



available at www.sciencedirect.com



journal homepage: www.elsevier.com/locate/jhydrol



Grasping the unavoidable subjectivity in calibration of flood inundation models: A vulnerability weighted approach

Florian Pappenberger^{a,d,*}, Keith Beven^a, Kevin Frodsham^b,
Renata Romanowicz^a, Patrick Matgen^c

^a Environmental Science/Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

^b JBA Consulting, The Brew House, Wilderspool Park, Greenall's Avenue, Warrington WA4 6HL, UK

^c Cellule de Recherche en Environnement et Biotechnologies, Centre de Recherche Public-Gabriel Lippmann, Luxembourg

^d European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK

Received 10 March 2006; received in revised form 14 August 2006; accepted 30 August 2006

KEYWORDS

Flood inundation model;
LISFLOOD-FP;
GLUE;
Raster map comparison;
Utility function;
Flood risk;
Flood hazard

Summary Quantitative modeling of risk and hazard from flooding involves decisions regarding the choice of model and goal of the modeling exercise, expressed by some measure of performance. This paper shows how the subjectivity in the choices of performance measures and observation sets used for model calibration inevitably results in variability in the estimation of flood hazard. We compare the predictions of a 2D flood inundation model obtained using different global and local evaluation criteria. It is shown that traditional area averaging performance measures are inadequate in the face of model imperfection, especially when such models are calibrated for flood hazard studies. In this study we include flood risk weighting into the performance measure of the model. This allows us to calibrate the model to places that are important, e.g. location of houses. The quantification of the importance of places requires the necessity of engaging stakeholders into the model calibration process.

© 2006 Elsevier B.V. All rights reserved.

Introduction

Accurate quantification of flood risk is important for forecasting, planning and in many decision making processes. The term 'risk' is interpreted in many different ways in

the context of natural hazards (for a comprehensive summary, see [Kelman, 2002](#)). Hazard in this study will be defined by a characteristic of or phenomenon from the natural environment which has the potential for causing damage to society, and will here be restricted to flood hazards such as water depth and velocity ([Kelman, 2002](#)). Vulnerability refers to a characteristic of society which indicates the potential for damage to occur as a result of hazards ([Kelman, 2002](#)). This paper will concentrate on

* Corresponding author.

E-mail address: florian.pappenberger@ecmwf.int (F. Pappenberger).

making estimates of flood hazard as the probability of inundation in the face of inadequate observational data and models that do not perform well everywhere.

The estimate of flood hazard involves estimating two types of probability: the probability of exceedence of an event of a given magnitude, and the probability of inundation (and consequent damage) during any particular event. The first is subject to considerable uncertainty (see, for example, Blazkova and Beven, 2004; Cameron et al., 2000) but in this paper, we focus on estimating the uncertainties associated with the hazard for any particular event. Historical records of flood events are scarce and thus usually flood inundation models are used to determine the hazard parameters such as water depth and flow velocity. 'Confidence' in the model outputs is in many cases established through calibration of the model on past flood events. However, in many situations models have to be used where no historical data are available or where there is a need to extrapolate to higher discharges.

In these cases, it is commonly assumed that all the approximations made are correct, or errors are negligible, and that the methodology and model are a valid and reasonable representation of the real physical system. This may not necessarily be the case in the face of different sources of uncertainty (see, for example Romanowicz and Beven, 2003). Current models are imperfect as they cannot reproduce satisfactorily measured data in every local location in time and space (for example Pappenberger et al., 2006b). Admittedly, many of those models have been designed for large scale applications, however, their results are also used on a local scale to determine flood hazard/risk. It then follows that model calibration with global performance measures will not necessarily provide accurate estimates of inundation probability everywhere. In what follows, the possibility of using local performance measures based on vulnerability is explored as one possible response to model inadequacy.

Past studies have shown that the choice of performance measure used for a flood inundation model can significantly influence flood hazard map predictions (Hunter, 2006). Here, we make use of that dependency to propose the use of performance measures related directly to the vulnerability of locations on the flood plain. The aim is to attach the most importance in assessing model predictions to those places that are of most interest. This then might give rise to an issue of model overfitting in order to get better results at locations of interest, with the danger that extrapolation to other conditions, even for those locations, may be less robust. This is, in part, mitigated by carrying out the calibration within the framework of the generalised likelihood uncertainty estimation (GLUE) methodology (Beven, 2006; Beven and Binley, 1992). The approach will be illustrated with examples from the 2003 flood on the river Alzette (Grand Duchy of Luxembourg) and the LISFLOOD-FP model.

Introducing the challenges

If a flood model could reproduce flood outlines without error, then the challenges in this paper would not exist. However, for large scale modelling exercises to estimate flood hazard or risk, this still seems to be a long way off. Currently, both models and the observational data available to evaluate model predictions are subject to significant error. This

then gives rise to an interesting problem: a model that gives a good overall fit to the available data may not give locally good results in locations that are of particular interest to flood planners and risk assessors. For such purposes, model predictions may well be looked at very closely at the local level, particularly in areas where the hazard may be high. We focus this paper on two questions: is global performance an adequate measure for the evaluation of local hazard and how can risk be included into the calibration process. Both points will be explained in detail in what follows.

Question 1: Is global performance the right measure to assess local hazard?

To date 2D inundation models have always been evaluated against inundation extent using global (average) performance measures that are obtained by averaging spatial performance over the entire flood domain (for example Aronica et al., 1998; Bates and De Roo, 2000). In the case of flood inundation models the calibration domain is usually constrained by data availability for both input data (topography, flood plain infrastructure, upstream discharges, effective roughness estimates) and calibration data such as recorded inundation extent for past events. Reliable inundation data can be exceptionally scarce for some of these problems, in terms of both events and spatial and temporal coverage during an event. The availability of both types of data, therefore, may give rise to bias and uncertainty in the predictions for high risk areas of interest.

The computation of global performance ignores the fact that flood inundation is a local as well as global phenomenon. Past experience in calibrating inundation models suggests that parameter sets which result in an acceptable model performance derived for global models may be entirely different for models calibrated on local performance. Ideally this would not be the case, a model based on a good representation of the physics, with good input data and adequate calibration data should, it would be hoped, give good results everywhere (Pappenberger et al., 2006b). This is not the case in many, if not all, applications of the current generation of 1D and 2D inundation models. Hunter (2006) has argued that local failure of globally calibrated models should be considered as part of the model precision and that calibration should aim to achieve a balance of bias and variance of the performance measure.

Such an approach, however, might still result in local error. The results of flood inundation models are usually used at a local scale (e.g. to determine the risk/hazard for a new development) and thus from the arguments given above, a global calibration alone will be inadequate. We will demonstrate this by using three different ways to compute model performance: a global performance measure over the entire domain; sub-domain performance measures (L_S), which will concentrate on sub-domains; and point performance measures which express model performance only in respect to the correctness to a certain cell (L_P).

Question 2: How can 'risk' be included in the evaluation process?

In most studies of predicting flood inundation it has been assumed that average global performance is the best way to

calibrate models. But if this might lead to error in predicting inundation in high risk areas, it leads to the question of whether local estimation of hazard might be improved by the use of local performance measures, with the danger, as in any calibration exercise, of overfitting the model to particular circumstances which might then lead to prediction error in other circumstances. The danger is always greatest when calibration data are scarce and model structure error or input data errors are significant. Some compromise between matching local detail and the danger of overfitting will always be needed in model calibration. Here, we show that global calibration can lead to error in predicting inundation in high risk areas and look at ways of incorporating risk into the calibration process.

Many ways to compute risk have been postulated (Alexander, 1991; European Environment Agency, 2006; Granger et al., 1999; Kelman, 2002; Smith, 2001). It has to be pointed out that a full risk assessment should include environmental and social impacts of floods (Bouma et al., 2005), pricing methods (Jonkman et al., 2003; Turner et al., 2001), preparedness (Thieken et al., 2005) and many other factors (for a summary see Thieken et al., 2005). In the same way that model performance functions influence the estimation of flood hazard, so different methods of quantifying vulnerability have similar impact on flood risk maps (an extensive comparison is given in Jonkman et al., 2003). In particular, the uncertainty in quantifying risk components such as cost, can be influential (Merz et al., 2004; Soetanto and Proverbs, 2004). In this paper, we will neglect the uncertainty in quantifying risk, or social vulnerability (Wu et al., 2002) and use only a simple assessment of relative risk for different locations on the flood plain. However, the methodology which is presented in this paper is general and the neglected factors could be easily included.

Methodologies

Inundation model and study region

Data from the River Alzette in Luxembourg for a medium scale flood event in January 2003 will be used. Discharge

potentially peaked at around $63 \text{ m}^3\text{s}^{-1}$ (see below) and the extent of inundation were recorded by the Synthetic Aperture Radar (SAR) sensors on board ENVISAT and ERS-2 satellites at a time close to this estimated peak discharge, which will be used for model calibration. The event has been modelled with the 2D raster based LISFLOOD-FP model (Bates and De Roo, 2000; Hunter et al., 2005). A detailed description of model set-up and implementation is given in Pappenberger et al. (2006b).

Calibration strategy

The model has been evaluated within the Monte Carlo based Generalized Likelihood Uncertainty Estimation (GLUE) framework. This methodology recognises that many different combinations of effective model parameters can lead to results which are acceptable representations of the available observations. In GLUE the model is run with multiple parameter sets and the performance for each evaluation computed (concept of equifinality Beven, 2006; Beven and Binley, 1992). All effective parameter sets which have an acceptable model performance are retained for further analysis (for an implementation with flood models see Pappenberger et al., 2005; Romanowicz and Beven, 2003; Romanowicz et al., 1996). For this analysis prior distributions from which the effective parameters are sampled have to be allocated (for a summary see Table 1 and for more details see Pappenberger et al., 2006b). Channel and floodplain friction have been assumed as constant over the reach. The downstream boundary condition was approximated by uniform flow and therefore required the additional specification of a roughness value. Channel widths along the reach were allowed to vary by $\pm 10\%$ from the values obtained from the channel surveys. In order to replicate the uncertainty that is believed to be inherent in using stream hydrographs as model inputs (Pappenberger et al., 2006b), a set of 20 different input hydrographs were prepared that were consistent with the available stage data via rating curves. The depth and slope of the channel bed have been derived from 73 surveyed cross-sections, which have been included in the LISFLOOD model. To allow

Table 1 Parameters included in the uncertainty analysis and ranges sampled

Parameter	Sampling range	Distribution	Additional description
Floodplain roughness	0.05–0.3	Log-normal	
Channel roughness	0.01–0.2	Log-normal	Channel friction always lower than Floodplain friction
Effective river width	$\pm 10\%$	Log-normal	
Outflow roughness	0.01–0.4	Log-normal	
Inflow magnitude	1–20	Uniform	A set of 20 contrasting hydrographs, that were consistent with the available stage data via rating curves have been prepared and used (after Pappenberger et al., 2006b)
Initial error on first cross-section	$\pm 15 \text{ cm}$	Uniform	For each model simulation an error has been assigned to the first cross-section
Standard deviation for cross-section error	0.01–0.1	Uniform	The error of the next cross-section has been derived from a normal distribution with the error of the previous cross-section as mean. A negative slope has been enforced by re-sampling until the condition has been met

for error in these cross-sectional data in representing the effective form of the channel in each reach, for each model simulation an error of the lowest elevation has been assigned to the first cross-section. The error of the next cross-section has been derived from a normal distribution with the error of the previous cross-section as mean. A positive slope has been enforced by re-sampling until the condition has been met. Thus only two parameters needed to be specified: the initial error and a variance. A series of ~28,000 simulations was performed using parameter values chosen at random from the designated ranges and results were presented as water depth maps at the time of satellite overpass.

Creating a spatial distributed flood hazard and flood risk map

The results of the uncertainty analysis can be used to create flood hazard maps. In this paper, we illustrate our arguments with a simple assumption: that the spatial predictions of a flood inundation model are transformed into a pattern of wet and dry cells and that this pattern is used for model evaluation. We accept that this may not be the best way to constrain flood inundation models (Werner et al., 2005). It would be better to use flood depths, particularly for events in which the flood inundates the valley floor or reaches flood defences. Flood depths are also important in assessing local vulnerability and hazard. However, only flood extent data were available for this particular example and, although this is generally related to depth of inundation, the transformation to the depth variable would involve significant interpretation and interpolation errors. Our arguments are illustrated on simple wet/dry patterns, but the methodology is readily applicable to calibrations based on flood depths or other variables if the calibration data were available for another application.

The possibility of a cell being hazardous (wet) can be computed by (see also Aronica et al., 2002):

$$p_{ij}^{flood} = \frac{\sum_{m=1}^n s_{ij,m} L_m}{\sum_{m=1}^n L_m} \quad (1)$$

where i, j is cell location, n is number of behavioural model simulations, s is binary state of the cell (wet = 1, dry = 0), P is possibility of cell being wet given assumptions, and L is the likelihood of simulation m .

It can be seen that different ways of computing the likelihoods, L_m , will lead to different flood hazard maps. In Hunter (2006) and Pappenberger et al. (2006b) this is demonstrated by comparing multiple performance measures.

Computation of the global, sub-domain and local performance

The *global performance* (L_G) is based on the fuzzy methodology introduced by Pappenberger et al. (2006b), which recognizes the uncertainty in the observations as well as the

model results. This method is based on the fuzzy inundation measure briefly described in Appendix A. For the *sub-domain performance* (L_S) the flood plain has been divided into seven tiles of 1 km² each (see Fig. 1). For each of these tiles the same performance measure as for the global computation has been used. The *point performance* (L_P) measure was devised based upon evaluating the shortest distance from a specified target cell to cells of a given inundation probability (Fig. 1). This effectively enables model performance to be evaluated with respect to a single cell as the distance from a critical target structure such as a hospital or emergency control centre to any number of possible shoreline scenarios (see Appendix B).

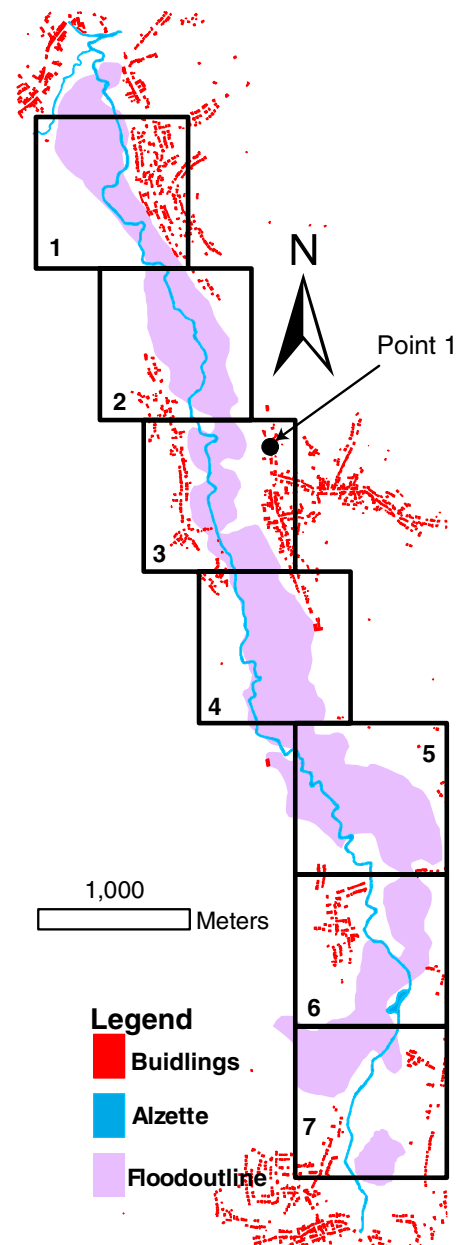


Figure 1 Outline of the Alzette catchment. The tiles indicate the area on which each subdomain performance has been calibrated. Point 1 shows the location of the point performance measure.

Results

Global, sub-domain and local performance

In Fig. 2, the results of the global, sub-domain and local performance have been plotted against each other. Each of the dots stands for one model simulation resulting from a randomly chosen input data series, channel geometry and roughness parameter values. It is apparent that the different measures do not always exhibit a positive relationship, in other words, a model run can perform well with one performance measure and under-perform with a different measure. Some of the more local performance measures (L_S and L_P) show a clear step response indicating that the water level reaches certain threshold values. For example, an entire area gets flooded at once as soon as a certain threshold is exceeded thus changing the performance measure significantly. This is particularly the case for the local point performance measure (L_P), which also suffers from the fact that 'steps' are always 50 m (cell size). These effects are lost as soon as the overall performance is computed and the distribution of the performance measures is without step responses.

A positive linear relationship between different performance measures would be expected if a high global performance always led to a high local performance. This is not the case for any of the graphs shown and indicates a significant difference between the local and global evaluation

criteria. For example, the comparison between $L(S1)$ and $L(S3)$ show that dots in the left top corner have a low local performance at square 1 but a high local performance in square 3 and vice versa in the bottom right corner. The histograms in the graph indicate the distribution of the performance measures. It can be observed that the global performance measure nearly exhibits a normal type of distribution whereas the local measures can be skewed. For example, the point measure has a large frequency at high performance values (Fig. 2, bottom right histogram). This means that this point performance measure has only a limited discriminatory power for predictions at that point and strongly reflects that this point is predominantly dry. However, the tail of this distribution specifies the residual risk and may be especially important when a risk averse policy is adopted for that location. The histograms indicate that areas in the floodplain exist which are predicted much better on average (for example, $L(S6)$) or much worse on average (for example, $L(S1)$). This difference is a combination of the amount of flooding and topography in each area.

These differences can also be seen if the model factors (parameters) which have been varied are plotted against the individual performances. Factor identifiability as well as factor sensitivity depends on the performance measure chosen (Hunter, 2006; Pappenberger et al., 2006b, which can be important if management decisions are based on these simulations.

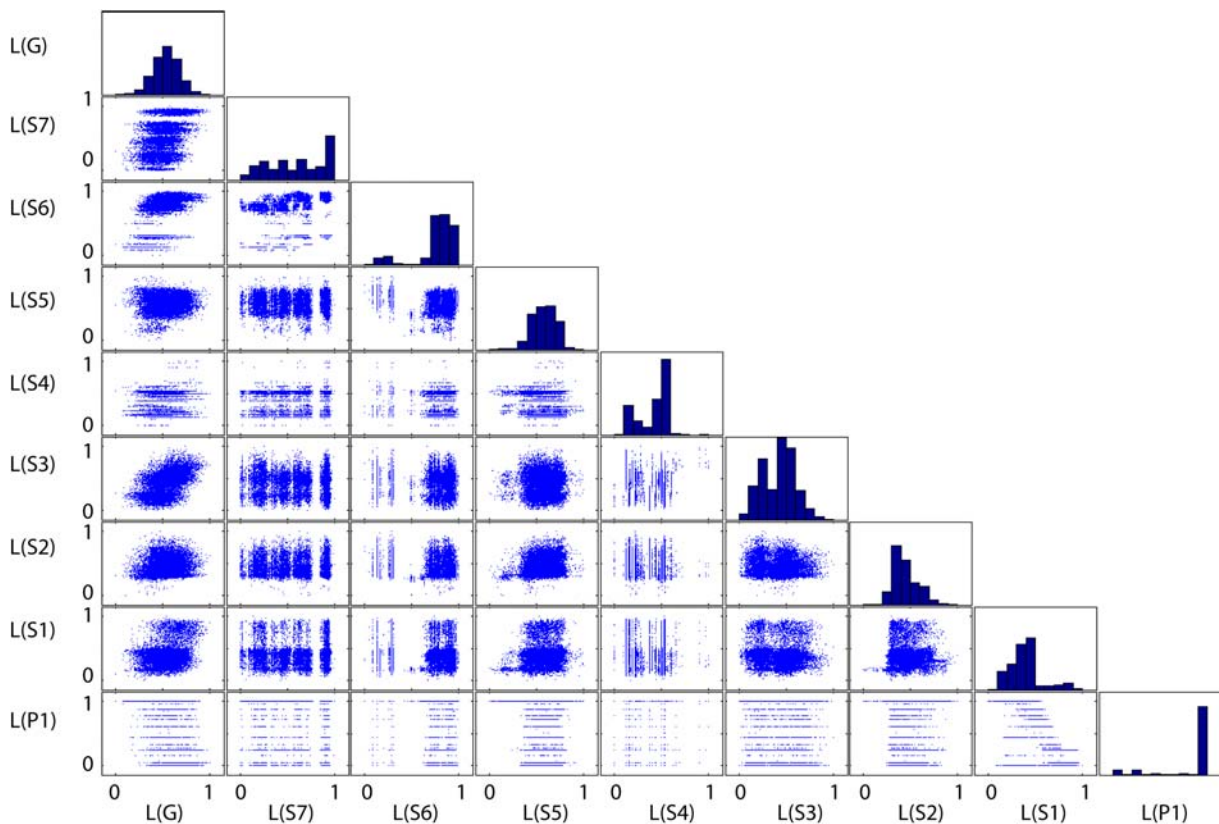


Figure 2 The performances of the different evaluation criteria are plotted against each other. Each dot represents one model simulation. $L(G)$ is the global measure and $L(S)$ the performance of each subdomain. $L(P1)$ is the performance measure for the location 1. The histogram of each performance measure is also presented. All performance measures have been normalized between 0 and 1 to allow for better comparison.

In Fig. 3, the relationship between global and local measures is investigated in more detail. The figure displays a histogram with the frequency distribution of the global performance. The lighter colour indicates the proportion (left ordinate) of models, which *underperform at all subdomain* evaluation criteria, namely $L(S1)$ to $L(S7)$. Underperformance is here defined as being in the lower one percentile of performance measures of one or more local criteria. The dotted line represents the percentage (right ordinate) of the underperforming models to the total number of models in this class. It is apparent that a model can perform well on a global performance and still fail on a sub-domain performance measure. Similar conclusions have been drawn by Pappenberger et al. (2006b) and Freer et al. (2003) for a one-dimensional flood inundation and a rainfall-

runoff model, respectively. Applying a multi-criteria evaluation in this application, if all parameter sets that are underperforming at the sub-domain level by this (very relaxed) definition are rejected, no models are retained as acceptable. The implication is that if the global performance measure is used as a basis for mapping flood hazard and flood risk, locally the map might be quite wrong.

In Table 2, the flood inundation possibility for three representative cells (one in each row) computed with all the measures introduced above is given for the calibration event of January 2003. The results of these three cells are consistent with many other cells. The table also includes an aggregation of the subregions by taking the maximum (L_{max}), the minimum (L_{min}), the product (L_{prod}) and the sum (L_{sum}) of all subregion likelihoods. This means that the possibility of

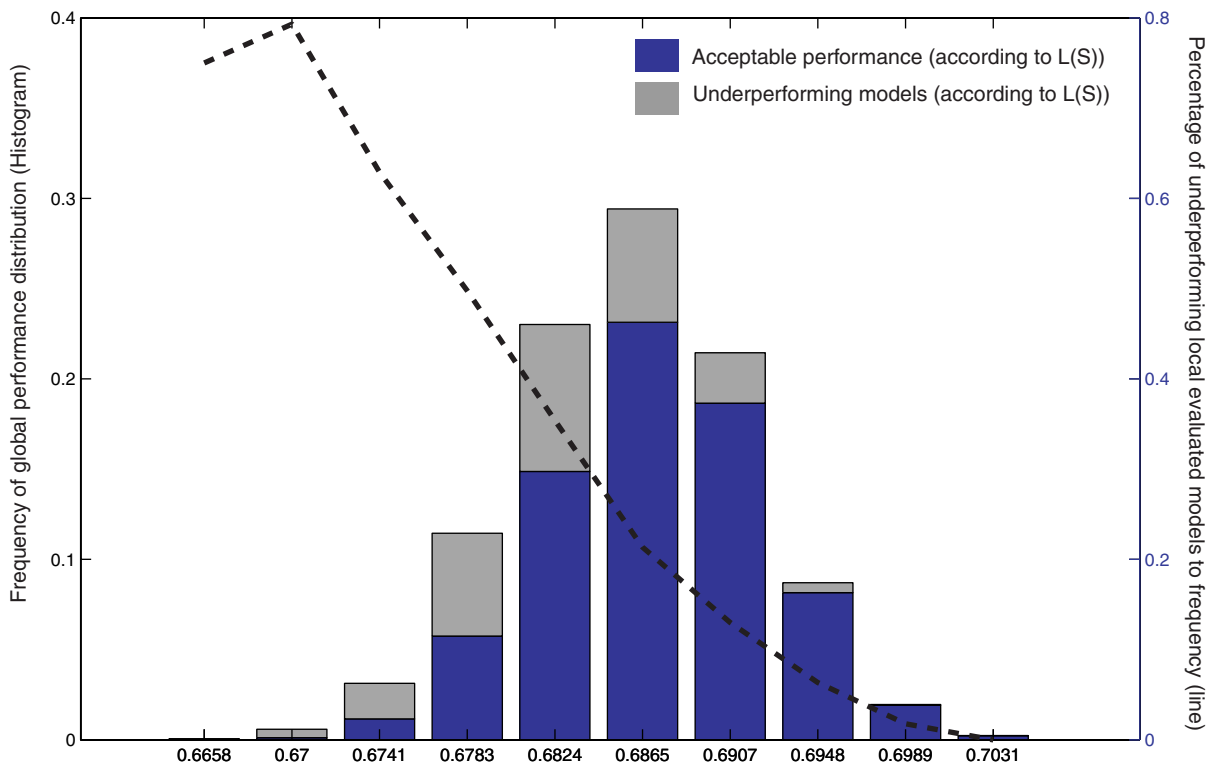


Figure 3 Histogram with the frequency distribution of the global performance (abscissa). The lighter colour indicates the proportion (left ordinate) of models, which under perform at all local evaluation criteria. Underperformance is here defined as being in the lower one percentile of the performance measures of one or more local criteria. The dotted line represents the percentage (right ordinate) of the underperforming models to the total number of models in this class.

Table 2 Flood inundation likelihoods calculated using Eq. (1) for three different cells in the floodplain as calculated using the observed inundation data for the January 2003 event

	p^{Flood} (L_G)	p^{Flood} (L_{S1})	p^{Flood} (L_{S2})	p^{Flood} (L_{S3})	p^{Flood} (L_{S4})	p^{Flood} (L_{S5})	p^{Flood} (L_{S6})	p^{Flood} (L_{S7})	p^{Flood} (L_P)	p^{Flood} (L_{max})	p^{Flood} (L_{min})	p^{Flood} (L_{prod})	p^{Flood} (L_{sum})
Location 1	0.15	0.12	0.12	0.12	0.12	0.12	0.12	0.23	0.03	0.15	0.16	0.18	0.12
Location 2	0.60	0.74	0.58	0.55	0.55	0.52	0.54	0.54	0.52	0.60	0.62	0.45	0.60
Location 3	0.91	0.89	0.97	0.87	0.85	0.88	0.88	0.88	0.89	0.93	0.93	0.72	0.96

The flood hazard has been computed with various performance values: L_G (global performance), L_S (subdomain performance, see Fig. 1), L_P (point performance, see Fig. 1) and aggregation of the subregions (maximum (L_{max}), the minimum (L_{min}), the product (L_{prod}) and the sum (L_{sum}) of all sub-region likelihoods).

flooding when computed with the global measure (column 2) is 15% at location 1, 50% at location 2 and 91% at location 3. For the performance measure of subdomain 5 ($L(S_5)$, column 7), the possibility of flooding at location 1 is 12%, at location 2 is 52% and at location 3 is 88%. The results are as expected: flood hazard depends on the performance measure chosen and its relationship to the location of the area of interest in the flood plain. Generally, the variance of all performance measures increases with decreasing probability of flooding. This means that areas with a low flood hazard have a higher uncertainty of the *true* flood hazard in comparison to areas which have a high probability of flooding. However, the concept of a *unique* flood hazard has to be rejected. Each of the performance measures can be reasoned logically according to a model aim. Moreover a higher order correlation between the performance measures, the location of the area of interest and all combinations of different performance measures cannot be attributed towards individual contributions.

The analysis presented so far has discussed only the evaluation of model performance with respect to one set of inundation observations for the January 2003 calibration event. An increase in the number of observations, or in the number of calibrated parameters (e.g. by using local roughness coefficients), will lead to an increased complexity of the model calibration problem. Further ambiguity will be introduced by the choice of how to combine multiple performance measures. The differences in flood hazard for a calibrated model (Table 2) also indicate that any value of flood hazard computed with an uncalibrated model, or one used outside its calibration range, may be subject to significant uncertainty.

These results are subject to the use of only two globally applied roughness coefficients and the implementation of LISFLOOD in this application using a 50 m grid. It could be argued that these limitations mean that the model cannot be expected to perform locally as well as globally. The most obvious solution would have been to introduce spatially disaggregated roughnesses or propose a nested model approach. A model with higher degrees of freedom may be able to reproduce local as well as global phenomena and thus avoid the problem of having behavioural global fits, which are unacceptable on the local scale. However, even with a maximum degree of freedom in the effective roughness values, we hypothesise that it maybe impossible to find models which fit everywhere. This is due to uncertainties in the observations used for calibration and the approximations inherent in the implementation of the model. Unfortunately, current limitations in CPU time and model complexity make it impossible to search the very high dimensional parameter spaces that a distributed roughness model involves, especially if finer grid scales are used. Therefore, this study (in common with all applications of flood inundation models in engineering practice) avoided this step. One possible solution may be a detailed analysis of local errors to cluster distributions of effective parameter values (Schumann and Matgen, 2006), but there will be no guarantee that a combination of local parameter distributions identified by local evaluation will produce globally acceptable models.

If model performance is such that global calibration might lead to predictions that are locally inadequate, one

response would be to concentrate on the local performance of the model (for others see discussion section). In particular, it suggests that we should concentrate our attention on models that do well in predicting inundation at important locations on the flood plain. In general, this will be where vulnerability is greatest. It follows that it may be necessary to use specific performance measures for specific locations that explicitly take vulnerability into account.

A vulnerability-based performance measure

Computing a local possibility of flood inundation based on vulnerability

In Eq. (1), the likelihood measure is easily modified to incorporate a weighting function that reflects local vulnerability. In this example, the performance measure of Eq. (A.3) (Appendix A) is multiplied by a relative vulnerability weight for each cell included in the evaluation, which results in a vulnerability weighted performance measure:

$$L(v) = \frac{\sum_{i,j=1}^n v_{i,j} S_{i,j}}{n} \quad (2)$$

where $v_{i,j}$ is the vulnerability weight of cell i, j (see Table 3), S is the similarity measure between the prediction and observed data in cell i, j , and n is the number of cells included. The similarity measure has been computed after Pappenberger et al. (2006a) and is illustrated in Eq. (A.2).

$$P(v)_{i,j}^{\text{flood}} = \frac{\sum_{m=1}^n s_{i,j,m} L(v)_m}{\sum_{m=1}^n L(v)_m} \quad (3)$$

where $P(v)$ is possibility of cell being wet given assumptions on vulnerability and s is the binary state of the cell (wet = 1, dry = 0).

A detailed description of how to calculate the similarity measure is given in Appendix A. This formula contains a number of simplifications in comparison to the more standard ways in which loss functions are computed. Future research could use a more integrated approach and adopt a nonlinear relationship between v and S for example by including water level and/or velocity into v and S in Eq. (2). These simplifications will not alter the methodology of this paper, but instead introduce an additional layer of complexity to the argument. In this example, the weights of the vulnerability are computed according to the length of road in each cell (ignoring the type and importance of structures), the types of building and a time component (see Table 3).

The time factor is added to demonstrate that for example buildings may exhibit different vulnerability depending on the time of day. This has been done to illustrate a particular point: when a flood inundation event is calibrated for a particular point in time, which is usually the time(s) of the available distributed observations, it is still necessary to focus on the purpose of the model in its practical use. It may for example well be that for emergency planning a time dependent flood hazard map is necessary. If vulnerability is defined in terms of potential loss of life, industrial buildings could have a relatively higher weight in the time between 9 a.m.

Table 3 Summary of targeted performance and weights of combinations

Targeted performance measures							
(a) Road km							
(b) Number of residential buildings							
(c) Number of industrial buildings							
(d) Number of agricultural buildings							
(e) Number of public buildings							
(f) Number of commercial buildings							
Name	Weights of combined performance measures for different time periods						
	(a)	(b)	(c)	(d)	(e)	(f)	(f)
Day time	0.05	0.05	0.3	0.1	0.2	0.3	0.3
Rush hour	0.25	0.15	0.15	0.15	0.15	0.15	0.15
Night time	0.1	0.7	0.05	0.05	0.05	0.05	0.05

The weighting only partially acknowledges the type of building as usually taken into account in computing damage functions (e.g. Penning-Rowsell and Chatterton, 1977). The classification has been derived from a land use map provided by the Administration du Cadastre du Luxembourg.

and 5 p.m. in comparison to residential areas, while roads subject to flooding or critical road junctions might have higher weight during rush hours.

In this scheme, non-target cells may be viewed as having a weighting factor of zero, but more complex, targeted measures can be devised to combine fit averaged across the target cells with fit across the remainder of the floodplain, via a suitable weighting scheme between target and non-target cells. Table 3 shows the targeted performance measures, which have been designed for this approach. The relative weighting of daytime, rush hour and night time periods is an arbitrary decision that we have used to demonstrate the concept, and could be changed as necessary.

The application of Eq. (2) embeds vulnerability directly into the model evaluation process. It provides a model performance measure that is weighted towards providing better predictions at pixels with high relative vulnerability (however that is estimated). The scaled weights are still applied as likelihoods for each set of model predictions using Eq. (3). Thus, as before, every pixel will have a predicted likelihood of inundation over all the behavioural models in the GLUE methodology, even those of low or zero vulnerability. For these pixels, however, there is less concern about getting accurate predictions of inundation. The possibility of overfitting the model to high vulnerability areas will be mitigated by applying Eq. (2) at the global level. This may still mean that the predictions may not provide accurate predictions for some high vulnerability pixels, but this should then induce a re-evaluation of the input data, model implementation or observation accuracy for those areas of model failure.

In Fig. 4, the relationship between the vulnerability based measures of Table 3 and the original global performance measure is displayed. Most of the performance measures indicate a step response, explained by the localisation of the measure as explained above. The fewer building types that a certain category has, the more clear the step response becomes. The limited number of agricultural buildings results in the existing buildings always being either flooded or not flooded, and thus has no distribution. The majority of buildings are residential buildings and a nearly linear relationship exists between the two weightings.

It can be seen that the introduction of the vulnerability weighted performance evaluation can lead to a significant change of model performance (Figs. 2 and 4). A well performing model for one weighting scheme does not necessarily perform well with another weighting. This change cannot be ignored in the creation of flood hazard maps.

In Fig. 5, two of the flood hazard maps generated are compared with the original global evaluation. The figure on the left shows the change of flood hazard between the global performance and the scheme weighted by road km. The figure on the right shows the change of flood hazard between the global performance and the scheme weighted by all buildings. The values are categorized for easier viewing and a significant increase or decrease indicates a change of more than 10%. A slightly reduced or increased flood hazard is equivalent to a change between 2% and 10%. Any change below 2% has been neglected as insignificant. It has to be pointed out that the weighting towards a feature is embedded in the global performance function (see Eq. (A.3)) such that in the results for any given cell each model simulation contributes a likelihood of inundation based on its global performance.

There is a visible difference between the global performance measure and each of these weighting schemes and between the two weighting schemes themselves. Several large areas exhibit a significant change in flood hazard, such as Area I and Area II marked in Fig. 5. In Area I, the flood hazard in a small village (Hunsdorf) changes significantly. There is a significant difference between the global performance measure and the building weighted measure (left figure) at the top end of the circle ('significantly increased'). In addition, the same area has a completely different value of flood hazard when a road km weighted scheme is used ('slightly reduced'). The entire Area II has approximately the same elevation and is all flooded at the same time. With the building weighted measure a significant reduction of flood hazard can be observed and with the road weighting a slight increase. Because this area is flooded at much the same time and therefore, the flood hazard changes over a large area rather than at the individual pixel scale. This illustrates the complex non-linear interactions between topography, model and performance measure.

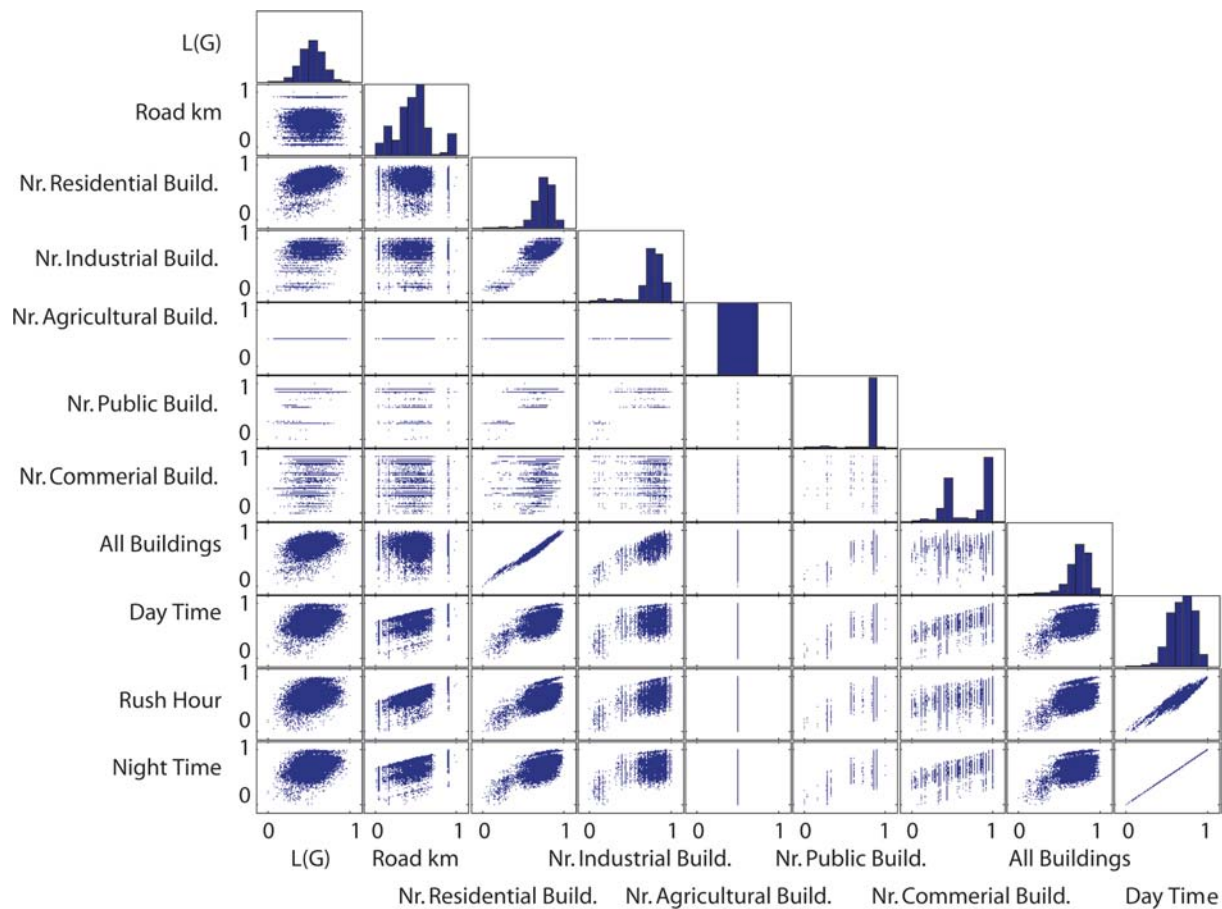


Figure 4 The performances of the different risk based evaluation criteria (Table 3) are plotted against each other. Each dot represents one model simulation. The histogram of each performance measure is also presented. All performance measures have been normalized between 0 and 1 to allow for better comparison.

Discussion and outline of future work

The difference in these flood hazard maps comes from explicitly acknowledging variation in the importance of different locations in the floodplain as expressed in terms of vulnerability measures. This differentiation of the importance of different parts of the floodplain is analogous to a multi-criteria analysis in rainfall–runoff modeling. For example, Boyle et al. (2003) subdivided a hydrograph into three different response types and evaluated the model on each of them. Three essential questions that have to be answered in this context are as follows: (i) what should be done if a model is acceptable for one performance measure but continuously underperforms for another evaluation criteria? (‘partial failure in model calibration’); (ii) how can local, global and combined performance measures be used to compute flood hazard maps and (iii) does the use of local measures lead to over fitting?

Partial failure in model calibration

The problem of distributed model simulations consistently performing well at location A whilst underperforming at location B. Similar has been introduced by Freer et al. (2004) for rainfall–runoff models and extensively discussed in Pappenberger et al. (2006b) for flood inundation models.

While we would hope that a physically-based hydraulic model should be capable of providing good predictions everywhere, the sub-domain analyses presented earlier have shown that in this application there is reason to reject all the model simulations run with global values of flood plain and channel roughness coefficients.

It can be argued that some models, such as that used in this paper, have been developed for applications at a certain scale and should not be expected to reproduce local phenomena. Indeed, it would be unrealistic to expect that LISFLOOD should reproduce point depth or velocity measurements. There is an expectation, however, that given appropriate floodplain and channel topography data it should reproduce the general pattern of inundation during an event: that is what the model has been designed for and it has been shown to perform well on hypothetical examples (Hunter et al., 2005).

The partial failure in model calibration allows for five possible responses (see Pappenberger et al., 2006b): (i) investigate those regions of the flow domain where there are consistent anomalies between model predictions and range of observations; (ii) avoid using data that are in some way doubtful; (iii) introduce local parameters if there are particular local anomalies; (iv) make error bounds wider in some way where data is doubtful; or, if none of the above can be justified (no reason to doubt anomalous data) by

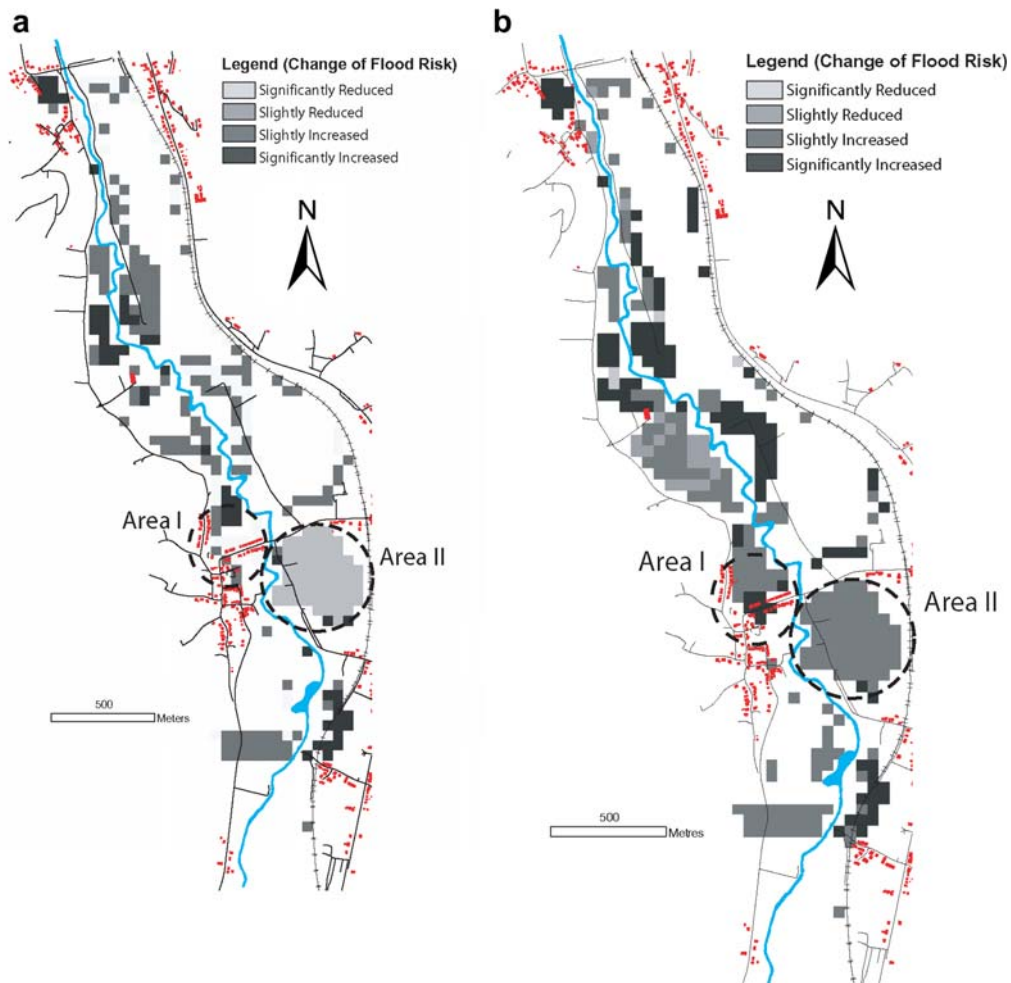


Figure 5 Change of flood hazard from maps calculated with the global performance measure to maps weighted by road km (left, a) and buildings (right, b). There is no change in flood risk in all white areas (land use map provided by the Administration du Cadastre du Luxembourg).

then (v) resort to local evaluations in assessing local uncertainties. In this paper, we have introduced a sixth solution, which is a vulnerability weighted global measure, which is an attempt to ensure that the behavioural models give more accurate predictions in those locations where good predictions are important.

The choice of vulnerability weight remains subjective. Even when we apply multiple criteria, the way in which measures are chosen and combined will be purpose specific. The use of vulnerability weighted measures makes this more explicit. There remains the possibility, of course, that even vulnerability weighted measures will result in the rejection of all the models tried, suggesting that either the input data or the model structure/parameterisation are inadequate for the goal stated.

Creation of flood hazard maps in a multi-objective framework

This problem mainly arises as the global performance measure necessarily averages performances and thus not always reproduces the local phenomena. This study explicitly recognizes the nature of this inadequacy and offers one possi-

ble fall-back position to get it right in places that matter. We recognize that different ways of computing model performance will influence the distribution of flood hazard and global, local and combined measures can create different estimates of flood hazard. This can be an important issue, when for example the influence of a new development on flood hazard for neighbouring houses is determined. Each of the buildings may have their 'own' flood hazard map. In principle, a decision of the choice and combination of performance measures should reflect the intended usage of the model, which implies that this multi-objective calibration problem cannot be entirely objective. This offers a real possibility of end-user engagement. The uncertainties associated with model calibration have to be considered more explicitly by flood inundation modellers while the importance of different sites or types of loss is a topic the end-user can engage with. Engaging the end-user cannot be seen as introducing additional subjectivity into an objective model evaluation, as past approaches have always been subjective in the choice of performance measure (or a priori choice of parameter values where no evaluation data are available). Of course, this raises questions about how stakeholders will agree about what is important in assessing vul-

nerability, or how to deal with multiple measures of 'importance'. The result may not be one single flood hazard map, but rather purpose specific hazard maps. We acknowledge that this will pose additional problems in the communication of flood hazard, but such a discussion is beyond the scope of this paper.

Over-fitting and predictability

The starting point for the use of vulnerability weighted likelihood weighting as used in the paper, was the potential for rejecting all the inundation models considered behavioural in terms of the normal global evaluation, when their local inundation predictions were evaluated. The use of local performance measures, or vulnerability weighted measures, however, will introduce the potential for overfitting the model to the errors in the modelling process for particular areas, which could lead to biased or wrong predictions for a different event of different magnitude. Romanowicz and Beven (2003), for example, show how distributions of effective channel and floodplain roughness values could be quite different when evaluated for events of different magnitude but the number of studies in which the predictions for different events could be assessed remains very small. To assess predictability, or the potential for overfitting, we need much more experience in assessing model extrapolations to either different events or different rivers.

Robustness in such extrapolations implies avoiding overfitting by a form of averaging of errors. There is a seamless transition between a local and global orientated performance evaluation as the more points one uses for fitting, the more global will be the properties of the evaluation. How 'local' an evaluation should be depends on the objectives of the modelling exercise. A balance between fitting the model to individual pixels and averaging the overall performance has to be found. Hence, the approach adopted here to give greater weight to those areas considered to be more vulnerability but to average the resulting measure globally over the full reach. In this way, the effects of very specific errors at specific locations might, to some extent, also be averaged out. The evaluation process should be based on a conscious decision about relative vulnerability, rather than a decision purely based on the size of the available image.

In this paper, we have not considered the possibility of allowing channel and flood plain roughness to vary spatially. In principle, we would expect local values of roughness to give more accurate local predictions (though we are not sure how great this effect might be). While recognising that varying only global values of roughness must inherently result in error for particular locations, however, allowing roughness to vary downstream will introduce its own overfitting problem. The more values of roughness that are included in the analysis, the more potential there is of fitting to local error, reducing predictability in extrapolation. The curse of dimensionality also arises: the greater the number of parameter dimensions to be considered, the greater the computational problem in making enough runs to evaluate the range of responses. Again, there is little experience available in the literature to enable the value of introducing varied roughness to be assessed in different circumstances. The study of Werner (2004) is interesting in this respect. He showed that introducing spatially variable flood

plain roughness made the distribution of the channel roughness (but not the floodplain roughness) for the behavioural models more identifiable. This result has yet to be confirmed elsewhere, and the resulting parameter estimates were not, in his study, subjected to an evaluation of predictions at another event magnitude.

Conclusion

Flood events causing major damage generally occur very rarely and thus historical records of areas liable to flooding are always scarce. Therefore, most flood hazard maps are based on numerical modelling and, in some cases, these models can be calibrated and evaluated on data from past events. A flood hazard map can be conditioned on the model performance in the calibration exercise. It has been documented that the different ways of computing the performance have a significant impact on the assessment of flood hazard for particular locations. This can be mainly attributed to the fact that the predictions of current flood models are subject to error in input data, model structure and the observations used in model evaluation. Therefore, the problem of deriving a correct flood hazard map is more fundamental: to date, 2D inundation models have always been evaluated against inundation extent using global performance measures that are obtained by averaging spatial performance across a suitable domain. In the case of flood inundation models this is normally an estimate of the flood prone area. Conditioning models on global performance has the disadvantage that the performance at a local scale is only considered indirectly. Here, it has been shown that *all* the models considered acceptable in terms of global performance can be rejected on the basis of equivalent sub-domain performance measures.

Should the affected cells contain buildings, transportation links or similar, then the potential economic and social consequences of the inadequate prediction of local hazard will be more severe than if the cells contain only undeveloped agricultural land. In order to support flood hazard management decisions with imperfect models, the calibration or conditioning of flood inundation models can be weighted in favour of the vulnerability of target structures such as buildings and roads rather than just relying on global performance in predicting inundation. Additionally, when a flood inundation event is calibrated for a particular point in time, which is usually the time of observation, it is still necessary to focus on the usage of the model. For example, for emergency planning a time dependent flood hazard map could be derived. Indeed, the potential consequences of false flood warnings to key buildings such as hospitals are sufficiently severe that it might even be considered prudent to condition a model solely in relation to the fit associated with a single structure, while remaining aware of the potential danger of less robust predictions in other conditions as a result of overfitting the calibration data. Consequently, in a setting in which the model is used to produce flood hazard maps, for e.g. a particular town, the vulnerability cannot be used as a simple add-on to the modelling process, but should be incorporated into the calibration procedure. This approach allows stakeholders to uniquely engage in the modeling process and offers the potential for new ways of calibrating flood inundation models in the future.

Acknowledgements

We thank Georges Müller of the Service de la Gestion de l'Eau for providing some of the data used in this study. The land use map has been provided by the Administration du Cadastre du Luxembourg for which we are grateful. This study would have been impossible without the help of Paul Bates (Professor of Hydrology, Bristol University) and Neil Hunter (Bristol University). The paper is based on the Master Thesis by Kevin Frodsham for his Master degree in Environmental Science (distinction). Florian Pappenberger and Renata Romanowicz have been funded by the Flood Risk Management Research Consortium (<http://www.flood-risk.org.uk>). Development of extensions to the GLUE methodology has been supported by NERC Grant NER/L/S/2001/00658 awarded to Prof. Keith Beven.

Appendix A. Fuzzy performance measure

Traditionally inundation model have been evaluated on by separating observations in each cell into binary categories of wet or dry (Aronica et al., 2002; Bates and De Roo, 2000; Horritt and Bates, 2001; Hunter, 2006). However, the true inundation extent can be estimated in many cases only with large uncertainties. Therefore, in Pappenberger et al. (2006b) a fuzzy mapping technique has been applied which takes account of this uncertainty. Fuzzy categories (high, medium, low, no flooding) have been created from the observed and modelled data for this computation, which reflect the certainty in the flooding of a particular cell. These maps have then been compared and a fuzzy performance value computed.

Four Fuzzy categories have been assigned to each cell of the observed and modeled maps a detailed description on deriving these categories has been presented by Pappenberger et al. (2006a). The categories of the observed map have been derived from the uncertainty in classifying the backscatter of a SAR image. The modeled categories are based on the uncertainty in the topography of surrounding cells.

$$V_{CAT} = \{V_{CAT_{high}}, V_{CAT_{medium}}, V_{CAT_{low}}, V_{CAT_{no}}\} \quad (A.1)$$

where $V_{CAT_{high}} = (1, 0.6, 0.3, 0)$, $V_{CAT_{medium}} = (0.6, 1, 0.6, 0.3)$, $V_{CAT_{low}} = (0.3, 0.6, 1, 0.6)$, $V_{CAT_{no}} = (0, 0.3, 0.6, 1)$. The similarity measure has then be computed by comparing the fuzzy categories of the observed (A) and modelled maps (B):

$$S(V_A, V_B) = \frac{|A_{CAT_{high}}, B_{CAT_{high}}|_{min} + |A_{CAT_{medium}}, B_{CAT_{medium}}|_{min} + |A_{CAT_{low}}, B_{CAT_{low}}|_{min} + |A_{CAT_{no}}, B_{CAT_{no}}|_{min}}{|A_{CAT_{high}}, B_{CAT_{high}}|_{max} + |A_{CAT_{medium}}, B_{CAT_{medium}}|_{max} + |A_{CAT_{low}}, B_{CAT_{low}}|_{max} + |A_{CAT_{no}}, B_{CAT_{no}}|_{max}} \quad (A.2)$$

$$L = \frac{\sum_{i,j=1}^n S_{i,j}(V_A, V_B)}{n} \quad (A.3)$$

where n representing the number of inundation prone cells for all simulations and the observed data set. i and j give the cell location.

Appendix B. Point performance method

For this study, the boundary between the zero possibility inundation category and the nearest cell of any other fuzzy category (see Appendix A) was selected to divide the maps

into wet and dry sectors. A detailed description of this method is given in Pappenberger et al. (2006a). This choice of boundary represents a conservative choice for the shoreline and was chosen to reflect a close to maximum hazard to the target cell as would have to be considered in a risk analysis. Other choices of boundary can be chosen from fuzzy category maps to represent different hazard levels; for example, analyzing performance with respect to the distance from target to nearest cell of a high inundation probability would be a suitable choice to condition the model on the hazard of severe flooding to the target cell. The performance measure L_{PA} was evaluated as the absolute difference between the observed ($dist_{A,OBS}$) and modelled ($dist_{A,MOD}$) distances from the target cell, A, to the nearest cell across the hypothetical shoreline. θ is the angular component as an approaching flood may be important depending on the direction of approach from all directions. For example, evacuation roads may be situated in only one direction. Such a measure could be sophisticated by mapping the entire access roads and include them into the analysis of Eq. (A.4). $w(t)$ is a weight for the time component as the evacuation during peak time of surgeries conducted may be more significant than otherwise. The angle as well as the time weight are included in Eq. (A.4) for completeness and are not used in this paper

$$L_P = \theta^* w(t)^* |dist_{A,OBS} - dist_{A,MOD}|. \quad (A.4)$$

An optimum fit for L_P , as defined in Eq. (A.4), of zero will occur when the target cells are both wet and the chosen shoreline is equidistant from the target cell on both observed and model predicted maps irrespective of the direction of nearest approach.

References

- Alexander, D., 1991. Natural disasters: a framework for research and teaching. *Disasters* 15 (3), 209–226.
- Aronica, G., Bates, P.D., Horritt, M.S., 2002. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrological Processes* 16 (10), 2001–2016.
- Aronica, G., Hankin, B., Beven, K.J., 1998. Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data. *Advances in Water Resources* 22 (4), 349–365.
- Bates, P.D., De Roo, A.P.J., 2000. A simple raster-based model for flood inundation simulation. *Journal of Hydrology* 236 (1-2), 54–77.
- Beven, K.J., 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* 320 (1–2), 18–36.
- Beven, K.J., Binley, A., 1992. The future of distributed models – model calibration and uncertainty prediction. *Hydrological Processes* 6 (3), 279–298.
- Blazkova, S., Beven, K., 2004. Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic. *Journal of Hydrology* 292 (1-4), 153–172.
- Bouma, J.J., Francois, D., Troch, P., 2005. Risk assessment and water management. *Environmental Modelling & Software* 20 (2), 141–151.
- Boyle, D.P., Gupta, H., Sorooshian, S., 2003. Multicriteria calibration of hydrologic models. In: Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), *Advances in Calibration of Watershed Models*. American Geophysical Union, Washington.

- Cameron, D., Beven, K., Tawn, J., Naden, P., 2000. Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation). *Hydrology and Earth System Sciences* 4 (1), 23–34.
- European Environment Agency, 2006. EEA multilingual environmental glossary, <http://glossary.eea.eu.int/EEAGlossary/>. European Environment Agency.
- Freer, J., Beven, K.J., Peters, N., 2003. Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure. In: Duan, Q.Y., Gupta, H., Sorooshian, S., Rousseau, A., Turcotte, R. (Eds.), *Calibration of Watershed Models*. American Geophysical Union, Washington, pp. 69–88.
- Freer, J.E., McMillan, H., McDonnell, J.J., Beven, K.J., 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology* 291 (3–4), 254–277.
- Granger, K., Jones, T., Leiba, M., Scott, G., 1999. Community Risk in Cairns: A Multihazard Risk Assessment. AGSO (Australian Geological Survey Organisation) Cities Project, Department of Industry, Science and Resources, Australia.
- Horritt, M.S., Bates, P.D., 2001. Predicting floodplain inundation: raster-based modelling versus the finite-element approach. *Hydrological Processes* 15 (5), 825–842.
- Hunter, N.M., 2006. Flood Inundation Modelling – PhD Thesis, School of Geography, Bristol University, Bristol.
- Hunter, N.M., Horritt, M.S., Bates, P.D., Wilson, M.D., Werner, M.G.F., 2005. An adaptive time step solution for raster-based storage cell modelling of floodplain inundation. *Advances in Water Resources* 28 (9), 975–991.
- Jonkman, S.N., van Gelder, P.H.A.J.M., Vrijling, J.K., 2003. An overview of quantitative risk measures for loss of life and economic damage. *Journal of Hazardous Materials* 99 (1), 1–30.
- Kelman, I., 2002. Physical Flood Vulnerability of Residential Properties in Coastal, Eastern England, PhD Thesis, University of Cambridge, UK. Available from: <http://www.ilankelman.org/phd.html>.
- Merz, B., Kreibich, H., Thielen, A., Schmidtke, R., 2004. Estimation uncertainty of direct monetary flood damage to buildings. *Natural Hazards and Earth System Sciences* 4 (1), 153–163.
- Pappenberger, F., Beven, K.J., Frodsham, K., Romanowicz, R., Matgen, P., 2006a. Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrology and Earth System Sciences (Discussions)*. Available from: <http://www.copernicus.org/EGU/hess/hessd/3/2243/hessd-3-2243.htm>.
- Pappenberger, F., Beven, K.J., Horritt, M.S., Blazkova, S., 2005. Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. *Journal of Hydrology* 302, 46–69.
- Pappenberger, F., Matgen, P., Beven, K.J., Henry, J.-B., Pfister, L., de Fraipont, P., 2006b. Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources* 29 (10), 1430–1449.
- Penning-Rowsell, E.C., Chatterton, J.B., 1977. *The Benefits of Flood Alleviation: Manual of Assessment Techniques*. Gower, Aldershot.
- Romanowicz, R., Beven, K.J., 2003. Estimation of flood inundation probabilities as conditioned on event inundation maps. *Water Resources Research* 39 (3), W01073. doi:10.1029/2001WR001056.
- Romanowicz, R., Beven, K.J., Tawn, J., 1996. Bayesian calibration of flood inundation models. In: Anderson, M.G., Walling, D.E., Bates, P.D. (Eds.), *Floodplain Processes*. John Wiley & Sons, New York, pp. 333–360.
- Schumann, G., Matgen, P., 2006. Personal communication.
- Smith, K., 2001. *Environmental Hazards: Assessing Risk and Reducing Disaster*. Routledge, London, 392 pp.
- Soetanto, R., Proverbs, D.G., 2004. Impact of flood characteristics on damage caused to UK domestic properties: the perceptions of building surveyors. *Structural Survey* 22 (2), 95–104 (special issue, Flooding: Implications for the Construction Industry).
- Thielen, A.H., Muller, M., Kreibich, H., Merz, B., 2005. Flood damage and influencing factors: New insights from the August 2002 flood in Germany. *Water Resources Research* 41 (12).
- Turner, R.K., Bateman, I., Adger, N., 2001. Economics of coastal and water resources: valuing environmental functions. *Studies in Ecological Economics*, vol. 3. Kluwer Academic, Dordrecht, London, vii, 342 pp.
- Werner, M., 2004. Spatial flood extent modeling: A performance-based comparison, PhD Thesis, Delft University of Technology/DUP Science/Delft Hydraulics Select Series 4 (ISBN 90-407-2559-4). Available from: http://www.wldelft.nl/rnd/publ/docs/We_2004a.pdf.
- Werner, M., Blazkova, S., Petr, J., 2005. Spatially distributed observations in constraining inundation modelling uncertainties. *Hydrological Processes* 19 (16), 3081–3096.
- Wu, S.Y., Yarnal, B., A., F., 2002. Vulnerability of coastal communities to sea-level rise: a case study of Cape May County, New Jersey, USA. *Climate Research* 22(3), 255–270.