the regression coefficients dimensionless, thus enabling the comparison of regression coefficients between variables with different units. It should be pointed out that no standardization was used to recalibrate model predictions.

Table VII shows the regression coefficients for the external variable-based recalibration. For a given external variable, a positive regression coefficient indicates that larger values of that variable are associated with the larger residuals. Based on the R^2 values in Table VII, we observe that external variable-based recalibration is again more successful for Q_D than for Q_{PK} . For Q_D , the strongest recalibration effect seems to come from the event duration Δt_{dur} : the positive regression slopes for all simulation scenarios except the SiltLm and Loam scenarios indicate that longer duration events are associated with greater underprediction by QPBRRM. Similarly, we see typically greater underprediction at larger rainfall depth (P_D for scenarios L90, Min) and at larger maximum rainfall intensity (P_{MX} for scenarios L90 and all Grid-scenarios), indicating that runoff depth for larger events is underpredicted, which may indicate that either infiltration is misrepresented or that other runoff-generating mechanisms play a role in generating stream flow. The

Table VII. Regression coefficients with standard errors (SE) and R^2 for external variable-based recalibration. Only the common subset without outliers was used ($n = 34$). Prior to regression, model errors and all covariates were standardized to have zero mean and unit standard deviation, to make regression coefficients dimensionless. Table I and Figure 2 explain the meanings of the symbols for the regressor variables

Simulation	P_D	SE	P_{MX}	SE	t_{MX}	SE	θ_{θ^*}	SE	Δt_{dur}	SE	Δt_{bet}	SE	R^2
Q_D													
SiltLm	-0.29	0.19	-0.20	0.18	0.31	0.18	0.12	0.17	-0.06	0.23	-0.16	0.16	0.39
Loam	-0.65	0.14	-0.13	0.13	0.21	0.14	0.04	0.13	-0.05	0.17	-0.18	0.12	0.65
LF85	0.25	0.17	0.13	0.16	0.27	0.17	0.06	0.15	0.38	0.20	-0.16	0.14	0.51
L90	0.31	0.14	0.37	0.13	0.13	0.14	0.07	0.13	0.51	0.17	-0.15	0.12	0.66
Max	0.22	0.17	0.12	0.16	0.26	0.17	0.07	0.15	0.40	0.20	-0.17	0.14	0.50
Min	0.32	0.15	0.27	0.14	0.19	0.15	0.08	0.14	0.46	0.18	-0.15	0.13	0.61
Dist	0.28	0.16	0.23	0.15	0.21	0.15	0.08	0.14	0.46	0.19	-0.15	0.13	0.58
Grid-L90	0.18	0.15	0.37	0.14	0.15	0.15	0.07	0.14	0.56	0.18	-0.19	0.13	0.60
Grid-Max	0.17	0.15	0.36	0.14	0.14	0.15	0.06	0.14	0.58	0.18	-0.20	0.13	0.61
Grid-Min	0.21	0.14	0.39	0.14	0.10	0.14	0.08	0.13	0.57	0.17	-0.19	0.12	0.64
Grid-Dist	0.19	0.15	0.39	0.14	0.11	0.14	0.07	0.13	0.58	0.17	-0.19	0.12	0.63
Q_{PK}													
SiltLm	-0.22	0.18	-0.42	0.17	0.25	0.17	0.18	0.16	-0.34	0.21	-0.07	0.15	0.46
Loam	-0.23	0.17	-0.46	0.16	0.16	0.17	0.16	0.15	-0.35	0.20	-0.13	0.14	0.50
LF85	0.30	0.21	-0.11	0.20	0.44	0.21	0.04	0.19	-0.29	0.26	-0.15	0.18	0.21
L90	0.37	0.19	0.38	0.18	0.37	0.19	0.13	0.17	-0.11	0.23	-0.14	0.16	0.38
Max	0.26	0.21	-0.13	0.20	0.43	0.21	0.11	0.19	-0.28	0.26	-0.13	0.18	0.21
Min	0.39	0.21	0.13	0.20	0.43	0.20	0.12	0.19	-0.23	0.25	-0.15	0.17	0.27
Dist	0.34	0.21	0.05	0.20	0.41	0.21	0.12	0.19	-0.23	0.25	-0.15	0.18	0.22
Grid-L90	0.21	0.20	0.34	0.19	0.31	0.20	0.16	0.18	0.02	0.24	-0.20	0.17	0.32
Grid-Max	0.20	0.20	0.31	0.19	0.29	0.20	0.17	0.18	0.05	0.24	-0.21	0.17	0.31
Grid-Min	0.24	0.19	0.41	0.18	0.28	0.19	0.19	0.17	0.04	0.23	-0.19	0.16	0.37
Grid-Dist	0.21	0.20	0.39	0.18	0.29	0.19	0.19	0.18	0.04	0.23	-0.19	0.16	0.35

latter hypothesis is in agreement with the hypothesis of Vanderkwaak and Loague (2001) that, in addition to Horton overland flow, Dunne overland flow may play a role in generating runoff at R-5. For Q_{PK} , only P_{MX} and time to maximum rainfall t_{MX} relate to the errors. For the SiltLm and Loam scenarios, events with larger P_{MX} values are associated with larger overpredictions, suggesting that these model simulation scenarios yield too much surface runoff by infiltration excess.

RMSE_r values for the external variable-based recalibrated prediction errors are given in the *b*-columns of Table VI. With the exception of the SiltLm and Loam scenarios for Q_D , external variable-based recalibration results in smaller RMSE_r than model output-based recalibration. For Q_{PK} , external variable-based recalibration RMSE_r values are of similar size to the RMSE_r after model output-based recalibration. The variability in RMSE_r after external variable-based recalibration is much smaller than before recalibration (except for scenarios SiltLm and Loam). Figure 6 allows the comparison of distributions of errors from the common event subset including the outliers for raw errors and errors remaining after both forms of recalibration. For all simulations of the R-5 events, the recalibrated error distributions for QPBRRM become more nearly centred around zero following model output-based recalibrations and more nearly symmetric after external variable-based recalibration.

VARIABILITY OF PERFORMANCE MEASURES

When comparing two simulations (e.g. LF85 and Min in the last column of Table VI with RMSE_r values of 67 and 69 respectively), we can conclude that, for this particular subset of 34 events, LF85 performs slightly better than Min. The question then arises as to whether this difference between 67 and 69 signifies a difference in performance larger than would be seen from event sampling variability.

Under the assumption that the subset of 34 events is a representative sample of events, we can quantify this sampling variability by bootstrapping (e.g. Willmott *et al.*, 1985; Efron and Tibshirani, 1993) using repeated resampling of the 34 events. Figure 11 shows the 95% bootstrap confidence intervals for the 66 RMSE_r values in the last three columns of Table VI. For smaller model RMSE_r values, confidence intervals for different simulation scenarios mostly overlap, indicating that the corresponding RMSE_r differences may be due to chance.

Table VI suggests that, when looking at the all-data column, for external variable-based recalibration, RMSE decreases with increasing level of information used through the 11 scenarios. This effect, however, is gone when looking at the other two subsets (without outliers, common subset without outliers), further stressing the need to remove outliers and to use the same set of events when comparing models.

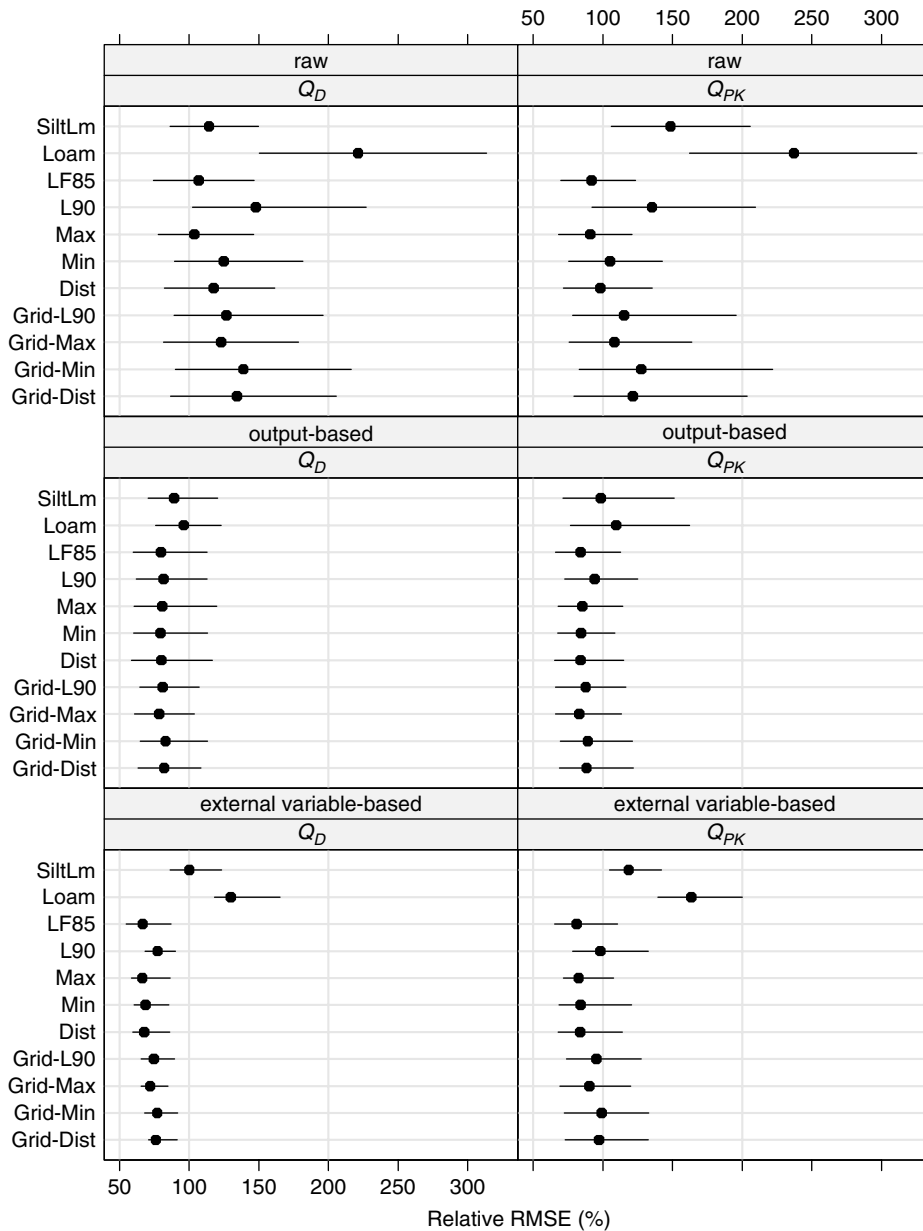


Figure 11. Estimates (bullets) and bias corrected and adjusted 95% bootstrap confidence intervals (solid lines) for relative RMSE for the 11 QPBRRM simulations of R-5 events. Top row: raw RMSE_r; middle row: RMSE_r after model output-based recalibration; bottom row: RMSE_r after external variable-based recalibration for Q_D (left) and Q_{PK} (right)

DISCUSSION

General

In this study, we do not consider whether we can formally reject the hypothesis that the model is ‘true’ (or ‘valid’) because formal statistical hypothesis testing (Mroczkowski *et al.*, 1997; Donnelly-Makowecki and Moore, 1999) can be misleading. The reasons for this are (i) any model is necessarily an approximation of reality at some level of detail and eventually every model will be rejected and (ii) not rejecting the model may give a false feeling of being on safe ground, because it may very well be a consequence of having inadequate data (i.e. insufficient or unsuitable data) for powerful testing (Mroczkowski *et al.*, 1997). Instead of hypothesis testing, we emphasize the estimation of model performance

measures, such as RMSE. To reflect the effect of event sampling variability we presented bootstrap confidence intervals for model RMSE given in Figure 11.

In the more operational sense, however, we can state that the QPBRRM has been rejected: simulations with a better alternative (Loague *et al.*, 2005) have shown that the QPBRRM has severe deficiencies (e.g. it lacks mechanisms for Dunne overland) that a better alternative model (InHM) does not have.

Previous discussions on model validation by Konikow and Bredehoeft (1992, and the rest of that issue) and Oreskes *et al.* (1994) are focused on whether a validated model implies that the model is ‘true’. In this respect, we agree with Addiscot *et al.* (1995), in that a model is regarded as validated when it suffices for its intended purpose.

There is a rich history of model performance testing for rainfall–runoff models, starting with the WMO model comparison study (WMO, 1975), followed, for example, by James and Burges (1982); recent perspectives are collected by Anderson and Bates (2001). However, there are no established standards for the performance evaluation of rainfall–runoff models (Loague and Vanderkwaak, 2004). Although quantitative standards for model performance are usually not specified as part of the modelling effort, we believe that it is important to report performance measures and diagnostics such as those described here. Validation and comparison of competing models will benefit from standard performance measures and diagnostics. Reporting performance measures in terms of confidence intervals helps to show the impact of limited data on the calculation of these performance measures.

In this paper, our primary interest in recalibration is not to use it for fixing models, but merely to use it as a diagnostic tool that exhibits the magnitude and direction of model errors and indicates whether these model errors are related to model inputs. Using recalibration statistics to adjust model simulations in a real-time setting would be interesting from an operational point of view, but we do not address real-time model recalibration here. Krzysztofowicz (1999) and Krzysztofowicz and Herr (2001) present a Bayesian framework that incorporates recalibration in a dynamic, real-time forecast system; their system allows for dependencies of model errors on model inputs.

Obviously, the measures of model performance and the recalibration statistics depend on the set of validation events that are used. Although the events selected here were considered large enough to be used for model comparison by Loague and Freeze (1985), many of them were very low rainfall events where observed and predicted values are all close to zero. For such events, models are not easily distinguished from one another, so one might consider taking these out of the model comparison. In addition, performance and recalibration measures are always subject to re-evaluation as new events enter the information stream.

In an alternative approach, errors could be defined as fractions (or relative errors), e.g. $e'_i = p_i/o_i$. Although this is attractive from the point of view that it would stabilize the variance of the errors (errors defined by Equation (1) tend to vary more when the rainfall event is larger), it would also inflate the importance of very low rainfall events. This alternative approach might reduce the need to remove outliers. As another alternative one could use robust or resistant regression methods for finding recalibration statistics based on sets of events that include outliers.

Model output-based recalibration as used in this study and discussed by Flavelle (1992) is equivalent to regressing the observed values on the predicted values, and Equation (5) is equivalent to

$$o_i = a_0 + (1 + a_1)p_i + \varepsilon_i \quad (9)$$

Krzysztofowicz and Watada (1986) used linear regression to model the conditional mean and variances of forecast errors for seasonal runoff forecasts, where they used observed values as the predictor variable. Their goal was the modelling of likelihood functions for the forecast errors at different stages of forecasting. For our goal of simple adjustment (recalibration) of model predictions, the form used here, Equation (5), is more relevant, because it has the potential to adjust model predictions, by the amount given in Equation (6).

How general is the procedure?

Recalibration statistics may be calculated for any model application where residuals can be calculated, and where predicted values and model inputs are available. The simple approach in this paper uses linear relations, assuming independence between events. Using more complex, possibly non-linear functions for recalibration may give useful information about more complex structures present in the errors. Examples of more complex functions include higher order linear regressions, interaction among external variables, additive models (Hastie and Tibshirani, 1990) or even artificial neural networks (Hastie *et al.*, 2001). This option is only viable when the number of observations is substantially larger than for the example used in this paper, and may also make the interpretation of recalibration coefficients more difficult.

For evaluating individual models that involve calibrated model parameters, the error analysis we propose should preferably be done on a set of observations that were not involved in the calibration. A useful scheme for classifying calibration–validation situations was given by Klemš (1986). For intercomparison of a set of models that involve a similar degree of parameter calibration, this point is less compelling.

Although the analysis in the case study of this paper involves a series of events, temporal correlation was not addressed because it was not evident in the series of prediction errors (e.g. Figure 4) for the event summary variables Q_D , Q_{PK} , and t_{PK} that were used. However, it may be possible to exploit autocorrelation in the full time-series of model errors within each event. For example, see Young (2001). Temporal correlation within events becomes evident when events are not summarized but are analysed as separate time-series of catchment discharge. Pebesma *et al.* (2005) provide an example where recalibration statistics are analysed using the full time-series structure of selected events.

Further useful extensions would include multivariate analysis of event summary variables, spatio-temporal analysis of model residuals, and analysis of stochastic model output where predictions are in the form of probability distributions.

Ranking the model simulation scenarios

Based on the RMSE values reported in Table VI, we can rank the model simulation scenarios with respect to their performance, both before and after recalibration.

Clearly, RMSE levels, as well as their ranks, depend to some extent on the chosen event series. For the R-5 series, the presence of outlier events and the differing availability of model input data have an influence on the ranking of model simulation scenarios. For these reasons we include comparisons for a common subset of events. After recalibration, different model simulation scenarios often have very similar RMSE values. When accounting for event sampling errors, the RMSE interval estimates in Figure 11 suggest that the set of events used here is too small to distinguish among some model simulation scenarios. This result is in accordance with what Beven and Freer (2001) call 'equifinality' (see also Beven (2001)).

Reproducing event-aggregated quantities

Our model evaluations are based on residuals of matched predicted and observed values. One might also compare the two distributions of observed and predicted values aggregated across the event series. This comparison has little value because it allows for cancellation of overpredictions and underpredictions. Thus, even a poorly performing model can provide a good distribution match. If the aim is to compare distributions, then a better approach is to compare distributions of predicted and observed values within groups having similar conditions. For example, it may be relevant whether *on average* the model performs well for large rainfall events, or for dry soil conditions. Still, whenever distributions for groups are compared, cancellations may occur.

SUMMARY AND CONCLUSIONS

This paper proposes a procedure for model performance evaluation. The procedure comprises (i) a detailed examination of model errors (residuals) using graphs and summary statistics, (ii) a recalibration of model predictions where errors are predicted by a simple linear adjustment based on model predictions or a set of external variables, and (iii) a comparison of recalibration errors with the original errors. This paper is about proposing model diagnostics and not about procedures for using diagnostics for model modification.

For the R-5 series of rainfall–runoff events previously modelled in 11 different simulation scenarios with QPBRRM, we found that model errors do not occur at random, but rather under specific circumstances. Figure 7 shows how errors depend on the model predicted value, and plots such as Figures 8 and 9 reveal the specific circumstances under which large over- or under-predictions occur. We show that model predictions can be easily improved by a linear adjustment based on a history of model predictions, and that there is room for improvement of the model using the available model predicted values and external variables.

It should be noted here that:

1. Linear adjustment of model predictions can be useful and can be done in a 'real-time' setting (e.g. for flood

forecasting). External variables that are needed for a real-time external variable-based recalibration should be available at the time of the prediction adjustment (unlike, for example, the initial soil-water contents of Loague (1992d) used in the study reported here). Recalibrated predictions should be used for events similar to those used for estimation of the recalibration coefficients.

2. External variable-based recalibration results, e.g. Tables V and VII, may suggest how the model can be improved (see the 'Model bias investigations using recalibration' section). Recalibration of errors can be viewed as a complementary effort to process modelling: it only captures the linear part of the variation that remains after the model has explained the physically meaningful part. As such, it helps to diagnose model deficiencies and enables simple adjustments of predictions.

Instead of using the coefficient of model efficiency E_f , we prefer model performance evaluation based on RMSE because RMSE has an absolute lower bound. In addition, comparison of 'raw' RMSE with RMSE values after recalibration informs the model user how much of the variation in model errors could have been predicted from model output itself or from external variables.

ACKNOWLEDGEMENTS

This paper was written while the first author was a visiting scholar at Stanford University in the Department of Geological and Environmental Sciences. The Netherlands Organization for Scientific Research (NWO) supported this visit with a travel stipend. The effort described here is a CESIR contribution. The computer facilities used in this study were provided by an NSF equipment grant to Steven Gorelick and Keith Loague.

REFERENCES

- Addiscot T, Smith J, Bradbury N. 1995. Critical evaluation of models and their parameters. *Journal of Environmental Quality* **24**: 803–807.
- Anderson MG, Bates PD (eds). 2001. *Model Validation: Perspectives in Hydrological Science*. Wiley: Chichester.
- Anderson-Sprecher R. 1994. Model comparison and R^2 . *The American Statistician* **48**: 113–117.
- Anscombe FJ, Tukey JW. 1963. The examination and analysis of residuals. *Technometrics* **5**: 141–160.
- Beven KJ. 2001. *Rainfall-Runoff Modelling: The Primer*. Wiley: Chichester.
- Beven KJ, Binley AM. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* **6**: 279–298.
- Beven K, Freer J. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* **249**: 11–29.
- Donnelly-Makowecki LM, Moore RD. 1999. Hierarchical testing of three rainfall–runoff models in small forested catchments. *Journal of Hydrology* **219**: 136–152.
- Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Chapman and Hall: London.
- Engman ET, Rogowski AS. 1974. A partial area model for storm flow synthesis. *Water Resources Research* **10**: 464–472.

- Flavelle P. 1992. A quantitative measure of model validation and its potential use for regulatory purposes. *Advances in Water Resources* **15**: 5–13.
- Garrick M, Cunnane C, Nash JE. 1978. A criterion of efficiency for rainfall–runoff models. *Journal of Hydrology* **36**: 375–381.
- Güntner A, Uhlenbrook S, Seibert J, Leibundgut Ch. 1999. Multi-criterial validation of TOPMODEL in a mountainous catchment. *Hydrological Processes* **13**: 1603–1620.
- Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models*. Chapman and Hall: London.
- Hastie T, Tibshirani R, Friedman J. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag: New York.
- James LD, Burges SJ. 1982. Selection, calibration and testing of hydrologic models. In *Hydrologic Modeling of Small Watersheds, chapter 14*. Haan CT, Johnson HP, Brakensiek DL (eds). St. Joseph, MI: American Society of Agricultural Engineers.
- Klemes V. 1986. Operational testing of hydrological simulation models. *Hydrological Sciences Journal* **31**: 13–24.
- Konikow LF, Bredehoeft JD. 1992. Ground-water models cannot be validated. *Advances in Water Resources* **15**: 75–83.
- Krzysztofowicz R. 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research* **35**: 2739–2750.
- Krzysztofowicz R, Watada LM. 1986. Stochastic model of seasonal runoff forecast. *Water Resources Research* **22**: 296–302.
- Krzysztofowicz R, Herr HD. 2001. Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *Journal of Hydrology* **249**: 46–68.
- Kuczera G, Mroczkowski M. 1998. Assessment of hydrologic parameter uncertainty and the worth of multi-response data. *Water Resources Research* **34**: 1481–1489.
- Loague K. 1990. R-5 revisited: 2. Re-evaluation of a quasi-physically based rainfall–runoff model with supplemental information. *Water Resources Research* **26**: 973–987.
- Loague K. 1992a. Impact of overland flow plane characterization on event simulations with a quasi-physically based model. *Water Resources Research* **28**: 2541–2545.
- Loague K. 1992b. Using soil texture to estimate saturated hydraulic conductivity and the impact on rainfall–runoff simulations. *Water Resources Bulletin* **28**: 687–693.
- Loague K. 1992c. Soil-water content at R-5. Part 1. Spatial and temporal variability. *Journal of Hydrology* **139**: 233–261.
- Loague K. 1992d. Soil-water content at R-5. Part 2. Impact of antecedent conditions on rainfall–runoff simulations. *Journal of Hydrology* **139**: 253–261.
- Loague KM, Freeze RA. 1985. A comparison of rainfall–runoff modeling techniques on small upland catchments. *Water Resources Research* **21**: 229–248.
- Loague K, Gander RA. 1990. R-5 revisited: 1. Spatial variability of infiltration on a small rangeland catchment. *Water Resources Research* **26**: 957–971.
- Loague K, Green RE. 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. *Journal of Contaminant Hydrology* **7**: 51–73.
- Loague K, Kyriakidis PC. 1997. Spatial and temporal variability in the R-5 infiltration data set: déjà vu and rainfall–runoff simulations. *Water Resources Research* **33**: 2883–2895.
- Loague K, Vanderkwaak JE. 2004. Physics-based hydrologic response simulation: platinum bridge, 1958 Edsel, or useful tool. *Hydrological Processes* **18**: 2949–2956.
- Loague K, Gander GE, Vanderkwaak JE, Abrams RH, Kyriakidis PC. 2000. Simulating hydrologic response for the R-5 catchment: a never ending story. *Floodplain Management* **1**: 57–83.
- Loague K, Heppner CS, Abrams RH, Carr AE, Vanderkwaak JE, Ebel BA. 2005. Further testing of the Integrated Hydrology Model (InHM): event-based simulations for a small rangeland catchment located near Chickasha, Oklahoma. *Hydrological Processes* **19**: 1373–1398.
- Mroczkowski M, Raper GP, Kuczera G. 1997. The quest for more powerful validation of conceptual catchment models. *Water Resources Research* **33**: 2325–2335.
- Nash JE, Suthcliffe JV. 1970. River flow forecasting through conceptual models. Part I—a discussion of principles. *Journal of Hydrology* **10**: 282–290.
- Oreskes N, Shrader-Frechette K, Belitz K. 1994. Verification, validation, and confirmation of numerical models in Earth sciences. *Science* **263**: 641–646.
- Pebesma EJ, Switzer P, Loague K. 2005. Error analysis for the evaluation of model performance: rainfall–runoff event time series data. *Hydrological Processes* **19**: 1529–1548.
- Philip JR. 1969. Theory of infiltration. *Advances in Hydrosciences* **5**: 215–296.
- Sharma ML, Luxmoore PJ. 1979. Soil spatial variability and its consequences on simulated water balance. *Water Resources Research* **15**: 1567–1573.
- Thiemann M, Trosset M, Gupta H, Sorooshian S. 2001. Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research* **37**: 2521–2535.
- Vanderkwaak JE, Loague K. 2001. Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model. *Water Resources Research* **37**: 999–1013.
- Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. 1985. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* **90**: 8995–9005.
- World Meteorological Organization. 1975. Inter-comparison of conceptual models used in hydrological forecasting, Geneva, Switzerland, *Operational Hydrological Report* **7**: 172.
- Young P. 2001. Data-based mechanistic modelling and validation of rainfall–flow processes. In *Model Validation: Perspectives in Hydrological Science*, Anderson MG, Bates PD (eds). Wiley: Chichester.