Calibration of flood inundation models using uncertain satellite observed water levels

Stephens, E.M; Bates, P.D; Freer, J and Mason, D.C.

Abstract

The performance of flood inundation models is often assessed using satellite observed data; however these data have inherent uncertainty. In this study we assess the impact of this uncertainty when calibrating a flood inundation model (LISFLOOD-FP) for a flood event in December 2006 on the River Dee, North Wales, UK. The flood extent is delineated from an ERS-2 SAR image of the event using an active contour model (snake), and water levels at the flood margin calculated through intersection of the shoreline vector with LiDAR topographic data. Gauged water levels are used to create a reference water surface slope for comparison with the satellite-derived water levels. Residuals between the satellite observed data points and those from the reference line are spatially clustered into groups of similar values. We show that model calibration achieved using pattern matching of observed and predicted flood extent is negatively influenced by this spatial dependency in the data. By contrast, model calibration using water elevations produces realistic calibrated optimum friction parameters even when spatial dependency is present.

To test the impact of removing spatial dependency a new method of evaluating flood inundation model performance is developed by using multiple random subsamples of the water surface elevation data points. By testing for spatial dependency using Moran's I, multiple subsamples of water elevations that have no significant spatial dependency are selected. The model is then calibrated against these data and the results averaged. This gives a near identical result to calibration using spatially dependent data, but has the advantage of being a statistically robust assessment of model performance in which we can have more confidence. Moreover, by using the variations found in the subsamples of the observed data it is possible to assess the effects of observational uncertainty on the assessment of flooding risk.

1. Introduction

Perhaps consider starting more like this – (note not written well here):

The quantification of flood inundation processes and the ability to effectively evaluate predictive models of such processes is a challenging research field. This is due to the nature of the flooding event, having complex and interacting spatial dynamics between the flood plain and the channel, that the event is extreme, that our data collection technologies … etc. These difficulties impact on the techniques we use to….

evaluate hydraulic model performance. Prior to the availability of remotely sensed data the majority of field measurements were available at the point scale (Bashford et al., 2000) with flood models primarily being 1D and validated using gauged records (Beven, 1987; Beven et al., 1984). However, the availability of LiDAR terrain data increased the feasibility of 2D flood inundation modelling (Marks and Bates, 2000) and along with these models came the clear need for distributed datasets to calibrate and validate predictions of the spatial pattern of flooding (Bates et al., 1997).

Remote sensing is an obvious source for such distributed data and early studies focussed on airborne and satellite optical sensors. Airborne photography of floods was limited by cost, requirements for suitable weather and the need for high altitude flights to view large floods (Usachev, 1983). Satellite data is a potential alternative and as this became more freely available monitoring floods from space became a viable option. Pioneering studies with satellite data focussed on visible/ infrared sensors (see Smith, 1997 for a discussion) however these sensors are limited to periods of clear skies and daylight (Melack et al., 2004; Rasid and Pramanik, 1990; Sanyal and Lu, 2004). Accordingly, research shifted towards active microwave sensors which have all weather and day-night capabilities.

Numerous studies presented at the First ERS Thematic Working Group Meeting on Flood Monitoring (ESA and ESRIN, 1995) found flood boundaries could be delineated accurately in synthetic aperture radar (SAR) data because of low radar backscatter from the surface of the water (Smith, 1997), which contrasted with higher returns from the relatively rough surrounding landscape. However, this delineation is complicated by emergent vegetation and buildings which cause multiple reflections, and wind or rain roughening of the water surface which increases backscatter, which cause flooded areas to look similar to those areas not flooded (Mason et al., 2009). With errors in satellite data likely to reduce the usability of these information sources for evaluation of model performance better image processing techniques were developed to remove or reduce errors in the extent delineation process. Horritt et al. (2001) describe the use of a statistical active contour model (snake), a method of SAR image processing which links together edge features from statistical properties of the image intensity data. Using aerial photography, this study the found the snake technique capable of identifying 75% of the flooded area correctly (Horritt et al., 2001), although it was noted that there were still errors caused by short vegetation adjacent to the flood which gave similar returns to open water. This was perhaps due to local wind-roughening of the surface or because the unflooded surface was wet due to rainfall. These features act to blur the margin between flooded and unflooded terrain.

In an attempt to overcome some of these problems, Mason et al. (2007) describe the use of airborne LiDAR data within a statistical active contour model to better delineate the SAR waterline. This method acts to reduce the errors in flood delineation by ensuring that the waterline varies smoothly in elevation along the reach. By including the assumption that changes in water surface elevations across the flood area usually vary only gradually it is possible to modify the snake algorithm (Horritt, 1999; Horritt et al., 2001) by weighting it so that the ground heights vary smoothly as well as the SAR intensity (Mason et al., 2007). The location of the SAR waterline can then be intersected with a LiDAR DEM to extract water surface elevations at the flood margin (Mason et al., 2009). When compared with water surface elevations derived by intersecting aerial photography with a LiDAR DEM (assumed to be ‘truth’) Mason et al. (2007) showed that the root mean square error (RMSE) in flood margin water elevations reduced from 221.1cm for the SAR data-only snake to 55.5cm for the LiDAR trained SAR snake (where points on slopes above 0.03 have been removed). However, despite the LiDAR constraint improving on some of the misclassification errors (Mason et al., 2007), considerable error still remains in these flood margin water elevations.

Water surface elevation data derived in the above way from SAR data can be used as an alternative method for assessing model performance by comparing actual and modelled water levels at the margin of the flood (Mason et al., 2009). An ideal model (using error-free observed data) would have no significant difference between its output and the observed

data, this can be tested using a T-test to evaluate the null hypothesis that the mean difference is not significantly non-zero (P(t>|to|)).

Prior to the availability of a method to improve the accuracy of flood shoreline elevation determination the most commonly used technique for comparing satellite derived flood extent with modelled data was to calculate a measure of fit between the modelled and observed binary pattern of flooding (e.g. Aronica et al., 2002; Bates and De Roo, 2000; Schumann et al., 2009; Yu and Lane, 2006). The extraction of water surface elevations at the margin provides scope for an alternative method of comparison which has yet to be fully evaluated and compared with the conventional pattern-matching method, although it has been shown to be more sensitive to modelled friction parameters than pattern matching methods (Mason et al., 2009). Hence, it potentially provides a more discriminatory measure of model performance.

Despite these advances errors in satellite-derived flood water levels still remain and there is a clear requirement to either identify factors which cause errors (and then remove points with these causal factors from any future analysis) or develop methods and performance measures which ensure that errors in observed data do not unduly influence any analysis. Errors in observed data are likely to be clustered since the factors which cause these errors, such as wind roughening of the water surface, will affect wide areas rather than specific points. This spatial pattern of errors could unduly influence an analysis if whole clusters of high error data points are included. However, given that it is unlikely to be possible to identify the cause of any particular cluster of errors it is difficult to create a systematic method to remove them. Instead, a method of analysis which removes the impact of this spatial dependency in data errors needs to be developed.

The objectives of this paper are therefore:

a) to assess errors and patterns of uncertainty in the water surface elevation data points derived by intersecting SAR-derived flood shorelines with high resolution DEM data

b) to develop a methodology for ensuring that spatial patterns of uncertainty in these observed data do not influence assessments of model performance, by ensuring that the stringent requirements of any statistical tests are met

c) to compare the calibration of flood inundation models using this new methodology, the original method using spatially-dependent waterline elevations and measures based on areal patterns of wet and dry pixels

d) to use the evaluated uncertainties to produce a probabilistic flood inundation map, and compare this with the deterministic flood inundation map when one optimum parameter set is used

2. Methodology

2.1. Study site and test data

The location of the study area is the confluence of the River Dee and the River Alyn in North Wales, UK. This study area was chosen due to the availability of satellite data as well as the frequency of overbank flow events, enabling studies of further flood events in future. The confluence location and the relatively low gradients in the area also provide a stern test for a flood inundation model due to the likely increased sensitivity to friction parameters.

The total area to be modelled covers an area of approximately 40km^2, and consists of the lowest 9km of the River Alyn and its confluence with an 11km section of the River Dee between Farndon (Holt Bridge) and Eaton Hall (Ironbridge). There are some tidal effects in this region which are captured by the downstream gauged stage. The River Alyn is on average 12m wide, and the River Dee 30m. The River Dee flows in a northerly direction, with the floodplain only about 150m in width for the first km of the study area. After this the floodplain opens out to, on average, 2km in width, narrowing to 0.5km for the final 3km of the study domain.

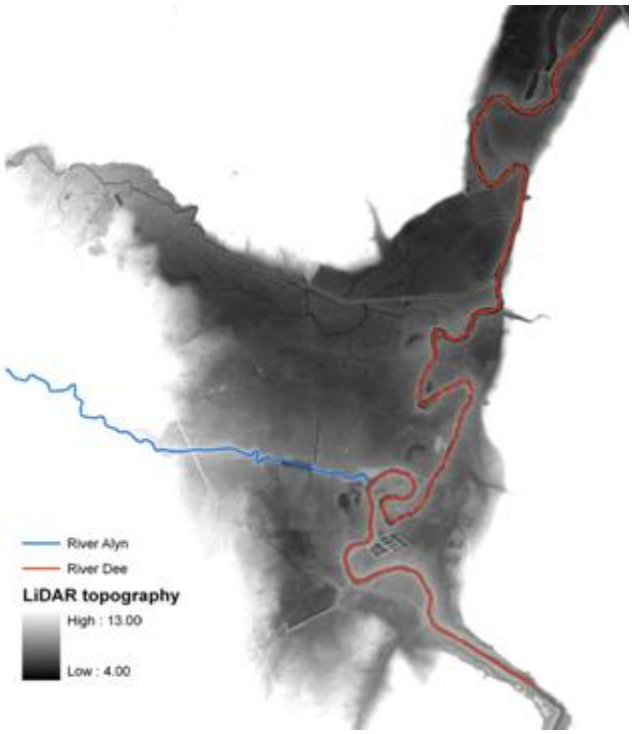

Figure 1: LiDAR DEM of catchment

The floodplain along this reach is rural, with the dominant land use being fields used for pasture but containing some isolated buildings. LiDAR data at 2m spatial resolution with a vertical accuracy of better than 15cm (see Figure 1) was acquired for the reach by the Environment Agency of England and Wales (EA), as well as ~1m vertical accuracy Interferometric Synthetic Aperture Radar (IfSAR) data at 5m spatial resolution. The study area floods on a frequent basis, with notable floods in 1964, 2006 and 2009.

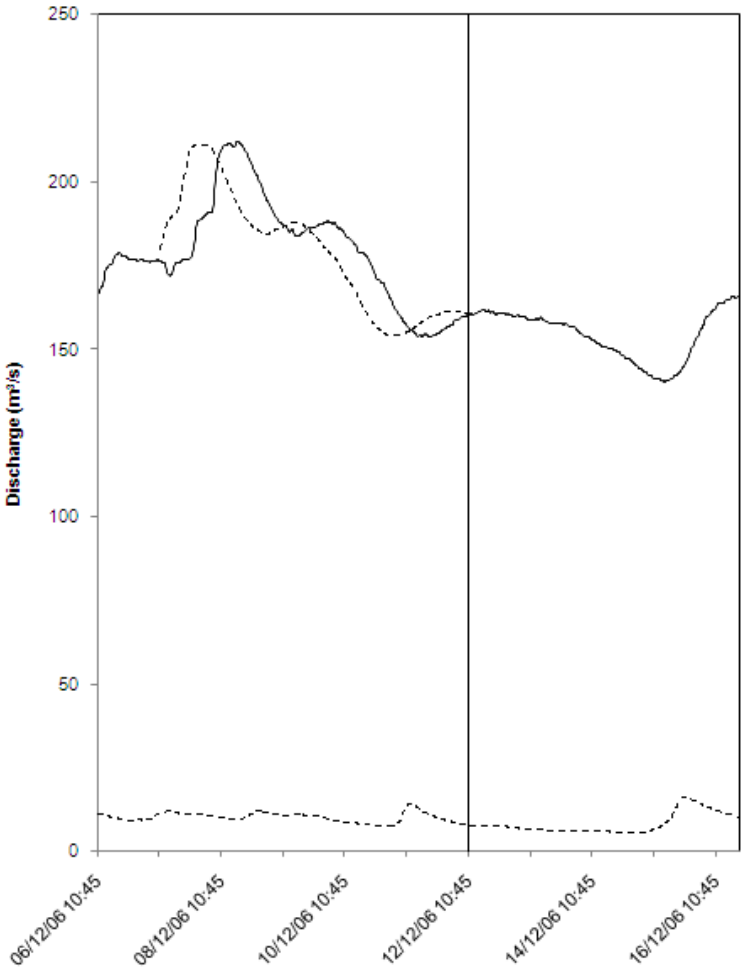

Figure 2: Gauged flow data for Ironbridge (solid line) and Pont-y-Capel (dashed line) with stage-derived flow for Farndon (dotted line). The vertical line represents the time of the satellite overpass on 12th December 2006.

Gauged flow and stage data are provided upstream for the River Alyn at Pont-y-Capel and downstream for the River Dee at the Ironbridge near Eaton Hall. Gauged stage data is provided upstream on the River Dee at Holt Bridge in Farndon. All these data have a temporal frequency of 15 minutes. The stage data at Farndon is converted to flows for use with the model using a flow rating curve which takes into account the downstream tidal effects. A 12.5m ERS-2 SAR image of a flood is available for 12th December 2006 at 11.07am. This corresponds to upstream gauged stage on the Dee of 8.77m a.s.l. at Farndon, and downstream at Eaton Hall of 6.675m a.s.l. Here we use the LiDAR guided snake algorithm described by Mason et al. (2009) to delineate a flood outline from the satellite image of the flood event. This algorithm weights the LiDAR topographic data with an 0.15 weighting factor relative to the SAR data (Mason et al. 2009). Water surface elevations at points along the margin of the flood are calculated by intersecting this outline with LiDAR topography data. This gives 232 elevation points along the margin of the flood, which will be referred to as ERS_{all}.

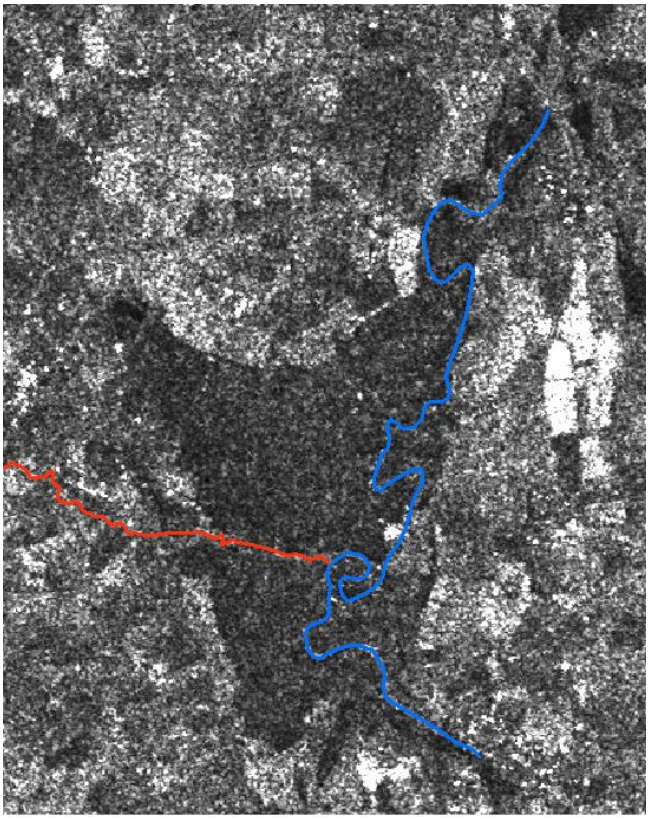

Figure 3: ERS-2 image of flooding on the River Dee, December 12th 2006

2.2. Identification of sources of uncertainty in the delineation techniques

Given the errors found in previous studies it is important to have some quantification of errors in the water surface elevation points derived from intersecting the SAR shoreline with the DEM. Unlike in Mason et al. (2007) for the 12th December 2006 event there is no available aerial photography to provide a high accuracy shoreline to which the SAR data can be compared. However, as this is a relatively short reach with known (gauged) water surface elevations upstream and downstream at the time of the flood event, validation of the satellite observed data can be achieved by creating a linear reference plane fitted to the upstream to downstream gauged elevations on the Dee. The slope from the Alyn was not included since the majority of the Alyn flooding resulted from backwater effects from the Dee and occurred around the confluence between the two rivers. The ERS_{all} points can be plotted against their distance downstream on the Dee, and then compared with the elevation at the same chainage on the reference line to produce a set of residuals between observed and reference data ($RESID_{all}$). This method assumes that the flood wave is linear, which has obvious limitations depending on the position of the flood wave within the reach (see Figure 2). However, given that the reach is relatively short, with gauges less than 10km apart, the deviation of the linear reference plane from the actual flood wave will be small. For example, Schumann et al. (2010) show a flood wave along a 98km reach of the River Po, Italy, which exhibits little deviation away from a linear plane for sections of river up to 30km. Using this validation method, despite its assumptions, makes it is possible to highlight points that are radically different to expectations, or which vary greatly from a smoothly varying water surface slope. Given that vertical accuracy in the LiDAR data is better than 15cm, and the actual flood wave is unlikely to vary greatly from the linear for this relatively small catchment, any large (i.e. > 0.25m) residuals flagged by this analysis are highly likely to be caused by

errors in the image processing techniques used to extract the flood shoreline from the SAR data.

Figure 4 plots ERS_{all} against the distance downstream (chainage), with the reference line shown. Clearly there is considerable variation from a smoothly varying water surface slope, of greater than 2m in places. Using the method outlined above we find the LiDAR guided method of Mason et al. (2007) to have a RMSE of 0.86m. Of the initial 232 points; 55 were removed since they were identified from aerial photography as being located in or along forested areas where the radar signal would not be able to penetrate below the tree canopy, A further 19 points were removed since the algorithm, using the LiDAR as a guide, delineated the flood along a drainage channel embankment rather than being able to identify flooding behind it. There are 158 points remaining after this initial reduction, with an average distance of 93.88m from one point to its nearest neighbour. A further qualitative analysis identified a cluster of 16 points with large (>2m) errors, close to the upstream gauge where the river emerges from a narrow valley. There were also removed since it was thought extremely unlikely that water levels could be that much greater than the nearby gauge. This quantitative elimination leaves 142 points, on average 95.79m apart. These 142 points, ERS_{cut}, have an RMSE of 0.70m if we assume that the reference line represents the true water surface profile at the time of the SAR overpass.

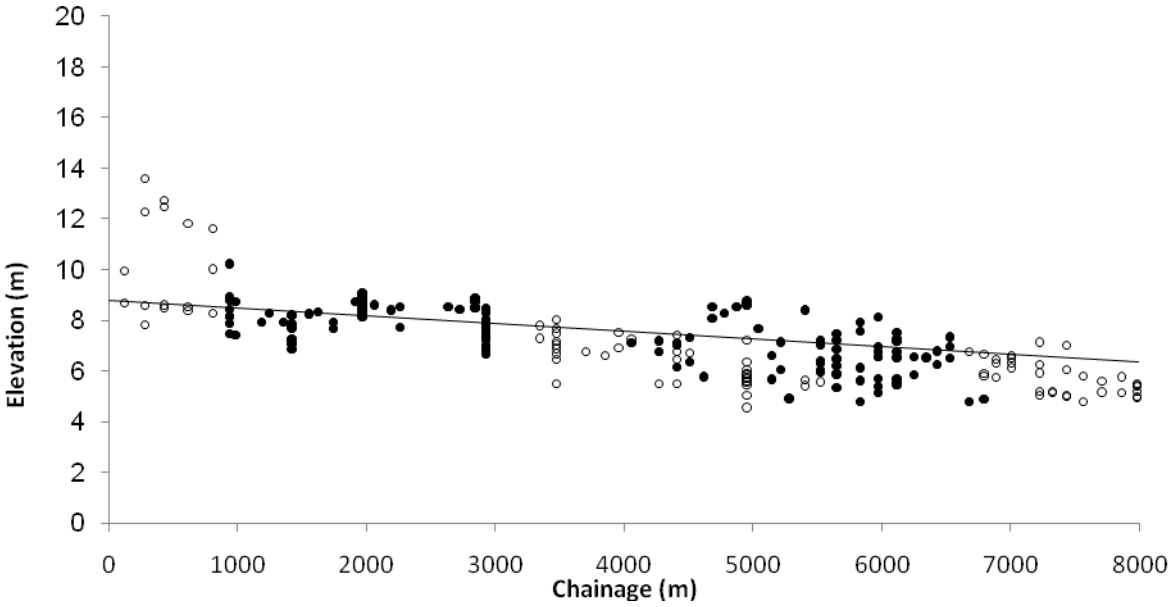

Figure 4: Elevation (m a.s.l) of ERS_{all} (black and white points) and ERS_{cut} (black points only) compared to the distance downstream (chainage). Solid line shows slope from upstream to downstream gauge on the River Dee.

2.3. Identifying spatial dependency

By mapping the residuals between the observed data and the reference data it is possible to highlight any spatial correlations. If observed data points were used to assess model performance which had significant spatial correlation in their errors, those point clusters with similar errors might unduly influence the results of the analysis. The spatial correlation in height errors in the satellite data can be visualised by mapping the ERS_{cut} residuals at each observed data point, with points of positive and negative errors greater than 0.25m (i.e. significantly above any likely LiDAR DEM vertical error) shown in Figure 5.

Figure 5 shows considerable spatial pattern to the residuals, with observed errors greater than expected for slopes on the western side of the catchment. Clearly, using all points in a comparison with modelled data could have a potentially significant impact on the analysis. The factors which lead to uncertainty in the observed data (wind roughening, vegetation, etc.) are likely to affect clusters of points, however as the causes of observed data uncertainty are difficult to determine it is not possible to remove these clusters from the analysis using any robust and systematic method. What can however be done is to ensure that there is no oversampling of points from clusters that all have similar errors. As such it is important to quantify and try to remove spatial dependency in the observed data.

Spatial dependency can be assessed using Moran's I (Moran, 1950), a statistical test used to evaluate whether observations of the same variable are significantly correlated with respect to their proximity from each other.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Equation 1

Where N is the number of spatial units indexed by i and j, x is the observed variable; $\bar{x}$ is the mean of x; and w_{ij} is the weighting of the distance between i and j. Moran's I values range from -1 (perfect dispersion) to +1 (perfect correlation, with values of 0 for a random spatial pattern. These values can be converted to a Z score (source: ESRI website) where $-1.96 < Z < 1.96$ represents values with no significant spatial pattern (dispersion or correlation) at the 5% level.

Although it is possible to carry out this test by only weighting adjacent cells, here we carry it out using an inverse distance weighting of the Euclidean distance between the xy coordinates of each pair in order to reflect the continuous rather than the categorised nature of the data. The 142 residuals from ERS_{cut} have a Moran's i value of 0.21 and Z value of 13.52 for a Mean Nearest Neighbour Distance (MNND) of 95.79m. Clearly it is necessary to increase the MNND until there is no significant correlation between the residuals at one point and those at its neighbours.

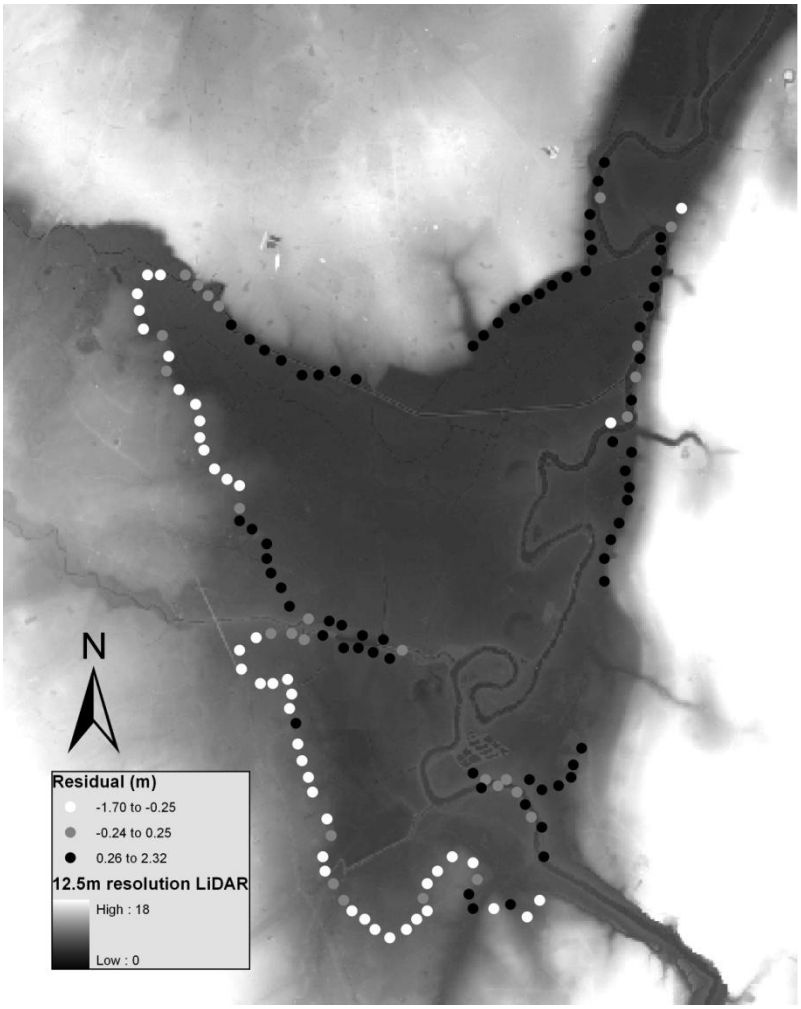

Figure 5: Elevation residuals for observed data points

2.4. Requirements for statistical tests

The requirements of a parametric test such as the paired T-test used by Mason et al. (2007) are that the values (here the residuals) have a normal distribution. Further, for any statistical test or model-data comparison there should be no significant correlation between values at adjacent points (spatial dependency). A plot of the residuals between ERS_{cut} and the reference line shows an approximate normal distribution, but a Kolmogorov-Smirnov test shows that this is not significant at the 5% level. Since it is then unlikely that smaller numbers of points will have a normal distribution it is not possible to use a parametric test for model evaluation in this case. Instead, we use RMSE as a performance measure.

Given that a Moran's I test has shown significant spatial dependency in the residuals between ERS_{cut} and a reference water surface slope there is a clear need for a method to remove the impact of this spatial dependency from any comparison with modelled data. There are perhaps two methods to achieve this; firstly, by using subsets of ERS_{cut} the MNND will effectively increase, and therefore groups of points can be found that do not exhibit spatial dependency. By using multiple subsets (there are over 3×10^{37} combinations of 50 points from the 142 original) it will be possible to look at multiple random combinations of all the observed data. This method uses subsampling of the original data, but a second method might be to instead retain all the data and weight each individual point based on how much it

is influenced by the characteristics of its neighbours. However, this will also penalise clusters of accurate as well as erroneous points, thereby preferentially weighting those in the analysis that are unique (and therefore perhaps more spurious). Developing a robust statistical method for this weighting is outside the scope of this study, and so we have chosen instead to develop the first method.

As a result of this requirement to increase the distance between neighbouring points it was necessary to investigate the relationship between the mean nearest neighbour distance (MNND) and the Z score for the Moran's i test, and in doing so determine the distance between points required in order to have no significant spatial dependency. This investigation was undertaken by performing the Moran's i test on multiple (10000) random configurations of different total numbers (between 20 and 142) of data points chosen from the set of 142 points. There is no significant spatial dependency when the Q value is less than 1.96 (or greater than -1.96).

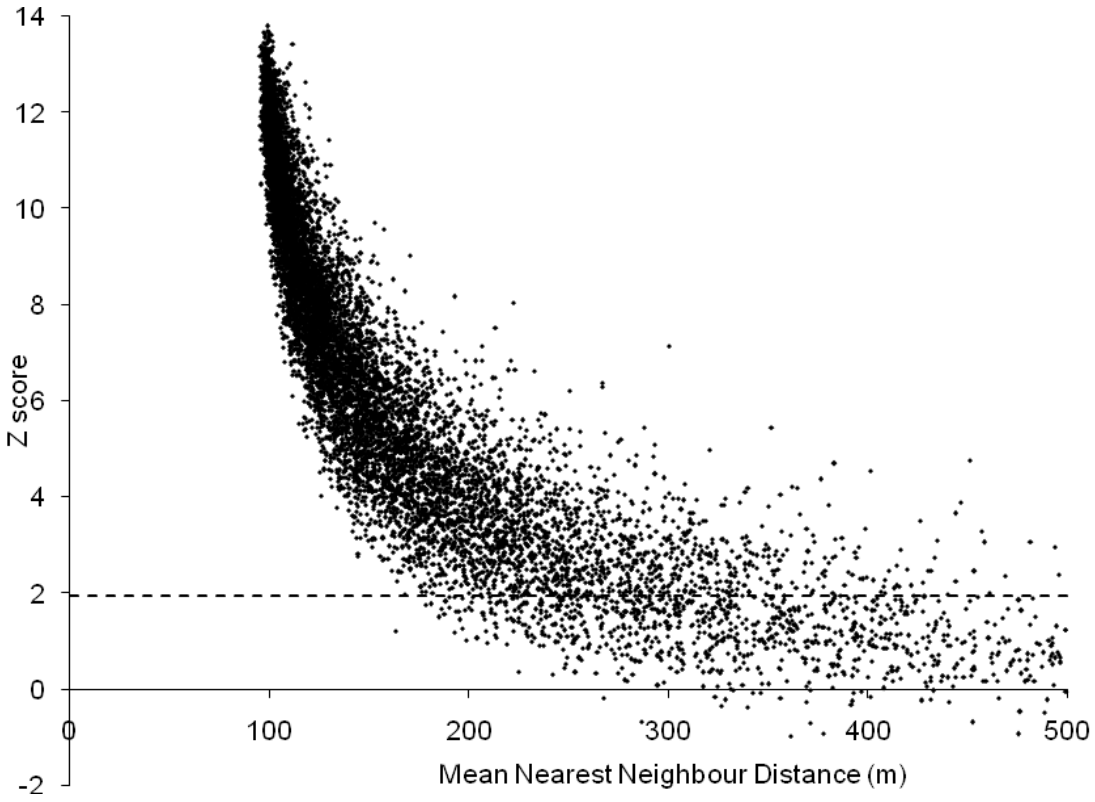

Figure 6: Z score for 10000 subsets plotted against Mean Nearest Neighbour Distance

Figure 6 shows that for MMNDs up to 150m there are no subsets of points from ERS_{cut} that show no significant spatial dependency, and only for MNNDs above 300m do approximately half of the subsets become non spatially dependent. The lowest MNND where subsets have no significant spatial dependency is 175m and this corresponds to subsets which contain, on average approximately 50 data points. By examining a large number of subsets containing 50 data points we find 1 in 300 of these to have no significant spatial dependency. Therefore out of the 3×10^{37} combinations there will be many which can be used within the calibration process.

2.5. Flow modelling and calibration

LISFLOOD-FP (Bates and De Roo, 2000), a simple two-dimensional flood inundation model, was used to model the December 2006 flood event. LISFLOOD-FP is a raster-based model originally designed to take advantage of high resolution LiDAR datasets. Channel flow is modelled using the diffusive wave approximation to the one-dimensional St. Venant equations solved using an implicit Newton-Raphson scheme. The channel parameters are specified at various points along the channel, these parameters are channel width, bed elevation and Manning's roughness coefficient (n) value. When combined with the floodplain DTM the bank full depth can be estimated from the bed slope, and the values of bed slope along the channel are used to define the channel slope. Floodplain flows are modelled using a storage cell method described by Cunge et al. (1980) implemented on a raster grid with fluxes between cells calculated using the inertial form of the shallow water equations described in Bates et al. (2010). A full description of the model is given in Bates and De Roo (2000), and the 1D channel diffusive solver in Trigg et al. (2009).

The model was implemented with a 12.5m resolution grid, to match the ERS-2 pixel size. The floodplain DEM was generated firstly by creating a 5m DEM from the 2m resolution LiDAR DTM to enable the missing data areas (<0.2km^2 out of the 40km^2 model domain) from the LiDAR to be patched with the IfSAR data. This 5m combined DEM was then resampled to a 12.5m cell size by taking the mean of each smaller cell value.

Model boundary conditions were prescribed using data from the gauges at Farndon Bridge (stage-derived discharge, upstream Dee), Pont-y-Capel (upstream Alyn) and Ironbridge, Eaton Hall (stage, downstream Dee). Channel geometry data were extracted from an ISIS model of the area provided by the Environment Agency, with channel widths checked using OS Digimap data and aerial photography.

The floodplain and channel friction parameters within the model require calibration. Here we run the model for 48 different parameter sets, with channel friction values from 0.01 to 0.15, and floodplain friction from 0.025 to 0.175. Calibration is carried out using three methods to evaluate the importance of removing spatial dependence from the process;

1) Binary comparison of the spatial pattern of flooding (Measure of Fit)

2) Water surface elevation comparison using all points in ERS_{cut} (RMSE)

3) Water surface elevation comparison using multiple subsamples of ERS_{cut} with no spatial dependency (RMSE)

3. Results

3.1. Composition of observed data for use in model calibration

It is possible to find many (50 point) subsamples of the original ERS_{cut} which have no spatial dependency, and therefore multiple random subsamples of ERS_{cut} were computed and tested using Moran's I until 1000 combination sets were found that had no significant spatial dependency. This number of subsets was determined as appropriate by repeating the analysis by using a different 1000 subsets, and a near identical outcome was found, whereas this was not the case for smaller numbers of subsamples. These 1000 combination sets enable a Monte Carlo sampling of points within the observed data. The LISFLOOD-FP model was calibrated by calculating the RMSE between the modelled water surface

elevation for each of the 48 different pairs of channel and floodplain friction parameters and every one of the 1000 subsets of observed water surface elevations individually. An average was then taken of the results. This calibration method showed an optimum parameter set of channel friction: 0.03, and floodplain friction 0.14, however, it is important to consider the variation within the subsamples of data. To assess the variation in optima choice across all 1000 subsamples we record the optimum parameter set for each. Figure 7 shows that there is considerable variation in the location of the optima within the parameter space depending on the particular subset of non-spatially dependent observed data points that is chosen. We find over 40% of the optima to be clustered at the parameter set chosen by the mean of the results (nch 0.03, nfp 0.14), however the parameter range in which at least 1 in 10 of the combination sets have an optima is nch: 0.01-0.04, nfp: 0.08-0.17.

This has important connotations for flood management, since studies might use a subsample of the original data to remove the likelihood of spatial dependency influencing the analysis. The selection of the subsample can have large influence over the final optimum parameter set chosen, and therefore over the model results themselves. It is clearly then important to understand how the uncertainty inherent in the observed data can influence assessments of flood risk, and to consider the number of subsamples required to achieve a true (consistent) average of the data.

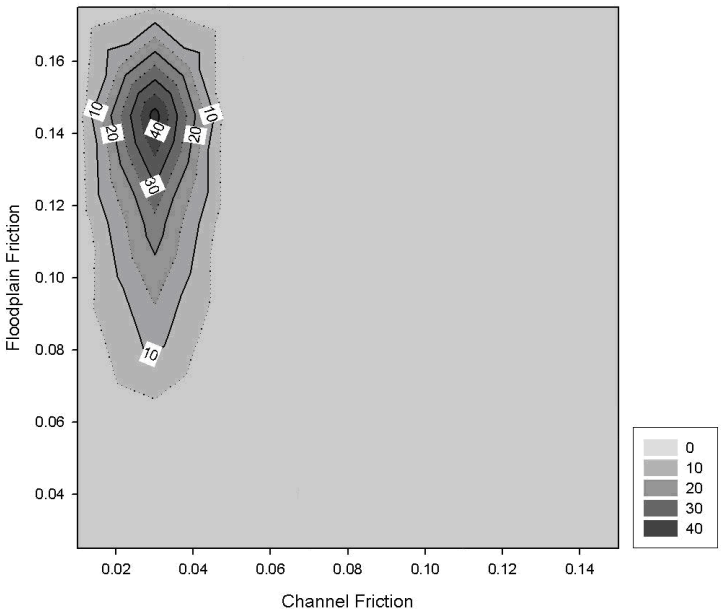

Figure 7: Percentage of combinations with optima at each friction combination

3.2. Comparison of methods of assessing model performance

Significant spatial dependency has been shown to be prevalent in the errors in the ERS-2 satellite image flood extent on the Dee in December 2006, therefore when using these data to assess model performance it is important to ensure that methods and performance measures are used that are not affected by spatial dependency. We used three methods of calibration to evaluate which methods showed particular sensitivity to the spatial dependency in this observed dataset, and to explore how the calibrated parameter space varies if a calibration method which removes spatial dependency is used.

The first method is a binary comparison of the flooded and observed extents using the Measure of Fit, F:

$$F = \frac{A}{A + B + C} \qquad \textbf{Equation 2}$$

Where A is the number of cells correctly predicted as flooded (wet in both observed and modelled image), B is the number of overpredicting cells (dry in observed but wet in modelled) and C is the number of underpredicting cells (wet in observed but dry in modelled).

The second method is the elevation comparison described by Mason et al. (2009), where the difference between the modelled and observed water surface elevations at the margins of the flood are assessed. However, here we use RMSE as a measure of model performance since parametric tests are not possible due to the data not having a significant normal distribution. The third method is that described in this study, with the elevation comparison carried out in the third method for multiple subsamples of the satellite observed data points, and an average taken of the RMSE of all of these subsamples.

Calibration using the binary pattern of flooding is clearly negatively influenced by spatial dependency in the observed data, with no clear optimum within what would be expected to be the parameter space (Figure 8: Calibration using comparison of the binary pattern of flooding (Measure of Fit %)Figure 8). This acute sensitivity to spatial dependency is not shown by the elevation comparison containing the spatially dependent data (Figure 9), which does highlight optima within the expected parameter space. In contrast to the binary pattern matching, elevation comparisons do not weight errors on lower gradients preferentially. The contrast between the two is likely to be due to the impact of different gradients across the catchment and how these gradients and spatial clusters of errors interact. Small errors in elevation on low gradients will lead to larger spatial errors than the same elevation errors on high gradients. Therefore spatial comparisons inadvertently weight errors on lower gradients preferentially.

This gradient influence explains the parameter space in Figure 8, which favours extremely high channel and floodplain friction parameters. Large overestimations in the observed flood extent to the west of the image dominate the calibration due to the low gradient of topography in their location (see Figure 5), with the model attempting to replicate this by increasing both channel and floodplain friction to physically unrealistic levels. As the water levels required by the observed data are impossible to attain given the flows entering the river there is no optimum parameter set found within the parameter space. A direct elevation comparison thereby avoids this gradient-related weighting, however, given the spatial dependency found in the satellite observed data it is important to assess if this has an effect on the elevation comparison. The calibration method developed here to remove spatial dependency (Figure 10) shows similar results to the elevation comparison using all points, however the lowest RMSE is 4cm lower, and this difference appears to be reasonably systematic across the parameter space.

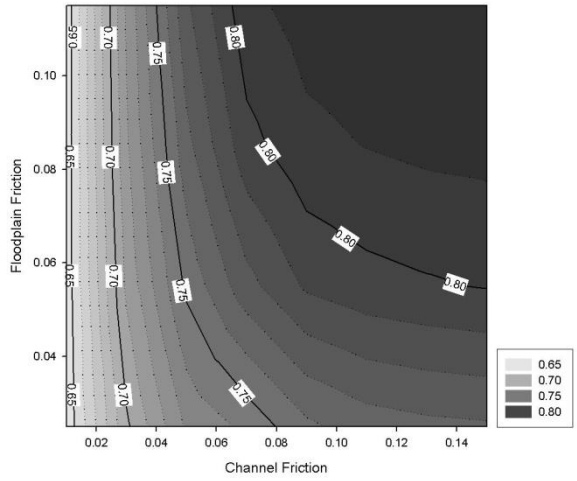

Figure 8: Calibration using comparison of the binary pattern of flooding (Measure of Fit %)

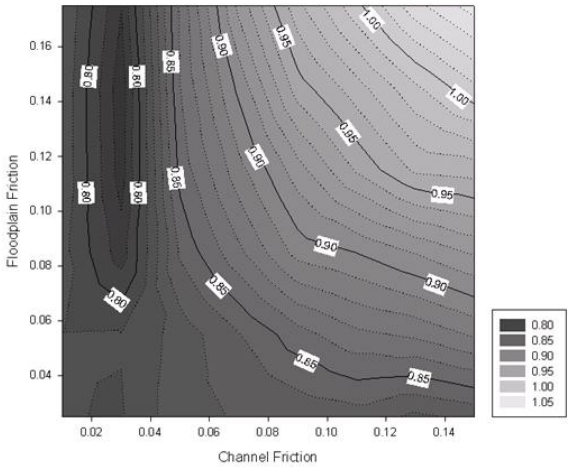

Figure 9: Calibration using comparison of all water margin elevations (RMSE)

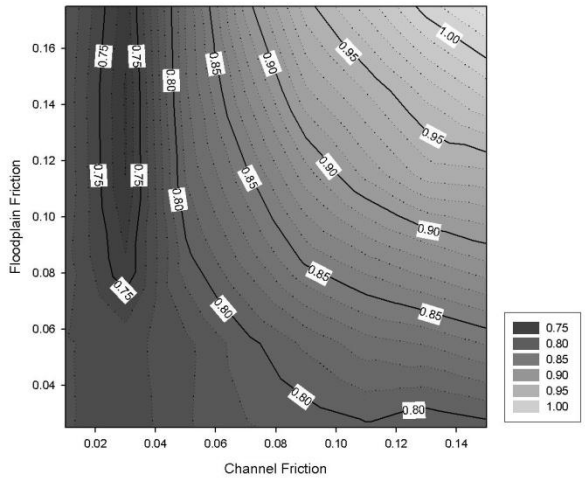

Figure 10: Calibration using multiple subsamples of water margin elevations (RMSE)

A closer comparison of the calibration using all points and using spatially independent subsamples suggests subtle differences between the two results. Firstly, if the RMSEs

across each parameter set are analysed, the lowest RMSE is lower by 4cm for the subsample method (0.88m and 0.84m), and the range in values is higher by 3cm (0.27m to 0.30m). This leads to a slightly better defined parameter space, with the subsample method being more sensitive to changes in channel and floodplain friction. A comparison of these two methods is visualised in Figure 11, with the inclusion, as a control, of results for when the subsample method is used but spatially dependent results are still included. This plot shows that when spatial dependence is still present in the data the 'all points' and subsample methods produce near identical results. The difference seen in the RMSE values can therefore be attributed to the removal of spatial dependency rather than the subsample method itself. The median value of the multiple subsample method is slightly lower than the interquartile range of the methods that have spatial dependency.

By using multiple spatially independent subsamples we find a result that provides similar benefits as a method of calibration, i.e. in terms of sensitivity to friction, but is statistically more robust. By using multiple subsamples of data we avoid any sampling bias towards the clusters of points identified previously, therefore giving us more confidence in the calibration results and the RMSE values calculated. Whilst the difference seen for the Dee catchment is minimal, it could be pronounced for different catchments where the spatial dependency is greater.

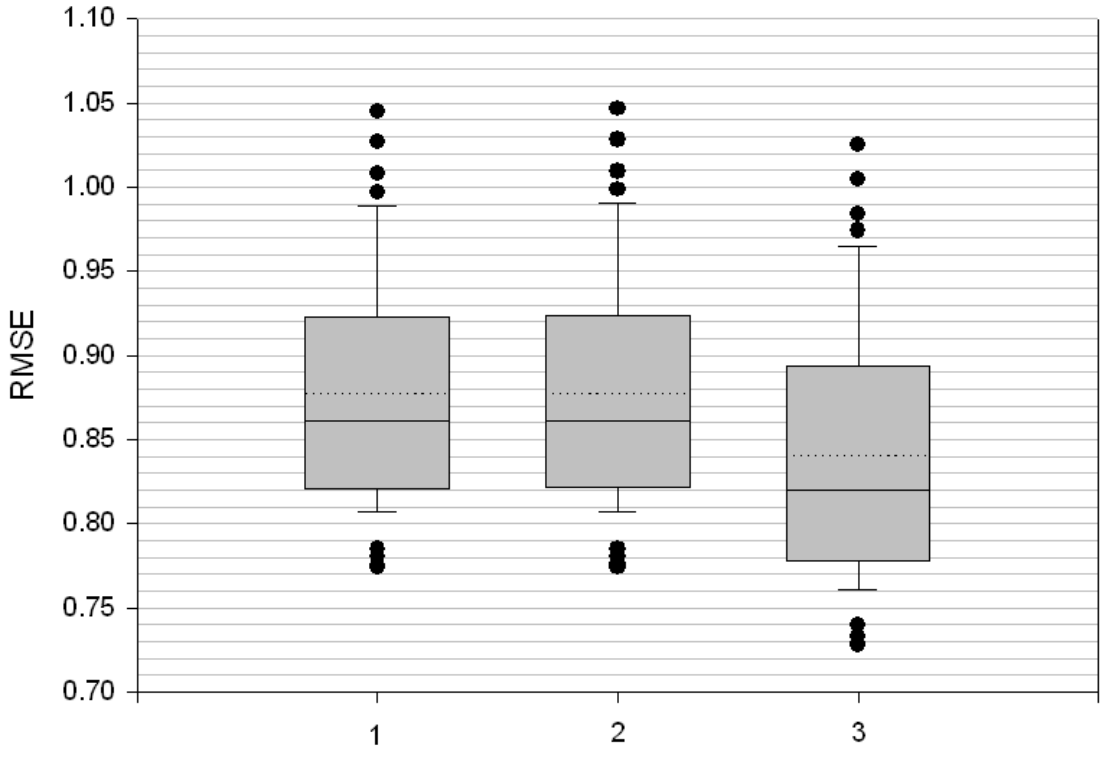

Figure 11: Distribution of RMSE across all of the parameter sets for 1) the method using all points; 2) the subsample method (with spatial dependency) and 3) the subsample method with spatial dependency removed.

A further assessment of the difference between these two methods is to compare the uncertain flood map when parameter uncertainty is assessed using the GLUE technique. Aronica et al. (2002) extended the generalized likelihood uncertainty estimation (GLUE) (Beven and Binley, 1992) technique to estimate spatially distributed uncertainty in models that are conditioned using the binary pattern of flooding extracted from the satellite data.

Modelled simulations using different parameter sets were weighted in a probabilistic assessment of flooding based on their ability to match an observed binary flood extent. Using GLUE it is possible to calculate (and then map) the probability (P_i^{flood}) that a given pixel is inundated.

$$P_i^{flood} = \frac{\Sigma_j f_{ij} W_j}{\Sigma_j W_j}$$

Where j is the number of model simulations, f is the flooded state of the pixel (1 = wet, 0 = dry). In this study we weighted the different parameter sets based on the RMSE,

$$\frac{}{}$$

Where W_j is the weighting of a parameter set, $RMSE_j$ the RMSE calculated for that parameter set, and $RMSE_{min}$ the lowest RMSE for all the parameter sets. Since we do not use the RMSE to discriminate non-performing parameter sets there is little difference between the P_i^{flood} maps for the original and subsample methods, with a maximum difference in the probability of a cell being flooded of 1%.

3.3. Uncertainty in observed data

Whilst the method described by Aronica et al. (2002) looked at parameter uncertainty, it is important to address the uncertainty in the observed data. The observed data points have been shown to have considerable variation from a smoothly varying water surface gradient. By using the variations found in the subsamples of the observed data, and incorporating this variation within a GLUE-style framework it is possible to assess the effect of observational data uncertainty on the assessment of flood risk.

Firstly, multiple combination sets of 50 points from the original 142 points are selected. For each of these combination sets the RMSE between the observed points and each modelled parameter set is calculated. Then, for each combination set the optimum parameter set is determined, and how often each parameter set is recorded as the optimum is counted. This count can then be converted to a percentage and used as the weighting within a GLUE-style uncertainty framework (as calculated in Figure 7) to determine the uncertainty in the observed data. Therefore, where a parameter set never appears as an optima W_j=0 and so it has no influence upon the uncertainty estimation. For example, a parameter set which appears in 1% of all observed data combinations will have a 40th of the influence upon one which appears in 40% of all observed data combinations. If the observed data were error-free the calculated $P_i^{observed}$ would be either 0 or 1, as each subsample would have the same optimum parameter set. The map of the probability of observed flood at each pixel is shown in Figure 12, which clearly highlights the large uncertainty around the confluence of the two rivers, where the $P_i^{observed}$ shows high uncertainty

For this study we have used only 48 combinations of friction parameters sampled at regular intervals as a means of developing this method. In practice it would be important to look at a much denser sampling of the parameter space. If a random rather than regular sampling of the parameter space were to be carried out it would be possible to calculate the weighting by sampling the number of optima found within areas or ‘bins’ within the parameter space.

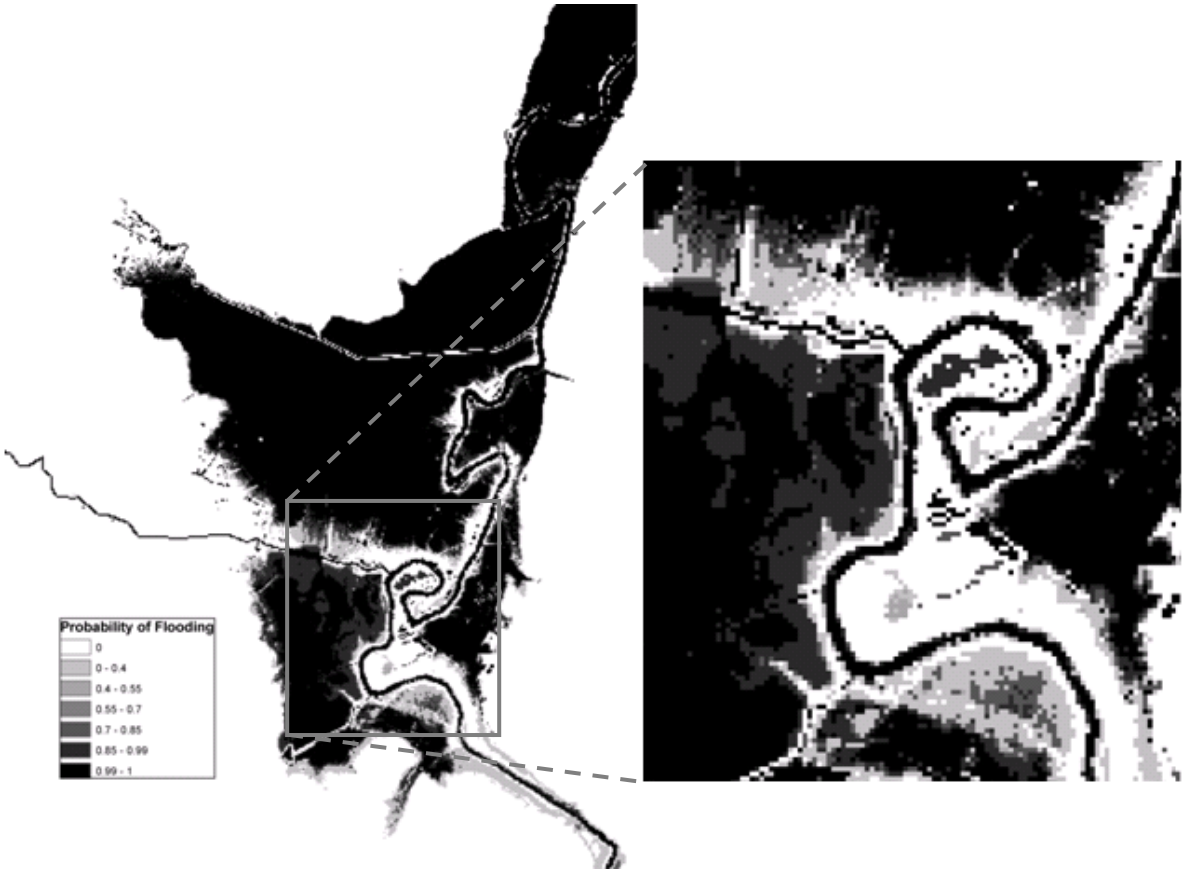

Figure 12: Probability of observed flood

4. Discussion

In this study, an analysis of the elevation comparison method used by Mason et al. (2009) underlines its robustness to spatial dependency compared to calibrating methods which compare the observed and modelled spatial extent of flooding. Further it is possible to develop the method of Mason et al. (2009) to remove the effect of spatial dependency on the calibration process, and therefore have more confidence in the method itself, whereas this is not possible for conventional comparisons carried out using the binary pattern of flooding (see Bates and De Roo, 2000). This extension of the Mason et al. (2009) method produces slight variations to the calibrated parameter space compared to the original method, but has the further benefit of enabling the observed data points to be sampled at random and the variation in these samples assessed. From this it is possible to map the effect that the uncertainty within the observed data has on the probability of flooding across the catchment.

We have shown that using different subsamples of the observed data can lead to large differences in the optimum calibration of the model. This highlights the importance of not choosing only one ‘optimum’ parameter set when assessing the risk of flooding, but using multiple subsamples to assess the range of likely friction parameter sets, and then mapping the probability of flooding based on this uncertainty within the observed data.

5. Conclusions

Spatial dependency in the observed data can have a considerable influence on the calibration process, with comparisons using the spatial pattern of flooding particularly

negatively influenced by this. For a robust statistical analysis spatial dependency can be removed by using a subsample of the data where the points are spaced further apart, this more robust method gives us more confidence in the results of the calibration. It is important to note that different subsamples can lead to wide variations in the calibration of an optimum parameter set. It is therefore necessary, as a minimum, to use multiple subsamples of the original observed dataset to ensure a sampling of all points, but ideally a thorough assessment should be carried out of the variation in the calibrated parameter space produced by all the subsamples.

This study has highlighted that for the Dee / Alyn catchment uncertainty in observed data leads to large potential differences in the predicted flood map in the area of the confluence. This has important implications for the testing and validation of hydraulic models, since the uncertainty in the observed data in these areas will not enable an adequate discrimination of model performance of this confluence feature. Further studies should aim to further understand and constrain the uncertainty in satellite images of flooding to enable a more precise model validation.

References

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.

Bashford, K.E., Beven, K.J., Young, P.C., 2000. Observational data and scale-dependent parameterizations: explorations using a virtual hydrological reality, Workshop on the Future of Distributed Modelling. John Wiley & Sons Ltd, Leuven, Belgium, pp. 293-312.

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.

Bates, P.D., Horritt, M.S., Fewtrell, T.J., 2010. A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. Journal of Hydrology, 387(1-2): 33-45.

Bates, P.D., Horritt, M.S., Smith, C.N., Mason, D., 1997. Integrating remote sensing observations of flood hydrology and hydraulic modelling. Hydrological Processes, 11(14): 1777-1795.

Beven, K., 1987. Towards the use of catchment geomorphology in flood frequency predictions. Earth Surface Processes and Landforms, 12(1): 69-82.

Beven, K., Binley, A., 1992. THE FUTURE OF DISTRIBUTED MODELS - MODEL CALIBRATION AND UNCERTAINTY PREDICTION. Hydrological Processes, 6(3): 279-298.

Beven, K.J., Kirkby, M.J., Schofield, N., Tagg, A.F., 1984. TESTING A PHYSICALLY-BASED FLOOD FORECASTING-MODEL (TOPMODEL) FOR 3 UK CATCHMENTS. Journal of Hydrology, 69(1-4): 119-143.

Cunge, J.A., Holly, F.M., Verwey, A., 1980. Practical aspects of computational river hydraulics. Pitman Publishing Ltd., London, 420pp pp.

ESA, ESRIN, 1995. ERS Thematic Working Group Meeting on Flood Monitoring, ESRIN, Frascati, Italy.

Horritt, M.S., 1999. A statistical active contour model for SAR image segmentation. Image Vis. Comput., 17(3-4): 213-224.

Horritt, M.S., Mason, D.C., Luckman, A.J., 2001. Flood boundary delineation from Synthetic Aperture Radar imagery using a statistical active contour model. International Journal of Remote Sensing, 22(13): 2489-2507.

Marks, K., Bates, P., 2000. Integration of high-resolution topographic data with floodplain flow models. Hydrological Processes, 14(11-12): 2109-2122.

Mason, D.C., Bates, P.D., Dall' Amico, J.T., 2009. Calibration of uncertain flood inundation models using remotely sensed water levels. Journal of Hydrology, 368(1-4): 224-236.

Mason, D.C., Horritt, M.S., Dall'Amico, J.T., Scott, T.R., Bates, P.D., 2007. Improving river flood extent delineation from synthetic aperture radar using airborne laser altimetry. Ieee Transactions on Geoscience and Remote Sensing, 45: 3932-3943.

Melack, J.M. et al., 2004. Regionalization of methane emissions in the Amazon Basin with microwave remote sensing. Global Change Biology, 10(5): 530-544.

Moran, P.A.P., 1950. Notes on Continuous Stochastic Phenomena. Biometrika, 37(1/2): 17-23.

Rasid, H., Pramanik, M.A.H., 1990. Visual Interpretation of Satellite Imagery for Monitoring Floods in Bangladesh. Environ. Manage., 14(6): 815-821.

Sanyal, J., Lu, X.X., 2004. Application of Remote Sensing in Flood Management with Special Reference to Monsoon Asia: A Review. Natural Hazards, 33(2): 283-301.

Schumann, G., Di Baldassarre, G., Alsdorf, D., Bates, P.D., 2010. Near real-time flood wave approximation on large rivers from space: Application to the River Po, Italy. Water Resources Research, 46.

Schumann, G., Di Baldassarre, G., Bates, P.D., 2009. The Utility of Spaceborne Radar to Render Flood Inundation Maps Based on Multialgorithm Ensembles. Ieee Transactions on Geoscience and Remote Sensing, 47(8): 2801-2807.

Smith, L.C., 1997. Satellite remote sensing of river inundation area, stage, and discharge: A review. Hydrological Processes, 11(10): 1427-1439.

Trigg, M.A. et al., 2009. Amazon flood wave hydraulics. Journal of Hydrology, 374(1-2): 92-105.

Usachev, V.F., 1983. Evaluation of flood plain inundations by remote sensing methods. In: Goodison, B.E. (Ed.), Hydrological applications of remote sensing and remote data transmission. International Association of Hydrological Sciences, Wallingford, Oxon.

Yu, D., Lane, S.N., 2006. Urban fluvial flood modelling using a two-dimensional diffusion-wave treatment, part 1: mesh resolution effects. Hydrological Processes, 20(7): 1541-1565.