Assessing the reliability of probabilistic flood inundation model predictions

Article

Accepted Version

Assessing the reliability of probabilistic flood inundation model predictions of the 2009 Cockermouth, UK

Elisabeth Stephens
School of Archaeology, Geography and Environmental Sciences
University of Reading, Reading, RG6 6AB
elisabeth.stephens@reading.ac.uk

Paul Bates
School of Geographical Sciences, University of Bristol
University Road, Bristol, BS8 1SS

December 2, 2014

Abstract

An ability to quantify the reliability of probabilistic flood inundation predictions is a requirement not only for guiding model development but also for their successful application. Probabilistic flood inundation predictions are usually produced by choosing a method of weighting the model parameter space, but this choice leads to clear differences in the prediction and therefore requires evaluation. However, a lack of an adequate number of observations of flood inundation for a catchment limits the application of conventional methods of evaluating predictive reliability. Consequently, attempts have been made to assess the reliability of probabilistic predictions using multiple observations from a single flood event.

Here, a LISFLOOD-FP hydraulic model of an extreme (>1 in 1000 year) flood event in Cockermouth, UK is constructed and calibrated using multiple performance measures from both peak flood wrack mark data and aerial photography captured post-peak. These measures are used in weighting the parameter space to produce multiple probabilistic predictions for the event. Two methods of assessing the reliability of these probabilistic predictions using limited observations are utilised; an existing method assessing the binary pattern of flooding, and a method developed in this paper to assess predictions of water surface elevation. This study finds that the water surface elevation method has both a better diagnostic and discriminatory ability, but this result is likely to be sensitive to the unknown uncertainties in the upstream boundary condition.

1 Introduction and Objectives

Broadly speaking, there are two different philosophies to uncertainty estimation in flood inundation (hydraulic) modelling; these are Bayesian approaches that use formal likelihood measures, and the Generalized Likelihood Uncertainty Estimation (GLUE) methodology, applied to hydrological predictions by Beven and Binley (1992) which uses pseudo-likelihood functions instead of formal likelihood functions.

The majority of flood inundation studies have used GLUE-based approaches (e.g. Romanowicz *et al.*, 1996; Romanowicz and Beven, 1998; Aronica *et al.*, 1998, 2002; Romanowicz and Beven, 2003; Bates *et al.*, 2004; Werner *et al.*, 2005; Horritt, 2006; Pappenberger *et al.*, 2007a,b; Schumann *et al.*, 2008; Di Baldassarre *et al.*, 2009b), although some studies have adopted Bayesian approaches, (see Romanowicz *et al.*, 1996; Hall *et al.*, 2011). These studies have addressed one or more of the types of the uncertainty in the modelling; model structural choice (e.g. Apel *et al.*, 2009), model friction and conveyance parameters (e.g. Aronica *et al.*, 1998; Romanowicz and Beven, 2003; Bates *et al.*, 2004; Werner *et al.*, 2005; Pappenberger *et al.*, 2007a), boundary conditions (e.g. Pappenberger *et al.*, 2006, 2007a), and the geometry of the floodplain (Werner *et al.*, 2005) and channel (e.g. Pappenberger *et al.*, 2006, 2007a) (including the representation of natural and man-made flow control structures such as vegetation and buildings (Beven *et al.*, 2012)), as well as the observed data used to condition the models (e.g. Pappenberger *et al.*, 2007a; Di Baldassarre *et al.*, 2009b).

The dominance of GLUE-based approaches perhaps reflects an acceptance of the 'effective' nature of the parameter values used in most inundation models; sub grid scale processes as well as unrepresented boundary condition and structural uncertainties are lumped into the parameterisation. It is usual that conditioning of model parameters on observed inundation data is used to produce uncertain predictions (e.g. Romanowicz and Beven, 2003; Pappenberger *et al.*, 2007b,a; Mason *et al.*, 2009, (among others)), with various pseudo-likelihood functions in use to weight the model parameters based on their agreement with these observed data.

In Stephens *et al.* (2012) a LISFLOOD-FP hydraulic model of the River Dee, UK was calibrated and uncertain flood inundation maps were produced using different performance measures to weight each parameter set. It was shown that the choice of performance measure for weighting the parameter space leads to differences in the final uncertain flood inundation map, with there being clear differences between a new uncertain measure (that implicitly takes into account the uncertainty in the observed water surface elevations), the RMSE and the Measure of Fit (Critical Success Index) used in studies such as that of Aronica *et al.* (2002). In this study the Measure of Fit will be referred to as the Critical Success Index as recommended by Stephens *et al.* (2014) to keep the terminology consistent with other disciplines.

Given the clear differences between uncertain flood inundation maps depending on how they are produced, there is a clear requirement for improving the ability to assess and quantify their reliability. This paper therefore focusses on the evaluation of uncertain flood inundation maps. In particular, two different methods are used to evaluate their reliability; the first method is that of Horritt (2006), and the second method is developed to account for the reliability of water surface elevation predictions (rather than the probability of a grid cell being wet / dry). Using these two different methods the reliability of the uncertain flood inundation maps and water surface elevation predictions produced using different methods of weighting the parameter sets is evaluated.

In this study the 2009 Cockermouth flood event on the River Derwent, UK is used as a case study. This allows for the method developed by Stephens *et al.* (2012), and the associated conclusions, to be tested on a different catchment, and is also a data-rich case study with a high spatial resolution (0.15m) aerial photography image that shows both the flood extent at the time of the photograph and enables identification of wrack marks to indicate water levels at peak flood.

1.1 Current methods for probabilistic evaluation of probabilistic flood inundation models

As Horritt (2006) notes, evaluation of a deterministic model prediction using data from a single event should be relatively straight forward (assuming any observed data of the flood to be perfect or the error distribution to be well constrained), but evaluation of uncertain model predictions is more problematic. Probabilistic evaluation of weather models is commonplace since ensemble forecasts have been used routinely since 1993 (NRC, 2006). This evaluation is largely enabled by a wealth of data as, for example, predictions of weather are made and realised on a daily basis. However, floods are rare events and consequently evaluating uncertain flood inundation model predictions using a (very) limited number of observations is problematic (Horritt, 2006).

Despite this, it is important for the applicability of probabilistic predictions to be able to state their accuracy: does an 80% chance mean that the event occurs

80% of the time? Therefore, even if the requirements of the formal probabilistic evaluation methods used in fields such as meteorology cannot be met because of data limitations, attempts should be made to evaluate probabilistic predictions using the few data that are available. Accordingly, modellers of extreme events and climate change, who have similar data limitation issues, have proposed the use of spatial patterns of predictions and outcomes to build sufficient datasets for evaluation (Horritt, 2006; Annan and Hargreaves, 2010).

Horritt (2006) proposed a method to validate inundation model predictions using a single observation of flood extent (hereby referred to as the Horritt method), in effect, aggregating observations of the flooded state within each grid cell to produce a large enough sample size. A LISFLOOD-FP model (Bates and De Roo, 2000) of a reach of the River Severn was set-up, and calibration / validation data were provided by two SAR images of flood events in October 1998 and November 2000. The model was calibrated using one dataset, and validated using the other, therefore allowing for some independence between model calibration and evaluation.

Horritt (2006) proposed that uncertain flood maps produced using multiple simulations that are weighted using different model parameter sets should be classified into regions of similar probability. By counting the number of observed wet cells in each of these regions it is possible to calculate reliability and visualise it using a reliability diagram. A perfectly reliable prediction would be one where, for a region of cells of similar inundation probability, the percentage of wet cells in this region is equal (or similar) to that probability. For example, if 15% of cells in the region characterised by 10-20% inundation probability are observed as flooded then this prediction could be considered reliable. The reliability can therefore be calculated as an average of the differences between the average forecast / predicted probability and the observed probability, and would take a value of 0 for a perfectly reliable forecast.

Although the Horritt (2006) paper maintains separation between the calibration and validation data, the Horritt method does not account for the co-dependence between the observations used in the analysis. For example, it is likely that if one cell on the floodplain has a predicted inundation probability of 50% and it is observed as being flooded, that any adjacent cells will have similar probabilities and observations. While Horritt (2006) suggests that the issue of only having single observations has been 'neatly sidestepped', it could be argued that by using observations from the same event on the same model domain leads to issues of co-dependence that could potentially bias the analysis.

To increase independence of observations it would be necessary to choose a subset of cells across the domain that are not related, and given a large enough number of cells this would be possible. However, a perhaps more sensitive and discriminatory measure might be to evaluate the water surface elevation predictions themselves, looking at where the observations fall within the predicted distribution of water depths. Unlike the Horritt method, a method that used observations of water surface elevations as the evaluation dataset would not require a continuous flood extent to be recorded, and therefore could be applied where there are discontinuous measurements such as wrack lines, or where the continuity of flood outlines derived from remote sensing is limited due to dense vegetation disguising the true flood edge in particular areas.

As well as using more 'independent' observations and being applicable for a larger variety of data sources, it is hypothesised that a method that evaluates probabilistic water surface elevation predictions will be more sensitive and therefore allow for better discrimination between the performance of different uncertain flood predictions. To judge this, different performance measures are used to weight water surface elevation predictions and produce predicted water elevation distributions for points across the domain. The objectives of this paper are therefore as follows:

1. To evaluate, for the 2009 flood event in Cockermouth, what performance measure / weighting method produces the more reliable probabilistic flood inundation predictions

2. To confirm the consistency of this conclusion by comparing results for calibrating / evaluating at time of peak flood and for the time of aerial photography overpass during flood recession, again using the Cockermouth dataset.

3. To compare the method for evaluating probabilistic predictions that is developed in this paper with the Horritt method, determining whether they produce the same outcomes, and which is more sensitive and therefore better for discriminating between these different weighting methods

4. To determine what can be learnt about the model from the two different methods for evaluating probabilistic predictions

2 Methodology

2.1 Study site and test data

The study site for this paper is the River Derwent in Cumbria, in the north-west of England (see Figure 1). The River Derwent flows west from Bassenthwaite Lake towards Cockermouth, where it meets the River Cocker and then continues on its westerly path to join the Irish Sea at Workington (see Figure 2).

An extremely large flood event occurred in the catchment in November 2009 after a prolonged period of rainfall over the mountains of the central Lake District. At the Seathwaite Farm raingauge in the upper reaches of the Derwent catchment a new UK record 24-hour rainfall record of 316.4mm was established for the 24-hour period up to 00:00 on the 20th November, and estimated to have a return period of 1862 years (Miller *et al.*, 2013). Due to the prolonged period of rainfall (10mm / hour average for 36 hours) (Miller *et al.*, 2013), levels of major lakes within the region reached new recorded maxima and consequently their buffering effect on downstream flows was reduced (Miller *et al.*, 2013). Using an improved Flood Estimation Handbook flood frequency analysis Miller *et al.* (2013) estimate that the discharge return period on the Derwent at Ouse Bridge was 1386 years, and 769 years on the Cocker at Southwaite Bridge. The combined flow at Camerton, estimated by the Environment Agency (EA) as $700m^3s^{-1}$ has a return period of 2102 years, with 95% confidence limits of 507 and 17706 years (Miller *et al.*, 2013).

The re-evaluation of return periods following the flood has led to increases in the estimates of the 1 in 100 year (21% increase) and 1 in 1000 year (38% increase) flows used to produce deterministic flood inundation maps for the Environment Agency, and subsequently used for planning purposes.

Gauged flow data (see Figure 3) are available for this flood event from Ouse Bridge on the Derwent (the outflow from Bassenthwaite lake), Southwaite Bridge on the Cocker (upstream of Cockermouth), and Camerton which is approximately 6km downstream from the confluence of the Cocker and Derwent as the crow flies. The flood is modelled from 12:00 on 17th November 2009, before water levels begin to rise, to 23:45 on 23rd November 2011, where water levels are nearly back to normal levels. Flow data for the River Marron have been provided by Professor Sear of Southampton University, by rescaling the flows in the Cocker using the comparative size of the catchments. For the Ouse Bridge gauge, the EA has provided metadata to advise that the stage at the peak of the flood has been edited using estimates of the maximum flood level from a wrack survey, with the time of peak and the infilled data estimated using correlation techniques. Further, for the conversion to flow data using a rating curve the Quality flag is given as 'Estimated' and 'Extrapolated Upper Part'. For the Southwaite gauge, the stage data is assigned a quality of 'Good' throughout, with approximately 17 hours at the peak of the flood where the information has been edited to use the back up data from the gauge due to float and weight issues that caused slight differences in the hydrograph. Accordingly, the Quality flag of the flow data is given as 'Good' throughout, and within the range of the rating curve for all but the 30 hours around the peak flood, where the data has been extrapolated.

The Camerton gauge was severely damaged during the event, with 'Good' readings only recorded up to 19th November 2009 at 20:30 (68.5 hours into the modelled flood). After this, the only available data are through correlation with the Southwaite gauge. The EA metadata also suggests that the river channel became 18m wider at the site of the Camerton gauge, thereby rendering useless the rating curve that existed for the site. For this study we ignore the data from the Camerton gauge, but make use of the data from the other gauges. Although the metadata reports show that there are some quality issues with the flow record for this flood, these are typical for such a large event. Ideally the uncertainty in the gauged data should be accounted for, however, this was considered as outside the scope of this paper, which aims to develop methods for assessing reliability, addressing in particular the different methods of weighting the parameter space examined in Stephens *et al.* (2012). Significant further work is required to look at the data in more detail to examine how to place upper and lower limits on the uncertainty envelope for the rating curve for an event such as this with a flow of twice the size of the next largest flood event. The implications of this boundary condition uncertainty are considered when drawing conclusions from this study.

LiDAR elevation data at 2m resolution are available for the reach from the Ouse Bridge gauge to a few kilometres downstream of the former Camerton gauge (see Figure 2). The Digital Elevation Model (DEM) used in this study is an almagamation of data from flights in 1998 and April / May 2009, with the majority sourced from a dataset collected in 1998. LiDAR data of this resolution from 1998 have a vertical Root Mean Square Error (RMSE) of approximately 0.25m (personal communication with Al Duncan, EA). The channel bed elevations have been burnt into the DEM using ground survey information from a 1D hydraulic model of the catchment provided by the EA.

Aerial photography of the flood is provided by the EA (see Figure 4 for an area of the image). According to the metadata provided the flight took place between 13.10 and 14.50 on November 20th, so for the purpose of comparing to model results the time is taken as 14:00, (86 hours into the flood event as modelled). These data have a horizontal resolution of 15cm. An outline of a flood extent derived from the aerial photography was provided by the EA, and this was edited using the imagery as a reference to improve its precision, and then converted to points. This dataset of points has then been cut down by removing points which would likely be erroneous (such as at the boundary of, or underneath, dense vegetation), as well as next to walls or other vertical features where an accurate delineation of the elevation at the edge of the flood could not be achieved. This results in a total of 3724 data points. Well defined wrack marks are visible along much of the extent of the flood in the aerial photograph (see Figure 5). Manual digitisation of these marks has provided a total of 177 maximum water elevations, intersected with the LiDAR topographic data to provide maximum water surface elevations for further comparison with model results. The aerial photography data will provide a stern test for the model on the falling limb of the flood. At the time of aerial photography overpass, flows still remained out of bank (as can be seen from the imagery), and so the floodplain is not considered to be draining at this point. However, it is worth noting that coarse resolution models have been shown to be poor at draining the floodplain (Bates *et al.*, 2006; Wright *et al.*, 2008; Neal *et al.*, 2011).

While in many studies aerial photography is used as a benchmark to assess the accuracy of satellite observed flood extents (Horritt *et al.*, 2001; Mason *et al.*, 2007), thereby assuming it to be accurate and precise, here this assumption is not made since these data will contain unknown errors. This is demonstrated in Figure 6, where there is obvious deviation from a smooth water surface for what should be an easy 200m stretch of floodplain to delineate the flood extent from. These deviations from a smooth water surface will be from two sources; the first being geolocational errors in the (manual or automatic) demarcation of the outline and the geocorrection of the data, and the second; errors in the LiDAR data used in the intersection of the flood extent and the topography. While it could be argued that the deviation would be smaller if the points were better digitised, these points have already been manually repositioned from the data as provided by the EA, and consequently any better recorrection of these 2000+ data points would be a significant time burden. Also, and as can be seen in Figure 6, there is some confusion over whether the edge of the water surface lies at the edge of the sediment-laden area of water, or whether it lies at the edge of the surrounding darker area of vegetation which could be the current flood level, emergent vegetation or simply wet vegetation that has been previously flooded. Further, the vertical height errors that are incorporated with the intersection with the LiDAR data could be in the region of 0.25m RMSE, and cannot be removed.

2.2 Model Set-Up and Calibration

A 2D LISFLOOD-FP model was set-up using the inertial formulation of the shallow water equations as decribed by Bates *et al.* (2010). The model incorporates the LiDAR topographic data outlined above rescaled to 20m resolution to enable multiple simulations to be run without unreasonable computational cost, and the gauged data as upstream boundary conditions. The gauged data for Camerton have not been used as a downstream stage-varying boundary condition due to the known poor data quality. Instead a free boundary condition has been imposed using test runs of the model to approximate the water surface slope at this part of the catchment, which was shown to vary slightly from the local valley slope. The model is run for 167.75 hours, from 12.00 on 17th November 2009 to 23:45 on the 23rd November 2009, across a domain 100km^2 in size (including No Data cells). A simulation of the model run on 4 processors of the University of Bristol's Blue

Crystal supercomputer takes between 1.5 and 2 hours depending on the friction parameters used, and the model runs with very small mass balance error.

The upland nature of the upper Cockermouth catchment means that channel friction values might be higher than lowland rivers such as the Dee due to a gravel bed, and consequently, floodplain friction values may possibly be lower than those for the channel due to the pastural land use which dominates the floodplain across the catchment. While it is expected that parameter values are effective, physically-based parameter ranges can be used to define the parameter space. According to Chow (1959) pasture with short grass would have a minimum Manning's n of 0.025, and a gravel bed would have a minimum of 0.030. Some areas of the catchment are heavily forested or have medium to dense brush, which might be expected to have a maximum Manning's n value of 0.12 (Chow, 1959). To ensure that the entire range of potential friction values are sampled, but also accepting that friction as specified in LISFLOOD-FP also acts as an 'effective' parameterisation (to account for unrepresented model structures such as sub-grid scale topographic features, and also unquantified uncertainties such channel topography and input flows), the parameter space is defined by channel and floodplain friction values of between 0.02 and 0.14. Calibration of the model was carried out by randomly sampling 300 parameter sets from the parameter space.

Four different measures are used to assess the performance of each of the three hundred parameter sets. The first is the water surface elevation comparison described by Mason *et al.* (2009), which is simply the Root Mean Square Error (RMSE) between the DEM elevation at each point on the observed flood margin, and the nearest water surface elevation in the model. If the cell that the observed point occupies is not flooded in the model, then an algorithm looks around adjacent cells (and then at cells of an increasing distance away) to this point until the water surface elevation is found. If multiple cells of an equal distance to the observed data point have a water surface elevation value then the value of the cell with the closest DEM elevation to the observed data point will be used. The second performance measure is the binary Critical Success Index (CSI):

$$CSI = \frac{A}{A+B+C} \tag{2.1}$$

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 third performance measure, Perc_50 is the percentage as optimum measure detailed in Stephens *et al.* (2012), developed to provide an (implicit) representation of the uncertainty in the observed data into the calibration process. For this measure, ten thousand subsets of fifty points are taken from the observed dataset, and the parameter set which produces the lowest RMSE for each subset is recorded. The frequency for which each parameter set occurs as the optimum is calculated, and converted into a percentage of the total number of subsets that have been evaluated.

The fourth performance measure, Perc_1 is similar to the third, except that it uses subsets of 1, i.e. just individual data points, and then records the optimum parameter set for each of the individual points. Again, the frequency for which each parameter set occurs as the optimum is recorded, and turned into a percentage of the total number of subsets that have been evaluated. It was decided to additionally use this measure (compared to Stephens *et al.* (2012)), since by sampling each point it may be possible to implicitly account for the full range of observed data uncertainty, with no averaging over observation errors. For example, a single observed water surface elevation, will contain some unknown uncertainty due to LiDAR data errors and potentially geocorrection errors when intersecting the observed outline with the topographic data, but provided that enough data points are used, the LiDAR topographic errors and any geolocational errors will be accounted for by combining the results from all of these points to look at the effect of the uncertainty on the modelled parameter space. This assumes that the errors are random rather than systematic.

The Perc measures allow for areas of the parameter space to be rejected, thereby acting as a behavioural threshold. One criticism of this measure could be that a model could be rejected by using this measure even if its performance compared to an optimal model could not be differentiated from the [estimated] observational error. There is no averaging of the observation errors in Perc_1, and so it provides an alternative approach to model rejection. To test whether it is this rejection criteria that influences reliability, or the measure itself, two more weighting methods are used based on a simple adjustment of the RMSE and CSI weightings. These RMSE* and CSI* inundation maps are constructed using a simple adjustment of the RMSE and CSI weightings by setting all weightings for the RMSE and CSI measures to 0 for parameter sets that are deemed non performing from the Perc_1 measure.

Other studies have represented the uncertainty in observational data more explicitly; Pappenberger *et al.* (2007a) use a fuzzy map of flood extent and a global fuzzy performance measure, and Di Baldassarre *et al.* (2009b) produced a 'possibility of inundation map' by looking at how the model calibration varies when different methods of determining the flood outline from two different SAR images of a flood event are used. However, these existing studies have focussed on the uncertainty in the pattern of flood extent. Such contingency table based performance measures have been shown to be problematic for model calibration given their sensitivity to spatial variations in topographic gradient (Stephens *et al.*, 2014), as such, research efforts should focus on the use of water surface elevation observations instead. Some studies have used an explicit representation of the uncertainty in satellite-derived water surface elevations for predicting flood wave propogation using a 1D model (Di Baldassarre *et al.*, 2009a) and discharge (Neal *et al.*, 2009), but this has yet to be addressed for (2D model) predictions of the pattern of flood inundation.

There is certainly a requirement for future inundation modelling studies to address explicit representations of uncertainty in water surface elevation observations, and these should also be tested using assessments of reliability. This was considered to be outside the scope for this study, as it would require a considerable amount of discussion on how best to address the multiple sources of error in the observed data, such as the affect of wind on the deposition of wrack marks or on the reflectance of the water surface for SAR imagery, error due to LiDAR resampling or registration errors in remotely sensed imagery. Accordingly, this study focusses on the behaviour of the Perc measures in comparison to the Critical Success Index and RMSE.

2.3 Probability of inundation maps

The generalized likelihood uncertainty estimation (GLUE) technique of Beven and Binley (1992) has been extended to estimate spatially distributed uncertainty in models that are conditioned using the binary pattern of flooding extracted from satellite data (e.g. Romanowicz *et al.*, 1996; Aronica *et al.*, 1998, 2002; Romanowicz and Beven, 2003). An ensemble of the model is run with each ensemble member using a different parameter set. These ensemble members are weighted in a probabilistic assessment of flooding based on their ability to match an observed binary flood extent. While these earlier studies conditioned uncertain predictions based on the model's ability to match the binary pattern of flooding, Mason *et al.* (2009) detailed how the weighting could also be based on a model's ability to match a set of observed water surface elevations, and Stephens *et al.* (2012), extended this water surface elevation comparison to use multiple subsets of these observed data. This percentage as optimum performance measure converts easily to a weighting because it sums to a percentage.

For the RMSE and CSI measures, parameter sets are weighted based on how they perform on a sliding scale from the best performing parameter set (weighting=1) to the worst performing parameter set (weighting=0). For example:

$$Weighting = \frac{RMSE_p - RMSE_{min}}{RMSE_{max} - RMSE_{min}} \qquad (2.2)$$

Using the GLUE procedure extended by Aronica *et al.* (2002) it is possible to calculate and then map the probability (P_i^{flood}) that a given pixel is inundated.

$$P_i^{flood} = \frac{\sum_j f_{ij} W_j}{\sum_j W_j} \tag{2.3}$$

Where j is the number of model simulations, f is the flooded state of the pixel (1 = wet, 0 = dry) and W_j is the weighting given to each model simulation.

2.4 Methods for evaluation of probabilistic predictions

Stephens *et al.* (2012) showed how these different methods of calculating the P_i^{flood} in each cell led to clear differences in the uncertain flood inundation maps produced. Consequently it is important to be able to evaluate how the use of different weighting methods influences predictive skill. It is possible to carry out such an evaluation by assessing the reliability of model predictions. Detailed below are two different methods of evaluating the reliability of uncertain flood inundation maps used for this study.

2.4.1 Assessing reliability using the Horritt method

A reliability diagram allows for a visual assessment to be made of whether the model is over or underestimating probabilities, by plotting the predicted probability on the x-axis, and the observed probability on the y-axis. A perfectly reliable prediction would lie on the 1:1 line. The reliability can be quantified as an average of the differences between the average forecast / predicted probability and the observed probability (Stephenson *et al.*, 2008):

$$Reliability = \frac{1}{N} \sum_{k=1}^{m} n(\bar{f}_k - \bar{o}_k)^2 \tag{2.4}$$

Where $\bar{f}_k$ is the mean of the probability forecasts of event k occurring (in each bin), and $\bar{o}_k$ is the observation of event k. N is the total number of observations, n is the number of events that fall into each bin m. Such an evaluation of reliability requires a wealth of event data which is problematic given the (very) limited number of observations of flood inundation (Horritt, 2006).

Despite this, it is important for the demonstration of the applicability of probabilistic predictions to be able to give some estimate of their reliability. Accordingly, modellers of extreme events and climate change, who have similar data limitation issues, have proposed the use of spatial patterns of predictions and outcomes to build sufficient datasets for evaluation (Horritt, 2006; Annan and Hargreaves, 2010). As such, Horritt (2006) proposed assessing reliability using the probabilities of inundation assigned to each cell.

For the Horritt method Equation 2.4 is adjusted such that $\bar{f}_k$ is the mean of the probability forecasts of a cell being flooded k (in each bin), and $\bar{o}_k$ is the observation of flooding k in each bin. N is the total number of observations, n is the number of events that fall into each bin m. Note that for the Horritt method model cells where the predicted probability of flooding = 0 are ignored in the calculation since they account for the vast majority of the domain and therefore would bias the result.

2.4.2 Assessing reliability of water surface elevation predictions

To achieve an assessment of the reliability using water surface elevation predictions rather than the probability of inundation in each cell the following methodology is proposed:

The first step is to calculate a predicted water surface elevation probability distribution for each cell, based on a weighting using the performance measures used in Stephens *et al.* (2012). It is important to sample from a large parameter space so that the limits of the probability distribution are not predetermined by a subjective choice of potential parameter sets. For observations where the modelled water surface elevation is zero an algorithm is used to search, with a increasing distance away from the observation cell, for the nearest water surface elevation. Where two cells of equal distance away from the observation contain water, the water elevation value from the cell with the closest topographic elevation to the observation cell is used.

The next step is, for each observation, to record where it lies within the predicted probability distribution. These records of observation location can be represented in a cumulative frequency plot, where the number of observations that fall within each bin of predictions is plotted. If the predictions are perfectly reliable the gradient of the line should be 1 since 10% of observations would fall within the first 10% of the probability distribution, 20% within the first 20%, and so on. Where the gradient is steeper than the 1:1 line then, in general, there has been an overestimation of the uncertainty in the model. Where the gradient is less steep than the 1:1 line there has been an underestimation of uncertainty, with observations having been made that lie outside of the predicted range.

An indication of bias within predictions, or where the full range of uncertainty has not been adequately captured, can be seen by identifying where the line intercepts with the vertical lines of x=0 (the y axis) and x=100. The intercept with the y axis is the percentage of observations that fall outside the lower bounds of the predicted probability distribution of water surface elevations. The intercept with the line x=100 can be substracted from 100 to give the percentage of observations that fall outside the upper bounds of the predicted probability distribution of water surface elevation predictions. The reliability of model predictions using this method can also be quantified using a calculation similar to Equation 2.4, by finding the difference between the expected and observed cumulative frequency of observations 2.5. For the wse reliability the cumulative reliability is calculated rather than an isolated comparison of the expected and actual number of observations in each bin to ensure that no model is penalised for bringing the probabilistic predictions back towards the expected 1:1 line. For example, if no observations fell within the first bin (0%-10% decile), then if 20% of observations fell in the (10%-20% decile), then the first bin should be penalised for a 10% difference, but the second bin should not be because it brings the overall percentage of observations in the first two bins back to the expected value. As such, for the WSE method E_m is the expected number of observations to have fallen up to and including bin m, and the O_m is the actual number of observations to have fallen up to and including bin m. If the bins were set as every 10%, then the total number of bins would be 10 and so the expected value for each individual bin inside the distribution would be 10%.

$$Reliability = \frac{1}{N}\sum^{m} n(E_m - O_m)^2 \tag{2.5}$$

3 Results

3.1 Modelled parameter space using different performance measures / data sources

Figure 8 shows the parameter space of the LISFLOOD-FP 2D model for different performance measures using the aerial photography data. The Perc measures provide well defined (perhaps spuriously precise) optimum friction values, whereas the drop-off in performance across the parameter space is less defined for RMSE and CSI. The RMSE measure (Plot a) and CSI (Plot b), show that these parameter spaces are unexpected or at least unusual compared to those for other catchments (such as the Dee), in that the model shows no real sensitivity to channel friction, only floodplain friction. This sensitivity is also seen in the calibration using the peak flood wrack mark data (Figure 9). This might be explained by putting this particular flood event into context - the flows during this extreme event are so large that the channel friction has little effect on the amount of water that flows out of bank, and also in some areas the floodplain becomes the channel as flood waters bypass river meanders. In effect, the entire valley floor is acting as a single channel unit in conveying the large flows; the channel is only a small proportion of the total flow area, and so floodplain friction is by far the dominant control on flood extent.

Optimum friction parameter sets for each measure and each dataset are shown in Table 1. For such an extreme event upstream boundary conditions are unlikely to be error-free, and as described previously, the friction parameters used in the modelling should also be considered as 'effective' given that they also compensate for subgrid scale processes. Accordingly, some deviation from physically realistic values for friction are to be expected, but a modeller that finds a 'physically realistic' parameterisation may have overconfidence in thinking that the model is robust with respect to other uncertainties. Here, the RMSE measure gives the most physically realistic floodplain friction optimum of around 0.03 for short pasture, the CSI measure finds higher than expected values, and the Perc measure does not find a well-defined optimum within the areas of the parameter space that might be considered to be physically realistic. However, it is important to assess whether these 'physically realistic' parameterisations produce reliable predictions.

It might be possible to conclude that there is no significant difference between the RMSE and CSI measures, given that the RMSE difference is less than the LiDAR data vertical error of 0.25m. However, care should be taken when drawing conclusions from averages of data. A histogram of the distribution of the two sets of model errors paints a more complete picture, giving an indication of the shift in the distribution of errors rather than just the difference between the means of each distribution. Figure 6 shows the error structure of two model parameter sets with RMSEs of 0.5624 (blue) and 0.4015 (red). It demonstrates that while the difference in RMSE is only 0.16m, a shift of approximately 0.4m would be required for the distributions to match, and this, backed up by the medians of each distribution (-0.0335 and 0.450083), is actually greater than the observed data error. Nevertheless, the observed data RMSE of 0.25m itself masks a distribution of errors, and therefore firm conclusions can not be drawn.

If a significant difference between the RMSE and CSI measures is assumed, it could be concluded that the CSI measure gives a much larger optimum value for floodplain friction than the other performance measures, while the broader pattern of non-sensitivity to channel friction remains the same. This comparison between parameter spaces can only be undertaken for the time of aerial photography overpass, since the CSI measure cannot be calculated for the discontinuous wrack marks dataset.

This optimum for higher floodplain friction parameters is investigated using a visual comparison between the observed dataset and the model output for two simulations with a fixed channel and different floodplain frictions (respectively of [0.027,0.026] and [0.027,0.057]). There are several areas across the domain where the higher floodplain friction simulation better matches a particular area of the observed extent than the low floodplain friction simulation (such as in the top right area of the catchment shown in Figure 10), but in doing so the higher floodplain friction simulation fails to match the areal pattern in nearby areas. These areas of unexpected inundation are not relics of observed data error, since there is strong agreement for multiple data points and they are clearly visible in the aerial photography. This suggests that higher floodplain friction simulation is perhaps correctly matching the observed inundation in specific areas for the wrong reasons. There are several possible explanations for the inability of the lower floodplain friction simulation to capture these flooded areas; the model may have a resolution too coarse to accurately capture bank heights, or processes not represented in the lower friction model such as bank failure might be important. Consequently, it is thought that the higher floodplain friction simulation is matching the pattern of flooding better, but for the wrong reasons.

Stephens *et al.* (2012) and Stephens *et al.* (2014) described the CSI measure's sensitivity to topographic slope, caused by it being more sensitive to correctly matching areas of the domain with low slope, where water elevation changes lead to greater changes in the areal pattern, rather than where gradients are steeper. Similarly, in this study calibration carried out using the CSI performance measure is more sensitive to (relatively) small parts of the model domain where there are large areal changes caused by tipping points (such as a bank being breached), than capturing the general pattern across the whole model domain. While for some applications it may be (more) important that the model correctly predicts these specific areas than the general pattern, caution should be exercised since the model could be capturing them for the wrong reasons or there could be observed data errors, therefore leading to a poorly calibrated model. While it is believed that for this case study the CSI might be showing the model matching the flood extent better but for the wrong reasons, it will be important to test this by evaluating the uncertain predictions produced when parameter sets are weighted using this and other performance measures.

In general there is more agreement in the form of the parameter space where the same performance measure is used for the two different datasets than between the measures themselves. This suggests that there is some consistency in parameter performance for two different times during the flood, but given that the interval between these datasets is relatively short, this consistency is less likely to occur for when flows are considerably different either during the same event or for different events.

The Perc_1 and Perc_50 plots distinguish areas of the parameter space that are non-performing, where parameter sets never appear as the optimum using multiple realisations of the observed data. Perc_50 shows (as would be expected) larger non-performing areas than Perc_1, since subsets of 50 act to average the range of uncertainty that can be represented using each individual point. The Perc measures hint that the optimum parameter sets sit to the margins of the parameter space, which suggests that the model (or at least its floodplain) contains too much water. This could be due to errors in the specification of the upstream flows, which is quite likely because of the potential errors in the gauged data detailed earlier in this paper, or alternatively due to geomorphological changes during the flood event that increased the capacity of the river channel. Such geomorphological changes can be identified in a post-flood LiDAR dataset of the event, and consequently will have some effect, although it is not possible without further modelling to be confident of whether this or incorrect upstream flows are the cause of the apparent bias in the model. Ignoring the CSI measure due to its known problems, it is interesting that the RMSE shows a well defined optimum within the parameter space, and this demonstrates the need for evaluating whether the Perc measures or the RMSE provides more reliable predictions. As mentioned earlier in this study in Section 2.4.2; it is important to ensure that the parameter space is large enough so that the limits of the predicted probability distribution are not predetermined by a subjective choice of potential parameter sets. The identification of optimum parameter sets at the margins of the parameter space for the Perc measures suggests that this may be an issue; however the lower bounds for the roughness parameters are limited by model stability rather than subjectivity, which is not untypical for hydraulic models and is not thought to affect the conclusions drawn in this study.

3.2 Uncertain Inundation Maps

The Probability of Inundation maps shown in Figure 11 demonstrate the effect that the choice of weighting method has on the mapping of flood hazard. Weighting measures that act to discard areas of the parameter space as non-performing mean that the flood margin becomes more certain / less blurred. This could lead to spurious precision, or could be an effective way of determining which parameter sets should be discarded or given low weighting: this can only be assessed by looking at the reliability of the predictions.

3.3 Reliability

A reliability plot using the Horritt method is shown in Figure 12, and the associated quantifications of this reliability can be found in Table 2. Note that the Horritt method requires a binary flood map of wet / dry areas, so can only be carried out using the aerial photography evaluation data since the wrack marks do not provide a continous boundary. Additionally, the reliability calculations for the Horritt method are strongly influenced by the number of cells predicted as having a 100% probability of flooding. Figure 12, Panel 2 does not use independent calibration and validation data, so the analysis here is focussed on Panel 1.

Figure 12, Panel 1 (calibration using wrack marks deposited at the time of peak flood) clearly demonstrates that the RMSE weighting overpredicts inundation probabilities, and that the Perc_50 method is an improvement on the RMSE, showing no bias but still some noise. As would be expected, the RMSE* method

[0.0087] performs considerably better than RMSE [0.0161] since it uses the Perc_1 method to discard non performing areas of the parameter space (parameter sets that never appeared as an optimum using multiple realisations of the observed data). Closest to the 1:1 line is the Perc_1 method [0.0070], which shows little bias or noise. There is only one non-performing point for the Perc_1 method that deviates far from the 1:1 line, and this could be due to the small number of data points in that category. Although drawing conclusions from Plot 2 should be done with caution because it uses the same dataset for calibration and validation data, it can clearly be seen that the CSI performance measure produces even less reliable predictions than RMSE.

The reliability plots using the new water surface elevation method are shown in Figure 13. In this Figure panels 1a) and 2b) use the same dataset for calibration and evaluation and so are not discussed. The WSE reliability plot for the time of flood peak (1b) reiterates the results of the Horritt method, showing that the CSI weighting produces the least reliable predictions, with RMSE also quite unreliable. These show that, on the whole, modelling using these weighting methods produces an overestimation of flood depths. The plotted line is always above the 1:1 line, showing that, in the case of CSI, 80% of observations fall within the first 20% of the predicted distribution of water depths. Discarding areas of the RMSE and CSI parameter spaces using Perc_1 enables a small improvement in reliability (RMSE* and CSI*), but the overestimation of flood depths remains. The Perc_50 method appears to have less bias than the RMSE or CSI, but should be penalised for the number of observations (approximately 20%) that fall outside the upper limit of the predicted range. The Perc_1 appears to be the best weighting method since it lies close to the 1:1 line and no observations fall outside the upper limits of the predicted WSE distribution. This conclusion is solidified by the calculated reliability shown in Table 2, where Perc_1 has clearly the best WSE reliability of 0.0133, and the RMSE* (0.1072) and CSI* (0.2120) measures do not perform better than even Perc_50 (0.0254). Markedly, the CSI measure (0.3028) has a poorer WSE reliability than an equal weighting (0.2361) would provide.

The WSE reliability plot for the time of aerial photography (2a) in general shows that the model is less reliable after the flood peak (1b) than before it, and this is backed up by an approximate halving of the (best) reliability score for Perc_1. It could also be argued that for the peak flood (1b) the model shows a tendency towards underpredicting flood depths (certainly for Perc_1), whereas for the aerial photography (2a) there is definite overprediction. Previous studies such as Wright *et al.* (2008) have shown model accuracy to diminish after peak flood, and this result is repeated for the 2009 Cockermouth event. The reliability plots used in this study suggest that the (effective) parameters used in LISFLOOD-FP modelling become less 'effective' post flood peak, in that they can no longer account for as much of the uncertainty in the modelling post flood peak. Consequently it will be important to account for these uncertainties explicitly.

It is possible to compare the Horritt and WSE reliability methods by looking at the evaluation for the time of aerial photography overpass calibrated using the wrack marks dataset (Plot 1 of Figure 12 and Plot 2a of Figure 13). While it appears at first that the two plots are 'switched' in that the points in the former lie mostly to the bottom right side of the diagonal, and in the latter the points lie to the top left, actually the plots show the same pattern. The WSE reliability plots give an indication as to what percentage of the observations have fallen within the corresponding cumulative percentile of the predicted distribution. As such, while (for example) the RMSE calibration is shown for the Horritt reliability to be overpredicting the probability of inundation, the WSE reliability plot shows that more observations than expected have occurred for a particular predicted cumulative percentile; e.g. the model has overestimated the likelihood of higher water surface elevations. The WSE reliability plot also provides additional information to the Horritt reliability plots; demonstrating the percentage of observations that fall outside the predicted distribution of water surface elevations.

It is clear that Perc_1 is the most reliable weighting method, but there is disagreement between the Horritt and WSE reliability methods over the worst performing weighting method. The WSE method suggests that it is Perc_50, but the Horritt method identifies RMSE. This is because the Horritt method does not penalise observations falling outside the range of predictions: the Perc_50 method for the time of aerial overpass shows only 60% to 70% of observations to fall within the predicted WSE distribution, and the line has a more shallow gradient than 1:1. The WSE method therefore makes clear that this Perc_50 method underestimates the full range of uncertainty, probably because it has discarded too many parameter sets as non-performing. RMSE is again quite an unreliable measure (note that there is no CSI measure for this because of the calibration using the discontinuous wrack marks), but RMSE* shows considerable improvement due to the link with the Perc_1 measure.

4 Discussion

One of the aims of this paper was to evaluate the most reliable performance measure for weighting parameter sets to produce uncertain flood inundation maps. As well as the conventional performance measures of RMSE and CSI, the Perc measure, developed in Stephens *et al.* (2012), was also used to address how observed data errors are accounted for in the calibration process. Unlike the Perc_50 measure, which uses multiple subsets of 50 data points, the Perc_1 measure records, using individual observed data points, the number of times that each parameter set appears as the optimum. This measure of agreement provides a parameter space that appears to give the best overall picture of the likelihood of each parameter set being the optimum.

Both methods of assessing model reliability show that the Perc_1 measure produces the most reliable predictions, and this result is consistent for the validation data at the time of peak flood and at the time of the aerial photography overpass. This is a surprising result as, up until now, observations are usually grouped together into a ‘global’ dataset for model calibration. While Pappenberger *et al.* (2007b) highlight the importance of a vulnerability-weighted model calibration to produce an improved local model performance, e.g. with respect to locations of critical infrastructure, we show that considering observations individually can actually improve the global performance. But RMSE, as a measure which uses an average of all the (uncertain) observed data, will be influenced by outliers. As there is no reason to discard such an outlying point (unlike points that are in densely vegetated areas), there is still a (perhaps very small) chance that it is correct, and that all other points are affected by some systematic error. Therefore with these outliers influencing model calibration, it is important that they are used proportionately.

In the Perc_1 measure an optimum parameter set that is only agreed upon by one data point will only be given a small weighting proportionate to the level of agreement, whereas for RMSE this data point will influence the characteristics of the entire parameter space. Perc_1 therefore reduces the influence of what are likely to be erroneous data points, but gives them some weighting based on their agreement with the rest of the observed dataset, such that if 10 out of 1000 observations point at a particular optimum parameter set, this parameter set will be given a weighting of 1%.

It could be argued that the Perc_1 measure should incorporate some kind of limits of acceptability approach so that a model is not rejected on this measure when its difference from an optimal model is less than the observational error. However, it is extremely rare to be able to adequately quantify the error in observations of flood extent, due not only to the availability of suitable validation datasets, but also because of the complexity of predicting the effect of wind on the deposition of wrack marks, or on the reflectance of the water surface for SAR imagery.

The Perc_1 methodology implicitly accounts for the potential uncertainty, arguably providing a different approach to acceptability rather than applying a subjective behavioural threshold based on a simple estimation of observed data uncertainty for the limit of acceptability. If there were observed data of multiple flood events on a catchment, and none showed a particular parameter set as an optimum, then this parameter set would surely be rejected. The Perc_1 measure applies this logic (albeit with assumptions) to multiple observations from the same flood event; in this approach each observation is treated as a separate observation, such that if a parameter set is never the 'optimum' the agreement or lack of in the Perc_1 measure is used to define acceptability. Ideally, this of course requires that all sources of uncertainty are accounted for, as potentially areas of the parameter space might be discarded that would otherwise be acceptable, if, for example, boundary condition uncertainty were taken into account.

Assessing reliability is a good way of testing the methodologies for defining acceptability and weighting the parameter space. In this study the focus was on the treatment of observed data for model calibration, and so the boundary condition uncertainty has not been taken into account. To provide a preliminary assessment of the sensitivity of the results described in this paper to upstream boundary condition uncertainty, a change in the hydrograph was simulated by taking / adding different amounts from the water surface elevations produced by the ensemble modelling Figure 14. These changes are commensurate with the changes seen when changing the hydrograph by a fixed percentage for a single parameter set, as indicated on the figure. The Brier reliability was recalculated for each applied change to give an indication of its sensitivity to boundary condition uncertainty. Figure 14 therefore demonstrates that if, in reality, the flows were consistently 10% lower then the choice of optimum weighting method would be different. Given that the uncertainty in the upstream boundary condition during this flood is unknown, this sensitivity urges caution when considering the robustness of these results.

Future work should explicitly incorporate boundary condition uncertainty into the analysis, as well as produce and test a methodology that incorporates a more detailed and explicit representation of observed data uncertainty, incorporating, for example, the resampling errors of the LiDAR data. Further studies are needed to confirm whether the conclusions are robust on different flood events with different magnitudes. Namely, does the Perc_1 measure produce the most reliable predictions for flood events of smaller magnitude, and can weighting using these smaller events still provide reliable inundation possibilities for extreme events such as the 1 in 1000 year return period flood? Further, would a more explicit representation of uncertainty in the observed data produce more reliable predictions?

The other main aim of this study was to develop a new method for evaluating uncertain flood inundation predictions, and then compare the results from this with those from the Horritt method. One of the advantages of the WSE method is that it can be used for discontinuous datasets (such as the wrack marks in this study), and it therefore has wider applicability. On top of this, and despite both reliability methods coming to the same overall conclusion, there are differences in the level of information provided by each that indicates that the WSE method is more discriminatory, since it produces a wider range of reliability scores, and also has wider diagnostic capabilities since it provides more information than the Horritt method. For example, the Horritt method does not show any bias when the Perc_50 measure is used, but the plots of cumulative reliability for the WSE method clearly show that this measure underestimates the range of uncertainty in the model. This underestimation is caused by discarding areas of the parameter space as 'non-performing' when they should still be taken into account when producing the uncertain estimates of flood hazard. Further, the WSE method can show whether and how many of the water surface elevation observations lie within the predicted range. If they do not, then this hints at epistemic uncertainty that needs to be addressed.

The Horritt method is poor at telling the modeller of model underprediction, and this is especially the case for cells that had a predicted probability of flooding of 0. Depending on how the domain is set up, large proportions of the cells in it would have predicted inundation probabilities of 0, including cells that lie well outside or above the floodplain. If some of these cells did in reality flood then the flooded percentage would be biased by the sheer number of cells that have a predicted probability of 0, therefore the Horritt method does not quantify how wrong these predictions are.

Similar problems can be seen for overprediction of flooding. Cells that have a probability of inundation of 1 (or perhaps even 0.9 or greater), and that are observed as flooded, may have considerably greater water surface elevations than were predicted, but this would not be recognised or penalised. The WSE method is be able to diagnose whether observations of water surface elevation fall outside the upper limit of the predicted distribution of water surface elevations. Further, it makes it possible to understand where the majority of observations lie within the predicted distribution.

Model evaluation using the WSE method has proved a useful diagnostic tool that provides more information about model performance than the Horritt method, giving an indication of the percentage of observations that fall outside the upper and lower limits of the probability distribution of water surface elevations. In the case of the Cockermouth flood it can be seen (using the Perc_1 measure which has been identified as producing the most reliable predictions), that at the time of the peak flood the model has around 12% of observations that fall below the lower limits of the range of water surface elevation predictions, which increases to around 22% at the time of the aerial photography overpass. Despite there being no other study for comparison, that 88% of peak flood observations fall within the predicted range could be considered good for a model that only takes into account parameter and observed data uncertainty, and especially for such an extreme flood event where the errors in the inflow and wrack mark data are likely to be high. The drop in model performance only a few hours after peak flood suggests that new sources of uncertainty need to be taken into account to produce a similar reliability to predictions made of the peak flood, and as mentioned previously the uncertainty in geomorphological change during the flood, or in the gauged flow data should be investigated.

Despite the apparent improvement in assessing reliability that the WSE method has over the Horritt method, this method is by no means a perfect test of probabilistic model performance. Such spatially-averaged approaches are problematic in that reliability is likely to be highly variable in space (Atger, 2003), and so an averaged estimate of reliability might hide local variations in model bias (Toth *et al.*, 2003). For example, the spatially-averaged reliability is likely to hide localised performance, for example, a perfect reliability might be recorded, but half of the domain might be overestimating probabilities and the other half underestimating them (Ferro, 2012). However, given the limited number of observations of flood inundation on a single catchment, the best that can be achieved is a careful analysis that requires a balance between achieving a sample size that is sufficient for a robust statistic, and being able to dissect localised variations in performance (Toth *et al.*, 2003).

5 Conclusions

This study aimed to determine which performance measure should be used to weight model parameter sets to produce reliable assessments of uncertain flood hazard. It was shown that the most reliable method is one that assesses the range in model performance across the parameter space by running multiple model calibrations using each of the observed data points individually. This result is in contradiction to current approaches used to map flood inundation, which generally group observed data points. However, an indicative assessment suggests that this conclusion may be sensitive to boundary condition uncertainty. Consequently it will be important to understand whether this conclusion is robust for flood events of different magnitude and in differenct locations.

This study has strong implications for the methodologies used for uncertain inundation mapping by practitioners; an uncertain treatment of observed data in the calibration process has been shown for the Cockermouth flood event to provide more reliable flood probabilities, and within or post-event surveyed water levels (where in abundance) are the best observed data to do this with because they will contain less uncertainty than water levels processed from remotely sensed extent data. In turn, these derived water levels have wider potential for use than binary maps of flood extent for model calibration and evaluation. It could be argued that these results reflect the better quality assurance carried out when processing extents to water levels, and to some extent this is true, but it is perhaps more reflective of the ability of water elevation comparisons to make better or broader use of the available data.

In assessing these weighting methods a new method of evaluating the reliability of uncertain flood inundation predictions has been developed by recording where observations lie within predicted probabilistic water surface elevation distributions. This method not only has the advantage over existing methods of being applicable for observations that are discontinuous, such as wrack marks or remote sensing images in vegetated areas, but it is also a more discriminatory technique with better diagnostic capabilities. It gives an indication of whether uncertainty is being under or over estimated, whether there is bias in the model, and also calculates the percentage of water surface elevation observations that fall within the predicted range.

Consequently, this WSE method has provided useful information about the LISFLOOD-FP model of the Cockermouth flood event. It demonstrates that, at peak flood, 88% of water surface elevation observations fall within the predicted model range, suggesting that the model does not take into account the full range of uncertainty seen in the observations (assuming the observations to be error-free), and as the 12% of observations outside the predicted range lie outside the lower limits of the distribution, the model is clearly biased towards over-predicting flood depths, and the source of this bias should perhaps be further examined. As some of the water surface elevation observations will be erroneous (for example the wrack marks could have been laid down after the peak flood), perhaps this figure is within the limits of acceptability for these data, and therefore it could be said that the model is performing well, but it would be interesting to observe how this figure might change if a higher resolution model were used, or model results were resampled onto higher resolution topography.

This study also shows model performance decreasing over the course of the flood, suggesting that the uncertainties that are not accounted for have greater influence after the flood peak. Further research could aim to improve model reliability by taking into account the uncertainties introduced into the modelling by gauged flow errors and geomorphological change, and evaluate whether different model complexities can better represent these uncertainties. It could also address how the resolution of the topographic data used in the model influences reliability, and whether improving the resolution of topographic data limits the number of observations that fall outside the predicted range of water surface elevations. Further investigation could also examine the potential for using the Perc measure as a discriminatory tool to identify subtle differences between the performance of different model structures and the benefits of including explicit representations of different sources of uncertainty.

6 Acknowledgements

The authors are extremely grateful to the Environment Agency for providing the LiDAR and aerial photography data used in this study as part of NERC Urgency Grant NE/I002219/1 awarded to David Sear at the University of Southampton. This work was supported by the European Union 'KULTURISK' project via grant FP7-ENV-2010-265280, a joint Great Western Research and Environment Agency studentship and a Leverhulme Early Career Fellowship. The authors are grateful to Matt Horritt, Hannah Cloke and two anonymous reviewers for their comments on the manuscript.

References

Annan, J. D. and Hargreaves, J. C. (2010). Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**(2), L02703.

Apel, H., Aronica, G., Kreibich, H., and Thieken, A. (2009). Flood risk analyses - how detailed do we need to be? *Natural Hazards*, **49**(1), 79–98.

Aronica, G., Hankin, B., and Beven, K. (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.

Aronica, G., Bates, P. D., and Horritt, M. S. (2002). Assessing the uncertainty in distributed model predictions using observed binary pattern information within glue. *Hydrological Processes*, **16**(10), 2001–2016.

Atger, F. (2003). Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, **131**(8), 1509–1523.

Bates, P. D. and 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., Aronica, G., and Beven, K. (2004). Bayesian updating of flood inundation likelihoods conditioned on flood extent data. *Hydrological Processes*, **18**(17), 3347–3370.

Bates, P. D., Wilson, M. D., Horritt, M. S., Mason, D. C., Holden, N., and Currie, A. (2006). Reach scale floodplain inundation dynamics observed using airborne synthetic aperture radar imagery: Data analysis and modelling. *Journal of Hydrology*, **328**(1-2), 306–318.

Bates, P. D., Horritt, M. S., and 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.

Beven, K. and Binley, A. (1992). The future of distributed models - model calibration and uncertainty prediction. *Hydrological Processes*, **6**(3), 279–298.

Beven, K., Leedal, D., Alcock, R., Hunter, N., Keef, C., and Lamb, R. (2012). Guidelines for good practice in flood risk mapping: The catchment change network.

Chow, V. T. (1959). *Open-Channel Hydraulics*. McGraw-Hill, New York.

Di Baldassarre, G., Schumann, G., and Bates, P. (2009a). Near real time satellite imagery to support and verify timely flood modelling. *Hydrological Processes*, **23**(5), 799–803.

Di Baldassarre, G., Schumann, G., and Bates, P. D. (2009b). A technique for the calibration of hydraulic models using uncertain satellite observations of flood extent. *Journal of Hydrology*, **367**(3-4), 276–282.

Ferro, C. (2012). Problems with 'distributed reliability': including forecast-observation data from multiple grid cells.

Hall, J. W., Manning, L. J., and Hankin, R. K. S. (2011). Bayesian calibration of a flood inundation model using spatial data. *Water Resources Research*, **47**.

Horritt, M. S. (2006). A methodology for the validation of uncertain flood inundation models. *Journal of Hydrology*, **326**(1-4), 153–165.

Horritt, M. S., Mason, D. C., and 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.

Mason, D. C., Horritt, M. S., Dall'Amico, J. T., Scott, T. R., and 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**(12), 3932–3943.

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

Miller, J., Kjeldsen, T., Hannaford, J., and Morris, D. (2013). An assessment of the magnitude and rarity of the november 2009 floods in cumbria. *Hydrology Research*.

Neal, J., Schumann, G., Bates, P., Buytaert, W., Matgen, P., and Pappenberger, F. (2009). A data assimilation approach to discharge estimation from space. *Hydrological Processes*, **23**(25), 3641–3649.

Neal, J., Schumann, G., Fewtrell, T., Budimir, M., Bates, P., and Mason, D. (2011). Evaluating a new lisflood-fp formulation with data from the summer 2007 floods in tewkesbury, uk. *Journal of Flood Risk Management*, **4**(2), 88–95.

NRC, N. R. C. (2006). Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts.

Pappenberger, F., Matgen, P., Beven, K. J., Henry, J.-B., Pfister, L., and Fraipont de, P. (2006). Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources*, **29**(10), 1430–1449.

Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R., and Matgen, P. (2007a). Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrol. Earth Syst. Sci.*, **11**(2), 739–752.

Pappenberger, F., Beven, K., Frodsham, K., Romanowicz, R., and Matgen, P. (2007b). Grasping the unavoidable subjectivity in calibration of flood inundation models: A vulnerability weighted approach. *Journal of Hydrology*, **333**(2-4), 275–287.

Romanowicz, R. and Beven, K. (1998). Dynamic real-time prediction of flood inundation probabilities. *Hydrological Sciences*, **43**(2), 181–196.

Romanowicz, R. and Beven, K. (2003). Estimation of flood inundation probabilities as conditioned on event inundation maps. *Water Resources Research*, **39**(3).

Romanowicz, R., Beven, K., and Tawn, J. (1996). Bayesian calibration of flood inundation models. In M. Anderson, D. Walling, and P. Bates, editors, *Floodplain Processes*. Wiley-Blackwell, London.

Schumann, G., Cutler, M., Black, A., Matgen, P., Pfister, L., Hoffmann, L., and Pappenberger, F. (2008). Evaluating uncertain flood inundation predictions with uncertain remotely sensed water stages. *International Journal of River Basin Management*, **6**(3), 187–199.

Stephens, E. M., Bates, P. D., Freer, J. E., and Mason, D. C. (2012). The impact of uncertainty in satellite data on the assessment of flood inundation models. *Journal of Hydrology*, **414-415**, 162–173.

Stephens, E. M., Bates, P. D., and Schumann, G. (2014). Problems with binary pattern measures for flood model evaluation. *Hydrological Processes*.

Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T. (2008). Two extra components in the brier score decomposition. *Weather and Forecasting*, **23**(4), 752–757.

Toth, Z Talagrand, O., Candille, G., and Zhu, Y. (2003). *Probability and Ensemble Forecasts*. John Wiley & Sons, Ltd., Chichester.

Werner, M., Blazkova, S., and Petr, J. (2005). Spatially distributed observations in constraining inundation modelling uncertainties. *Hydrological Processes*, **19**(16), 3081–3096.

Wright, N. G., Asce, M., Villanueva, I., Bates, P. D., Mason, D. C., Wilson, M. D., Pender, G., and Neelz, S. (2008). Case study of the use of remotely sensed data for modeling flood inundation on the river severn, uk. *Journal of Hydraulic Engineering-Asce*, **134**(5), 533–540.

Table 1: Optimum parameter sets of channel (ch) and floodplain (fp) friction identified using different performance measures for both aerial photography and wrack marks

	Aerial Photography			Wrack Marks		
Measure	ch	fp	Value	ch	fp	Value
CSI	0.026	0.057	83.67% (0.61m)	-	-	-
RMSE	0.038	0.029	0.40m	0.034	0.036	0.28m
Perc_50	0.054	0.022	12.42% (0.41m)	0.034	0.036	29.1% (0.28m)
Perc_1	0.047	0.02	20.76% (0.47m)	0.047	0.02	12.99% (0.48m)

Table 2: Brier Reliability for Different Uncertain Calibrations of the Cockermouth Model. Numbers in italics indicate where calibration / validation data are the same.

	Aerial Photography		Wrack Marks	
Weighting Method	Horritt	WSE	Horritt	WSE
Wrack RMSE	0.0157	0.038	-	*0.1304*
Wrack RMSE*	0.0079	0.053	-	*0.0279*
Wrack RMSE**	0.0133	0.128	-	*0.0255*
Wrack Perc_50	0.0157	0.1106	-	*0.0581*
Wrack Perc_1	0.0098	0.0221	-	*0.0130*
AP RMSE	*0.0157*	*0.0991*	-	0.1304
AP RMSE*	*0.0126*	*0.0460*	-	0.1072
AP RMSE**	*0.0115*	*0.2467*	-	0.0235
AP Perc_50	*0.0170*	*0.0435*	-	0.0254
AP Perc_1	*0.0087*	*0.0201*	-	0.0133
AP CSI	*0.0265*	*0.2467*	-	0.3028
AP CSI*	*0.0213*	*0.1998*	-	0.2120
Equal	*0.0268*	*0.2262*	-	0.2361

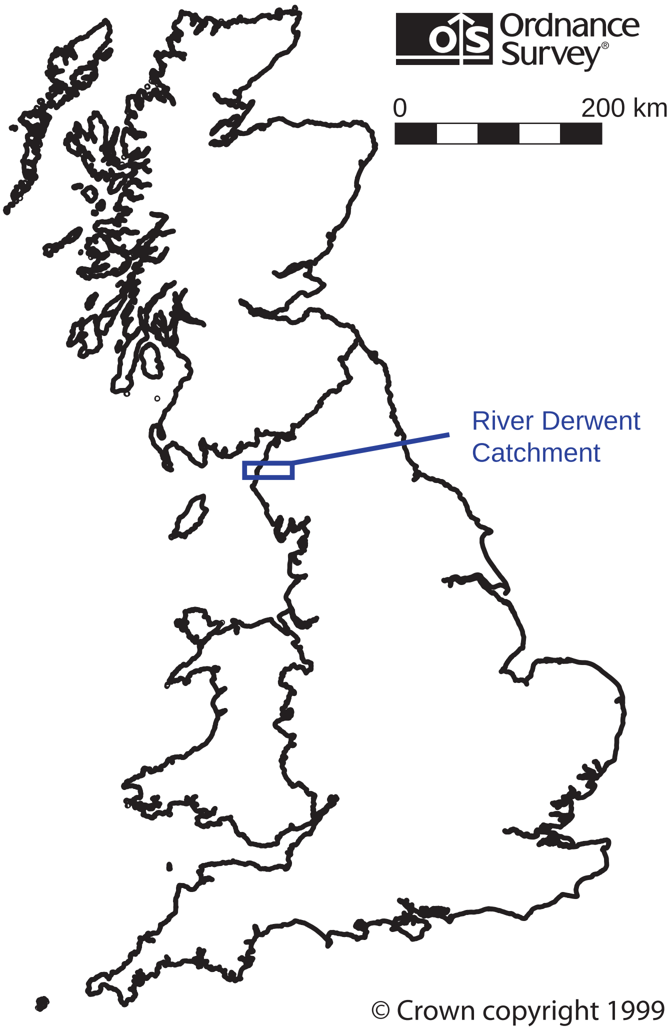

Figure 1: Location map showing the River Derwent catchment in the north-west of England. Source: Ordnance Survey

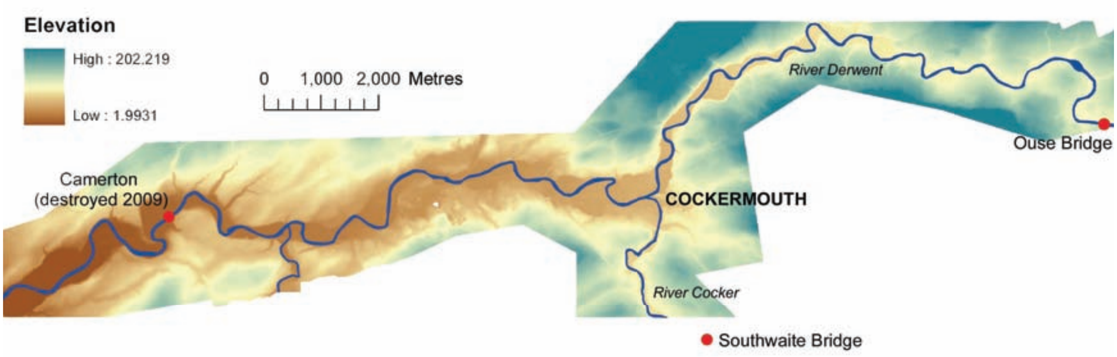

Figure 2: Topographic map of the River Derwent using LiDAR data at 2m resolution, showing location of gauges (red points). Source: Environment Agency

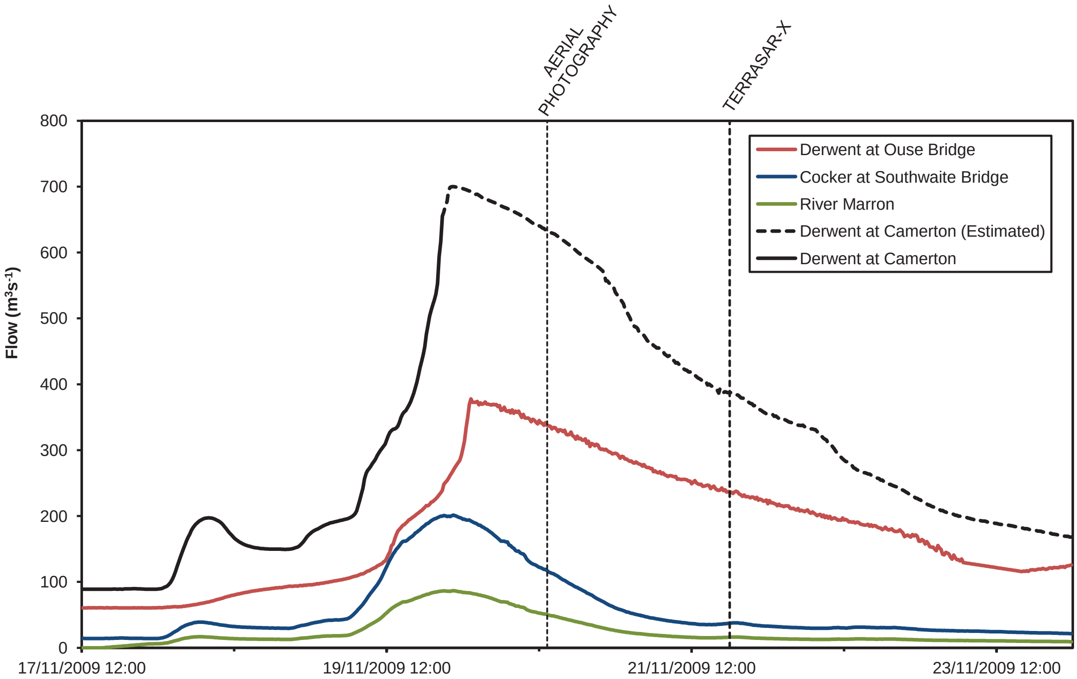

Figure 3: Gauged upstream flows for the River Derwent at Ouse Bridge, the River Cocker at Southwaite Bridge and the River Marron, with gauged downstream flows for the River Derwent at Camerton. Source: Environment Agency

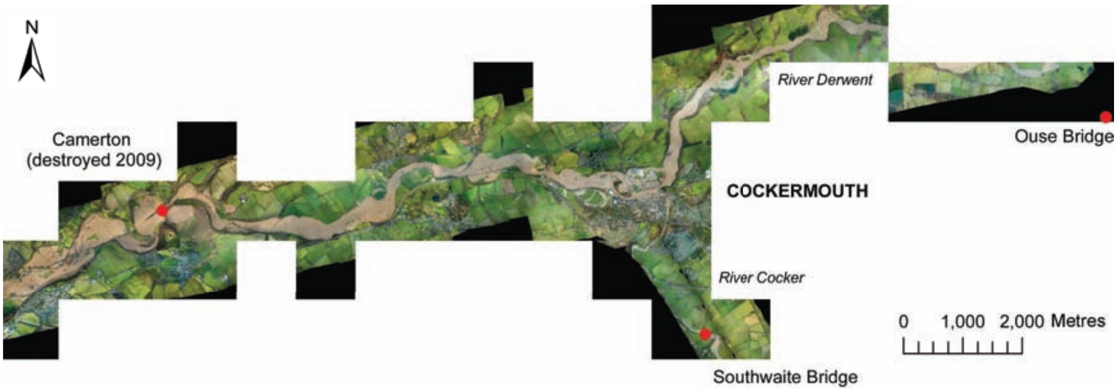

Figure 4: Extent of the aerial photography flown during the flood event. Source: Environment Agency

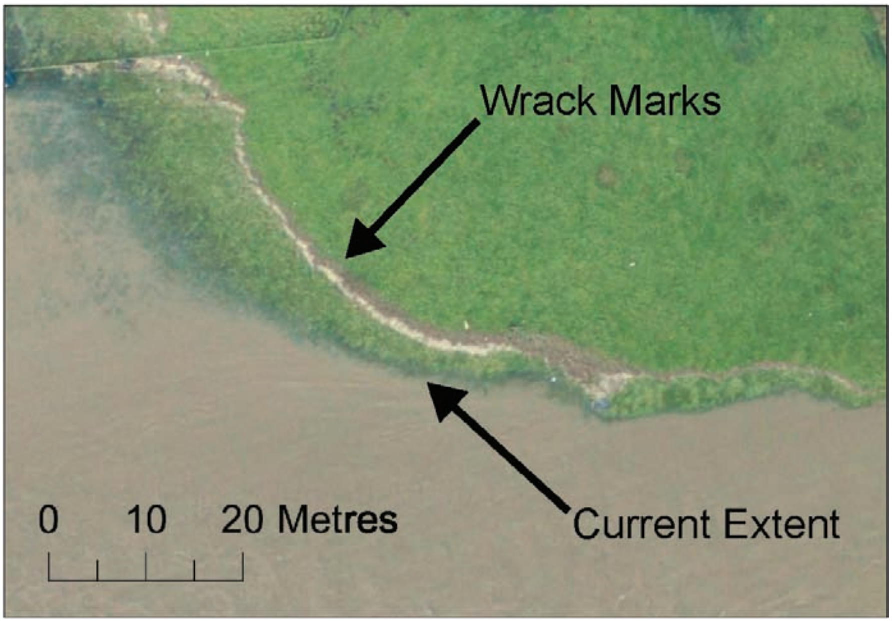

Figure 5: Example of wrack marks visible in the aerial photography adjacent to the then-current flood extent. Source: Environment Agency

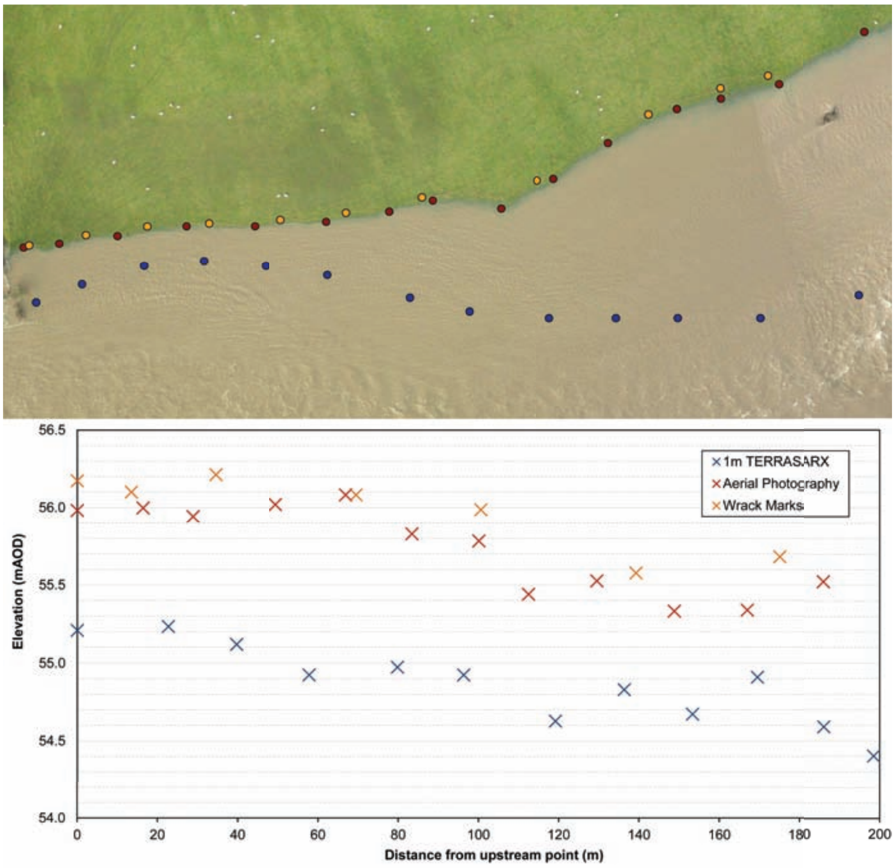

Figure 6: Demarked points along the margin of the flood along a field, with associated elevations derived by intersecting with LiDAR topographic data.

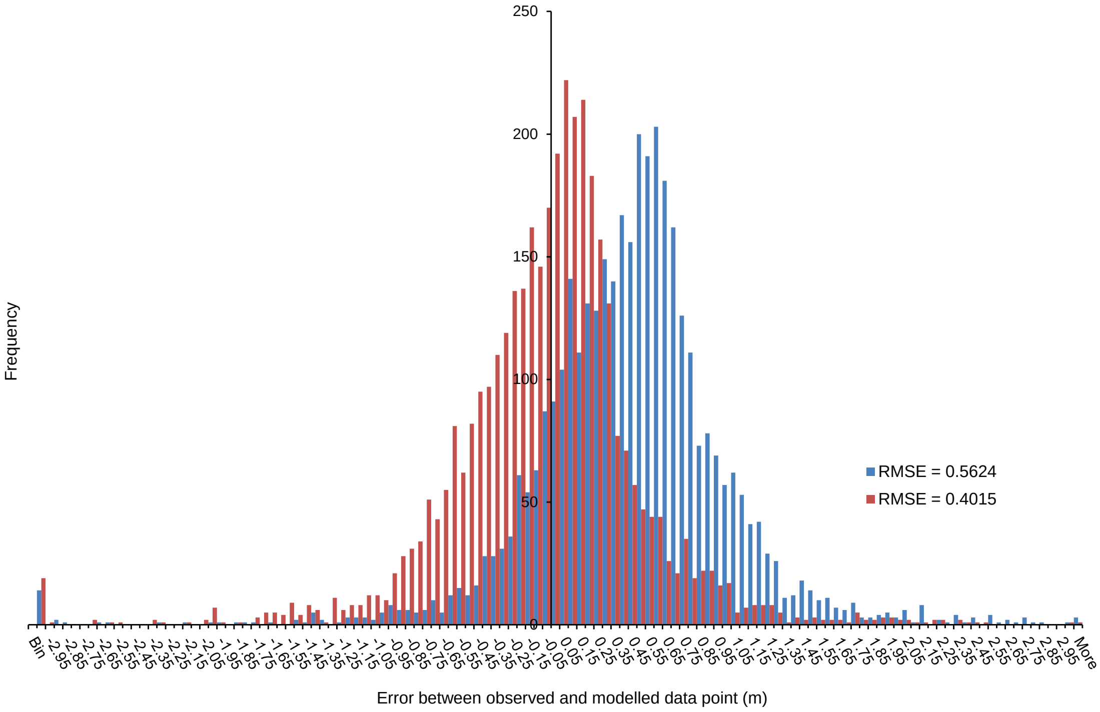

Figure 7: Frequency of error between individual observed and modelled data points, for two parameter sets with RMSEs of 0.5624 (blue) and 0.4015 (red).

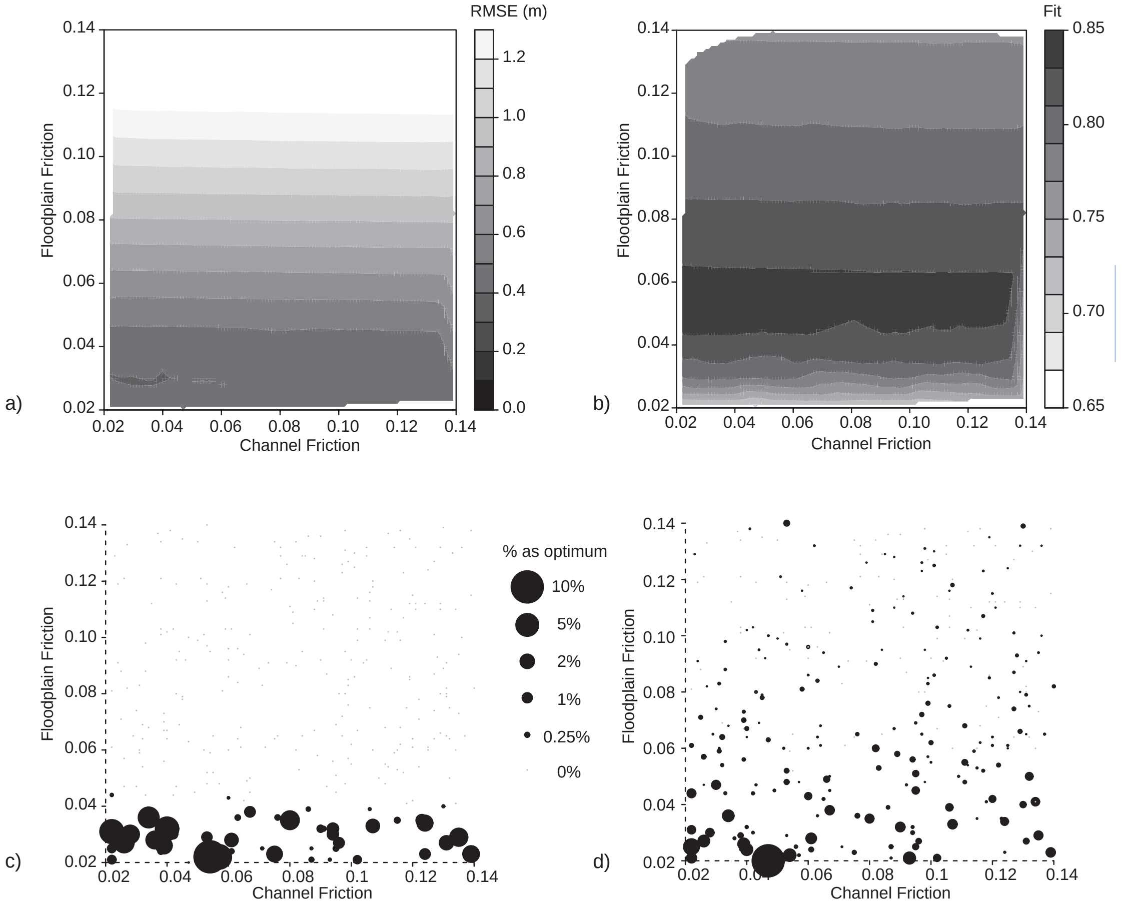

Figure 8: Parameter spaces for calibration of channel (x-axis) and floodplain (y-axis) friction parameters using Aerial Photography with the performance measures of: a) RMSE; b) CSI; c) Percentage as optimum parameter set for subsets of 50 points; and d) c) Percentage as optimum parameter set for all individual points (subsets of 1).

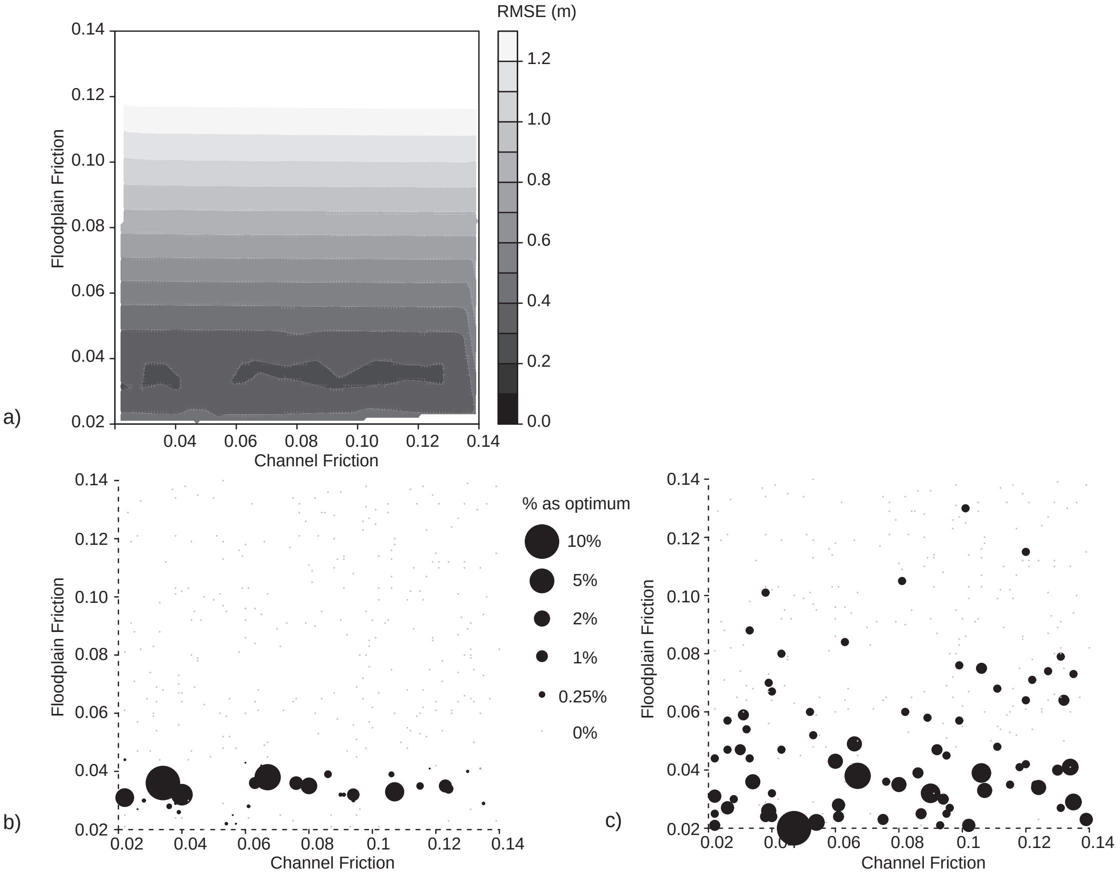

Figure 9: Parameter spaces for calibration of channel (x-axis) and floodplain (y-axis) friction parameters using Wrack Marks with the performance measures of a) RMSE; b) Percentage as optimum parameter sets for subsets of 50 points, and; c) Percentage as optimum parameter set for all individual points (subsets of 1).

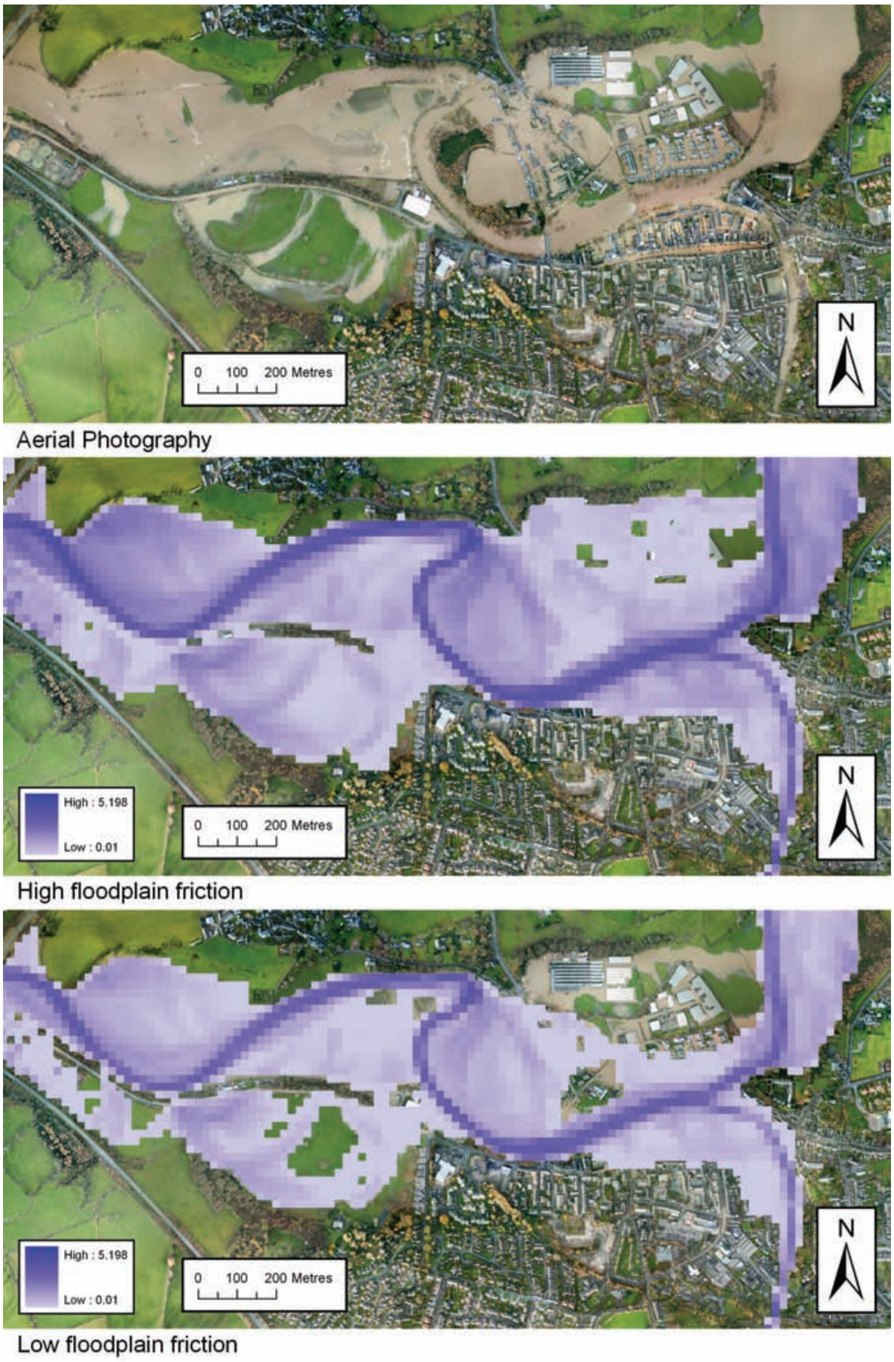

Figure 10: Difference in modelled extent compared to aerial photography for a high and low floodplain friction parameter sets on a subsection of the domain covering the Cockermouth area.

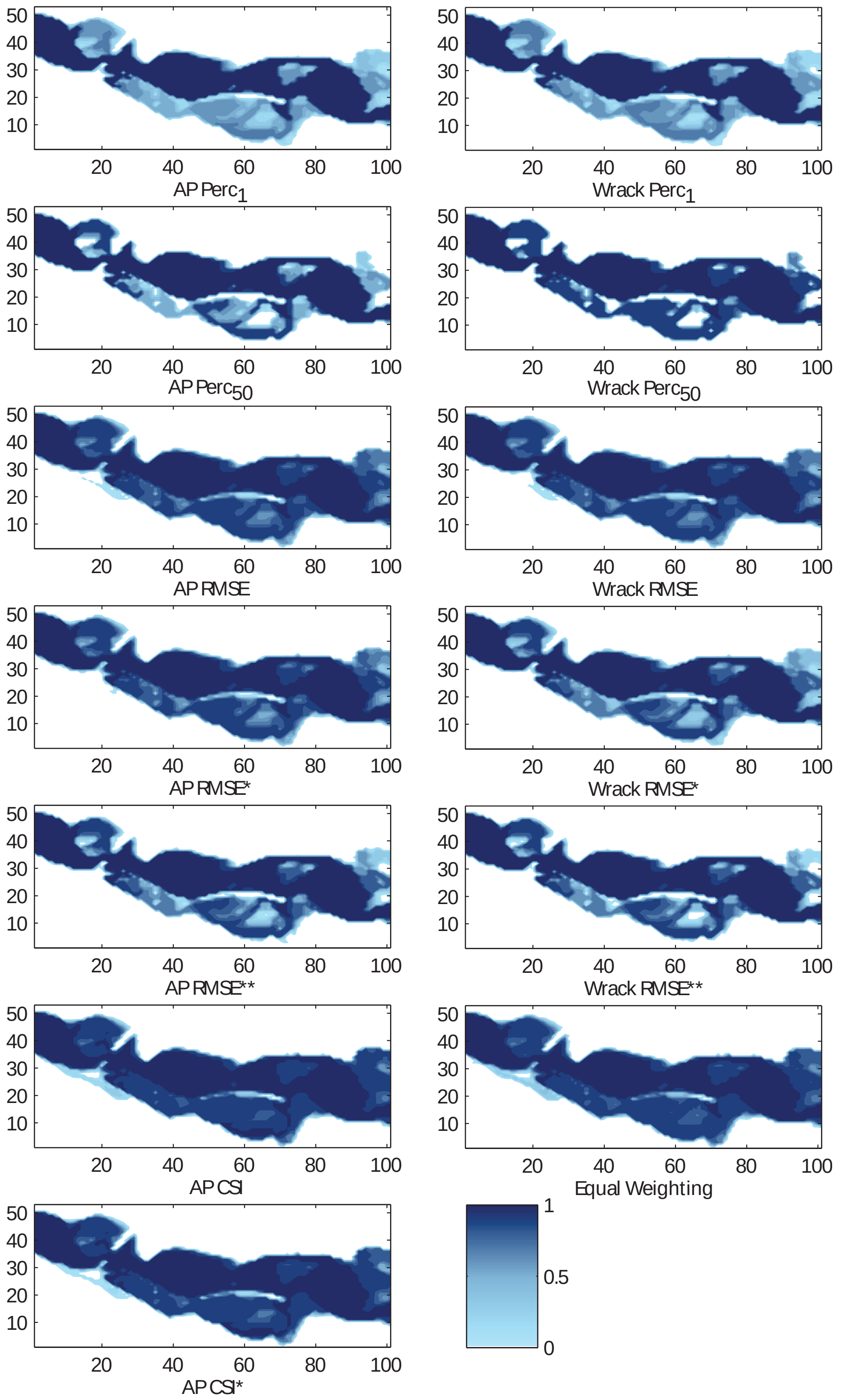

Figure 11: Cut-out from Probability of Inundation maps for the time of a Terrasar-X overpass (see 3). Showing the subtle differences in the mapped probabilities with the different weighting methods used for their construction.

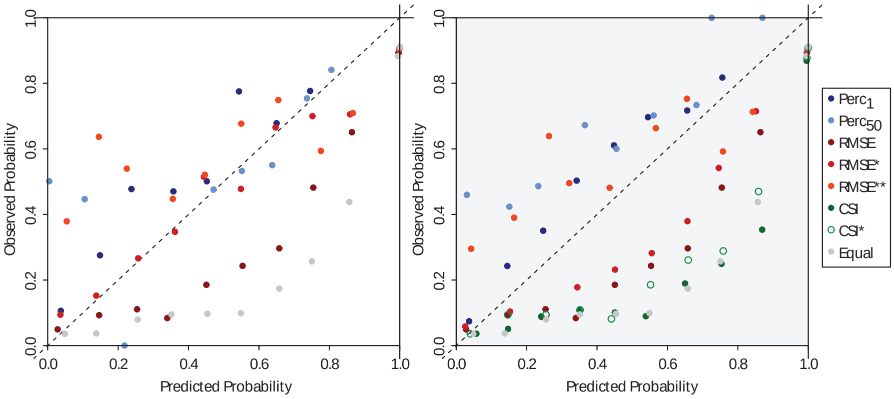

Figure 12: Horritt Reliability at the time of aerial photography overpass using calibrated weightings from 1) peak flood (wrack marks) and 2) aerial photography extent elevations. Greyed out plot indicates where the calibration / validation data are the same.

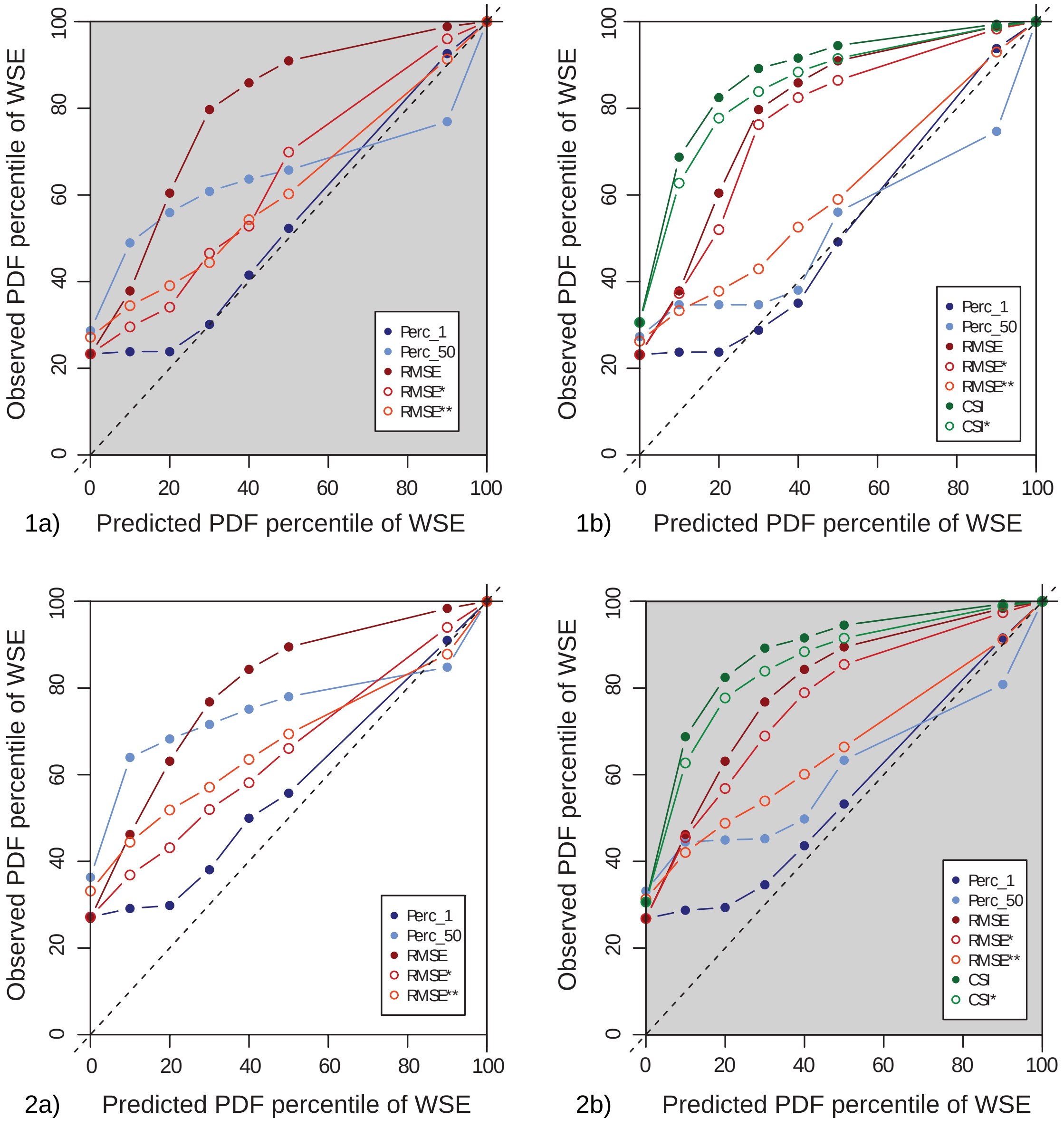

Figure 13: WSE Reliability for 1) Flood Peak using a) Wrack Marks, b) Aerial Photography, and 2) Time of Aerial Photography using a) Wrack Marks and b) Aerial Photography. Greyed out plots indicate where the calibration / validation data are the same

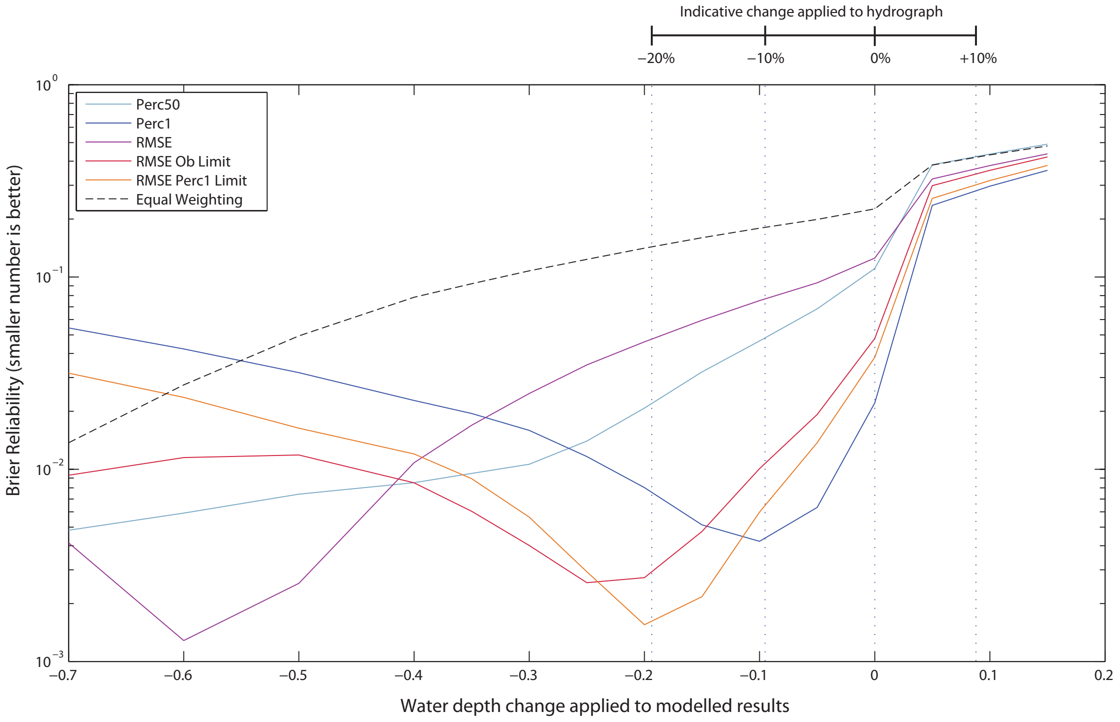

Figure 14: Change in Brier Reliability for different weighting methods if water depths are added / taken from the model results to represent boundary condition uncertainty. Bar along top gives indication of change in depths for different percentage change to flows.