Citation for published version:
Kjeldsen, T & Prosdocimi, I 2018, 'Assessing the element of surprise of record-breaking flood events', Journal of Flood Risk Management, vol. 11, no. S1, pp. S541-S553. https://doi.org/10.1111/jfr3.12260

DOI:
10.1111/jfr3.12260

Publication date:
2018

Document Version
Peer reviewed version

Assessing the element of surprise of record-breaking flood events

[1]Thomas Rodding Kjeldsen and [2]Ilaria Prosdocimi

[1]Department of Architecture and Civil Engineering, University of Bath, Bath, BA2 7AY, UK, t.r.kjeldsen@bath.ac.uk

[2]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK

Abstract

The occurrence of record-breaking flood events continuous to cause damage and disruption despite significant investments in flood defences, suggesting that these events are in some sense surprising. This study develops a new statistical test to help assess if a flood event can be considered surprising or not. The test statistic is derived from annual maximum series (AMS) of extreme events, and Monte Carlo simulations were used to derive critical values for a range of significance levels based on a Generalized Logistic distribution. The method is tested on a national dataset of AMS of peak flow from the United Kingdom, and is found to correctly identify recent large event that have been identified elsewhere as causing a significant change in UK flood management policy. No temporal trend in the frequency or magnitude of surprising events was identified, and no link could be established between the occurrences of surprising events and large-scale drivers. Finally, the implications of the findings for future research needs into the most extreme flood events are discussed.

1. Introduction

Despite substantial human endeavours and financial investments in flood protection infrastructure, the occurrence of floods continues to cause widespread damage and disruption around the World (Kron, 2015). Large flood events are, of course, not a unique contemporary phenomenon, and accounts of several past events have been published in the scientific literature (Macdonald and Black, 2010) in some cases dating back millennia (England et al., 2010). The notion of flood risk management accepts the inability to determine the exact magnitude of future floods and therefore design and planning decisions are often based on pre-specified levels of probability, such as the flood magnitude with a return period of 100-years (Plate, 2002). It is therefore implicitly acknowledged that a larger event can occur. When a large-scale extreme event does occur it is therefore relevant from an operational perspective to determine if such an event should be considered a surprise, or if it was within the range of events that could have been reasonably anticipated based on the information on the flood risk available just before an event. For example, Miller et al. (2013) reported that a large flood observed in November 2009 in the English Lake District had a return period between 33,400 years and somewhere in excess of 50,000 years when based on the available 50 years of at-site annual maximum peak flow data only. This suggests a very rare event indeed, but would it be reasonable at all to expect an event of this magnitude given the past record of flood events? Similar problems of assessing the rarity of very extreme hydro-meterological events from relatively short records were discussed by Coles and Pericchi (2003) and Viglione et al. (2013). These examples demonstrate the difficulty of using traditional flood frequency methods for assessing the rarity of extreme events and to assess if these events could reasonably have been anticipated based on available records, or if the magnitude of the event was a surprise. According to Itti and Baldi (2009) *First, surprise can exist only in the presence of uncertainty.*

Uncertainty can arise from intrinsic stochasticity, missing information, or limited computing resources. A world that is purely deterministic and predictable in real-time for a given observer contains no surprises. Second, surprise can only be defined in a relative, subjective, manner and is related to the expectations of the observer. Fiering and Kindler (1984) discussed the potential for developing a surprise criterion for use in the analysis of water resources systems and included aspects such as, for example, institutional surprises due to changing legislative requirements or structural collapse of components under stress. Interestingly, they argued that the occurrence of a very extreme events should not necessarily be considered a *surprise* but merely as an instance of *bad luck*, as it can be interpreted as a manifestation of an event located far out on the tail of the flood distribution. However, this argument appears to suggest that the flood distribution is correctly specified, whereas in practise it will have been estimated based on the available (and often short) flood records which might not consider sufficient information to capture the true flood risk. For example, a short flood series might not contain information on all possible types of events that can occur at the specific location. Bier et al. (1999) refer to 'counter expected' and 'unexpected events', where the former type of events have previously been rejected as being impossible, whereas the latter events were never even anticipated (unknown unknowns). With reference to the definition of a surprise offered above by Itti and Baldi (2009), we argue here that a reasonable man could indeed be surprised by a large event if previous evidence suggests that an event of this magnitude could occur with a very small probability akin to, for example, the chance of winning the main prize on a single lottery ticket, even if it somehow could be related to a point on the far end of the tail of an estimated flood distribution. This is an important consideration as flood management policy is often developed in response to public demands for action following large-scale severe and disruptive events (e.g. Samuels et al., 2006), exceeding the design specifications of the existing infrastructure installations and inundating communities not previously considered at risk of flooding. For example, Johnson et al. (2005) argued that recent flood policy developments in England and Wales were developed in response to public demands for action following large-scale severe and disruptive events. Others have highlighted the importance of evaluating the performance of existing emergency response procedures following surprisingly large events (e.g. Litman, 2006) and to produce evidence-based future improvements in flood management policies (Thieken et al., 2007). Others again have studied the change in attitude towards flood risk among communities previously flooded, and attributed reductions in flood damage to lessons learned from previous events (e.g. Wind et al., 1999; Burn, 1999).

Following the discussion of what constitutes a surprising event, it is natural to ask if there is evidence of such events becoming more frequent (i.e. less surprising) as a result of climate change. Another related question is if the surprising events are a result of a particular set of circumstance. For example, Lavers et al. (2011) showed that the largest winter flood events in selected British catchments coincided with the occurrence of atmospheric rivers influencing the rainfall. If the surprising events can be attributed to particular mechanisms, then the will cease to surprising.

The objective of this paper is to develop a simple and operational index to help assess if an event can be considered a surprising event based on the magnitude relative to previously observed events. Using a national dataset consisting of annual maximum series of instantaneous peak flow, the objective of this study is to investigate which flood events captured by the gauging network in the United Kingdom (UK) could be considered surprising events: (i) at the time of their occurrence, and (ii) if the same events happened today. The analysis will be based on a relatively simple index of surprise and the results compared to the timing of recent Government flood management policy initiatives to assess the degree to which such policy are drawn-up in response to surprising events. The index will also investigate if the frequency of surprising events has increased, and if they can be linked to large-scale drivers.

2. Measuring the level of surprise

The starting-point for this analysis is that in order for a large event to be considered a surprising event it should be larger than any previously observed events, i.e. it must be a record-breaker. A simple way of classifying a record-breaker is by using order statistics. Consider a sample of n annual maximum events is available x_i, i=1,…,n with the associated ordered series ($x_{[1]} \leq x_{[2]} \leq \ldots x_{[n]}$). A new observation y is considered a record-breaker if it is larger than the previous record, i.e.

$$y \geq x_{[n]} \tag{1}$$

One possible mathematical definition of a surprise could measure if y can be considered an outlier of the distribution responsible for generating the available annual maximum events available so far. According to Hawkins (1980) outliers can be caused by a number of different mechanisms: (i) the annual maximum data originate from an *outlier prone* distribution (Green, 1976), i.e. unexpected large event can occur especially if only short samples are available, (ii) a different mechanism is responsible for the occurrence of y not previously observed in the sample (e.g. Rossi et al., 1984). A possible addition (which might be considered a subset of ii) is that the distribution of the annual maximum data are changing over time (e.g. as a result of climate change) so that the probability of observing y becomes more likely as time progresses. Finally, the reported magnitude of the record-breaker might be inaccurate, as the most extreme events are often the most difficult to measure, but this aspect is not pursued further here.

There is, of course, an abundant literature on the identification of outliers in statistical analysis (e.g. Hawkins, 1980; Hodge and Austin, 2004), which mostly define a point as an outlier as compared to a (parametric) model which is assumed to be underlying the process under study. In this study, the focus is on the identification of events which might be considered surprising, rather than on the identification of outliers in a statistical sense; a relatively simple non-parametric method was chosen to enable a transparent application to national datasets of annual maximum instantaneous peak flow observations.

2.1 An operational definition of a surprise

Solow and Smith (2005) introduced an index of surprise r where the surprise of the new record-breaker y exceeding the previous record $x_{[n]}$ events is measured relative to the previous record margin, i.e.

$$r = \frac{y - x_{[n]}}{x_{[n]} - x_{[n-1]}} \tag{2}$$

If x is Gumbel distributed, then the random variable R is distributed as

$$P(R > r) = \frac{2}{r+2} \tag{3}$$

from which a critical value can be derived for the null-hypothesis that a new record-breaker y is generated from the same distribution as gave rise to the previous values of $x_{[n]}$ and $x_{[n-1]}$ against the alternative hypothesis that y originates from a process that gives raise to larger events than previously observed, i.e. a different underlying flood distribution in this case. Solow and Smith (2005) also introduced a version of the test statistic which made use of the top k ranked events $x_{[n-k]} \leq \cdots \leq x_{[n]}$ as

$$t_k = \frac{y - x_{[n]}}{\left(y - x_{[n]} + \sum_{j=0}^{k-2} (j+2)\left(x_{[n-j]} - x_{[n-j-1]}\right) \right)} \quad (4)$$

The random variable T_k is distributed as a beta distribution:

$$P(T_k > t_k) = (1 - t_k)^{k-1} \quad (5)$$

which, again, can be utilised to derive a critical value for a given significance level. This result is exact when the events follow an exponential distribution. This version of the index was used to assess the surprise of an athletic records and the age of a newly discovered cave painting (Solow and Smith, 2005) and to assess if the recent sighting of a presumed extinct type of wild cat could be the result of animals being released into the wild or not (Solow et al., 2006). In this study the index will be used to identify past flood events in the UK which could reasonably have been labelled as surprising given the observed series. However, as discussed in the next section, the distributional assumptions underpinning Eq. (5) are not fulfilled when considering annual maximum series of peak flow, and thus the test must be modified accordingly.

2.2 Response surface for critical values

As annual maximum series of flood events in the UK are routinely modelled using a Generalised Logistic (GLO) distribution (Institute of Hydrology, 1999), the critical level of the surprise index t_k derived from Eq.(4) was not considered suitable. Therefore, a set of Monte Carlo experiments were conducted to derive a set of regression models enabling prediction of critical values for selected significance levels ($\alpha = 20\%, 15\%, 10\%, 5\%$ and 1%) under the GLO assumption, for a range of record-lengths and shape parameters.

Without loss of generality, samples were generated from GLO distributions with location and scale parameters set to 0 and 1, respectively and with shape parameters assigned the following values κ= -0.4, -0.3, -0.2, -0.1, -0.05, +0.05, +0.10, +0.20, +0.30, +0.40. For a given parameter set, a total of 100,000 samples were generated with sample size of n=10, 15, 20, 25, 30, 35, 40, 50, 100. For each sample the critical value was determined as a specified quantile in the empirical sampling distribution of t_k estimates. Following the procedure Tolikas and Heravi (2008) and Heo et al. (2013), to avoid fitting individual regression models for each individual value of the shape parameter and to allow interpolation, a linear response surface was fitted to the entire simulation output

$$t_k(\alpha) = \beta_0 + \beta_1\left(\frac{1}{n}\right) + \beta_2\left(\frac{1}{n^2}\right) + \beta_3\kappa + \beta_4\kappa^2 \tag{6}$$

where $t_k(\alpha)$ is the critical value for chosen significance level, n is record-length, and κ is the shape parameter. The model parameters are reported for a range of significance levels for the GLO (Table 1), and Figure 1 shows an example of Eq.(6) fitted to the critical values obtained for the GLO distribution using Monte Carlo simulations. The hatched horizontal line represent the critical value as derived from Eq.(5). Note that all model parameters are significantly different from zero.

TABLE 1

FIGURE 1

As most GLO distributions fitted to UK flood series have negative shape parameters significantly different from zero, the use of critical values derived from Eq.(5) are generally too low and therefore will too readily accept an event as being surprising. The need to evaluate the statistical test based on distributional assumptions using Eq.(6) is less appealing than the elegant analytical solution provided by Eq.(5). But given the widespread acceptance of the

GLO distribution for flood frequency analysis in the UK, the use of Eq.(6) rather than Eq.(5) is considered only a minor inconvenience necessary to avoid high rates of incorrect detections.

3. Case study: Surprising events in the UK

The surprise index in Eq.(4) for k=5 was applied to a database of annual maximum series of instantaneous peak flow contained in the HiFlows-UK database v.3.3.4 available from the National River Flow Archive.

3.1 Annual maximum peak flow data

A total of 852 annual maximum series of peak flow are considered of sufficiently high quality to be used in flood frequency analysis are available from the HiFlows-UK database hosted by the NRFA. The version of the database used in this study include annual maximum events up-to and including the water year 2011, except for gauging stations located in Scotland where data are only available up-to (and including) the water-year 2007. The locations of the gauging stations are shown in the map on Figure 2 indicating a reasonably even geographical spread throughout the country with the exception of the relatively sparsely populated areas such as, for example, the Scottish Highlands.

FIGURE 2

FIGURE 3

A time series plot showing the number of events available in each year is shown in Figure 3 There was considerable growth in the number of gauging station from the mid-1960s onwards, reaching a reasonably stable number from the mid-1970 onwards.

3.2 Past surprises

The index of surprise was estimated for each of the 852 annual maximum series using the following approach. First, the largest recorded event on record y was identified for each series together with the year of occurrence. Next, the largest observations $\left(x_{[n]}, x_{[n-1]}, \ldots x_{[n-k]}\right)$ were identified in the n years preceding the year in which y occurred. The years following y were discounted as the analysis is designed to represent the level of surprise assigned to each event at the time of occurrence. Finally, the index of surprise is estimated using Eq.(2) and Eq.(4) for k=5, and the results summarised in Figure 4. A minimum record-length of 7-years prior to a record-breaker was imposed to the analysis resulting in a reduction from 852 to 791 catchments.

FIGURE 4

From Figure 4 it is clear that the version of the index in Eq.(2) based on only the two previous highest values is not suitable for application to a large-scale national dataset. The range of values obtained using this version of the index is substantial, and large values are often caused by a tie (or almost a tie) of the two previously highest values $x_{[n]}$ and $x_{[n-1]}$. This problem disappears when using the version of the index based on the $k=5$ previous values. For the remainder of this study, the index with k=5 was chosen; similar to Solow and Smith (2005) and Solow et al. (2006).

Comparing the sample values of t_5 obtained for each of the 852 series (Eq. 4) with the critical interval for a significance level as derived based on record-length and estimated shape-parameter (Eq. 6) a subset of surprising events was identified. Initial experiments highlighted that the sampling variability of the GLO shape parameter, κ, was causing excessive variability in the estimates of the critical interval. More reliable estimates of the shape-parameter was obtained by deriving the regional averages of L-skewness. For each gauging station, the corresponding geographical region, as defined by the Flood Studies Report (NERC, 1975), was identified and the regional average L-skewness parameter derived using only observations up-to (but not including) the year in which the record-breaking event was recorded. Thus, the dataset used for estimating the shape parameter is uniquely defined for each record-breaking event. Finally, the GLO shape parameter is estimated using the regional L-skewness as outlined by Hosking and Wallis (1997). Next, the events at the individual gauging stations were grouped together into *events* by combining all series where the surprising events occur within the same 7-day window. Figure 5 shows the geographical location of gauging stations where a surprising event was identified for four different levels of significance: 0.15, 0.10, 0.05 and 0.01.

FIGURE 5

Events where four or more gauging stations record a surprising event within the same 7-day window are highlighted in colour, whereas stations with a grey dot experienced a surprising event, but the event was recorded at less than four locations. As expected, the higher the significance level, the more events are classified as being surprising. At a significance level of 0.01, there are relatively few events classified as surprising, and no surprising event recorded at four or more sites simultaneously. Conversely, for higher significance levels such as $p = 0.10$ and $p = 0.15$, there are numerous events highlighted. To identify an operational definition of a surprise, a list of events was created based on evidence that these events had resulted in some form of change in UK flood management policy. Table 2 shows the correspondence between the Johnson et al. (2005) events and the automatically identified events, including a short description of the resulting policy change. This list is mostly based on the list of catalyst events discussed by Johnson et al. (2005). The Table includes the event that occurred in March 1947 but as evident from Figure 3, only very few gauging stations were operational at that time. Thus, despite the important role of this event in changing flood management at the time, it is not considered further in this study. The June and July flood events of 2007 happened after Johnson et al. (2005) published their results, but as this event has been an important driver for change in flood policy (Pitt, 2008), it has been included in this study. Notably, events such as: September 1968, December 1979, October 1987 are all classified as surprising but were not considered by Johnson et al. (2005). The November 2009 (Stewart et al., 2012; Miller et al., 2013) was not considered either, but again, this event occurred after the study of Johnson et al. (2005) was published. In addition to the catalyst-events listed in Table 2, there might be changes to flood policy that were initiated for reasons other than as a response to a major flood event and therefore not considered. Finally, any link between the specific location of the flooding and the initiation of a policy change is considered outside the scope of this study.

From Figure 5 it can be seen that for $p = 0.10$, all the events in Table 2 (April-1998, November-December 2000 and July-2007) have been highlighted in colour (along with September-1968, December-1979, October-1987 and November-2009), flagging that these events have been identified as surprising at four or more gauging stations. Adopting the $p = 0.05$ or 0.01 levels, the criterion for a surprise is too strict to highlight these events over other more localised events. Notably, both the September-1968 and the December-1979 events have been identified for $p = 0.10$ as a surprising and widespread events, yet the authors could not identify published reviews containing details of this event. For the remaining parts of this study a critical threshold corresponding to $p = 0.10$ and records recorded simultaneously at a minimum of four gauging stations is therefore chosen here as defining a surprising event. This resulted in 121 surprising records across the 852 gauging stations. Of the 121 surprising record-breakers, 39 were recorded at a single gauging station only within 7-days, 10 were recorded at two gauging stations, 4 were recorded at three gauging stations, and 9 were recorded at four or more gauging stations, resulting in a total of 62 individual events.

3.3 Contemporary surprises

Next, a numerical experiment was conducted by moving the record-breaking event at each station from its current location in the sample, to the end of the sample. This is synonymous with assessing the level of surprise of the same events if they were to occur at a time where all contemporary information is available. As in the previous assessment, a minimum record-length of 7-years was imposed, and the shape parameter of the GLO distribution is estimated using the average regional L-skewnness from each hydrometric region using all available annual maximum data, but excluding the year of the record-breaker itself. This experiment resulted in a total of 62 surprising record-breakers from 834 gauging stations (with more than 7-year of data). As expected the increased length of the data series available prior to the record-breaking event has resulted in an overall reduction in the amount of surprising events (down 59 from 121 to 62), highlighting the value of maintaining a flood flow monitoring and archiving programme. The location of the surprising events is shown in Figure 6, highlighting events where four or more surprising events were recorded in the same 7-day window.

FIGURE 6

Comparing Figures 5 and 6 it can be seen that four of the initial nine large-scale events (see map for p=0.10 in Figure 5) would still be considered a surprising (Sep-1968, Dec-1979, Jun-2007, Jul-2007) when based on contemporary experience of past floods. Interestingly, more sites recorded surprising events in 1968 when considering the complete record. This is due to the required availability of a minimum of 7-year record prior to the event which excluded a number of gauging stations in the first analysis.

Notably, most of the surprising events shown in both Figures 5 and 6 have been recorded in the southern part of the UK. It is not clear to what degree this is caused by differences in the density of the gauging network, or regional differences in the flood hydrology making the southern part of the country more prone to surprisingly large events.

4. Non-stationarity of surprising events

This section will investigate if changes in the magnitude and frequency of the record-breaking events can be detected over the recent time period. Figure 7 shows the number of gauging stations within each of the 62 surprising past record-breaking events plotted against the timing of the event. Blue coloured bars indicate a winter event (Oct-Mar) and red bars indicate a summer event (Apr-Sep).

FIGURE 7

Using only data from 1975 onwards to minimise the effect of varying data availability across years (as shown in Figure 3), a Poisson regression model was fitted to the data shown in Figure 7, describing the number of sites recording a surprise within each event using time as an exploratory variable. Three different models were considered: (i) using all events, (ii) winter events only, and (iii) summer events only. No significant relationship (trend) was found at the 0.05 confidence level when using all events nor for either winter or summer events only. It is therefore not possible to conclude from this analysis alone that the number of surprising events has increased or decreased over the considered time window.

5. Review of external drivers of surprising events

As evident from Figure 7, surprising events are recorded almost every year at one or more gauging stations in the United Kingdom. While a detailed investigation of the exact meteorological and hydrological circumstances characterising each of these events is beyond the scope of this study, it is none the less of interest to try to link the occurrence of surprising events to large-scale drivers. Previous studies have suggested that elevated flood levels might be connected to phenomena such as: the North Atlantic Oscillation (e.g. Hannaford and Marsh, 2008), solar magnetic activity (Macdonald, 2014) and atmospheric rivers (Lavers et al., 2011; 2012).

For example, in a study of extreme winter flood events at selected gauging stations in the UK, Lavers et al. (2011) found that the largest winter flood events at selected gauging stations coincided with atmospheric rivers. However, the annual maximum flow data available at the gauging station for which results were reported by Lavers et al. (2011) did not report a surprising event in this study. Furthermore, most of the events (7 out of 10) identified by Lavers et al. (2011) as being driven by atmospheric rivers did not result in a surprising events at any gauging stations across the UK; notable exceptions were the 03 January 1982, 07 January 2005 and 19 November 2009. Interestingly, none of the nine flood records used in the follow-up study by Lavers et al. (2012) recorded a surprising event in this study. These results do not suggest that the results by Lavers et al. (2011) are not valid, but rather that the effect of atmospheric rivers is most likely subsumed within the general year-to-year variability of the annual maximum peak flow series and therefore falls within the range of events expected from the GLO distribution. Clearly, further research is needed to better understand the implications of these findings for flood frequency analysis practise, and if more sophisticated modelling tools should be developed to better represent known atmospheric drivers, helping to better anticipate events such as the November 2009 event within flood risk analysis.

6. Discussion and conclusion

This study has attempted to derive a simple but operational index for identifying a surprising flood event by combining a national-scale data set of extreme floods with evidence of flood policy changing as a result of large-scale flooding. The results shows that in order for an event to be classified as surprising it needs to be both unexpectedly large and occurring in several locations simultaneously. Based on the ability to highlight particular flood events, simple statistical test of whether an event is surprising or not was developed and applied at a significance level of p = 0.10 while also being recorded at a minimum of four gauging stations within a common 7-day period. The threshold of four stations used in this study was found to be appropriate for the density of the gauging network in the United Kingdom to define large-scale events driving policy change. It is likely that other regions with more or less dense gauging network might find other threshold values more suitable.

It is noteworthy that for a significance level of p = 0.10, a total of 121 surprising events were identified out of a possible of 852, corresponding to 14.2% of all gauging stations reporting a surprising record-breaker. The most spatially extensive of these events coincide with the most recent policy-changing events. However, the fact that 10% of gauging stations were expected to report a surprising event even if all events are derived from an underlying GLO distribution suggesting a small tendency to observe more surprising events than expected. This result could indicate the existence of flood generating processes causing more extreme events in some years than others. However, no temporal trend in the occurrence of surprising events was identified in this study. Likewise, an attempt in this study to link the occurrence of surprising events to the impact of atmospheric rivers was inconclusive. This does not suggest that no link exists between the presence of atmospheric rivers and flood magnitude, but merely that the year to year variability of annual maximum peak flow data used in this study might be too large or the records too short to allow such links to be identified for the largest events. This conclusion was also echoed by Prosdocimi et al. (2015) who advocated the use of more advanced data structures and statistical models to better capture aspects of non-stationarity in flood risk. The results presented here therefore suggest that despite a relatively extensive archive of past flood events from across the UK, it is still very difficult to predict the flood risk with any degree of precision, and thus we continue to be surprised by large events. There are several research avenues that should pursued to further improve the ability to predict flood risk. Notably, the use of historical and documentary evidence is considered useful and valuable across Europe and beyond (e.g. Kjeldsen et al., 2014; O'Connor et al. 2014) and has the potential to reduce the surprising aspect of large events. Another promising approach is to develop new and more advanced statistical models with more explicit links between flood magnitude and external drivers such as climate and land-use change (e.g. Renard and Lall, 2014; Prosdocimi et al., 2015). Modelling systems coupling stochastic rainfall generators with rainfall-runoff models have also been used for estimating very rare events, e.g. for dam safety (Lawrence, 2014). However, such systems suffer from the same fundamental limitations as the statistical approach that they must be calibrated to a dataset of already observed events which might or might not include any surprisingly large events.

The surprise index was deliberately developed as a simple tool to enable identification of surprising events. It has been shown that these events largely correspond to moments in which flood management policies in the UK were amended, suggesting that very large unexpected events can be catalysts for changes in practice. However, the index did also identify events (Septemner-1968, December-1979, and October-1987) where the authors were unable to link the events to policy changes. Finally, it must be acknowledged that not all policy changes are necessarily driven by surprising events, and such changes therefore cannot be identified using an index based on flow records only. For example, the EU Floods Directive must be implemented in all EU member states regardless of whether they have recently experienced a surprising event or not. Also, the index cannot, in the present form, consider the relative importance of the flood location in relation to policy change. However, the gauging network shown in Figure 2 appears to be relatively denser in the more populated areas, and thus the index might have an implicit bias towards identifying surprising events more easily in these areas. In contrast, the Scottish highlands have a relatively low population density and also relatively fewer gauges. It is therefore less likely that a surprising event is identified in this area.

Finally, it should be acknowledged that this study has adopted a definition of surprise from the perspective of an analyst and based purely on flood magnitude. It is possible that a more comprehensive method could be developed by considering surprise in term of both likelihood and vulnerability of communities at risk of flooding. For example, relatively high likelihood events causing large damage might be considered surprising from the perspective of the impacted communities. Surprise could also be defined in terms of sequences of high-flow events, such as experiencing floods in excess of the 100 year event in a relatively short time period.

Acknowledgements: The authors would like to thank Dr Neil Macdonald and another anonymous reviewer for their helpful comments on an earlier version of the manuscript. The HiFlows-UK data used in this study is available from the National River Flow Archive (http://www.ceh.ac.uk/data/nrfa/peakflow_overview.html). The support provided for this research while the second author was employed by the Centre for Ecology & Hydrology is kindly acknowledged.

References

Bier, V. M., Haimes, Y. Y., Lambert, J. H., Matalas, N. C., and Zimmerman, R. (1999) A survey of approaches for assessing and managing the risk of extremes. *Risk analysis*, 19(1), 83-94.

Burn, D. H. (1999), Perceptions of flood risk: A case study of the Red River Flood of 1997, *Water Resour. Res.*, 35(11), 3451–3458.

Coles, S. and Pericchi, L. (2003) Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 52(4),240 405.

England, J. F., Godaire, J. E., Klinger, R. E. and Bauer, T. R. (2010) Paleohydrologic bounds and extreme flood frequency of the Upper Arkansas River, Colorado, USA. *Geomorphology*, 124(1), 1-16.

Fiering, M. and Kindler, J. (1984) Surprise in water-resources design. *International Journal of Water Resources Development*, 2(4), 1-10.

Green, R. F. (1976) Outlier-prone and outlier-resistant distributions. *Journal of the American Statistical Association*, 71(354), 502-505.

Hannaford, J. and Marsh, T. J. (2008), High-flow and flood trends in a network of undisturbed catchments in the UK. *Int. J. Climatol.*, 28: 1325–1338

Hawkins, D. M. (1980) *Identification of outliers*, Volume 11, Springer.

Heo, J.-H., Shin, H., Nam, W., Om, J. and Jeong, C. (2013) Approximation of modified Anderson-Darling test statistics for extreme value distributions with unknown shape parameter. *Journal of Hydrology*, 499, 41-49.

Hodge, V. J. and Austin, J. (2004) A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85-126.

Horner, M. W. and Walsh, P.D. (2000) Easter 1998 floods. *Water and Environment Journal*, 14(6), 415-418.

Institute of Hydrology (1999) *Flood Estimation Handbook*, 5 Volumes, Institute of Hydrology, Wallingford, UK.

Itti, L. and Baldi, P. (2009) Bayesian surprise attracts human attention. *Vision research*, 49(10), 1295-1306.

Johnson, C. L., Tunstall, S. M. and Penning-Rowsell, E. C. (2005) Floods as catalysts for policy change: historical lessons from England and Wales. *Water Resources Development*, 21(4), 561-575.

Kjeldsen, T. R., Macdonald, N., Lang, M., Mediero, L., Albuquerque, T., Bogdanowicz, E., Brazdil, R., Castellarin, A., David, V., Fleig, A., Gul, G. O., Kriauciuniene, J., Kohnova, S., Merz, B., Nicholson, O., Roald, L. A., Salinas, J. L., Sarauskiene, D., Sraj, M., Strupczewski, W., Szolgay, J., Toumazis, A., Vanneuville, W., Veijalainen, N. and Wilson, D., (2014) Documentary evidence of past floods in Europe and their utility in flood frequency estimation. *Journal of Hydrology*, 517, 963-973.

Kron, W. (2015) Flood disasters – a global perspective. *Water Policy*, 17(1), 6–24.

Lavers, D. A., Allan, R. P., Wood, E. F., Villarini, G., Brayshaw, D. J. and Wade, A. J. (2011) Winter floods in Britain are connected to atmospheric rivers. *Geophysical Research Letters*, 38(23).

Lavers, D. A, Villarini, G., Allan, R. P., Wood, E. F. and Wade, A. J. (2012) The detection of atmospheric rivers in atmospheric reanalyses and their links to British winter floods and the large-scale climatic circulation. *Journal of Geophysical Research*: Atmospheres (1984{2012), 117(D20).

Lawrence, D., Paquet, E., Gailhard, J., and Fleig, A. K.: Stochastic semi-continuous simulation for extreme flood estimation in catchments with combined rainfall–snowmelt flood regimes, *Nat. Hazards Earth Syst. Sci.*, **14**, 1283-1298

Litman, T. (2006) Lessons from Katrina and Rita: What major disasters can teach transportation planners. *Journal of Transportation Engineering*, 132(1), 11-18.

Macdonald, N. (2014) Millennial scale variability in high magnitude flooding across Britain. *Hydrol. Earth Syst. Sci. Discuss.*, 11, 10157–10178.

Macdonald, N., and Black, A. R. (2010) Reassessment of flood frequency using historical information for the River Ouse at York, UK. *Hydrological Sciences Journal*, 55(7), 1152-1162.

Marsh, T. J. and Hannaford, J. (2007) The summer 2007 floods in England and Wales – a hydrological appraisal. Centre for Ecology & Hydrology. 32pp.

Marsh, T. J. and Dale, M. (2002) The UK floods of 2000{2001: a hydrometeorological appraisal. *Water and Environment Journal*, 16(3), 180-188.

McEwen, L., Hall, T., Hunt, J., Dempsey, M. and Harrison, M. (2002) Flood warning, warning response and planning control issues associated with caravan parks: the April 1998 floods on the lower Avon floodplain, Midlands region, UK. *Applied Geography*, 22(3), 271-305.

Miller, J. D., Kjeldsen, T. R., Hannaford, J. and Morris, D. G. (2013) A hydrological assessment of the November 2009 floods in Cumbria, UK. *Hydrology Research*, 44(1), 180-197.

O'Connor, J.E., Atwater, B.F., Cohn, T.A., Cronin, T.M., Keith, M.K., Smith, C.G., and Mason, R.R. (2014) Assessing inundation hazards to nuclear powerplant sites using geologically extended histories of riverine floods, tsunamis, and storm surges: *U.S. Geological Survey Scientific Investigations Report* 2014–5207, 66 p.

Paranjothy, S., Gallacher, J., Amlôt, R., Rubin, G. J., Page, L., Baxter, T., Wight, J., Kirrage, D., McNaught, R. and Palmer, S. R. (2011) Psychosocial impact of the summer 2007 floods in England. *BMC public health*, 11(1), 145.

Pitt, M. (2008) The Pitt review: Learning lessons from the 2007 floods. London: Cabinet Office, 2008.

Plate, E. J. (2002). Flood risk and flood management. *Journal of Hydrology*, 267(1), 2-11.

Prosdocimi, I., Kjeldsen, T. R., and Miller, J. D. (2015). Detection and attribution of urbanization effect on flood extremes using nonstationary flood frequency models. *Water Resources Research*, 51(1)

Renard, B. and Lall, U. (2014). Regional frequency analysis conditioned on large-scale atmospheric or oceanic fields. *Water Resources Research*, **50**(12), 9536-9554.

Risk Management Solutions (2007) 1947 U.K. River Floods: 60-Year Retrospective, *RMS Special Report*, 14pp. http://riskinc.com/Publications/1947_UKRiverFloods.pdf (accessed 08 June 2015)

Rossi, F., Fiorentino, M. and Versace, P. (1984) Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, 20(7), 847-856.

Samuels, P., Klijn, F., and Dijkman, J. (2006) An analysis of the current practice of policies on river flood risk management in different countries. *Irrigation and Drainage*, 55(S1), S141-S150.

Solow, A. R. and Smith, W. (2005), How surprising is a new record? *The American Statistician*, 59(2), 153-155.

Solow, A. R., Kitchener, A. C., Roberts, D. L. and Birks, J. D. S. (2006) Rediscovery of the Scottish polecat, mustela putorius: Survival or reintroduction? *Biological conservation*, 128(4), 574-575.

Stewart, E. J. Morris, D. G., Jones, D. A. and Gibson, H. S. (2012) Frequency analysis of extreme rainfall in Cumbria, 16-20 November 2009. *Hydrology Research*, 43(5), 649-662.

Thieken, A. H., Kreibich, H., Müller, M. and Merz, B. (2007) Coping with floods: prepared ness, response and recovery of flood-affected residents in Germany in 2002. *Hydrological Sciences Journal*, 52(5), 1016-1037.

Tolikas, K. and Heravi, S. (2008) The Anderson-Darling goodness-of-fit test statistic for the three-parameter lognormal distribution. *Communications in Statistical Theory and Methods*, 37(19), 3135-3143.

Viglione, A., Merz, R., Salinas, J. L. and Blöschl, G. (2013) Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, 49(2), 675-692.

Wind, H. G., T. M. Nierop, C. J. deBlois, and J. L. deKok (1999), Analysis of flood damages from the 1993 and 1995 Meuse Floods, *Water Resour. Res.*, 35(11), 3459–3465.

Table 1: Response function for the t5 critical values at the 20%, 15%, 10%, 5% and 1% significance levels for the GLO distribution.

	Coefficients					
Significance level	β_0	β_1	β_2	β_3	β_4	R^2
20%	0.392	-2.368	7.908	-0.428	0.215	0.998
15%	0.436	-2.449	7.884	-0.464	0.225	0.998
10%	0.492	-2.477	7.487	-0.501	0.227	0.999
5%	0.574	-2.439	6.803	-0.535	0.213	0.999
1%	0.713	-1.965	3.700	-0.519	0.145	0.995

Table 2: List of large-scale identified a catalysts for policy change

Date	Description	Policy change	Reference
1947 March	Extensive floods resulting from heavy rainfall combined with rapid snowmelt in early March 1947 following one of the coldest and snowiest winters ever recorded. Inundated almost 3000 km^2 of land	The 1947 floods resulted in policies aimed at improving the structured defence agricultural land.	Johnson et al. (2005) RMS (2007)
1998 April	Heavy rainfall on already saturated soil in early April 1998 caused extensive flooding across the English Midlands. Damage to towns, villages and agricultural lands was estimated to have caused £500million of damage, including five deaths.	The Easter 1998 floods were catalysts for policy change with regards to flood warning and public awareness raising	Horner and Walsh (2000) McEwen et al. (2002) Johnson et al. (2005)
2000 November	Widespread and prolonged flooding in the Winter of 2000 resulted in 10,000	The winter 2000 floods were catalysts for policy change with regards to spatial	Marsh and Dale (2002) Johnson et al. (2005)

	homes being flooded Damages estimated to be in excess of £1000 million	planning, resulting to the introduction of the PPG25 planning documents	
2007 June / July	Three storms in June and July of 2007 caused widespread flooding across most of the UK. More than 55000 homes and 6000 businesses were affected, resulting in insurance claims in excess of £3bn.	Following the 2007 summer flood events, a review commissioned by the UK government and carried-out by Pitt (2008) who drew-up a list of 15 urgent recommendation (out of 107 actions) for improving flood management in the UK.	Marsh and Hannaford (2007) Pitt (2008) Paranjothy et al. (2011)

FIGURE LABELS

Figure 1: Comparison of critical values of t5 obtained from Monte Carlo simulations (•) and the polynomial in Eq. (6)

Figure 2: Location of HiFlows-UK gauging stations with rating curves considered suitable for flood estimation by the gauging authorities.

Figure 3: Number of AMAX data available within each water-year.

Figure 4: Comparison of sample values of the index of surprise for k=2 (Eq. 2) and k=5 (Eq. 4) for 852 annual maximum series.

Figure 5: Comparison of surprising events identified for p = 15%, 10%, 5% and 1%.

Figure 6: Cluster of surprising events recorded at four or more sites when the largest events is located as the most recent event (contemporary assessment).

Figure 7: Number of gauging stations recording a record as a function of time. Summer events marked in red (broken lines) and winter events in blue (solid lines).

Figure 1:

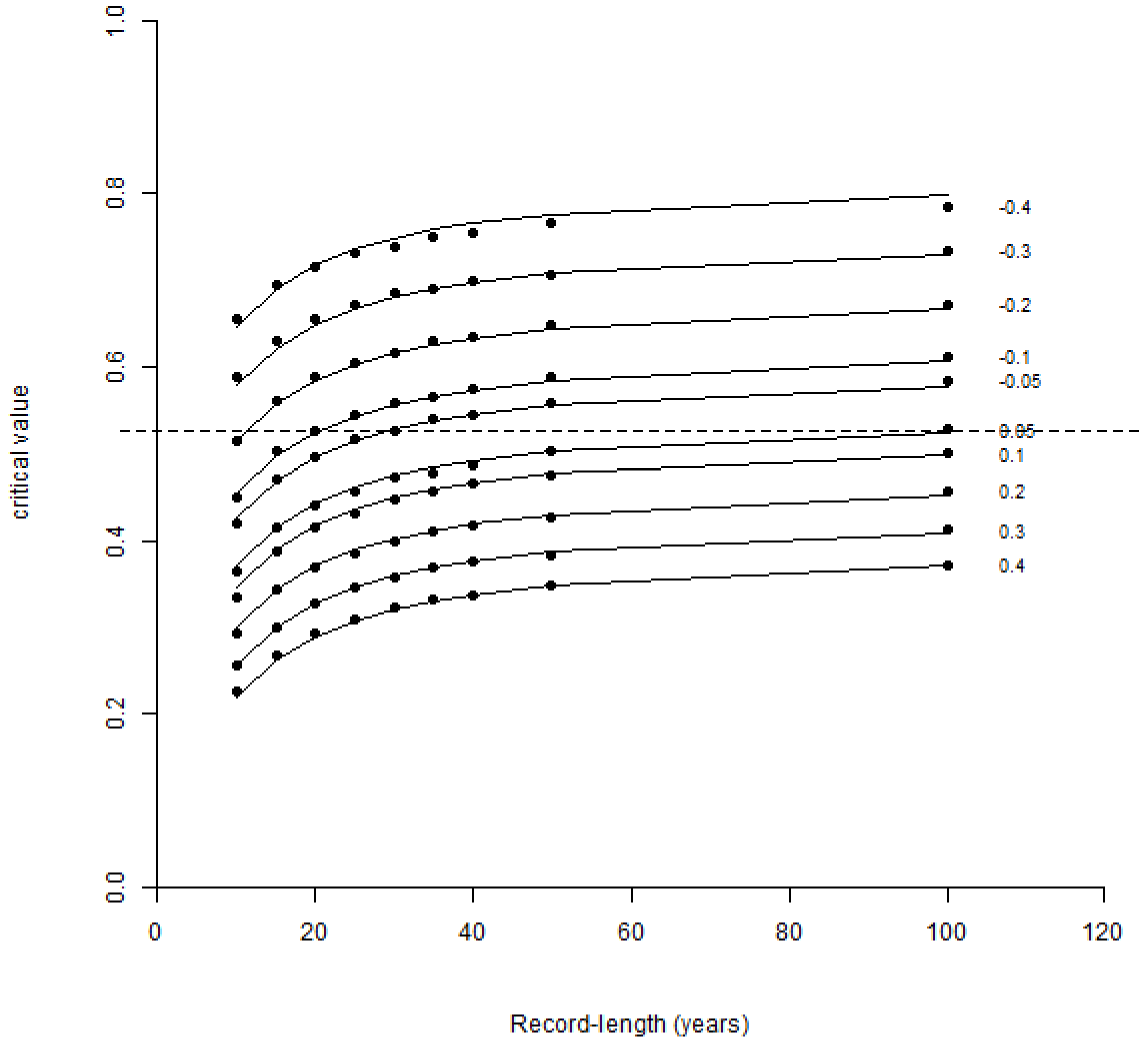

critical value
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Record-length (years)
-0.4
-0.3
-0.2
-0.1
-0.05
0.05
0.1
0.2
0.3
0.4

Figure 2

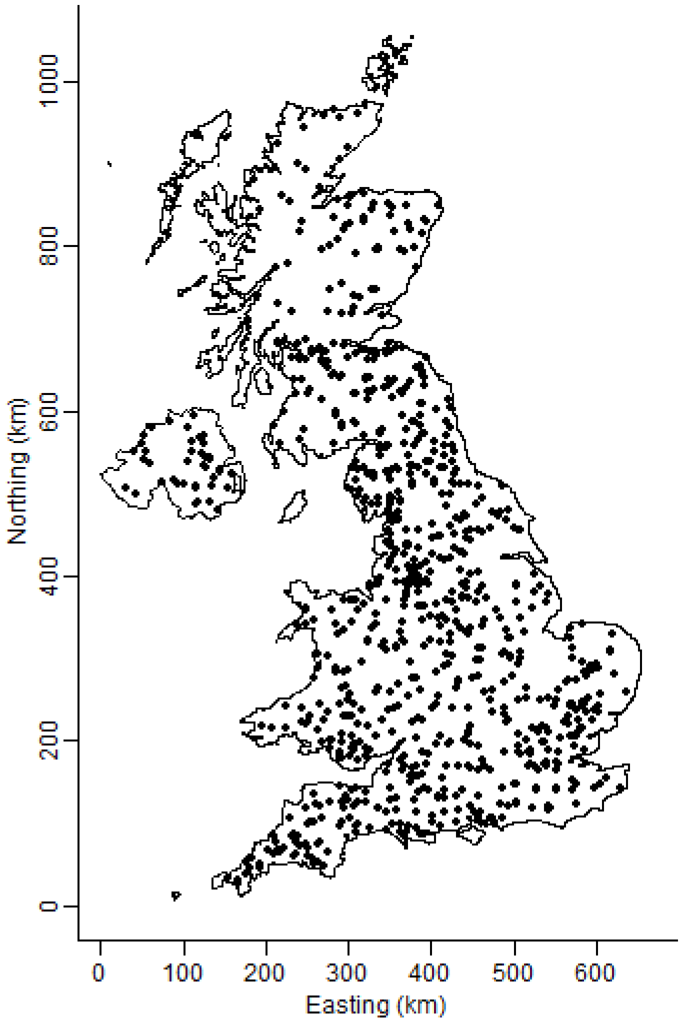

Northing (km)
0
200
400
600
800
1000
0
100
200
300
400
500
600
Easting (km)

Figure 3

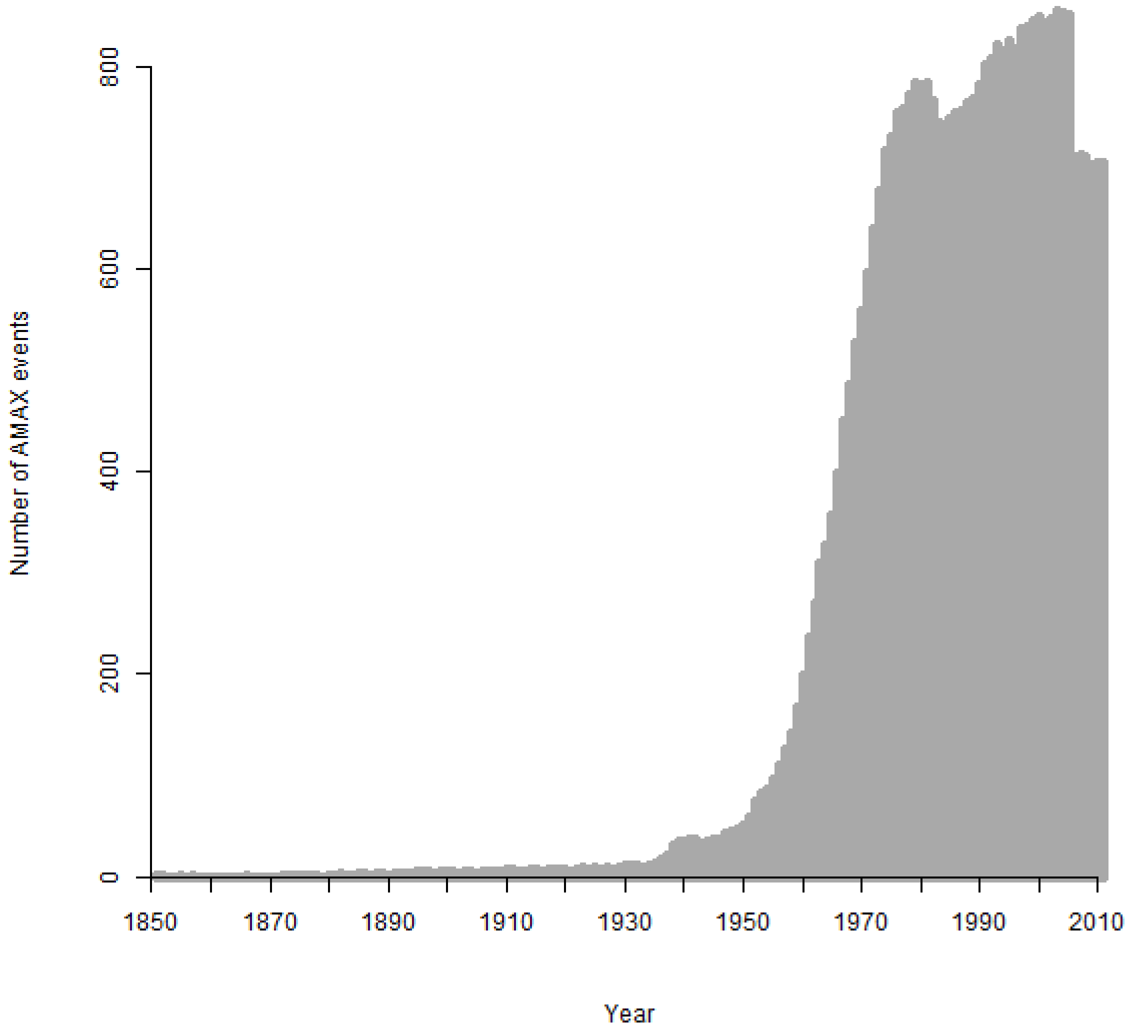

Number of AMAX events
0
200
400
600
800
1850
1870
1890
1910
1930
1950
1970
1990
2010
Year

Figure 4

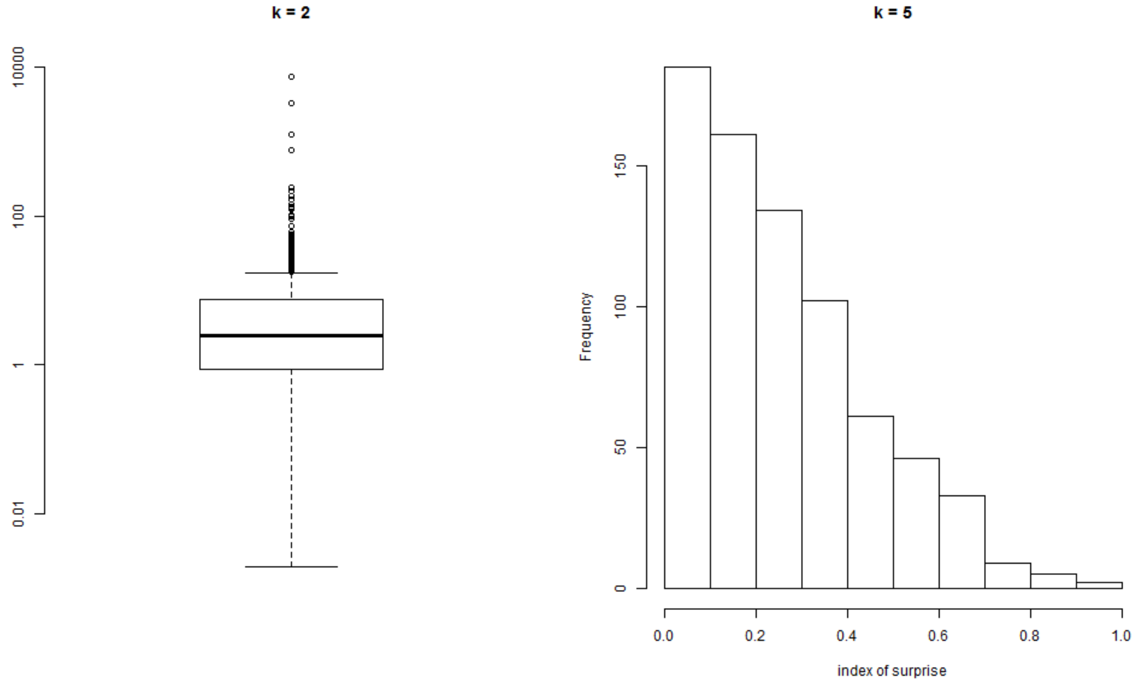

k = 2
10000
100
1
0.01
k = 5
Frequency
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
index of surprise

Figure 5

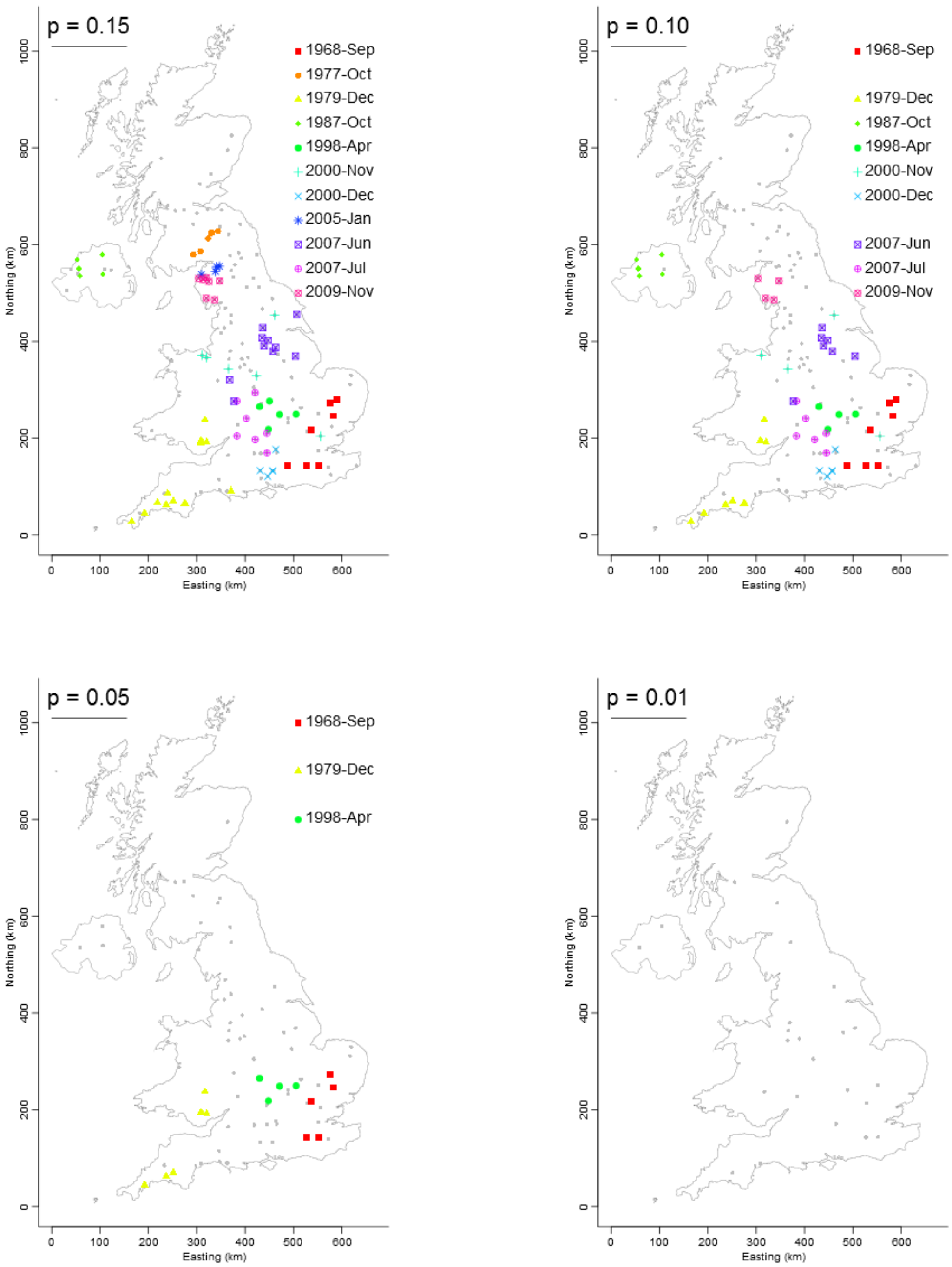

p = 0.15
1968-Sep
1977-Oct
1979-Dec
1987-Oct
1998-Apr
2000-Nov
2000-Dec
2005-Jan
2007-Jun
2007-Jul
2009-Nov
p = 0.10
1968-Sep
1979-Dec
1987-Oct
1998-Apr
2000-Nov
2000-Dec
2007-Jun
2007-Jul
2009-Nov
p = 0.05
1968-Sep
1979-Dec
1998-Apr
p = 0.01
Northing (km)
Easting (km)
0
100
200
300
400
500
600
800
1000

Figure 6

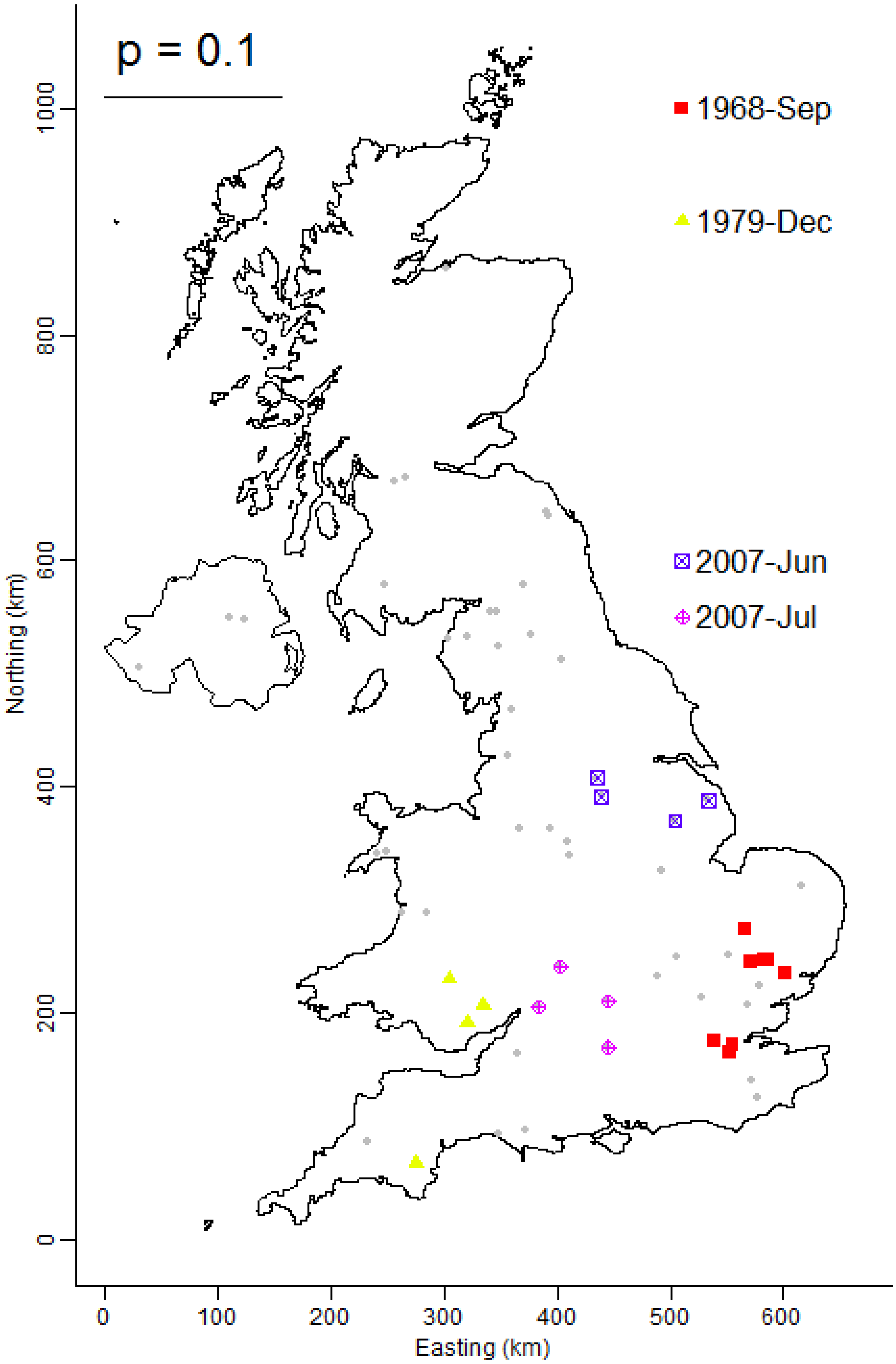
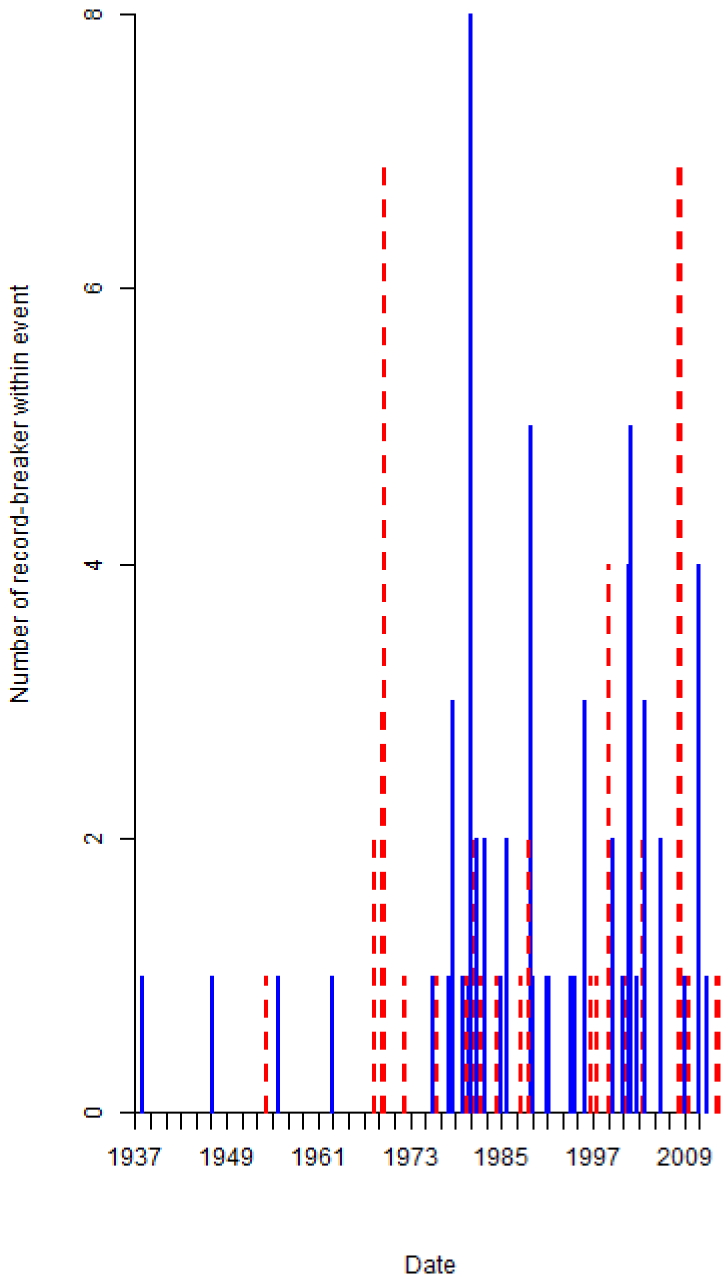

p = 0.1
1968-Sep
1979-Dec
2007-Jun
2007-Jul
Northing (km)
0
200
400
600
800
1000
Easting (km)
0
100
200
300
400
500
600

Figure 7


Number of record-breaker within event
0
2
4
6
8
1937
1949
1961
1973
1985
1997
2009
Date