Utility of different data types for calibrating flood inundation models within a GLUE framework

N. M. Hunter, P. D. Bates, M. S. Horritt, A. P. J. de Roo, M. G. F. Werner

▶ **To cite this version:**

N. M. Hunter, P. D. Bates, M. S. Horritt, A. P. J. de Roo, M. G. F. Werner. Utility of different data types for calibrating flood inundation models within a GLUE framework. Hydrology and Earth System Sciences Discussions, 2005, 9 (4), pp.412-430. hal-00304850

HAL Id: hal-00304850
https://hal.science/hal-00304850

Submitted on 18 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hydrology and Earth System Sciences, 9(4), *412–430* (2005) © EGU

Hydrology & Earth System Sciences

Utility of different data types for calibrating flood inundation models within a GLUE framework

Neil M. Hunter[1], Paul D. Bates[1], Matthew S. Horritt[2], P.J. De Roo[3] and Micha G.F. Werner[4]

[1]School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS, UK

[2]Department of Civil Engineering, University of Bristol, Queen's Building, University Walk, Bristol, BS8 1TR, UK

[3]European Commission, Joint Research Centre, Institute for Environment and Sustainability (IES), Weather Driven Natural Hazards Action; LM Unit, Via E. Fermi, TP 261, 21020 Ispra (Va), Italy

[4]WL | Delft Hydraulics, P.O. Box 5048, 2600 GA, Delft, The Netherlands

E-mail for corresponding author Neil.Hunter@bristol.ac.uk

Abstract

To translate a point hydrograph forecast into products for use by environmental agencies and civil protection authorities, a hydraulic model is necessary. Typical one- and two-dimensional hydraulic models are able to predict dynamically varying inundation extent, water depth and velocity for river and floodplain reaches up to 100 km in length. However, because of uncertainties over appropriate surface friction parameters, calibration of hydraulic models against observed data is a necessity. The value of different types of data is explored in constraining the predictions of a simple two-dimensional hydraulic model, LISFLOOD-FP. For the January 1995 flooding on the River Meuse, The Netherlands, a flow observation data set has been assembled for the 35-km reach between Borgharen and Maaseik, consisting of Synthetic Aperture Radar and air photo images of inundation extent, downstream stage and discharge hydrographs, two stage hydrographs internal to the model domain and 84 point observations of maximum free surface elevation. The data set thus contains examples of all the types of data that potentially can be used to calibrate flood inundation models. 500 realisations of the model have been conducted with different friction parameterisations and the performance of each realisation has been evaluated against each observed data set. Implementation of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology is then used to determine the value of each data set in constraining the model predictions as well as the reduction in parameter uncertainty resulting from the updating of generalised likelihoods based on multiple data sources.

Keywords: floods, hydraulic modelling, model calibration, uncertainty analysis

Introduction

Flood events across Europe in the summer of 2002 and during previous years have raised public, political and scientific awareness of flood risk and flood protection (Becker and Grünewald, 2003). Flooding is now widely acknowledged as an issue of strategic importance at a trans-national level, with major economic and social implications for citizens of many European countries (Samuels, 2003; Collier, 2003). In the absence of sufficient observations of flood extent, areas at risk from flooding are usually identified using numerical hydraulic models. This requires a dynamic approach to represent transient storage effects (Wheater, 2002), and various methods based on one- and two-dimensional hydrodynamic modelling have been presented with proven ability to simulate inundation extent, water depth and velocity for river and floodplain reaches up to 100 km in length (Bates *et al.*, 1998; Bates and De Roo, 2000; Werner, 2001; Werner, 2002a; Horritt and Bates, 2002).

In all but the simplest cases, some form of calibration is required to apply these models successfully to a particular reach for a given flood event. Calibration is undertaken to identify appropriate values for parameters such that the model is able to reproduce observations and, in the inundation case, typically considers roughness coefficients assigned to the main channel and floodplain. Though these

values may sometimes be estimated expertly in the field with a high degree of precision (Cunge, 2003), it has proven very difficult to demonstrate that such 'physically-based' models are capable of providing accurate predictions from single realisations for reasons discussed in the critiques of Beven (1989, 1996, 2001a) and Grayson *et al.* (1992). As such, values of parameters calculated by the calibration of models should be recognised as lacking a physical interpretation outside the model structure within which they were calibrated (Beven, 2000). Typically, data available for this process include water level and bulk discharge measured at flow gauging stations and, more rarely, flood extent data from satellites or air photos and distributed ground measurements of water level taken during or after a flood event. Given that the number of degrees of freedom in even the simplest of numerical models is relatively large, it is no surprise that many different combinations of effective parameter values may fit sparse validation data equally well (see, for example, Romanowicz *et al.*, 1994, 1996; Aronica *et al.*, 1998, 2002; Romanowicz and Beven, 2003; Bates *et al.*, in press). In response to this problem, uncertainty analysis techniques, often based on the Generalised Likelihood Uncertainty Estimation (GLUE) methodology of Beven and Binley (1992), have been developed.

To date, only very limited attempts to calibrate distributed models against more than one particular data type have been made. Horritt and Bates (2002) tested the predictive performance of three industry-standard hydraulic codes on a 60-km reach of the River Severn, UK, using independent calibration data from hydrometric and satellite sources. They found that all models were capable of simulating inundation extent and floodwave travel times to similar levels of accuracy at optimum calibration. However, due to the different model responses to friction parameterisations, differences emerged according to the calibration data used when the models were used in predictive mode. Horritt and Bates (2002) did not consider either the potential for combining both data sources in the calibration process or assessing the uncertainties associated with the model predictions. The value of incorporating additional data in the calibration process has been explored tentatively in other areas of distributed modelling, notably by Franks *et al.* (1998) in a catchment hydrology context. In this study, the authors used Synthetic Aperture Radar (SAR) imagery to obtain soil saturation maps to compare with predictions of soil moisture from the catchment hydrology code, TOPMODEL (Beven and Kirkby, 1979). Using these data as supplementary information to constrain the model predictions of discharge for the catchment, they showed that the addition of this information enabled the rejection of many previously acceptable parameterisations resulting in the improved prediction of some discharge events.

Multiple observational data sets for historical flood events are still exceedingly rare, so the potential value of additional observations in the calibration process has yet to be explored in the case of distributed inundation models. Furthermore, the use of uncertainty estimation techniques during this conditioning process allows the relative value of individual (sets of) observations to be quantified precisely in terms of the reduction in uncertainty over effective parameter specification (c.f. Beven and Binley, 1992). Interpretation of these uncertainty measures may also provide guidance over *how much* and *of what type* of observational data would be required to achieve given levels of uncertainty reduction in simulated variables.

There is thus a clear need to develop methods for assessing the relative utility of different observational data types for the calibration of distributed floodplain inundation models, both in terms of quantification of (1) the uncertainties associated with the simulation of various distributed hydraulic variables and (2) the (hopeful) reduction in uncertainty over effective parameter specification. Fulfilment of such aims is an important component of the European Flood Forecasting System (EFFS) project where ultimately the wish is to translate 10-day ahead forecasts of flood discharge into maps of inundation probability. This requires a thorough understanding of the response of hydraulic models to a variety of different types of calibration data, objective functions and rejection criteria.

Study site, data availability and model description

The identification of an appropriate data set, which encompasses both multiple observations for model output evaluation and a commensurate specification of boundary conditions, is central to the development of methods proposed in this paper. To test these techniques, the LISFLOOD-FP distributed hydraulic model was applied to a 35-km reach of the River Meuse between the gauging stations at Borgharen (near the city of Maastricht) in The Netherlands and Maaseik in Belgium for the January 1995 flood event. Commencing at 00:00 hours on 22 January and continuing for 20 days (or 480 hours), this severe flood had an estimated return period of 63 years and resulted in extensive inundation of the river valley. Initial data collected and available for study have been described and used in Bates and De Roo (2000) and Werner (2002b). These data, in addition to other flow observations made available recently, contain examples of all the types of information that can potentially be used to calibrate flood inundation models and are summarised in Table 1 and Fig. 1.

The availability of four hydrometric gauges, two external (Borgharen and Maaseik) and two internal (Elsloo and Grevenbicht) to the model domain, should provide a good description of the spatially-integrated response of the study reach to the passage of the flood wave. It is also likely that these observations, collected at hourly intervals, will have sufficient temporal resolution to discriminate between competing models. The data consist of discharges and levels at the external and internal gauges respectively. Given the relatively short interval between stations and the inherent uncertainties of stage-discharge relationships at each station, water levels recorded internal to the model domain are likely to prove a more valuable asset than discharges in conditioning model performance, particularly in the detection of erroneous compensating errors.

To have more than one observation of inundation extent per flooding episode is highly unusual. However, for the 1995 event, the inundation of the Meuse floodplain was captured in both air photo imagery and by an overpass of the ERS-1 SAR satellite system (De Roo *et al*., 1999. These multiple observations of flood extent should provide a rigorous test of the model's ability to simulate a dynamic flood shoreline. The SAR overpass occurred on 30 January at 1033 hours when the discharge at Borgharen was 2631 $m^3 s^{-1}$. The air photo survey was conducted on 27 January when the discharge at Borgharen was 2645 $m^3 s^{-1}$ and water levels along the reach measured by the Dutch Water Authorities (RWS) were approximately 0.1 m lower than those recorded at peak flow. Unfortunately in this instance, despite the different sampling periods for the observed data, variations between data sets due to different hydraulic conditions are likely to be small. That there are considerable classification discrepancies in regions where the two 'observations' coincide is a direct result of the different processing strategies used to derive extent shorelines from the two image sources. The air photo mosaic was digitised manually by RWS and has an approximate horizontal accuracy of 25 m while the SAR imagery was processed automatically using the statistical active contour model or 'snake' (Horritt, 1999). The central problem with SAR image processing is how to combat the high level of noise (or speckle) without the degradation in spatial resolution associated with many local averaging techniques (for a full review see Horritt, 1999). The snake algorithm deals with this by measuring local speckle statistics along the shoreline and is thus able to segment the shoreline to an accuracy of ~1 pixel (12.5 m for ERS-1 SAR). However, problems may still occur as increased back scattering of the radar signal by wind roughening of the water surface and particular land use types can lead to misclassification of flooded areas. In general, misclassification errors will be greater with the SAR data than the low altitude airborne survey, and so the latter is likely to be the data set closest to the true shoreline. From Fig, 1, it is also obvious that there are several regions, particularly in the vicinity of Elsloo, where both observations of inundation extend beyond the boundaries of the available DEM. Such areas will never be simulated as flooded by the model and so cannot reasonably be considered in the evaluation process.

The final source of data available for the 1995 flood event is 86 point observations of maximum free surface elevation surveyed post-event by RWS using traditional ground survey methods. While this type of data is subject to well-documented limitations (e.g. Beven, 1989; Lane *et al.*, 1999), their broad spatial distribution across the Dutch floodplain (right bank) and non-binary nature (i.e. possession of a quantity beyond simply 'wet' or 'dry') make them potentially very valuable in the model conditioning process. However, because of the inherent sensitivity to topographic description within the model, simulated variables are not always compatible with these data in every instance. In two cases, the surveyed levels are actually below

Table 1. Summary of observational data sources available for the January 1995 flood event on the River Meuse.

Observational data type	*Source*	*Description*
Internal bulk flow time series	Stage at internal gauge 1, Elsloo Stage at internal gauge 2, Grevenbicht	Hourly gauged stage hydrographs
External bulk flow time series	Discharge at downstream boundary, Maaseik	Hourly gauged stage hydrograph converted using rating curve
Vector polygons	Satellite radar-derived inundation shoreline	ERS-1 SAR imagery converted into a shoreline using the statistical active contour model of Horritt (1999)
	Air photo-derived inundation shoreline	Image mosaic converted into a shoreline by the Dutch Water Authorities (RWS)
Points	Maximum free surface elevation survey	Systematic ground survey conducted by RWS

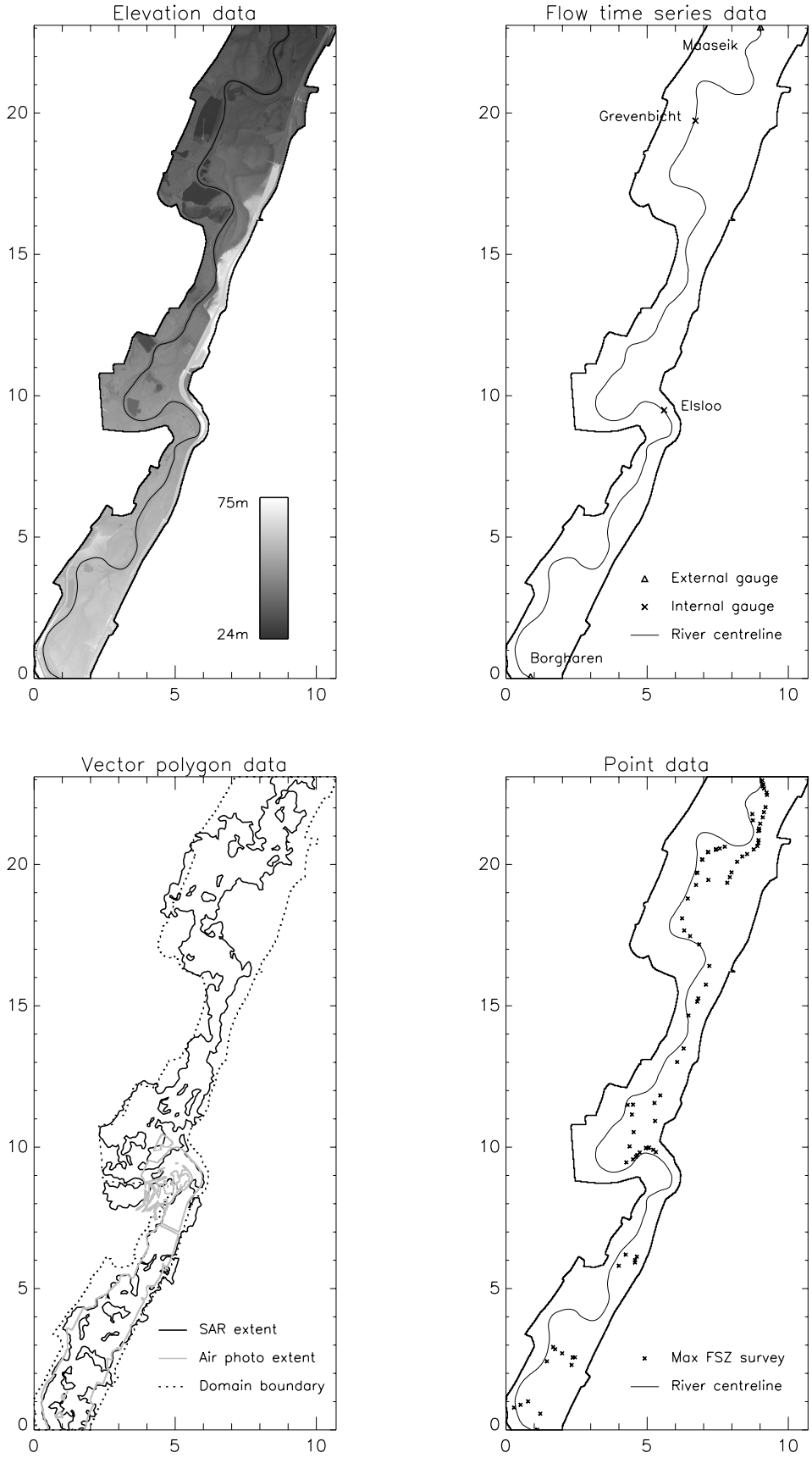

Fig. 1. *Spatial distribution of observational data sources available for the January 1995 flood event on the River Meuse overlain on the 50 m resolution Digital Elevation Model.*

the 'dry' floodplain as represented in the DEM and subsequently could not be used in model evaluation.

Due to the difficulties in collecting measurements in the field, there is a trend in distributed environmental modelling to ignore the errors associated with observational sources. However, errors and uncertainties associated with each type of observation data may have a significant impact on the values of effective parameters estimated within distributed models using some or all of these sources. The sets of values estimated for freely-varying parameters will ultimately dictate the predictive performance of the model and it is, therefore, important to assess how the aggregate of all these uncertainties (although not explicitly those relating to boundary condition specification and model structure) propagate through the model calibration process. To demonstrate a methodology capable of making such assessments requires a distributed hydraulic model with which to perform the necessary simulations. There is evidence that when evaluating against sparse observational data, fully dynamic models do not necessarily produce better results than models with simplified dynamics (Bates and De Roo, 2000; Horritt and Bates, 2001, 2002). Moreover, the computational efficiency afforded by simpler approaches makes them highly appealing for evaluating multiple model realisations within a Monte Carlo framework.

LISFLOOD-FP is a raster-based inundation model specifically developed to take advantage of high resolution topographic data sets (Bates and De Roo, 2000) and is based on the storage cell concept of Cunge *et al.* (1980). Similar approaches have been proposed by Bechteler *et al.* (1994), Estrela and Quintas (1994) and Romanowicz *et al.* (1996) although LISFLOOD-FP differs from these schemes by incorporating a one-dimensional wave routing model for channel flow (rather than using uniform flow formulae) and by the method and scale of floodplain discretisation. The model is fully described by Bates and De Roo (2000) and Horritt and Bates (2001), so only the main features and assumptions are given here.

Channel flow is handled using a one-dimensional kinematic routing procedure that is capable of capturing the downstream propagation of a flood wave and the response of the free surface slope, which can be described in terms of continuity and momentum equations as:

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = q \tag{1}$$

$$S_0 - \frac{n^2 P^{4/3} Q^2}{A^{10/3}} = 0 \tag{2}$$

Q is the volumetric flow rate in the channel, A, cross sectional area of the flow in the channel, q, the flow into the channel from other sources (i.e. from the floodplain or tributary channels), S_0, the down-slope of the bed, n, the Manning's coefficient of friction, P, the wetted perimeter of the flow, and h, the flow depth. Equations 1 and 2 are solved using an implicit non-linear first order finite difference scheme. A flow rate is imposed at the upstream end of the reach, which for a kinematic wave model is sufficient as a boundary condition, as wave effects can only propagate downstream and any backwater effects are ignored. Additional gauging stations, both internal to the model domain and at the downstream outlet, are thus fully independent of the model and may be used as evaluation data in the calibration process. The channel parameters required to run the model are its width, bed slope, depth (for linking floodplain flows) and Manning's n value. Width and depth are assumed to be uniform along the reach, their values assuming the average effective values taken from field surveys of the channel. The Manning's n roughness is left as a calibration parameter to be estimated.

Floodplain flows are described similarly in terms of continuity and momentum equations, discretised over a grid of square cells, which allows the model to represent two-dimensional dynamic flow fields on the floodplain. Flow between two cells is simply calculated as a function of the free surface slope between those cells (Estrela and Quintas, 1994):

$$\frac{\partial h^{i,j}}{\partial t} = \frac{Q_x^{i-1,j} - Q_x^{i,j} + Q_y^{i,j-1} - Q_y^{i,j}}{\Delta x \, \Delta y} \tag{3}$$

$$Q_x^{i,j} = \frac{h_{\text{flow}}^{5/3}}{n} \left(\frac{h^{i-1,j} - h^{i,j}}{\Delta x} \right)^{1/2} \Delta y \tag{4}$$

where $h^{i,j}$ is the water free surface height at the node (i, j), Δx and Δy are the cell dimensions, n is the Manning's friction coefficient for the floodplain, and Q_x and Q_y describe the volumetric flow rates between floodplain cells. Q_y is defined analogously to Eqn. 4. The flow depth, h_{flow}, represents the depth through which water can flow between two cells, and is defined as the difference between the highest water free surface in the two cells and the highest bed elevation (this definition has been found to give sensible results for both wetting cells and for flows linking floodplain and channel cells). While this approach does not actually simulate diffusive wave propagation on the floodplain, due to the decoupling of the x and y components of the flow, it is computationally simple and has been found to make a negligible difference when compared with results simulated using a faithful finite difference discretisation of the diffusive wave equation (Horritt and Bates, 2001).

Equation 4 is also used to calculate flows between

floodplain and cells, allowing floodplain cell depths to be updated using Eqn. 3 in response to flow from the channel. These flows are also used as the source term in Eqn. 1, effecting the linkage of channel and floodplain flows. Thus only mass transfer between channel and floodplain is represented in the model, and this is assumed to be dependent only on relative water surface elevations. While this neglects effects such as channel-floodplain momentum exchange mechanisms and the effects of advection and secondary circulation on mass transfer, it is the simplest approach to the coupling problem and should reproduce some of the behaviour of the real physical system (Aronica *et al.*, 2002).

Model calibration and uncertainty estimation within a GLUE framework

Some form of inverse modelling procedure is now required to turn observed data into estimates of model parameters. In a classical calibration study, this would involve the identification of an optimum parameter set that maximises the fit between model predictions and observations. However, such a deterministic scheme effectively ignores any uncertainties in the modelling process. The aim of this paper is to assess the worth of observed data in an uncertain framework, and this calibration process needs to be recast in an uncertain form. Given the imperfect knowledge of the statistical properties of the observed data, it is difficult to formulate an error model for the observations, and also difficult to justify making strong assumptions about the observed data's statistics (e.g. independence, distribution) as this may affect the outcome of the calibration. A procedure is thus required in which uncertain observed data can be used to calibrate a non-error free model and allow the uncertainty in model parameters to be assessed, without making restrictive assumptions about errors in the observed data.

The GLUE methodology of Beven and Binley (1992) is one such method, and provides a simple, flexible approach to parameter sensitivity analysis, model conditioning and uncertainty estimation. It is based on rejecting the idea that there is a unique optimum parameter set in a model calibration in favour of identifying the many different combinations of parameter values that may be equally acceptable in simulating the system under study. In this situation it is possible to evaluate only the relative likelihood of a given non-error free model and parameter set in reproducing the non-error free data available to test the model. 'Consistency' in some sense, defined with respect to the application in mind, is central to the concept of equifinality (Beven, personal communication). However, models that are deemed unacceptable or non-behavioural may be rejected and removed from subsequent evaluation by being assigned a generalised likelihood of zero.

A number of decisions must be made when implementing the GLUE methodology (Beven, 2001b): (1) a decision about posing a computational tractable calibration problem (i.e. how best to simplify the characteristically high dimensional parameter spaces of distributed models); (2) a decision about the feasible range and sampling strategy for each parameter; (3) a decision about appropriate generalised likelihood measure(s) for each source of observational data; (4) a decision about criteria for acceptance or rejection of models; and, (5) a decision about the methodology for updating (or combining) generalised likelihood measures.

Because of their subjective nature, it is important that decisions made at each stage of the GLUE procedure be transparent and unambiguous. However, it should also be noted that transparency in the decision making does not eliminate this subjectivity.

In distributed inundation models, roughness coefficients may theoretically be specified at each computation node in the model domain and, with recently published models having upwards of ten thousand grid elements (Aronica *et al.*, 2002), this poses a formidable calibration problem. Moreover, despite the fact that friction values vary markedly over the floodplain in space due to vegetation changes and in time during the flood event with changes in stage, typically available observational data have not been sufficiently detailed to require such sophistication from the model (Bates *et al.*, 1998). Thus some simplification of this high dimensional calibration problem can undertaken, particularly since it is known that the storage cell codes can be relatively insensitive to floodplain roughness specification (Romanowicz *et al.*, 1996; Horritt and Bates, 2001). Initial simulations exhibited little response to floodplain friction, but a crudely distributed channel friction calibration was found to be worthwhile. The distribution of gauging stations along the reach allowed channel roughnesses to be aggregated, reasonably, into single effective values for three distinct sub-sections (reach 1, 'Upstream' — Borgharen to Elsloo; reach 2, 'Midsection' — Elsloo to Grevenbicht; and reach 3, 'Downstream' — Grevenbicht to Maaseik), while a single roughness value was assigned to each floodplain grid cell uniformly.

The next step of the GLUE procedure is to decide the range of the parameter space to be examined, which relies upon an 'informed knowledge' of the system. However, this initial decision can exert an influence on resulting predicted uncertainties as parameter values outside this range are effectively assigned a generalised likelihood of zero. Importantly, calibration of friction coefficients for the channel and floodplain was not undertaken in the Bates and

De Roo (2000) study which used values selected on a physical basis from available literature (e.g. Chow, 1959). Following a preliminary investigation, friction coefficients (Manning's n) for the calibration process were distributed randomly, independently and uniformly between 0.02 $m^{1/3} s^{-1}$ and 0.05 $m^{1/3} s^{-1}$ for each channel sub-reach, with floodplain friction fixed at 0.06 $m^{1/3} s^{-1}$. Such an approach is relatively standard in inundation modelling and so is evaluated here. This gave a three-dimensional parameter space that was sampled by Monte Carlo methods using 500 realisations of the model.

Many different statistical measures exist to evaluate the 'goodness of fit' of a model simulation. Selection of an appropriate generalised likelihood measure will depend primarily on what observational data are available to evaluate the model but also on the purposes of the study. Different output variables will also demand different types of measure to facilitate evaluation. In this study, three independent measures are required for quantifying errors in simulating: (1) at-a-point time series of levels and discharges (hydrometric records); (2) spatially distributed, binary pattern data (flood extent); (3) spatially distributed, continuous point data (maximum water levels).

There are several methods for fitting simulated and observed hydrographs and these invariably have a bias towards one specific characteristic of the hydrograph (e.g. accurate recession or peak prediction). Some well-established 'traditional' measures, e.g. the Nash-Sutcliffe model efficiency, have achieved widespread usage, often without due consideration of their limiting assumptions concerning the probability distributions of the residual errors (Clarke, 1973; Green and Stephenson, 1986; Beven, 2001b; Christiaens and Feyen, 2002). Based on the sum of error variances, the Nash-Sutcliffe efficiency is sensitive to differences in both maximum values and timing of flood peak. It is known from statistical theory that the error variance is most suitable as a performance measure when errors between the observations and predictions are of mean zero, are normally distributed with constant variance and are not correlated (Beven, 2001b). Very often, hydrometric data violate these assumptions and so the Heteroscedastic Maximum Likelihood Estimator (HMLE) measure of Sorooshian and Dracup (1980) was proposed to account properly for the presence of either autocorrelation (non-independence) or heteroscedasticity (changing variance) of data errors. It has the form:

$$\mathrm{HMLE} = \frac{\frac{1}{NT}\sum_{t=1}^{NT} w_t \left[Q_t - \hat{Q}_t(\Theta, Y)\right]^2}{\left[\prod_{t=1}^{NT} w_t\right]^{1/NT}} \qquad (5)$$

where Q_t is the observed value at time index t; $\hat{Q}_t(\Theta, Y)$ is the simulated variable given by parameters Θ and input data Y, and NT is the number of time steps. The weights w_t are defined as $w_t = Q_t^{2(\lambda-1)}$ where l is an unknown shaping parameter linked to the variances. By using a form of weighted least-squares approach, the HMLE measure tends to weight the fit of the model more towards recession periods than high flow periods and generally provides a better overall measure than the Nash-Sutcliffe efficiency which gives more weight to higher flows. Initial evaluation of LISFLOOD-FP's downstream discharge predictions indicated that the HMLE measure also gives better discrimination between competing parameters sets than the Nash-Sutcliffe efficiency, and is therefore used throughout. The HMLE measure is minimised for an optimal fit between data, so it is subtracted from 1 and rescaled to sum to unity to give a generalised likelihood value (Wagener *et al*., 2001).

A spatially-distributed, binary pattern (wet/dry)or flood is essentially being dealt with for flood extent data. However, comparing such data with a modelled binary inundation pattern is not straightforward, with potential problems arising when models of different reaches or magnitudes of flood event are intercompared (Aronica *et al*., 2002). Model predictions of inundation extent can be compared with the synoptic observations using measures of fit based on a contingency table that shows the frequency of 'wet' and 'dry' predictions and observations as described in Table 2.

The measure used in Aronica *et al*. (2002):

$$F^{\langle 1\rangle} = \frac{\sum_{i=1}^{NC} P_i^{M_1 D_1}}{\sum_{i=1}^{NC} P_i^{M_1 D_1} + \sum_{i=1}^{NC} P_i^{M_1 D_0} + \sum_{i=1}^{NC} P_i^{M_0 D_1}} \qquad (6)$$

has been modified in this study to penalise, additionally, overprediction of the flood extent:

$$F^{\langle 2\rangle} = \frac{\sum_{i=1}^{NC} P_i^{M_1 D_1} - \sum_{i=1}^{NC} P_i^{M_1 D_0}}{\sum_{i=1}^{NC} P_i^{M_1 D_1} + \sum_{i=1}^{NC} P_i^{M_1 D_0} + \sum_{i=1}^{NC} P_i^{M_0 D_1}} \qquad (7)$$

Table 2. Matrix of possible model/data combinations for a binary classification scheme.

	Present in data (D_1)	*Absent in data* (D_0)
Present in model (M_1)	$M_1 D_1$	$M_1 D_0$
Absent in model (M_0)	$M_0 D_1$	$M_0 D_0$

where i is the grid cell (or pixel) index and NC represents the total number of grid cells in the model. This formulation was found to give enhanced discrimination between inundation patterns and reduce the tendency of the original measure (Eqn. 6) to preferentially weight overprediction. The measure gives a value of 1 when the observed and predicted inundation patterns coincide exactly.

The maximum water elevations are compared using a sum of absolute errors, with special consideration given to points that are observed as being wet, but predicted as dry. For example, for points observed and modelled as wet, the absolute error in the free surface elevation is calculated:

$$\mathrm{SAE} = \sum_{p=1}^{NP} \left| FSZ_p - \hat{FSZ}_p(\Theta, Y) \right| \tag{8}$$

where FSZ_p and $\hat{FSZ}_p(\Theta, Y)$ are the observed value and simulated variable at point observation p respectively; and NP represents the total number of points in the evaluation data set. For points that are observed wet but predicted dry, the difference between the observed water elevation and the model's DEM elevation is calculated. This measure effectively gives the absolute predicted depth error, corrected for potential discrepancies in the representation of floodplain topography in the model.

Objective function values for each parameter set, Θ, were then transformed into generalised likelihoods, $L(\Theta)$, according to the methodology proposed by Wagener *et al*. (2001). This ensures all generalised likelihood values are positive and sum to unity (i.e. $\sum L(\Theta) = 1$). This measure can be treated as analogous to a true probability, but cannot be used for formal statistical inference.

The generalised likelihood measures defined above can now be used to weight each model realisation and hence each parameter set with a value corresponding to the confidence in that parameter set as a good predictor of the system behaviour. Furthermore, the observed data sets can be combined, by calibrating on one data set, then using another to 'update' the weights assigned to each parameter set. This is synonymous with the more common practice in GLUE of updating an existing generalised likelihood estimate with a new measure calculated for the prediction of an additional set of observations from a second flood event (e.g. Romanowicz and Beven, 2003). One way of combining generalised likelihoods is proposed by Lamb *et al*. (1998), where a Bayes-type equation is expressed in the following form:

$$L_p(\Theta \mid Y) = \frac{L_o(\Theta)\ L(\Theta \mid Y)}{C} \tag{9}$$

where $L_o(\Theta)$ is the prior generalised likelihood of parameter set Θ, $L(\Theta \mid Y)$ is the generalised likelihood calculated for the current evaluation given the set of observations Y, $L_p(\Theta \mid Y)$ is the posterior generalised likelihood, and C is a scaling constant to ensure that the cumulative posterior generalised likelihood is unity. This method of combining generalised likelihoods assumes that they behave in the same way as probabilities and is consistent with the definitions of generalised likelihood and confidence within the GLUE procedure. Whilst the same assumption could perhaps not be made for an uncertainty method based on formal statistical inference, in this context it is a practically expedient and commonly used approach to the problem. In the absence of any observations, the prior generalised likelihood is assumed subjectively to reflect expert prior knowledge of parameter distributions as in this case, where a uniform distribution has been assumed. A number of updating equations can be chained together to combine many data sources, but care is required as the multiplicative nature of the Bayes equation may lead to assigning zero weight to all models in cases where multiple observations are available. In this case either a better model or less stringent likelihood criteria are required.

Results and discussion

The above generalised likelihood measures were applied to each of the 500 model realisations according to the GLUE methodology outlined above. An impression of the sensitivity of individual calibration parameters can be gained by plotting scatter diagrams of parameter value against performance measure evaluated for a single observational data type. Dotty plots project the goodness of fit response surface onto individual parameter dimensions with each dot representing one run of the model with friction parameter values chosen randomly by uniform sampling across the ranges of each parameter.

Figure 2 shows dotty plots of objective function (2a–i) and generalised likelihood (2j–r) values using the Heteroscedastic Maximum Likelihood Estimator as the criterion of model performance evaluated for the available hydrometric observations. It can be seen that as model performance or goodness of fit to the available observational data increases, the value of the objective function and corresponding likelihood decreases and increases respectively. For predictions of stage at the internal gauges, parameter sensitivity is well-defined and dominated by channel friction values assigned to the reach upstream of the respective gauge. This localised calibration response is intuitively reasonable given the assumptions and structure of the LISFLOOD-FP model. By using a kinematic wave approximation to represent channel flow, dynamic effects arising from varying parameter specification in this region

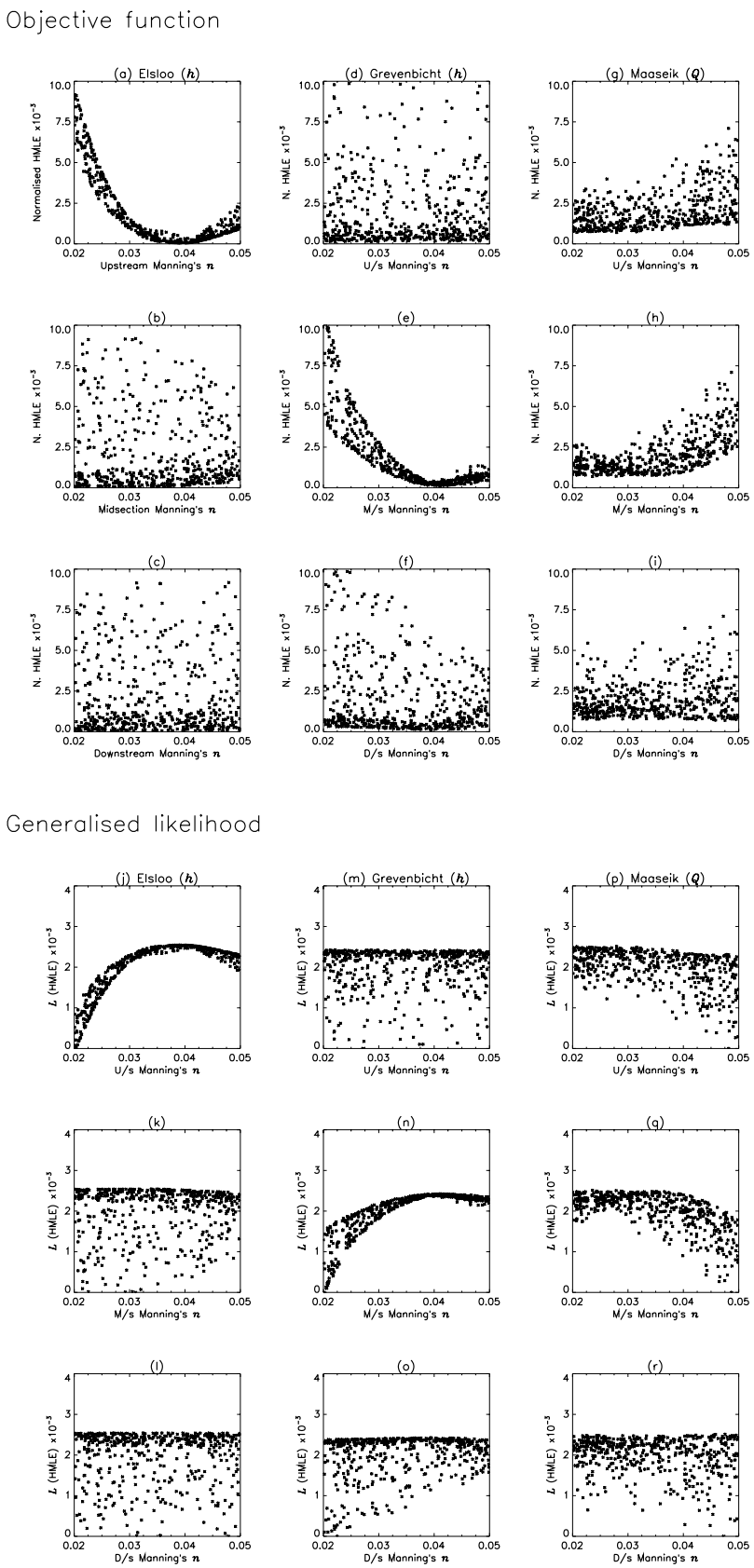

Fig. 2. *Dotty plots of objective function (2a–i) and generalised likelihood (2j–r) values using the Heteroscedastic Maximum Likelihood Estimator as the criterion of model performance.*

may only propagate downstream unless they are explicitly linked to upstream areas by adjacent near-channel floodplain topography. Furthermore, the well-constrained nature of these distributions increases the temptation to (attempt to) identify 'optimum' parameter values for each calibration parameter. However, it is well-documented that this optimum will be non-stationary when evaluated on alternative measures and data sets (e.g. Gupta *et al*., 1998). For the discharge evaluations at Maaseik, it can generally be seen that for each parameter there are good simulations across the whole range of feasible values. The apparent increase in sensitivity to parameter values specified for the upstream and midsection reaches reflects their greater length within the model domain. It is thus reasonable to expect that the evaluation of internal predictions of stage and, to a lesser degree, the external predictions of discharge should offer considerable potential for reducing uncertainty over effective parameter specification.

As well as weighting each parameter set according to the generalised likelihood measures developed above, it is also possible to construct prediction quantiles by applying each generalised likelihood weight to model-predicted variables. This allows the uncertainty in continuous time-evolving predictions (e.g. hydrometric variables) to be visualised by the construction of a cumulative generalised likelihood distribution at each model timestep. Thus the dynamic behaviour of parameter uncertainty and its manifestation in the simulated variables can be assessed.

In keeping with the subjective nature of the GLUE procedure adopted so far, rather than attempting to define probabilities, relative confidence measures (RCM) for model predicted variables are derived. These express belief that a prediction is a true representation of the system behaviour for the single model structure used, but do not express any measure of confidence in that model over competing structures. These relative confidence measures can be expressed analogously to cumulative probabilities:

$$RCM\left(\hat{Q}_t < q\right) = \sum_{j=1}^{NR} L\left[\left(\Theta_j\right) \mid \left(\hat{Q}_{j,t} < q\right)\right] \qquad (10)$$

where $\hat{Q}_{j,t}$ is the variable of interest predicted by the j^{th} Monte Carlo sample. Prediction quantiles, $RCM\left(\hat{Q}_t < q\right)$, obtained in this way are thus conditioned on inputs to model, the model responses for the particular sample of parameter sets used, the subjective choice of generalised likelihood measure and the observations used in the calculation of the likelihood measure (Beven, 2001). In Fig. 3, the 5 and 95% quantiles are considered, resulting in a 90% uncertainty 'envelope'. The uncertainty for each gauge is then evaluated using the HMLE measure to compare predictions and observations at that gauge.

Figure 3 shows these quantiles, along with the observations of stage at Elsloo and Grevenbicht (internal gauges) and discharge at Maaseik (downstream boundary of the reach). The most striking feature is the difference in uncertainty between the stage and discharge measurements. This is a result of the model structure and calibration process: different friction coefficients will produce different predicted water levels but with approximately the same downstream

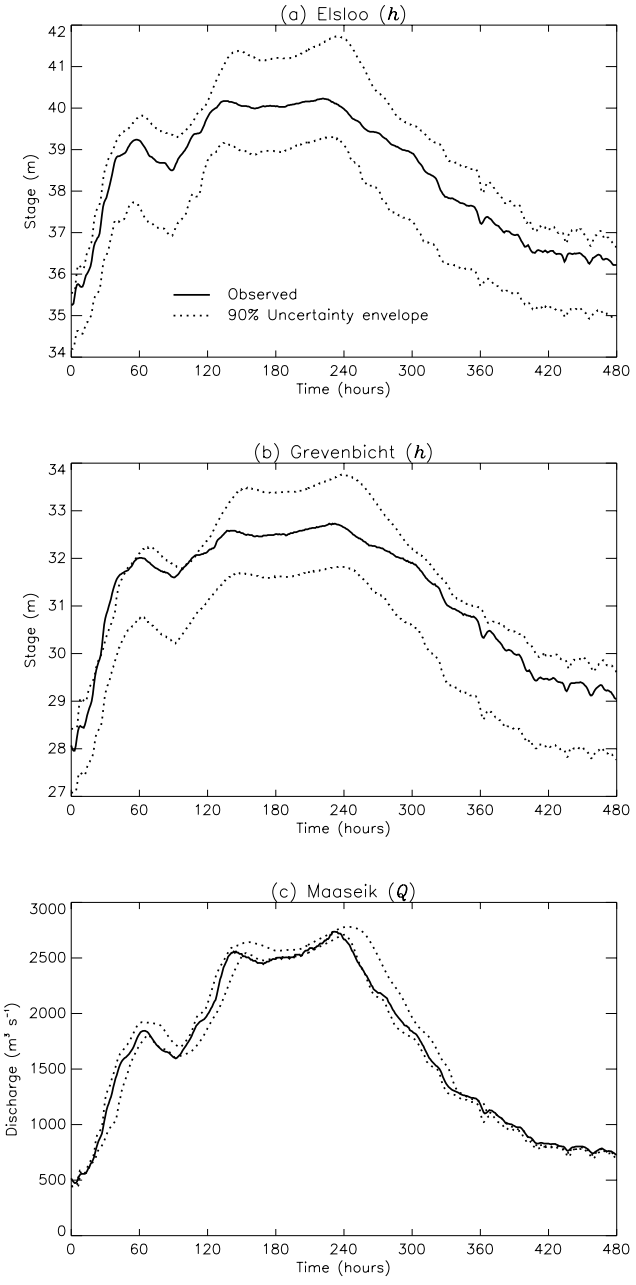

Fig. 3. *Prediction bounds for (3a) stage at Elsloo; (3b) stage at Grevenbicht; and (3c) discharge at Maaseik after conditioning on respective hydrometric observations only.*

discharge. This effect is exacerbated by the dynamic behaviour of the reach and hydrograph, as the reach response time is short compared to the duration of the flood event. Thus the model behaviour could be approximated by a series of steady states, and the downstream hydrograph is affected little by the friction parameterisation. That the uncertainty envelope does not always bracket the observations at Grevenbicht may be a result of the model failing to represent adequately the complex flow dynamics (e.g. backwater or inertial effects) associated with the wide, shallow floodplain at this point. Alternatively, this may indicate that some of the subjective choices made during the GLUE procedure are inadequate and need to be reconsidered. At Elsloo, in contrast, floodplain flow is well constrained and the bulk flow behaviour is more easily represented by the simple dynamics of the model. The relationship between the quantiles and the observed stages also changes throughout the flood event. At peak flow, the observed data lie approximately midway between the quantiles, whereas on the receding limb, especially for the Grevenbicht gauge, the observations lie toward the 95th percentile. Model response to the calibration process is thus different for high and low flows.

The use of inundation extent observations to calibrate model performance is explored in Figs. 4 and 5. In Fig. 4, it can be seen that for the upstream section of the reach the flood was essentially a 'valley-filling' event — i.e. once the valley is filled any value of n will produce acceptable results when compared with the binary pattern data. For the mid- and downstream reaches, there is a steady decrease in performance as Manning's n increases. This is a result of the generalised likelihood measure penalising over-prediction in regions where the flood is not bounded by steep slopes or defence structures. The combination of the

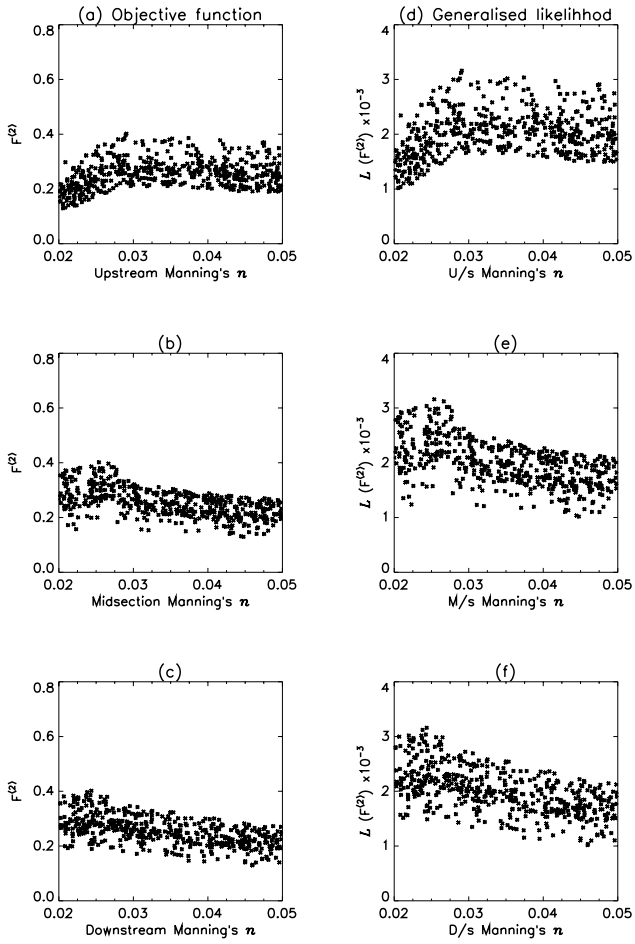

Fig. 4. D*otty plots of objective function (4a–c) and generalised likelihood (4d–f) values using the* $F^{(2)}$ *performance measure to compare model predictions of inundation extent with the shoreline derived from satellite radar data.*

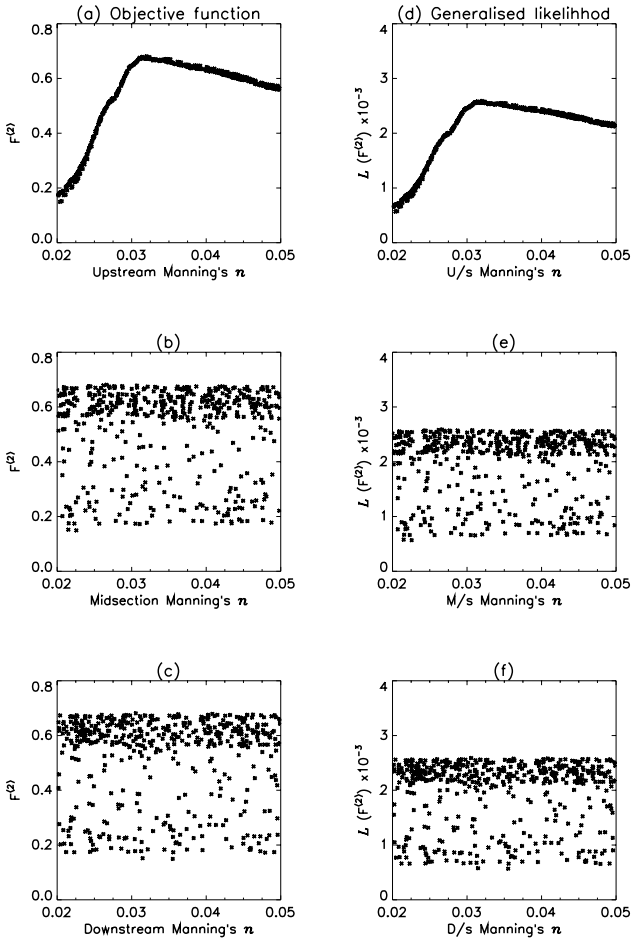

Fig. 5. *Dotty plots of objective function (5a–c) and generalised likelihood (5d–f) values using the* $F^{(2)}$ *performance measure to compare model predictions of inundation extent with the shoreline derived from air photo data. The predominant sensitivity to variations in friction value assigned to the upstream reach is reasonable given the spatial extent of the observational data.*

generalised likelihood measure with the SAR inundation extent data is, however, relatively insensitive to changes in Manning's *n*, and potentially there will be only a small reduction in parameter uncertainty from calibrating with these data and objective function. Figure 5 shows model response when calibrated against air photo flood extent data, with a marked difference when compared to calibration with the SAR data, despite the same generalised likelihood measure being used. The upstream Manning's *n* value is well constrained, with a clearly defined optimum around 0.032 $m^{-1/3}$ s, and the drop in performance apparent as Manning's *n* increases from this value is not apparent in the far noisier SAR data. The mid- and downstream reaches show no sensitivity to Manning's *n* as the air photo data does not extend this far down the reach. These two figures show that model response to the calibration process depends not only on the choice of generalised likelihood measure, but also on the errors in the data set: the much higher quality air photo flood extent has given a more clearly defined calibration response than the noisy SAR data.

To unearth the spatial uncertainty in model predictions, the uncertainty in the two-dimensional binary inundation field must be visualised. Here, too, a relative confidence measure is derived which expresses belief that a given pixel will be flooded, given the uncertainty in model parameters. This is done by taking the flood state as predicted by the model for each pixel for each realisation and weighting it according to the measure of fit $F^{\langle 2\rangle}$, to give a RCM of flooding for each pixel *i*, RCM_i^{flood}:

$$RCM_i^{flood} = \frac{\sum_{j=1}^{NR} f_i \, F^{\langle 2\rangle}(\Theta_j, Y)}{\sum_{j=1}^{NR} F^{\langle 2\rangle}(\Theta_j, Y)} \tag{11}$$

where f_j takes a value of 1 for a flooded pixel and is zero otherwise and $F^{\langle 2\rangle}(\Theta_j, Y)$ is the global performance measure for model realisation h:*j* and data set *Y*. RCM_i^{flood} will assume a value of 1 for pixels that are predicted as flooded in all simulations and 0 for pixels always predicted as dry, as the generalised likelihoods are renormalised to sum to unity. Model uncertainty will manifest itself as a region of pixels with intermediate values, maximum uncertainty being indicated by pixels with $RCM_i^{flood} \approx 0.5$. Such a RCM_i^{flood} map is shown in Fig. 6 for models calibrated against the SAR and air photo data, along with the respective inundation shoreline. It is evident that variations in RCM_i^{flood} follow the microtopography and paleo-features of the floodplain to be identified as these are inundated in some cases depending on the value of Manning's *n* specified for a particular reach. Thus RCM_i^{flood} exhibits a high degree of spatial heterogeneity and the calibration process has generated a continuous gradient from regions of $RCM_i^{flood} = 1$ (e.g. channel) to $RCM_i^{flood} = 0$ (e.g. high ground). Furthermore, the localised model response is clearly demonstrated by the RCM_i^{flood} evaluated on the air photo observation, which shows a marked increase in uncertainty away from the conditioning observation.

Figure 7 shows the objective function and generalised likelihood for model predictions when compared with point maximum free surface elevation data. Model response is seen to lie somewhere between the responses for SAR and air photo data. An optimum for the upstream Manning's *n* is broadly identifiable in the same region of the parameter space as for the air photo data. This model response is mostly reconcilable with the response for the air photo data, as both are essentially comparing water surface elevations, and the results are replicated for the mid- and downstream reaches of the model.

The effectiveness with which each data source or combination of data sources constrains the parameter sets can be assessed by quantifying the uncertainty in these weighted parameters, for example by treating them as a probability distribution. One measure of the uncertainty in a distribution is the Shannon entropy, defined in this case by:

$$H = -\sum_{j=1}^{NR} L(\Theta_j \mid Y) \, \log_2 L(\Theta_j \mid Y) \tag{12}$$

where *j* is the model realisation index and *NR* represents the total number of model realisations (i.e. 500). This can be used to quantify the spread in a distribution, irrespective of its form (e.g. multimodal). The variance of a distribution can also be used to quantify uncertainty, which may give good results for unimodal distributions. The Shannon entropy for parameter distributions calibrated against the various data sets available is given in Table 3. Although the absolute differences between the entropy values are small, the relative differences reflect the differences in model calibrations observed above. For example, the parameter distribution entropy for the internal stage gauges is lower than for the downstream discharge, showing that the stage data have constrained the friction parameters more effectively. The entropy values can also be divided broadly into two classes: accurate stage measurements or their derivatives (from internal gauges or air photo data), with $H <= 8.9$, and the other data (SAR and downstream discharge).

The previous discussion has been limited to calibrations using a single set of observations. Attention is now turned to combining two or more data sets in the updating process of Eqn. 9. The methodology for the updating process implemented in this study is outlined in Table 4. A prior uniform distribution is conditioned firstly against data from

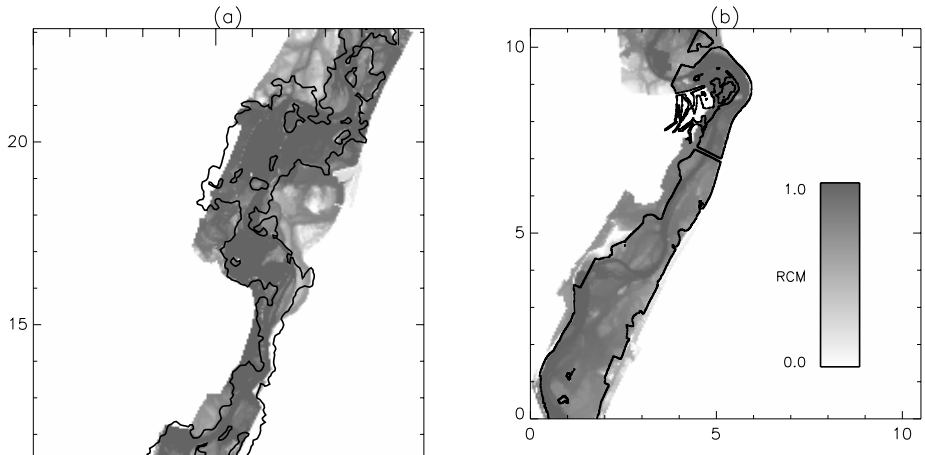

Fig. 6. *Relative confidence map of predicted inundation,* RCM_i^{flood}, *using the* $F^{\langle 2 \rangle}$ *performance measure for (6a) the full Monte Carlo ensemble conditioned on the satellite radar data and (6b) the air photo data. A contiguous flood shoreline can be extracted from each observational data source and is also shown.*

Table 3. Shannon entropy measure, *H*, based on the individual performance measures evaluated against available observational data sources. *H* is a maximum (= $\log_2 NR \approx 8.966$) for the case of the uniform prior distribution.

Observational data	*Shannon entropy measure, H*
Stage at internal gauge 1, Elsloo	8.871
Stage at internal gauge 2, Grevenbicht	8.900
Discharge at downstream boundary, Maaseik	8.921
Satellite radar-derived inundation shoreline	8.935
Air photo-derived inundation shoreline	8.890
Maximum free surface elevation survey	8.887

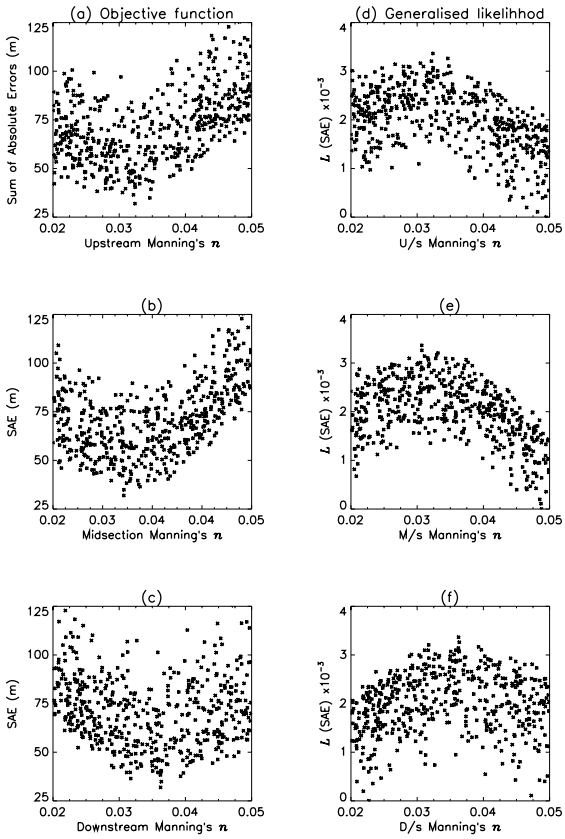

Fig. 7. *Dotty plots of objective function (7a–c) and generalised likelihood (7d–f) values using the sum of absolute errors to compare grid-scale model predictions with point maximum free surface elevation data surveyed along the floodplain. As goodness of fit to the available observational data increases, the value of the objective function and corresponding generalised likelihood decreases and increases respectively.*

the internal gauge at Elsloo, then updated by the Grevenbicht and finally by the Maaseik data. The updating process thus uses data moving sequentially down the reach. Figure 8 shows the results of this updating procedure in the form of histograms of the resulting parameter distributions. The first conditioning data set (Elsloo stage) affects only the upstream

Table 4. Combination sequence of generalised likelihood measures associated with individual evaluations of observational hydrometric data calculated using the form of Bayes equation expressed in Eqn. 9.

Prior generalised likelihood, $L_o(Q)$	*Current evaluation given observational data, Y*	*Posterior generalised likelihood, $L_p(Q \mid Y)$*
Uniform prior distribution	Stage at internal gauge 1, Elsloo	$L_p(Q \mid Y_1)$
$L_p(Q \mid Y_1)$	Stage at internal gauge 2, Grevenbicht	$L_p(Q \mid Y_{1,2})$
$L_p(Q \mid Y_{1,2})$	Discharge at downstream boundary, Maaseik	$L_p(Q \mid Y_{1,2,3})$

Manning's *n* significantly, the other distributions remaining largely uniform. This is a further example of the model's localised response to friction coefficients. Updating with the Grevenbicht data serves to reduce uncertainty in the mid-section Manning's *n*, but does not significantly affect the other parameters, apart from adding some noise. The addition of the Maaseik discharge data further reduces uncertainty for the upstream and mid-section, especially for the tails at high Manning's *n* values.

Generalised likelihoods previously evaluated for each source of observation data are then combined according to their generally prevailing availability for inundation model evaluation. For example, hydrometric data would be expected to be more commonly available for model calibration than SAR data, with air photo and ground-surveyed water elevations rarer still. The value of each data

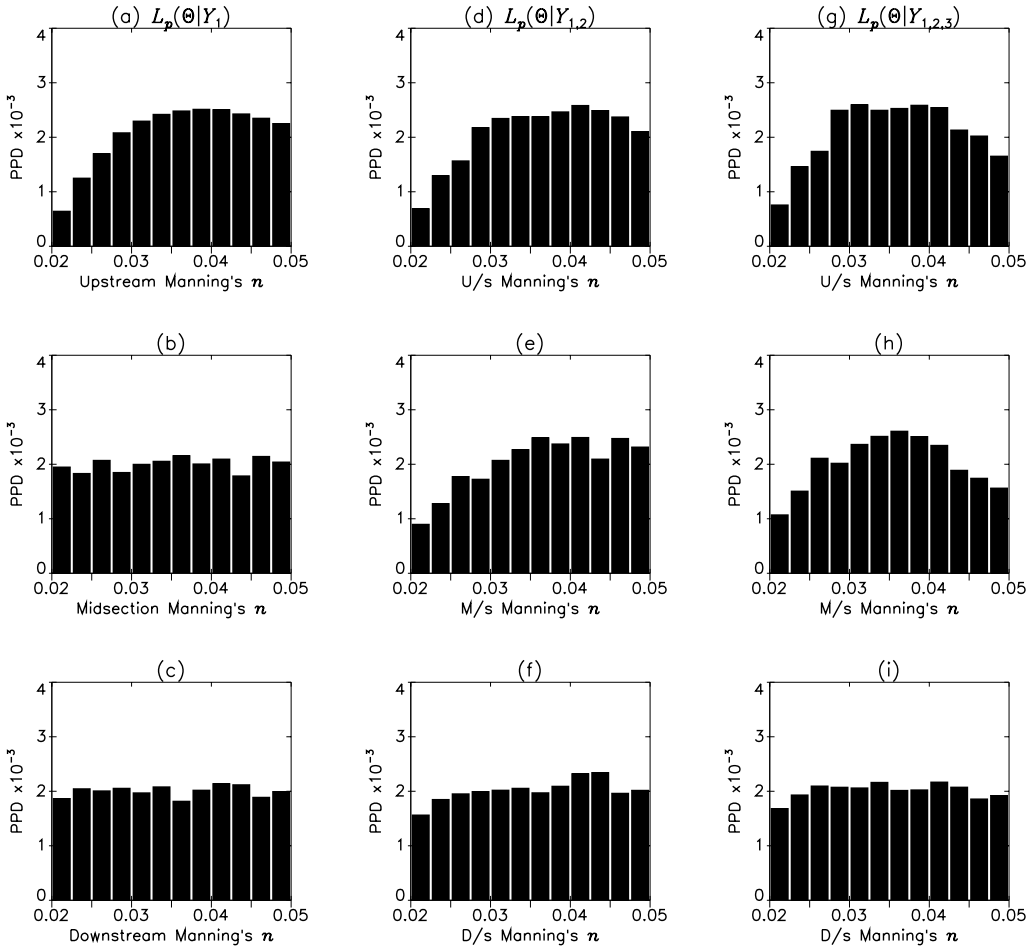

Fig. 8. A posteriori *parameter distributions (PPD) of selected LISFLOOD-FP parameters after conditioning on hydrometric observations evaluated using the Heteroscedastic Maximum Likelihood Estimator as the criterion of model performance. Figures 8a–c correspond to parameters conditioned on stage data recorded at internal gauge 1, Elsloo, only; 8d–f correspond to parameters conditioned on a combination of individual evaluations of model predictions of stage at Elsloo and internal gauge 2, Grevenbicht; and 8g–i to parameters conditioned on all available hydrometric observations (i.e. combination of individual evaluations of stage at the two internal gauges and discharge at Maaseik). The resulting distribution values are plotted as bars (remembering that the initial population was uniformly distributed).*

Table 5. Combination sequence of generalised likelihood measures associated with individual evaluations of all available observational data calculated using the form of Bayes equation expressed in Eqn. 9.

Combination	*Prior generalised likelihood,* $L_o(\Theta)$	*Current evaluation given observational data Y*	*Posterior generalised likelihood,* $L_p(\Theta \mid Y)$
1	Uniform prior distribution	$L_p(\Theta \mid Y_{1,2,3})$ (combined hydrometric generalised likelihood)	$L_p(\Theta \mid Y_{\text{Hyd}})$
2	$L_p(\Theta \mid Y_{\text{Hyd}})$	Satellite radar-derived inundation shoreline	$L_p(\Theta \mid Y_{\text{Hyd, SAR}})$
3	$L_p(\Theta \mid Y_{\text{Hyd, SAR}})$	Air photo-derived inundation shoreline	$L_p(\Theta \mid Y_{\text{Hyd, SAR, Air}})$
4	$L_p(\Theta \mid Y_{\text{Hyd, SAR, Air}})$	Maximum free surface elevation survey	$L_p(\Theta \mid Y_{\text{Hyd, SAR, Air, FSZ}})$

source at each stage of the updating process is considered in terms of: (1) conditioning the *a posteriori* parameter distributions (PPD) and the subsequent reduction in variance (i.e. increase in parameter identifiability) of these updated distributions; and (2) global uncertainty reduction according to the Shannon entropy measure. The combination sequence is described in Table 5.

Figure 9 shows the development of the *a posteriori* parameter distributions and reduction in variance at each stage of the updating process. The results show that the addition of SAR data to the distribution already conditioned against hydrometric observations only affects the upstream friction, and then only slightly. The addition of air photo data reduces uncertainty in the upstream Manning's *n* further, considerably more than the SAR data. The effect of the air photo data on the other friction coefficients is minimal, as would be expected given the model's evident localised response and the air photo coverage, which is limited to the upstream third of the reach. Inclusion of the surveyed water surface elevations further reduces uncertainty in all three Manning's *n* distributions, reflecting the global nature of the data. Again, it appears that accurate water surface elevations or their surrogates are more effective in reducing parameter uncertainty. Figure 10 shows the reduction in global Shannon entropy as a result of this process, which again reflects the varying worth of the observed data. Comparison of the plots in Figs. 9 and 10 show that the reduction in variance is not reflected well in the reduction in Shannon entropy, and in this case the Shannon entropy is not a particularly sensitive measure of the reduced parameter uncertainty evident in the parameter distributions.

Conclusions

This paper has sought to develop methods for assessing the relative utility of different observational data types for the calibration of distributed floodplain inundation models. For a 35-km reach of the River Meuse below the gauging station at Borgharen a benchmark data set for the 1995 flood event was assembled consisting of hydrometric data at the model boundaries and two internal measurement points, air photo and SAR images of flood extent and a post-event survey of 84 maximum water levels. Appropriate generalised likelihood measures were constructed for each data set and evaluated for 500 realisations of a simple two-dimensional hydraulic model, LISFLOOD-FP. The realisations differed in terms of the friction values assigned to each of three channel sub-reaches to give a three-dimensional parameter space. Channel friction values were sampled uniformly within this space and floodplain friction values were held constant. Comparison and combination of the various generalised likelihood measures were then conducted to quantify (1) the uncertainties associated with the simulation of various distributed hydraulic variables and (2) the reduction in uncertainty over effective parameter specification afforded by particular data sources and combinations of data sources.

The preceding discussion has highlighted a number of important points. First, whilst previous studies have highlighted the utility of flood extent data in constraining model predictions (e.g. Bates and De Roo, 2000), this paper has shown that the evaluation of internal predictions of stage also offer considerable potential for reducing uncertainty over effective parameter specification. The air photo data has similar properties as it provides an effective surrogate measure of water surface elevation when combined with an accurate DEM. Discharge data, on the other hand, is much less effective because of the essentially mass conservative nature of the model and the flow dynamics of this particular event.

Second, the analysis has shown that model response to the calibration process is different for high and low flows, thereby lending support to the conclusions of Romanowicz

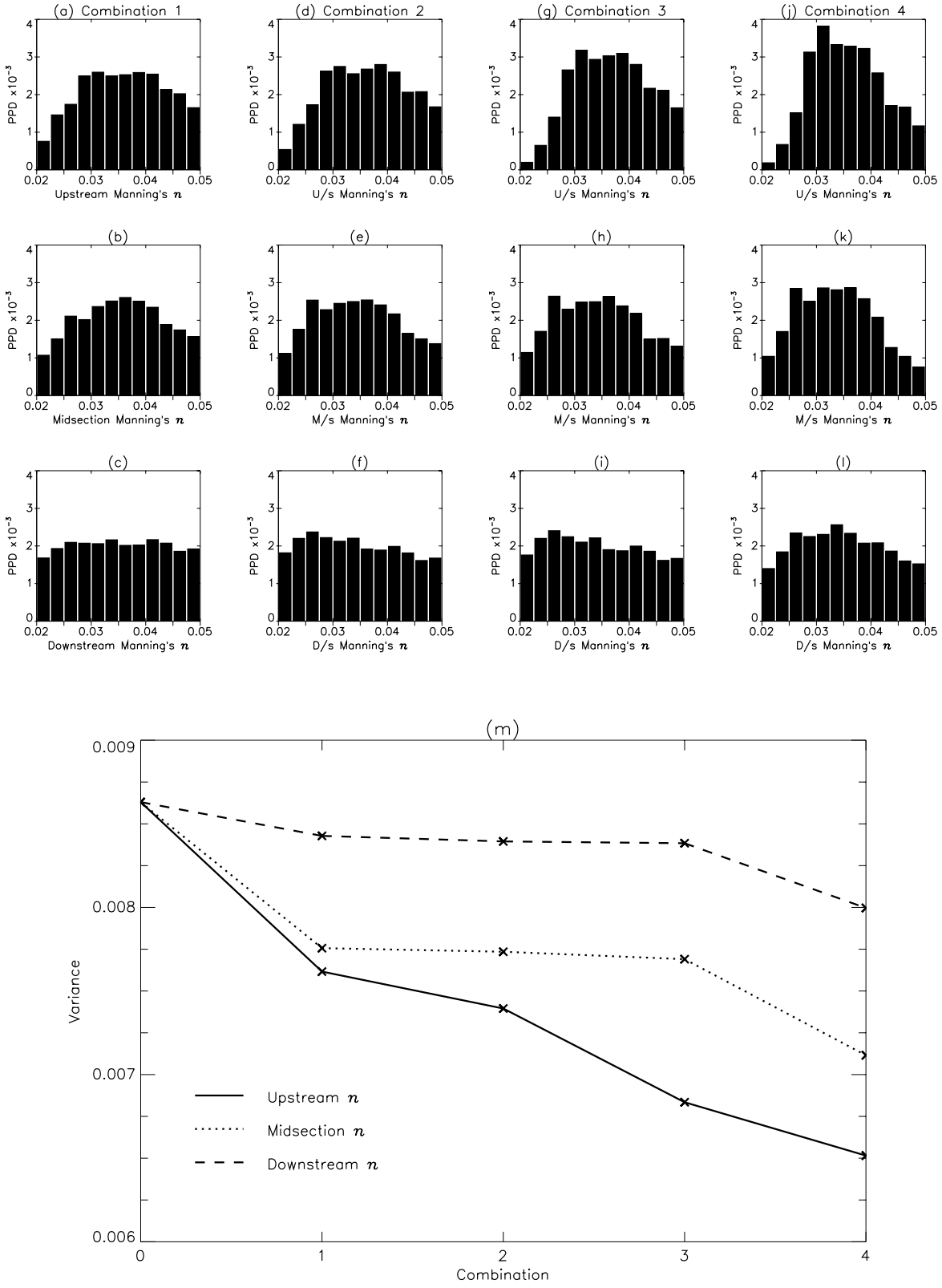

Fig. 9. A posteriori *parameter distributions (9a–l) of selected LISFLOOD-FP parameters after conditioning on a combination of individual evaluations of all available observational data. Individual model performance measures have been combined using a form of the Bayes equation (Eqn. 9) according to the sequence in Table 4. The reduction in variance from an initial uniform distribution (Combination 0) at each stage in the process of parameter conditioning is shown in Fig. 9m. In this instance, variance may be used as an indice of parameter uncertainty reduction because PPDs are unimodal throughout.*

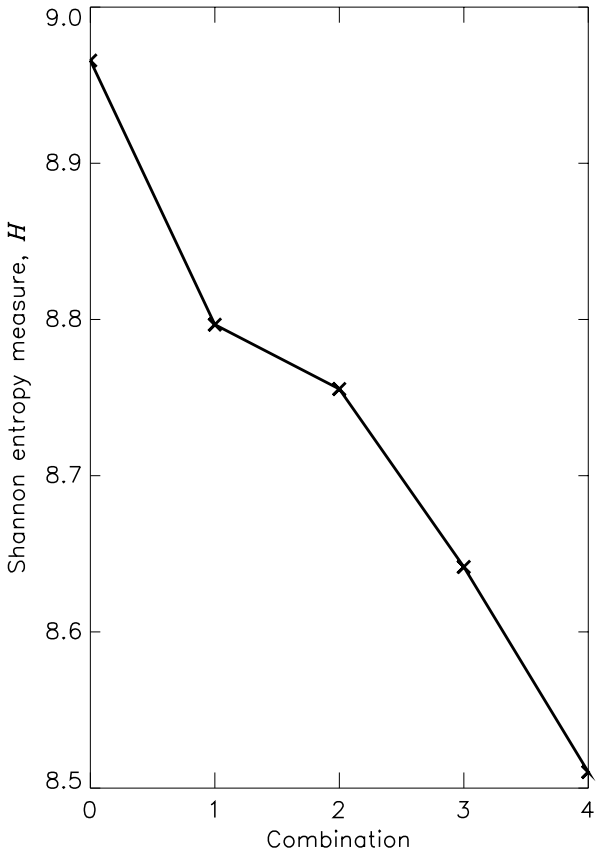

Fig. 10. *Change in Shannon Entropy measure, H, based on the a posteriori generalised likelihoods for the sequence of combining individual performance measures in Table 4. H is a maximum (= $\log_2 NR \approx 8.966$) when all the realisations are equally likely (the case for a uniform prior distribution, Combination 0) and a minimum of 0 when one single realisation has a generalised likelihood of 1 and all others have a generalised likelihood of zero (Beven and Binley, 1992).*

and Beven (2003) who found marked differences in effective friction parameters for events of different magnitude. Here, it has been demonstrated that a similar effect can occur on an intra- as well as an inter-event scale. However, the present study also demonstrates that model response to the calibration process depends not only on the (subjective) choice of generalised likelihood measure, but also on the errors in data used, with, for example, the air photo data being much more effective than the coincident SAR image in discriminating between parameter sets. This somewhat contradicts the results of Horritt and Bates (2002) who found no difference in ability to constrain uncertainty in LISFLODD-FP between SAR flood extent images of markedly different quality. This may relate to the manner in which the SAR data were processed for this application and the breakdown of the snake algorithm for urban areas on the floodplain in this case. Differing response to data may also be considered an example of model overfitting to uncertain data and highlights the need not to treat observations as if they were error-free. This compares with the study of Horritt and Bates (2002) where errors in the observations were taken into account in an uncertain classification procedure that led to more consistent results between data sets with different accuracies.

Third, the localised model response is clearly demonstrated by the RCM_i^{flood} maps, which show a marked increase in uncertainty away from the conditioning observation. Such behaviour is likely typical of distributed models and reinforces the need to map uncertainties back into real space in order to understand this spatial response (c.f. Aronica *et al*., 2002).

Lastly, analysis of the ability of different data combinations to reduce the entropy of the simulation ensemble reflects the properties of the calibration process discussed above, although the Shannon entropy is not found to be a particularly sensitive measure in this instance. An updating process using increasing amounts of information has been shown here to lead to a monotonic decrease in parameter uncertainty, although in many environmental applications this may not be the case due to irreconcilable differences between data sets. Data consistency here stems from the fact that all observations are of water level or approximations to water level, such as flood extent (a strong surrogate) and discharge (a weak surrogate).

The steady decrease in parameter uncertainty also results from the choice of generalised likelihood measures, which give a large proportion of simulations with high generalised likelihood values. This would not be the case if arbitrary rejection criteria (e.g. a thresholding of a continuous numerical performance measure to reject 'unphysical' simulations) or more discriminatory objective functions were used. Such a process might lead to decreased parameter uncertainty for a single observed data set, but lead to parameter distributions so constrained as to be incompatible between data sets. Rejection criteria may be used in some cases to reduce parameter uncertainty, but this approach would require careful application and justification, for example by adopting physically meaningful criteria for rejection. Put another way, in this study consistency of the data sets has been ensured by rejecting the rejection approach.

Future research in this area should examine the significance of rejection in models conditioned in an uncertainty analysis framework to define criteria that are physically sensible and allow the combination of multiple data sets in a consistent fashion. Additional studies could also examine the significance of subjective assumptions made in applying the GLUE method to environmental

problems. In particular, this should examine the importance of initial parameter ranges in cases, such as river inundation, which are more linear than those typically found in rainfall–runoff modelling for which the GLUE methodology was originally developed.

Acknowledgements

The authors wish to thank Rolf Van Deer Veen and the Dutch Ministry of Public Works and Water Resources (RWS Directie Limburg and RIZA) for providing access to the air photo inundation extent data, DEM, surveyed channel cross-sections and flow discharge and stage data at Borgharen, Elsloo and Grevenbicht. The Belgian Water Authority is also thanked for allowing access to gauging station records from Maaseik. Anna Ghelli of ECMWF and Jean-Philippe Renaud of the University of Bristol are thanked for providing clarification on the proper use and interpretation of various statistical measures. Keith Beven, Hoshin Gupta and two anonymous reviewers provided stimulating discussion and insightful comments on early drafts of the paper which improved the final version significantly. Finally, this work was supported financially by the EU Framework 5 Project EVG1-CT-1999-00011, EFFS: A European Flood Forecasting System (http://effs.wldelft.nl/).

References

Aronica, G., Bates, P.D. and Horritt, M.S., 2002. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrol. Process.* **16**, 2001–2016.

Aronica, G., Hankin, B.G. and Beven, K.J., 1998. Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data. *Adv. Water Resour.,* **22**, 349–365.

Bates, P.D. and De Roo, A.P.J., 2000. A simple raster-based model for flood inundation simulation. *J. Hydrol.,* **236**, 54–77.

Bates, P.D., Horritt, M.S, Aronica, G. and Beven, K.J., in press. Bayesian updating of flood inundation likelihoods conditioned on flood extent data. *Hydrol. Process.*

Bates, P.D., Stewart, M.D., Siggers, G.B., Smith, C.N., Hervouet, J.-M. and Sellin, R.H.J., 1998. Internal and external validation of a two dimensional finite element model for river flood simulation. *Proc. Inst. Civil Eng.-Water Mar.* **130**, 127–141.

Bechteler, W., Hartmann, S. and Otto, A.J., 1994. Coupling of 2D and 1D models and integration into Geographic Information Systems (GIS). In: *2nd International Conference on River Flood Hydraulics*, W.R. White and J. Watts (Eds.), Wiley, Chichester, UK. 155–166.

Becker, A. and Grünewald, U., 2003. Flood risk in Central Europe. *Science,* **300**, 1099.

Beven, K.J., 1989. Changing ideas in hydrology - The case of physically-based models. *J. Hydrol.,* **105**, 157–172.

Beven, K.J., 1996. A discussion in distributed modelling. In: *Distributed Hydrological Modelling*, J.-C. Refsgaard and M.B. Abbott (Eds), Kluwer Academic Publishers, The Netherlands. 289–295.

Beven, K.J., 2000. Uniqueness of place and process representations in hydrological modelling. *Hydrol. Earth Syst. Sci.,* **4**, 203–213.

Beven, K.J., 2001a. How far can we go in distributed hydrological modelling? *Hydrol. Earth Syst. Sci.,* **5**, 1–12.

Beven, K.J., 2001b. *Rainfall-Runoff Modelling: The Primer*. Wiley, Chichester, UK. 372pp.

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

Beven, K.J. and Kirkby, M.J., 1979. A physically based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.,* **24**, 43–69.

Chow, V.T., 1959. *Open Channel Hydraulics*. McGraw-Hill, USA. 680pp.

Christiaens, K. and Feyen, J., 2002. Constraining soil hydraulic parameter and output uncertainty of the distributed hydrological MIKE SHE model using the GLUE framework. *Hydrol. Process.* **16**, 373–391.

Clarke, R.T., 1973. A review of some of the mathematical models used in hydrology, with observation on their calibration and use. *J. Hydrol.,* **19**, 1–20.

Collier, C.G., 2003. On the lessons learnt from EU FPV flood-related projects. *International Conference on Advances in Flood Forecasting in Europe*, Rotterdam, The Netherlands, March 2003.

Cunge, J.A., Holly, F.M. and Verwey, A., 1980. *Practical Aspects of Computational River Hydraulics*. Pitman Publishing, UK. 420pp.

De Roo, A.P.J., Van Der Knijff, J., Horritt, M.S., Schmuck, G. and De Jong, S., 1999. Assessing flood damages of the Oder flood and the Meuse flood. *2nd International Symposium on Operationalization of Remote Sensing*, Enschede, The Netherlands, August 1999.

Estrela, T. and Quintas, L., 1994. Use of a GIS in the modelling of flows on floodplains. In: *2nd International Conference on River Flood Hydraulics*, W.R. White and J. Watts (Eds.). Wiley, Chichester, UK. 177–190.

Franks, S.W., Gineste, P., Beven, K.J. and Merot, P., 1998. On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas in the calibration process. *Water Resour. Res.,* **34**, 787–797.

Green, I.R.A. and Stephenson, D., 1986. Criteria for comparison of single event models. *Hydrol. Sci. J.,* **31**, 395–411.

Gupta, H.V., Sorooshian, S. and Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.,* **34**, 751–763.

Horritt, M.S., 1999. A statistical active contour model for SAR image segmentation. *Image Vis. Comput.,* **17**, 213–224.

Horritt, M.S. and Bates, P.D., 2001. Predicting floodplain inundation: raster-based modelling versus the finite-element approach. *Hydrol. Process.,* **15**, 825–842.

Horritt, M.S. and Bates, P.D., 2002. Evaluation of 1-D and 2-D numerical models for predicting river flood inundation. *J. Hydrol.,* **268**, 87–99.

Lamb, R., Beven, K.J. and Myrabø, S., 1998. Use of spatially-distributed water table observations to constrain uncertainty in a rainfall-runoff model. *Adv. Water Resour.,* **22**, 305–317.

Lane, S.N., Bradbrook, K.F., Richards, K.S., Biron, P.M. and Roy, A.G., 1999. The application of computational fluid dynamics to natural river channels: three-dimensional versus two-dimensional approaches. *Geomorphology,* **29**, 1–20.

Romanowicz, R.J. and Beven, K.J., 2003. Estimation of flood inundation probabilities as conditioned on event inundation maps. *Water Resour. Res.,* **39**, 1061–1073.

Romanowicz, R.J., Beven, K.J. and Tawn, J.A., 1994. Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian Approach. In: *Statistics for the Environment (2), Water Related Issues*, V. Barnett and F. Turkman (Eds.), Wiley, Chichester, UK. 297–318.

Romanowicz, R.J., Beven, K.J. and Tawn, J., 1996. Bayesian calibration of flood inundation models.In: *Floodplain Processes*, M.G. Anderson, D.E. Walling and P.D. Bates (Eds.), Wiley, Chichester, UK. 333–360.

Samuels, P.G., 2003. Lessons from previous EC research on floods. *International Conference on Advances in Flood Forecasting in Europe*, Rotterdam, The Netherlands, March 2003.

Sorooshian, S. and Dracup, J.A., 1980. Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resour. Res.,* **16**, 430–442.

Wagener, T., Lees, M.J. and Wheater, H.S., 2001. A toolkit for the development and application of parsimonious hydrological models. In: *Mathematical Models of Large Watershed Hydrology – Volume 1*, V.P. Singh and D. Fervert (Eds.). Water Resources Publishers, USA. 87–136.

Werner, M.G.F., 2001. Impact of grid size in GIS based flood extent mapping using a 1D flow model. *Phys. Chem. Earth Pt B-Hydrol. Oceans Atmos.,* **26**, 517–522.

Werner, M.G.F., 2002a. Uncertainties in floodplain inundation modelling. *Proc. Fifth Int. Conf. Hydroinformatics*, Cardiff, UK, July 2002.

Werner, M.G.F., 2002b. The value of floodplain land-use data in constraining uncertainties in flood extent estimation. *Proc. XIVth Int. Conf. Computational Methods in Water Resources (CMWR XIV)*, Delft, The Netherlands, June 2002.

Wheater, H.S., 2002. Progress and prospects for fluvial flood modelling. *Phil. Trans. Roy. Soc. Lond. A,* **360**, 1409–1431.