

Assessing the reliability of probabilistic flood inundation model predictions

Article

Accepted Version

Stephens, E. ORCID: <https://orcid.org/0000-0002-5439-7563> and Bates, P. (2015) Assessing the reliability of probabilistic flood inundation model predictions. *Hydrological Processes*, 29 (19). pp. 4264-4283. ISSN 08856087 doi: 10.1002/hyp.10451 Available at <https://centaur.reading.ac.uk/46597/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

Published version at: <http://onlinelibrary.wiley.com/doi/10.1002/hyp.10451/full>

To link to this article DOI: <http://dx.doi.org/10.1002/hyp.10451>

Publisher: Wiley Online Library

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

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

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

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

11 December 2, 2014

12 **Abstract**

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

23 Here, a LISFLOOD-FP hydraulic model of an extreme (>1 in 1000 year)
24 flood event in Cockermouth, UK is constructed and calibrated using multiple
25 performance measures from both peak flood wrack mark data and aerial
26 photography captured post-peak. These measures are used in weighting the
27 parameter space to produce multiple probabilistic predictions for the event.
28 Two methods of assessing the reliability of these probabilistic predictions
29 using limited observations are utilised; an existing method assessing the

25 binary pattern of flooding, and a method developed in this paper to assess predictions of water surface elevation. This study finds that the water
26 surface elevation method has both a better diagnostic and discriminatory
27 ability, but this result is likely to be sensitive to the unknown uncertainties
28 in the upstream boundary condition.
29

30 1 Introduction and Objectives

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

37 The majority of flood inundation studies have used GLUE-based approaches
38 (e.g. Romanowicz *et al.*, 1996; Romanowicz and Beven, 1998; Aronica *et al.*, 1998,
39 2002; Romanowicz and Beven, 2003; Bates *et al.*, 2004; Werner *et al.*, 2005; Horritt,
40 2006; Pappenberger *et al.*, 2007a,b; Schumann *et al.*, 2008; Di Baldassarre *et al.*,
41 2009b), although some studies have adopted Bayesian approaches, (see Romanow-
42 icz *et al.*, 1996; Hall *et al.*, 2011). These studies have addressed one or more of the
43 types of the uncertainty in the modelling; model structural choice (e.g. Apel *et al.*,
44 2009), model friction and conveyance parameters (e.g. Aronica *et al.*, 1998; Ro-
45 manowicz and Beven, 2003; Bates *et al.*, 2004; Werner *et al.*, 2005; Pappenberger
46 *et al.*, 2007a), boundary conditions (e.g. Pappenberger *et al.*, 2006, 2007a), and the
47 geometry of the floodplain (Werner *et al.*, 2005) and channel (e.g. Pappenberger
48 *et al.*, 2006, 2007a) (including the representation of natural and man-made flow
49 control structures such as vegetation and buildings (Beven *et al.*, 2012)), as well as
50 the observed data used to condition the models (e.g. Pappenberger *et al.*, 2007a;
51 Di Baldassarre *et al.*, 2009b).

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

60 In Stephens *et al.* (2012) a LISFLOOD-FP hydraulic model of the River Dee,
61 UK was calibrated and uncertain flood inundation maps were produced using
62 different performance measures to weight each parameter set. It was shown that

63 the choice of performance measure for weighting the parameter space leads to
64 differences in the final uncertain flood inundation map, with there being clear
65 differences between a new uncertain measure (that implicitly takes into account the
66 uncertainty in the observed water surface elevations), the RMSE and the Measure
67 of Fit (Critical Success Index) used in studies such as that of Aronica *et al.* (2002).
68 In this study the Measure of Fit will be referred to as the Critical Success Index
69 as recommended by Stephens *et al.* (2014) to keep the terminology consistent with
70 other disciplines.

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

81 In this study the 2009 Cockermouth flood event on the River Derwent, UK is
82 used as a case study. This allows for the method developed by Stephens *et al.*
83 (2012), and the associated conclusions, to be tested on a different catchment,
84 and is also a data-rich case study with a high spatial resolution (0.15m) aerial
85 photography image that shows both the flood extent at the time of the photograph
86 and enables identification of wrack marks to indicate water levels at peak flood.

87 **1.1 Current methods for probabilistic evaluation of prob- 88 abilistic flood inundation models**

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

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

101 80% of the time? Therefore, even if the requirements of the formal probabilistic
102 evaluation methods used in fields such as meteorology cannot be met because of
103 data limitations, attempts should be made to evaluate probabilistic predictions
104 using the few data that are available. Accordingly, modellers of extreme events
105 and climate change, who have similar data limitation issues, have proposed the
106 use of spatial patterns of predictions and outcomes to build sufficient datasets for
107 evaluation (Horritt, 2006; Annan and Hargreaves, 2010).

108 Horritt (2006) proposed a method to validate inundation model predictions us-
109 ing a single observation of flood extent (hereby referred to as the Horritt method),
110 in effect, aggregating observations of the flooded state within each grid cell to
111 produce a large enough sample size. A LISFLOOD-FP model (Bates and De Roo,
112 2000) of a reach of the River Severn was set-up, and calibration / validation data
113 were provided by two SAR images of flood events in October 1998 and Novem-
114 ber 2000. The model was calibrated using one dataset, and validated using the
115 other, therefore allowing for some independence between model calibration and
116 evaluation.

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

129 Although the Horritt (2006) paper maintains separation between the cali-
130 bration and validation data, the Horritt method does not account for the co-
131 dependence between the observations used in the analysis. For example, it is
132 likely that if one cell on the floodplain has a predicted inundation probability of
133 50% and it is observed as being flooded, that any adjacent cells will have similar
134 probabilities and observations. While Horritt (2006) suggests that the issue of
135 only having single observations has been ‘neatly sidestepped’, it could be argued
136 that by using observations from the same event on the same model domain leads
137 to issues of co-dependence that could potentially bias the analysis.

138 To increase independence of observations it would be necessary to choose a
139 subset of cells across the domain that are not related, and given a large enough
140 number of cells this would be possible. However, a perhaps more sensitive and dis-

141 criminatory measure might be to evaluate the water surface elevation predictions
142 themselves, looking at where the observations fall within the predicted distribu-
143 tion of water depths. Unlike the Horritt method, a method that used observations
144 of water surface elevations as the evaluation dataset would not require a contin-
145 uous flood extent to be recorded, and therefore could be applied where there are
146 discontinuous measurements such as wrack lines, or where the continuity of flood
147 outlines derived from remote sensing is limited due to dense vegetation disguising
148 the true flood edge in particular areas.

149 As well as using more ‘independent’ observations and being applicable for a
150 larger variety of data sources, it is hypothesised that a method that evaluates
151 probabilistic water surface elevation predictions will be more sensitive and there-
152 fore allow for better discrimination between the performance of different uncertain
153 flood predictions. To judge this, different performance measures are used to weight
154 water surface elevation predictions and produce predicted water elevation distri-
155 butions for points across the domain. The objectives of this paper are therefore
156 as follows:

- 157 1. To evaluate, for the 2009 flood event in Cockermouth, what performance
158 measure / weighting method produces the more reliable probabilistic flood
159 inundation predictions
- 160 2. To confirm the consistency of this conclusion by comparing results for cali-
161 brating / evaluating at time of peak flood and for the time of aerial photog-
162 raphy overpass during flood recession, again using the Cockermouth dataset.
- 163 3. To compare the method for evaluating probabilistic predictions that is de-
164 veloped in this paper with the Horritt method, determining whether they
165 produce the same outcomes, and which is more sensitive and therefore bet-
166 ter for discriminating between these different weighting methods
- 167 4. To determine what can be learnt about the model from the two different
168 methods for evaluating probabilistic predictions

169 2 Methodology

170 2.1 Study site and test data

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

175 An extremely large flood event occurred in the catchment in November 2009
176 after a prolonged period of rainfall over the mountains of the central Lake District.
177 At the Seathwaite Farm raingauge in the upper reaches of the Derwent catchment a
178 new UK record 24-hour rainfall record of 316.4mm was established for the 24-hour
179 period up to 00:00 on the 20th November, and estimated to have a return period
180 of 1862 years (Miller *et al.*, 2013). Due to the prolonged period of rainfall (10mm
181 / hour average for 36 hours) (Miller *et al.*, 2013), levels of major lakes within the
182 region reached new recorded maxima and consequently their buffering effect on
183 downstream flows was reduced (Miller *et al.*, 2013). Using an improved Flood
184 Estimation Handbook flood frequency analysis Miller *et al.* (2013) estimate that
185 the discharge return period on the Derwent at Ouse Bridge was 1386 years, and
186 769 years on the Cocker at Southwaite Bridge. The combined flow at Camerton,
187 estimated by the Environment Agency (EA) as $700m^3s^{-1}$ has a return period of
188 2102 years, with 95% confidence limits of 507 and 17706 years (Miller *et al.*, 2013).

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

193 Gauged flow data (see Figure 3) are available for this flood event from Ouse
194 Bridge on the Derwent (the outflow from Bassenthwaite lake), Southwaite Bridge
195 on the Cocker (upstream of Cockermouth), and Camerton which is approximately
196 6km downstream from the confluence of the Cocker and Derwent as the crow
197 flies. The flood is modelled from 12:00 on 17th November 2009, before water
198 levels begin to rise, to 23:45 on 23rd November 2011, where water levels are nearly
199 back to normal levels. Flow data for the River Marron have been provided by
200 Professor Sear of Southampton University, by rescaling the flows in the Cocker
201 using the comparative size of the catchments. For the Ouse Bridge gauge, the EA
202 has provided metadata to advise that the stage at the peak of the flood has been
203 edited using estimates of the maximum flood level from a wrack survey, with the
204 time of peak and the infilled data estimated using correlation techniques. Further,
205 for the conversion to flow data using a rating curve the Quality flag is given as
206 'Estimated' and 'Extrapolated Upper Part'. For the Southwaite gauge, the stage
207 data is assigned a quality of 'Good' throughout, with approximately 17 hours at
208 the peak of the flood where the information has been edited to use the back up
209 data from the gauge due to float and weight issues that caused slight differences in
210 the hydrograph. Accordingly, the Quality flag of the flow data is given as 'Good'
211 throughout, and within the range of the rating curve for all but the 30 hours
212 around the peak flood, where the data has been extrapolated.

213 The Camerton gauge was severely damaged during the event, with 'Good'
214 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 amalgamation 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

255 will provide a stern test for the model on the falling limb of the flood. At the time
256 of aerial photography overpass, flows still remained out of bank (as can be seen
257 from the imagery), and so the floodplain is not considered to be draining at this
258 point. However, it is worth noting that coarse resolution models have been shown
259 to be poor at draining the floodplain (Bates *et al.*, 2006; Wright *et al.*, 2008; Neal
260 *et al.*, 2011).

261 While in many studies aerial photography is used as a benchmark to assess
262 the accuracy of satellite observed flood extents (Horritt *et al.*, 2001; Mason *et al.*,
263 2007), thereby assuming it to be accurate and precise, here this assumption is
264 not made since these data will contain unknown errors. This is demonstrated
265 in Figure 6, where there is obvious deviation from a smooth water surface for
266 what should be an easy 200m stretch of floodplain to delineate the flood extent
267 from. These deviations from a smooth water surface will be from two sources; the
268 first being geolocational errors in the (manual or automatic) demarcation of the
269 outline and the geocorrection of the data, and the second; errors in the LiDAR
270 data used in the intersection of the flood extent and the topography. While it
271 could be argued that the deviation would be smaller if the points were better
272 digitised, these points have already been manually repositioned from the data as
273 provided by the EA, and consequently any better recorrection of these 2000+
274 data points would be a significant time burden. Also, and as can be seen in
275 Figure 6, there is some confusion over whether the edge of the water surface lies
276 at the edge of the sediment-laden area of water, or whether it lies at the edge of
277 the surrounding darker area of vegetation which could be the current flood level,
278 emergent vegetation or simply wet vegetation that has been previously flooded.
279 Further, the vertical height errors that are incorporated with the intersection with
280 the LiDAR data could be in the region of 0.25m RMSE, and cannot be removed.

281 **2.2 Model Set-Up and Calibration**

282 A 2D LISFLOOD-FP model was set-up using the inertial formulation of the shal-
283 low water equations as described by Bates *et al.* (2010). The model incorporates
284 the LiDAR topographic data outlined above rescaled to 20m resolution to enable
285 multiple simulations to be run without unreasonable computational cost, and the
286 gauged data as upstream boundary conditions. The gauged data for Camerton
287 have not been used as a downstream stage-varying boundary condition due to the
288 known poor data quality. Instead a free boundary condition has been imposed
289 using test runs of the model to approximate the water surface slope at this part of
290 the catchment, which was shown to vary slightly from the local valley slope. The
291 model is run for 167.75 hours, from 12:00 on 17th November 2009 to 23:45 on the
292 23rd November 2009, across a domain 100km² in size (including No Data cells).
293 A simulation of the model run on 4 processors of the University of Bristol's Blue

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

296 The upland nature of the upper Cockermouth catchment means that channel
297 friction values might be higher than lowland rivers such as the Dee due to a gravel
298 bed, and consequently, floodplain friction values may possibly be lower than those
299 for the channel due to the pastural land use which dominates the floodplain across
300 the catchment. While it is expected that parameter values are effective, physically-
301 based parameter ranges can be used to define the parameter space. According to
302 Chow (1959) pasture with short grass would have a minimum Manning's n of 0.025,
303 and a gravel bed would have a minimum of 0.030. Some areas of the catchment are
304 heavily forested or have medium to dense brush, which might be expected to have
305 a maximum Manning's n value of 0.12 (Chow, 1959). To ensure that the entire
306 range of potential friction values are sampled, but also accepting that friction as
307 specified in LISFLOOD-FP also acts as an 'effective' parameterisation (to account
308 for unrepresented model structures such as sub-grid scale topographic features,
309 and also unquantified uncertainties such channel topography and input flows), the
310 parameter space is defined by channel and floodplain friction values of between
311 0.02 and 0.14. Calibration of the model was carried out by randomly sampling
312 300 parameter sets from the parameter space.

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

$$CSI = \frac{A}{A + B + C} \quad (2.1)$$

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

328 The third performance measure, Perc_50 is the percentage as optimum measure
329 detailed in Stephens *et al.* (2012), developed to provide an (implicit) representation
330 of the uncertainty in the observed data into the calibration process. For this
331 measure, ten thousand subsets of fifty points are taken from the observed dataset,

332 and the parameter set which produces the lowest RMSE for each subset is recorded.
333 The frequency for which each parameter set occurs as the optimum is calculated,
334 and converted into a percentage of the total number of subsets that have been
335 evaluated.

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

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

363 Other studies have represented the uncertainty in observational data more ex-
364 plicitly; Pappenberger *et al.* (2007a) use a fuzzy map of flood extent and a global
365 fuzzy performance measure, and Di Baldassarre *et al.* (2009b) produced a ‘pos-
366 sibility of inundation map’ by looking at how the model calibration varies when
367 different methods of determining the flood outline from two different SAR im-
368 ages of a flood event are used. However, these existing studies have focussed
369 on the uncertainty in the pattern of flood extent. Such contingency table based
370 performance measures have been shown to be problematic for model calibration
371 given their sensitivity to spatial variations in topographic gradient (Stephens *et al.*,

372 2014), as such, research efforts should focus on the use of water surface elevation
373 observations instead. Some studies have used an explicit representation of the
374 uncertainty in satellite-derived water surface elevations for predicting flood wave
375 propagation using a 1D model (Di Baldassarre *et al.*, 2009a) and discharge (Neal
376 *et al.*, 2009), but this has yet to be addressed for (2D model) predictions of the
377 pattern of flood inundation.

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

388 2.3 Probability of inundation maps

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

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

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

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

$$P_i^{flood} = \frac{\sum_j f_{ij} W_j}{\sum_j W_j} \quad (2.3)$$

408 Where j is the number of model simulations, f is the flooded state of the pixel
 409 (1 = wet, 0 = dry) and W_j is the weighting given to each model simulation.

410 2.4 Methods for evaluation of probabilistic predictions

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

418 2.4.1 Assessing reliability using the Horritt method

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

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

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

430 Despite this, it is important for the demonstration of the applicability of proba-
 431 bilistic predictions to be able to give some estimate of their reliability. Accordingly,
 432 modellers of extreme events and climate change, who have similar data limita-
 433 tion issues, have proposed the use of spatial patterns of predictions and outcomes
 434 to build sufficient datasets for evaluation (Horritt, 2006; Annan and Hargreaves,
 435 2010). As such, Horritt (2006) proposed assessing reliability using the probabilities
 436 of inundation assigned to each cell.

437 For the Horritt method Equation 2.4 is adjusted such that \bar{f}_k is the mean of
 438 the probability forecasts of a cell being flooded k (in each bin), and \bar{o}_k is the

439 observation of flooding k in each bin. N is the total number of observations, n is
440 the number of events that fall into each bin m . Note that for the Horritt method
441 model cells where the predicted probability of flooding = 0 are ignored in the
442 calculation since they account for the vast majority of the domain and therefore
443 would bias the result.

444 **2.4.2 Assessing reliability of water surface elevation predictions**

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

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

458 The next step is, for each observation, to record where it lies within the pre-
459 dicted probability distribution. These records of observation location can be rep-
460 resented in a cumulative frequency plot, where the number of observations that
461 fall within each bin of predictions is plotted. If the predictions are perfectly re-
462 liable the gradient of the line should be 1 since 10% of observations would fall
463 within the first 10% of the probability distribution, 20% within the first 20%, and
464 so on. Where the gradient is steeper than the 1:1 line then, in general, there has
465 been an overestimation of the uncertainty in the model. Where the gradient is
466 less steep than the 1:1 line there has been an underestimation of uncertainty, with
467 observations having been made that lie outside of the predicted range.

468 An indication of bias within predictions, or where the full range of uncertainty
469 has not been adequately captured, can be seen by identifying where the line inter-
470 cepts with the vertical lines of $x=0$ (the y axis) and $x=100$. The intercept with
471 the y axis is the percentage of observations that fall outside the lower bounds of
472 the predicted probability distribution of water surface elevations. The intercept
473 with the line $x=100$ can be substracted from 100 to give the percentage of obser-
474 vations that fall outside the upper bounds of the predicted probability distribution
475 of water surface elevation predictions. The reliability of model predictions using
476 this method can also be quantified using a calculation similar to Equation 2.4, by
477 finding the difference between the expected and observed cumulative frequency of

478 observations 2.5. For the wse reliability the cumulative reliability is calculated
 479 rather than an isolated comparison of the expected and actual number of observa-
 480 tions in each bin to ensure that no model is penalised for bringing the probabilistic
 481 predictions back towards the expected 1:1 line. For example, if no observations fell
 482 within the first bin (0%-10% decile), then if 20% of observations fell in the (10%
 483 20% decile), then the first bin should be penalised for a 10% difference, but the
 484 second bin should not be because it brings the overall percentage of observations
 485 in the first two bins back to the expected value. As such, for the WSE method E_m
 486 is the expected number of observations to have fallen up to and including bin m ,
 487 and the O_m is the actual number of observations to have fallen up to and including
 488 bin m . If the bins were set as every 10%, then the total number of bins would be
 489 10 and so the expected value for each individual bin inside the distribution would
 490 be 10%.

$$Reliability = \frac{1}{N} \sum_{m=1}^M n(E_m - O_m)^2 \quad (2.5)$$

491 3 Results

492 3.1 Modelled parameter space using different performance 493 measures / data sources

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

510 Optimum friction parameter sets for each measure and each dataset are shown
 511 in Table 1. For such an extreme event upstream boundary conditions are unlikely
 512 to be error-free, and as described previously, the friction parameters used in the

513 modelling should also be considered as ‘effective’ given that they also compensate
514 for subgrid scale processes. Accordingly, some deviation from physically realistic
515 values for friction are to be expected, but a modeller that finds a ‘physically
516 realistic’ parameterisation may have overconfidence in thinking that the model is
517 robust with respect to other uncertainties. Here, the RMSE measure gives the most
518 physically realistic floodplain friction optimum of around 0.03 for short pasture,
519 the CSI measure finds higher than expected values, and the Perc measure does not
520 find a well-defined optimum within the areas of the parameter space that might
521 be considered to be physically realistic. However, it is important to assess whether
522 these ‘physically realistic’ parameterisations produce reliable predictions.

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

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

543 This optimum for higher floodplain friction parameters is investigated using
544 a visual comparison between the observed dataset and the model output for two
545 simulations with a fixed channel and different floodplain frictions (respectively of
546 [0.027,0.026] and [0.027,0.057]). There are several areas across the domain where
547 the higher floodplain friction simulation better matches a particular area of the
548 observed extent than the low floodplain friction simulation (such as in the top
549 right area of the catchment shown in Figure 10), but in doing so the higher flood-
550 plain friction simulation fails to match the areal pattern in nearby areas. These
551 areas of unexpected inundation are not relics of observed data error, since there is
552 strong agreement for multiple data points and they are clearly visible in the aerial

553 photography. This suggests that higher floodplain friction simulation is perhaps
554 correctly matching the observed inundation in specific areas for the wrong reasons.
555 There are several possible explanations for the inability of the lower floodplain friction
556 simulation to capture these flooded areas; the model may have a resolution
557 too coarse to accurately capture bank heights, or processes not represented in the
558 lower friction model such as bank failure might be important. Consequently, it is
559 thought that the higher floodplain friction simulation is matching the pattern of
560 flooding better, but for the wrong reasons.

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

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

584 The Perc_1 and Perc_50 plots distinguish areas of the parameter space that are
585 non-performing, where parameter sets never appear as the optimum using multi-
586 ple realisations of the observed data. Perc_50 shows (as would be expected) larger
587 non-performing areas than Perc_1, since subsets of 50 act to average the range of
588 uncertainty that can be represented using each individual point. The Perc mea-
589 sures hint that the optimum parameter sets sit to the margins of the parameter
590 space, which suggests that the model (or at least its floodplain) contains too much
591 water. This could be due to errors in the specification of the upstream flows, which
592 is quite likely because of the potential errors in the gauged data detailed earlier in

593 this paper, or alternatively due to geomorphological changes during the flood event
594 that increased the capacity of the river channel. Such geomorphological changes
595 can be identified in a post-flood LiDAR dataset of the event, and consequently
596 will have some effect, although it is not possible without further modelling to be
597 confident of whether this or incorrect upstream flows are the cause of the apparent
598 bias in the model. Ignoring the CSI measure due to its known problems, it is inter-
599 esting that the RMSE shows a well defined optimum within the parameter space,
600 and this demonstrates the need for evaluating whether the Perc measures or the
601 RMSE provides more reliable predictions. As mentioned earlier in this study in
602 Section 2.4.2; it is important to ensure that the parameter space is large enough so
603 that the limits of the predicted probability distribution are not predetermined by
604 a subjective choice of potential parameter sets. The identification of optimum pa-
605 rameter sets at the margins of the parameter space for the Perc measures suggests
606 that this may be an issue; however the lower bounds for the roughness parameters
607 are limited by model stability rather than subjectivity, which is not untypical for
608 hydraulic models and is not thought to affect the conclusions drawn in this study.

609 **3.2 Uncertain Inundation Maps**

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

617 **3.3 Reliability**

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

626 Figure 12, Panel 1 (calibration using wrack marks deposited at the time of
627 peak flood) clearly demonstrates that the RMSE weighting overpredicts inundation
628 probabilities, and that the Perc_50 method is an improvement on the RMSE,
629 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

670 for as much of the uncertainty in the modelling post flood peak. Consequently it
671 will be important to account for these uncertainties explicitly.

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

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

700 4 Discussion

701 One of the aims of this paper was to evaluate the most reliable performance mea-
702 sure for weighting parameter sets to produce uncertain flood inundation maps. As
703 well as the conventional performance measures of RMSE and CSI, the Perc mea-
704 sure, developed in Stephens *et al.* (2012), was also used to address how observed
705 data errors are accounted for in the calibration process. Unlike the Perc_50 mea-
706 sure, which uses multiple subsets of 50 data points, the Perc_1 measure records,
707 using individual observed data points, the number of times that each parameter set

708 appears as the optimum. This measure of agreement provides a parameter space
709 that appears to give the best overall picture of the likelihood of each parameter
710 set being the optimum.

711 Both methods of assessing model reliability show that the Perc_1 measure pro-
712 duces the most reliable predictions, and this result is consistent for the validation
713 data at the time of peak flood and at the time of the aerial photography over-
714 pass. This is a surprising result as, up until now, observations are usually grouped
715 together into a ‘global’ dataset for model calibration. While Pappenberger *et al.*
716 (2007b) highlight the importance of a vulnerability-weighted model calibration to
717 produce an improved local model performance, e.g. with respect to locations of
718 critical infrastructure, we show that considering observations individually can ac-
719 tually improve the global performance. But RMSE, as a measure which uses an
720 average of all the (uncertain) observed data, will be influenced by outliers. As
721 there is no reason to discard such an outlying point (unlike points that are in
722 densely vegetated areas), there is still a (perhaps very small) chance that it is
723 correct, and that all other points are affected by some systematic error. Therefore
724 with these outliers influencing model calibration, it is important that they are used
725 proportionately.

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

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

742 The Perc_1 methodology implicitly accounts for the potential uncertainty, ar-
743 guably providing a different approach to acceptability rather than applying a sub-
744 jective behavioural threshold based on a simple estimation of observed data un-
745 certainty for the limit of acceptability. If there were observed data of multiple
746 flood events on a catchment, and none showed a particular parameter set as an
747 optimum, then this parameter set would surely be rejected. The Perc_1 measure

748 applies this logic (albeit with assumptions) to multiple observations from the same
749 flood event; in this approach each observation is treated as a separate observation,
750 such that if a parameter set is never the 'optimum' the agreement or lack of in the
751 Perc_1 measure is used to define acceptability. Ideally, this of course requires that
752 all sources of uncertainty are accounted for, as potentially areas of the parameter
753 space might be discarded that would otherwise be acceptable, if, for example,
754 boundary condition uncertainty were taken into account.

755 Assessing reliability is a good way of testing the methodologies for defining acceptability
756 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
757 uncertainty has not been taken into account. To provide a preliminary assessment
758 of the sensitivity of the results described in this paper to upstream boundary condition
759 uncertainty, a change in the hydrograph was simulated by taking / adding
760 different amounts from the water surface elevations produced by the ensemble
761 modelling Figure 14. These changes are commensurate with the changes seen
762 when changing the hydrograph by a fixed percentage for a single parameter set,
763 as indicated on the figure. The Brier reliability was recalculated for each applied
764 change to give an indication of its sensitivity to boundary condition uncertainty.
765 Figure 14 therefore demonstrates that if, in reality, the flows were consistently 10%
766 lower than the choice of optimum weighting method would be different. Given that
767 the uncertainty in the upstream boundary condition during this flood is unknown,
768 this sensitivity urges caution when considering the robustness of these results.

770 Future work should explicitly incorporate boundary condition uncertainty into
771 the analysis, as well as produce and test a methodology that incorporates a more
772 detailed and explicit representation of observed data uncertainty, incorporating, for
773 example, the resampling errors of the LiDAR data. Further studies are needed to
774 confirm whether the conclusions are robust on different flood events with different
775 magnitudes. Namely, does the Perc_1 measure produce the most reliable predictions
776 for flood events of smaller magnitude, and can weighting using these smaller
777 events still provide reliable inundation possibilities for extreme events such as the
778 1 in 1000 year return period flood? Further, would a more explicit representation
779 of uncertainty in the observed data produce more reliable predictions?

780 The other main aim of this study was to develop a new method for evaluating
781 uncertain flood inundation predictions, and then compare the results from this
782 with those from the Horritt method. One of the advantages of the WSE method
783 is that it can be used for discontinuous datasets (such as the wrack marks in this
784 study), and it therefore has wider applicability. On top of this, and despite both
785 reliability methods coming to the same overall conclusion, there are differences in
786 the level of information provided by each that indicates that the WSE method
787 is more discriminatory, since it produces a wider range of reliability scores, and

788 also has wider diagnostic capabilities since it provides more information than the
789 Horritt method. For example, the Horritt method does not show any bias when the
790 Perc_50 measure is used, but the plots of cumulative reliability for the WSE method
791 clearly show that this measure underestimates the range of uncertainty in the
792 model. This underestimation is caused by discarding areas of the parameter space
793 as ‘non-performing’ when they should still be taken into account when producing
794 the uncertain estimates of flood hazard. Further, the WSE method can show
795 whether and how many of the water surface elevation observations lie within the
796 predicted range. If they do not, then this hints at epistemic uncertainty that needs
797 to be addressed.

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

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

814 Model evaluation using the WSE method has proved a useful diagnostic tool
815 that provides more information about model performance than the Horritt method,
816 giving an indication of the percentage of observations that fall outside the upper
817 and lower limits of the probability distribution of water surface elevations. In the
818 case of the Cockermouth flood it can be seen (using the Perc_1 measure which
819 has been identified as producing the most reliable predictions), that at the time
820 of the peak flood the model has around 12% of observations that fall below the
821 lower limits of the range of water surface elevation predictions, which increases to
822 around 22% at the time of the aerial photography overpass. Despite there being
823 no other study for comparison, that 88% of peak flood observations fall within the
824 predicted range could be considered good for a model that only takes into account
825 parameter and observed data uncertainty, and especially for such an extreme flood
826 event where the errors in the inflow and wrack mark data are likely to be high.
827 The drop in model performance only a few hours after peak flood suggests that

828 new sources of uncertainty need to be taken into account to produce a similar
829 reliability to predictions made of the peak flood, and as mentioned previously the
830 uncertainty in geomorphological change during the flood, or in the gauged flow
831 data should be investigated.

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

845 5 Conclusions

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

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

866 reflective of the ability of water elevation comparisons to make better or broader
867 use of the available data.

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

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

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

905 6 Acknowledgements

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

914 References

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

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

919 Aronica, G., Hankin, B., and Beven, K. (1998). Uncertainty and equifinality in
920 calibrating distributed roughness coefficients in a flood propagation model with
921 limited data. *Advances in Water Resources*, **22**(4), 349–365.

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

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

928 Bates, P. D. and De Roo, A. P. J. (2000). A simple raster-based model for flood
929 inundation simulation. *Journal of Hydrology*, **236**(1-2), 54–77.

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

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

937 Bates, P. D., Horritt, M. S., and Fewtrell, T. J. (2010). A simple inertial formula-
938 tion of the shallow water equations for efficient two-dimensional flood inundation
939 modelling. *Journal of Hydrology*, **387**(1-2), 33–45.

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

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

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

946 Di Baldassarre, G., Schumann, G., and Bates, P. (2009a). Near real time satellite
947 imagery to support and verify timely flood modelling. *Hydrological Processes*,
948 **23**(5), 799–803.

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

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

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

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

958 Horritt, M. S., Mason, D. C., and Luckman, A. J. (2001). Flood boundary delin-
959 eation from synthetic aperture radar imagery using a statistical active contour
960 model. *International Journal of Remote Sensing*, **22**(13), 2489–2507.

961 Mason, D. C., Horritt, M. S., Dall’Amico, J. T., Scott, T. R., and Bates, P. D.
962 (2007). Improving river flood extent delineation from synthetic aperture radar
963 using airborne laser altimetry. *Ieee Transactions on Geoscience and Remote
964 Sensing*, **45**(12), 3932–3943.

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

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

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

974 Neal, J., Schumann, G., Fewtrell, T., Budimir, M., Bates, P., and Mason, D.
975 (2011). Evaluating a new lisflood-fp formulation with data from the summer
976 2007 floods in tewkesbury, uk. *Journal of Flood Risk Management*, **4**(2), 88–95.

977 NRC, N. R. C. (2006). Completing the forecast: Characterizing and communicat-
978 ing uncertainty for better decisions using weather and climate forecasts.

979 Pappenberger, F., Matgen, P., Beven, K. J., Henry, J.-B., Pfister, L., and
980 Fraipont de, P. (2006). Influence of uncertain boundary conditions and model
981 structure on flood inundation predictions. *Advances in Water Resources*, **29**(10),
982 1430–1449.

983 Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R., and Matgen, P.
984 (2007a). Fuzzy set approach to calibrating distributed flood inundation models
985 using remote sensing observations. *Hydrol. Earth Syst. Sci.*, **11**(2), 739–752.

986 Pappenberger, F., Beven, K., Frodsham, K., Romanowicz, R., and Matgen, P.
987 (2007b). Grasping the unavoidable subjectivity in calibration of flood inundation
988 models: A vulnerability weighted approach. *Journal of Hydrology*, **333**(2-4),
989 275–287.

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

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

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

997 Schumann, G., Cutler, M., Black, A., Matgen, P., Pfister, L., Hoffmann, L., and
998 Pappenberger, F. (2008). Evaluating uncertain flood inundation predictions with
999 uncertain remotely sensed water stages. *International Journal of River Basin*
1000 *Management*, **6**(3), 187–199.

1001 Stephens, E. M., Bates, P. D., Freer, J. E., and Mason, D. C. (2012). The impact
1002 of uncertainty in satellite data on the assessment of flood inundation models.
1003 *Journal of Hydrology*, **414-415**, 162–173.

1004 Stephens, E. M., Bates, P. D., and Schumann, G. (2014). Problems with binary
1005 pattern measures for flood model evaluation. *Hydrological Processes*.

1006 Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T. (2008). Two extra compo-
1007 nents in the brier score decomposition. *Weather and Forecasting*, **23**(4), 752–757.

1008 Toth, Z Talagrand, O., Candille, G., and Zhu, Y. (2003). *Probability and Ensemble*
1009 *Forecasts*. John Wiley & Sons, Ltd., Chichester.

1010 Werner, M., Blazkova, S., and Petr, J. (2005). Spatially distributed observations in
1011 constraining inundation modelling uncertainties. *Hydrological Processes*, **19**(16),
1012 3081–3096.

1013 Wright, N. G., Asce, M., Villanueva, I., Bates, P. D., Mason, D. C., Wilson, M. D.,
1014 Pender, G., and Neelz, S. (2008). Case study of the use of remotely sensed data
1015 for modeling flood inundation on the river severn, uk. *Journal of Hydraulic*
1016 *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

Measure	Aerial Photography			Wrack Marks		
	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.

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

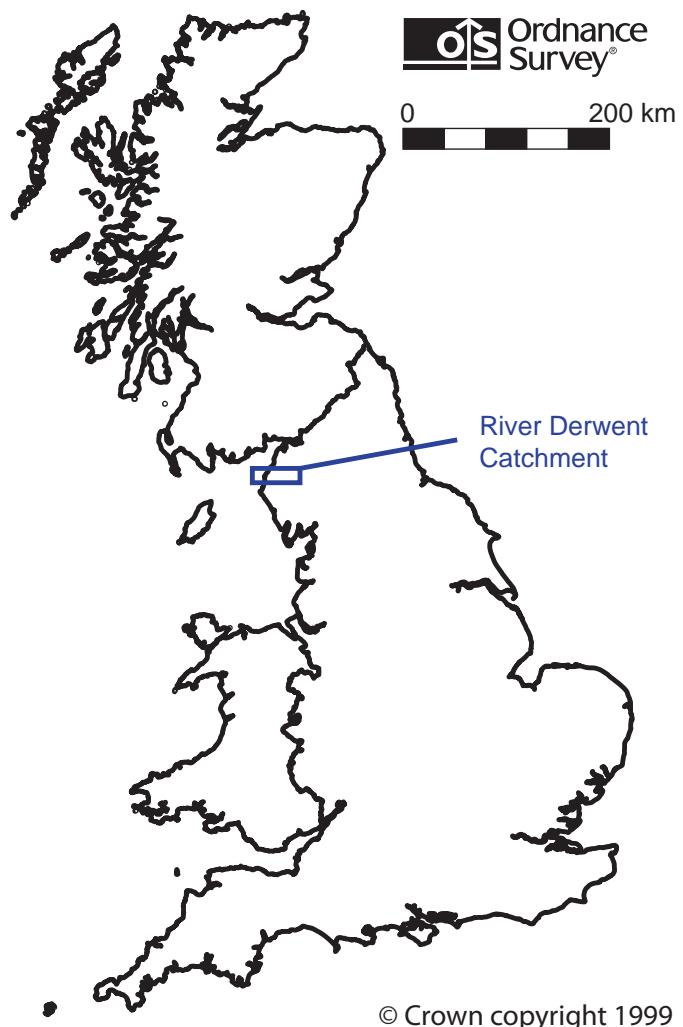


Figure 1: Location map showing the River Derwent catchment in the north-west of England. Source: Ordnance Survey

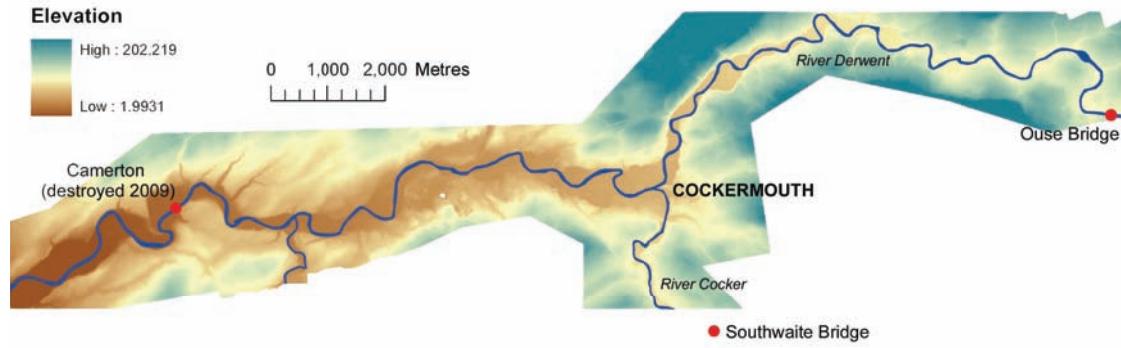


Figure 2: Topographic map of the River Derwent using LiDAR data at 2m resolution, showing location of gauges (red points). Source: Environment Agency

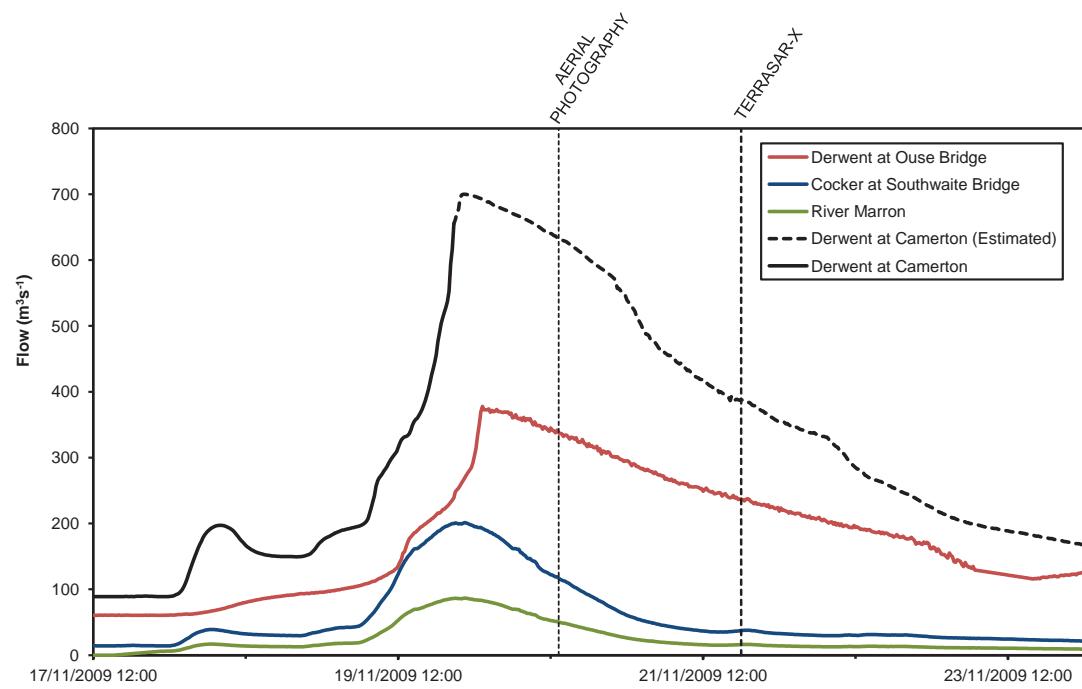


Figure 3: Gauged upstream flows for the River Derwent at Ouse Bridge, the River Cocker at Southwaite Bridge and the River Marron, with gauged downstream flows for the River Derwent at Camerton. Source: Environment Agency

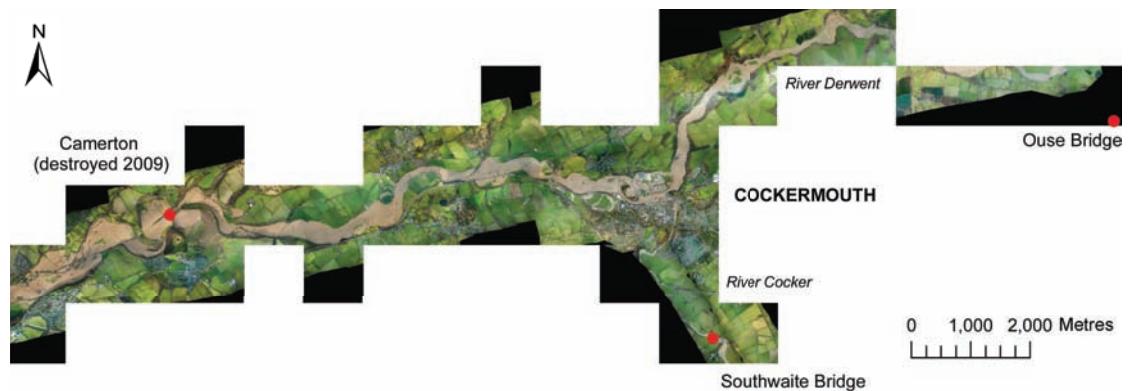


Figure 4: Extent of the aerial photography flown during the flood event. Source: Environment Agency

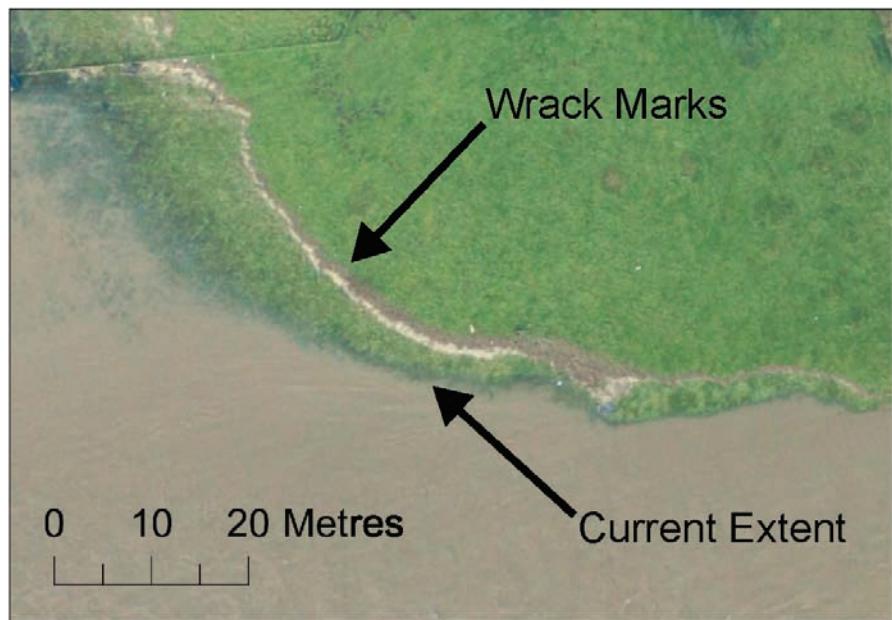


Figure 5: Example of wrack marks visible in the aerial photography adjacent to the then-current flood extent. Source: Environment Agency

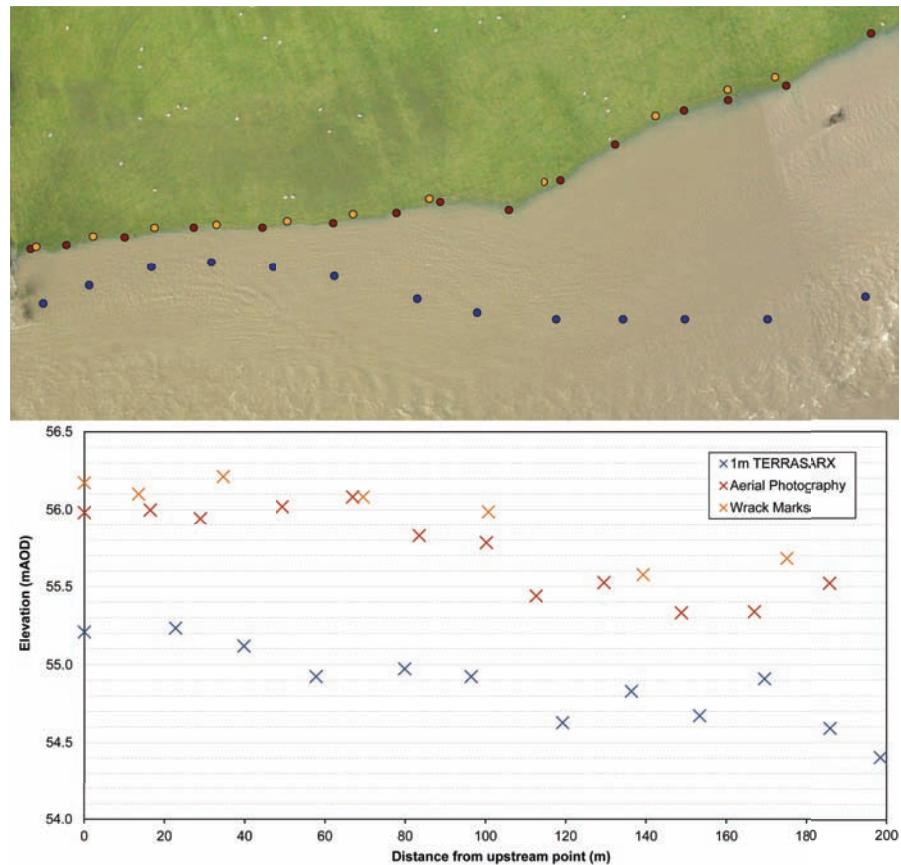


Figure 6: Demarked points along the margin of the flood along a field, with associated elevations derived by intersecting with LiDAR topographic data.

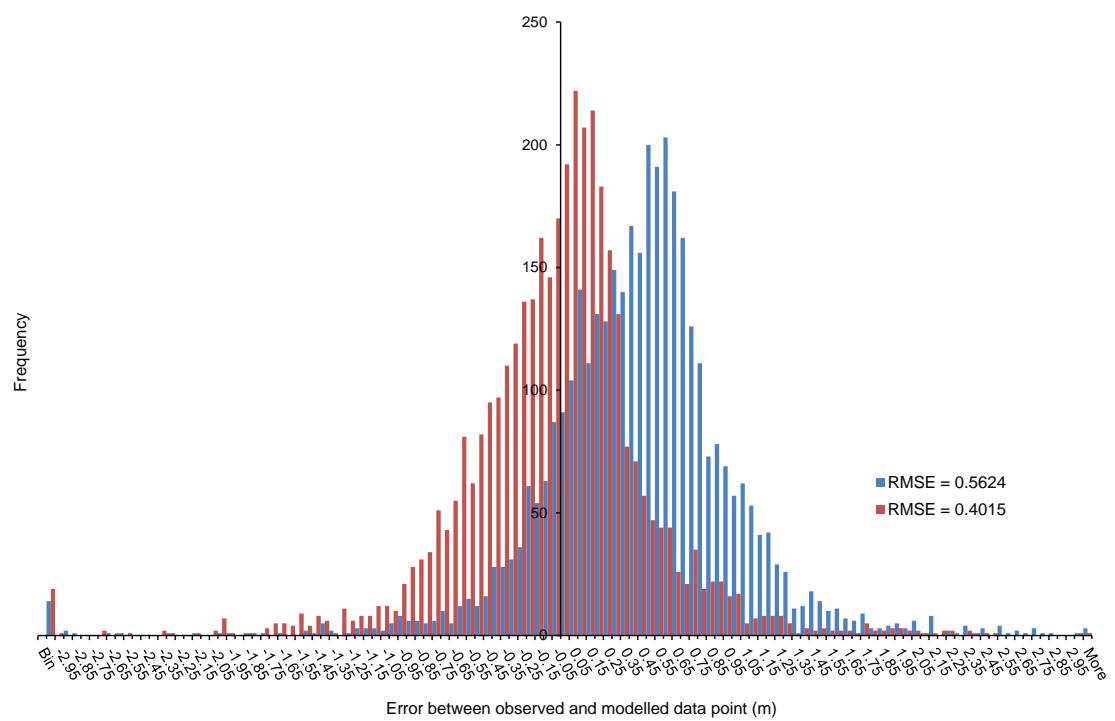


Figure 7: Frequency of error between individual observed and modelled data points, for two parameter sets with RMSEs of 0.5624 (blue) and 0.4015 (red).

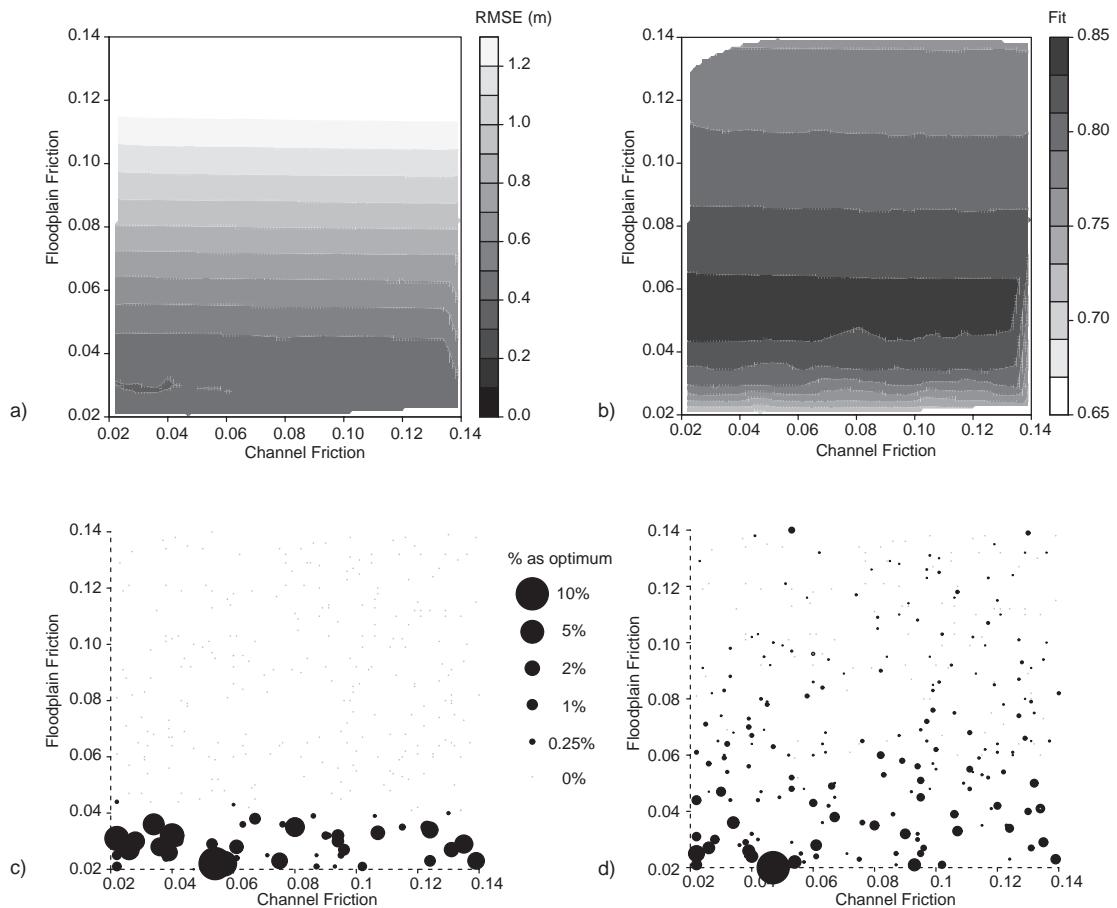


Figure 8: Parameter spaces for calibration of channel (x-axis) and floodplain (y-axis) friction parameters using Aerial Photography with the performance measures of: a) RMSE; b) CSI; c) Percentage as optimum parameter set for subsets of 50 points; and d) c) Percentage as optimum parameter set for all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

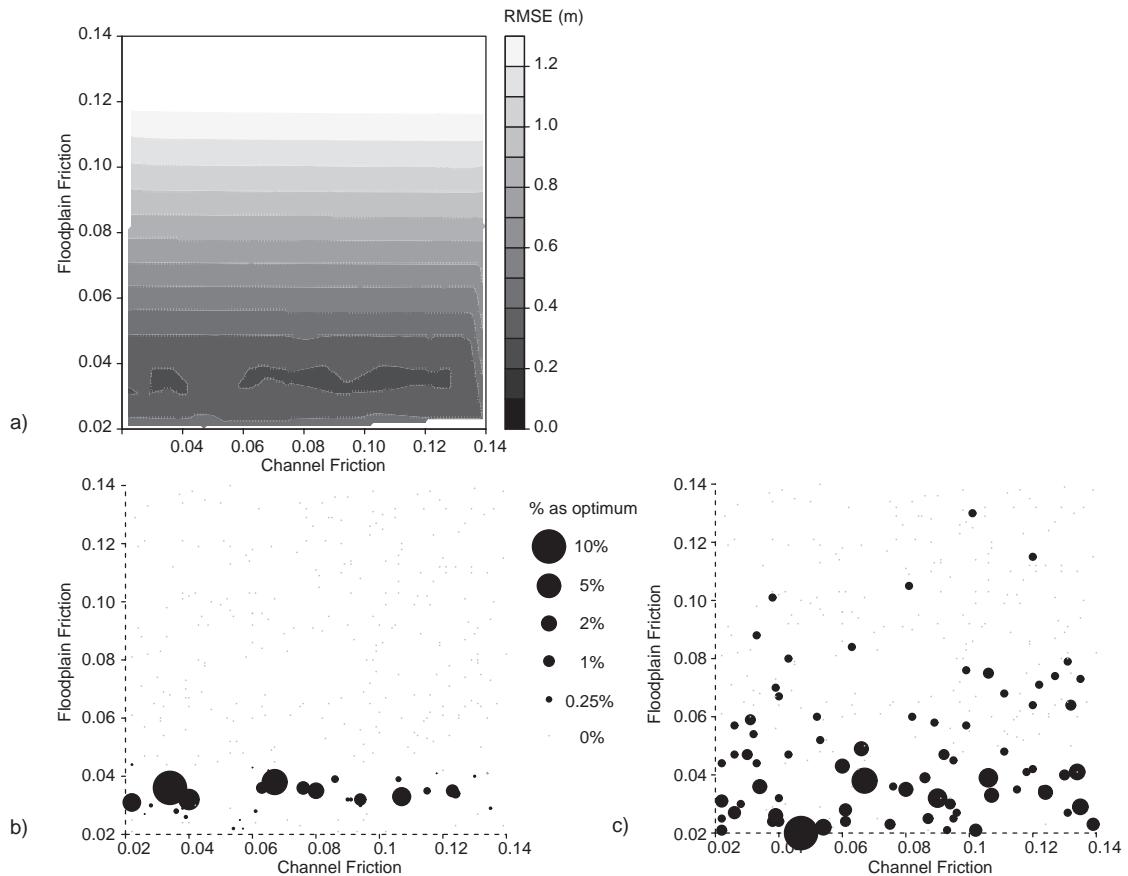


Figure 9: Parameter spaces for calibration of channel (x-axis) and floodplain (y-axis) friction parameters using Wrack Marks with the performance measures of a) RMSE; b) Percentage as optimum parameter sets for subsets of 50 points, and; c) Percentage as optimum parameter set for all individual points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

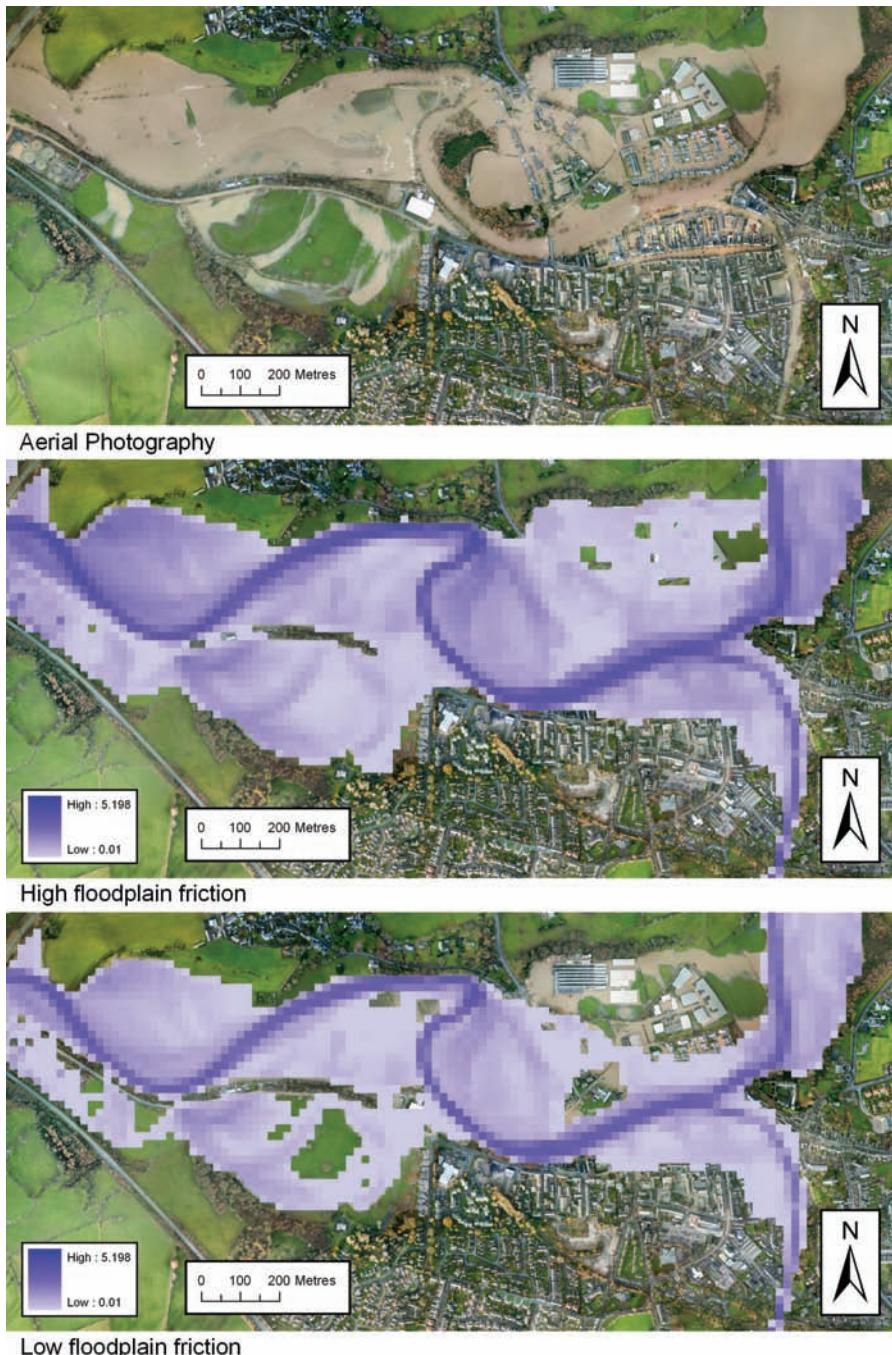


Figure 10: Difference in modelled extent compared to aerial photography for a high and low floodplain friction parameter sets on a subsection of the domain covering the Cockermouth area.

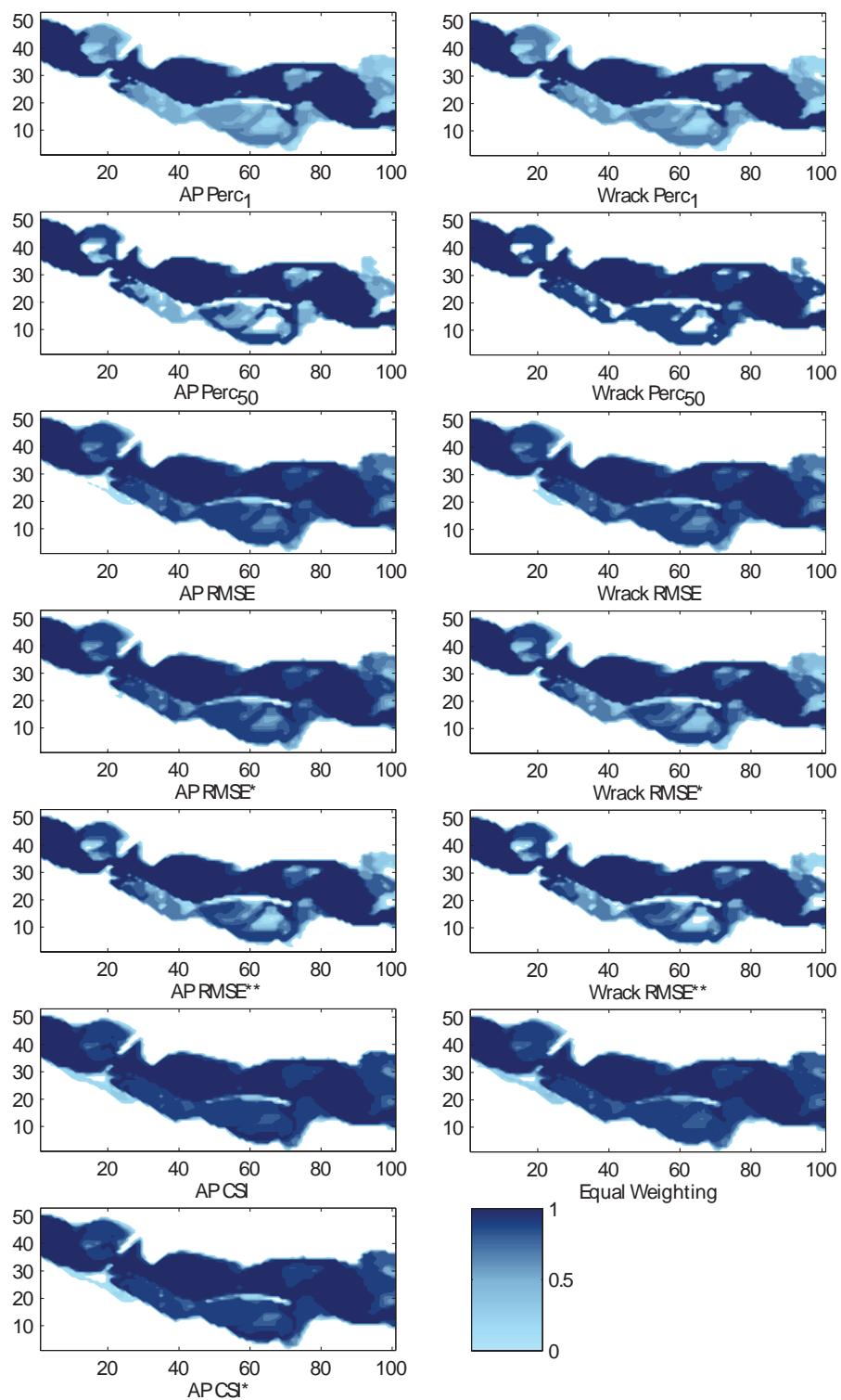


Figure 11: Cut-out from Probability of Inundation maps for the time of a Terrasar-X overpass (see 3). Showing the subtle differences in the mapped probabilities with the different weighting methods used for their construction.

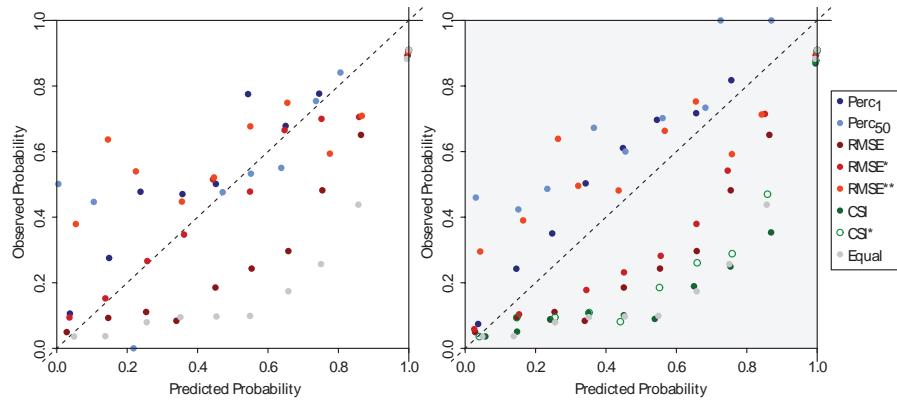


Figure 12: Horritt Reliability at the time of aerial photography overpass using calibrated weightings from 1) peak flood (wrack marks) and 2) aerial photography extent elevations. Greyed out plot indicates where the calibration / validation data are the same.

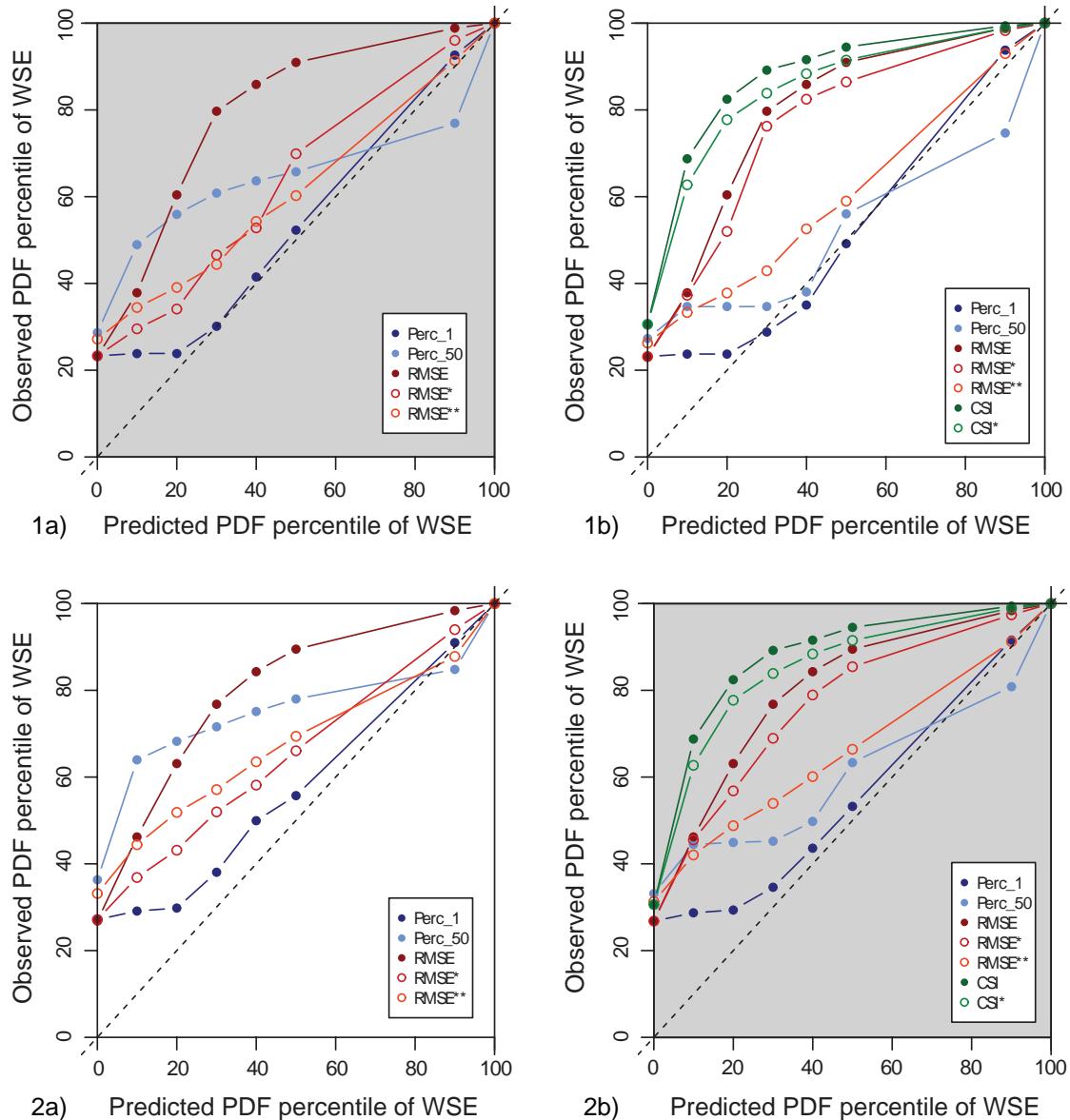


Figure 13: WSE Reliability for 1) Flood Peak using a) Wrack Marks, b) Aerial Photography, and 2) Time of Aerial Photography using a) Wrack Marks and b) Aerial Photography. Greyed out plots indicate where the calibration / validation data are the same

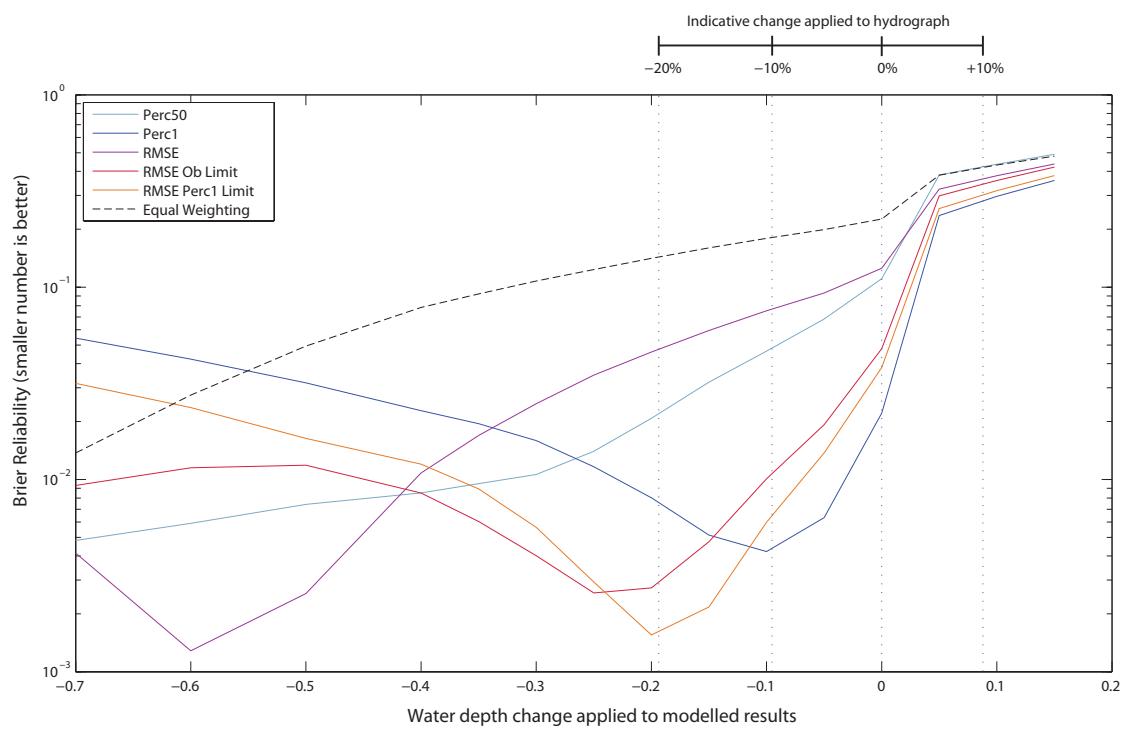


Figure 14: Change in Brier Reliability for different weighting methods if water depths are added / taken from the model results to represent boundary condition uncertainty. Bar along top gives indication of change in depths for different percentage change to flows.