Department for Environment Food & Rural Affairs

Llywodraeth Cymru Welsh Government

Cyfoeth Naturiol Cymru Natural Resources Wales

Environment Agency

delivering benefits through evidence

Development of interim national guidance on non-stationary fluvial flood frequency estimation – science report

FRS18087/IG/R1

Flood and Coastal Erosion Risk Management Research and Development Programme

We are the Environment Agency. We protect and improve the environment.

Acting to reduce the impacts of a changing climate on people and wildlife is at the heart of everything we do.

We reduce the risks to people, properties and businesses from flooding and coastal erosion.

We protect and improve the quality of water, making sure there is enough for people, businesses, agriculture and the environment. Our work helps to ensure people can enjoy the water environment through angling and navigation.

We look after land quality, promote sustainable land management and help protect and enhance wildlife habitats. And we work closely with businesses to help them comply with environmental regulations.

We can't do this alone. We work with government, local councils, businesses, civil society groups and communities to make our environment a better place for people and wildlife.

Published by:
Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH

http://www.gov.uk/government/organisations/environment-agency

ISBN: 978-1-84911-467-7

Email: fcerm.evidence@environment-agency.gov.uk

Further copies of this report are available from our publications catalogue:
http://www.gov.uk/government/publications

or our National Customer Contact Centre:
T: 03708 506506

Email: enquiries@environment-agency.gov.uk

Authors:
Duncan Faulkner; Dr Adam Griffin
Jamie Hannaford, Dr Paul Sharkey
Sarah Warren, Dr Kay Shelton,
Gianni Vesuviano, Nikos Mastrantonas,
Lisa Stewart

Dissemination status:
Publicly available

Keywords:
Stationarity, non-stationarity, flood frequency estimation, climate change, trend, clustering

Research contractor:
JBA Consulting
1 Broughton Park, Old Lane North, Broughton,
Skipton, North Yorkshire, BD23 3FD
www.jbaconsulting.com

Environment Agency's Project Manager:
Dr Sean Longfield

Theme Manager:
Dr Sue Manson, Incident Management and Modelling Theme Manager

Collaborator:
Centre for Ecology and Hydrology
Professor Jonathan Tawn
Dr Janet Heffernan

Evidence at the Environment Agency

Scientific research and analysis underpins everything the Environment Agency does. It helps us to understand and manage the environment effectively. Our own experts work with leading scientific organisations, universities and other parts of the Defra group to bring the best knowledge to bear on the environmental problems that we face now and in the future. Our scientific work is published as summaries and reports, freely available to all.

This report is the result of research commissioned and funded by the Joint Flood and Coastal Erosion Risk Management Research and Development Programme. The Joint Programme is jointly overseen by Defra, the Environment Agency, Natural Resources Wales and the Welsh Government on behalf of all risk management authorities in England and Wales:
http://evidence.environment-agency.gov.uk/FCERM/en/Default/FCRM.aspx

You can find out more about our current science programmes at:
https://www.gov.uk/government/organisations/environment-agency/about/research

If you have any comments or questions about this report or the Environment Agency's other scientific work, please contact research@environment-agency.gov.uk

Professor Doug Wilson
Director, Research, Analysis and Evaluation

Executive summary

Reports of ‘record-breaking’ or ‘unprecedented’ floods rarely seem out of the news in England and Wales. This has raised questions about whether the probability of floods has changed over the 40 to 60 years for which river flow data are available, and whether probability might change in the future.

Flood frequency analysis tells us what flood flows are expected to occur with a given probability. This is a fundamental part of cost-benefit analysis which is used to make decisions on investment in flood protection. Flood frequency analysis is also important for other areas of flood risk management such as mapping flood risk for planning, long-term investment planning, national flood risk assessment, setting insurance premiums, designing river structures and for reservoir safety.

In UK flood frequency analysis practitioners use the methods in the Flood Estimation Handbook. These techniques assume that in a data series each value, for example, each annual maximum flow or rainfall, is independent and has the same probability distribution as all the other values. If this probability distribution is not constant over time (**non-stationary**), the peak flows are not identically distributed and so this assumption cannot be made. This could have significant implications for capital investment decisions now and in the future if it means that flood risk has been over or underestimated.

The main aim of this project was to develop interim guidance for dealing with non-stationarity in annual maximum river flow series. The objectives were to:

- develop methods for identifying non-stationarity in annual maximum flow series
- develop a scientifically robust process for carrying out non-stationary flood frequency analysis
- outline how to take account of future climate change under non-stationary conditions
- provide guidance and tools for practitioners to carry out non-stationary flood frequency analysis
- make a high-level assessment of the impact of allowing for non-stationarity on flood frequency estimation across England and Wales

The focus was on fluvial flood frequency using methods that analyse annual maximum river flows. It is worth noting that flood studies on some rivers, all reservoirs and all surface water flooding investigations are based on rainfall-run-off models that use rainfall frequency statistics. The project has not investigated non-stationarity in rainfall, but similar methods could potentially be applied.

This report summarises all the methods applied and results obtained during the project. The project has also produced a separate Environment Agency technical guidance document for practitioners (), ‘Allowing for non-stationarity in flood frequency estimation’ (Ref: FRS18087/IG/R2) and a package of computer code written in the R language, nonstat (Ref: FRS18087/IG/R3).

Method development

Methods of non-stationary flood frequency analysis are widely applied in a research context, but there are few examples of flood management authorities using them in decision-making. The project has included an extensive literature review, and these findings have been used to develop non-stationary methods that practitioners can readily use. The method development is a follow-up to an earlier investigation of trends and non-stationarity in north-west England commissioned by the Environment Agency following the floods of December 2015.

The non-stationary methods applied in this project introduce covariates to help explain trends in flood flow. These are other variables that are expected to be related to the variable of interest. The simplest statistical models use time as a covariate, in other words, the probability distribution of flood flows is modelled as changing over time. Other models incorporate physical covariates such as measures of rainfall or climatic indices. This project has made some breakthroughs in applying non-stationary methods, including innovative techniques for extracting design flood estimates from statistical models that include physical covariates.

Some of the method development tasks were more exploratory in nature, and aspects such as pooled non-stationary analysis will require further development before being suitable for routine application. Non-stationary methods can currently only be applied at sites where peak flow data is available, over a suitably long record length.

Findings from national-scale analysis

River flow records show general but not universal evidence of an increase in flood peaks. Two thirds of gauging stations in England and Wales show upward trends in peak flows when tested using the non-parametric Mann-Kendall test. 13% of stations show an upward trend that is significant at the 5% level, and this increases to 21% if the significance threshold is relaxed to 10%. Positive trends are seen across much of England and Wales, with some of the strongest and most statistically significant trends in the north and west. Some areas of central and eastern England also display negative trends. The analysis included data up to September 2017. The degree of upward trend would be expected to increase if the tests were repeated using data that included the extensive and severe floods of winter 2019 to 2020.

When non-stationarity is modelled only in relation to changes over time, around 22% of stations preferred a non-stationary over a stationary model. This is similar to the findings from the non-parametric trend testing, although there is not complete overlap between the stations identified in the 2 different analyses.

The proportion of stations best fitted by a non-stationary model increases to 36% when physical covariates are added along with water year. Physical covariates are almost always beneficial to the fit of non-stationary flood frequency models, with annual or seasonal rainfall totals proving more beneficial than the indices of atmospheric circulation that were tested. This finding indicates that the increase in model complexity is nearly always outweighed by the increase in goodness of fit provided by the physical covariates.

On average, across England and Wales, including non-stationarity makes little difference to estimating design flows. In individual cases, it can make a large difference, leading to an increase in present-day estimates. There are large local variations in the comparison between stationary and non-stationary estimates, making it difficult to generalise the results across regions.

The evidence of general increases in flood magnitude is consistent with projections of the impacts of climate change. This project has not attempted to attribute trends, which may occur due to changes in catchments (such as urbanisation), changes in river channels or climate changes. It is quite possible that there is a cyclical element to recent trends. However, it would seem unwise, in the face of a warming climate, to expect trends to reverse in the near future.

Non-stationary methods can potentially provide more credible answers that can be more easily justified to interested groups. On the other hand, a shift to non-stationary techniques can lead to an increase in uncertainty. The methods developed in this project all include quantification of uncertainty, and this is output from the *nonstat* package in the form of confidence limits.

Acknowledgements

Particular thanks are extended to the Environment Agency's project team: Richard Macilwaine, Peter Spencer, Sean Longfield and Parveen Mann.

The project team is grateful to Professor Emeritus Donald Burn of the University of Waterloo, Canada for his inputs to the work on pooled non-stationary flood frequency analysis and to Professor Bruno Merz of the Helmholtz Centre Potsdam, Germany for advice on cluster analysis.

Contents

List of figures

List of tables

1 Introduction

1.1 Background

Reports of ‘record-breaking’ or ‘unprecedented’ floods rarely seem out of the news in England and Wales. 2019 to 2020 saw extreme flooding in many areas, including Lincolnshire, Yorkshire, Lancashire, South Wales and the English Midlands. North-west England was badly hit with severe floods in 2005 (Carlisle), 2009 (much of Cumbria) and 2015 (much of Cumbria and Lancashire). This severe flooding has raised questions about whether the probability of these events is now higher than it was and how probability might change in the future.

Answers to these questions are needed to help plan investment in flood alleviation. Decisions on investment are made based on cost-benefit analysis, which needs information on the likelihood of flood damage occurring. This comes, in part, from flood frequency analysis, which tells us what flood flows are expected to occur with a given probability. This analysis is also important for other areas of flood risk management such as mapping flood risk for planning, long-term investment planning, national flood risk assessment, setting insurance premiums, designing river structures and for reservoir safety.

UK practice in flood frequency analysis is to use methods in the Flood Estimation Handbook (FEH) (Institute of Hydrology 1999) and its updates (Kjeldsen and others, 2005; 2008). The techniques in these publications assume that in a data series each value, for example, each annual maximum flow or rainfall, is independent and has the same probability distribution as all the other values. If the flood frequency behaviour of a catchment is not constant over time (non-stationary), peak flows are not identically distributed and so this assumption cannot be made.

Non-stationarity in flood time series may occur due to changes in catchments (such as urbanisation), changes in river channels (such as dredging) or climate changes. If non-stationarity is present, then the flood frequency estimates from commonly applied methods are called into question. This could have significant implications for capital investment decisions now and in the future.

1.2 Project scope

1.2.1 Scope

This project builds on work carried out by JBA Consulting in 2017, which investigated trends and non-stationarity in north-west England following the floods of December 2015. A peer review of that work by Dr Ilaria Prosdocimi (then at the University of Bath) recommended further areas of investigation, which helped define the scope of the present project.

The main aim of this project was to develop interim national guidance for dealing with non-stationarity in annual maximum river flow series. The objectives were to:

- develop methods for identifying non-stationarity in annual maximum flow series
- develop a scientifically robust process for carrying out non-stationary flood frequency analysis
- outline how to take account of future climate change under non-stationary conditions

- provide guidance and tools for practitioners to carry out non-stationary flood frequency analysis
- carry out a high-level assessment of the impact of allowing for non-stationarity on flood frequency estimation at a national scale

It was envisaged that the interim guidance would remain in place for at least 2 years, until superseded by future work, and would initially be introduced for the planning and appraisal of flood risk management schemes.

The report did not investigate the causes of non-stationarity, but just considered fluvial flood frequency using methods that analyse annual maximum flow data. It is worth noting that flood studies on some rivers, all reservoirs and all surface water flooding investigations are based on rainfall-run-off models that use rainfall frequency statistics. This project has not investigated non-stationarity in rainfall.

1.2.2 Related project: Rapid evidence assessment

In parallel with this project, the Environment Agency commissioned a rapid evidence assessment (REA) on non-stationarity in sources of UK flooding, including surface water and tidal flooding as well as rivers. The REA addressed one primary question:

- What is the evidence for stationarity or non-stationarity in sources of UK flooding?

Three secondary questions were addressed in less detail:

- What can cause non-stationarity in the sources of UK flooding?
- What techniques are used to detect and account for non-stationarity in the sources of UK flooding?
- To what extent does assuming stationarity or non-stationarity alter the outcome of flood risk analysis?

These questions were answered from a comprehensive review of published articles. The evidence showed a general, but not universal, consensus that both precipitation and flood flows on rivers are increasing. Most of these studies analysed series of measured data, but about a third included an investigation of future changes, generally using modelling techniques.

The report is called, Rapid Evidence Assessment of Non-Stationarity in Sources of UK Flooding (FRS18087/REA/R1).

1.3 Project outputs

This report summarises all the methods applied and results obtained during the project. As well as this main report, the project has also produced guidance for practitioners (FRS18087/IG/R2, ‘Development of interim national guidance on non-stationary fluvial flood frequency estimation – practitioner guidance’) and a package of computer code written in the R language, nonstat. The nonstat package comes with its own user guide. The package provides functions for trend testing and fitting of non-stationary flood frequency distributions, with user-friendly input formats and outputs.

Detailed results of the national-scale tests for trend and non-stationarity are provided digitally ‘FRS18087-IG-D2-digital_outputs.zip’.

1.4 Report structure

The main part of this report summarises the work carried out and the results obtained. The appendices contain more in-depth reporting.

Chapter 2 describes how the project data set was developed, and the various screening steps carried out, including the application of change point tests.

Chapter 3 outlines methods used to detect and account for non-stationarity, including trend tests. The project included extensive investigation of non-stationary methods of flood frequency estimation, including development of new techniques and consideration of how to account for climate change. Chapter 3 also gives an overview of clustering of floods. Appendices A to F support chapter 3.

Chapter 4 presents results at a national scale, including a summary of the different types of analysis applied.

Finally, chapter 5 draws the strands of investigation together and makes some suggestions for further research, including potentially replacing the interim guidance in due course.

1.5 Terminology

Some of the main technical terms used in this report are:

- **Model** – in this report model refers to a statistical description of the flood frequency relationship, unless otherwise stated.
- **Covariate** – another variable that is included because it is related to the variable of interest. Here, covariates help to explain trends in flood flow. The most straightforward non-stationary models have a single covariate, which is time, meaning that the probability distribution changes from year to year. Other covariates might represent physical quantities such as measures of the climate or the catchment.
- **Return period** – for a non-stationary model, the return period T is the reciprocal of the annual exceedance probability (AEP) at a particular point in time, t:

$$T_t = \frac{1}{AEP_t}$$

- **Return level** – the value (in this analysis, the river flow) associated with a particular return period or probability. Hydrologists often refer to this as the design flood, or as a point on the flood frequency curve. The practitioner guidance avoids using the terminology of return level because of its potential for confusion with water levels. It is used in this project report for consistency with the scientific literature.
- **Quantile** – a point on a probability distribution; essentially the same as a return level when considering annual maximum data.
- **Conditional return level** – the return level that would be expected if the covariates had a particular combination of values (this means it is conditional on those values). For instance, if the covariate was the water year, there would be a return level conditional on the water year being 2019 to 2020.
- **Marginal return level** – Expected value of a variable (such as flow) for a particular probability, without any conditionality on covariate values (contrast with conditional return level, above). The marginal return level is calculated by averaging the probabilities corresponding to the conditional return levels over a sample or a statistical distribution of covariate values. In the practitioner guidance, the marginal return level is referred to as an 'integrated flow estimate'. This concept is explained further in this report.
- **Design life level** – similar to marginal return level, it is the return level corresponding to a particular exceedance probability (the **encounter probability**) over a design period. The concept is presented in Yan and others (2017). Rather than being a return level for a particular year (past, present or future), it is associated with a period of time, such as the design life of a flood alleviation scheme.

Other terminology is defined as it is introduced throughout the report.

2 Data

2.1 Requirements for data

To analyse non-stationarity flood peak data sets should be as long as possible. This is because any underlying trends tend to be masked by variability over short to medium timescales. A minimum record length of 30 years was imposed, while recognising that at least 40 years would be desirable. The multi-temporal trend tests explicitly considered the effect of record length, applying the tests to all combinations of start and end year within the period of record.

The vast majority of the analysis was based on annual maximum flow (AMAX) data. Although peaks over a threshold (POT) data might be expected to provide a more complete picture of flood characteristics, there are several reasons why POT data were not thought appropriate to use in this project. These are listed below:

1) In the UK, AMAX data have been used and reviewed a great deal more than POT data and so form a more reliable flood peak data set.
2) The reliability of POT data in the pre-digital period is variable. There are several sources of data and they are not always consistent with each other or with the AMAX series.
3) About 13% of gauging stations in England do not have POT data, generally those on groundwater-dominated rivers.
4) POT records are shorter than AMAX records at some stations.
5) There are gaps in POT records but they can be difficult to detect.
6) The average number of events/year varies between stations. POT extraction rules may have been different at different times.
7) POT time series are more likely than AMAX series to show serial correlation, which is undesirable for trend analyses.

The exception was for a limited investigation of clustering, for which it was possible to review the quality of a small number of POT records.

Some of the analytical methods applied in this project only looked at whether successive AMAX flows were increasing or decreasing, without accounting for the magnitude of the changes. The Mann-Kendall trend test is an example of this. Other methods accounted for the magnitude of change, such as the fitting of non-stationary flood frequency distributions. The latter methods demanding higher quality data, and so were applied on a higher quality subset of the data.

The project included gauges throughout England and Wales.

2.2 Data screening

The National River Flow Archive (NRFA) Peak Flow Dataset version 7, containing data up to the end of September 2017, was used as the basis for this project. The current 753 stations in England and Wales in this data set were collated, together with another 5 stations that were expected to be added to the NRFA Peak Flow Dataset in future, making 758 stations in total.

As a starting point, the data quality for the 2 types of analyses were considered to be roughly equivalent to the 'indicative suitability' classes in the NRFA data set. Stations that were 'OK for QMED' were initially listed for the trend analyses, and those 'OK for pooling' were listed for the flood frequency analyses.

When testing for non-stationarity it is important that the tests reflect genuine changes (or lack of changes) in flow. It is common for measurement structures or rating equations to change over the period of record, and it is important that these changes do not introduce false changes in the recorded flows.

For this reason, the gauging stations included in the project were screened carefully. The steps involved in the screening were as follows:

1) The Environment Agency and Natural Resources Wales initially assessed the list of gauging stations to identify where the length, quality and consistency of the data was good enough to use in the statistical analyses. Gauging authority staff with local knowledge of the stations carried out this assessment, following consistent guidance. The individual station assessments were reviewed by one experienced flood hydrologist to provide national consistency.

2) JBA Consulting and the National River Flow Archive (NRFA) carried out a second stage of screening, considering whether any gauges should be added to or removed from the initial list. This assessment included applying statistical tests for change points and using knowledge gained from detailed hydrological projects.

3) The measuring authorities commented on the revised list.

4) The Environment Agency and contractors worked together to produce a near-finalised list.

5) The Environment Agency subsequently made a small number of changes to this list throughout the project. The trend testing but not the flood frequency analysis was revised to account for these alterations.

A number of authors (for example, Wilby and others, 2017) note the importance of sound data and describe some of the possible data biases and errors. Key ways in which data may not be consistent are discussed below.

a) Different ratings, applied to different periods in the record, may vary in their applicability to high flows. For example, some ratings may be based on gaugings at high flows, others may be based on gaugings at low flows and just extended to higher flows. This commonly occurs where stations have changed (for example, location, from open-channel to a structure, change in structure type, or changes to a structure) or where an open-channel station has a moving bed so that there are frequent changes in ratings.

b) Drowning of weirs may be handled differently over different parts of the record. For example, an early record may not have drowning adjustments, some parts of the record may use a pressure head, and the later period may be adjusted using a downstream level recorder. Drowning may be consistent, may vary with

seasonal changes in downstream channel roughness, or may depend on flows at a downstream river confluence.

c) Some stilling-wells and inlet pipes are prone to siltation and blockage. At some stations, the under-recording may be slight, but at others this can mean that large events are effectively missed. Historically, attempts have been made to reduce this by pumping out, but the frequency and effectiveness of this is rarely known. In more recent times, in-river pressure transducers are increasingly used to provide a check and back-up to the primary in-well recorder.

d) Pre-digital data: The measuring authorities' digital archives have powerful tools to review data. Plots enable stations to be compared and gaps in the data to be identified; the listed validity of ratings identifies the rating source of a flow value. In comparison, the quality assurance of pre-digital data is harder - gaps in the data may not be apparent, copies of charts showing hydrographs may not be readily accessible, and the ratings used may not be clear. There may be several sources of pre-digital data, with different flows.

Stations and periods of record were not excluded due to changes in the catchment, because the project aimed to identify non-stationarity whatever its cause. An exception to this was when a reservoir with a significant effect on the flood regime was constructed during the period of record. In this case, only the longer of the pre- and post-reservoir periods was retained. This screening process did not, in general, account for the construction of smaller flood storage schemes.

Stations with gaps in their records were not excluded from the project data set. However, some of the statistical tests excluded stations with long gaps, as discussed in the relevant sections of this report. Appendix H shows the reasons why individual gauges were excluded.

2.3 Change point tests

2.3.1 Purpose

Change point tests were applied to help screen the data sets, to identify gauges where apparent trends in peak flows may be false, for example due to changes in rating equations or alterations to a gauging structure or river channel that have not been accounted for in the rating. It is expected that, in some cases, change points may represent genuine sudden changes in peak flows, for example due to rapid urbanisation, clear felling of forestry or a shift between a flood-rich and flood-poor period, or vice versa.

2.3.2 Pettitt's test

Pettitt's test is designed to detect a sudden change in the mean of a time series. It outputs the time of the shift as well as the significance level. It is a non-parametric test, which means it makes no assumption about the distribution followed by the data.

The null hypothesis H0 is that there is no difference between the means of the earlier and later portions of each annual maximum (AMAX) flow series. The tests output a p-value, or probability, and if this is less than a chosen significance level then H0 is rejected. The conventional approach is then to (provisionally) accept a single alternative hypothesis H1, in other words, that there is a sudden change.

2.3.3 PELT test

One limitation of Pettitt's test is that it can only detect a single change point. It has been criticised for its tendency to classify gradual trends as sudden changes (Rougé and others, 2013).

An alternative to Pettitt's test is the PELT (Pruned Exact Linear Time) test. Like with most change point algorithms, PELT (Killick and others, 2012) tries to find the optimal segmentation in a time series. Optimality is usually determined by minimising a cost function. Where a parametric model, for example, a normal distribution, is assumed, the cost function is typically related to the likelihood function, which expresses the probability that the observed data arise from a particular assumed distribution. The change point algorithm finds the segmentation, with varying parameters across segments, that minimises the cost function. The likelihood is usually combined with a penalty term to prevent overfitting to the data; it stops too many unrealistic change points being identified.

PELT is based on the idea of optimal partitioning by recursive minimisation of segment-wise cost functions in a time series. This is an exact algorithm in that it always finds the optimal solutions, but typically these approaches come at a high computational cost. PELT requires a pruning step that discards candidate change points that will not lead to optimal segmentations, which leads to a computational complexity that scales linearly with the number of data points. As a result, PELT can give optimal segmentations in a reasonable length of time.

PELT is implemented through the 'changepoint' R package (Killick and Eckley, 2014) and allows for detecting one or more changes in the mean, the variance or both. The approach requires an assumption that the data follow a known distributional form. Since the flood time series are not generated from any known distribution, a log transformation is applied to transform the data to approximate normality (Box and Cox, 1964). In effect, it was assumed that the underlying AMAX flows follow a log-normal

distribution, as assumed in previous studies (for example, Prosdocimi and others, 2013). It would be desirable in future to modify the PELT test so it can assume specific extreme value distributions such as GEV or GLO.

A minimum segment length is required, which prevents additional overfitting and false positive changes at short time scales. Ten years was chosen to be a suitable length for local stationarity pre- and post-change.

2.3.4 Data sets for change point tests

Both the change point tests were applied twice: first to all the stations initially proposed to be included in the project, to help screen stations, and then again to a near-final project data set.

All AMAX flows marked as rejected were excluded from the analysis.

2.3.5 Results

Table 2.1 summarises the results of the change point tests applied to a near-final version of the project data set, which comprises 471 stations, plus 6 files containing alternative versions of the AMAX data set at 5 stations. A further 4 stations were excluded from the analysis because their record length was under 30 years.

Table 2.1 Summary of change point test results

	% of stations with positive change (at 5% significance level)	**% of stations with negative change (at 5% significance level)**
Pettitt test: Change in mean	10%	1%
PELT test: Change in mean	4%	1%
PELT test: Change in standard deviation	3%	2%

In summary, the vast majority of stations in the final project data set showed no significant step changes. The PELT test is more stringent than the Pettitt test and so detected a smaller number of stations with significant changes. It did not find any stations with more than one change point.

Figure **2.1** shows the timing of the change points. From the PELT test, many of the change points occur between 1995 and 2003. Similarly, from Pettitt's test, the highest density of change points is between 1996 and 1998. This period is sometimes referred to as the time when much of the UK transitioned from a flood-poor to a flood-rich period, with the widespread events of Easter 1998 and autumn 2000 being the most obvious signs of that transition.

Figure **2.1** also shows a tendency for longer records to show earlier change points. The opposite would be impossible, because nearly all shorter records cover recent decades and so could not contain earlier change points. The longest record that contains a

change point is the 131-year data set on the Ouse at Skelton (York), for which an upward shift in flood magnitude occurred in 1943. There have been several changes in rating at this gauge, to account for alterations in the channel and flood defences. It is difficult therefore to be confident that the change in 1943 represents a genuine step up in flood magnitude. Similar comments may apply for some of the other change points that have been detected. However, in each case, a decision has been made to retain the gauge in the project, balancing the findings from the change point tests with the knowledge of staff from gauging authorities and the NRFA.

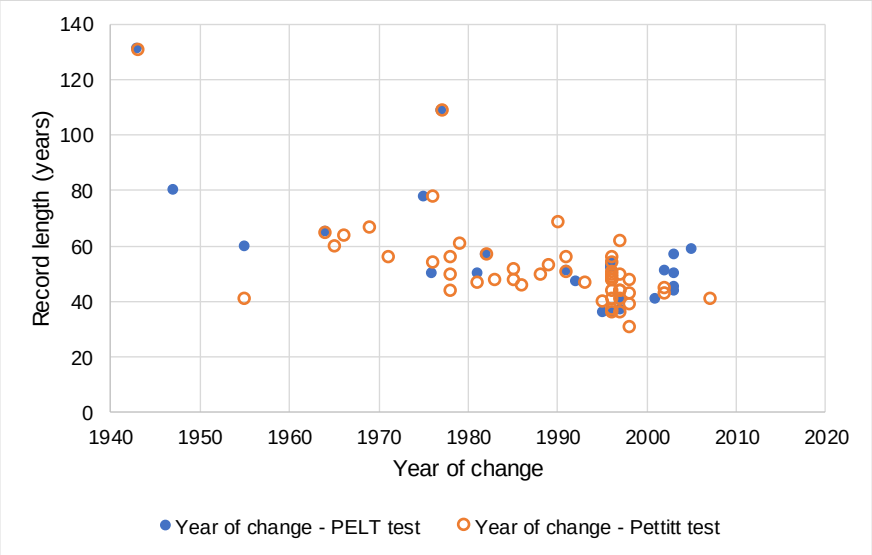


Figure 2.1 Years corresponding to change points

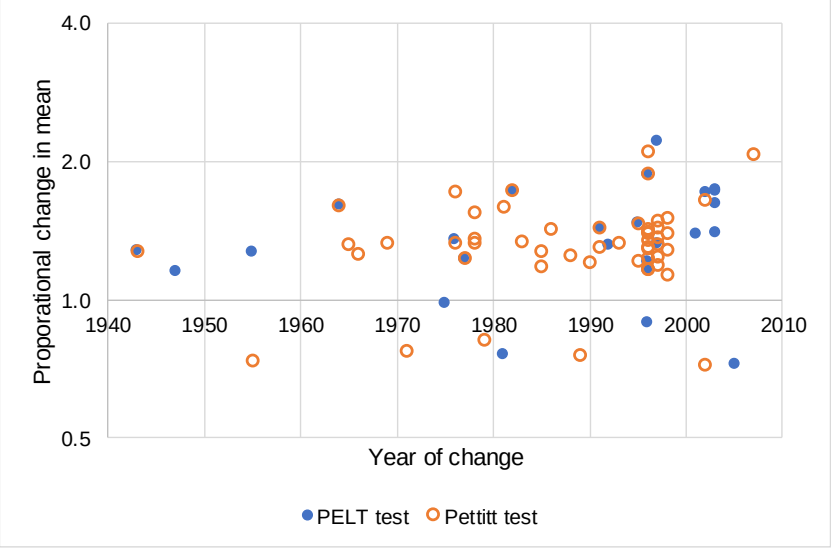


Figure 2.2 Magnitude of changes in the mean annual maximum flow

Figure **2.2** indicates that the vast majority of changes in the mean annual maximum flow are positive. The typical magnitude of change is an increase in the mean by a factor of 1.2 to 1.6 (this is, an increase of 20 to 60%). The largest changes are seen at a small number of stations in 1996, 1997 or 2007. Where there are reductions in the mean, they are generally -20 to -30% and do not appear to be clustered in time.

2.3.6 Example: the most extreme change

The most extreme step change in both mean and standard deviation occurs at station 27055, the Rye at Broadway Foot. This catchment drains the North York Moors. Both the Pettitt and PELT tests detected the largest change in mean AMAX flow at this station; the change point dates differing by a year (1996 for Pettitt and 1997 for PELT).

Figure **2.3** shows the flood peak series and the change point. The mean AMAX flow more than doubles at the change point. The large increase in the mean is mainly caused by the outstanding flood of June 2005. However, the floods of March 1999 and November 2000 were also exceptional compared with the earlier record, and, in fact, all top 10 floods in the record occur after the change point.

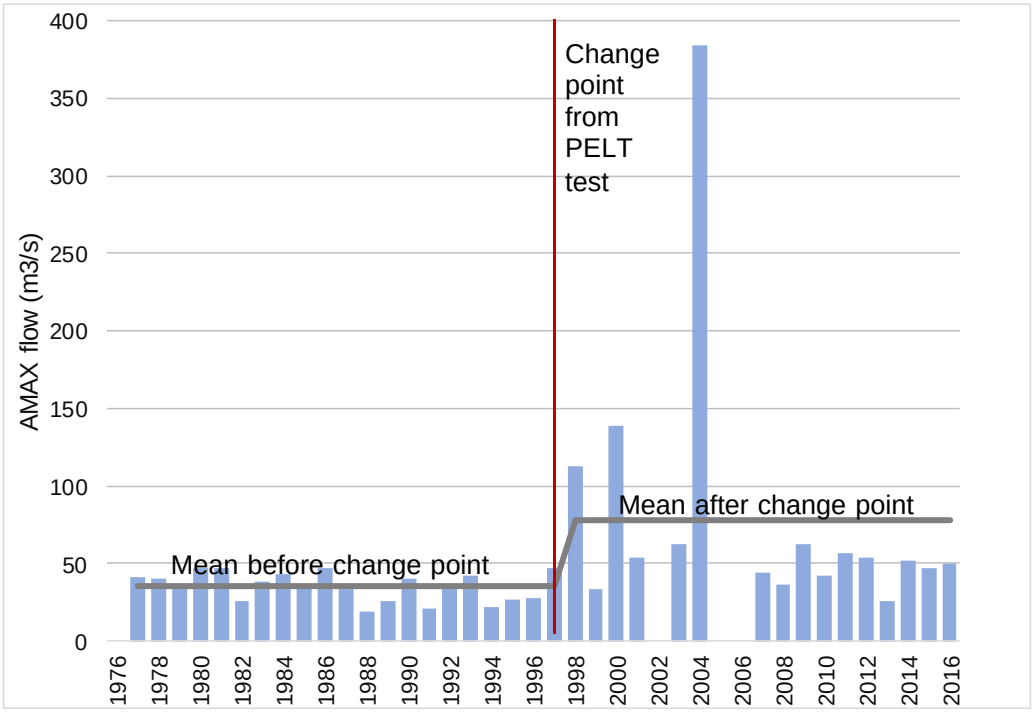


Figure 2.3 Change point for the Rye at Broadway Foot

The standard deviation increases by a factor of 8 at the change point as a result of the exceptional floods mentioned above.

The PELT and Pettitt tests do not detect changes in the median. Flood Estimation Handbook (FEH) methods use the median rather than the mean because the former is more robust in the presence of outliers. At Broadway Foot, the median annual maximum flow (QMED) is 46% higher after the change point than before it. This is a more modest increase than in the mean, but still substantial.

Local Environment Agency staff have suggested that the change seems likely to be due to climatic reasons, exacerbated by the Corallian limestone geology of the catchment.

2.4 Final project data set

A total of 375 gauges across England and Wales were selected as suitable for the flood frequency analysis. A further 100 were used just for trend testing. The gauges are listed in Appendix H.

The AMAX series from the NRFA Peak Flow Dataset (version 7) was selected at most stations. Replacement series were produced for 80 stations. Of these, compared to the NRFA Peak Flow Dataset, 31 had different ratings and different periods of record, 13 had new or amended data, 3 were combined with earlier stations, and at 2 stations post-reservoir data were used because the data length was longer than the pre-reservoir data selected in the NRFA.

88 stations, which are classed as 'Not OK for pooling', were included within the flood frequency analyses. This was considered acceptable because the aim of the project was to study the effect of any changes over time. Therefore, a lower standard of accuracy of high flows than normally applied in flood frequency estimation by the Flood Estimation Handbook (FEH) statistical method was acceptable, as long as the methods of deriving the flows were consistent over the period applied.

The median length of record of the selected stations was 49 years, with the longest record (Thames at Kingston) being 134 years, and 95% of the stations having 37 years or more of record.

The Environment Agency made some amendments to the data set during the project. The trend tests and split sample tests were rerun using this revised data set. The non-stationary frequency analysis was not repeated. This was not thought necessary as the frequency analysis was never intended to be applied to an identical data set given the less stringent data quality requirements imposed for gauges.

3 Methods

3.1 Trend tests

The methodology for trend testing was based on the National River Flow Archive (NRFA) trend testing toolkit as described by Harrigan and others (2018). There is a detailed account of the tests and their results in Appendix A. The appendix also includes information on how the tests handle gaps in the data series.

To test for trends, the non-parametric Mann-Kendall (MK) test was applied. This is a very widely used method for monotonic trend testing which has been applied extensively in hydrological change applications in the UK and elsewhere. It is a non-parametric test, in that it makes no assumption about the statistical distribution of the data. The test is not dependent on the magnitude of the data, but is based on the proportion of increases and decreases between pairs of values. A consequence of this is that it tests the statistical significance of the trend but does not directly measure the strength of the trend.

The test produces a score, known as MKZ. Positive values of MKZ indicate increasing trends, while negative ones refer to decreasing trends. MKZ scores are standardised, in order to compare the different periods of interest and stations. For identifying whether the results are statistically significant at a 5% significance level, a two-tailed MK test was chosen, meaning that if the absolute value of MKZ exceeds 1.96, the null hypothesis H0 of no-trend is rejected. To test for significance at a 10% level, the critical value of MKZ is 1.645.

If H0 is rejected, the conventional approach is then to (provisionally) accept a single alternative hypothesis H1, in other words, that a statistically significant trend exists. This is not always the correct conclusion to draw: the discrepancy of the observations from H0 may actually be due to factors not included in the formulation of H0 and different from H1 (Serinaldi and others, 2018). One particular factor could be dependence between the observations, as the MK test assumes that each data value is independent of the others. This assumption may not hold due to the occurrence of flood-rich and flood-poor periods, or on some groundwater-dominated catchments where high baseflow persists for more than one year. To satisfy this assumption, the AMAX time series were first analysed for significant lag-1 serial correlation using the autocorrelation function. For instances with significant lag-1 serial correlation, block bootstrapping was applied for the significance testing.

The magnitude of trends was calculated using the Theil-Sen approach (TSA). This helps focus attention on the direction and strength of changes and not entirely on statistical significance relative to arbitrary p-value thresholds (Nicholls 2001).

The Theil-Sen (sometimes referred to as Kendall-Theil) robust line is widely used for quantifying trend magnitude, and is similar to the gradient of a least-squares linear regression line, but is preferred due to being less sensitive to the presence of outliers (for example, Stahl and others, 2012).

For a data set $(t_i, Q_i : i = 1, \dots N)$ with all different values of Qi, the Theil-Sen estimator of the slope of $Q = (Q_1, \dots, Q_N)$ is given by:

$$TSA = median\left\{\left(\frac{Q_j - Q_i}{t_j - t_i}\right) : i \neq j = 1, \dots, N\right\} \quad (1)$$

TSA is the median of all pairwise slopes between all points with different times.

To make a relative comparison between sites, the trend magnitude $\mathrm{TSA_{rel}}$ (%) for each time series was expressed as a percentage of the long-term mean annual maximum

flow μ over the period of record of n years where β is the TSA slope, given by Stahl and others (2012) as:

$$\mathrm{TSA_{rel}} = \left(\frac{\beta \times n}{\mu}\right) \times 100 \qquad (2)$$

Hannaford and Buys (2012) found this approach preferable compared to expressing trend magnitude as a simple percentage change over the full record, which can yield larger changes in the presence of abnormally large start or end values.

Given the confounding effect of hydrological variability over decades, trend tests do not necessarily provide evidence of long-term variation. The analysis has been set in context by applying in a multi-temporal framework, whereby trends are analysed for all possible combinations of start and end points rather than just for the entire period of record at each gauge. The results are visualised using heatmaps.

The results are summarised in section 4.1.

3.2 Split sample tests

3.2.1 Purpose and significance tests

Split sample tests were carried out as requested by the Environment Agency. The flood peak series at each gauging station was split into 2 portions, before and after a fixed change point. Statistics of the earlier and later portions of the record were calculated, and tested for significant differences.

The Mann-Whitney U test was used to test for significance of changes between the distributions of AMAX flows in the earlier and later periods. The test determines whether 2 independent samples were selected from populations having the same distribution. The null hypothesis is that the distributions of the 2 populations are identical.

The Brown-Forsythe test (Brown and Forsythe, 1974) was used to test for significance of changes between the variances of AMAX flows in the earlier and later periods. The null hypothesis is that the samples of AMAX flows are drawn from populations with equal variance.

The results are given in section 0.

3.2.2 Determining the split point

The Pettitt and PELT change point tests both determine the point in time (if any) at which a change occurs in a data set. For the split sample tests, the timing of the split point was predetermined. It was calculated as the typical midpoint of the flood peak series, that is, the point in time that lies midway between the median start year (1967) and median end year (2016). This gave a midpoint of approximately 1991. The annual maximum (AMAX) flows up to the water year 1990 were taken as the earlier portion of the record, and those from 1991 onwards were taken as the later portion.

Gauges were included in the tests if they had at least 15 AMAX flows before and after the split point. 400 AMAX series met this criterion.

All years of record were included in the tests; this means that for long record stations the earlier period may be much longer than the later period. All AMAX flows marked as rejected were excluded from the analysis.

3.3 Non-stationary flood frequency analysis

3.3.1 Scope of method development tasks

The scope of the project was to investigate and develop methods of non-stationary flood frequency analysis, following the earlier trial of non-stationary methods in north-west England. To do this, the project team would:

- carry out a literature review
- investigate methods to identify best model fit and test the realism of the flood frequency results
- investigate suitable distributions for non-stationary flood frequency analysis
- investigate a non-stationary model form with correlation between the location and scale parameters
- investigate the application of covariates other than time
- investigate methods for reconciling non-stationary analysis with FEH pooling techniques
- investigate the feasibility of deriving generalised allowances for non-stationarity based on spatial statistics

3.3.2 Literature review

The objective was to review literature relevant to methods of applying non-stationary frequency analysis. The review did not, in general, include literature on trend testing or investigations of trend in UK flood peak data, the latter of which was covered by the rapid evidence assessment project (see section 1.2.2). There is some literature that recommends avoiding non-stationary analysis; this was not included in the review apart from where it appears to provide useful pointers for practitioners. The review was not limited to UK literature.

References were particularly sought on the topics of:

- regional/pooled non-stationary analysis, including work on spatial trend/non-stationary analysis
- non-stationary analysis using physical covariates, both techniques for incorporating covariates and suggestions of useful covariates to try

34 papers, research reports or presentations were included in the review. Relevant papers are referred to throughout this section and in the supporting appendices.

Non-stationary frequency estimation is a popular topic for research, with new papers appearing every few days. This report includes reference to a few papers that appeared during the course of the research, as potential sources of ideas for further development of the methods.

3.3.3 General approach to analysis

Extreme value distributions were fitted to annual maximum flow series using the method of maximum likelihood estimation (MLE). This involves calculating a statistic known as the 'log-likelihood' and attempting to find its maximum value by varying the parameters of the model over their feasible ranges. It is a numerical method and convergence to a true maximum is not guaranteed. In contrast, FEH methods use L-moments to fit extreme value distributions. These are not readily adapted to work in non-stationary conditions.

Section 3.3.5 discusses choice of distribution. In conventional flood frequency analysis, for a particular gauging station, the distribution parameters are thought of as fixed quantities that need to be estimated. In non-stationary analysis, one or more of the parameters is not fixed. It might be changing over time, or changing in response to changes in some variable other than flow. These variables that may affect flood frequency are known as covariates (refer to section 3.3.6).

When developing methods, analysis using the GEV distribution was carried out using the extRemes package in R (Gilleland and Katz, 2016) and separate code was written to implement the GLO distribution. When implementing in the nonstat package, the texmex package (Southworth and others, 2020) was used in preference for statistical analysis, to enable full consistent implementation of the GEV and GLO distributions. This was also used for the national analysis.

3.3.4 Methods for judging goodness of fit

In non-stationary flood frequency analysis, there can be a large variety of models to choose from. Even if there is only one covariate, there is a need to choose between models in which only the location, only the scale, both or neither vary. If there are several potential covariates, the number of candidate models can grow rapidly.

The following methods are suitable for selecting between models, and practitioners are recommended to apply them all. The nonstat package in R produces outputs to help with the first 4 methods and the final one. The practitioner guidance provides more pointers and examples of how to apply them.

Likelihood ratio testing

Likelihood ratio testing is a hypothesis test that is carried out under a pre-specified significance level (usually 5%, as applied in this project). It can only be applied to sets of nested models, that is, the parameters of one model must be a subset of the parameters of the other models (Coles, 2001). The test statistic, or deviance, is calculated as:

$$D = -2(y - x)$$

- where y is the negative log-likelihood for the more complex model and x is the same for the simpler model
- the null hypothesis is that D=0, in which case the simpler model would be preferred. If the null hypothesis is rejected, the more complex model can be selected

This is a preferred approach when comparing a small number of candidate models, when the likelihood ratio can be calculated for each nested pair of models. It is impractical when comparing hundreds, which can be the case when several covariates are being considered.

AIC (Akaike information criterion)

The AIC establishes a trade-off between the goodness of fit and the simplicity of the model, measured by the number of parameters. It can be readily compared across a large number of candidate models, the lowest AIC indicating the preferred model. AIC and BIC (below) are derived from the information-theoretic approach (Burnham and Anderson, 2002).

BIC (Bayesian information criterion)

BIC gives more weight than AIC to model simplicity. In calculating AIC, the penalty for the number of parameters k is $2k$; for the BIC the penalty is $ln(n)k$ where n is the sample size, so if n >8 then the penalty for a more complicated model is greater and so there is a preference for simpler models.

Visual inspection of P-P and Q-Q plots

Visual inspection of model fit plotted on **probability-probability (P-P) and quantile-quantile (Q-Q) plots**

A probability-probability (or P-P) plot compares the following 2 quantities, calculated for each annual maximum flow in a series:

a) the value of the distribution function (that is, the non-exceedance probability) of the flow value, estimated from a statistical model
b) equally spaced points spanning the interval (0,1)

   More formally, it is a plot of the points:

$$\left\{\left(\hat{F}(x_{(i)}), \frac{i}{n+1}\right) : i = 1, \dots n\right\}$$

where $x_{(i)}, i = 1, \dots n$ is an ordered sample of independent observations and $\hat{F}$ is a candidate model for the true probability function F. The quantity $i/(n+1)$ corresponds to the empirical distribution function evaluated at $x_{(i)}$, that is, a plotting position. If $\hat{F}$ is a reasonable model for the true distribution, then the points in the probability plot will lie close to the unit diagonal.

A quantile-quantile (or Q-Q) plot compares the following, again calculated for each annual maximum flow:

a) the flow estimated using the statistical model from the empirical probability at step (b) above
b) the measured flow

   More formally, it is a plot of the points:

$$\left\{\left(\hat{F}^{-1}(\frac{i}{n+1}), x_{(i)}\right) : i = 1, \dots n\right\}$$

The quantity $\hat{F}^{-1}(\frac{i}{n+1})$ gives a model-based estimate of the $\mathrm{i}/(\mathrm{n}+1)$ quantile provided by the candidate distribution $\hat{F}$, while $x_{(i)}$ itself provides an empirical estimate of this quantile. Again, a well-fitting model would provide points on this plot lying close to the unit diagonal.

These visual comparisons are feasible when a small number of candidate models is being compared.

Hydrological reasoning

It is important that the statistical model makes physical sense. For instance, if rainfall is included as a covariate, its coefficient should be positive so that higher rainfall is associated with higher peak flows. It would be all too possible to fit a model with covariates that were correlated with peak flow, without any causal relationship being present. Another consideration is that the covariates should not be too correlated with each other.

Consistency of model form across locations

A consistent choice of covariates and type of relationship between covariates and parameters is expected for nearby and similar catchments, and particularly for gauges on the same river. For example, if all neighbouring gauges have a significant trend in the location parameter, it is likely that the trend should be included for a nearby gauge even if the trend is not statistically significant at that site. We expect results should have a level of spatial cohesion and this should be considered ahead of the random (stochastic) nature of hypothesis testing.

Consistency has been considered when it comes to choosing covariates. It is desirable to have a model that uses the same covariates at all locations, even if at some sites some of the estimated regression coefficients are not statistically significantly different from zero. Therefore, the research has given priority to covariates that are important at many locations.

Visual inspection of the return levels in comparison with the recorded flood peak data

A final check is to see if the model outputs look sensible. Often it is interesting to examine the exceedance probability of the largest flood(s) that have been observed. This judgment can be conceptually more difficult in a non-stationary setting, where a flood that occurred in a particular year might have a different exceedance probability if it occurred earlier in the record, or in a year in which the annual rainfall or NAO, for example, were different. The concept of the marginal return level can help.

For further considerations on model choice, refer to Xavier and others (2019), which was published towards the end of this project. Of the methods they tested, one finding was that the BIC performs best, when the true distribution is non-stationary with varying location, but the AIC should be preferred if the scale is varying.

3.3.5 Distributions for non-stationary analysis

Under stationary conditions the **generalised extreme value (GEV)** is the only possible limit distribution of maximum values of a sequence of independent and identically distributed random variables. A limit distribution models how large (or small) a variable will probably get. Therefore, there is a strong mathematical justification for the GEV distribution, and it is widely used by statisticians for extreme value analysis.

The **generalised logistic (GLO)** distribution is recommended by the FEH, where it is typically fitted using L-moment estimation.

The GEV and GLO are 3-parameter distributions. The definitions of both the GEV and GLO in this section represent how the distributions are implemented in the nonstat package. The notation here is slightly different from that of the FEH, and the shape parameter (ξ) has been set to be equal to $-k$ in the FEH definition, so that the finite upper end point occurs with negative ξ and infinite upper tail with positive ξ.

The **kappa** distribution is a generalisation of the GEV and GLO distributions with an additional, fourth parameter (h) that influences the shape of the body of the distribution.

In accordance with most previous work (for example, O'Brien and Burn, 2014), the third (shape) parameter of the GEV and GLO distributions was assumed to be constant because there is too much error in its estimation to include a covariate. This leaves the location and/or the scale parameters that can vary with covariates. The scale parameter was log-transformed to ensure it remains positive for all possible covariate values.

GEV

The GEV distribution function is of the form:

$$F(x) = exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \quad (3)$$

where μ is the location parameter, σ is the scale parameter, ξ is the shape parameter, $y_+ = \max\{y, 0\}$, $\xi \neq 0$ and $\sigma > 0$.

For simplicity, it was assumed that the location μ and the logarithm of the scale σ vary linearly with covariates. This assumption of linearity, while allowing for easily interpretable results and straightforward model-fitting, may not always be suitable. Additive models, where parameters are only restricted to be smooth functions of the covariate (Chavez-Demoulin and Davison 2005, Jonathan and others, 2014) offer greater flexibility despite reduced extrapolation capability.

For a vector of covariates x:

$$\mu(x) = \mu_0 + \mu_1 x_1 + \mu_2 x_2 + \cdots \quad (4)$$

$$\sigma(x) = \exp(\phi_0 + \phi_1 x_1 + \phi_2 x_2 + \cdots) \quad (5)$$

Therefore, for a non-stationary fit there are 2 or more elements of the location parameter to estimate; a constant component μ_0 and μ_1, μ_2, which represent the influence of the covariates on the parameter, and the same for the scale parameter.

There are methods intended to estimate more realistic values of the shape parameter, which can be poorly estimated from small samples. The 'geophysical prior' of Martins and Stedinger (2000) is sometimes used to restrict the range of shape parameters based on previous hydrological experience (Renard and others, 2013). This is a beta distribution bounded to the interval (-0.5, +0.5). This method can be adapted to the case of MLE, where the prior information is added as a penalty to the likelihood function. It is implemented in the extRemes package as an option known as generalised maximum likelihood estimation (GMLE).

The effect of GMLE on the estimation of the shape parameter was tested at a sample of stations. It was found to lead to an increase in the parameter; in some cases, a large increase when the estimate from the standard MLE is close to zero. This can lead to a large and, sometimes unrealistic, increase in the estimated return levels for long return periods.

It appears that, far from helping to constrain the range of shape parameters, GMLE, as implemented in extRemes, can exaggerate the shape parameter. For the present study, the shape parameter was therefore not constrained, that is, the GEV was fitted using the standard MLE method. It was subsequently found that the default implementation in the extRemes package may be using a beta distribution bounded to the interval (0, 1), rather than (-0.5, +0.5), which may be pushing the estimates towards more positive values. Further investigation could look into determining a penalty weight

for the GEV shape parameter from UK flood peak data, using an approach given by Gabda and Tawn (submitted).

GLO

The GLO distribution function is of the form:

$$F(x) = \frac{1}{1+\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}} \quad (6)$$

where all parameters are defined in the same way as for the GEV distribution, and the same assumption was made that location and scale vary linearly with covariates for non-stationary fits (Equations 4 and 5).

Kappa

Kjeldsen and others (2017) recommend wider use of the kappa distribution, stating that it renders 3-parameter distributions 'obsolete' on most UK catchments. A drawback is that the fourth parameter introduces another degree of freedom, leading to the danger of overfitting. In their analysis, Kjeldsen and others (2017) fixed the value of the fourth parameter (h) to a national value of -0.40. Therefore, the distribution for fitting became a 3-parameter distribution lying somewhere between a GLO (h=-1) and a GEV (h=0). Further, Kjeldsen and others (2017) fitted this distribution in a pooled analysis, in which enough data are available to obtain relatively stable estimates of the parameters.

This project included development of code for fitting both stationary and non-stationary versions of the kappa distribution. Early results of stationary fitting produced some unrealistic extrapolations when all 4 parameters were allowed to vary, an example of the overfitting problem mentioned above. Further work would be needed to fit a version of the kappa distribution with a fixed fourth parameter, but it was decided not to take this further within the project.

Comparison of GEV and GLO fit

Tests at 5 trial gauging stations compared the fit of stationary and non-stationary (varying location parameter) versions of the GEV and GLO models. Model fits were compared using AIC, BIC and inspection of P-P, Q-Q and return level plots. Figure 3.1 shows an example. Overall, there was little to choose from between the 2 distributions. The fit statistics were generally slightly better for the GLO, which is in accordance with findings in the FEH.

The nonstat package offers functionality to fit either the GEV or GLO distributions, so practitioners are able to choose an appropriate distribution for their data.

Correlation of distribution parameters

An alternative formulation of the GEV distribution was investigated, which accounts for the correlation of the location and scale parameters. This was suggested in Ilaria Prosdocimi's peer review of the non-stationarity project in north-west England. The idea of linking the variation of the scale to that of the location parameter was also explored in Coles and Tawn (1990) and suggested by Steirou and others (2019), although the latter authors ended up assuming a constant scale parameter.

In the FEH method, the coefficient of variation (L-CV) is pooled across stations after the peak flow data is standardised by the index flood, a measure of the scale. Under this approach, the scale parameter would be linked to the location:

$$\mu(x) = \exp(\mu_0 + \mu_1 x_1 + \mu_2 x_2 + \cdots) \quad (7)$$

$$\sigma(x) = \tau \exp(\mu_0 + \mu_1 x_1 + \mu_2 x_2 + \cdots) \quad (8)$$

where τ is a measure of the coefficient of variation. Unlike in the standard formulation (Equations 4 and 5), an exponential function is used for both the location and the scale parameter, so that the scale can be more straightforwardly related to the location. This could be improved by imposing a constraint to ensure that τ remains positive.

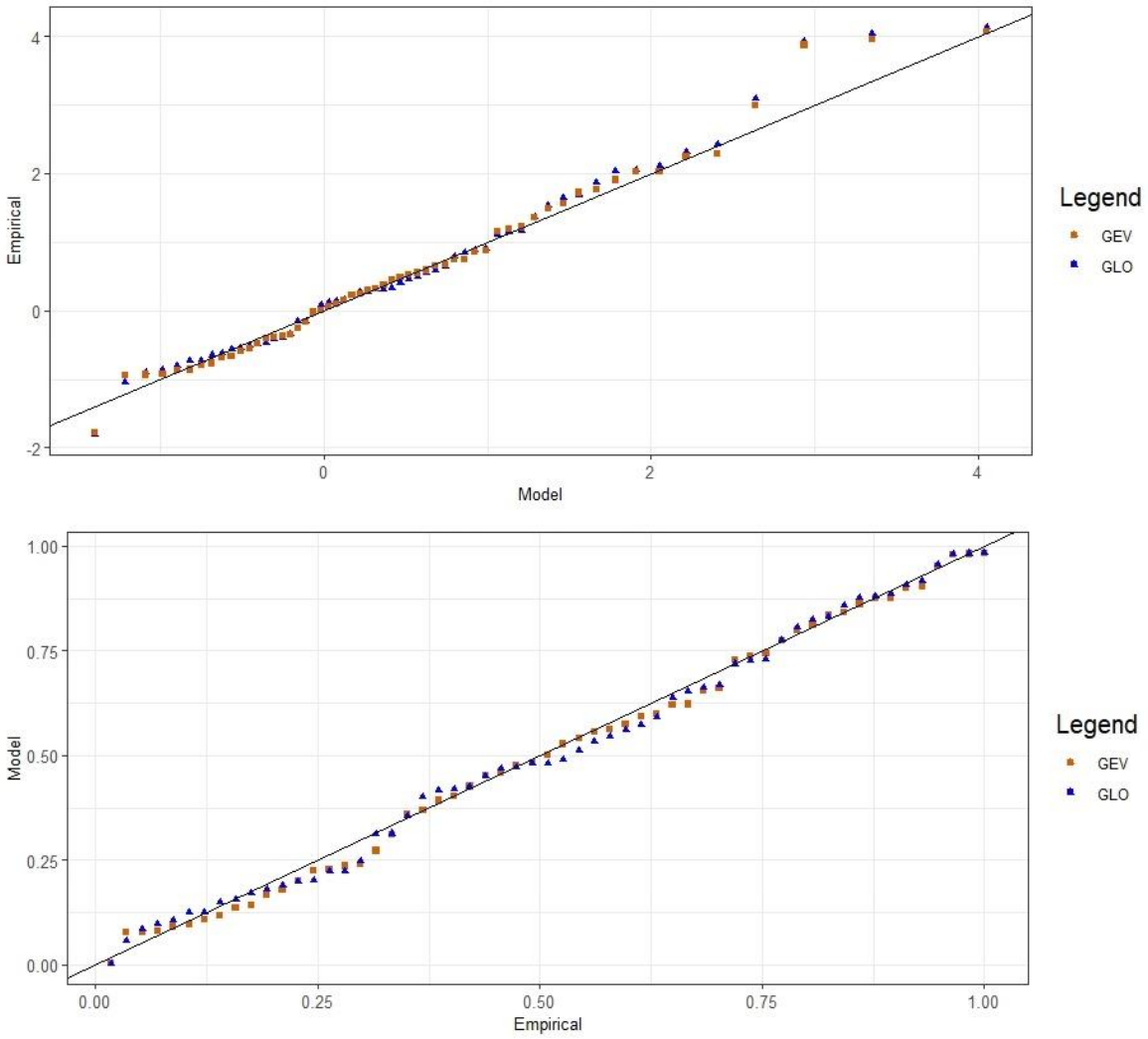


Figure 3.1: Comparing non-stationary GEV and GLO models fitted to station 24004, Bedburn Beck at Bedburn, using a Q-Q plot (upper) and P-P plot (lower)

An alternative with more orthogonal parameters might be:

$$\sigma(x) = \tau \exp(\mu_1 x_1 + \mu_2 x_2 + \cdots). \quad (9)$$

There was an expectation that one of these formulations might help avoid some of the unrealistic extrapolations that could occur when both the location and scale parameters independently vary with time. A more complex formulation would allow τ to vary with time.

To investigate this relationship between location and scale, non-stationary GEV models were fitted at 5 trial gauges using both the standard form and the linked version, defined in Equations 7 and 8. Model fits were compared using AIC and diagnostic plots.

Overall, the 2 versions of the GEV model showed a similar fit. For 3 out of 5 gauges, the standard parameterisation model had a better fit as judged by AIC. At one gauge, there was a convergence problem with the linked model. This can occasionally happen given the exponential form in the location parameter (Equation 7); any poor starting value could make subsequent maximisation difficult due to the high correlation between the scale and location forms.

In light of these results, it was decided that the linked model did not offer a clear benefit and so subsequent investigation used the standard non-stationary GEV model, as defined in Equations 4 and 5. A linked version of the GLO model was not investigated.

3.3.6 Covariates for modelling non-stationarity

One of the largest areas of the investigation was to select physical covariates, investigate ways of incorporating them into non-stationary models and, crucially, develop methods for extracting useful results for practitioners. Appendix B has a full report of this work. A summary is given here.

Reasons for incorporating covariates

Two reasons are sometimes given for modelling non-stationarity of floods using physically-based covariates. The reasons are:

1. Physical covariates help remove some of the year-to-year variability in AMAX flows, allowing time-based trends to be better identified and better fit of the distribution. For example, in Equation 4, if covariate x_1 represents time and covariate x_2 is a measure of rainfall, μ_1 would express how floods change over time, once variations in rainfall are taken into account by μ_2 (Prosdocimi and others, 2014).
2. They provide a more physically meaningful model of non-stationarity, since time on its own has no physical influence on flooding. As a covariate, time is merely a substitute for some other physical quantity that is influencing floods. Some physical covariates may open up the prospect of predicting the future evolution of the flood frequency curve (Sraj and others, 2016).

Reason (1) leads to models that include both time and physical quantities as covariates. To avoid correlation between the covariates, it is desirable that any trend in the physical covariates is removed before they are included in the non-stationary model. The time covariate will then represent the presence of any temporal trend in the flood peak series.

Reason (2) tends to lead to a rather different approach in which the physical variables replace time as a covariate. Within this approach, to model a flood series that has a trend over time, it would be preferable to include at least one physical covariate in the model which exhibits a time trend[1]. This then introduces a need to model the trend in that covariate in order to understand how flood magnitudes are changing over time. The hope here is that a covariate can be found for which the trend is easier to

[1] Without this, the fitted model is another another version of a stationary model, one with covariates that do not have a time trend.

model than that in the flood series, perhaps because it has less variability and more predictability into the future. An example might be the extent of urbanisation in a catchment, which can be typically expected to show a monotonic increase over time. Additionally, urbanisation can be reasonably predicted into the future under a range of scenarios.

A risk associated with this second approach is confusing correlation with causation. In principle, it would be possible to include any covariate with a trend, whether or not it had any physical connection with the processes that cause floods. This could lead to a false sense of confidence about the ability to estimate the future evolution of the flood frequency curve. It might be possible to end up with a covariate for which future values can be confidently predicted, but which is no more useful than the date as a way of explaining observed trends in flood magnitudes. Therefore, it is necessary to demonstrate a strong causal relationship for physical covariates if they are to be used for predictions.

Choice of physical covariates

In light of findings from the literature review, the following were selected as trial covariates:

- catchment-average rainfall, calculated over the water year, the autumn and the winter seasons. Rainfall accumulations were calculated from the Centre for Ecology & Hydrology Gridded Estimates of Areal Rainfall (CEH-GEAR) data set, which provides daily rainfall on a 1 km grid across the UK from 1890 (Tanguy and others, 2016).
- North Atlantic Oscillation (NAO) index, averaged over the winter, summer and autumn
- East Atlantic (EA) index, averaged over the winter
- global mean temperature anomaly, averaged over the year and over the winter. This covariate was discarded after trend analysis showed a near-linear trend over the period of flow records. Although superficially it might be thought that temperature is a useful covariate as its future evolution can be predicted with reasonable confidence by climate models, this reasoning could equally apply to using time as a covariate, since its future values can be predicted perfectly. Since there is no clear causal connection between global temperature and UK flood magnitudes, it was not considered further as a covariate.

These covariates, although physically-based, do not directly represent the physical processes that cause floods. For instance, none of the covariates measures the strength and direction of atmospheric rivers, which have been linked with the occurrence of winter flooding in the UK (Lavers and others, 2011). However, they represent a potential step forward from the simplistic approach of modelling non-stationarity as a change over time.

The research considered only covariates that are expected to be significant across many catchments in preference to those that represent locally-specific effects such as urbanisation or changes in forest cover. The nonstat package is flexible, allowing practitioners to include any covariates they can obtain.

The 7 remaining candidate physical covariates were included in fitting of non-stationary models on some trial catchments in accordance with the following options:

1. In a group, allowing any number of physical covariates plus water year to be included, with the physical covariates being detrended.

2. Up to 2 covariates per model, with a maximum of one being a physical covariate, the other being water year, with the physical covariate being detrended.
3. Up to one physical covariate per model, with no detrending.
4. Only time allowed as a covariate.

In all these cases, covariates were considered for modelling either or both of the location and scale parameters.

This leads to a large number of candidate models. Even when only one physical covariate is allowed at a time, there are 22 models fitted (7 physical covariates times 3 for location, scale and both varying, plus one with no covariates). With up to 2 covariates, there are 88 models fitted. The number can grow to many thousands when more covariates are considered, which can lead to unfeasibly long run times.

Appendix B presents results on the trial catchments.

Incorporating physical covariates into estimating design flows

One of the biggest challenges the project faced was how practitioners could extract estimates of design flows from models fitted to physical covariates. Design flow estimates from a non-stationary model that uses physical covariates will change not only over time, if water year is included as a covariate, but also with the value of the other covariates. For instance, if the covariate is annual rainfall, then the 1% AEP design flow given 1,200 mm of rainfall is the expected flow under the (clearly hypothetical) conditions that the annual rainfall is always 1,200 mm. This quantity is known as the **conditional flow estimate**, or more formally as the **conditional return level**.

The conditional flow estimate may be useful when examining the probability of past floods, but it is less informative when thinking about design.

The **integrated flow estimate**, formally the **marginal return level** (Eastoe and Tawn, 2009) removes the dependence on a particular value of the covariates. It is defined as the return level corresponding to the encounter probability[2] averaged over covariates in a period of interest. The marginal return level should be understood as applying over a period rather than instantaneously. This is a useful concept for planning investment decisions in flood risk management, which need to consider the probability of floods occurring over the period of the planning horizon. Appendix B provides a formal mathematical definition of the marginal return level.

The marginal return level is normally calculated by averaging over the covariate values observed during the period of flow record. In theory, it can also be calculated by averaging over a different distribution of covariate values, for instance one that is intended to represent future conditions. This is not recommended for practical application because it is only valid if the physical covariates provide a complete causal description of the non-stationarity in peak flows. For example, if the covariate was annual rainfall, this calculation would assume, wrongly, that future changes in peak flow can be entirely explained by changes in annual rainfall. Although it is expected that climate change may affect annual rainfall, and therefore catchment wetness, it can also be expected to influence other factors that control flood magnitudes. These include storm intensity and evapotranspiration (which influences soil moisture).

[2] An encounter probability is the probability of an event occurring in a given number of years. The concept is explained further in the practitioner guidance.

If the covariates include both water year and physical variables, it is possible to calculate a marginal return level by averaging the probabilities corresponding to the observed physical covariate values, but setting the water year covariate to a single value, such as the final year of record. This gives what the project has termed a **single-year integrated flow estimate**. If the flow record runs up to the present day, this estimate represents the present-day expected flow for a particular exceedance probability, without being conditional on any particular value of a covariate. The single-year integrated flow estimate can be more easily compared with alternative estimates such as those from a model that uses only water year as a covariate.

The nonstat package can calculate the various statistical outputs mentioned above, along with suitable graphical outputs. The single-year integrated flow estimate is calculated only for the final year of record. The output from the package refers to this quantity as the **present-day marginal return level.**

Confidence limits

Confidence limits are estimated using a parametric bootstrapping procedure. This is a method of deriving confidence limits in situations where the underlying statistical population is unknown or where an analytical solution is impractical. For the integrated flow estimate and single-year integrated flow estimate, the process involves resampling the covariates (drawing randomly from the covariates with replacement), keeping the time covariate fixed, then re-estimating the parameters of the fitted distribution. New process values (flows) are then sampled from the distribution with the new parameter estimates, to generate a new flow time series. Finally, the model is refitted to the new sampled data and return levels are estimated from this final estimated model. This process is repeated a large number of times and confidence intervals are extracted from appropriate quantiles of the results.

For the stationary and conditional return levels, the process is the same, but missing the first 2 steps (that is, omitting the resampling of covariates and subsequent refitting of the model with the resampled covariates). Process values (flows) are resampled from the original fitted model (which either has no covariates, in the stationary case, or has fixed covariates, in the conditional case), then model parameters are re-estimated by refitting the model to the new resampled data, and return levels are estimated from this model.

3.3.7 Pooled non-stationary analysis

Pooled analysis was another major part of the research, since it is considered important to build bridges between non-stationary analysis and the widely-used FEH method. A particular aspiration was to be able to apply non-stationary methods at ungauged locations. This section gives an overview, and the details are in Appendix C.

The research considered how to apply non-stationarity within the pooling-group framework. This included 2 parts:

- how to account for trend when forming pooling groups, developing a new similarity distance metric
- how to estimate index floods and flood growth curves that incorporate trend

Both parts compared the developed methods with existing approaches to check for improved performance.

The first part investigated including trend descriptors as components in the similarity distance metric (SDM) currently used in the UK to form pooling groups. Alternative

similarity distance metrics were investigated, based on using the Theil-Sen estimator of trend as a component. These were based on fitting L-CV and L-SKEW models, and recalibrated using existing pooled uncertainty optimisation. The SDM based on an L-CV model performed similarly to the existing FEH. The Theil-Sen estimates of slope were slightly more accurately predicted when using a similarity metric which included trend as a component. The 20-, 50- and 100-year floods seem to be predicted with similar accuracy with the old and new SDMs.

The second part built on the first, looking into the most appropriate method of using index flood methods and growth curve formulations within pooling group methods. This was primarily to improve flood frequency curves where non-stationarity may be present, and secondly to improve estimates of trend (or confirm the absence of trend) at locations with short or no gauged records of flow. Methods of choosing stationary and non-stationary index floods and growth curves were investigated. When trialled on observed data, mixed signals were observed, though trends in the scale parameter led to consistently larger estimates in the 20-year flood. By using simulations with realistic dependence structures, it was observed that correctly modelling trend is important for accurate parameter estimates, particularly in modelling GLO scale parameters.

The investigation recommended that the current FEH SDM should still be used for forming pooling groups, but if regions or pooling groups are generated to account for trend, care should be taken to select stations with like trend (positive, negative or no trend). Non-stationarity in growth curves can be helpful in very specific cases with consistent pooling groups in terms of like trend. However, non-stationary analysis should be used with caution, due to problems in fitting of such curves through maximum likelihood methods on short records, and should not be used for extrapolation beyond the present until further work is conducted.

More work must be done before it can be recommended that practitioners solely use non-stationary index floods or growth curves. If all stations within a pooling group have similar trend, then incorporating a trend into the growth curve is reasonable, and should be considered and compared with the stationary growth curve.

In the meantime, practitioners will face challenges in reconciling the results of non-stationary analysis, at individual gauging stations, with those of pooled analysis using FEH methods.

3.3.8 Spatial statistics

The research included some exploratory work to fit a spatial statistical model. The motivation was to boost the strength of the trend signal by combining data from multiple sites, helping to remove some of the ‘noise’ introduced by year-to-year variability in peak flows. A Bayesian hierarchical model was fitted, applying a gauge-specific GEV distribution with time covariate, whose parameter μ_1 (defined in Equation 4) is modelled as a normal distribution with shared variance across a predefined region such as a hydrometric area.

Appendix D contains a description of this investigation. The model shows promise but is not ready for incorporating into the interim guidance for practitioners.

3.4 Allowing for the impacts of climate change

3.4.1 Scope of investigation

The project included an investigation of different approaches for applying climate change allowances to the results of non-stationary flood frequency analysis. Appendix E contains a detailed report of this investigation. The sections below provide a summary.

The report discusses various ways in which climate change allowances are, or could be, applied. It compares how different extrapolations to 2025, 2050, and 2080 are affected on a regional scale, depending on which baseline is chosen and whether the baseline is assumed to be stationary.

3.4.2 Existing guidance on climate change allowances for England

Agencies across the UK have been providing guidance on the potential impacts of climate change on floods for many years, so that these can be accounted for by flood management authorities and local planners aiming to reduce flood risk (Reynard and others, 2017). The most recent guidance adopts a regional risk-based approach (Environment Agency 2016 a,b), and is based on combining the UK Climate Projections 2009 (UKCP09) (Murphy and others, 2009) with a sensitivity-based approach to modelling the impacts of climate change on peak flows (Kay and others, 2011; 2014).

The guidance for flood management authorities (Environment Agency, 2016a) provides a set of 5 numbers (lower, central, higher central, upper and H++) for each of 11 regions covering England for 3 future time slices (2020s, 2050s and 2080s). The 'lower', 'central' and 'upper' numbers represent the main range of estimated impacts of climate change on flood peaks from the UKCP09 projections. The H++ numbers represent plausible but unlikely high-end impacts of climate change.

The guidance recommends that the central estimate of change should be used to define the risk over the decision lifetime, with the upper and lower estimates provided to encourage the options required to manage the fuller range of risk to be considered, for example, building flexibility into the plan to allow future adjustments, if necessary (Reynard and others, 2017).

3.4.3 Issues when applying the guidance

Current guidance on applying climate change allowances (CCAs) is somewhat open to interpretation. Some of the issues are listed below.

- Climate change allowances are derived from climate projections from a 1961 to 1990 baseline, typically using hydrological modelling for the baseline period 1961 to 2001 (Kay and others, 2014). 1961 to 1990 continues to be the standard baseline in new research updating CCAs using UKCP18 climate projections.
- The impacts for the 2020s time slice are based on the potential climate change between the baseline period and the period 2010 to 2039, therefore raising the question of whether some of the climate change has 'already happened'. If so, is applying the full allowance still valid?

- Even if a clear trend is apparent in the AMAX data for a particular catchment, it could be for a range of reasons other than climate change, including natural climate variability.
- The impacts of any of these effects can be difficult to spot given the large natural variability seen in peak flow data.
- While anthropogenic climate change may be a major driver of non-stationarity in peak flow data, it is very important that global change impacts are attributed reliably and the risk of 'climatisation' is avoided by taking non-climatic factors into account (Wine and Davison 2019).

3.4.4 Methods investigated

Five methods of accounting for climate change were trialled, and these are listed below. Each method was applied across the full project data set of gauging stations suitable for non-stationary frequency analysis.

Table 3.1 Methods of accounting for climate change

Code	Period of record analysed	Type of flood frequency analysis	Baseline year for extraction of results	Method for adjusting for potential future conditions	Comment
STFULL	Full	Stationary	n/a (any)	Current climate change (CC) allowances	Current approach used by practitioners
ST6190	1961 to 1990	Stationary	n/a (any)	Current CC allowances	Results intended to be representative of 1961 to 1990 baseline
NSTREP	Full	Non-stationary	1990	Current CC allowances	
NST6190	1961 to 1990	Non-stationary	1990	Current CC allowances	
NSTEXT	Full	Non-stationary (varying location with time)	n/a	Extrapolation of non-stationary model	Purely for comparative purposes. Not to be applied in practice.

3.4.5 Findings from investigation

- On average, using full periods of record with climate change allowances applied (STFULL) leads to the largest estimates of flow among those methods considered. This means that the present method of applying CCAs to the full record tends to lead to the most conservative estimates, since this report shows that Q50 and Q100 estimates are typically largest in most regions using this method.
- The non-stationary representative method (NSTREP) gives a reasonable compromise; using non-stationary peak flow estimates representative of 1961 to 1990 (the baseline period for CCAs), but also accounting for trends in more recent data within the analysis. This method should be considered to use if a more precise estimate of trend is important, even if that gives smaller future

estimates of Q_T, but only in circumstances where trends are thought to be driven by climate change.

- Stations with significant positive trend tend to show a mixed picture; STFULL does not always produce the largest estimates of flow for these stations. This is most evident in the south west and Wales. Multiple methods with and without non-stationary approaches should be consulted to get a fuller picture of possible future flow.
- Methods that analyse only 1961 to 1990 data are not recommended as the more recent data are invaluable in giving a more accurate present-day picture. Always try to use a method of future flow estimation which includes as much good quality data as are available.

3.4.6 Implications for practitioners

Although it is very difficult to make general recommendations from the results, the following comments may help practitioners choose an approach in cases where climate change rather than any other factor is believed to be a driver of non-stationarity:

- No evidence was found to suggest that the existing climate change allowances should be revised other than to apply the new UKCP18 probabilistic climate projections. This update is already under way based on the outputs of the project 'Providing more locally-appropriate information on potential impacts of climate change on flood peaks in England and Wales' (Kay and others, 2019). The intention is for the outputs of the project to be made available via a web tool. However, the Environment Agency has not yet made any decisions on updates to guidance on flooding and climate change.
- To assess whether climate change has already started to affect flood frequency, QMED and/or a higher quantile estimated over the 1961 to 1990 baseline period could be compared to that estimated from the full record. If the two differ substantially, the effect of applying the climate change allowances to the baseline and full record estimate could be explored. The split sample tests carried out within this project make this type of comparison (section 0).

The guidance for practitioners includes 2 suggested approaches for adjusting non-stationary flood frequency estimates for future climate change, based on the comments above.

For any applications where it is only the short-term future that is of interest, an alternative approach would be to adopt the present-day estimate from a non-stationary model, treating this as a new baseline, which might be representative of the next few years, into the 2020s. This approach would amount to an updated stationary model, assumed to be applicable over the short term. Research in the USA using split-sample testing (Luke and others, 2017)[1] has shown that this can be a better assumption than either extrapolation of a non-stationary fit into the future or fitting of a stationary model in the presence of changes in flood frequency (albeit caused by changes in land use in that case).

Ultimately, it will not be possible to develop definitive guidance without tackling the challenge of trend attribution.

3.5 Clustering

The primary focus of the project was on long-term trends in flood magnitude. It can be notoriously difficult to distinguish between persistent trends and cycles of flood-rich or flood-poor periods, some of which can operate over several decades. As well as clusters of flood-rich years, shorter-term clusters are also seen, for example in wet years such as 2012, winter 2015 to 2016 and winter 2019 to 2020. These sequences of events can pose challenges for managing flood incidents.

Understanding temporal clustering better may help the attribution of trends and therefore understanding whether and how they might continue. It could help determine whether practitioners are relying on an unrepresentative period spanned by peak flow data sets. Knowledge of clustering on a shorter timescale may help plan response and communication in the aftermath of flood incidents, and perhaps have implications for designing storage-based schemes.

Within the current project, the scope for investigating temporal clustering was limited to an initial exploratory analysis. This involved:

- literature review covering both clustering and also identifying flood-poor and flood-rich periods using longer-term sources of information
- developing data sets, screening long-term series of POT (peaks over threshold) flow data. Fourteen gauges were selected for the investigation, all with at least 60 years of reliable POT data
- quantifying the degree and duration of clustering using the index of dispersion

Appendix F contains a report on the investigation.

Statistically significant clustering was found at all gauges. The strongest clustering tended to occur between 2 and 6 years, and up to 10 years on some rivers.

If clustering is present successive POT data may not be independent, despite applying rules of thumb intended to ensure that POT data are independent, as described in the FEH and modified for the NRFA data set[3]. It may even be the case that some successive annual maximum flows are not independent. The methods of both stationary and non-stationary flood frequency analysis described in this report all assume that peak flows are independent. Attributing any dependence may help in fitting more appropriate statistical models.

There are several possible avenues for follow-up work. A more in-depth investigation could answer some of the following questions:

- How does the degree of clustering vary with flood magnitude, with catchment type or location?
- Is clustering primarily due to hydrological or meteorological causes?
- What information do we have on the typical duration of flood-rich or flood-poor periods?
- How much more likely is a flood to occur in the aftermath of another large flood?
- How can we quantify the probability of a sequence of floods as opposed to a single event?

[3] https://nrfa.ceh.ac.uk/peaks-over-threshold, accessed 25 March 2020. Accessed 12 May 2020

- Do the rules of thumb for identifying independent POT data need to be improved?

Appendix F provides some pointers as to how these questions could be answered.

4 National results

4.1 Introduction

This chapter presents the results of the methods described in the previous chapter, applied to the set of gauging stations described in chapter 2. Results are presented for trend tests, split sample tests and non-stationary flood frequency analysis. Section 0 discusses some of the similarities and differences between the findings of all these analyses.

4.2 Trend tests

4.2.1 Results

Records of suitable length and completeness were available at 471 gauging stations. Appendix A and the digital outputs (FRS18087-IG-D2-digital_outputs.zip) contain the full results. A summary is provided here.

Three periods were selected to compare results from all stations and assess the spatial variability of the trends in England and Wales. These periods are short (starting 1987), long (starting 1967) and full. For full consistency, results were also produced using a fixed start and end year (that is, 1987 to 2016 and 1967 to 2016). In this case, all stations have an identical record length and so can be compared fairly.

For the 5 fixed periods over which trends were analysed, most stations have increasing trends (

Table 4.21). There are more than twice as many stations with increasing trends as decreasing trends. Depending on the record length examined, 10 to 21% of gauges show increasing trends that are significant at a 10% or 5% level.

Table 4.1 Summary of MKZ trends for 5 periods of interest. Each cell shows the number of gauges within the category

	Short (1987-)	**Long (1967-)**	**Full**	**1987 to 2016**	**1967 to 2016**
Number of gauges with suitable data	435	272	471	403	202
Direction of trend					
Positive (MKZ>0)	287	186	318	261	132
None (MKZ=0)	14	6	4	14	5
Negative (MKZ<0)	134	80	149	128	65
Significance of trend					
Positive, 10%	48	52	99	42	26
Positive, 5%	30	37	63	25	22
Negative, 5%	5	8	12	5	7

Negative, 10%	10	13	22	10	12

Similar conclusions are found by examining the results for all the combinations of start and end years in Appendix A. The median percent of positive trends is over 74%, meaning that half of the stations have more than 74% of positive trends for all the examined combinations. There are some stations with persistent positive/negative trends for all the combinations. For example, at station 54040 (the Meese at Tibberton), about 40% of combinations of start and end years show a significant negative trend. There are 29 stations with positive trends for all combinations of start and end years, while 5 of them have over 80% of these trends classed as significant.

However, the multi-temporal analysis, the results of which are provided digitally, more typically illustrates how trends in fixed periods are not representative of the full range of hydrological variability. There are often changes in the magnitude and even direction of trends over the course of the period of record. This sensitivity to start and end years is a very widely known issue and discussed at length in the literature (see Hannaford, 2015 and references for a UK context on this issue).

A map of the Theil-Sen approach (TSA) measure of trend magnitude for the full period of record at each gauge (Figure 4.1) shows the propensity for positive trends in peak flows across much of England and Wales. Large areas of central and eastern England also display negative trends, but these are often non-significant (except for a coherent cluster in the Thames catchment and some in the north-east and Midlands at the 10% significance level). In comparison, significant increasing trends are prevalent across some areas, especially in northern England, Wales and parts of western central England.

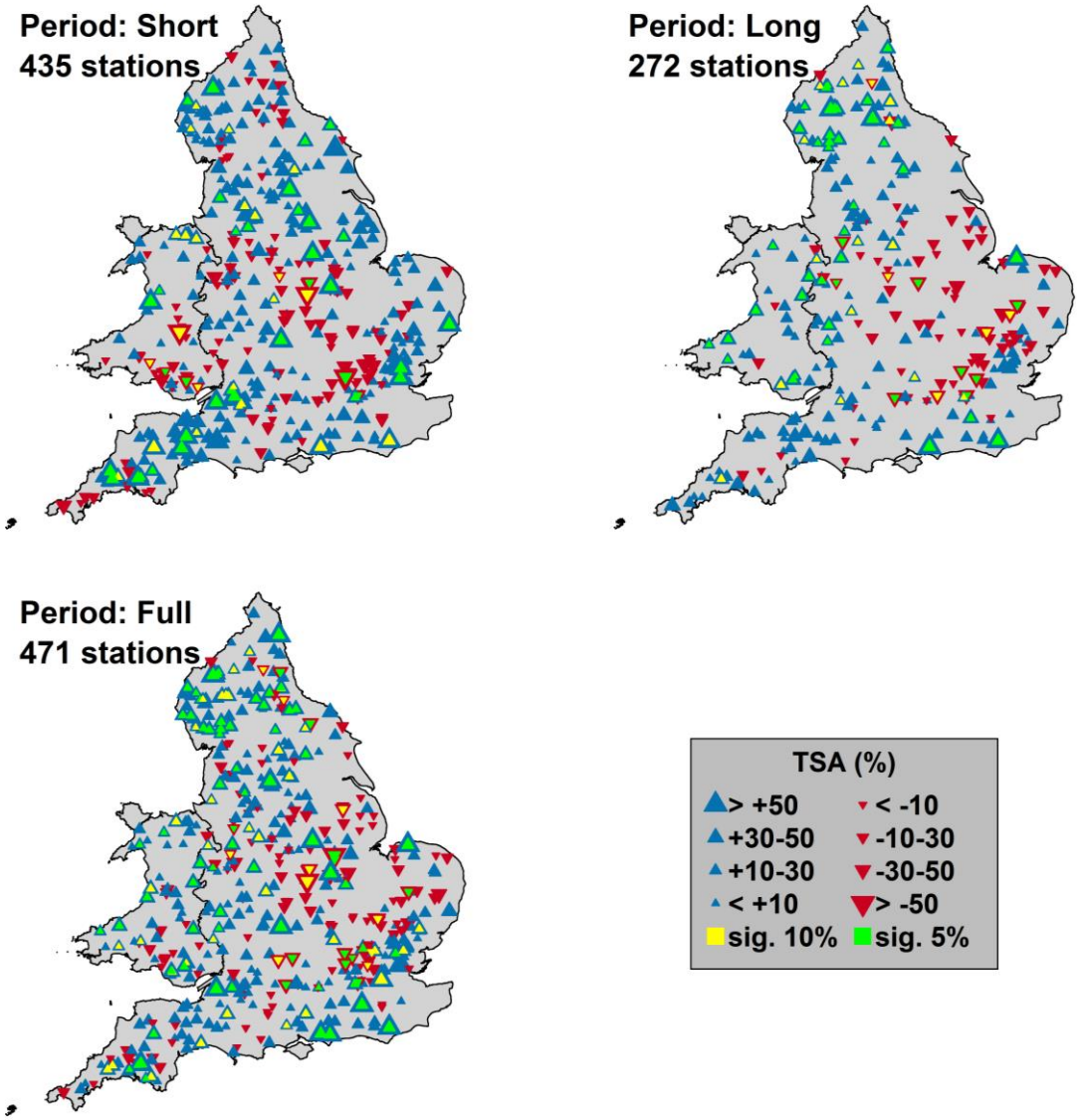


Figure 4.1 Maps of trend results for the full period of record and the short and long periods

For the fixed periods of record, short (1987 onwards) and long (1967 onwards), the patterns are broadly similar, despite the sparser coverage.

Cumbria has been a focus of attention regarding trends in peak flows, and hydrometric areas 73 to 76 cover Cumbria and parts of surrounding counties. Looking at the full period of record, all 32 stations in Cumbria have a positive trend, of which 56% are significant at a 10% significance level and 41% are significant at a 5% level.

4.2.2 Discussion

If we take the long period as the most robust outcome (balancing the length of record, while also being a fixed period allowing comparison between sites) to compare with other work, we can conclude that the national picture broadly agrees with previously published research on trends in AMAX and other flood indicators (Hannaford, 2015). That is, there has been a tendency towards higher flows in northern and western areas over the last 4 to 5 decades. The majority of that work was carried out in study periods ending in the mid to late 2000s, so the current study provides an update of around a decade.

Echoing other studies published in the last few years (for example, Brady and others, 2019; Prosdocimi and others, 2019), it appears that the previously identified gross patterns of change in the UK are fairly resilient, that is, these tendencies have not been countered by adding new data. Indeed, if anything, the results show an increase in positive trends and in the proportion of significant trends. To a degree, this is unsurprising given that the recent decade includes some very major flood events (for example, the winter 2015 to 2016 floods, which have a strong influence on the number of significant positive trends in northern England).

The results accord with Harrigan and others (2018), who, using the same testing methodology reported primarily positive trends in high flows (the Q5 flow in each year), with significant trends in northern and western areas. However, the current study uses peak flow data as opposed to daily flow data. It also uses the NRFA Peak Flows (v7) data, which includes AMAX data up to the 2016 water year, whereas Harrigan and others (2018) featured data up to 2014. The higher (relative) number of significant trends in England and Wales in the current study may reflect the addition of the 2015 to 2016 floods.

The current study also features the entire peak flows data set (that meets the agreed study criteria) rather than focusing on near-natural catchments. The agreement with Harrigan and others (2018) is encouraging, as that study deliberately focused on near-natural, high quality stations to prevent spurious trends arising from poorer quality data or human effects. Here, there are similar geographical patterns using the whole peak flows data set, including very heavily influenced catchments. This suggests that, at the national scale, a similar ‘headline’ picture emerges even when all catchments, of varying properties and degrees of disturbance, are mixed together.

One important spatial contrast with Harrigan and others (2018) is that there are more negative trends in the present study, particularly in central England, than in the results for the UKBN. This is discussed in Appendix A.

The general dominance of positive trends for the UK agrees with several recently published studies of spatially coherent trends in flooding at the national scale (Brady and others, 2019, Prosdocimi and others, 2019). These studies use a novel Bayesian approach to characterise regional and national-scale trends, rather than focusing on at-site trends. While these studies show that the signal towards positive trends in flooding is prevalent at the large (national) scale, Prosdocimi and others (2019) also show significant regional variations in the strength of trends, with broadly similar patterns to those shown here.

4.2.3 Trends over the longer term

There is considerable discussion in the literature about how long a period is needed to adequately quantify trends, with 50 years sometimes cited (for example, Kundzewicz and Robson, 2000). Arguably, no single period can reliably be used to characterise trends given the presence of interdecadal variability.

Such variability between 'flood-rich' and 'flood-poor' periods is clearly prevalent in UK hydrometric records, as shown in the smoothed LOESS (local polynomial regression fitting) series presented in the trend outputs in Appendix A. Even in a 50-year record, longer periods than the extant record could give different results still. A range of studies has shown that historical flood events reconstructed from epigraphic, documentary or other sources are often larger than events in contemporary gauging station records, and flood-rich and flood-poor periods exist in reconstructed records covering centuries (for example, Wilby and Quinn, 2013; MacDonald and Sangster, 2017).

For this reason, the multi-temporal analysis provides important context. It is important to underline that a trend in a fixed period should be seen as a descriptor of non-stationarity in that period alone. Trends should not be extrapolated, because, as the multi-temporal analysis makes abundantly clear, variability over a range of scales (years to decades) means that the strength and direction of trends could change in future.

The multi-temporal analysis also helps overcome the issue of 'stopping rule bias', which refers to the tendency for analyses such as this project to take place in the aftermath of large events. This tends to mean that stronger trends are found than if the timing of the analysis were unrelated to the occurrence of recent floods.

An important question is whether a trend for any given period is representative of a long-term change rather than short-term variation. This question is difficult to avoid if we want to use non-stationary techniques to support investment in future flood alleviation schemes, as discussed in section 0. The question cannot be answered by any of the techniques for detecting trend or estimating non-stationary flood frequency applied in this project. It should be addressed by moving from detection to attribution, in other words, identifying the processes driving the trend Merz and others (2012).

This is, however, a much more challenging scientific task and one that requires blending observational data sets with hydrological models. This is typically extremely resource intensive even at the catchment scale (for examples, see Harrigan and others, 2015; Prosdocimi and others, 2017), and therefore is beyond the scope of the present study.

4.3 Split sample tests

4.3.1 Change in location of the distribution

400 AMAX series have enough data both before and after the split point of 1991 to be included in the split sample tests.

Figure **4.2** is a histogram summarising the variety of ratios of the median of AMAX flows, QMED, calculated from the earlier and later parts of the record. The categories on the x axis are calculated using a geometric scale, so as to give equal visual prominence to ratios representing decreases and those representing increases. Ratios greater than 1 (that is, bars to the right of the red line) indicate an increase in the median in the later part.

71% of gauges show an increase in QMED. The lowest ratio is 0.73 and the highest 1.52. These opposite extreme changes occur at gauges 29009 (Ancholme at Toft Newton) and 41023 (Lavant at Graylingwell) respectively.

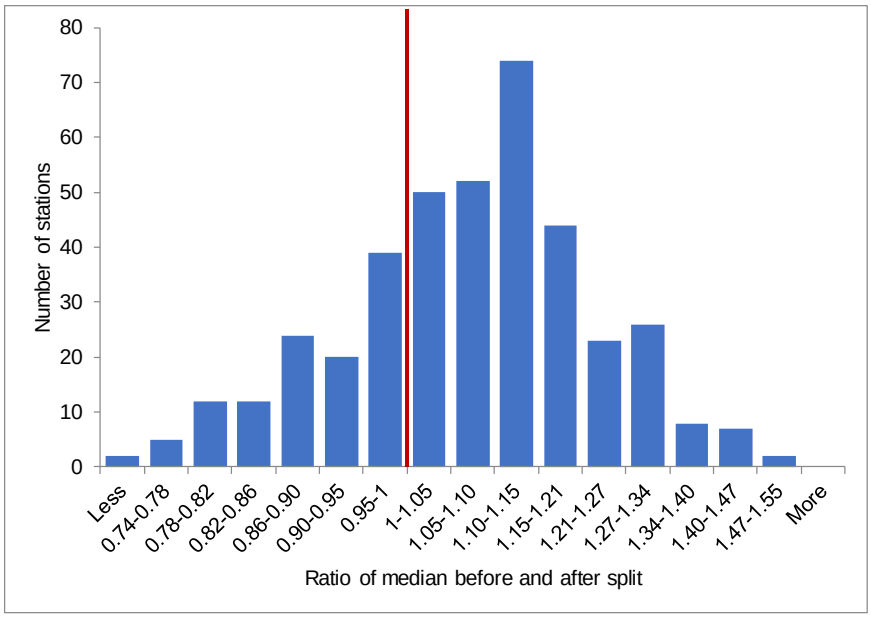


Figure 4.2 Histogram of changes in the median AMAX flow, QMED

At a 5% significance level, the null hypothesis of no significant change in distribution was rejected for 52 gauges, representing 13% of the total with long enough records. Provisionally, an alternative hypothesis can be accepted for these gauges, that there is a difference in the distributions of the populations of AMAX flows represented by the periods before and after 1991. At 47 of these 52 gauges, QMED is higher for the later period.

At a 10% significance level, the null hypothesis was rejected for 23% of gauges.

There is a wide range in the results between Environment Agency Areas. The north-east, the Solent and South Downs, and Cumbria and Lancashire have more than 90% of stations where QMED is higher in the later series. In Cumbria (as defined in section 4.2.1), all but one of the 28 stations have a higher QMED in the later series.

4.3.2 Change in spread of the distribution

Figure **4.3** summarises the ratio of variances of AMAX flows calculated from the earlier and later parts of the record. As for QMED, the variance is higher for the later period at the majority of gauges (70%). The implication is that the general increase in QMED is not just an upward shift in the location of the distribution; rather, the large floods are tending to become larger in the later part of the records.

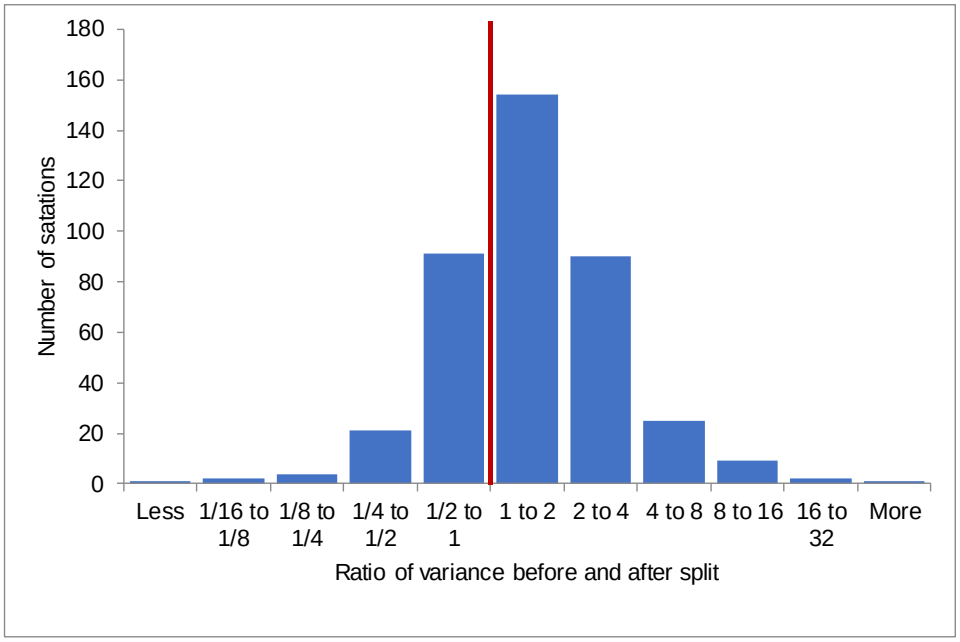


Figure 4.3 Histogram of changes in the variance of AMAX flows

The variance is highly sensitive to the presence of outliers, which is one reason why flood frequency estimation techniques, such as those used in the FEH, avoid using conventional statistical moments such as the variance in fitting flood frequency curves. A single outstanding flood in the earlier or later parts of the record can be enough to change the variance by an order of magnitude or more.

The lowest ratio of variances estimated from the earlier and later parts of the record is 0.04 and the highest is 41. These opposite extreme changes occur at gauges 40012 (Darent at Hawley) and 27023 (Dearne at Barnsley) respectively and can be explained by outliers in 1968 and 2007 respectively.

At a 5% significance level, the null hypothesis of no significant change in variance was rejected for 40 gauges, representing 10% of the total with long enough records. Provisionally, an alternative hypothesis can be accepted for these gauges, that there is a difference between the variances of populations of AMAX flows represented by the periods before and after 1991. At 36 of these 40 gauges, the variance is higher for the later period.

At a 10% significance level, the null hypothesis was rejected for 18% of gauges.

These results complement those produced from the PELT and Pettitt tests and also the results of the multi-temporal trend testing.

4.4 Non-stationary flood frequency analysis

4.4.1 Methods applied

The following methods were applied to estimate a flood frequency curve at each of the 375 stations:

1. **SS-STAT-MLE:** Single-site stationary analysis, fitted using MLE. These results help identify how much of the difference between the FEH results and the non-stationary analysis are due to differences in the methodology other than the assumption of stationarity. Both GLO and GEV distributions were fitted, for comparison.
2. **SS-NONSTAT-MLE:** Single-site non-stationary analysis, using the following approaches. In each case, both a GLO and GEV distribution was fitted.
    a. Using time (the water year) as a covariate. Four model variants were fitted, in which the location, scale, neither and both parameters were allowed to vary with time. The results from the best fitting model, judged by likelihood ratios, were output. In some cases, this is the stationary model, in which case the results are identical to those of (2).
    b. Using physical quantities as covariates, either as well as or instead of time. Up to 2 covariates are allowed per model, with a maximum of one being a physical covariate (the other being water year). The results from the best-fitting model, judged by the BIC, were output. In some cases, this is the stationary model, in which case the results are identical to those of (2). In others, the preferred model has only time as a covariate, in which case the results are identical to those of (3a).

    Note that the criterion is up to one physical covariate per model, not per model parameter. So, models with one physical covariate for the location parameter and a different physical covariate for the scale parameter are not considered.
3. **P-FEH:** Using the pooled FEH method, in which the flood growth curve is estimated from a pooling group composed of gauge sites of similar catchment characteristics, with additional weight given to data at the subject site. Also known as the enhanced single-site approach, this is commonly applied when estimating flood frequency at gauging stations within the FEH methodology. The growth curve is a GLO distribution fitted using L-moment ratios as most commonly applied with the FEH. As with all FEH methods, this is a stationary analysis.

The 7 physical covariates are explained in section 3.3.6.

4.4.2 Results and ways of comparing them

The first type of comparison was to note at which sites a non-stationary model was preferred (as judged by likelihood ratios).

Results, that is, estimates of flow, were output for a range of annual exceedance probabilities (AEPs). The comparisons in this report focus on 3 AEPs: 50%, 10% and 1%. The 10% AEP is expected to be most influential in the economic assessment of flood damages, and the 1% most influential in calculating costs for flood alleviation scheme development.

For the non-stationary models with time as a covariate, results were output for each year of record, and the results for the most recent year were used in the comparison.

For the non-stationary models with physical covariates, integrated flow estimates have been output. Section 4.4.7 shows a discussion of these results.

Maps of the results are provided in Appendix G. Results at each gauge are provided in the digital outputs (see section 4.6).

4.4.3 Extent of non-stationarity across the data set when considering only time as a covariate

At about 80% of gauges, a single-site stationary model (SS-STAT-MLE) was preferred over any of the non-stationary models that use time as a covariate. See

Table 4.2 for the breakdown of results, which differ slightly between the GEV and GLO distributions.

At the gauges where a non-stationary model gave a better fit, the trends could mostly be modelled using either a varying location parameter or a varying scale. Only 2% of the national data set ended up being modelled, with both the location and scale parameters varying.

Table 4.2 Types of flood frequency model selected: time as covariate

Flood frequency model (GEV distribution)	**Number of gauges where model is preferred**	**%**
Stationary (SS-STAT-MLE)	287	76%
SS-NONSTAT-MLE (varying location)	48	13%
SS-NONSTAT-MLE (varying scale)	32	9%
SS-NONSTAT-MLE (varying location and scale)	8	2%
Flood frequency model (GLO distribution)	**Number of gauges where model is preferred**	**%**
Stationary (SS-STAT-MLE)	302	81%
SS-NONSTAT-MLE (varying location)	47	13%
SS-NONSTAT-MLE (varying scale)	18	5%
SS-NONSTAT-MLE (varying location and scale)	8	2%

As mentioned earlier, the best model fit was judged using likelihood ratios. The findings were extremely similar if the BIC statistic was used to select the preferred model. At 97% of gauges it gave the same preference as the likelihood ratio method.

The finding can be compared with that from Faulkner and others (2019), who applied a similar analysis to a large data set of 509 gauges in England, Wales and Scotland, with much less screening than was applied in the present project. A stationary model was found to fit best at two thirds of gauges in Great Britain.

The findings from this analysis can be interpreted as another type of trend test, to complement those carried out elsewhere in this project. Where a non-stationary model is found to fit better, this is evidence of a statistically significant trend. Unlike the Mann-Kendall test, the non-stationary distribution fitting is a parametric trend test, which accounts for the magnitude of the trend rather than only its direction. The results of the different types of test are compared in section 0.

The direction of trend is indicated by the sign of the GEV or GLO model parameters, μ_1, which indicates how the location parameter varies with time, and ϕ_1, which indicates how the logarithm of the scale parameter varies with time. Equations 3 to 6 contain the relevant definitions.

Table 4.3 shows the findings. Trends in the location parameter, where present, are overwhelmingly positive. Trends in the scale parameter are mainly positive, but there are a fair number of gauges with a negative trend, indicating that, at those gauges, the variability of floods appears to be decreasing over time.

Table 4.3 Direction of trends in flood frequency model parameters

Direction of trend in GEV location parameter		**Direction of trend in GEV scale parameter**	
Positive	47	Positive	29
No trend	319	No trend	335
Negative	9	Negative	11
Direction of trend in GLO location parameter		**Direction of trend in GLO scale parameter**	
Positive	45	Positive	17
No trend	320	No trend	349
Negative	10	Negative	9

Figure G-1 **Spatial distribution of preferred model fit**

in Appendix G) shows the geographical distribution of the types of flood frequency model that were found preferable. There are few clear geographical patterns in the results. Stationary models are generally preferred in all areas. Non-stationary models are seen in the north of England, but also in the Home Counties, Wales and the south-west. Large numbers of non-stationary models are found in hydrometric areas 39 (Thames, 10 out of 25 gauges) and 54 (Severn, 9 out of 22 gauges).

There are 23 gauges within Cumbria (as defined in section 4.2.1). Looking at these 23 gauges, the majority (61%) show the stationary model as fitting better than non-stationary models with time as the covariate. The stationary models include all gauges on the River Eden and also on the River Derwent and its tributaries. This finding

implies that although there may be non-stationarity in peak flows across Cumbria, at about 60% of stations the non-stationarity is not strong enough to justify the increase in model complexity associated with adding time as a covariate. (When physical covariates are added, the picture changes, with non-stationary models providing the best fit at 75% of stations).

4.4.4 Comparison of FEH and stationary at-site results

Before comparing the SS-NONSTAT-MLE results with those obtained from the P-FEH method, it is worth examining other ways in which the results might differ, apart from including non-stationarity. There are several other methodological differences, which are inevitable given that non-stationary analysis cannot currently be carried out using the type of methods generally applied within the FEH methodology:

a) The analysis carried out using the methods developed in this project is single-site; the FEH growth curves are derived using pooling (albeit giving extra weight to the at-site data).

b) The non-stationary analysis uses MLE; FEH uses L-moment ratios.

c) The non-stationary analysis does not standardise the annual maximum flow data. The FEH method standardises the data by the median, QMED, fits a dimensionless growth curve and then multiplies it by QMED, which is estimated as the median of the annual maximum flows at the gauge. This is the reason why the 2 approaches can give different results for an AEP of 50%: one method estimates it as the sample median and the other as a point on the probability distribution. Standardisation by an index flood in a non-stationary context is awkward since the index flood may not take a constant value.

d) The analysis fits both the GEV and GLO distributions; the FEH results created for this project use the GLO only.

These variations between the 2 methods may be expected to lead to differences in the results even when non-stationarity is not modelled. Although the project team has not attempted to quantify the relative contributions of all 4 factors listed above, the first one can be expected to be quite significant at some short-record gauges where most of the weight in the FEH pooling is given to other gauges in the pooling group.

Table 4.4 provides some summary statistics of the differences. On average across the whole data set there is very little difference between the results of the 2 methods when applied in a stationary framework. The mean ratios between the 2 sets of design flows are very close to 1. This is encouraging, indicating that there is little bias associated with any of the methodological differences listed above.

At individual gauges, some large differences are seen, and these increase with reducing AEP, as might be expected. Differences are nearly all within ±10% for small floods (50% AEP).

For more extreme floods (1% AEP), there are 13 gauges where the GLO stationary estimate is over twice as high as the FEH estimate. These tend to be locations where outliers are present in the AMAX series. As expected, the influence of the outliers is moderated by the FEH pooled analysis, leading to the large difference between the gradient of the flood frequency curves. Places where the single-site stationary estimates greatly exceed the FEH stationary estimates tend to correspond to areas that have experienced exceptional floods.

Table 4.4 Summary statistics calculated over the full data set: Ratios of SS-STAT-MLE to P-FEH estimate

	AEP (%)	50	10	1
GEV distribution	**Maximum ratio**	1.11	1.26	2.15
	Geometric mean ratio	1.00	1.01	0.95
	Minimum ratio	0.79	0.82	0.60
GLO distribution	**Maximum ratio**	1.10	1.44	2.86
	Geometric mean ratio	1.00	1.02	1.11
	Minimum ratio	0.89	0.82	0.62
	Note: GLO model results from station 37031 have been excluded in calculating these ratios: the model fit at that station is not believable, with standard errors of zero for all parameters.			

It is typical for the GLO distribution to fit a steeper flood frequency curve than the GEV distribution. This effect can be seen in all the tables of results, where for lower AEPs the single-site GLO tends to show higher ratios than the GEV when compared with the pooled results from the FEH (which are all from the GLO).

4.4.5 Comparison of stationary and non-stationary results with time as covariate

One of the important questions that this report addresses is the extent to which non-stationary analysis gives different estimates of design flows compared with conventional methods.

The project team carried out 2 sets of comparisons: one in which the baseline results were from an equivalent stationary model fitted using the same techniques as the non-stationary model (SS-NONSTAT-MLE), and the other in which the baseline results were from the FEH method (P-FEH). In both cases, the results that were compared with these baselines were those from the ‘preferred model’, that is, the one chosen using likelihood ratios. This is stationary at some gauges and non-stationary at others. All non-stationary results are evaluated for the most recent year of record at the gauge; this is 2016 to 2017 at most sites.

Table 4.5 summarises the findings.

Table 4.5 Summary statistics calculated over the full data set: Ratios of flood estimates from preferred model (time covariate only) to estimates from stationary models

		Preferred model compared with SS-STAT-MLE			Preferred model compared with P-FEH		
	AEP (%)	50	10	1	50	10	1
GEV	Maximum ratio	1.65	1.51	1.64	1.74	1.92	2.73
	Geometric mean ratio (all gauges)	1.02	1.02	1.02	1.02	1.03	0.96
	Geometric mean ratio (only gauges with non-stationary model preferred)	1.09	1.10	1.07	1.09	1.11	1.00
	Minimum ratio	0.71	0.62	0.52	0.61	0.63	0.41
GLO	Maximum ratio	1.62	1.80	2.67	1.65	1.99	4.04
	Geometric mean ratio (all gauges)	1.01	1.02	1.03	1.01	1.04	1.14
	Geometric mean ratio (only gauges with non-stationary model preferred)	1.08	1.12	1.15	1.07	1.14	1.28
	Minimum ratio	0.66	0.65	0.60	0.61	0.62	0.55

The table shows 2 sets of mean ratios: one evaluated over all gauges and the second only over gauges at which a non-stationary model was the preferred fit. The means evaluated over the entire data set indicate the overall effect of accounting for non-stationarity in flood frequency analysis. The data set includes many gauges where the preferred model is a stationary model, and so the ratio of model results is 1. This pulls the average close to 1.

The means evaluated only over gauges where non-stationary models are chosen indicate the average effect of moving from a stationary to a non-stationary model at locations where there is enough evidence of non-stationarity to justify preferring that model. On average, this leads to an increase in design flows of about 10% when comparing with the equivalent stationary model, SS-STAT-MLE.

At individual locations, the non-stationary models can give results that differ greatly from the equivalent stationary model, up to a factor of 2.67 higher or a factor of 1.9 lower (1/0.52).

The last 3 columns of

Table 4.5 compare the preferred model results with those from the pooled FEH analysis. Differences between these 2 sets of results are influenced both by including non-stationarity in some models and by the other methodological differences mentioned earlier. Taken together, these lead to some larger differences, up to a factor of 4.0 higher or 2.4 lower (1/0.41).

The frequency distribution of the ratios between the various sets of results is shown in

Figure **4.4**.

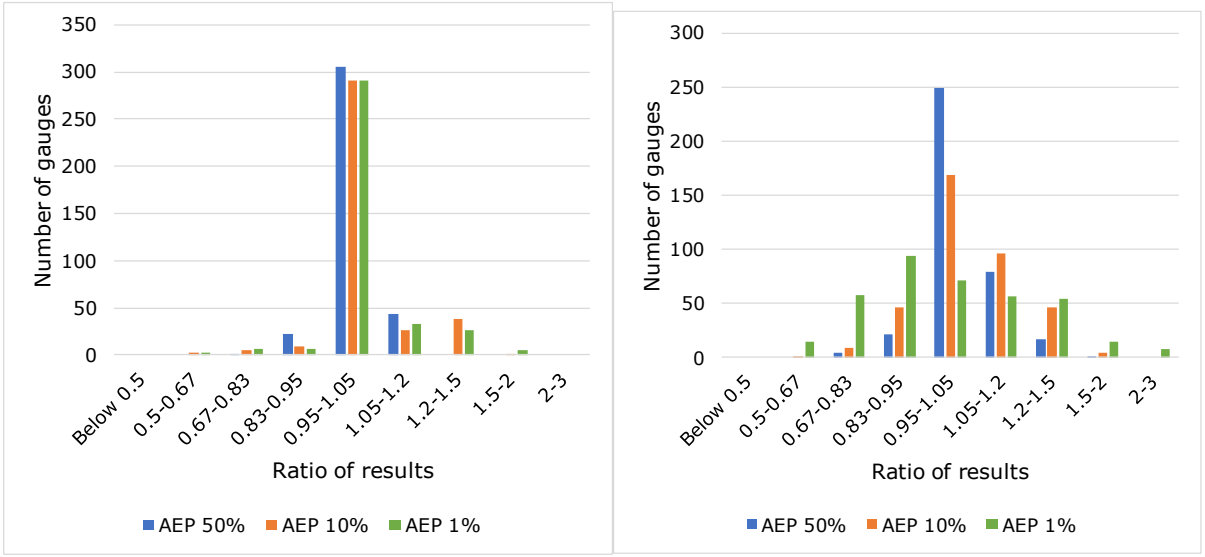


Figure 4.4 Distribution of ratios of results across the data set: (left) comparing the preferred version of the GEV model with an equivalent stationary model (SS-NONSTAT-MLE); (right) comparing the preferred GEV model with the P-FEH results

The left-hand plot on

Figure **4.4** compares results from equivalent models, the only difference being that one may be non-stationary whereas the other is always stationary. That is why there are many gauges where the ratio is 1. Ratios above 1 are more common than those below, that is, the non-stationary analysis, where it is preferred, tends to increase design flows. There are more gauges at which the results for lower AEPs increase when non-stationarity is modelled, compared with those for higher AEPs.

If the non-stationarity is due to climate change, then it is worth noting that at some gauges the estimated design flows for the most recent year of record have increased above their stationary equivalents by more than 20%. This is the upper (90th percentile) climate change factor currently recommended for 6 river basin districts (Environment Agency, 2016a).

The plot on the right compares with the FEH results. There is a wider spread of ratios, particularly for the lower AEPs where the contrast between at-site and pooled analysis becomes more influential.

Figure G-2 to Figure G-4 (in Appendix G) compare the design flows estimated from the preferred model at each gauge with those from the equivalent stationary model, for 3 AEPs. At most gauges, the results are identical because the stationary model is the preferred one. Gauges where the preferred model gives increased flows are widely scattered across England and Wales. The smaller number of gauges that see a decrease are concentrated in East Anglia.

In interpreting the maps, it is worth remembering that the results are plotted where the gauges are located, not catchment centroids. Two adjacent gauges may have contributing catchments with headwaters that are far away from each other. There is a large variation between some results even at some nearby gauges. For example, around London there are gauges showing both an increase and a decrease when the non-stationary and stationary models are compared. Further investigation would be needed to find out the reasons for this, but likely explanations include differences in geology, record lengths and urban influence.

Figure G-5 to Figure G-7 (in Appendix G) compare the design flows estimated from the preferred model at each gauge (whether stationary or non-stationary) with those from the FEH pooled analysis for 3 flood probabilities. The ratios shown on these maps are an important output of the project, because they compare (within the constraints of a national, automated analysis) the current best estimate of design flows using stationary methods with an alternative non-stationary estimate.

The geographical patterns on these maps are the combination of those seen on the previous set of maps and those summarised in section 0. There is little sign of spatial variation in the ratios that is consistent nationally: local factors appear to dominate. Local clusters are evident, such as in Cumbria at the 1% AEP. This finding is consistent with that of the National Flood Resilience Review (HM Government, 2016), that natural variability is expected to dominate underlying trends over the next 10 years.

4.4.6 Preferred physical covariates

Including physical covariates in the non-stationary model adds an extra dimension to comparing results.

The results are summarised as follows:

- numbers of gauges showing different types of model fit as preferable, that is, stationary, physical covariates only, time as covariate only, or both time and physical covariates (Table 4.6 and Figure G-8 in Appendix G)
- which covariates proved to be the most ‘popular’ across the gauge network (
- Table **4.7** and Figure G-9 in Appendix G), that is, those that were selected in the largest numbers of models
-

Table 4.6 Proportions of model types preferred across the data set (GEV distribution)

Stationary: No covariates	**Temporally stationary with physical covariates**	**Non-stationary: time the only covariate**	**Non-stationary: time and physical covariates**
2%	61%	1%	36%

The findings of this assessment raise the question of what is meant by non-stationarity. There are many gauges where the preferred model is one that includes only a physical covariate, with the parameters showing no dependence on time. The findings for these gauges indicate that flood magnitude is statistically associated with the value of the covariate. If the covariate is, for example, winter rainfall, then this may not be surprising

news to most hydrologists, and does not necessarily imply any non-stationarity over time.

In fact, these models, without time as a covariate, should be regarded as temporally stationary. All the physical covariates are detrended by the model fitting routine, even when fitting models that involve only physical covariates. This is statistically advisable since for the model comparison to be valid it is necessary to be fitting to the same input covariates. It would not be necessary to detrend the covariates if fitting models only using physical covariates, although the concept of the integrated flow estimate (see section 4.4.7) would not be valid in that case as it relies on an unchanging distribution of the covariates.

In nearly all cases, the chosen physical covariate is water year rainfall. The project team has not attempted to test the individual catchment-average rainfall series for trend, but that would be an interesting avenue for further investigation.

A model with only water year as a covariate is preferred at only 1% of gauges. This finding indicates that physical covariates are adding useful information in nearly all cases. The increase in model complexity is outweighed by the increase in goodness of fit provided by the physical covariates. At the very few gauges where the parameters of the flood frequency distribution are found to vary with time but not with the physical covariates, one possible explanation is that the non-stationarity is driven by effects not represented by the covariates, such as changes in land use or flood alleviation works. These 4 gauges, widespread across the country (Figure G-8 in Appendix G), are:

- 23017 (Team at Team Valley) - decreasing trend, part-urban catchment
- 37019 (Beam at Bretons Farm) - increasing trend, substantial urban development in lower catchment
- 68018 (Dane at Congleton Park) - increasing trend, largely rural
- 76008 (Irthing at Greenholme) - increasing trend, rural, upland headwaters, afforestation and felling

Further investigation of these exceptional results would be of interest.

When physical covariates are selected, the annual rainfall is the most popular choice (

Table **4.7**). It is included as a covariate for the location parameter at 65% of all gauges, and for the scale parameter at 21% of all gauges (the scale is modelled as fixed, with no covariate, at most gauges). In some cases, annual rainfall is included alongside time as a covariate, and in others it is the only covariate.

The second most useful physical covariate was the winter rainfall, chosen at 22% of gauges for the location parameter and at 4% for the scale parameter.

Table 4.7 Most commonly selected covariates (GEV distribution)

Covariates for the location parameter			**Covariates for the scale parameter**		
Rank	Covariate	% of gauges where chosen	Rank	Covariate	% of gauges where chosen
1	Annual rain	65	1	None (where parameter is fixed)	57
2	Time	28	2	Annual rain	21

3	Winter rain	22	3	Time	15
4	Autumn rain	5	4	Winter rain	4
Note: Percentages in the columns add up to more than 100, because models can include both time and a physical covariate together. More models included covariates for the location parameter than for the scale parameter.					

Figure G-9 (in Appendix G) shows that, while the annual rainfall is widespread as a preferred physical covariate across the country, the winter rainfall tends to be preferred in some parts of southern England. There appears to be some correspondence between these gauging stations and the locations of chalk outcrops, that is, they are concentrated in central southern England and along a band running up into East Anglia. This makes sense physically since flood flows on groundwater-dominated catchments are expected to be strongly linked with the level of the water table, which is determined mainly by the volume of winter recharge. On chalk catchments, rainfall outside the winter recharge season may be lost due to evaporative demands and so have little impact on flood flows. Chalk aquifers, which have much lower specific yield than sandstone for instance, tend to respond more rapidly to recharge over a single winter season.

It is worth noting that the rainfall covariates associated with each annual maximum flow are calculated by accumulating rainfall within the same water year as the annual maximum flow. For annual total rainfall, this means rainfall between 1 October one year and 30 September the following year. It is likely that some of this rainfall will have occurred after the annual maximum flood. Despite this, the annual rainfall appears to be a widely-preferred covariate.

The small number of catchments where autumn rainfall is chosen as a covariate are widely scattered.

4.4.7 Comparison of stationary and non-stationary results with time and/or physical covariates

Results from models that include physical covariates have been calculated in the form of integrated flow estimates. These represent conditions during the whole period of recorded flow data, rather than being associated with one particular point in time.

The integrated flow estimate for a particular AEP can be compared with the stationary estimate, although it is important to take care in interpreting it. For instance, if the integrated flow estimate is 20% higher than the stationary result for a particular AEP, this does not mean that the **current** flood magnitudes for that exceedance probability are expected to be 20% higher. Rather, it means that over the **observed period of record**, the flow with some particular probability of being exceeded is 20% higher than the flow of the same probability if the non-stationarity is ignored.

This probability associated with a period of N years is referred to as the ‘encounter probability’. The AEP is the encounter probability for a period of one year.

The results in

Table **4.8** and Figure 4.5 compare the integrated flow estimates with the stationary SS-STAT-MLE results. The integrated estimates for each station are calculated from the model with lowest BIC. This may be:

- a stationary model, in which the ratio is 1
- a model with only time as a covariate (there are very few), in which case the integrated flow estimate is calculated using the same procedure as for physical covariates. It is representative of the whole period of record rather than any particular year
- a model with only a physical covariate, in which case the integrated estimate is representative of the observed sample of covariate values
- a model with both water year and a physical covariate, in which case the integrated estimate is representative of the observed sample of covariate values and water year values

Table 4.8 Summary statistics calculated over the full data set: Ratios of integrated flow estimates from preferred model (time and/or physical covariates) to estimates from SS-STAT-MLE model

	Preferred model (GEV) compared with SS-STAT-MLE model (GEV)		
AEP (%)	50	10	1
Maximum ratio	1.15	1.31	3.13
Geometric mean ratio (all gauges)	1.00	0.98	1.03
Geometric mean ratio (only gauges with temporally non-stationary model preferred)	1.00	0.99	1.06
Minimum ratio	0.90	0.79	0.43

As for the comparison of results with time as the covariate, there are 2 sets of mean ratios provided in the table: one evaluated over all gauges and the second only over gauges at which the preferred fit was a temporally non-stationary model, that is, one with time as one of the covariates. The means evaluated over the entire data set indicate the overall effect of accounting for non-stationarity in flood frequency analysis. In this case, there is little difference between the 2 sets of means. Both show that, on average, there is little difference between the preferred model and the stationary estimates. In individual cases, there can be a large difference, particularly for low-probability floods, where the integrated flow estimate is, in one case, 3 times larger than the stationary estimate; this is for gauge 47020, Inny at Bealsmill, which is an outlier (see Figure 4.5), the second highest ratio being 1.78.

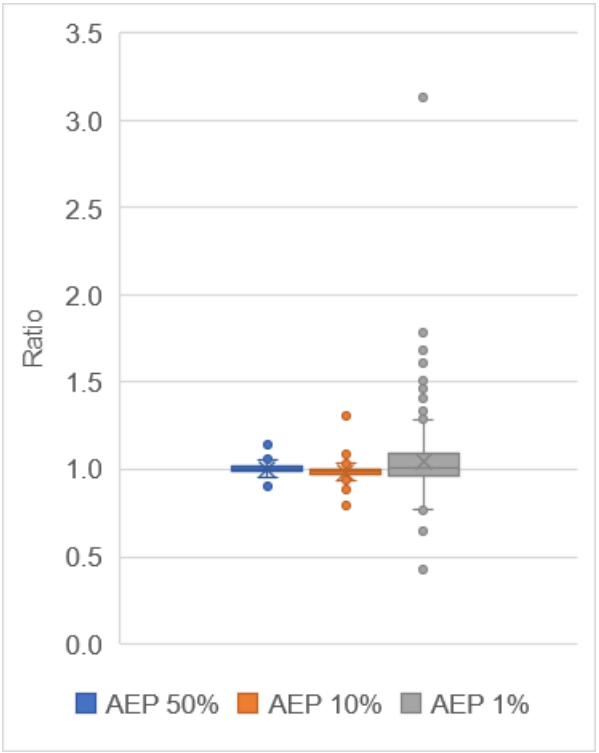


Figure 4.5 Box and whisker plot showing ratios of integrated flow estimates from preferred model (time and/or physical covariates) to estimate from SS-STAT-MLE model

A more meaningful comparison could involve calculating the single-year integrated flow estimate, defined in section 3.3.6. This could be more readily compared with the results of the analysis using time as a covariate. This was not included in the scope of the project.

4.4.8 Concluding remarks on non-stationary flood frequency results

One headline message is that when non-stationarity is modelled only in relation to changes over time, a stationary model is found to fit best at about 80% of gauging stations. At the remaining gauges where a non-stationary model gave a better fit, the trends were mostly modelled using either a varying location (representing the typical size of floods) or a varying scale (representing the variability in flood sizes). Only 2% of the national data set gave the best fit, with both the location and scale parameter varying. The proportion of stations fitted by a non-stationary model increases to 37% when physical covariates are added to the fitting in addition to water year.

Another headline is that on average, across England and Wales, including non-stationarity makes little difference to estimating design flows. In individual cases, it can make a large difference, usually but not always, leading to an increase in present-day estimates.

The big local variations in the findings of this national analysis make it difficult to generalise the results across regions, for example, Environment Agency areas or hydrometric areas. This project has included trials of 2 pooled versions of non-stationary analysis, but neither is currently recommended as ready to be practically applied. In the meantime, it seems wise to consider each case of apparent non-stationarity individually, looking at local circumstances, trying to find physical reasons for the trends, and asking whether nearby gauges support or oppose the hypothesised attribution. For instance, if trends are thought to be caused by increases in rainfall, yet are not apparent on some neighbouring and similar catchments, why is this? The case study in the practitioner guidance on the River Eden uses annual rainfall as a covariate. This decision is supported by the finding that non-stationary models fitted to peak flows on other nearby rivers also find rainfall to be a useful covariate.

Physical covariates are almost always beneficial to the fit of non-stationary flood frequency models, with annual or seasonal rainfall totals proving more beneficial than indices of atmospheric circulation. The annual rainfall is the most commonly chosen covariate. The time-varying model is preferred over those that include physical covariates at only 4 gauges. This finding indicates that the increase in model complexity is nearly always outweighed by the increase in goodness of fit provided by the physical covariates.

4.5 Overall comparison of results

The various analyses have produced a wealth of results that are worth exploring further. As would be expected, there is much consistency between the results of the different statistical tests and model fitting procedures. Where there are differences between the results they are mostly due to:

- the contrast between non-parametric methods, like the MK test, which measure the direction of trends, and parametric methods like the TSA and non-stationary frequency analysis, which also account for the strength of trends
- the fact that some tests detect changes in the magnitude of typical floods (measured by the mean, the median or the location parameter of a fitted distribution) and others are more tailored to detecting changes in the variability of floods (measured by the variance or the scale parameter)
- the different approaches that the analytical methods take for assessing statistical significance, where applicable

Results from the various analyses have been compared, for example as summarised in

Table **4.9**. Comparisons with the non-stationary analysis are limited to the set of 375 gauges classed as suitable for frequency analysis. It was not possible to produce results at every one of these gauges from the MK and TSA tests due to gaps in the record. In addition, the split sample tests could only be applied at gauges with at least 15 years of data both before and after 1991.

Table 4.9 Comparison of national average results from a range of analyses

Preferred version of GEV distribution fit (with time as covariate)	Results from trend tests over full period of record				Geometric means of results from split sample tests: ratio of statistic after split point to statistic before split	
	Mean MKZ	Mean absolute MKZ	Mean TSA (%)	Mean absolute TSA (%)	Ratio of medians	Ratio of variances
Stationary (n=287)	0.40	0.91	6.7	15.8	1.06	1.47
Non-stationary (varying location) (n=48)	1.48	2.17	20.9	34.7	1.14	1.39
Non-stationary (varying scale) (n=32)	0.62	1.08	12.0	20.0	1.09	1.77
Non-stationary (varying location and scale) (n=8)	3.42	3.42	51.0	51.0	1.28	2.28

For simplicity, this cross-comparison only considers non-stationary models that incorporate time as a covariate. Care is needed in interpreting some of the averages in

Table **4.9** because it would, in theory, be possible to obtain an average indicating little overall change from a data set that included many gauges with large positive changes and many with large negative changes. The mean absolute MKZ and TSA statistics avoid this potential pitfall, considering only the degree of change rather than its direction.

As would be expected, the gauges where the stationary model is the preferred fit (as judged by the likelihood ratio) are those that show, on average, the lowest absolute MKZ scores, the lowest absolute TSA scores and the smallest differences between medians from the split sample tests.

The gauges where the varying-location model is preferred show much higher MKZ and TSA scores, and an average difference of +14% between the pre-1991 and post-1991 estimates of the median, QMED. There is a good correlation between both the MKZ and TSA scores and the ratio of change in the median from the split sample tests, as shown in Figure 4.6. The correlation with TSA is stronger.

Also consistent with expectations, the gauges where the varying scale or varying location and scale models are those that show, on average, the largest increase in variance in the split sample tests. At the 8 gauges where the varying location and scale model fits best, the MKZ and TSA scores are particularly high, indicating the most extreme (and all positive) trends. There is little correlation between either MKZ or TSA scores and the ratio of change in the variance from the split sample tests (plots not shown here). This is because the trend tests are not designed to measure changes in the variance.

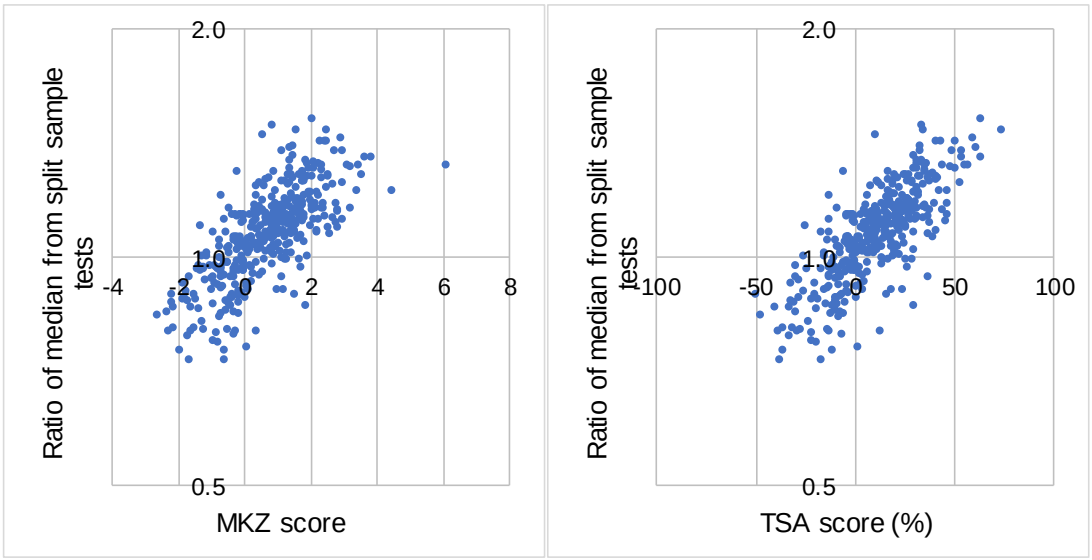


Figure 4.6 Comparison of trend test and split sample test results for change for median

One way to extract a numerical result from the non-stationary frequency analysis is to examine the value of the parameter that controls the rate at which the location of the distribution changes over time (μ_1 in Equation 4) and the equivalent for the scale of the distribution (ϕ_1 in Equation 5). These parameters are compared with the TSA score for each gauge in Figure 4.7. The many points lying on the x axis are for gauges where the best-fitting model did not include any change in the location or scale parameters, respectively for the left and right plots. There is close correlation between the TSA and

the rate of change in the location, but much more scatter in the right-hand plot that shows changes in the scale parameter.

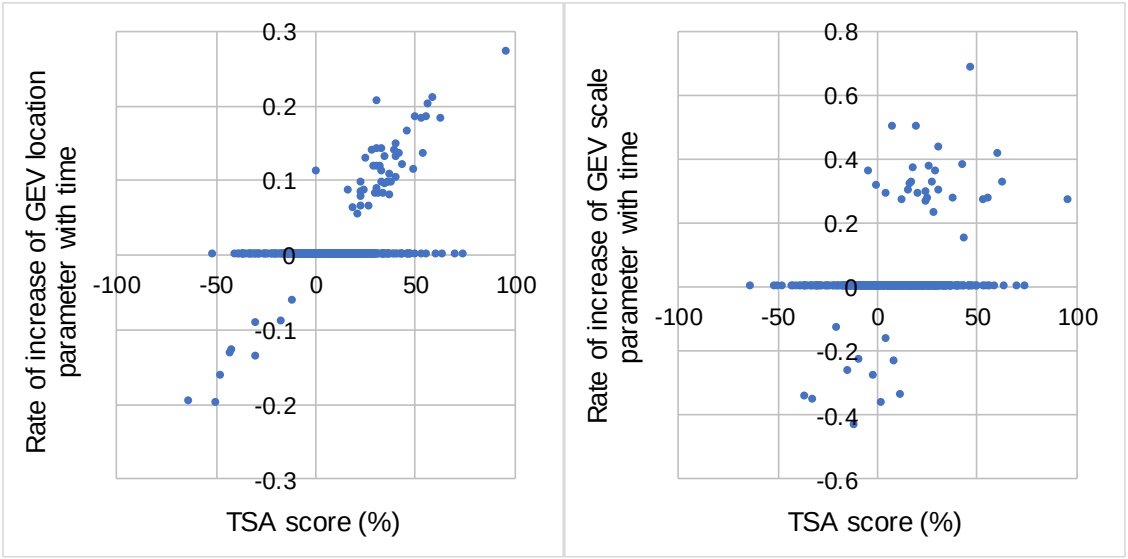


Figure 4.7 Comparison of trend test (TSA score) and results of the non-stationary GEV fitting with time as covariate

Figure 4.7 also shows results from some gauges with large positive TSA scores, indicating a strong upwards trend, and yet no temporal change in the GEV parameters, that is, a stationary model was found to fit best, according to the likelihood ratio method. A brief investigation found no obvious factors that might explain this apparent paradox. Conversely, there are a few sites with apparently little trend as measured by TSA and at which a non-stationary model was fitted.

When non-stationarity is modelled only in relation to changes over time, a non-stationary model is found to be the preferred fit at about 20% of stations. This proportion closely matches the 21% proportion of gauges for which the Mann-Kendal test shows increasing trend significant at a 10% level, over the full period of record. There is not complete overlap between the lists of stations identified in the 2 different analyses, partly because the Mann-Kendal test was applied on a larger data set.

To conclude this section, the project team considered 2 questions:

1) Can non-parametric trend tests be a useful screening measure to decide whether non-stationary flood frequency analysis is worth trying?

   The MK trend test provides only a partial picture of trend. As a non-parametric test, it does not account for the magnitude of the trend, nor can it detect changes in the variance of floods. It is recommended that non-stationary analysis is considered on its own merits rather than only after screening using trend tests. Thanks to the tools developed in this project, non-stationary analysis can now be applied rapidly.

2) Can non-parametric trend tests be a helpful addition in choosing between a stationary or a non-stationary model?

   On balance, it seems that the answer may be yes. Non-parametric analysis has an advantage in terms of robustness compared with parametric methods, and so it does appear to be useful to add this to the decision-making process. For example, if numerical measures like AIC or BIC favour a more complex non-stationary model and yet the MKZ score is not far from zero, this might prompt a

more in-depth review of the results, taking into account some of the other factors suggested in section 3.3.4.

4.6 Digital outputs

The digital results (FRS18087-IG-D2-digital_outputs.zip) that accompany this report comprise:

1) outputs of the change point tests:
    a) a spreadsheet containing the full outputs of the Pettitt and PELT tests at all gauges
    b) two image files per gauge containing plots showing the time series of annual maximum flows and the position of the change point(s) detected, if any
2) outputs of the multi-temporal trend testing:
    a) a spreadsheet giving results at all gauges
    b) a shapefile enabling mapping of the results
    c) a set of image files, one per gauging station, showing the detailed results of the multi-temporal test visually, alongside a time series plot and a table of statistical outputs
3) outputs of the split sample tests:
    a) a spreadsheet containing the full outputs of the Mann-Whitney and Brown-Forsythe tests at all gauges
4) outputs of the non-stationary flood frequency analysis, all as text files in comma-separated format:
    a) with only water year considered as a covariate:
        i) a pair of files containing distribution parameters and measures of goodness of fit for each version of the GEV and GLO models fitted with time as a covariate at all gauges
        ii) a set of files, 2 per gauge, containing flow estimates for each year of record at that gauge for each version of the GEV and GLO models fitted with time as a covariate
        iii) a pair of files containing flow estimates for the final year of record for each version of the GEV and GLO models fitted with time as a covariate at all gauges. This concatenates the final line of all the individual files at 4(a)(ii)
    b) considering both time and physical quantities as covariates:
        i) a set of files, one per gauge, containing measures of goodness of fit and distribution parameters for all 88 versions of the GEV model fitted to different combinations of covariates
        ii) a file containing the covariates chosen and the distribution parameters for the best-fitting version of the GEV model, at all gauges
        iii) a set of files containing flow estimates in the form of stationary, conditional and marginal return levels, along with confidence limits, for the best-fitting version of the GEV model (based on BIC)

iv) a file containing flow estimates, in the form of marginal return levels (without confidence intervals), for the best-fitting version of the GEV model, at all gauges

v) a set of images, one per gauge, containing plots of the best-fitting version of the GEV model, as a time series showing the stationary and conditional return levels

vi) a set of images, one per gauge, containing plots of the best-fitting version of the GEV model, comparing the stationary and marginal return levels in the form of an encounter probability plot

c) various other plots showing diagnostic information for the fitted models

d) various other files containing any error or warning messages generated by the code

5 Conclusions and recommendations

5.1 Conclusions

Some communities in England and Wales have experienced severe flooding many times in the last few years. For example, Calderdale in South Yorkshire was badly flooded in 2012, 2015 and 2020. There is a widespread perception that floods are increasing in these areas and in other parts of the UK. In light of this, there is understandably concern over making decisions about investment in flood protection based an assumption that there has been no change in the probability of flooding.

The perceptions of an increase in flood magnitude and/or frequency are consistent with projections of the impacts of climate change. This project has not attempted to attribute trends. It is quite possible that there is a cyclical element to recent trends. However, it would seem unwise, in the face of a warming climate, to expect trends to reverse in the near future.

Using data up to September 2017, river flow records show general but not universal evidence of this perceived increase. Two thirds of gauges in England and Wales show upward trends in peak flows. Nationally, 13% of gauges show upward trend that is significant at the 5% level, with another 8% showing significance at the 10% level but not the 5% level. Positive trends are seen across much of England and Wales, with some of the strongest and most significant trends in the north and west. Some areas of central and eastern England also display negative trends. The analysis included data up to September 2017. The degree of trend would be expected to increase if the tests were repeated using data that included the extensive and severe floods of winter 2019 to 2020.

This project has made some breakthroughs in the applicability of non-stationary methods of flood frequency estimation, including some innovative techniques for extracting design flood estimates from statistical models that include physical covariates. Non-stationary methods can now be used in practice and can potentially provide more credible answers that can be more easily justified to interested groups. On the other hand, a shift to non-stationary techniques can lead to an increase in uncertainty, as well as a need to choose between numerous analytical techniques (which distribution, which covariates, which method of selecting the preferred model?).

Identifying trends and fitting non-stationary models is complicated due to temporal clustering in flood series. The investigation of clustering detected clusters at all 14 long-record gauges, with a typical cluster duration of 2 to 6 years, and up to 10 years.

Non-stationary fluvial flood frequency estimation remains an active area of research. It is expected that the interim guidance developed in this project will need to be updated to account for scientific developments. A particular challenge, on which this project has made a start, is to develop techniques that can be applied at ungauged locations.

5.2 Recommendations for practitioners

It is recommended that non-stationary analysis is adopted alongside conventional methods, and that the uncertainty of the results from both types of analysis is considered when deciding on a preferred approach.

Specific recommendations for practitioners are provided in the separate Environment Agency guidance document produced as part of this project.

5.3 Recommendations for further research

Despite the wealth of research worldwide into trend analysis and non-stationary flood frequency estimation, there are many areas that are worth investigating further. This includes developing approaches that practitioners in the UK can routinely apply. Other research would examine trend and non-stationarity in other aspects of flood hazards, such as the frequency of floods (using POT data) or the characteristics of extreme rainfall. Other areas of research are relatively minor updates, such as including the flow records from the many record-breaking floods during winter 2019 to 2020.

Table **5.1** summarises the recommendations, including the immediacy of relevance to practitioners, the amount of time and money that might be needed, and a suggested rank to help prioritise further research.

Table 5.1 Summary of recommendations for further research

Recommendation (NS = non-stationarity/non-stationary as appropriate)	Immediacy of relevance to practitioners	Amount of time and resources needed	Suggested rank
1. Develop practical method of pooled NS analysis, including for ungauged catchments.	High	High	High
2. Analyse NS in extreme rainfall and develop NS rainfall frequency estimation.	High	High	High
3. Attribution of trends.	High	High	High
4. Seamless integration of modelling past and future NS.	High	High	High
5. Further analysis of national-scale results.	Moderate	Low	High
6. Update analysis to include floods up to winter 2019 to 2020 (taking due account of stopping rule bias).	High in some areas?	Low	High
7. Further work on clustering, including examining how occurrence of a flood changes probability of another occurring over the short term.	Moderate	Moderate	High
8. Continue investigating spatial statistics to boost signal strength of trends.	Less obvious	Moderate	Medium
9. Adding dates of AMAX flows into NS models.	Less obvious	Moderate	Medium
10. Investigation into use of the Markov Chain Monte Carlo method for fitting models and calculating confidence intervals.	Less obvious	Moderate	Medium
11. Further work on constraining shape parameter of GEV or GLO distributions.	Less obvious	Low	Medium
12. Analyse trends and NS in POT data.	Moderate	High	Lower
13. Analyse NS in flood frequency, duration, spatial extent as well as peak magnitude.	Less obvious	High if national scale	Lower

The recommendations are expanded on below. Other recommendations may emerge from an ongoing piece of work commissioned by the Environment Agency that aims to investigate overseas practice and related research initiatives in applying non-stationary methods.

1. Develop practical method of pooled non-stationary analysis, including for ungauged catchments.

   This would continue the investigations reported in Appendices C and D. It would be desirable to develop either a spatial model of non-stationarity or an equation that would allow a measure of non-stationarity to be estimated from catchment descriptors. Either of these could enable non-stationary analysis at ungauged locations. The investigation would benefit from including physical covariates into pooled analysis, taking a spatially coherent view of model acceptance or rejection in which the same covariate(s) are included at all locations in a group or region.

2. Analyse trends in extreme rainfall and (if necessary) develop non-stationary rainfall frequency estimation.

   Flood studies on some rivers, all reservoirs and all surface water flooding investigations are based on rainfall-run-off models that use rainfall frequency statistics as input. If extreme rainfall is non-stationary, as would be expected from understanding of climate change impacts, then it could be important to account for such non-stationarity in rainfall frequency statistics.

3. Attribution of trends.

   The need for this has been mentioned several times in this report. Essentially, without knowing what is driving trends, there is little chance of knowing how they will evolve in the future. Robust attribution takes much effort. It is necessary to demonstrate that the observed trends are consistent with the proposed cause, that they are inconsistent with alternative causes, and to provide a measure of confidence in the attribution (Merz and others, 2012).

   A first step towards attribution would be to test for trends in extreme rainfall over the same period of record as peak flows (see recommendation 2 above).

   A related suggestion is given in recommendation 4 below, attributing past trends is broadened to cover dynamic modelling of both past and future trends.

4. Seamless integration of modelling past and future non-stationarity.

   This project has taken a statistical approach to modelling past non-stationarity and recommends that future climate change is accounted for using the standard approach based on the output of climate models. A more seamless approach might be desirable, merging the statistical analysis of past floods with the physics-based modelling of future changes. One approach would be to apply dynamic statistical models, which could potentially account for the physical drivers of change, including both climatic effects and catchment land use (for example, Slater and others, 2019).

5. Further analysis of national-scale results.

   This could follow up the work of this project, examining the national results that have already been produced. It could include investigating any relationships between non-stationarity and catchment properties, which would complement the analysis of spatial patterns in the results and could help develop a method for ungauged catchments (see recommendation 1 above).

It would also be interesting to examine the results of the non-stationary models at all gauges, that is, including those where the stationary models were judged to fit best.

6. Update analysis to include floods up to winter 2019 to 2020 (taking due account of stopping rule bias).

   This could be achieved with minimal effort using the R package.

7. Further work on clustering.

   Further work could address questions such as:

   a) How does the degree of clustering vary with flood magnitude, with catchment type or location?

   b) Is clustering primarily due to hydrological or meteorological causes?

   c) What information is there on the typical duration of flood-rich or flood-poor periods?

   d) How much more likely is a flood to occur in the aftermath of another large flood? This question has come up several times during recent sequences of storms, including in February 2020.

   e) How can we quantify the probability of a sequence of floods as opposed to a single event?

   f) Do the rules of thumb for identifying independent POT data need to be improved?

   Any larger-scale investigation of clustering would need to address the major challenge of developing a fit for purpose archive of UK POT data. See also recommendation 12, below.

8. Continue investigating spatial statistics to boost signal strength of trends.

   Appendix D suggests some next steps.

9. Adding dates of annual maximum flows into non-stationary models.

   The dates of flood peaks provide a valuable source of information to complement their magnitudes. Dates can provide evidence of spatial dependence and may help identify physically meaningful covariates, for example, winter floods being related to winter or autumn rainfall. Dates are readily available and generally accurate for most AMAX series. This type of analysis would be a step towards analysis of POT data (see recommendation 12 below) but would avoid the need for a major data-cleaning effort.

10. Investigation into use of the Markov Chain Monte Carlo (MCMC) method for fitting models and calculating confidence intervals.

    A potential problem with using MLE for fitting statistical distributions is that sometimes there can be non-convergence during the optimisation step. This can have a particular impact when using bootstrapping to estimate confidence intervals as this involves a large number of calls to the maximum likelihood estimator function, some of which may not converge. Non-convergence is more likely to happen when including lots of covariates, and seems to occur more frequently when fitting the GLO distribution than when fitting the GEV distribution. For bootstrap estimation to work properly, all of the bootstrap samples need to converge. One solution is to work out good starting points for the optimiser for each bootstrap sample, but this is not practical for lots of

model fits. Another option would be to use the MCMC method for fitting the models and calculate posterior distributions of return levels, and from this, credible intervals. Using MCMC would mean however that there would be a number of extra technical considerations before the results can be accepted.

11. Further work on constraining shape parameter of GEV or GLO distributions.

    As discussed in section 3.3.5, the shape parameter can be poorly estimated for small samples. It would be worthwhile looking into determining a penalty weight for the shape parameter from UK flood peak data.

12. Analyse trends and non-stationarity in peaks over a threshold (POT) data.

    POT data can be expected to provide a more complete picture of flood characteristics than AMAX data. Analysis of POT offers an increased sample size and the opportunity to examine non-stationarity in aspects such as the frequency of floods. Examples of non-stationary analysis using POT are given by Burn and others (2016) in Canada and Eastoe (2019) in the UK.

    Reasons why the analysis focused on AMAX data are given in section 2.1. Any nationwide analysis of POT data would require a major effort to clean up the archive: detecting and infilling gaps, checking for consistency with AMAX series and across different portions of the record, particularly before the start of digital data.

13. Analyse non-stationarity in flood frequency, duration, spatial extent as well as peak magnitude.

    Aspects of this could perhaps follow on from recommendations 8, 9 and/or 12.

One important question that this project has not attempted to address is the impact that adopting non-stationary analysis might have on the Environment Agency's programme of capital spending on flood protection. There are likely to be effects both on calculating scheme costs and scheme benefits. The Environment Agency is planning to investigate this question internally.

References

Akaike H. 1974 ‘A new look at the statistical model identification’ IEEE Transactions on Automatic Control. https://doi.org/10.1109/TAC.1974.1100705 (viewed on 13 May 2020)

Asquith WH. 2018 ‘Lmomco - L-Moments, Censored L-Moments, Trimmed L-Moments, L-Comoments, and Many Distributions’ Lubbock, Texas.: CRAN

Barlow AM, Rohrbeck C, Sharkey P, Shooter R and Simpson ES. 2018 ‘A Bayesian spatio-temporal model for precipitation extremes—STOR team contribution to the EVA2017 challenge’ Extremes, 21(3), 431 to 439

Box GE and Cox DR. 1964 ‘An analysis of transformations’ Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211 to 243

Brady A, Faraway J and Prosdocimi I. 2019 ‘Attribution of long-term changes in peak river flows in Great Britain’ Hydrological Sciences Journal, 64:10, 1159-1170, DOI: 10.1080/02626667.2019.1628964

Brown MB, Forsythe AB. 1974 ‘Robust Tests for the Equality of Variances’ Journal of the American Statistical Association, 69(346), 364 to 367
doi: 10.1080/01621459.1974.10482955

Burn DH, Whitfield PH and Sharif M. 2016 ‘Identification of changes in floods and flood regimes in Canada using a peaks over threshold approach’ Hydrological Processes. DOI: 10.1002/hyp.10861

Burnham KP and Anderson DR. 2004 ‘Multimodel inference: Understanding AIC and BIC in Model Selection’ Sociological Methods & Research, 33, 261 to 304 https://doi.org/10.1177/0049124104268644 (viewed on 13 May 2020)

Chavez-Demoulin V and Davison AC. 2005 ‘Generalized additive modelling of sample extremes’ Applied Statistics 54(1):207 to 222

Coles S. 2001 ‘An Introduction to Statistical Modeling of Extreme Values’ London: Springer https://doi.org/10.1007/978-1-4471-3675-0 (viewed on 13 May 2020)

Coles SG and Tawn JA. 1990 ‘Statistics of coastal flood prevention’ Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences, 332(1627), 457 to 476

Cooley D and Sain SR. 2010 ‘Spatial Hierarchical Modeling of Precipitation Extremes From a Regional Climate Model’ Journal of Agricultural, Biological, and Environmental Statistics, 15(3), 381 to 402

Cooley D, Nychka D and Naveau P. 2007 ‘Bayesian Spatial Modeling of Extreme Precipitation Return Levels’ Journal of the American Statistical Association, 102(479), 824 to 840

Cunderlik, Juraj M and Burn Donald H. 2003 ‘Non-stationary pooled flood frequency analysis’ Journal of Hydrology 276 (1 to 4): 210–23. https://doi.org/10.1016/S0022-1694(03)00062-3 (viewed 13 May 2020)

Cunderlik, Juraj M and Ouarda, Taha BMJ. 2006 ‘Regional flood-duration-frequency modeling in the changing environment’ Journal of Hydrology 318 (1 to 4): 276 to 91 https://doi.org/10.1016/j.jhydrol.2005.06.020 (viewed on 13 May 2020)

Deser C, Hurrell JW, Phillips AS. 2017 ‘The role of the North Atlantic Oscillation in European climate projections’ Climate Dynamics 49:3141 to 3157

Diggle PJ, Tawn JA and Moyeed RA. 1998 ‘Model-based geostatistics’ Journal of the Royal Statistical Society: Series C (Applied Statistics), 47(3), 299 to 350

Eastoe EF. 2019 ‘Nonstationarity in peaks-over-threshold river flows: A regional random effects model’ Environmetrics, e2560

Eastoe EF and Tawn JA. 2010 ‘Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series’ Water Resources Research

Eastoe EF and Tawn JA. 2009 ‘Modelling non-stationary extremes with application to surface level ozone’ Applied Statistics, 58:22 to 45

El Adlouni S, Ouarda TBMJ, Zhang X, Roy R and Bobee B. 2007 ‘Generalized maximum likelihood estimators for the nonstationary generalized extreme value model’ Water Resources Research, 43, W03410

Environment Agency. 2016a ‘Adapting to Climate Change: Advice for Flood and Coastal Erosion Risk Management Authorities’ http://www.gov.uk/government/publications/adapting-to-climate-change-for-risk-management-authorities (viewed 13 May 2020)

Environment Agency. 2016b ‘Flood risk assessments: climate change allowances’ http://www.gov.uk/guidance/flood-risk-assessments-climate-change-allowances (viewed 13 May 2020)

Faulkner D, Warren S, Spencer P and Sharkey P. 2019 ‘Can we still predict the future from the past? Implementing non-stationary flood frequency analysis in the UK.’ Journal of Flood Risk Management https://doi.org/10.1111/jfr3.12582

Fawcett L and Walshaw D. 2006 ‘A hierarchical model for extreme wind speeds’ Journal of the Royal Statistical Society: Series C (Applied Statistics), 55(5), 631 to 646

Francois B, Schlef K, Wi S and Brown C. 2019 ‘Design Considerations for Riverine Floods in a Changing Climate – A Review’ Journal of Hydrology. 574. 557 to 573. 10.1016/j.jhydrol.2019.04.068

Franks S, White C and Gensen M. 2015 ‘Estimating extreme flood events – assumptions, uncertainty and error’ Proceedings of the International Association of Hydrological Sciences, 369, 31 to 36

Friedman Jerome Hastie Trevor and Tibshirani Robert. 2010 ‘Regularization Paths for Generalized Linear Models via Coordinate Descent’ Journal of Statistical Software 33 (1): 1 to 22. https://doi.org/10.18637/jss.v033.i01 (viewed 13 May 2020)

Gabda D and Tawn J. ‘Submitted to Extremes. Marginal extreme value inference from small sample sizes in environmental contexts’

Gilleland E and Katz RW. 2016 ‘extRemes 2.0: An Extreme Value Analysis Package in R’ Journal of Statistical Software, 72(8), 1 to 39

Griffin Adam, Vesuviano Gianni, Stewart Lisa. 2019 'Have trends changed over time? A study of UK peak flow data and sensitivity to observation period' Natural Hazards and Earth System Sciences, 19 (10). 2,157 to 2,167

Gu, X and others. 2016 'Temporal clustering of floods and impacts of climate indices in the Tarim River basin, China' Global and Planetary Change

Hanel Martin, Buishand T Adri and Ferro Christopher AT. 2009 'A nonstationary index flood model for precipitation extremes in transient regional climate model simulations' Journal of Geophysical Research Atmospheres 114 (15): 1 to 16 https://doi.org/10.1029/2009JD011712

Hannaford J and Buys G. 2012 'Trends in seasonal river flow regimes in the UK' Journal of Hydrology, 475, 158 to 174

Hannaford J. 2015 'Climate-driven changes in UK river flows: A review of the evidence' Progress in Physical Geography. 39, 29 to 48

Hannaford J, Buys G, Stahl K, Tallaksen LM. 2013 'The influence of decadal-scale variability on trends in European streamflow records' Hydrology and Earth Systems Sciences 17, 2,717 to 2733

Harrigan S, Hannaford J, Muchan K and Marsh TJ. 2018 'Designation and trend analysis of the updated UK Benchmark Network of river flow stations: the UKBN2 dataset' Hydrology Research. 49 (2), 552 to 567

Harrigan S, Murphy C, Hall J, Wilby RL and Sweeney J. 2014 'Attribution of detected changes in streamflow using multiple working hypotheses' Hydrology Earth System Sciences, 18(5), 1,935 to 1,952

HM Government. 2016 'National Flood Resilience Review'

Hosking JRM and Wallis JR. 2005 'Regional frequency analysis: an approach based on L-moments' Cambridge University Press

Institute of Hydrology. 1999 'Flood Estimation Handbook' Wallingford: Institute of Hydrology

Jonathan P, Ewans K and Randell D. 2014 'Non-stationary conditional extremes of northern North Sea storm characteristics' Environmetrics. 125:172 to 188

Jones David A. 2013 'On an extension of the L-moment approach to modelling distributions which include trend' Hydrology Research 44 (4): 571. https://doi.org/10.2166/nh.2012.081 (viewed 13 May 2020)

Kay AL, Crooks S, Davies HN, Prudhomme C, Reynard NS. 2011 'Practicalities for implementing regionalised allowances for climate change on flood flows' Report to Department for Environment, Food and Rural Affairs, Technical Report FD2648, CEH Wallingford, May 2011, 209pp

Kay AL, Crooks SM, Davies HN, Prudhomme C, Reynard NS. 2014 'Probabilistic impacts of climate change on flood frequency using response surfaces I: England and Wales' Regional Environmental Change, 14(3), 1,215 to 1,227

Kay AL, Rudd AC, Fry M, Nash G. 2019 'Climate change and fluvial flood peaks' Report to Environment Agency/Scottish Environment Protection Agency, SC150009 WP2 Final Report, CEH Wallingford

Killick R, Eckley IA. 2014 ‘changepoint: An R package for Changepoint Analysis’ Journal of Statistical Software 58(3) 1 to 19

Killick R, Fearnhead P, Eckley IA. 2012 ‘Optimal Detection of Changepoints With a Linear Computational Cost’ Journal of the American Statistical Association 107(500) 1,590 to 1598

Kjeldsen TR, Ahn H and Prosdocimi I. 2017 ‘On the use of a four-parameter kappa distribution in regional frequency analysis’ Hydrological Sciences Journal, 62(9), 1,354 to 1363

Kjeldsen TR, Jones DA, Bayliss AC. 2008 ‘Improving the FEH statistical procedures for flood frequency estimation: Science Report: SC050050’ Environment Agency, Bristol

Kjeldsen Thomas R and Jones David A. 2009 ‘A formal statistical model for pooled analysis of extreme floods’ Hydrology Research 40 (5): 465 to 480 https://doi.org/10.2166/nh.2009.055 (viewed 13 May 2020)

Komsta Lukasz. 2019 ‘Mblm: Median-Based Linear Models. Version 0.12.1’ CRAN. https://cran.r-project.org/package=mblm (viewed 13 May 2020)

Lavers DA, Allan RP, Wood EF, Villarini G, Brayshaw DJ and Wade AJ. 2011 ‘Winter floods in Britain are connected to atmospheric rivers’ Geophysical Research Letters, 38, L23803, doi:10.1029/ 2011GL049783

Liu J and Zhang Y. 2017 ‘Multi-temporal clustering of continental floods and associated atmospheric circulations’ Journal of Hydrology

Liu J and others. 2017 ‘Nonstationarity and clustering of flood characteristics and relations with the climate indices in the Poyang Lake basin, China’ Hydrological Sciences Journal

Luke A and others. 2017 ‘Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States’ Water Resources Research, DOI 10.1002/2016WR019676

Lumley Thomas. 2017 ‘Leaps: Regression Subset Selection’ CRAN. https://cran.r-project.org/package=leaps (viewed 13 May 2020)

Macdonald N and Sangster H. 2017 ‘High-magnitude flooding across Britain since AD 1750’ Hydrology and Earth System Sciences, 21(3), 1,631 to 1,650

Mallakpour and others. 2017 ‘On the use of Cox regression to examine the temporal clustering of flooding and heavy precipitation across the central United States’ Global and Planetary Change

Martins ES and Stedinger JR. 2000 ‘Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data’ Water Resources Research, 36(3), 737 to 744

Merz B, Vorogushyn S, Uhlemann-Elmer S, Delgado J and Hundecha Y. 2012 ‘More efforts and scientific rigour are needed to attribute trends in flood time series’ Hydrology and Earth System Sciences, 16, 1,379 to 1,387

Merz and others. 2016 ‘Temporal clustering of floods in Germany: Do flood-rich and flood-poor periods exist?’ Journal of Hydrology

Met Office Hadley Centre. 2018 'UKCP18 Probabilistic Projections by UK River Basins for 1961-2100' Centre for Environmental Data Analysis, April 2019. http://catalogue.ceda.ac.uk/uuid/10538cf7a8d84e5e872883ea09a674f3 (viewed 13 May 2020)

Moore David S and Marcus C Spruill. 1975 'Unified Large-Sample Theory of General Chi-Squared Statistics for Tests of Fit' The Annals of Statistics 3 (3): 599 to 616 https://www.jstor.org/stable/2958431 (viewed 13 May 2020)

Murphy JM, Sexton DMH, Jenkins GJ, Booth BBB, Brown CC, Clark RT, Collins M, Harris GR, Kendon EJ, Betts RA, Brown SJ, Humphrey KA, McCarthy MP, McDonald, RE, Stephens A, Wallace C, Warren R, Wilby R, Wood RA. 2009 'UK Climate Projections Science Report: Climate Change Projections' Met Office Hadley Centre, Exeter, UK

Nam W, Kim S, Kim H, Joo K and Heo J.-H. 2015 'The evaluation of regional frequency analyses methods for nonstationary data' PIAHS 371: 9 to 98 https://doi.org/10.5194/piahs-371-95-2015 (viewed 13 May 2020)

Nicholls N. 2001 'Commentary and analysis: The insignificance of significance testing' Bulletin of the American Meteorological Society, 82(5), 981 to 986

O'Brien NL and Burn DH. 2014 'A nonstationary index-flood technique for estimating extreme quantiles for annual maximum streamflow' Journal of Hydrology, 519, 2,040 to 2,048

Prosdocimi I, Kjeldsen TR and Svensson C. 2014 'Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK' Natural Hazards and Earth System Sciences,14, 1,125 to 1,144

Prosdocimi I, Dupont E, Augustin N Kjeldsen T, Simpson D and Smith T. 2019 'Spatially consistent trend detection in peak river flow across Britain: an approach based on areal models' Technical report

Prosdocimi I, Kjeldsen TR and Miller JD. 2015 'Detection and attribution of urbanization effect on flood extremes using nonstationary flood-frequency models' Water Resources Research, 51, 4,244 to 4,262

Prosdocimi Ilaria, Dupont Emiko, Augustin Nicole H, Kjeldsen Thomas R, Simpson Dan P, Smith Theresa R. 2019 'Areal models for spatially coherent trend detection: the case of British peak river flows' Geophysical Research Letters, 46, 13,054 to 13,061

R Core Team. 2016 'R: A Language and Environment for Statistical Computing' Vienna. https://www.r-project.org/ (Viewed 14 May 2020)

Renard B, Kochanek K, Lang M, Garavaglia F, Paquet E, Neppel L and Borchi F. 2013 'Data-based comparison of frequency analysis methods: A general framework' Water Resources Research, 49(2), 825 to 843

Reynard NS, Kay AL, Anderson M, Donovan B, Duckworth C. 2017 'The evolution of climate change guidance for flood risk management' Progress in Physical Geography, 41(2), 222 to 237, doi:10.1177/0309133317702566

Rougé C, Ge Y and Cai X. 2013 'Detecting gradual and abrupt changes in hydrological records' Advances in Water Resources, 53, 33 to 44

Salas Jose D and Obeysekera Jayantha. 2014 ‘Revisiting the Concepts of Return Period and Risk for Nonstationary Hydrologic Extreme Events’ Journal of Hydrologic Engineering 19 (3): 554 to 68. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000820 (viewed 14 May 2020)

Sang H and Gelfand AE. 2009 ‘Hierarchical modeling for extreme values observed over space and time’ Environmental and Ecological Statistics, 16(3), 407 to 426

Sen Pranab Kumar. 1968 ‘Estimates of the Regression Coefficient Based on Kendall’s Tau’ Journal of the American Statistical Association 63 (324): 1379 to 1389 https://doi.org/10.1080/01621459.1968.10480934 (viewed 14 May 2020)

Serinaldi F, Kilsby CG and Lombardo F. 2018 ‘Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology’ Advances in Water Resources, 111, 132 to 155

Sharkey P and Winter HC. 2019 ‘A Bayesian spatial hierarchical model for extreme precipitation in Great Britain’ Environmetrics, 30(1), e2529

Slater L and Villarini G. 2017 ‘On the impact of gaps on trend detection in extreme streamflow time series’ International Journal of Climatology, 37(10), 3,976 to 3,983

Slater LJ, Villarini G, Bradley A. 2019 ‘Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA’ Climate Dynamics, 53(12), pp. 7,381 to 7,396. doi: 10.1007/s00382-016-3286-1

Southworth H, Heffernan JE, Metcalfe PD, Papastathopoulos Y, Stephenson A and Coles S. 2020 ‘texmex R package’ accessible from https://github.com/harrysouthworth/texmex

Spencer P, Faulkner D, Perkins I, Lindsay D, Dixon G, Parkes M, James R. 2018 ‘The floods of December 2015 in northern England: description of the events and possible implications for flood hydrology in the UK’ Hydrology Research, 49,568 to 596 https://doi.org/10.2166/nh.2017.092 (viewed 27 May 2020)

Šraj M, Viglione A, Parajka J and Blöschl G. 2016 ‘The influence of non-stationarity in extreme hydrological events on flood frequency estimation’ Journal of Hydrology and Hydromechanics, 64 (4), 426 to 437

Stahl K, Tallaksen LM, Hannaford J and van Lanen HAJ. 2012 ‘Filling the white space on maps of European runoff trends: estimates from a multi-model ensemble’ Hydrology and Earth System Sciences, 16(7), 2,035 to 2,047

Steirou E, Gerlitz L, Apel H, Sun X and Merz B. 2019 ‘Climate influences on flood probabilities across Europe’ Hydrology and Earth System Sciences, 23, 1,305 to 1,322, https://doi.org/10.5194/hess-23-1305-2019 (viewed 27 May 2020)

Tanguy M, Dixon H, Prosdocimi I, Morris DG and Keller VDJ. 2016 ‘Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2015) [CEH-GEAR]’ NERC Environmental Information Data Centre. https://doi.org/10.5285/33604ea0-c238-4488-813d-0ad9ab7c51ca (viewed 27 May 2020)

Tibshirani Robert. 1996 ‘Regression Shrinkage and Selection Via the Lasso’ Journal of the Royal Statistical Society: Series B (Methodological) 58 (1): 267 to 88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (viewed 27 May 2020)

Towe R, Tawn J, Eastoe E and Lamb R. 2019 ‘Modelling the Clustering of Extreme Events for Short-Term Risk Assessment’ Journal of Agricultural, Biological and Environmental Statistics’ https://doi.org/10.1007/s13253-019-00376-0 (viewed 27 May 2020)

Varadhan Ravi and Gilbert Paul. 2015 ‘BB : An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function’ Journal of Statistical Software 32 (4): 1 to 26 https://doi.org/10.18637/jss.v032.i04 (viewed 27 May 2020)

Villarini and others. 2012 ‘On the temporal clustering of US floods and its relationship to climate teleconnection patterns’ International Journal of Climatology

Wallingford HydroSolutions. 2019 ‘WINFAP 4’ Wallingford: Wallingford HydroSolutions Ltd

Wang Y and So MK. 2016 ‘A Bayesian hierarchical model for spatial extremes with multiple durations’ Computational Statistics & Data Analysis, 95, 39 to 56

Wilby RL and Quinn NW. 2013 ‘Reconstructing multi-decadal variations in fluvial flood risk using atmospheric circulation patterns’ Journal of Hydrology, 487, 109 to 121

Wine ML, Davison JH. 2019 ‘Untangling global change impacts on hydrological processes: Resisting climatization’ Hydrological Processes, doi:10.1002/hyp.13483

Woollings T and Blackburn M. 2012 ‘The North Atlantic Jet Stream under Climate Change and Its Relation to the NAO and EA Patterns’ Journal of Climate, 25 (3). pp. 886 to 902

Wu and Lye. 1994 ‘Identification of temporal scaling behaviour of flood: A study of fractals’ Fractals

Xavier Ana Carolina Freitas, Blain Gabriel Constantino, Morais Marcos Vinicius Bueno de and Sobierajski Graciela da Rocha. 2019 ‘Selecting “the best” nonstationary Generalized Extreme Value (GEV) distribution: on the influence of different numbers of GEV-models’ Bragantia, 78(4), 606-621. Epub December 13, 2019 https://doi.org/10.1590/1678-4499.20180408 (viewed 27 May 2020)

Yan L, Xiong L, Guo S, Xu C, Xia J, Du T. 2017 ‘Comparison of four nonstationary hydrologic design methods for changing environment’ Journal of Hydrology 551, 132 to 150, doi:10.1016/j.hydrol.2017.06.001

List of abbreviations

AEP	Annual exceedance probability.
AIC	Akaike information criterion. A measure of the quality of a statistical model, which establishes a trade-off between the goodness of fit and the simplicity of the model.
AMAX	Annual maximum (for example, the highest river flow in a water year).
BIC	Bayesian information criterion. A measure of the quality of a statistical model, which establishes a trade-off between the goodness of fit and simplicity.
EA	East Atlantic pattern: an index of atmospheric variability, like a southwards shifted version of the NAO.
FEH	Flood Estimation Handbook.
GEV	Generalised extreme value: a statistical distribution fitted to extremes such as floods.
GLO	Generalised logistic: another statistical distribution.
GMLE	Generalised maximum likelihood estimation.
LOESS	Local polynomial regression fitting
MK	Mann-Kendal, a non-parametric test for trend.
MKZ	Mann-Kendall Z score.
MLE	Maximum likelihood estimation. A way of fitting a statistical model by maximising something known as the likelihood function.
NAO	North Atlantic Oscillation: an index of the north-south difference in air pressure between the north and central Atlantic Ocean, associated with changes in the direction and strength of the jet stream.
NRFA	National River Flow Archive.
PELT	Pruned exact linear time: a test for a sudden step change in a time series.
POT	Peaks over a threshold.
P-P	Probability-probability plot.
QMED	Median annual maximum flood.
Q-Q	Quantile-quantile plot.
REA	Rapid evidence assessment.
SDM	Similarity distance metric.
TSA	Theil-Sen approach to determining strength of a trend.
UKBN	United kingdom Benchmark Network
UKCP09	UK Climate Projections 2009.
UKCP18	UK Climate Projections 2018.

Appendix A: Multi-temporal trend tests

Introduction

The objective of this part of the project was to quantify trends (specifically in annual maximum peak flow) in flooding in England and Wales. This was to support the wider aim of developing approaches and guidance for flood frequency estimation in the presence of non-stationarity in observed flood records.

There is a large body of previous work on non-stationarity in flooding in the UK. Hannaford (2015) reviewed the literature up to the early 2010s in detail and this is a useful starting point for appraising past work. Since then, a number of other studies have also quantified trends in flood series, including Prosdocimi and others (2014), Harrigan and others (2018), Spencer and others (2018), Brady and others (2019), Griffin and others (2019), Prosdocimi and others (2019) and Faulkner and others (2020).

The distinctive feature of the present study is that it provides a site-by-site analysis of non-stationarity for a large number (>480) of gauging stations across England and Wales, a significant proportion of the NRFA Peak Flows Dataset, using more up-to-date peak flow data than any previous study.

The main purpose of the study was to characterise monotonic (that is, changing in the same direction, either increasing or decreasing) trends over the period of record, in other words, the full operational period, for each gauging station. However, another defining feature of this study is the use of a 'multi-temporal' analysis to examine the sensitivity to changing timeframes for the analysis. This follows the recommendation of Hannaford and others (2013) who argue that trends in any fixed period need to be put into a longer term context, given the confounding role of decadal-scale hydrological variability that hampers the interpretation of linear trends. These authors, along with many others, advocate a multi-temporal approach, where trends are evaluated for all possible study periods, that is, varying the start and end year of the analysis and looking at the sensitivity of the results to such changes.

This appendix contains a brief description of the methodology used for the analysis, and basic results are presented and discussed. Accompanying this report is a series of analytical outputs, including a spreadsheet containing trend results for all featured stations, and a graphical output containing a range of results plots, one for each featured gauging station. An annex describes these outputs and provides a sample results page, with information about layout and interpretation.

Methodology

The methodology followed in this work is based on the standard National River Flow Archive (NRFA) trend testing toolkit as described by Harrigan and others (2018), who examined trend characteristics for the near-natural catchments of the UK Benchmark Network version 2 (UKBN2). In this section, the project team describes the data set, methods for trend assessment and the multi-temporal approach.

Data set

The data set used was the agreed project data set (finalised on 17 Feb 2020), consisting of 471 stations. This was the data set agreed on after review by the

Environment Agency, the UK Centre for Ecology & Hydrology (CEH) and JBA Consulting.

The NRFA Peak Flow Dataset (v7, released October 2018) was the default data set used, as being the most up to date available at the time analysis began (spring to summer 2019). The Environment Agency provided new time series for some stations in accordance with its ongoing reviews of peak flow time series.

Changes were made to the data set at numerous points, with the final analysis runs eventually carried out in February 2020. While a new Peak Flows Dataset (v8, Sept 2019) was available at this time, the V7 data were retained as the source data, for consistency with other analysis carried out within the wider project. A spreadsheet of the data set used and audit trail of the various changes is available on request.

Before analysis, the annual maximum flow (AMAX) time series for the stations of interest (whether from NRFA version 7 data set, or provided by the Environment Agency) underwent a missing-value criterion analysis. Missing data can result in misleading outcomes of trend analysis, so it is important to set some sensible criteria to include in the study. Ideally, visual inspection and infilling would be applied to minimise gaps, but this was not possible when applying analysis to such a large data set given available resources. An automated approach was therefore needed.

In common with Harrigan and others (2018), 30 years is considered the minimum time series required for trend analysis, and 10% as an acceptable criterion for missing data. Periods with less than 27 years non-missing data or with missing values totalling more than 10% of the expected length of the data were excluded from any of the analyses described in the following section. For longer records, for the multi-temporal analysis, periods within the record that do not meet this criteria were also excluded.

Primarily, the missing data criteria were selected for consistency with past NRFA studies (Harrigan and others, (2018)) rather than any particular standard. Any automated procedure involves arbitrary decisions to some extent, and there is no readily available standard or guidance on missing data criteria. Slater and Villarini (2016) conducted a simulation study to examine the impacts of gaps on trend detection. The project team could not directly apply their guidance as it requires assumptions to be made about acceptable test accuracy and trend strength (that theoretically could not be known), but in general the 10% criteria for a 30+ year record is reasonable given their results. They acknowledge also that the location of gaps is important, with gaps in the middle of series having less impact than gaps at the start and end. The project team's criteria ensure that long gaps are unlikely at start or end points of relatively short records.

Trend significance

The trend analysis method used is the Mann-Kendall test, a very widely used method for monotonic trend testing, which has been applied extensively in hydrological change applications in the UK and elsewhere (for example, as recommended by World Meteorological Organisation Guidance: Kundzewicz and Robson, 2000). The method is not described in detail here as it is outlined in these reports and various standard statistical textbooks. The equations are succinctly presented in the 'R' statistical package for Mann Kendall trend detection which was used in the present study:

https://cran.r-project.org/web/packages/trend/vignettes/trend.pdf

The Mann-Kendall test is a test for statistical significance of trends, that is, whether an observed trend is significantly different from zero (a null hypothesis of 'no trend'). Here, 2 significance thresholds are used: 5% and 10%.

For identifying whether the results are statistically significant at 5% (or 10%) significance level, a two-tailed MK test was chosen, meaning that, for 5% significance, at $|MKZ| > 1.96$ (for 10%, > 1.645) the null hypothesis of no-trend is rejected.

As the MK test requires serial independence of data, the AMAX time series were first analysed for significant lag-1 serial correlation using the autocorrelation function (ACF). Because of the potential for confounding trend and serial correlation, the serial correlation test was applied to detrended series, specifically to the residuals from the Theil-Sen trend line (see below) for estimating trend magnitude.

For instances with significant lag-1 serial correlation, block bootstrapping (BBS) was applied. More specifically, following Harrigan and others (2018), a block length of 4 time steps was selected to preserve any short-term autocorrelation structure, and 1,000 resamples were generated and their MKZ values calculated. For 5% significance, if the MKZ value of the original data set for the selected temporal coverage lay outside the 25th and 975th ranked values of the MKZ calculated from the resamples, this was interpreted as indicating a statistically significant trend in the original data set.

Trend magnitude

As the focus here is on the direction and strength of changes and not entirely on statistical significance relative to arbitrary p-value thresholds (Nicholls 2001), trend magnitudes were also estimated in order to corroborate and map the strength and regional coherence of trends.

The magnitude (strength or steepness) of trends was quantified by the Theil-Sen approach (TSA). The Theil-Sen (sometimes referred to as Kendall-Theil) robust line is widely used for quantifying trend magnitude, and is similar to the gradient of a least-squares linear regression line (Beta, β), but is preferred due to being less sensitive to the presence of outliers (for example, Stahl and others, 2012).

For a data set $(t_i, Q_i : i = 1, \dots N)$ with all different values of Qi, the Theil-Sen estimator of the slope of $Q = (Q_1, \dots, Q_N)$ is given by:

$$TSA = \ median\left\{\left(\frac{Q_j - Q_i}{t_j - t_i}\right) : i \neq j = 1, \dots, N\right\}$$

In p, the TSA is the median of all pairwise slopes between all points with different times.

To make a comparison between sites, the trend magnitude $\mathrm{TSA_{rel}}$ (%) for each time series was expressed as a percentage of the long-term mean flow μ over the period of record of n years where β is the TSA slope, given by Stahl and others (2012) as:

$$\mathrm{TSA_{rel}} = \left(\frac{\beta \times n}{\mu}\right) \times 100$$

Hannaford and Buys (2012) found this approach preferable compared to expressing trend magnitude as a simple percentage change over the full record, which can yield larger changes in the presence of exceptionally large start or end values.

Study periods and the multi-temporal approach

Three fixed study periods were selected to compare results from all stations and assess the spatial variability of the trends in England and Wales. These periods are: short (starting in 1987), long (starting in 1967) and full. The latter means that for each station the results are derived by applying trend analysis to all the available data (period of record). For the short and long period, the trend statistics are derived by having a fixed start year, and using all the data from that point up to the last available data for each station. For full consistency for spatial comparisons, the project team also used a fixed start and end year (that is, 1987 to 2016 and 1967 to 2016). In this case, all stations have an identical record length and so can be compared fairly.

The periods were selected by Harrigan and others (2018) after an appraisal of record lengths on the NRFA, to reflect the trade-off between spatial coverage (optimised in the 'short' period) and record length (optimised in the latter).

The project team also carried out a multi-temporal analysis for all stations. Harrigan and others (2018) carried out a limited multi-temporal analysis in high flows (rather than AMAX) data for the UKBN2 by varying the start date of UK river flow trends. Here, trends are analysed in all possible start/end points, and visualised using heatmaps – as used in the European-scale study of Hannaford and others (2013). Figure A-1 shows an example.

For all non-rejected periods, the non-parametric Mann-Kendall (MK) test was implemented. The results of this test are standardised (Mann-Kendall Z score, MKZ) in order to compare the different periods of interest and stations. Positive values indicate increasing trends, while negatives ones refer to decreasing trends.

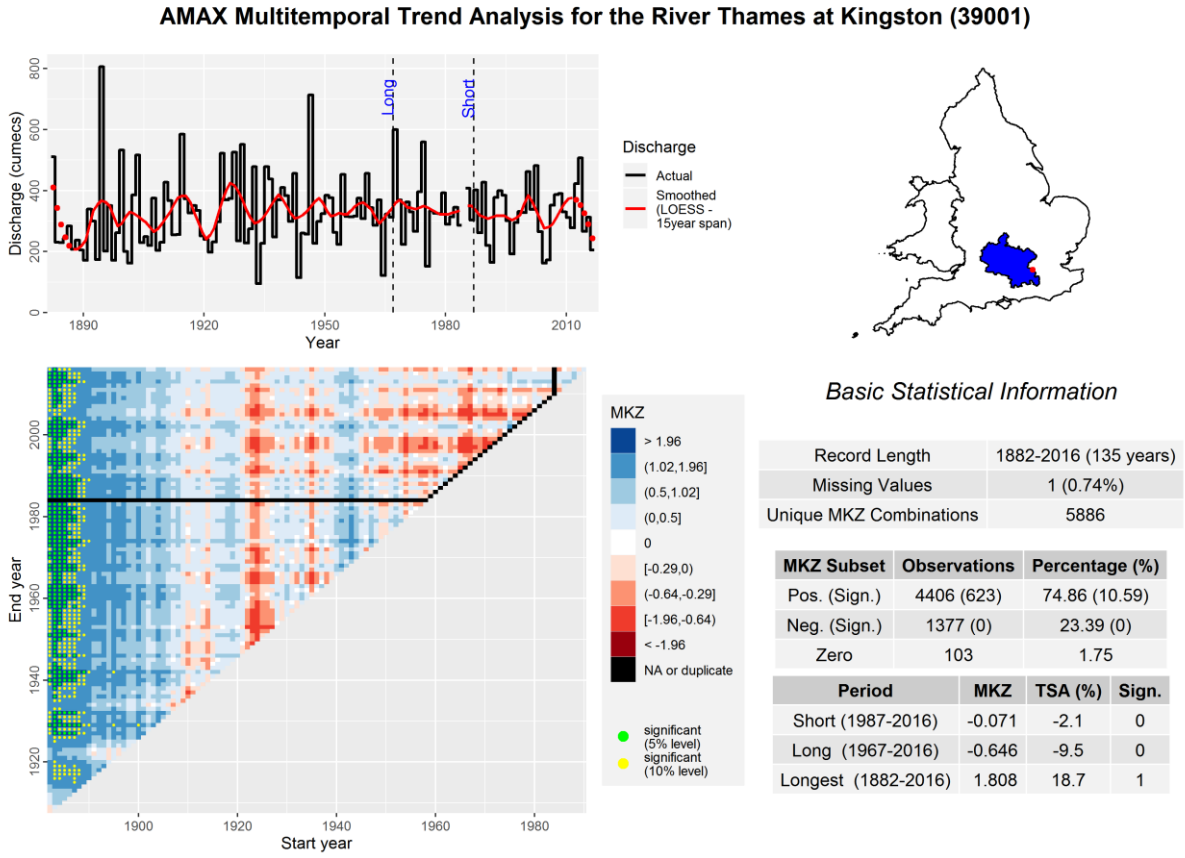


Figure A-1 Sample figure of the multi-temporal trend analysis for a long record, the Thames at Kingston (39001).

The bottom left plot of Figure A-1 shows a multi-temporal analysis, where each cell in the plot shows the result of a trend analysis with a given start year (as shown on the x axis) and end year (show in on the y axis). The colour scale shows the trend output (the MK Z statistic, with blue = positive and red = negative, and darker colours showing stronger trends. Significance values are shown with green (5% significance level) and yellow (10% significance) dots. Note that the black cells indicate where trend analysis periods are excluded because of duplication (that is, due to missing data in 1984,

analysis periods starting or ending in 1984 must be excluded as they contain the same information). For further information on the tables, see Appendix 1, Annex 1.

Results and discussion

Summary of trends across data set

For the 5 fixed periods of interest, it is evident that the majority of the stations had increasing trends (Table A-1). There were more than double the number of stations with increasing trends compared with decreasing trends. Nevertheless, while there is a clear dominance of positive trends, it must be emphasised that a majority of these trends are non-significant (around a quarter of positive trends were significant at either 5% or 10%).

Table A-1 Summary of MKZ trends for 5 periods of interest. Each cell shows the number of cases

	Short	Long	Full	1987 to 2016	1967 to 2016
>0	287	186	318	261	132
>0 & Sign. 10%	18	15	36	18	7
>0 & Sign. 5%	30	37	63	24	19
=0	14	6	4	14	5
<0	134	80	149	128	65
<0 & Sign. 10%	5	5	10	5	5
<0 & Sign. 5%	5	8	12	5	7
SUM	435	272	471	403	202

By analysing the results for all the combinations of start~end year, the conclusions are similar to above (Table A-2). The median percent of positive trends is over 74%, meaning that half of the stations have more than 74% of positive trends for all the examined combinations. On the other hand, there are considerably fewer negative trends, and only around a quarter of the stations (3rd quantile – 57%) have more negative trends than positive.

Table A-2 Summary statistics of MKZ for all combinations of start~end year

%	Positive (%)	Sign. Pos. (10%)	Sign. Pos. (5%)	Negative (%)	Sign. Neg. (10%)	Sign. Neg. (5%)	Zero (%)
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1st Qu.	39.9	0.0	0.0	4.8	0.0	0.0	0.3

Median	74.2	1.8	0.1	23.8	0.0	0.0	1.1
Mean	65.6	5.6	9.5	32.9	1.7	1.6	1.5
3rd Qu.	94.8	8.6	9.1	57.6	0.8	0.0	2.2
Max.	100.0	100.0	97.5	100.0	38.3	43.1	10.9
NAs	4.0	4.0	4.0	4.0	4.0	4.0	4.0

Multi-temporal analysis

From the same table, it can be concluded that there are stations with persistent positive/negative trends for all the combinations.

Examples of stations showing negative trends are 54040, 27010, 31004, 30015 and 28086. Especially for station 54040 (the Meese at Tibberton), about 40% of combinations of start and end years show a significant negative trend (Figure A-2).

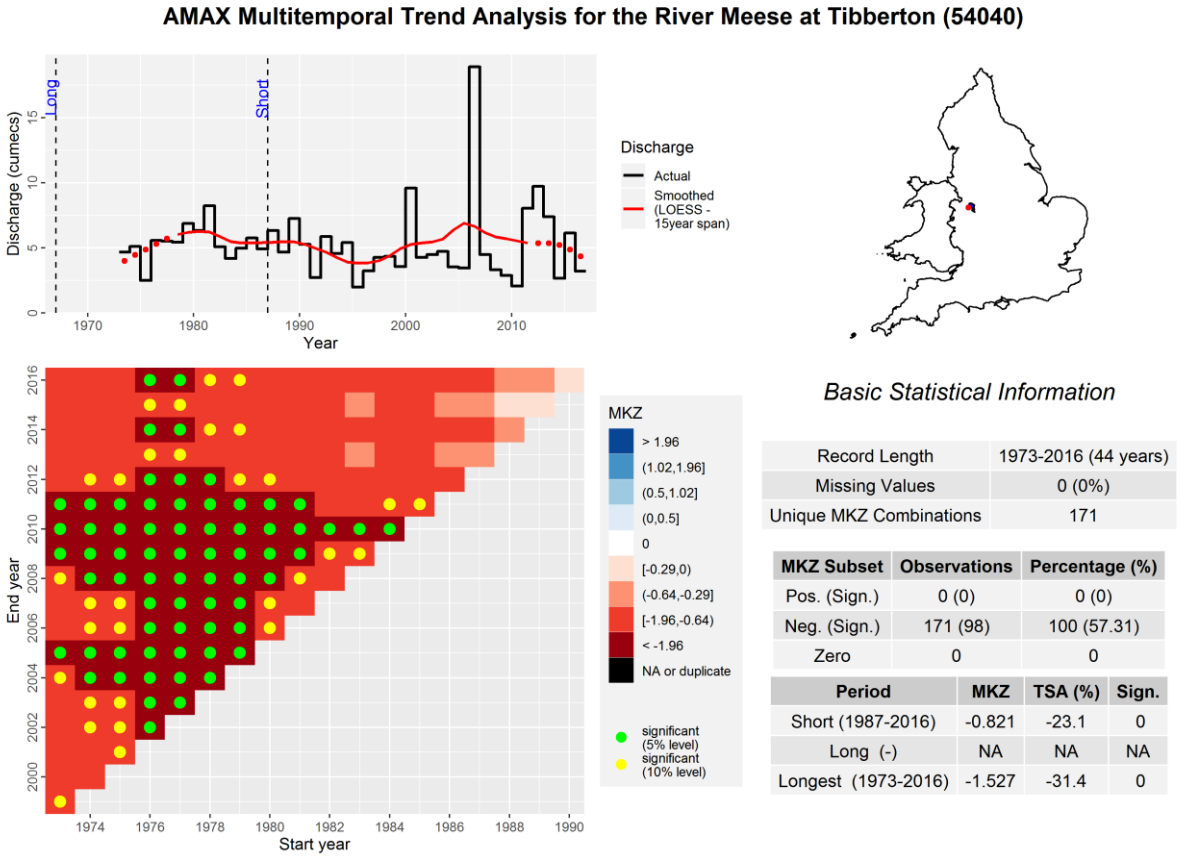


Figure A-2 Multi-temporal trend for station 54040 (the Meese at Tibberton)

There are 29 stations with positive trends for all combinations of start and end years, while 5 of them have over 80% of these trends being significant (for example, station 58007, Llynfi at Coytrahen, Figure A-3).

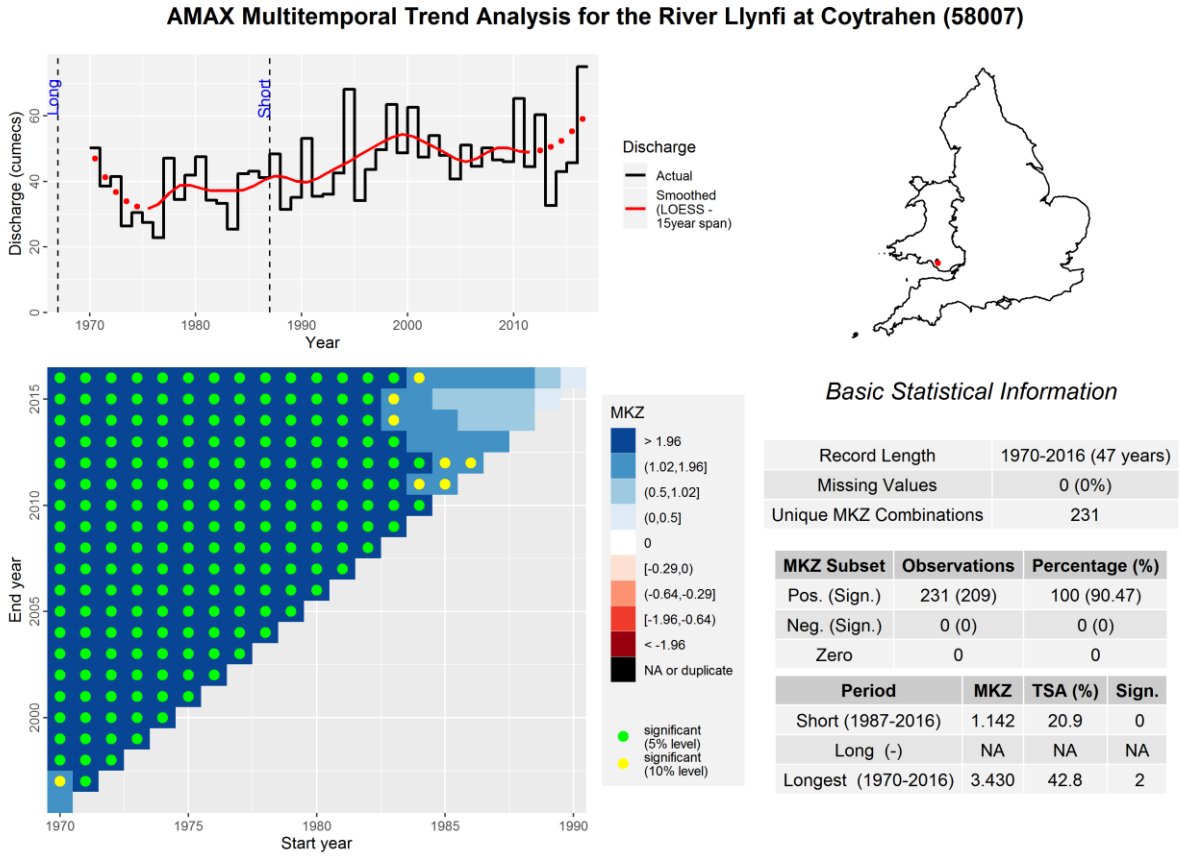


Figure A-3 Multi-temporal trend for station 58007 (Llynfi at Coytrahen)

However, the multi-temporal analysis more typically illustrates how trends in fixed periods are not representative of the full range of hydrological variability. There are often changes in the magnitude and even direction of trends over the course of the period of record (as can be seen for the long Thames series in Figure A-5).

This sensitivity to start and end years is a very widely known issue and discussed at length in the literature (see Hannaford, 2015 and references to a UK context).

There is considerable discussion in the literature about how long a period is needed to adequately quantify trends, with 50 years sometimes cited (for example, Kundzewicz and Robson, 2000). However, while as long a record as possible is favoured, arguably no single period can reliably be used to characterise trends given the presence of interdecadal variability.

Such variability between ‘flood rich’ and ‘flood poor’ periods is clearly prevalent in UK hydrometric records, as shown in the smoothed LOESS series presented in the trend outputs. Even in a 50-year record, longer periods than the existing record could give different results still. A range of studies has shown that historical flood events reconstructed from epigraphic, documentary or other sources are often larger than events in contemporary gauging station records, and flood-rich and flood-poor periods exist in centennial scale reconstructed records (for example, Wilby and Quinn, 2013; MacDonald and Sangster, 2017).

As such, the multi-temporal analysis provides important context for the fixed periods (which allow fairer spatial comparisons between sites and therefore are better suited to national-scale ‘headline’ results) and should be used in association with those results.

It is important to underline, therefore, that a trend in a fixed period should be seen as a descriptor of non-stationarity in that period alone. Trends should not be extrapolated into the future, because, as the multi-temporal analysis makes abundantly clear, variability over a range of scales (between years to between decade) means that the strength and direction of trends could change in future.

Variability in trends over time can influence the outcomes of flood frequency estimates (Griffin and others, 2019), which underlines the importance of contextualising results from any fixed period analysis and communicating sensitivity to changing study period in any outcomes. It also supports the case for investigating non-stationary frequency estimation (Faulkner and others, 2020).

An important question is whether a trend for any given period is representative or meaningful, in terms of reflecting a long-term change rather than short-term variation. However, this cannot be answered by trend detection using linear, monotonic trend tests, and should be addressed by moving from detection into attribution, that is, identifying the processes driving the trend (Merz and others, 2012).

This is, however, a much more challenging scientific task and one that requires observational data sets to be combined with hydrological models. It is typically extremely resource intensive even at the catchment scale (for examples see Harrigan and others, 2015, Prosdocimi and others, 2017), and therefore is well beyond the scope of the present study.

Spatial patterns

The maps (Figure A-4) for the full period again show the propensity for positive trends in AMAX across much of England and Wales.

Large areas of central and eastern England also display negative trends, but these are often non-significant (except for a coherent cluster in the Thames catchment and some in the north-east and Midlands at the 10% significant). In comparison, significant increasing trends are prevalent across the UK, especially in northern England, Wales and parts of western central England.

For the long period, the patterns are broadly similar, despite the sparser coverage. For the short period, again results are similar. Note that in Figure A-4 the results for the short and long periods are shown with no fixed end year, because the results for the same periods, but ending in 2016, were so similar, but there are more sites included in the former. These alternative periods are provided in an accompanying map, see Appendix1, Annex 1.

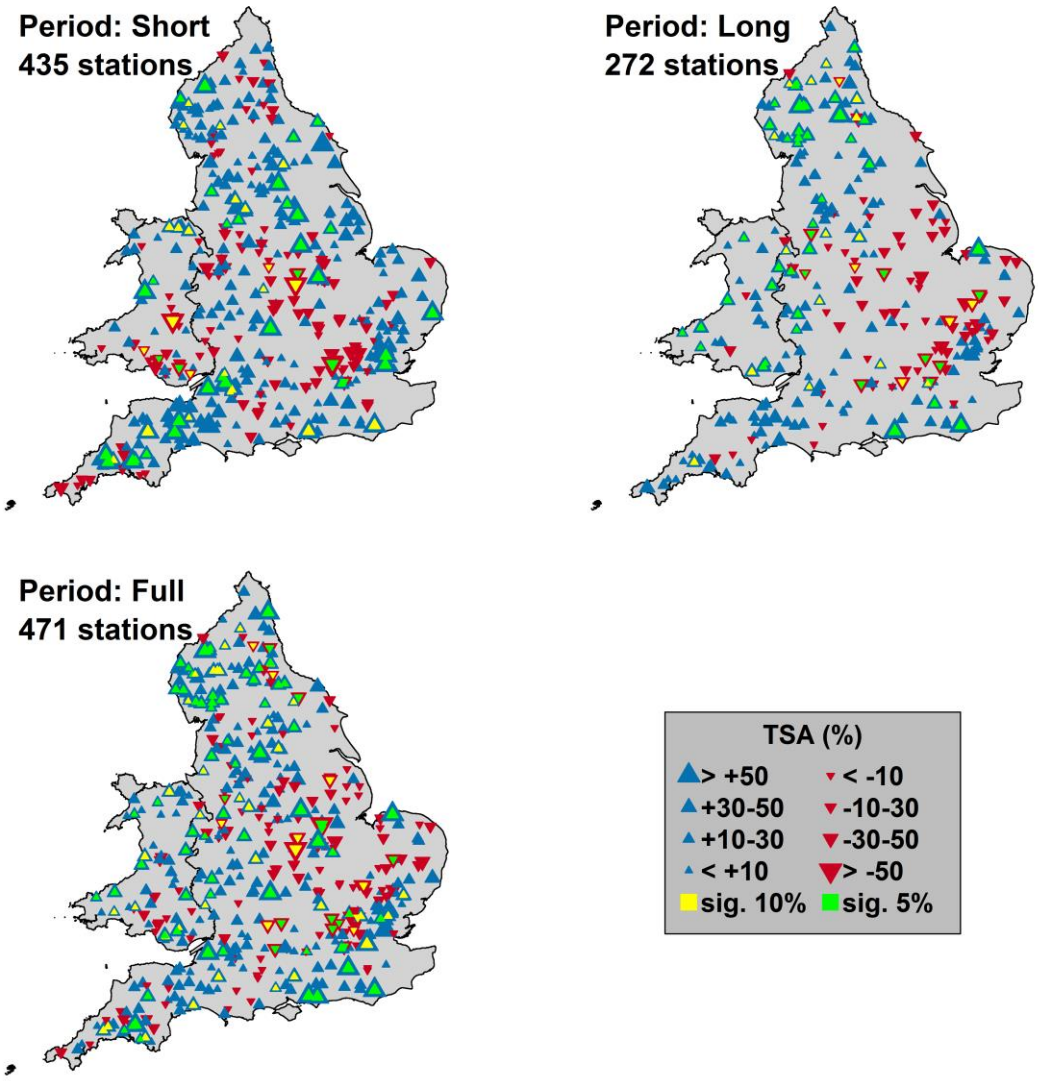


Figure A-4 Maps of trend results for the full period of record and the short and long periods

If the long period is assumed to have the most robust outcome (balancing the length of record while also being a fixed period allowing comparison between sites) to compare with other work, it is possible to conclude that the national picture agrees with previously published research on trends in AMAX and other flood indicators (Hannaford, 2015). That is, there has been a tendency towards higher flows in northern and western areas over the last 4 to 5 decades. The majority of that work was carried out in study periods ending in the mid to late 2000s, so the current study provides an update of around a decade.

Echoing other studies published in the last few years (for example, Brady and others, 2019, Prosdocimi and others, 2019), it appears that the previously identified gross patterns of change in the UK appear to be fairly resilient. In other words, these tendencies have not been countered by the addition of new data. Indeed, if anything, the results suggest an increase in the prevalence of and in the proportion of significant positive trends. To a degree, this is not surprising given that the recent decade includes

some very major flood events (for example, the winter 2015 to 2016 floods, which have a strong influence on the number of significant positive trends in northern England).

The results accord with Harrigan and others (2018), who, using the same testing methodology reported primarily positive trends in high flows (the Q5 flow in each year), with significant trends in northern and western areas. However, the current study uses peak flow data as opposed to daily flow data. It also uses the latest NRFA Peak Flows (v7) data, which includes AMAX data up to the 2016 water year, whereas Harrigan and others (2018) featured data up to 2014. The higher (relative) number of significant trends in England and Wales in the current study may reflect the addition of the 2015 to 2016 floods.

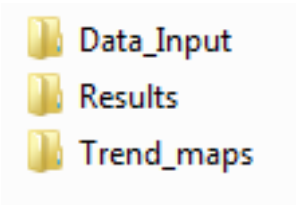
The current study also features the entire peak flows data set (that meets the agreed study criteria) rather than focusing on near-natural catchments. The agreement with Harrigan and others (2018) is encouraging, as that study deliberately focused on near-natural, high quality stations to prevent spurious trends arising from poorer quality data or anthropogenic effects. Here, similar geographical patterns are shown using the whole peak flows data set, including very heavily influenced catchments. This suggests that, at the national scale, a similar 'headline' picture emerges even when all catchments, of varying properties and degrees of disturbance, are mixed together. Of course, at individual catchments, trends may still be influenced by data quality issues or human disturbances rather than hydro-climatic variations, so the picture from the full peak flows data set is inevitably a 'noisier' one than for the benchmark network.

One important spatial contrast with Harrigan and others (2018) is that there are more negative trends in the present study, particularly in central England, than in the results for the UKBN. A straightforward comparison is difficult because the studies used different indicators and there is a different end year (while there are only 2 years, major record-defining floods in both these years, but in different parts of the country, could impact on spatial comparisons). Nevertheless, the lack of agreement warrants further investigation to determine whether the negative trends in the current study affect catchment disturbances (which seems unlikely given their spatial coherence) or whether the absence of negative trends in the UKBN reflects a bias in that set. Most importantly, however, these differences in central and eastern areas are mostly between non-significant trends and so could reflect chance differences. However, the greater feature of this study are the increases in northern and western areas, which is in agreement with both Harrigan and others (2018) and other published research.

The present study also shows that there are positive trends in other areas, and that some of these trends are significant (for example, in the far south-east of England and parts of east Anglia). The general dominance of positive trends for the UK agrees with several recently published studies of spatially coherent trends in flooding at the national scale (Brady and others, 2019, Prosdocimi and others, 2019). These studies use a novel Bayesian approach to characterise regional- and national-scale trends, rather than focusing on at-site trends. While these studies show that the signal towards positive trends in flooding is prevalent at the large (national) scale, Prosdocimi and others (2019) also show significant regional variations in the strength of trends, with broadly similar patterns to those shown here.

Annex 1: Outputs provided

The results of the analysis are provided in the 'FRS18087-IG-D2-digital_outputs.zip' file, in the '2. Multi-temporal trend testing' directory. This file contains various subfolders and files as shown on the adjacent figure. More specifically, there is one folder with the processed data used for the analysis (Data_Input), one folder with the results for the multi-temporal analysis (Results) and one folder with compiled figures for England and Wales (Trend_maps).



Data input

The 'AMAX_Info.csv' provides basic statistical data for the AMAX time series of the stations, while the 'AMAX_matrix.csv' is a compiled file with all the AMAX time series for the stations of interest.

Results

The 'Multi-temporalPlots.zip' provides a figure for each station of interest. A sample of this figure is presented in Figure A-5 and a brief explanation about the various sections of the figure follows:

- A: Title providing information about the river, the gauge location, and the station ID based on the NRFA website.
- B: Time series with the actual AMAX data (black) and smoothed data (by locally weighted regression smoothing (LOESS) using a span of 15 years; Harrigan and others (2018).
- C: Location of catchment and gauge on the map of England and Wales.
- D: Heatmap with the results of the MKZ multi-temporal analysis. Each pixel shows the result of the trend analysis for the corresponding start year (x axis) and end year (y axis), with the colour reflecting the direction and strength of the trend according to the MK Z statistic. The values for the MKZ 'bins' in the colour scale are derived from the quantiles of the MKZ scores for every combination of start and end year across every gauge (positive data are used for the positive quantiles, and negative for the negative ones). The values correspond to the 25^{th} quantile, 50^{th} quantile (median) and +/- 1.96. The value of 1.96 is selected since this indicates statistical significance at the 5% level. Coloured dots also show cells significant at the 5% or 10% level. The significance is based on the MKZ value derived from either conventional significance testing or using the block bootstrap, as explained in the methodology section.
- E: Information about the record length and the percent of missing values for the station.
- F: The information in this table is derived from all the multiple combinations of start~end year for which the station of interest has sufficient data (as shown on the heatmap, section B). This particular table summarises these results, and shows how many of them are positive, negative, both significant and positive, and both significant and negative.
- G: This table provides information about the MKZ and TSA for 3 particular periods of interest, as described in the methodology section.

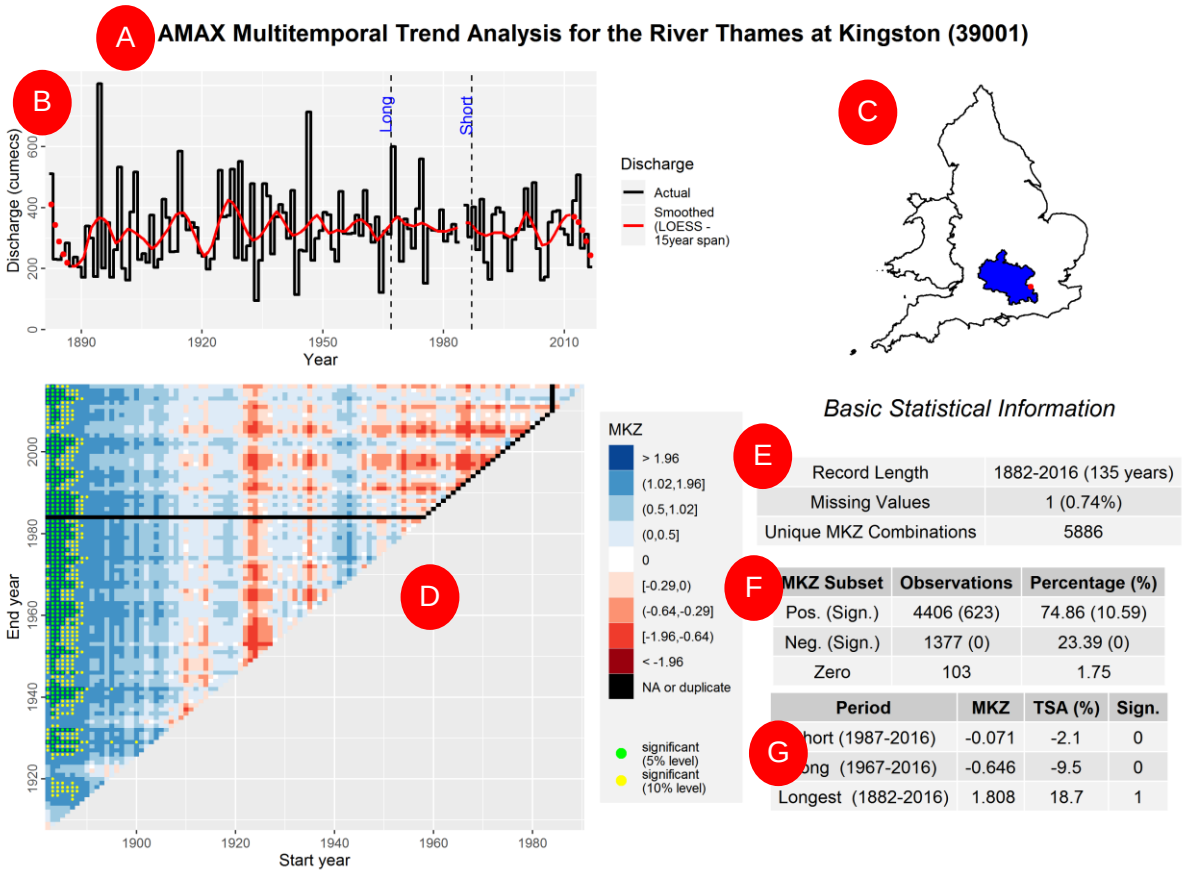


Figure A-5 Sample figure of the multi-temporal trend analysis

The 'AMAX_StationsSummaryTable.csv' file provides basic statistical information for the trend testing. More specifically, there is information about the total number of years (columns B, C), the percent of missing data (column D), and the percent of positive, significantly positive, negative and significantly negative combinations of start~end year for each station (columns E:H respectively).

Finally, there is information related to the MKZ value for 3 particular periods and the significance of these values (columns I:N). This file is essentially a summary of the tables E, F, G from the above figures. Where results are marked as 'NA' for some periods, it is generally because there is too much missing data for the testing to take place.

The 'Statistics.xlsx' file provides summary information about the MKZ results for the 5 fixed periods of interest (short, long, full, 1967 to 2016, 1987 to 2016) on the 'FixedPeriods' spreadsheet. On the 'Multi-temporal' spreadsheet, statistics from all the combinations of start~end year are provided.

Trend maps

This file has a range of maps of England and Wales with the MKZ results for each station. These maps are also reproduced here in the results and discussion.

Appendix B: Developing methods for incorporating physical covariates in non-stationary analysis

Trial set of catchments

The effectiveness of physical covariates was tested on 8 trial catchments, listed in Table B-1. All provide high-quality measurements of flood flows with long records. They were chosen to show a variety of trend behaviours, catchment locations and types. The climate index data run from 1950, and so the tests described in this section are restricted to the years since 1950. The list deliberately includes one catchment with known land use change, Bedburn Beck, where it is possible that trends are influenced by afforestation and felling.

Table B-1 Trial catchments for testing physical covariates

Station number	River	Station	Record length (years)[4]	Mann-Kendall test result at 5% significance level[5]	PELT change point test
24004	Bedburn Beck	Bedburn	57	Upward, significant	Positive change in 1982
27034	Ure	Kilgram Bridge	50	Upward, significant	No change point
33034	Little Ouse	Abbey Heath	48	Downward, significant	No change point
42010	Itchen	Highbridge & Allbrook Total	59	Upward, not significant	No change point
45001	Exe	Thorverton	61	Upward, not significant	No change point
55002	Wye	Belmont	109	Upward, significant	Positive change in 1977

[4] Excluding missing years and years classed as rejected.

[5] The Mann-Kendall test results are taken from the multi-temporal trend testing carried out as part of this project. Results quoted here are those for the full period of record. The PELT tests were carried out as part of the data screening. Several catchments have a trend that is not significant according to the non-parametric Mann-Kendall test, but is expected to be more significant when judged using parametric tests.

60002	Cothi	Felin Mynachdy	56	Upward, not significant	No change point
76005	Eden	Temple Sowerby	53	Upward, not significant	No change point

Choice of covariates

The first task was to determine which covariates to focus on and to obtain the data needed to create a covariate observation to pair up with each AMAX flow in each trial catchment.

Example physical covariates in the literature include:

- annual rainfall (Sraj, 2016, Yan and others, 2017)
- 99th percentile of daily rainfall values during a year or season (Prosdocimi and others, 2014)
- urban extent (Prosdocimi and others, 2015)
- population of the catchment (Yan and others, 2017)
- climatic indices such as the NAO (Steirou and others, 2019), EA (Francois and others, 2019), El Nino Southern Oscillation (El Adlouni and others, 2007) or Interdecadal Pacific Oscillation (Franks and others, 2015).

The research considered only covariates that are expected to be significant across many catchments rather than those that represent locally-specific effects such as urbanisation or changes in forest cover.

On the face of it, **rainfall** is an attractive covariate. Rainfall covariates can be useful for identifying interannual variation in extreme flows. Rainfall is expected to be well correlated with fluvial flood flows, particularly when it is accumulated over a duration similar to the critical storm duration of the catchment. However, this can lead to a dead end in that if a covariate is very closely correlated with peak flow, this can result in the circular logic of needing to know the frequency distribution of something like peak flow in order to estimate the frequency distribution of peak flow. The problem of estimating high quantiles has been shifted to another part of the hydrological cycle rather than being solved. A more useful role for rainfall as a covariate is to consider longer duration totals (annual or seasonal), which can be associated with catchment conditions such as soil moisture and groundwater storage.

Global mean **temperature** was also considered as a covariate, although for reasons discussed later it was not considered further. The intention was to include an index of the changing climate.

River flows can vary according to **large-scale patterns of atmospheric circulation**, sometimes referred to as ‘teleconnections’. These patterns can be described by modes which indicate the position and magnitude of large-scale atmospheric waves. Over Europe and the north Atlantic the modes control the strength and location of the northern hemisphere jet stream, therefore strongly affecting near-surface climate conditions (Steirou and others, 2019).

A recent comprehensive study (Steirou and others, 2019) investigated the influence of 5 climatic covariates on flood flows across Europe: North Atlantic Oscillation (NAO), east Atlantic pattern (EA), east Atlantic–western Russian pattern (EA/WR), Scandinavia pattern (SCA) and polar–Eurasian pattern (POL). The authors fitted GEV distributions, whose parameters varied according to these 5 covariates, at 600 river flow gauges in Europe on catchments larger than 200 km^2. The authors allowed only the location parameter to vary, after finding similar results when allowing both location

and scale to vary. They found a generally better fit if seasonal average indices are used, rather than indices from the month in which the peak discharge occurs. This may be because catchment wetness is more influenced by the seasonally-averaged climate state.

The results, visualised in Figure B-1 showed that for Great Britain the most influential covariate in winter is the EA. There is also a large influence, in most seasons, from the NAO. In spring, there is relatively little influence of atmospheric circulation indices on flood magnitudes, perhaps due to local convective rainfall processes starting to dominate.

Steirou and others (2019) found that the effect on estimated quantiles is that between years with medium and high values of the NAO or EA, peak flows can differ by 10 to 20% in Britain for some seasons.

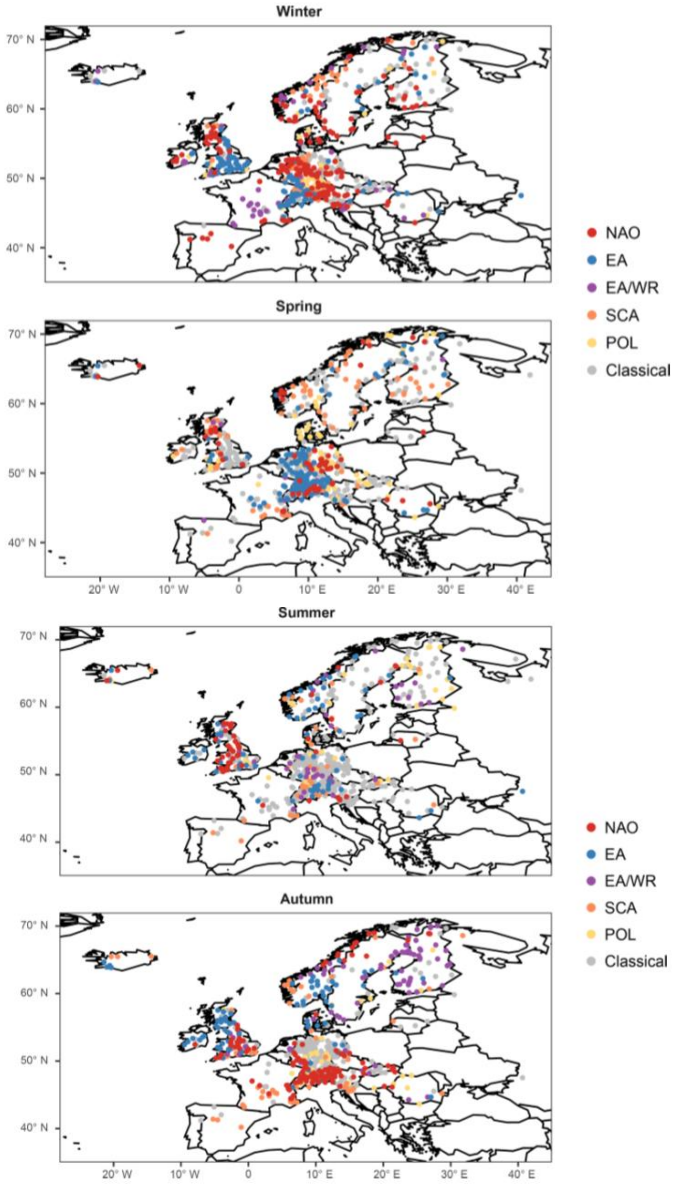


Figure B-1 The maps, from Steirou and others (2019), show best overall models for each season, using mean seasonal covariates. 'Classical' refers to a model without covariates

The results are consistent with the findings of Brady and others (2019) who found a clear association between the EA index and annual maximum flows in Great Britain. In contrast, although NAO on its own was influential, its influence vanished when time was included as a covariate because of the strong temporal trend in NAO.

In light of these results, the project team considered the following as covariates:

- catchment-average rainfall (annual, autumn and winter)
- NAO index (winter, summer, autumn)
- EA index (winter)
- temperature (global mean, annual and winter) (later discarded)

Catchment-average rainfall accumulations were calculated from the CEH-GEAR data set, which provides daily rainfall on a 1 km grid across the UK from 1890 (Tanguy and others, 2016).

NAO and EA indices were obtained from NOAA[6]. The indices are calculated from anomalies that are standardised by monthly means and standard deviations calculated over a 1950 to 2000 baseline period.

Global mean temperature was obtained from the Hadley Centre[7], in the form of anomalies from a 1961 to 1990 baseline period.

In all cases, annual values of the covariates were calculated using water years.

The rainfall data were centred and scaled, subtracting the mean from each observation and dividing the result by the standard deviation. This transformation is expected to reduce computational problems in which the likelihood optimisation algorithm converges to local maxima; each of the covariates is essentially on the same scale, so the regression parameters are more comparable over covariates for numerical purposes. The other covariates were already standardised in a similar way.

These covariates, although physically-based, do not directly represent the physical processes that cause floods. For instance, none of the covariates measures the strength and direction of atmospheric rivers, which have been linked with the occurrence of winter flooding in the UK (Lavers and others, 2011). However, they represent a potential step forward from the simplistic approach of modelling non-stationarity as a change over time.

Screening covariates using trend analysis

One important consideration is the effect of collinearity and confounding variables on results when including multiple covariates. Although it does not matter for fitting if there is dependence between the covariates, the greater the dependence between covariates, then (i) the harder it is to interpret the regression coefficients and (ii) the more numerical instability/convergence issues that will arise.

Therefore, to avoid or reduce these problems, it is desirable, although not essential, if the covariates are orthogonal (uncorrelated). In order to look for causal relationships through the regression modelling, the fitted models should be as meaningful and interpretable as possible. This therefore leads to having as orthogonal covariates as possible. When time is one of the covariates, this is achieved by detrending the physical covariates before including them in the statistical model to ensure that they

[6] https://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.shtml (viewed 14 May 2020)

[7] https://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/download.html#regional_series (viewed 14 May 2020)

are orthogonal. For example, if physical covariates include a time trend, they are correlated with the water year which may be another covariate.

Figure B-2 *to* Figure B-4

Figure B-2 Time series of annual and seasonal rainfall over an example catchment (the Itchen), with linear trend lines fitted

show how the covariates change over the water years 1950 to 2016. A common colour scheme is used across the 3 figures to indicate the season over which the covariate is calculated. Table B-2 summarises the statistical significance of the trend in each covariate. Some of the chosen physical covariates show a strong trend over time; for others, there is no trend evident during the observed period.

There is a strongly significant trend in the mean global temperature, as would be expected. Since the trend appears to be reasonably linear over the period of record, it may be that this covariate provides little more information than would be provided by adopting the year of each flood as a covariate. Although superficially it might be thought that temperature is a useful covariate as its future evolution can be predicted with reasonable confidence by climate models, this reasoning could equally well apply to using time as a covariate, as its future values can be predicted perfectly. Since there is no clear causal connection between global temperature and UK flood magnitudes, global temperature was no longer considered as a covariate. This leaves 7 candidate physical covariates. However, in the future, water year and global temperature are likely to become non-linearly related, so it is possible that global temperature could become a more useful covariate for future periods.

The NAO and EA indices show a mixed picture in terms of trend. The EA (annual and winter) and the NAO (winter) show trends towards more positive values over the period of record. When these indices are included alongside time as a covariate, they need to be detrended. In contrast, there is little or no evidence of significant trend (at the 5% level) in the annual, summer or autumn NAO.

There is no evidence of trend in annual, winter or autumn rainfall over most of the example catchments. These rainfall covariates can be incorporated alongside time in a non-stationary model, with no need for detrending. The exceptions are the Ure and Eden, both upland northern catchments. Over the Ure there is an upward trend in winter rainfall; over the Eden there is an upward trend in both annual and winter rainfall.

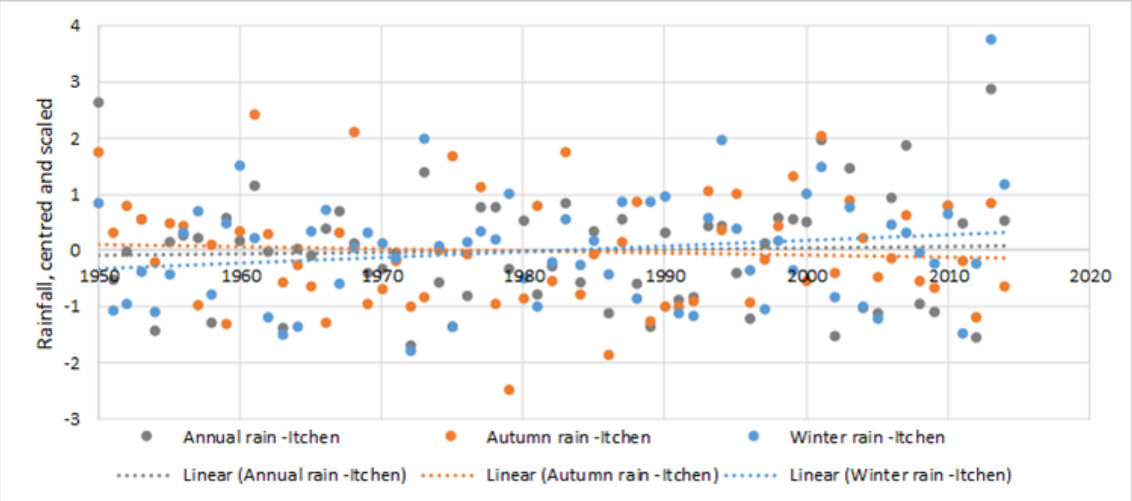


Figure B-2 Time series of annual and seasonal rainfall over an example catchment (the Itchen), with linear trend lines fitted

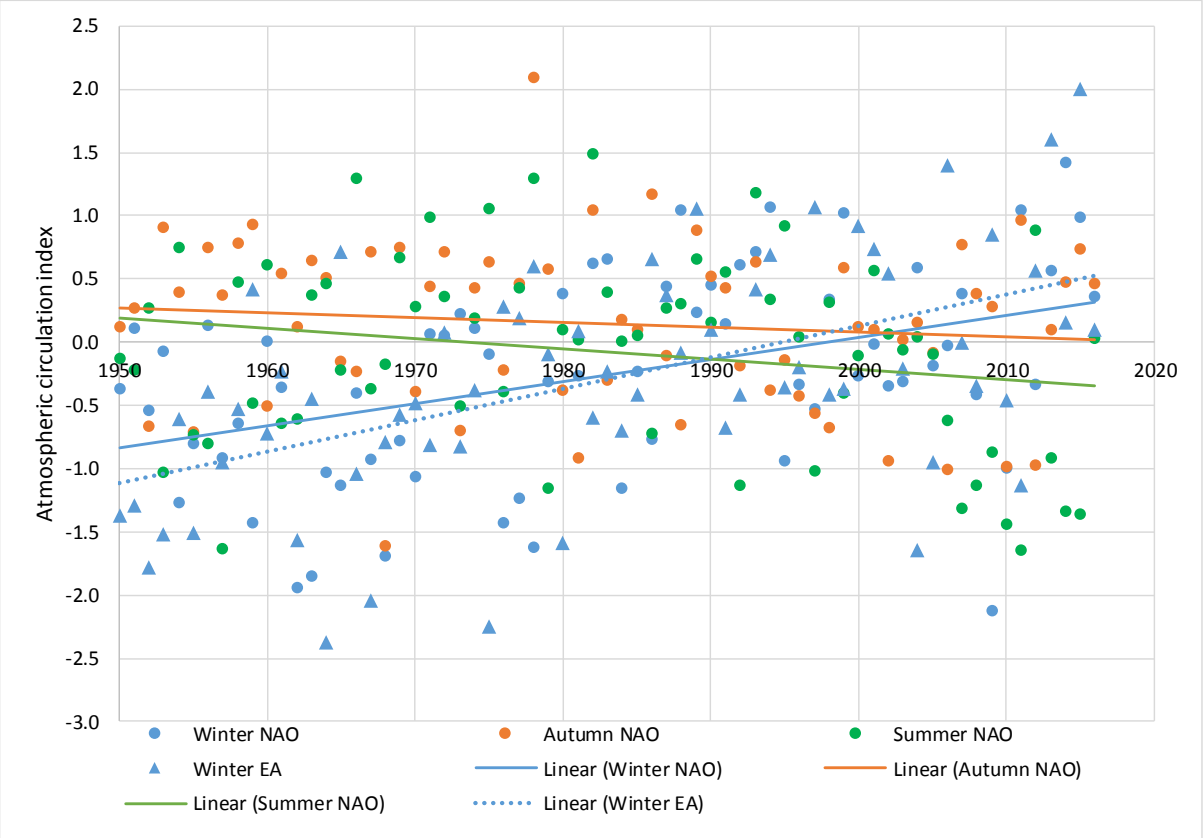


Figure B-3 Time series of atmospheric circulation indices, with linear trend lines fitted

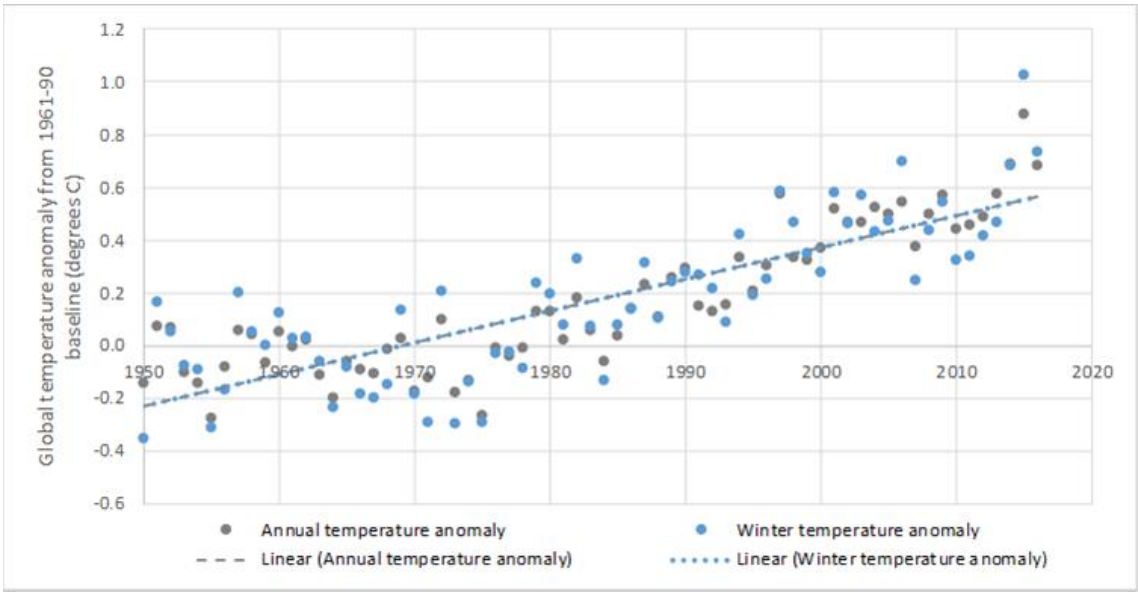


Figure B-4 Time series of global temperature anomaly, with linear trend lines fitted

Table B-2: Trend tests of covariates

Covariate	**Annual NAO**	**Annual EA**	**Winter NAO**	**Winter EA**	**Summer NAO**	**Autumn NAO**
Mann-Kendal Z statistic	1.943	6.326	3.453	4.524	-1.634	-1.001
2-sided p-value	5.2%	0.0%	0.1%	0.0%	10.2%	31.7%
Provisional interpretation	Borderline trend	Trend	Trend	Trend	No trend	No trend
Covariate	**Annual temperature anomaly**	**Winter temperature anomaly**	**Winter rainfall, Ure catchment**	**Annual rainfall, Eden catchment**	**Winter rainfall, Eden catchment**	
Mann-Kendal Z statistic	8.139	7.230	2.392	2.738	2.500	
2-sided p-value	0.0%	0.0%	1.7%	0.6%	1.2%	
Provisional interpretation	Trend	Trend	Trend	Trend	Trend	

For catchment rainfall: only showing results where there is trend evident at the 5% significance level

Screening covariates using correlations

A wide variety of combinations of covariates has been investigated for the 8 catchments listed in Table B-1. As a first step, correlations were calculated between covariates and annual maximum flows, and cross correlations between covariates. The results for correlations of the following covariates with AMAX flow (not tabulated here) indicated:

- annual rainfall: moderate to strong positive correlations (strongest on the Itchen, with a coefficient of 0.70, where river flow is dominated by the groundwater level, which will be strongly influenced by annual rainfall)
- winter rainfall: moderate correlations on most catchments, an exception being Bedburn Beck, with little correlation
- autumn rainfall: low to medium correlations on most catchments
- NAO: little correlation between AMAX flow and annual NAO. Winter NAO shows little to moderate positive correlation for most catchments, strongest on the Wye. Summer and autumn NAO show little to moderate negative correlation, strongest on the Bedburn Beck (summer) and Exe (autumn)

- EA: little to moderate positive correlation for both annual and winter EA, strongest on the Wye (annual) and Itchen (winter). Less correlation between summer EA and AMAX flow
- water year (and global temperature anomaly): positive correlation on most catchments, strongest on the Wye and Bedburn Beck. Negative on the Little Ouse

As expected, strong cross correlations were observed between annual and seasonal versions of the same variables, such as rainfall or NAO. Correlations were also seen between atmospheric circulation indices and rainfall for most catchments, when averaged over the same season, in particular between winter rainfall and winter EA. It is desirable that the covariates included in a non-stationary statistical model are orthogonal, so highly correlated covariates should not be included. This is one consideration, along with the others listed in section 3.3.4, that helps select covariates.

Incorporating physical covariates into estimating design flows

One of the biggest challenges the project faced was how practitioners could extract estimates of design flows from models fitted using physical covariates to inform design specifications, given the different ways of defining risk. Typically, flood risk estimates are defined using return periods, but in non-stationary models, this becomes more complicated. Given a model with time covariates, one could use the **conditional return level** (or **conditional flow estimate**) given the current year, that is, the design flow is conditional on the year being 2019.

This is less useful when using physical covariates, as for example, the 100-year return level given the 2019 rainfall amount is the expected return level under the (clearly hypothetical) conditions that the rainfall always takes the value that is observed in 2019. Therefore, the conditional flow estimate may be useful when examining the probability of past floods, but it is less informative when thinking about design.

In this case, an alternative approach would be to define a **marginal return level**, as used in Eastoe and Tawn (2009). This is also referred to as an **integrated flow estimate**. It is defined as the return level corresponding to the averaged encounter probability[8], where averaging is over covariates in a period of interest. Specifically, it is the return level where the average encounter probability is equal to the reciprocal of the return period.

This is distinct from the stationary estimate of return level, even though the two might appear superficially similar because both will plot as horizontal lines on a time series graph. The marginal return level could also be calculated separately for different portions of the record, for example, 10-year blocks. It is also possible to calculate the marginal return level by integrating over a sample of covariates that spans a period different from that covered by the river flow data. For instance, the record of the covariates might be longer, enabling a more confident estimate of the distribution of the covariates.

Another situation might be incorporating information on predicted future covariates such as scenarios of land use or climate. This may be useful for planning future flood protection, although it would only be justifiable in situations where there is a demonstrated causal link between the covariates and flood peaks, and so is not recommended for current application. If this approach were to be applied, it would be possible to use the Eastoe and Tawn (2009) approach to derive levels corresponding

[8] An encounter probability is the probability of an event occurring in a given number of years.

to the expected annual exceedance probability over the design period. This is equivalent to the 'average design life level' presented in Yan and others (2017). Rather than aiming to estimate a flood frequency curve for a particular year (past, present or future), they estimate quantities that are associated with a period of time, such as the design life of a flood alleviation scheme. The output from the analysis is similar to a probability of a particular flood flow being exceeded during the lifetime of a scheme or a development.

The candidate covariates vary in their future predictability. Annual and seasonal rainfalls are included in the UKCP09 and UKCP18 data sets. Large-scale atmospheric circulation indices are more difficult to predict. Changes in the latitude and speed of the jet stream over the North Atlantic can be described using a combination of the NAO and EA patterns. Climate models show large bias in predictions of the NAO and EA over the baseline period. While there is some agreement between models on a future northern shift of the jet stream, there is considerable spread between different model projections (Woollings and Blackburn, 2012). Some studies project a slight positive shift in the probability distribution of NAO phase and a small north-eastward displacement of its centre by the end of this century (Deser and others, 2017). The UKCP18 reports (Met Office, 2019) mention future evolution of the NAO index but not of the EA.

In summary, it does not appear to be currently feasible to obtain scenarios of future atmospheric circulation indices that will be useful to practitioners. Scenarios of future rainfall total are available, and were used in the non-stationary project in north-west England. However, in that project it was decided that the information was not helpful for predicting future flood frequency because impacts of climate change on peak flows cannot be simply modelled by a statistical relationship between peak flow and seasonal rainfall. This calculation would assume, wrongly, that future changes in peak flow can be entirely explained by changes in seasonal rainfall. Although climate change is expected to affect rainfall, and therefore catchment wetness, it can also be expected to influence other factors that control flood magnitudes. These include storm intensity and evapotranspiration (which influences soil moisture). Given that using future predictions of covariates over the design life of a flood alleviation scheme is only valid if the physical covariates provide a complete causal description of the non-stationarity in peak flows, this is not recommended for practical application.

In light of the cautionary notes about extrapolation into the future, there may be a desire to extract a ‘present-day’ return level from these non-stationary models. This can be done by calculating the marginal return level, setting the time covariate to the most recent water year for which data are available, and integrating over the full observed distribution of the physical detrended covariate.

Such an estimate would represent the present-day expected return level for a particular exceedance probability, without being conditional on any particular value of a covariate such as annual rainfall. It could probably represent the short-term future too. For the longer term, it could perhaps be adjusted using outputs from climate models, as long as an appropriate baseline period can be identified and the concerns about causal relationships can be overcome.

It would also be possible to set the water year covariate to other years within the record. However, this functionality is not available in the nonstat package. The project has termed these estimates **single-year integrated flow estimates**. The single-year integrated flow estimate can be more easily compared with alternative estimates such as those from a model that uses only water year as a covariate.

Calculating return level estimates

This section sets out the methods for calculating both marginal and conditional return levels, in each case for a non-exceedance probability value of *p*.

Notation:

- Let $F(y|x;\theta)$ be the distribution function of annual maximum flow, where x is a vector of covariates and θ is the vector of parameters of the distribution, including the regression coefficients for the covariates.
- Define the marginal return level for probability *p* (the marginal *p*th quantile) by y_p and the conditional return level for probability *p* (the conditional *p*th quantile), that is, conditional on a particular set of values for the covariates, by $y_p(x)$.
- Let $f(x)$ be the joint density of the covariates for the time period of interest. For the purposes of this work, this is the period spanned by the river flow record, but it could also be some future period if information is available about how the covariates might evolve, and if the covariates provide a complete causal description of the non-stationarity in peak flows.
- Let Φ be the set of possible covariates.

The marginal distribution function for AMAX flow in the period of interest is $F(y;\theta)$. Note that this distribution function does not depend on the covariates, that is, unlike $F(y|x;\theta)$ there is no conditionality on covariates, as this distribution is required irrespective of the covariates that occur in the period of interest.

The marginal distribution can be obtained by integrating out the covariates:

$$F(y;\theta) = \int_{\Phi} F(y|x;\theta)\, f(x)\, dx. \qquad \text{(B1)}$$

The marginal quantile for the period of interest is then y_p, which is such that $F(y_p;\theta) = p$.

Inverting F (which needs to be solved numerically), $y_p = F^{-1}(p;\ \theta)$ is obtained.

In practice, the parameters θ of this distribution are not known, nor is the true density $f(x)$ of the covariates. Therefore, to get an estimated marginal quantile (return level), they need to be replaced with estimates in Equation B1. Here, θ is estimated by $\hat{\theta}$, the maximum likelihood estimate of the parameters, and the joint density of the covariates $f(x)$ needs to be replaced by some estimate $\hat{f}(x)$, for example, the empirical density of the data or a kernel density estimate (which smooths the empirical estimate).

The estimated marginal *p*th quantile $\widehat{y_p}$ is found from $\widehat{y_p} = \hat{F}^{-1}(p;\ \hat{\theta})$. Similarly, the estimated conditional *p*th quantile is obtained from $\widehat{y_p}(x) = \hat{F}^{-1}(p|x;\hat{\theta})$.

Calculating confidence limits for return level estimates

Confidence limits are calculated using a parametric bootstrapping procedure. This is a method of deriving confidence limits in situations where the underlying statistical population is unknown or where an analytical solution for return levels is impractical. The following bootstrap method is proposed by Eastoe and Tawn (2009).

The following is a general algorithm for parametric bootstrapping from a parametric model with assumed distribution function $F(y;\theta)$, with parameters θ and an estimate $\hat{\theta}$ of the parameters derived from the independent and identically distributed data sample $y_1, \ldots, y_n$.

1. Generate data sample of size n from the fitted model, by using $F(y;\hat{\theta})$.
2. Fit the model for the data simulated in step 1 to give a new estimate $\hat{\theta}^{(1)}$.
3. Repeat steps 1 to 2 k times to give a set of estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(k)}$, called a bootstrapped sample.
4. Use the sample of k estimates to derive the confidence interval for each element of θ, by ranking the bootstrapped sample for each element and picking, for example, the 2.5% and 97.5% quantiles to give a 95% confidence interval.
5. If interest is in some function of θ, say g(θ), then the bootstrap sample $g(\hat{\theta}^{(1)}), \ldots, g(\hat{\theta}^{(k)})$ can be used to construct the confidence interval for g(θ).

Notes:

- This work has used k = 200, so the selected quantiles from the bootstrap sample for a 95% confidence interval, for example, would correspond to the 5th and 195th ordered values in the sample.
- When there are covariates in the model, step 1 differs depending on the type of confidence interval required. For conditional return levels, covariates remain fixed in step 1, but for marginal return levels, the covariates are sampled first (with replacement or from $\hat{f}(x)$) and then the data are simulated conditional on the covariates. In all other cases, the bootstrap method remains unchanged.

Confidence limits for the conditional return levels

Let $\hat{y}_p^{(i)}(x) = F^{-1}(p|x;\hat{\theta}^{(i)})$ for $i = 1, \ldots, k$, where during the bootstrapping procedure, the covariates are not resampled in step 1 before estimating the model parameters $\hat{\theta}^{(i)}$.

Then $\hat{y}_p^{(i)}(x)$, for $i = 1, \ldots, k$, gives a sample to build confidence intervals, taking, for example, the 2.5% and 97.5% quantiles of this sample to obtain the 95% confidence interval.

Confidence limits for the stationary and conditional return levels are calculated using the in-built bootstrapping code in the texmex R package.

Confidence limits for the marginal return levels

There are 2 sources of uncertainty in the marginal distribution: the parameter estimates $\hat{\theta}$ and the estimated distribution of the covariates $\hat{f}(x)$. Firstly, ignoring the latter source

gives $\hat{y}_p^{(i)} = F^{-1}(p; \hat{\theta}^{(i)})$ for $i = 1, \ldots, k$, from which confidence limits are derived as in step 5 above.

If the uncertainty in the distribution of the covariates is to be included, the sample of distribution estimates $\hat{f}^{(1)}, \ldots, \hat{f}^{(k)}$, needs to be used, either from the same bootstrap sample or from some independent source if the covariates are being set to represent a future period.

The following is then obtained:

$F\left(y; \hat{\theta}^{(i)}, \hat{f}^{(i)}\right) = \int_{\Phi} F\left(y|x; \hat{\theta}^{(i)}\right) \hat{f}^{(i)}(x)\, dx$ for $i = 1, \ldots, k,$

where in each of the k bootstrap samples, uncertainty in the covariate distribution is incorporated in the uncertainty of the estimated model parameters via the sampled covariates in step 1, and by including the same sampled covariates in the terms in the above integral. The resulting sample of estimated return levels (from the following equation) is used to build confidence limits in the usual way.

$\hat{y}_p^{(i)} = F^{-1}(p; \hat{\theta}^{(i)}, \hat{f}^{(i)})$ for $i = 1, \ldots, k.$

Notes:

- The uncertainty in $\hat{f}(x)$ can be ignored when time is the only covariate, since $f(x)$ in that case forms a uniform distribution with no uncertainty.
- One challenge with this approach is that it involves a large number of calls to the maximum likelihood estimator function, some of which may not converge.

Covariates selected for trial catchments

All non-stationary frequency analyses reported here use the GEV distribution. Results from alternative distributions are not expected to show any significant differences in terms of covariates selected.

The 7 candidate physical covariates were included in fitting of non-stationary models on some trial catchments in accordance with the following options:

1. In a group, allowing any number of physical covariates plus water year to be included, with the physical covariates being detrended.
2. Up to 2 covariates per model, with a maximum of one being a physical covariate (the other being water year), with the physical covariate being detrended.
3. Up to one physical covariate per model, with no detrending.
4. Only time allowed as a covariate.

In all these cases, covariates were considered for modelling either or both of the location and scale parameters.

This leads to a large number of candidate models: even when only one physical covariate is allowed at a time, there are 22 models fitted (7 physical covariates, times 3 for location, scale and both varying, plus one with no covariates). With up to 2 covariates, there are 88 models fitted. The number can grow to many thousands when more covariates are considered, which can lead to long run times.

The AIC and BIC were calculated for each model. Testing indicated that the BIC is a more useful measure because it tends to give preference to simpler models (see section 3.3.4). However, even the BIC is not an appropriate way to select models without considering other factors such as physical interpretability and spatial coherence.

For example, on the Little Ouse, when all covariates are considered to be included, the model with the lowest BIC includes the following covariates:

- for location: water year, winter EA, winter NAO, autumn NAO, summer NAO, annual rain, winter rain and autumn rain (that is, all 8 covariates)
- for scale: water year, summer NAO, annual rain and winter rain

In contrast, the lowest BIC model for the Wye includes:

- for location: water year and annual rain
- for scale: none

There is therefore large variation between catchments with respect to the covariates included in the model with the lowest BIC. This varies from 2 to 8 in the location parameter, and from 0 to 5 in the scale parameter (the total number of covariates varies from 2 to 12). This implies inconsistency across the catchments and those with more covariates produce warnings during the model fitting and are likely to be too complex to be useful. Therefore, to guide the model selection towards simpler versions with greater interpretability and spatial consistency, subsequent analysis was restricted to options 2 and 3 above, that is, a maximum of one physical covariate.

Table B-3 summarises the choice of covariates for each trial catchment; in each case listing the covariates corresponding to the model with the lowest BIC from the combinations available. Results are given for options 2 and 3 listed above, that is, with and without water year as a potential covariate. In addition, the final 2 columns show:

- the lowest BIC from models that use only water year as a covariate (for either location, scale or both parameters)
- the BIC for a stationary model, with no covariates

The BIC values in Table B-3 should be compared only within each row, not between the catchments.

Table B-3 Choice of covariates for each trial catchment (lowest BIC in bold)

Catch-ment	Lowest-BIC model considering time (water year) plus up to one physical covariate (detrended)			Lowest-BIC model considering up to one physical covariate (not detrended)			BIC for model with only time as co-variate	BIC for statio-nary model
	Covariates for location	Covariates for scale	BIC	Covariate for location	Covariate for scale	BIC		
Bedburn Beck	Time, autumn NAO	Time	**439.1**	Autumn NAO	None	452.2	439.9	452.3
Ure	Time, annual rain	None	542.8	Annual rain	None	**538.9**	558.6	557.1
Little Ouse	Winter rain	Time	**314.2**	Annual rain	None	320.5	334.4	337.2
Itchen	Time, winter rain	None	264.8	Winter rain	Winter rain	**263.2**	294.2	293.8
Exe	Time	Annual rain	**659.5**	Winter NAO	None	662.7	665.2	667.1
Wye	Time, annual rain	None	**754.2**	Annual rain	None	775.6	769.6	792.2
Cothi	Time, annual rain	None	613.5	Annual rain	None	**612.5**	628.5	628.8
Eden	Time, annual rain	None	638.9	Annual rain	None	**635.0**	659.4	658.2

There are some interesting patterns that emerge from the results in Table B-3. Firstly, it is generally the case that the location parameter includes more covariates than the scale parameter. Annual rainfall is by far the most commonly chosen physical covariate. It is the preferred physical covariate in 10 out of 16 models that include physical covariates (2 per catchment, one with and one without water year included as a potential covariate). Winter rainfall is chosen in 3 models, autumn NAO in 2 (for the Bedburn Beck catchment) and winter NAO for the other one. There are no cases where EA is chosen. On the basis of these results, the project team suggests that the covariates representing atmospheric circulation patterns are no longer considered. As discussed in section 3.3.4, it is desirable to have a model that uses the same covariates at all locations.

In fact, there is an argument for retaining only annual rainfall as a single physical covariate worth considering at a national scale. On the Bedburn Beck, Itchen and Exe, the models using annual rainfall as a covariate achieve a BIC extremely close to the models with the lowest BIC reported above.

There are no cases where the model with only time as a covariate has a lower BIC than any model that uses a single physical covariate. For half of the catchments, the lowest BIC model is one that includes both water year and a physical covariate. For the Bedburn Beck catchment, however, the physical covariate appears to add little to the quality of model fit, since the model with only water year achieves a similar BIC. This is

as expected since the non-stationarity is thought to be due mainly to changes in tree cover.

On most catchments, there is little difference in BIC between a stationary model and one that uses water year as a covariate (the exceptions being the Wye and Bedburn Beck catchments). The Itchen is a case in point. The Itchen is the catchment that shows the greatest benefit (in terms of lowering BIC) from including physical covariates, probably because the groundwater dominance results in a stronger link between long-term rainfall accumulations and peak flow.

In many cases, there is also only a small difference in BIC between models that consider only physical covariates and those that consider both physical and time-based covariates. As previously discussed, if time is excluded as a covariate to model a non-stationary flood series, it would be preferable to include at least one physical covariate in the model which ideally exhibits some trend. The covariate that gives the lowest BIC is generally annual rainfall, and yet over most of the catchments this does not show a significant trend during the period of record. In this case, a model fitted using only annual rainfall as a covariate gives conditional return levels that show no evidence of trend, although the marginal return levels differ from the stationary estimates, at least for longer return periods.

This type of model, with no time-variation of parameters, has the theoretical advantage of not having to resort to crude extrapolation to estimate return levels for a future period. It could, in theory, be applied together with an estimated future distribution of annual rainfall, from a climate model. However, as discussed earlier, this sort of application would only be valid in the case of a causal relationship between annual rainfall and peak flow. Future changes in peak flow would need to be entirely explained by changes in annual rainfall. Although climate change may be expected to affect annual rainfall, and therefore catchment wetness, it can also be expected to influence other factors that control flood magnitudes, including storm intensity and evapotranspiration (which influences soil moisture). These influences are not included in the current non-stationary statistical models considered.

In conclusion, physical covariates (generally, annual rainfall) appear to add useful information to the models on all trial catchments. As hoped, the physical covariates are helping to remove some of the year-to-year variability in AMAX floods, allowing better time-based trends to be better identified. In some cases, however, the model performance is not much improved compared with a model fitted only with water year as a covariate. Because time is retained as a useful covariate on some catchments, it is not advisable to apply those models outside the observed period. There are reasons for being similarly cautious in extrapolating even the models based only on physical covariates.

If covariates representing atmospheric circulation patterns were to be ignored, there would no longer be a need to restrict the analysis to the period starting in 1950. Some of the trial catchments have AMAX flow data from earlier decades that could be incorporated. This would require data sets to be reworked and reanalysed, and so was not attempted.

Model results for trial catchments

Some example results are provided in the practitioner guidance.

Appendix C: Exploring methods of pooled non-stationary analysis

C.1 Outline of problem

The current version of the FEH statistical method (Kjeldsen, Jones, and Bayliss 2008) can be used to assign return periods or average recurrence intervals to peak flows at any location on the digital river network of the UK. The method is based on the index flood technique, where the index variable is the median annual flood (QMED), and this is combined with a statistical growth curve relating flood peak to return period. For ungauged sites, QMED can be estimated via a regression equation based on hydrological catchment descriptors, while the dimensionless growth curve is derived from a flexible regionalisation procedure within which annual maximum flow data (AMAX) from hydrologically similar sites (pooling groups) are pooled together.

This method does not take into account any changes in the flood regime. More specifically, the concept of the index flood, and the methods for estimating it, assume stationarity in the average behaviour of the AMAX series, as does the choice of probability distribution, the generalised logistic distribution (GLO). Secondly, the growth curve method makes use of L-moments (Hosking and Wallis 1997) which rely on stationary time series. Finally, the hydrological similarity distance metric used in the formation of pooling groups does not account for variability in flow regime change spatially.

It would be beneficial to know how and where to apply non-stationarity within the pooling group framework. This includes how to account for trend when forming pooling groups, and what to do in terms of calculating flood frequency curves, index floods and growth curves, incorporating trend once the pooling group has been formed.

The present work focused on 2 aspects. Firstly, it investigated whether including trend as a criteria for pooling groups can improve current stationary estimates of T-year floods. To do this, a new similarity distance metric (SDM) to determine pooling group members was investigated which included descriptors of trend. This was developed in a similar way to the previous SDM. It was then compared to the currently used SDM, and previously published results. Secondly, the use of non-stationarity was investigated for its usefulness in computing index floods and growth curves to improve pooled estimates of flood frequency. To achieve this, new pooling group methods were trialled on real and synthetic data, drawing on various methods currently in peer-reviewed literature.

Outline of work

This report consists of 2 parts: developing a new similarity distance metric, and comparing choices of index flood method and growth curve generation. Both compare the developed methods to existing approaches to check for improved performance. Figure C-1 demonstrates the pooling group method graphically.

The first part investigates including trend descriptors such as the Theil-Sen estimator (Sen 1968) as components in the similarity distance metric (SDM) currently used in the UK to determine pooling groups for estimating regional flood frequency.

In previous work, the SDM was constructed using a linear regression to model L-CV to select components. Components were sequentially weighted to minimise a pooled uncertainty measure (PUM). The new SDM is compared to the old SDM using the PUM as a performance metric.

The second part of the current work looks into the most appropriate method of using index flood methods and growth curve formulations within UK pooling group methods to improve estimates (or confirm the absence of) trend at locations with short or no gauged records of flow, and generally to improve the estimates of long return period flood events.

To do this, the work compares a set of different approaches to applying a non-stationary index flood method to pooling groups (using both the current pooling SDM and that devised in the present work). To obtain the non-stationary index flood, a time-varying location parameter (for GLO) is used. For the non-stationary pooled growth curve, both location and scale were modelled as potentially changeable over time. Parameters were estimated using maximum likelihood methods. Since the 'true' flood frequency curves are unknown, it is useful to offer a method that can show how well methods perform on an artificial data set. Therefore, the pooling methods were trialled on the simulated data set to give a baseline to compare methods. The existing method was also compared to the new methods on the real observed data set.

Technical points

Non-stationary generalised logistic distribution

Throughout, a modified version of the generalised logistic distribution (Hosking and Wallis 1997) will be used which allows for the parameters to vary over time. This leads to the following formulation of the cumulative distribution function of flow $Q(t)$:

$$F(x,t) = Prob[Q(t) \leq x] = \left(1 + \left(1 - \frac{\kappa(t)}{\alpha(t)}\left(x - \xi(t)\right)\right)^{\frac{1}{\kappa(t)}}\right)^{-1} \qquad (1)$$

where the location parameter $\xi(t)$, the scale parameter $\alpha(t)$, and the shape parameter $\kappa(t)$ are functions of time which take the following forms:

$$\xi(t) = \xi_0 + \xi_1 t$$

$$\alpha(t) = \exp(\alpha_0 + \alpha_1 t)$$

For this piece of work, the shape parameter $\kappa(t) = \kappa$ is taken to be a constant value over time, since the associated uncertainty linked to estimating the shape parameter is much greater than that associated with the other 2 parameters. The short length of the AMAX records available makes estimating a non-stationary shape parameter with any useful amount of uncertainty unachievable in this work.

Other covariates could be used in place of or as well as time. The present work does not include these in order to focus on one issue at a time; alternative covariates are discussed elsewhere in this project.

Non-stationary return periods

Flood frequency curves in all parts of this work are plotted by fixing a reference year, and computing return periods based on an expected number of events definition (Salas and Obeysekera 2014). A reference year, y0 is the year against which extreme events are estimated. For example, the 100-year flood measured against reference year 2000 is the flood expected to occur once between 2000 to 2099.

In this case, the 'T-year event' is defined as the largest event which is expected to be exceeded just once in the next T years. In practice, the 'T-year event' depends on the reference year selected, and the length of time for which any particular model is appropriate; that is, for the length of time one can possibly expect the catchment to follow the modelled trends. For example, one would not expect a linear growth in extreme flows indefinitely, and so such a model would not be valid for all time.

For the present work, the size of the T-year event relative to a reference year y0 is computed to be the value Q which solves the equation:

$1 = \sum_{t=0}^{T}\left(1 - F(Q, y_0 + t)\right)$ (2)

where F is the cumulative density function for the GLO in year (y0 + t). Under stationary conditions where no trend is present, this agrees with the stationary definition of a T-year return period event.

Theil-Sen estimator

The Theil-Sen estimator of slope (Sen 1968) describes the average of the set of pairwise slopes. For a data set $(t_i, Q_i : i = 1, \dots N)$ with all different values of Qi, the Theil-Sen estimator of the slope of $Q = (Q_1, \dots, Q_N)$ is given by:

$\beta = TSE = median\left\{\left(\frac{Q_j - Q_i}{t_j - t_i}\right) : i \neq j = 1, \dots, N\right\}$ (3)

In words, the TSE is the median of all pairwise slopes between all points with different times. This is typically seen as more robust to outliers than a standard linear regression of flow against time. It has been used in several other studies, such as those analysing trends in the UK Benchmark Network (Harrigan and others, 2017). It is implemented in this section using the mblm R package (Komsta 2019).

For the present work, the focus is on the normalised TSE (TSE_{norm}) which uses $Q = AMAX/QMED$. This allows simple comparison between sites of different sizes, and gives a value of 'average percent change relative to QMED.'

Data used

The data used in this work are the same as for the rest of this project. Within this data set, catchment descriptors are as published on the FEH Web Service (fehweb.ceh.ac.uk) and the National River Flow Archive (NRFA) website (nrfa.ceh.ac.uk). AMAX records are primarily from the NRFA with some additions, alterations and exclusions from measuring authorities and other parties. Each record was considered over the period 1977 to 2016, with only the 296 stations used which were suitable for flood frequency analysis and had fewer than 2 missing values in the 1977 to 2016 period.

The main catchment descriptors considered for pooling are outlined in Table C-1. In addition to this, the Theil-Sen estimator has been computed for each catchment for each station, both normalised and unnormalised.

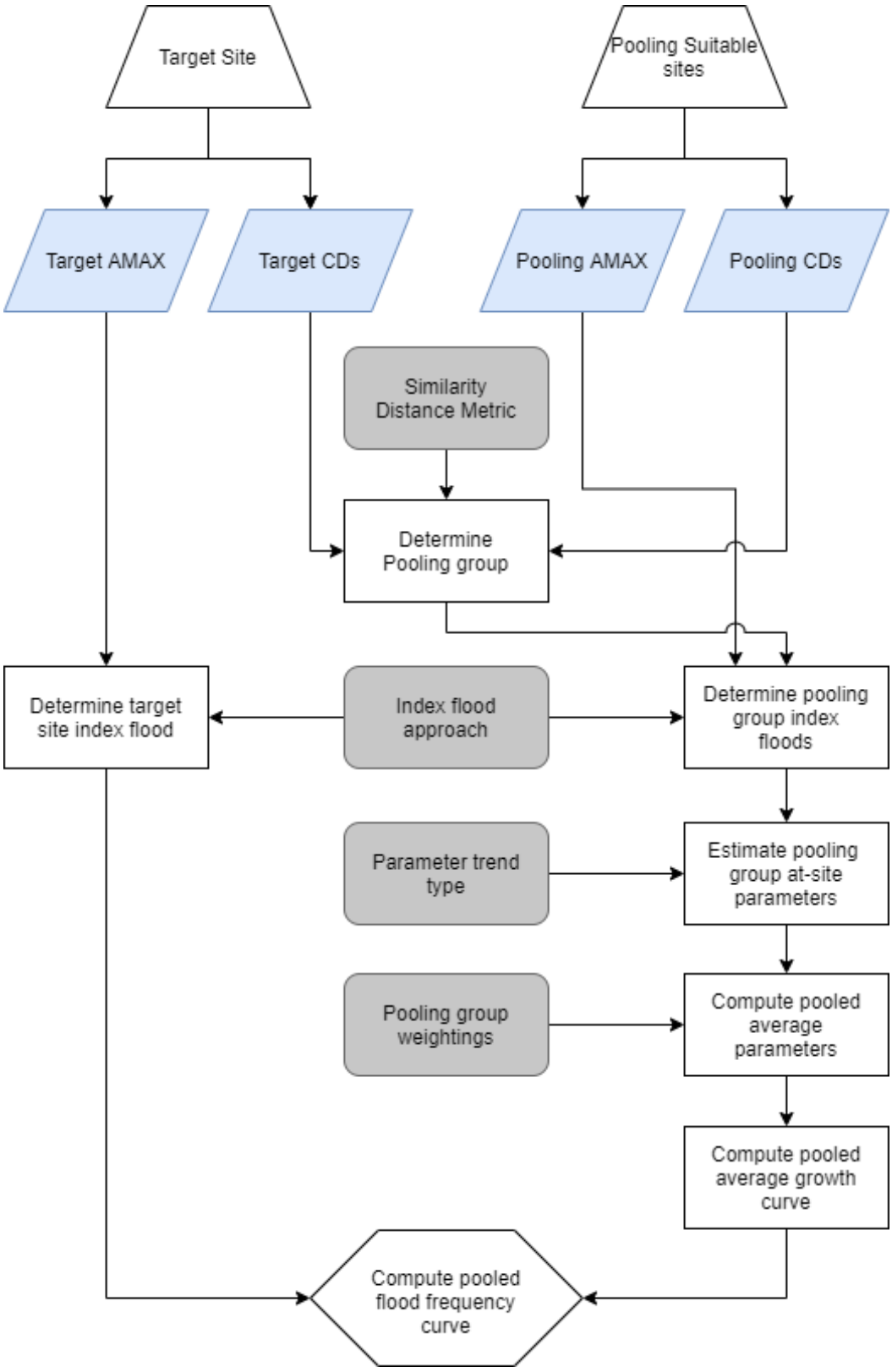


Figure C-1 Flowchart indicating pooling group index flood method, incorporating non-stationarity. Blue regions indicate data, grey regions indicate choices to be made by the practitioner

Table C-1 Summary of catchment descriptors used in this study

Descriptor	Units	Range	Definition
AREA	km^2	$x \geq 0$	Catchment area (as defined by DTM)
SAAR	mm	$x \geq 0$	Standard-period average annual rainfall (1961 to 1990)
FPEXT	-	$0 \leq x \leq 1$	Flood plain extent as fraction of catchment
FARL	-	$0 \leq x \leq 1$	Flood attenuation due to lakes and reservoirs
BFIHOST	-	$0 \leq x \leq 1$	Baseflow index derived from HOST soil classes
SPRHOST	-	$0 \leq x \leq 100\%$	Standard percentage run-off derived from HOST soil classes
$URBEXT_{2000}$	-	$0 \leq x \leq 1$	Combined fraction of urban and suburban within catchment, based on Land Cover Map 2000
DPSBAR	m/km	$x > 0$	Mean catchment drainage pathway slope
DPLBAR	km	$x > 0$	Mean channel length to catchment outlet
PROPWET	-	$x > 0$	Proportion of time soil moisture deficit is below 6 mm (1961 to 1990, based on MORECS)
CEasting CNorthing	m		Catchment centroid easting (northing) as derived from British National Grid
TSE	-	-	Theil-Sen estimator based on a given length of record

C.2 Developing new similarity distance metrics

Currently, the FEH pooling method in England and Wales groups stations of sufficient quality (as determined by the NRFA and measuring authorities) by hydrological similarity, measured using a weighted average of area, average annual rainfall, flood plain extent, and attenuation due to storage. This allows the generation of pooling groups of catchments with similar coefficients of variation (L-CV) and skew (L-SKEW). It does not, however, account for possible trends in the AMAX data at each station.

This section investigates the effect of including trend in estimating long return period events by using pooling groups. To this end, a new metric to measure similarity of catchments is developed, one which includes trend as a component. This is then compared to the currently recommended version (Kjeldsen and Jones 2009) in terms of estimating the at-site Theil-Sen estimator and in terms of estimating the 20, 50 and 100-year return period events.

Similarity distance metrics

The current method for forming pooling groups uses FEH catchment descriptors and transformations of them (for example, log(AREA)) to compute the similarity between catchments. The SDM formulation developed in (Kjeldsen, Jones, and Bayliss 2008) took a weighted average of the distance in difference catchment descriptors between 2 stations, scaled by the variance of that catchment descriptor. The SDM between 2 stations *i* and *j* with *n* descriptors is defined by:

$$SDM(i,j) = \sqrt{\sum_{k=1}^{n} \omega_k \left(\frac{x_i^k - x_j^k}{\sigma_k} \right)^2} \quad (4)$$

where x_i^k and x_j^k are the k'th catchment descriptor values at stations *i* and *j*, ω_k is the weight of the *k*'th catchment descriptor and σ_k is the variance of the *k*'th catchment descriptor.

Catchment descriptors to include in the similarity metric were selected by applying linear regression models to L-CV (the ratio of the second and first L-moments) and L-SKEW (the ratio of the third and second L-moments) to determine the best catchment descriptors for explaining the variance in these L-moment ratios. Despite the relatively poor performance of these linear regression models in the original work (Institute of Hydrology, 1999) (R^2 = 0.375 for L-CV, and R^2 < 0.09 for L-SKEW), this led to the following catchment descriptors being selected: log(AREA), log(SAAR), FPEXT and FARL for the SDM.

To determine the weights (ω_k in equation (4)) to associate to each catchment descriptor, the pooled uncertainty measure (PUM) was used to compare pooled and at-site estimates of various points of the growth curve. The pooled uncertainty metric for a set of M catchments for a return period of T years is defined by equation (5):

$$PUM_T = \left(\frac{\sum_{i=1}^{M} h_i \left(\log z_{T,i} - \log z_{T,i}^{(P)} \right)^2}{\sum_{i=1}^{M} h_i} \right)^{\frac{1}{2}} \quad (5)$$

where h_i are pooling group member weightings, $z_{T,i}$ are at site-*i* growth factors for the *T*-year event, and the superscript (*P*) indicates a pooled estimate for the same. In Kjeldsen, Jones, and Bayliss (2008), the SDM coefficients were determined sequentially to minimise uncertainty. This lead to the final similarity metric of:

$$SDM\ (i,j) = \sqrt{3.2 \left(\frac{\ln AREA_i - \ln AREA_j}{1.28} \right)^2 + 0.5 \left(\frac{\ln SAAR_i - \ln SAAR_j}{0.37} \right)^2 + 0.1 \left(\frac{FARL_i - FARL_j}{0.05} \right)^2 + 0.2 \left(\frac{FPEXT_i - FPEXT_j}{0.04} \right)^2} \quad (6)$$

which is the currently recommended version, and the one implemented in WINFAP 4 (Wallingford Hydrosolutions 2019). It will be denoted SDM_{08} from now on.

To determine the number of catchments required, investigation suggested a '5T rule-of-thumb' (Institute of Hydrology 1999) for the number of station-years (the sum of the length of record in all stations in the pooling group). Subsequent development decided on a uniform 500 station-years for flood frequency estimates up to a return period of 100 years (Kjeldsen, Jones, and Bayliss 2008). The present work uses the fixed 500 station-year requirement to determine the number of stations for a pooling group.

Method

New prospective similarity metrics (SDMs) have been developed using roughly the same methodology as described above. Descriptors were selected to be included based on their importance in linear regression models describing L-CV and L-SKEW. These linear models were determined using a stepwise regression model as implemented in the *leaps* R package (Lumley 2017); for this work the Theil-Sen estimator (TSE) was forced to be included in all models.

For each station consistent values of TSE, L-CV, L-SKEW and stationary estimates of Q_{20}, Q_{50}, and Q_{100} were used throughout model development, and similar patterns were observed in all cases.

The coefficients ω_k were optimised simultaneously to minimise pooled uncertainty, rather than sequentially as was the case in previous work, to avoid an exhaustive search of all possible coefficient combinations. Pooled uncertainty (PUM) was calculated for the 20-year event to determine the final coefficients. This was implemented using *optim* from the base R software (R Core Team 2016), and *BBoptim* from the BB R package (Varadhan and Gilbert 2015).

Throughout this section, stationary estimates for growth curves are used. Section C.3 addresses whether non-stationary growth curves are indeed an improvement.

Results

Covariate selection

Performance of the new models including normalised-TSE as a component was similar to previously documented work; Table C-2 summarises the final components chosen for each prospective similarity metric.

Figure C-2 and Figure C-3 show examples of the subset diagram used to select the models for investigation. Each row shows a prospective model, with shaded squares indicating a component's inclusion. Darker shading shows a higher adjusted R^2 value, labelled on the vertical axis.

Under the stepwise regression models, similar catchment descriptors were obtained when trying to fit a model to explaining L-CV across all record periods. The best in terms of maximising the adjusted R^2 value for each of the choices of record length were FARL, $URBEXT_{2000}$, log(AREA), log(SAAR) and normalised-TSE. This matches up fairly well with the catchment descriptors in SDM_{08}, which highlighted FPEXT rather than $URBEXT_{2000}$. This may be down to the fact that the original development restricted itself to only considering rural catchments, where there is limited variation in $URBEXT_{2000}$. Although TSE was forced to be included in the model by user choice, it produced model coefficients that were not significantly different from zero (at the 10% significance level). When the inclusion of TSE was not enforced, it was never selected for the L-CV model. It would be worth seeing if trend would be included in a covariate if the fitting target (instead of L-CV or L-SKEW) had non-stationarity incorporated in some way.

Including TSE did not seem to reduce the success of the linear models used to describe L-moment ratios. Table C-3 shows the fitting statistics for the final models chosen for each record period. At best, the L-CV model achieved an adjusted-R^2 value of 0.39, which is similar to the value reported in Kjeldsen, Jones, and Bayliss (2008). For comparison, the 2008 model was also fitted to L-CV and reported in Table C-3, showing slightly poorer statistics of fit on the present data set (adjusted-R^2 values of 0.33-0.35).

When trying to fit a model explaining the variance in L-SKEW, it was noted that even the best performing model showed an adjusted R^2-value of less than 5% and so it was expected that these models would not provide suitable SDM components. This fits with Kjeldsen and others (2008) who found that the model explained 8% of the observed variation. However, the model selected was fairly consistent, identifying FARL, BFIHOST, SPRHOST and FPEXT as the best catchment descriptors.

Table C-2 Summary of components in different SDM models

L-CV	**L-SKEW**	**SDM_{08}**
• FARL	• FARL	• log(AREA)
• $URBEXT_{2000}$	• BFIHOST	• log(SAAR)
• log(AREA)	• SPRHOST	• FARL
• log(SAAR)	• FPEXT	• FPEXT
• Normalised TSE	• Normalised TSE	

Table C-3 Fitting statistics for final models chosen for SDM calibration

Model	**Model components**	**Record period**	**R^2**	**RMSE**	**significance of TSE (p-value)**
SDM_{LCV}	FARL, $URBEXT_{2000}$, log(AREA), log(SAAR), TSE	77-16	0.37	0.055	0.353
SDM_{LSKEW}	DPLBAR, BFIHOST, DPSBAR, SPRHOST,TSE	77-16	0.040	0.068	0.975
SDM_{08} (to describe L-CV)	Log(AREA), log(SAAR), FARL, FPEXT	77-16	0.33	0.056	NA

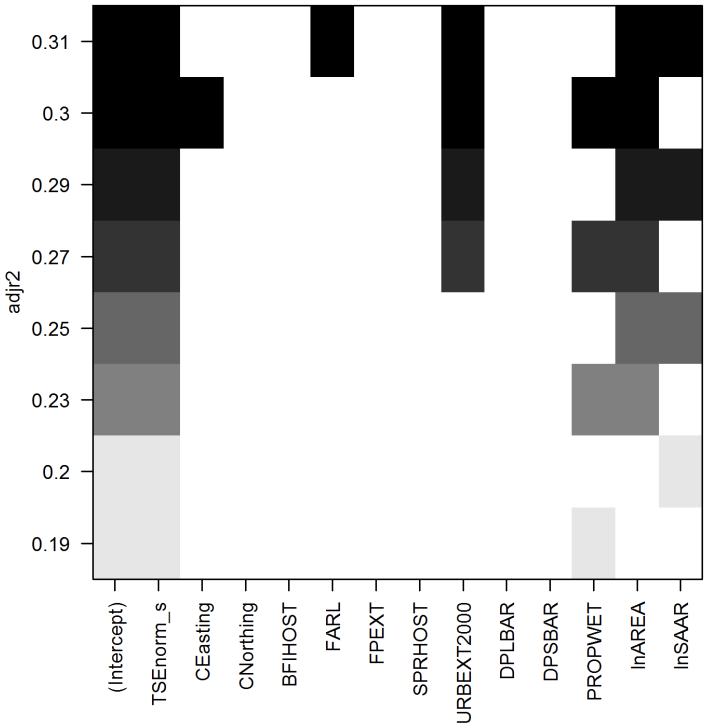


Figure C-2 Subset diagram showing performance of different L-CV regression models under stepwise regression. Shows the top 2 models of each model from 2 to 5 terms, excluding Intercept

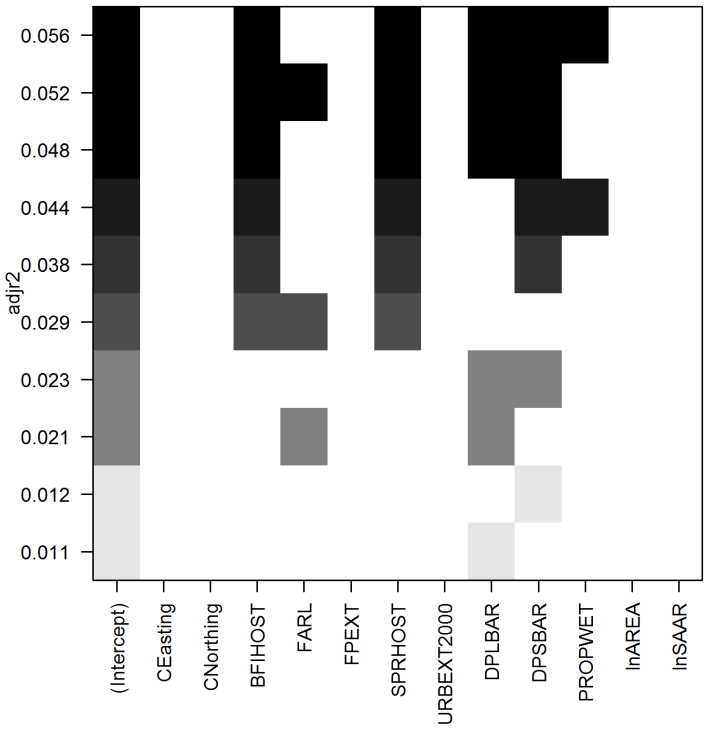


Figure C-3 Subset diagram showing performance of different L-SKEW regression models under stepwise regression. Shows the top 2 models of each model length from 2 to 5 terms, excluding Intercept

Calibration of similarity metrics

Each of the models in Table C-3 was calibrated by selecting coefficient values using the simultaneous coefficient optimisation mentioned earlier. This included the SDM_{08} model, since the new coefficient selection method and the new data set may lead to different results. Pooling group members were weighted (weights h_i, equation (5)) as in (Kjeldsen, Jones, and Bayliss 2008), using weights based on record length. Final estimates for the growth curves were also computed using weighted averages for L-CV and L-SKEW as in Kjeldsen, Jones, and Bayliss (2008), based on similarity of L-moments within each pooling-group.

Table C-4 Pooled uncertainty values for different models using the long record period

Model	PUM (T=20)	PUM (T=50)	PUM (T=100)
L-CV	0.224	0.370	0.540
L-SKEW	0.233	0.384	0.544
SDM_{08}	0.207	0.346	0.503

Table C-4 highlights that the recalibrated SDM_{08} model still outperforms both the L-CV and L-SKEW models in terms of pooled uncertainty. For longer return periods, pooled uncertainty increases in all cases. All 3 models give very similar results under the different record periods, so results from other record periods are not presented here. Equations (7) to (9) show the calibrated similarity metrics based on the 20-year PUM calibration, using the long record period (1957 to 2016). AREA and FARL are weighted strongly where they appear (as for SDM_{08}), but TSE is also weighted similarly strongly for the L-CV model. The recalibrated version of SDM_{08} shows that the difference in the method of optimisation does not have a significant effect on the final results; compare to equation (6).

Since the L-CV model performs similarly to the L-SKEW model in terms of pooled uncertainty, one could argue that they both perform well in determining pooling groups. However, the poor performance in selecting the L-SKEW model suggests that the similarity metric generated from this model may not work as well in practice, as the L-SKEW linear model is almost no better at prediction than a constant value. Additionally, the increased measurement uncertainty inherent in estimated values like BFIHOST and SPRHOST means that including these terms may not be appropriate in regions where baseflow and soil data are poor.

$$SDM_{LCV}(i,j) = \sqrt{\begin{aligned} &0.592\left(\frac{FARL_i - FARL_j}{0.044}\right)^2 + 0.882\left(\frac{URBEXT_i - URBEXT_j}{0.086}\right)^2 + 2.230\left(\frac{\ln(AREA)_i - \ln(AREA)_j}{1.29}\right)^2 + \\ &1.194\left(\frac{\ln(SAAR)_i - \ln(SAAR)_j}{0.367}\right)^2 + 0.035\left(\frac{TSE_{norm,i} - TSE_{norm,j}}{0.0063}\right)^2 \end{aligned}} \quad (7)$$

$$SDM_{LSKEW}(i,j) = \sqrt{0.990\left(\frac{DPLBAR_i - DPLBAR_j}{0.217}\right)^2 + 1.03\left(\frac{BFIHOST_i - BFIHOST_j}{0.164}\right)^2 + 1.02\left(\frac{DPSBAR_i - DPSBAR_j}{0.053}\right)^2 + 0.990\left(\frac{SPRHOST_i - SPRHOST_j}{0.110}\right)^2 + 0.962\left(\frac{TSE_{norm,i} - TSE_{norm,j}}{0.0063}\right)^2}$$

(8)

$$SDM_{08}(i,j) = \sqrt{2.94\left(\frac{\ln AREA_i - \ln AREA_j}{1.29}\right)^2 + 0.46\left(\frac{\ln SAAR_i - \ln SAAR_j}{0.367}\right)^2 + 0.11\left(\frac{FARL_i - FARL_j}{0.044}\right)^2 + 0.49\left(\frac{FPEXT_i - FPEXT_j}{0.0017}\right)^2}$$

(9)

Variations in Theil-Sen estimators within pooling groups

Here, the pooled average Theil-Sen estimators are compared to the at-site versions, based on the pooling groups derived from SDM_{08} and SDM_{LCV}. In the interests of clarity, and based on the observations above, the L-SKEW-based similarity metric is not pursued further. The weightings of pooling group members (h_i, equation (5)) are based on record length:

$$h_i = \frac{n_i}{1+\frac{n_i}{16}} \quad (10)$$

where n_i is the record length at site *i*, and TSE pooled estimates are calculated using these weights. This is based on discussion in Kjeldsen, Jones, and Bayliss (2008).

In Figure C-4 and Figure C-5 it can be seen that the results are fairly similar between the 2 similarity metrics SDM_{LCV} and SDM_{08}. As expected, the pooled estimates of normalised TSE are much smaller than the large at-site estimates, as the significant trends are many fewer in number than the non-significant trends. Additionally, in general one sees less intragroup variance in TSE and less difference between at-site and pooled TSE estimates when using SDM_{LCV}. There is more intragroup variation in the north-west, and more difference between the at-site and pooled estimates of trend in this area. The effect is less pronounced as one moves south and east. Additionally, it can be seen that many of the pooled estimates gave positive trend for both SDM_{08} and SDM_{LCV}. Note that most of these pooled TSE estimates are **not** significant trends over the 1977 to 2016 period. Due to the lack of significant negatively trending stations, pooling groups typically consist of near-zero trend station, leading to near-zero pooled estimates of trend across the region.

The more geographically consistent patterns in variance and accuracy when using the L-CV similarity metric match up with some opinions that positive trend is more likely to be observed in the north-west.

See additional note 1 for a brief investigation into the prospect of a catchment descriptor SDM to make pooling groups with matching TSE, making use of Easting and Northing.

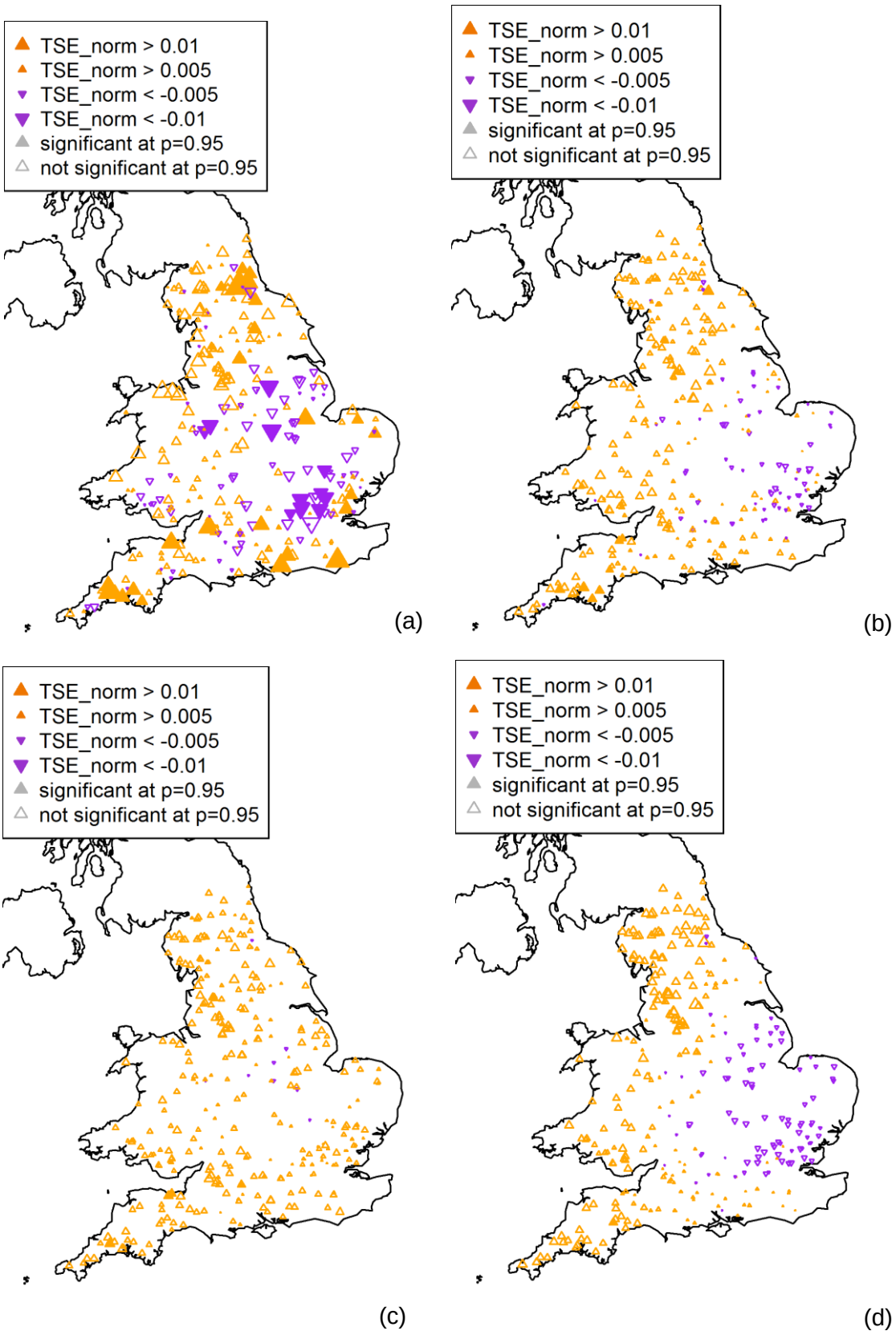


Figure C-4 a) At-site normalised TSE, (b) Pooled normalised TSE using SDM08, (c) Pooled normalised TSE using new SDMLCV, (d) Pooled normalised TSE using TSE distance metric (see additional note 1)

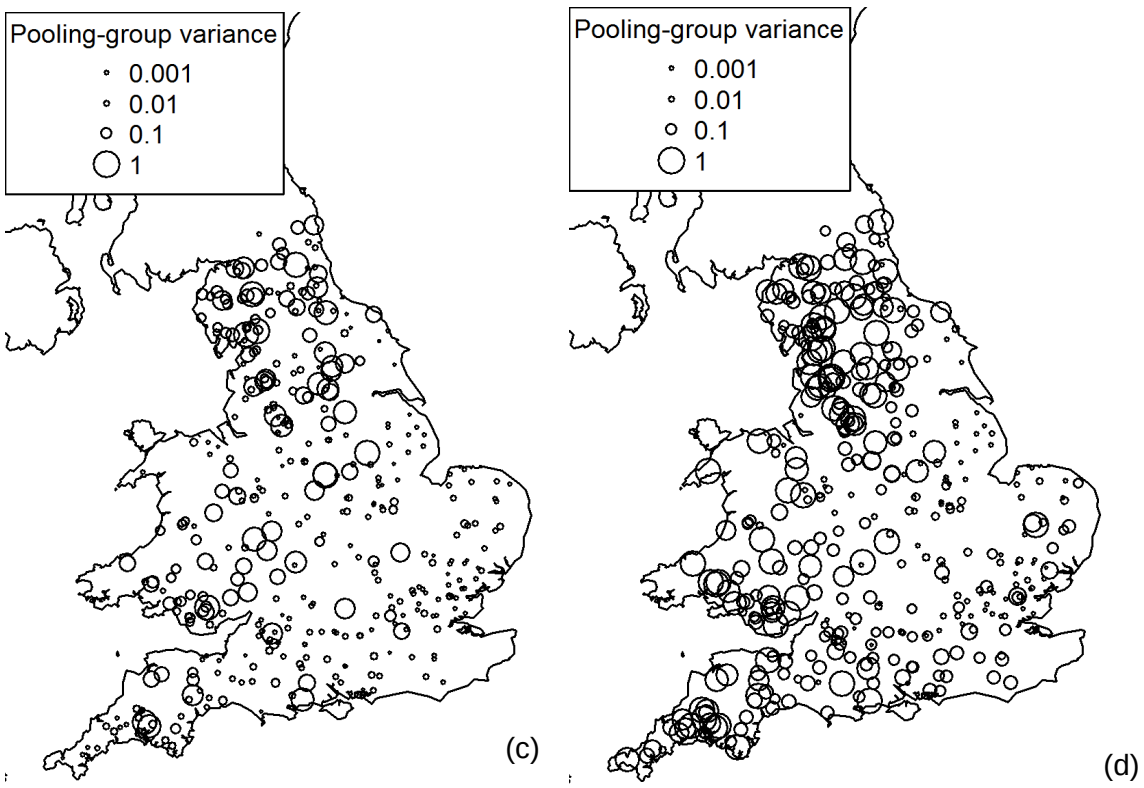


Figure C-5 Within-pooling-group variance of Theil-Sen estimators using (left) SDMLCV and (right) SDM08 metrics, based on 1977 to 2016 data

Example pooling group

To demonstrate the proposed similarity metric as calibrated on the L-CV model (SDM_{LCV}), it is applied to a specific station. The pooled parameter estimates and T-year event estimates are reported for the different methods. This includes the FEH at-site estimate, and both the SDM_{LCV} and SDM_{08} including the target site, and excluding it (treating it as ungauged).

Here, it is applied to NRFA station 76005, Eden at Temple Sowerby. The single-site estimates are based on the 1977 to 2016 period, and use SDM_{08} to determine the pooling group. The results are summarised in Table C-5. In all cases, the growth curve location parameter ξ is given, rather than QMED, for comparability.

For this specific example, compared to the single-site, the parameter estimates when the target site is gauged are fairly close under both choices of SDM. For both choices of similarity metric, the ungauged estimates are slightly worse, as expected, particularly for the scale parameter. The $SDM_{LCV,}$ however, seems to offer closer estimates to the at-site analysis. Note this is only one example, other locations show much worse estimates for both SDMs when considering the site as ungauged.

The Gini coefficient reported in Table C-5 describes the homogeneity of the pooling group: how similar the pooling group members are to each other based on the value of L-CV. (0 is perfect homogeneity, 1 is total heterogeneity). It can be seen that ungauged pooling groups are typically less homogeneous than the gauged pooling groups. This can be thought of as pooling group members being similar to the target, but not

necessarily each other. Between choices of SDM, the Gini coefficient is fairly similar, but is smaller for the SDM_{LCV}.

Table C-6 and Table C-7 show a comparison of the constituent catchment descriptors in the pooling groups formed under the 2 methods (including the target site). A large overlap is clearly visible, although difference, particularly in ordering in terms of SDM (decreasing SDM from top to bottom) is clear. Smaller, drier catchments appear to rank higher, possibly due to the emphasis on like trend in the SDM.

Table C-5 Summary of parameters and significant growth curve estimates starting from 2000.

Method	ξ	α	κ	Q_{20}	Q_{50}	Q_{100}	Gini of LCV
Single site	1.006	0.212	-0.315	2.04	2.63	3.19	NA
SDM_{08} (gauged)	1.011	0.210	-0.245	1.91	2.37	2.79	0.180
SDM_{08} (ungauged)	1.008	0.215	-0.237	1.93	2.38	2.80	0.200
SDM_{LCV} (gauged)	1.010	0.205	-0.257	1.91	2.38	2.81	0.147
SDM_{LCV} (ungauged)	1.012	0.219	-0.259	1.98	2.48	2.95	0.148

Concluding remarks on developing a new similarity distance metric

This section investigated the scope for including trend as a catchment descriptor to inform pooling group formation, and its effects on pooled estimates of trend and long return period events. Overall, the work seems to suggest that although including trend provides a slightly more spatially consistent estimate of trend, the estimates for parameters and flood frequency curves are not improved. This lack of conclusive evidence towards using a new similarity metric suggests that, for the time being, the currently published SDM_{08} is still recommended for generating pooling groups, even in the presence of trend. This should be compared to other literature on trend. O'Brien and Burn (2014) conclude that, "The results indicate that there is less uncertainty in quantile estimates found through the application of the trend centred pooling approach when compared to a regional stationary analysis of the same regions."

It does, however, show the potential for improvement. This could be further extended to include making use of non-stationary return periods and using different weighting methods. Additionally, it would be useful to use some sort of spatial model to estimate trends at an ungauged location, and reanalyse the pooled data.

Using peaks-over-threshold data in computing at-site trends may also be worth investigating, as previously lost 'second-biggest' floods may provide valuable extra data into quantifying trend.

Table C-6 Pooling-group for 33034 from SDM_{LCV}

Station	AREA	BFIHOST	FARL	FPEXT	SAAR	SPRHOST	$URBEXT_{2000}$	PROPWET	TSE	TSE_{norm}
22001	578.25	0.393	0.993	0.0403	850	42.53	0.002	0.44	1.039	0.006
27029	340.75	0.455	0.931	0.0256	1257	38.53	0.0477	0.57	1.364	0.010
27043	430.01	0.366	0.975	0.0351	1385	46.68	0.0036	0.62	0.306	0.001
43008	448.17	0.937	0.976	0.0518	830	6.92	0.0102	0.35	-0.243	-0.002
45001	608.16	0.526	0.985	0.0314	1249	36	0.0061	0.46	-0.150	-0.0008
53008	305.17	0.622	0.988	0.093	804	27.97	0.009	0.34	0.162	0.004
56001	913.25	0.597	0.98	0.0445	1367	28.86	0.0064	0.56	0.594	0.002
71011	203.18	0.382	0.998	0.0987	1446	46.04	0.0071	0.61	0.090	0.0007
76005	618.21	0.474	0.998	0.06	1142	37	0.0039	0.66	1.727	0.006
76017	1371.7	0.509	0.955	0.0615	1273	36.89	0.0049	0.65	3.297	0.007

Table C-7 Pooling-group for 33034 from SDM_{08}

Station	AREA	BFIHOST	FARL	FPEXT	SAAR	SPRHOST	$URBEXT_{2000}$	PROPWET	TSE	TSE_{norm}
27007	912.58	0.42	0.981	0.0674	1120	43.66	0.0078	0.41	2.685	0.010
27028	687.01	0.408	0.968	0.0586	1048	38.51	0.1051	0.49	0.231	0.002
27034	510.94	0.386	0.99	0.0452	1337	46.93	0.0043	0.63	1.544	0.006
43008	448.17	0.937	0.976	0.0518	830	6.92	0.0102	0.35	-0.243	-0.002
50002	664.26	0.425	0.996	0.0496	1184	40.52	0.0036	0.49	1.367	0.005
54005	2026.77	0.47	0.977	0.0919	1147	38.49	0.0042	0.5	1.338	0.004
54008	1124.62	0.612	0.994	0.0635	841	28.53	0.006	0.36	0.477	0.003
56001	913.25	0.597	0.98	0.0445	1367	28.86	0.0064	0.56	0.594	0.001
76005	618.21	0.474	0.998	0.06	1142	37	0.0039	0.66	1.727	0.006

C.3 Non-stationary pooling methods

This section looks into the most appropriate method of using pooling groups within the UK to improve estimates of flood frequency in the presence of potential non-stationarity in flow regimes.

To do this, a set of different approaches to applying a non-stationary index flood method to pooling groups were investigated (using the SDM_{LCV} as it includes at-site trend as a pooling covariate). The methods which were compared are:

1. (ALLSTA) Using a stationary index flood and stationary growth curve.
    - This is the existing stationary method that is compared against.
2. (NSTGC) Using a stationary index flood and non-stationary growth curve.
    - This assumes trends are regional in nature, and so the description of at-site index floods can be kept simple.
3. (NSTIF) Using a non-stationary index flood and stationary growth curve.
    - This makes the assumption that trends are site-specific and only observed in median behaviour. Therefore, once this at-site trend is accounted for, the dimensionless growth curve is regionally consistent and stationary.

4. (ALLNST) Using a non-stationary index flood and non-stationary growth curve.
    - This is the most generalised form of including non-stationarity; all the above are special cases of this one. It allows for regional patterns in trend, along with at-site variability of this trend.

Figure C-6 provides a graphical summary.

These 4 methods are compared using the 381 stations appropriate for flood frequency estimation. The focus is on the 'short fixed period' of 1977 to 2016. Only stations with fewer than 2 missing values are included.

A simulation study using artificially constructed pooling groups of theoretical stations has also been carried out to test the methods, to allow the assumption that the 'true' values of normalised annual maximum series are known. The different pooling methods are used to estimate the at-site GLO parameters, under simulated trend and stationarity.

Literature review

FEH pooling method

The FEH pooling group method, as previously mentioned, was developed by the Institute of Hydrology (1999), and updated by Kjeldsen, Jones, and Bayliss (2008). It uses a stationary index flood and stationary growth curve, and provides stationary estimates for T-year events. Pooled estimates of parameters are based on pooled estimates of L-moment ratios, specifically L-CV and L-SKEW (Hosking and Wallis 1997), weighting L-moment ratios according to station record lengths and intragroup similarity of L-CV and L-SKEW.

This stationary index flood method has been used globally with different distributions, such as the generalised extreme value (GEV) distribution and the Gumbel distribution. The pooling group selection is also frequently referred to in the literature as a 'region of influence' approach. This is an alternative to fixed geographical regions within which growth curves are assumed to be consistent, as in the FSR (1975).

Non-stationary index floods

Several works in academic literature describe regional flood frequency by adopting a non-stationary index flood and a stationary growth curve, typically by including a trend in the location parameter in the distribution used to fit the at-site data, $\xi(t) = \xi_0 + \xi_1 t$, and applying this to a preformed region or group of stations.

Cunderlik and Burn (2003) apply a non-stationary index flood to a group of stations in South British Columbia by fitting a linear trend in the mean and variance and obtaining a pooled shape parameter from the detrended data, finding that neglecting the slight negative trend in mean could lead to overestimation of the 200-year design value by up to 13% by 2020. This was only performed on a single region, however. O'Brien and Burn (2014) analysed 4 regions in Canada, split by regional trend. Linear trends were fitted to the location parameter for generalised normal and GEV distributions. Including trends in scale and shape led to less well fitting models in terms of AIC (a goodness of fit statistic based in information theory, (Akaike 1974)). It is also noted that having pooling group members with at-site parameter estimates with mixed scale trends (that is, some positive, some negative) may lead to an incorrect estimate of trend at-site. Both studies were only involving sites with previously identified significant trends in

AMAX behaviour, and how effective it is to apply trends to an entire country, for example, is unknown.

Non-stationary growth curves

As an alternative to having a non-stationary index flood, one could consider trends to be regional in nature, with non-stationary growth curves scaled by at-site index floods. Similar to the present work, Nam and others (2015) trialled putting a trend in the GEV location parameter of the index-flood, the growth curve, and both in a basic simulation study using Monte Carlo simulation. This suggested that, as record length increased, including non-stationarity in the fitted model reduced model error in RMSE.

Including trend in the index flood and growth curve is a fourth option (Nam and others, 2015). As mentioned above, the rationale behind this is more complex, but allows for large-scale regional trends, with at-site adjustment. Cunderlik and Ouarda (2006) investigate a flood-duration-frequency model which incorporated non-stationarity in index-floods and growth curves, making use of Theil-Sen estimates for a single homogeneous region in Québec over a 30-year period. Significant trends in mean and variance were observed, leading to smaller T-year event magnitudes for the 5-day duration events. When assuming a stationary 'truth', leaving out 10 years of data (leaving estimates based on a 20-year period) gave an overestimate with bias, and RMSE of up to 5%. If one assumes the non-stationary data to be more 'true', then the stationary estimate based on 20 years of data overestimates with bias of up to 29%.

Hanel, Buishand, and Ferro (2009) make use of seasonal global temperature, rather than linear time, to fit trends to. Applying trend to the at-site location parameter and the regional dispersion parameter (ratio of location and scale), 5 regions in the Rhine basin were fitted with non-stationary flood frequency curves for winter and summer AMAX series. Spatial pooling reduced the uncertainty in all the parameters across all the regions by at least 30%. However, with no 'true' value to compare to, as with most of the other studies, the accuracy cannot be well assessed. In practice, fitting large scale trends first across pooling groups is challenging, and has to happen before the index flood is computed.

Methods

Comparing non-stationary pooling methods against the FEH method

The pooling methodology has a number of choices. This includes SDM, the weighting function for pooling group members, and the choices of stationary and non-stationary index floods and growth curves. In this study, the weight function will be restricted to weights scaling with record length (equation (10)), based on scaling methods outlined in Kjeldsen, Jones, and Bayliss (2008). The pooling metric used was SDM_{LCV}, due to its inclusion of TSE, based on the literature strongly suggesting that like-trended catchments are key when computing regional flood frequency estimates with non-stationary models (O'Brien and Burn 2014)

For each of the 4 methods above and summarised in Figure C-6 (ALLSTA, NSTIF, NSTGC, ALLNST), an index flood will be generated, and a pooling group will be generated using the 'normalised' data (AMAX divided by the index flood).

To obtain the non-stationary index flood, a time-varying location parameter (for GLO) is used, $\xi(t) = \xi_0 + \xi_1 t$. These parameters are estimated using normalised TSE for ξ_1 and maximum likelihood estimation methods (MLE) for the other parameters, as L-moments are not currently applicable under non-stationarity; see (Jones 2013) for a discussion of attempting to define non-stationary L-moments. For consistency, MLE

methods are also used for estimating the stationary index flood, rather than using the median (QMED) or the first L-moment. In this case, the stationary index flood is equal to the fitted at-site GLO location parameter (ξ).

For the stationary pooled growth curve, the normalised (and detrended in the case of NSTIF) at-site values are fitted to a GLO distribution using MLE methods. In this case, detrending refers to the fact that the flow is divided by $(\xi_0 + \xi_1 t)$, theoretically removing the trend if such a trend is linear in the original time series. Regional estimates for the growth curve are determined by taking averages of the at-site parameters weighted by a function of record length (equation (10)). This is a diversion from the FEH method where the L-moment ratios are averaged. Since there is no obvious analogue for these ratios in the non-stationary setting, and due to the use of maximum likelihood estimators rather than L-moment estimators, averaging parameters is a common approach in much of the literature. For consistency, all pooled averages are computed this way in this section.

To obtain a non-stationary growth curve, trends in the location and scale parameters are considered using a linear form for the location, and an exponential form, $\alpha(t) = \exp(\alpha_0 + \alpha_1 t)$ for the scale parameter. Growth curves are estimated using the annual maximum series normalised by the location parameter, be that stationary or non-stationary. So, this means that in the non-stationary case, the AMAX flow for each year is normalised by the location parameter for that year Throughout this work, the shape parameter κ is assumed constant in time both at-site and regionally. An alternative where the scale parameter is also stationary at-site and regionally is also considered.

For each choice of index flood and growth curve, pooled-estimate flood frequency curves are compared to at-site estimates (computed for the reference year 2000). The pooled estimates under the current FEH statistical methodology framework are also documented, to verify that the stationary index flood and growth curves coincide with currently accepted methods. Pooling groups in this section consist only of those stations that were both determined suitable for flood frequency analysis in this study and suitable for pooling by the NRFA. Therefore, they may not match pooling groups as generated in practice by, for example, WINFAP 4.

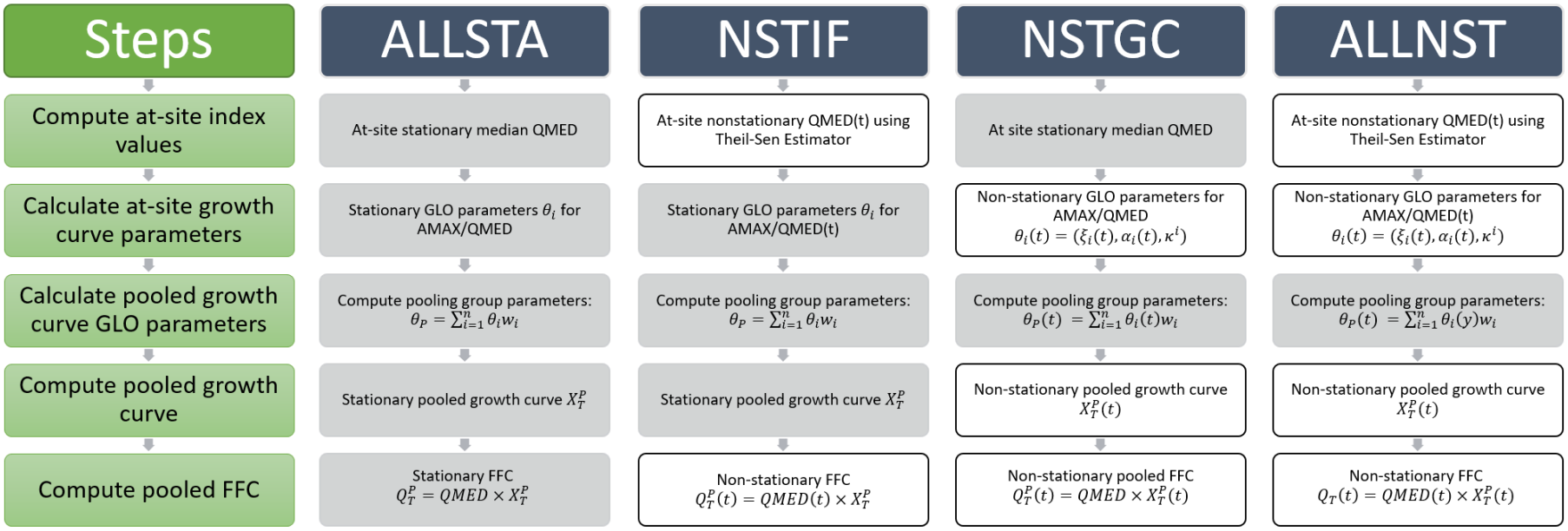


Figure C-6 Step-by-step for the 4 methods. White highlights indicate steps involving non-stationarity

Simulation study

As mentioned above, it is not possible to compare the methods to a known 'true' value when applying them to real data, as the 'true' values are not known. To allow a fairer comparison, the methods presented now are compared to each other when they are applied to an artificial data set. To generate this artificial data set with known trends in the location and scale parameters, a set of simulated pooling groups was generated, and for each pooling-group, the at-site flood frequency curve was compared to the pooled average.

In order to make this simulation worthwhile and appropriate, each simulated pooling group was given a dependence structure which reflects real-world dependence between AMAX series at gauging stations. A simple approach would be to treat the AMAX series for all the stations to be totally independent, but this would be unrealistic as, for example, if a very large storm hits in a particular year, all the stations in that region may record a very high AMAX value for that year. Similarly, if one catchment is nested in another, one might expect both catchments to have similar AMAX series, unless the scale of the catchments is very different.

To achieve this, an empirical copula is used. Copulas are commonly used to describe the relationship between random variables. Under some simple conditions on the functions, every multivariate cumulative distribution function can be decomposed into a copula and a set of marginal univariate distributions (Moore and Spruill 1975). In the case of this project, the team looked at the rankings of the events in each pooling group member's AMAX series. From these, it determined an empirical copula that describes how similar and dissimilar the rankings of the pooling group members are; see Box 1 for details.

The empirical dependence structure copulas were determined from 3 good-quality stations and their pooling groups in different areas of England: north-west England, south-east England and south-west England. Once the copula was estimated, each station within a pooling group was simulated with a 100-year record.

Since only the target site needs to have an index flood, the simulation method only needs to estimate the 'normalised flow'; the flow following scaling by the at-site index flood. Under the index flood method, growth curves are assumed to be regional in nature, so all stations in the pooling group were treated as identically distributed after normalisation. This study used 'true' parameters which are the same for each pooling group. Three alternatives were simulated:

- stationary simulation: all growth curves have $\xi = 1, \alpha = 0.25, \kappa = -0.3$
- positive trend: all growth curves have $\xi(t) = 1 + 0.03t$, $\alpha(t) = \exp(-1.38 + 0.01t)$, $\kappa = -0.3$
- negative trend: all growth curves have $\xi(t) = 1 - 0.007t$, $\alpha(t) = \exp(-1.38 - 0.01t)$, $\kappa = -0.3$.

The simulation tested the following cases, summarised in Table C-8. Pooling groups were simulated under stationary conditions; with a positive trend and with a negative trend. Then estimates were made under stationary and non-stationary assumptions. This gives 6 cases, where stations were correctly or incorrectly classified as having trend.

Plots are shown for the parameter estimates in each case, along with uncertainty estimates (5% and 95% quantiles) and boxplots.

Algorithm for simulation of catchment normalised AMAX values

1. For each timestep with complete observations for the pooling group, calculate the ranks of the scaled observations $R_i(t)$.
2. For each timestep s to be simulated, draw one uniformly at random from the observed timesteps, and note the ranks associated to that time $R_i(s)$.
3. Draw N values $U_1,\ldots, U_N$ from a Unif[0,1] distribution, where N is the number of pooling group members.
4. Order the U_n by size, and assign them to the station with the corresponding rank.
5. Using the simulation distribution for the normalised AMAX values, compute $Q_i(s) = Q(F=U_i, s)$.
6. Repeat for each timestep.
7. Estimate the GLO distribution parameters $\theta_i(s)$ for each site, based on the simulated $Q_i(s)$.
8. Using a record-length based weighting, determine the pooled estimate for the GLO parameters $\theta^*(s)$ at the target site.

Box 1: Details of empirical copula algorithm

Table C-8 Simulation methods and estimation approaches. Parameters are (location, scale, shape) in order. t is number of years since simulation year zero

		Simulated 'true' behaviour with parameters		
		Stationary (1,0.25,-0.3)	Positive trend (1+0.03t, $e^{-1.38+0.01t}$, -0.3)	Negative trend (1-0.007t, $e^{-1.38-0.01t}$, -0.3)
Estimated behaviour	Stationary	Correct	Misclassified	Misclassified
	Non-stationary	Misclassified	Correct	Correct

Results

Comparison of non-stationary pooling methods against the FEH method

For all the stations determined for this study to be suitable for flood frequency estimation (381 stations) the index floods and at-site growth curve estimates were computed for each of the 4 methods outlined at the start of section C.3.

Figure C-7(a) shows the difference between the at-site estimates for a 20-year return period under stationary and non-stationary assumptions (trend in the location parameter). For this short return period, little difference can be seen between the stationary and non-stationary at-site estimates at most locations. This seems reasonable, as even for the most extreme trends, the amount of change in flood regime in 20 years is quite small. To put this in perspective, most stations have a trend parameter ξ_1 of less than 0.01, so one might expect a change in the median flood of 1 m^3/s per 100 years based on this. There are a handful of stations with big differences, both positive and negative. Typically, the stationary estimate is, in some sense, an average of the non-stationary estimates (assuming a linear trend). For longer return periods, the patterns are magnified, but are ultimately similar. Places where positive trend has been observed tend to show an increase in at-site estimate when including non-stationarity; likewise, negative trends occur in the same places as decreased estimates.

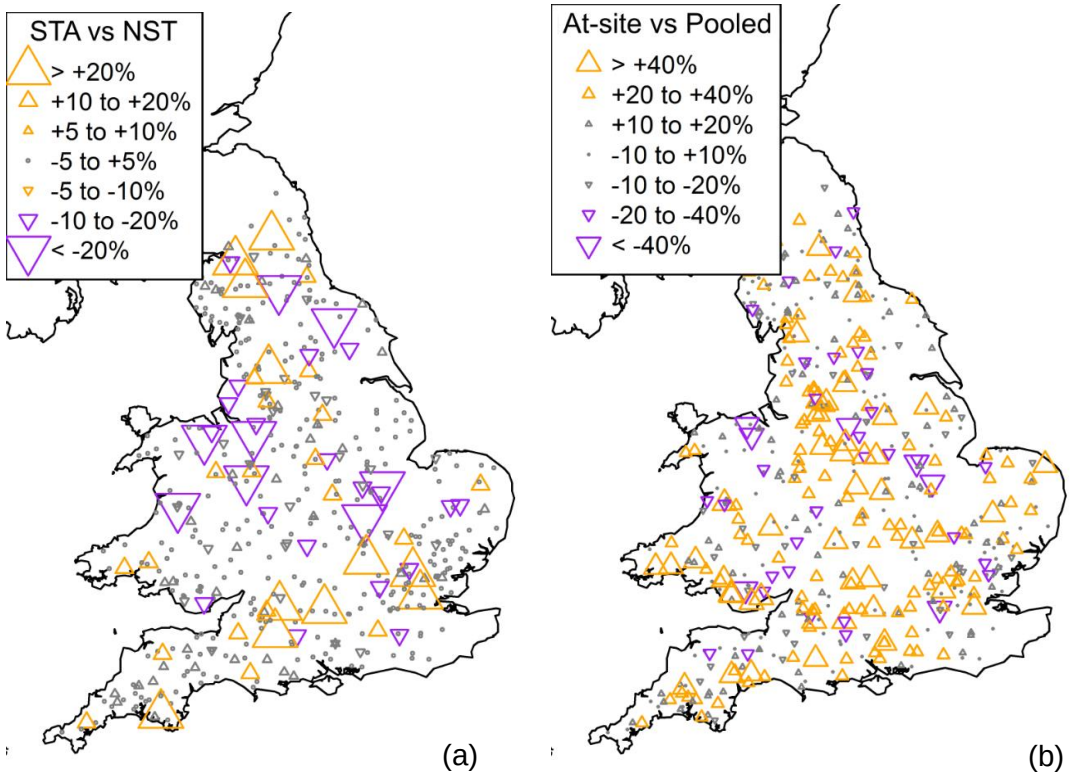


Figure C-7 (a) Comparison between at-site Q20 under stationary and non-stationary calculations. Positive percentage indicates larger estimates from the non-stationary fitting. (b) Comparison between at-site and stationary pooled estimates. Positive values indicate pooled estimates are larger. Please note the different scales in (a) and (b) for readability

Once the pooled estimates are included in Figure C-7(b), one sees a big difference between at-site and stationary pooled estimates of Q_{20} (ALLSTA). The other pooling methods perform relatively similarly to ALLSTA. For NSTGC, 2 versions are presented in Figure C-8: one including a trend in scale and growth curve location (a), and one including a trend in growth curve location only (b). Here, the combined location-scale trends create a greater amount of difference (up to 20%, usually increasing Q_{20}) than the location trend alone, which matches very well with ALLSTA. Including a trend in scale is not pursued further, and is advised against at this time, due to the much greater uncertainty involved in estimating such a trend in scale.

Using a non-stationary index-flood alone (NSTIF, Figure C-8(c)) gives more variable results compared to NSTGC (Figure C-8(b)). Here, there are a lot of smaller estimates, but nearly all are within 15% of the stationary pooled estimate. ALLNST (Figure C-8(d)) only used a trend in the location parameter for index flood and growth curve, and quite neatly combined the differences of NSTGC and NSTIF, leading to more estimates smaller than those under ALLSTA.

This may be due to inconsistencies in the trends present at the sites within the pooling group. However, there may be 'double-fitting' of trends in the location parameter, once at the index flood stage, once at the pooling stage. If this is the case, the uncertainty in the trend estimated in the first stage is compounded by the uncertainty in the second. It may be the case that the 'true' curve is much less extreme. Note that in all cases the members of each pooling group are the same, only the parameter estimates change.

In all methods, the fitted trends in the location and scale parameters were quite sensitive to the record period chosen. Due to the short length of records, there was also high uncertainty in the parameter estimates.

Example station

In this section, the focus returns to the example station on the Eden at Temple Sowerby (76005). Figure C-9 shows the flood frequency curves for the years 1970, 2000 and 2020 under the 4 methods, alongside the at-site estimate. Only key return periods were calculated (10, 20, 30, 50, 75 and 100 years), but for the purpose of presentation, they have been unrealistically connected with straight lines. In reality, one cannot expect a true 'GLO-like' flood frequency curve due to the changing probabilities.

Here, one can see that, for 2020 in particular, the non-stationary index flood has the biggest impact in future flood frequency estimates, with the growth curve actually reducing estimates. ALLSTA (stationary index flood and growth curve) is also, as expected, the closest to the at-site estimates in 1970, but not for the later estimates. The fact that methods NSTGC and ALLNST show more inconsistent growth curves may be partly due to the complexity in a) finding maximum likelihood estimates for the non-stationary parameters and b) in solving the return period equation (equation (2)). See Table C-9 for the parameters used to calculate the flood frequency curves.

It is probable that within pooling groups, positive trends in scale were underestimated and negative trends overestimated; this is investigated in the following section.

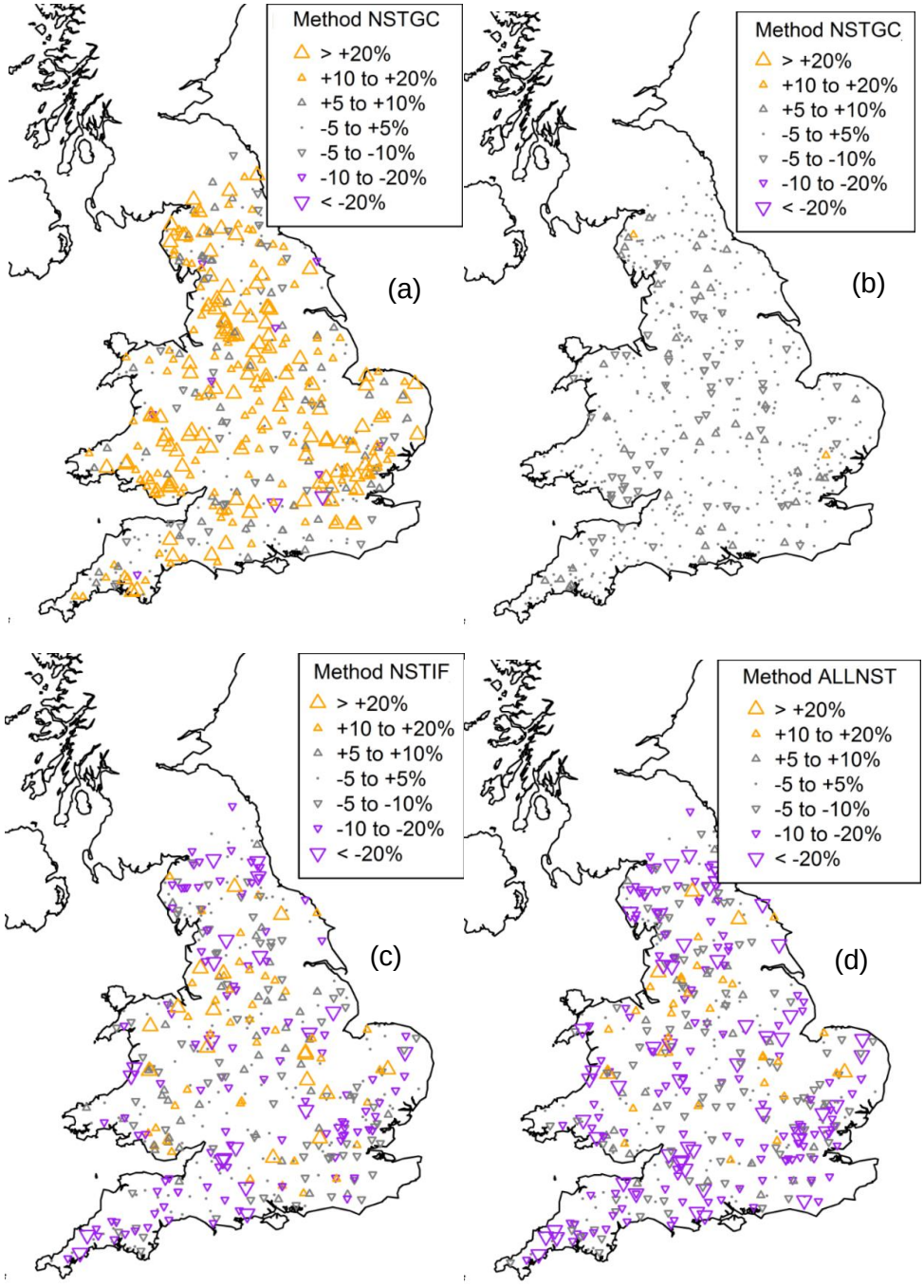


Figure C-8 Comparison of NSTGC, NSTIF and ALLNST against ALLSTA (the standard FEH method), showing percentage difference in pooled estimate of Q20: (a) non-stationary growth curve with varying location only, (b) non-stationary growth curve with varying location and scale, (c) non-stationary index flood with varying location only, (d) non-stationary growth curve and non-stationary index flood with independently varying location

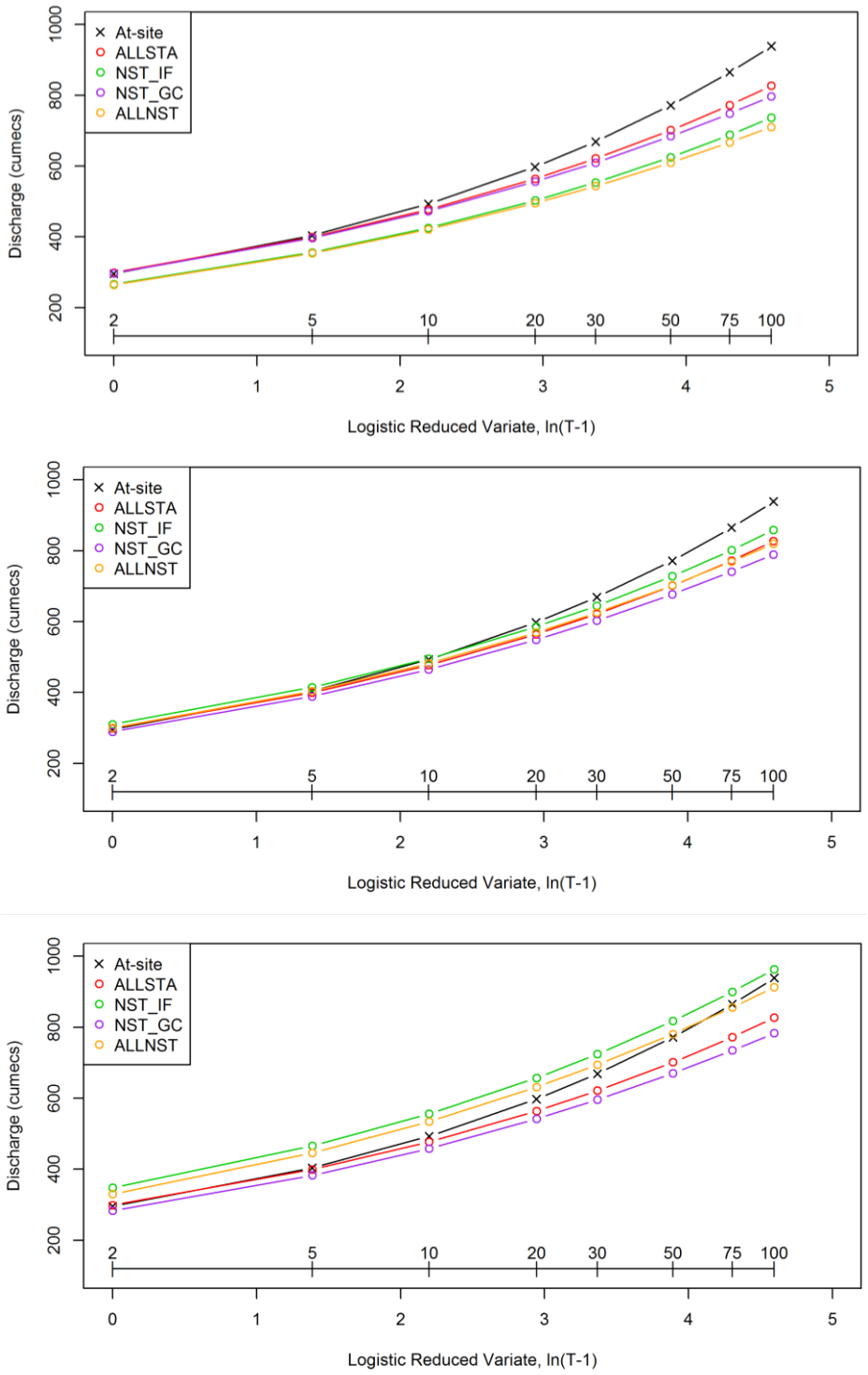


Figure C-9 Flood frequency curves for station 76005 on the Eden under different pooling methods as calculated for (a) 1977, (b) 2000 and (c) 2020. Trends are only included in location parameters, using the TSE

Table C-9 Index flood and growth curve parameters under different methods

	Pooled STA $(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$	**Pooled NST** $(\boldsymbol{\xi(t)}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$	**STA Index flood**	**NST Index flood**
1977	(1.01, 0.205, -0.255)	(1.005, 0.206, -0.245)	295.5	263.3
2000	(1.01, 0.205, -0.255)	(0.980, 0.206, -0.245)	295.5	306.6
2020	(1.01, 0.205, -0.255)	(0.958, 0.206, -0.245)	295.5	344.3

Simulation study

Figure C-10 illustrates the effectiveness of correctly identifying whether a trend needs fitting in a pooling group; the box representing the inter-quartile range, and the points representing the 5%, 50% and 95% quantiles. These quantiles were calculated by using the same 'true' distribution and re-simulating pooling groups 100 times to generate confidence intervals.

Each plot shows the estimates for one parameter (notice that one cannot estimate ξ_1 or α_1 using stationary models). Each block of 3 bars shows the 3 dependency structures (simulated pooling groups) under identical conditions. From left to right, the bars correspond to the stationary model (bars 1 to 6), the positive trend simulations (bars 7 to 12) and the negative trend simulations (bars 13 to 18). Orange bars indicate the model was fitted with stationary parameters, purple with non-stationary parameters.

All 3 dependency structures behave very similarly, as hoped. For all parameters, the stationary fitting is typically much tighter, giving smaller confidence intervals. In all cases, the stationary model was fairly accurately modelled when it was correctly specified. When the stationary model was incorrectly specified as non-stationary, one can see that ξ_0 and α_0 are fitted less successfully. This is due to the fact that the model is trying to fit these along with the slope, and so both are compromised.

When the non-stationary models are fitted as stationary, a compromise seems to occur. The single location (or scale) parameter fits to the average of the whole data set, and so seems to fit only with the behaviour towards the middle of the record, since both location and scale are fitted with linear functions of time. If classed correctly, ξ_0 and α_0 should describe the behaviour of the time series at the start of the record, t=0. Consequently, this leads to overestimation of ξ_0 and α_0 for positive trends, and underestimation for negative trends.

In all the simulations with trend (positive and negative) the scale parameter α_0 is overestimated and the slope α_1 is underestimated. This problem with the fitting method is likely present in the true data set as well, and so trends in that section should be treated with more caution. 'Significant' negative trends in scale may still be incorrect, if the 90% confidence points in Figure C-10 are representative of the larger observed data set. On the whole however, trends in location parameter were produced with much smaller confidence intervals, and showed fairly good agreement with the truth.

For all the methods, there is much greater uncertainty in fitting the shape parameter κ. This is a known issue in flood frequency analysis: for a fixed record length, the shape parameter will have a greater confidence interval than either the location or scale. Despite this, most of the methods had a 90% confidence interval containing the true value of κ. The only exception was fitting a non-stationary distribution to the stationary simulation.

Finally, it should be noted that these are based on 100-year records. In the observed data set, most records are less than 60 years, and so the uncertainty and accuracy in practice will be worse. This should be accounted for when making use of such analyses.

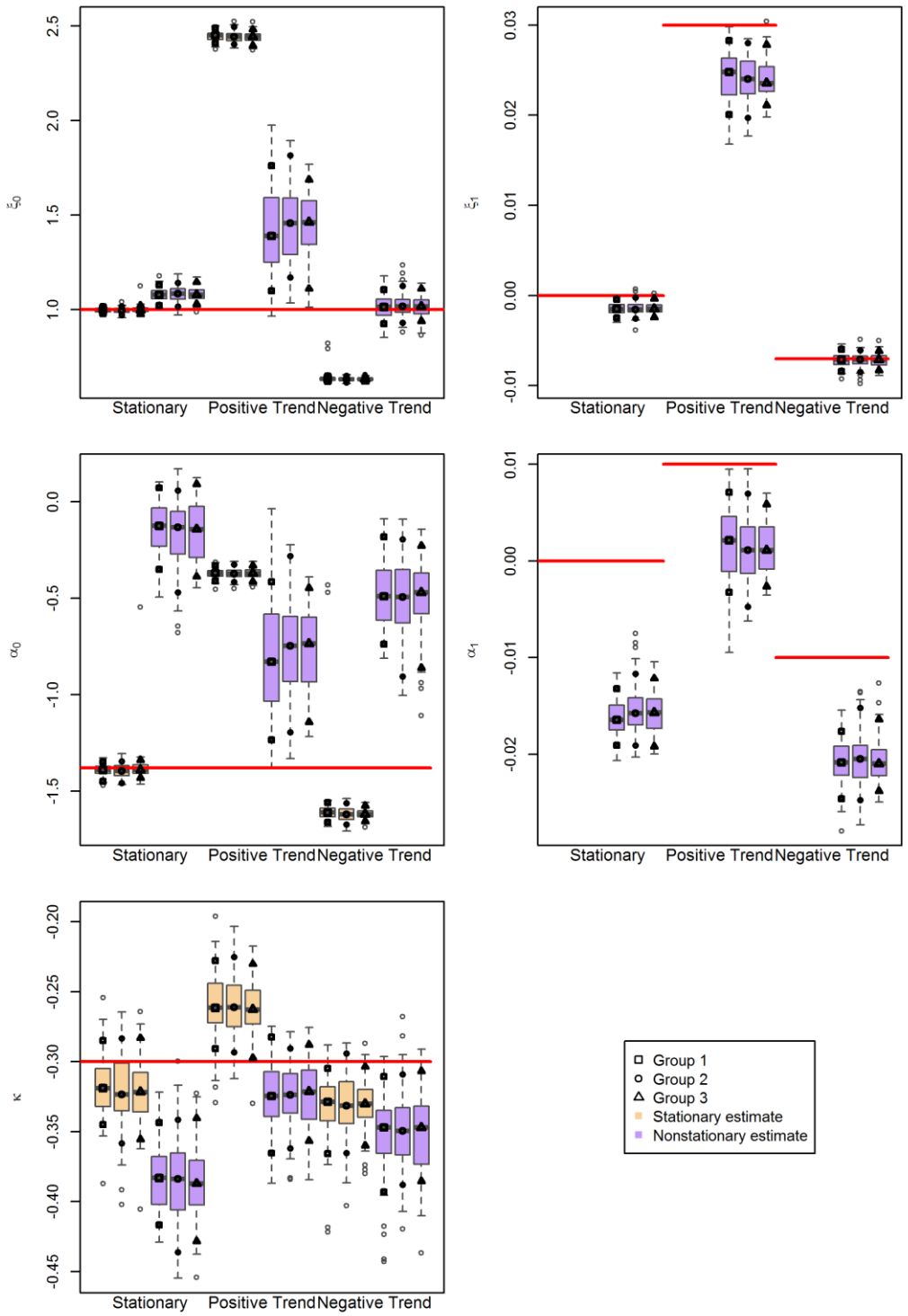


Figure C-10 Boxplots of parameter estimates from simulated data under different pooling approaches. Red lines indicate true parameter values

Concluding remarks on non-stationary pooling methods

Pooling groups can be valuable in improving the estimates at ungauged or poorly gauged locations. However, correctly assessing whether a trend is present across a pooling group can be challenging. Under current similarity metrics (SDM_{08}), the stations in the pooling group are not guaranteed to exhibit the same trends. Indeed, most stations do not show significant trends at all.

Under the 4 approaches, consisting of combinations of index-flood and growth curve choices, pooling seems to give moderately consistent results, possibly due to the lack of strong trend in many of the stations investigated. The difference between choosing a stationary or non-stationary index flood seems to be important, as can be seen in Figure C-7, where a strong difference in Q_{20} can be observed.

When the 'truth' is known, much more can be said about the effectiveness of the pooling approaches and the ability of the maximum-likelihood methods used. When correctly specified, pooling methods can give good estimates of the true value of at-site parameters. However, when models are not correctly specified (assigning trend where there is none, or vice versa), this can lead to quite poor fitting of the parameters of interest (Figure C-10). To address this, it would be interesting to specifically force different models depending on the observed at-site trend, or the trend of the pooling group members, but that would be a more substantial project than the present one.

One argument against the proposed method of investigation was described in (O'Brien and Burn 2014), where it was discussed that mixed messages regarding trends in the pooling group can be quite detrimental. It was also suggested that, where possible, restriction to like-trended pooling groups would help, or only fitting stationary distributions where this trend may be uncertain. However, restricting this present study just to those stations with significant trends would have vastly reduced the sample space and led to potentially unrepresentative pooling groups due to the small number of stations to choose from. The SDM_{LCV} developed above can improve the consistency of trend, but not ensure it.

C.4 Conclusions

This chapter summarises the findings and recommendations, and suggests future work that would be most beneficial.

Alternative similarity distance metrics were investigated, based on using the normalised Theil-Sen estimator (TSE) as a pooling variable. These were based on fitting the pooling groups using stationary estimates of L-CV and L-SKEW. In terms of improving stationary flood frequency estimates, no method performed better than the existing method as developed in Kjeldsen, Jones, and Bayliss (2008).

The Theil-Sen estimates of slope were slightly more accurately predicted when using a similarity metric which included trend as a component. This was not however, a statistically significant improvement and using TSE in the SDM components may be impractical, requiring observed flow of sufficient length.

Methods of choosing stationary and non-stationary index floods and growth curves were investigated. When trialled on the observed data, mixed signals were observed, but it seemed as though trends in the scale parameter led to consistently larger estimates in, for example, the 20-year event. Just using trend in location, rather than both location and trend, gave more consistent results with less associated model uncertainty.

By using simulations with empirically derived correlation and covariance structures, it was observed that correctly modelling trend is important. Incorrectly assuming the presence or non-presence of significant trend at a site gave strongly different results. Scale parameters seemed to be most affected by this, but seemed to show poor fit in modelling trend over time, even when correctly specified. Trend in location parameter (median behaviour) seemed to be quite accurate for observed records of 100 years; in practice, this good performance should not be expected, due to shorter records, and less well quantified uncertainty in real records.

Continued research is needed. Pooling groups making use of non-stationary index floods seem most promising, but a new distribution test under the influence of non-stationarity is an important step. The present work used the GLO as the recommended distribution, but it is not known whether there is a better distribution to choose when trend is observed. Moreover, the existing Hosking-Wallis method of distribution choice based on L-moments does not apply in the non-stationary case. Overall, this work recommends that the current SDM (SDM_{08}) should still be used for forming pooling groups, but if regions or pooling groups are generated to account for trend, care should be taken to select stations with like trend (positive, negative or no trend). Non-stationarity in growth curves should be used with caution, due to problems in fitting these curves through maximum likelihood methods. For now, stationary growth curves and index floods are the recommended method, using all available data. However, if a given pooling group is of entirely like trend, then incorporating a trend into the growth curve is reasonable, and should be considered and compared to the stationary growth curve. It is not recommended to use this non-stationarity for future projection, as the validity of a linear trend into the future cannot be assured. At this point, more work must be done before solely using non-stationary index floods or growth curves can be recommended. Specifically, more informed trend functions, such as climate-model-informed trend functions, should be investigated to improve the model for future prediction.

A spatial model of trend, a catchment descriptor equation, or both, to describe the likely trend at ungauged locations would be invaluable. The present work can only provide estimates at gauged locations, although could be of benefit if the record for the target site is short. A more comprehensive spatial model could possibly make use of large scale meteorological trends in, for example, the North Atlantic Oscillation.

The provided description of return period is complex and unintuitive. One alternative that is already used in many applications would be to use stationary return periods based on the moving parameters. These would give, for example, a ‘2020 level for a design life of 50 years.’ This does not account for change over time, but is simpler to calculate, and easier to understand. It could however, be simply calculated for different horizons to see change over time. A design life level, as described in (Yan and others, 2017), could be a plausible alternative.

Additional note: Linear catchment descriptor modelling of at-site TSE

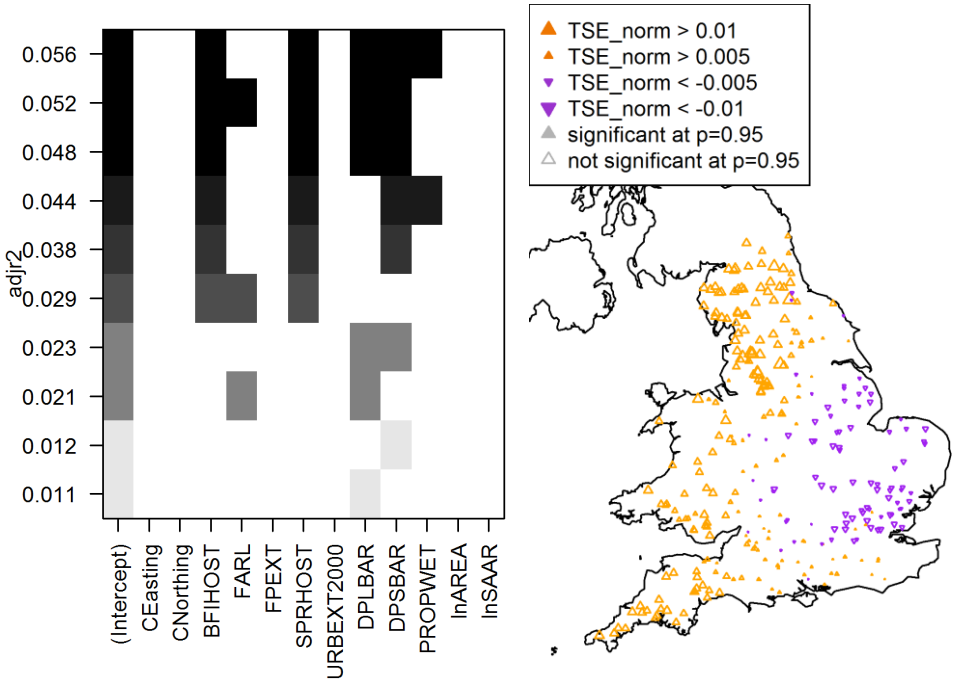


Figure C-11 Stepwise model selection diagram (left) and modelled pooled estimate of TSEnorm (right)

Using the same methods to develop SDMs to have pooling groups with similar L-CV or L-SKEW, a distance measure was developed to generate pooling groups with similar TSE (see Figure C-11 above). The stepwise regression method highlighted FPEXT, URBEXT2000 and PROPWET as possible factors to include, but Euclidean distance (EUCL) was also included, due to the spatial pattern seen in at-site TSE.

This gave the following distance metric:

$$dist_{TSE}(i,j) = \sqrt{0.724\left(\frac{FPEXT_i - FPEXT_j}{0.041}\right)^2 + 0.347\left(\frac{URBEXT_{2000,i} - URBEXT_{2000,j}}{0.086}\right)^2 + 1.106\left(\frac{PROPWET_i - PROPWET_j}{0.130}\right)^2 + 0.823\left(\frac{EUCL(i,j)}{115.8}\right)^2}$$

The model for estimating TSE_{norm} had an adjusted R^2 value of less than 10%, even weaker than the L-SKEW model. This leads to the conclusion that the model should not be used for any trend estimation at ungauged sites at present, and more work needs to be done.

As can be seen in Figure C-11, the pooled estimates led to extremely strong spatial patterns, stronger than is to be reasonably expected. Also, as before, the pooling procedure led to much less strong values for TSE.

Appendix D: Exploring spatial statistics

Introduction

This project has investigated spatial aspects of non-stationary flood frequency analysis in 2 ways:

- reconciliation of non-stationary flood frequency estimation and the FEH pooling approach, that is, regional frequency analysis
- exploring fitting a spatial statistical model that weights information by distance

The first has been investigated in much more depth, since it is considered important to build bridges between non-stationary analysis and the widely-used FEH method. This appendix records the trialling of a spatial statistical model, which shows promise, but is not proposed as something that is ready for incorporating into the interim guidance for practitioners.

Spatial variation in extreme values

When considering the distribution of extreme flows at multiple gauges, comparison of model parameters should reveal similarity between neighbouring gauges, implying that neighbouring gauges are exposed to similar physical processes. Fitting independent models between gauges relies on the data to reveal this similarity. However, with short record lengths often associated with annual maximum series, this signal can be hidden by sampling uncertainty.

An alternative approach is to explicitly incorporate information from other sites into an extreme value analysis. This is typically done using 2 methods:

- assuming statistical homogeneity over all sites within a region, that is, information is shared equally from all sites
- defining a spatial structure that weights information by distance, that is, the probability of extreme flood events will be more similar at neighbouring gauges than for those separated by large distances

This appendix explores some approaches for incorporating spatial information into extreme value models and outlining the merits and disadvantages of each. Extensions of these models for non-stationary processes and issues regarding parameter uncertainty are discussed. Finally, 3 models are proposed to explore and develop further and an example is presented to illustrate the benefits of spatial extreme value models. All discussion assumes the generalised extreme value (GEV) distribution is being used to model annual maximum flows.

Regional frequency analysis

One approach that incorporates spatial information into extreme value modelling is regional frequency analysis (RFA). This is the approach taken in the FEH pooling approach, in which the regions are defined using catchment similarity rather than proximity.

RFA pools information across gauged catchments with homogeneous statistical behaviour, with the aim of reducing uncertainty in parameter estimates (Hosking and Wallis, 2005). Failure to account for spatial dependence between sites in a pooling group can lead to

artificial and false reductions in uncertainty. As mentioned in Appendix C, simulations with realistic dependence structures have been used to overcome this difficulty.

Bayesian hierarchical models

More recently, Bayesian hierarchical models have been used as a way of incorporating spatial information into an extreme value analysis. Bayesian methods infer the probability distribution of the parameters of a statistical model, assuming that the data are fixed. This is in contrast to frequentist methods, which focus on finding fixed estimates of parameters that best describe the probability distribution of the data. Bayes' theorem states that the posterior probability distribution of a parameter is proportional to the product of a prior probability distribution and the likelihood. The prior distribution can be used to define the user's prior belief about the parameter that is independent of the data being analysed. This can be especially useful for incorporating expert domain knowledge about the process being studied. For complex models, inference is carried out using Markov Chain Monte Carlo (MCMC) methods, which consists of iterative sampling from the posterior distribution of interest.

Hierarchical models are statistical models written on multiple levels that can be used to model grouped or nested data. Historically, these methods have been typically used for modelling spatial count and binary data (Diggle and others, 1998) before being extended for modelling extreme values at multiple locations. Fawcett and others (2006) proposed a hierarchical model for wind speeds with exchangeable spatial effects, that is, that all sites provide equally-weighted information regardless of distance. Cooley and others (2007), Sang and Gelfand (2009) and Cooley and Sain (2010) all assume a Gaussian process prior distribution for the model parameters, explicitly weighting information from other sites based on distance. Recent studies following this general approach include Wang and So (2016), Barlow and others (2018) and Sharkey and Winter (2019).

These approaches tend to produce a smooth and spatially cohesive map of parameter estimates and return levels that are perhaps more representative of the underlying physical process. Incorporating spatial information helps to prevent neighbouring sites from having very different statistical models of extreme flow.

Spatial models for non-stationarity

The studies mentioned above are focused mainly on modelling spatial variation in extreme values, but recent studies have also looked at modelling temporal non-stationarity on regional scales. Eastoe (2019) assumes regional homogeneity in the inter-annual variability of the model parameters, which improves estimation of site-specific non-stationarity through spatial pooling. Prosdocimi and others (2019) model the test statistic of the regression coefficient of a log-normal distribution using a random effect for the hydrometric area corresponding to each gauge, which is assumed to be homogeneous nationally. This model allows for an overall UK trend, a regional trend for each hydrometric area and a random variation specific to each gauge.

To be consistent with the scope of this project, which calls for an investigation of feasibility, the project team explored a GEV model where only the location parameter μ is time-dependent, and where the parameter that expresses this time variance (μ_1) has some sort of spatial structure. Three approaches to Bayesian spatial models are presented here that would be useful to explore further and possibly develop in the future, but only option 1 has been implemented as part of this project. Any of these models could be adapted to impose a spatial structure on other GEV parameters such as the scale. The 3 models considered were:

1. A gauge-specific GEV distribution with time covariate, whose parameter μ_1 can be modelled as a normal distribution with shared variance across a pre-defined region such as a hydrometric area. All gauges are assumed to be statistically homogenous over the region. Given the parameters, the observations in a year over sites are assumed to be independent.
2. A gauge-specific GEV distribution with time covariate, whose parameter μ_1 can be modelled as a normal distribution with structure imposing that gauges that are spatially or hydrologically similar are more likely to have similar coefficients. Let $X_{s,t}$ denote the annual maximum flow at gauge s in year t. One then assumes:

$$X_{s,t} \sim GEV(\mu_s^{(0)} + \mu_s^{(1)} t, \sigma_s, \xi_s),$$

where

$$\mu_s^{(1)} \sim \mathcal{N}\left(\frac{\sum_{s' \neq s} \mu_{s'}^{(1)} d_{s,s'}}{\sum_{s' \neq s} d_{s,s'}}, \frac{1}{\tau_{\mu^{(1)}} \sum_{s' \neq s} d_{s,s'}} \right)$$

and

$$d_{s,s'} = \exp\{-\|x_s - x_{s'}\|\}$$

Where x_s is the spatial location, or some measure of catchment similarity, for gauges. This approach is theoretically the same as kriging, except the spatial smoothing becomes part of the inference rather than a post-hoc step.

3. A regional GEV random effects model with regression coefficient estimated for a pre-defined region such as a hydrometric area. Let $X_{j,s,t}$ denote the annual maximum flow at gauge s in hydrometric area j in year t. One then assumes:

$$X_{j,s,t} \sim GEV(\mu_{j,s}^{(0)} + \mu_j^{(1)} t, \sigma_{j,s}, \xi_{j,s}).$$

This approach assumes a gauge-specific intercept term in the location parameter but a regional trend term. Spatial pooling of this type would help to increase the signal across a region.

Results

Option 1 has been implemented on a set of 12 gauges in north-west England. Modelling the parameter μ_1 as following a normal distribution that is common across all gauges has the effect of pooling information from all gauges within the region and producing a spatially smooth set of trend estimates. This is shown in Figure D-1, where posterior means from the spatial model are compared with maximum likelihood estimates from analysis of the individual sites. μ_1 varies between 0 and 1.2 when estimated separately at each gauge. When estimated via the Bayesian spatial process, μ_1 varies over a smaller range, between 0.2 and 0.6. The more extreme estimates are moderated.

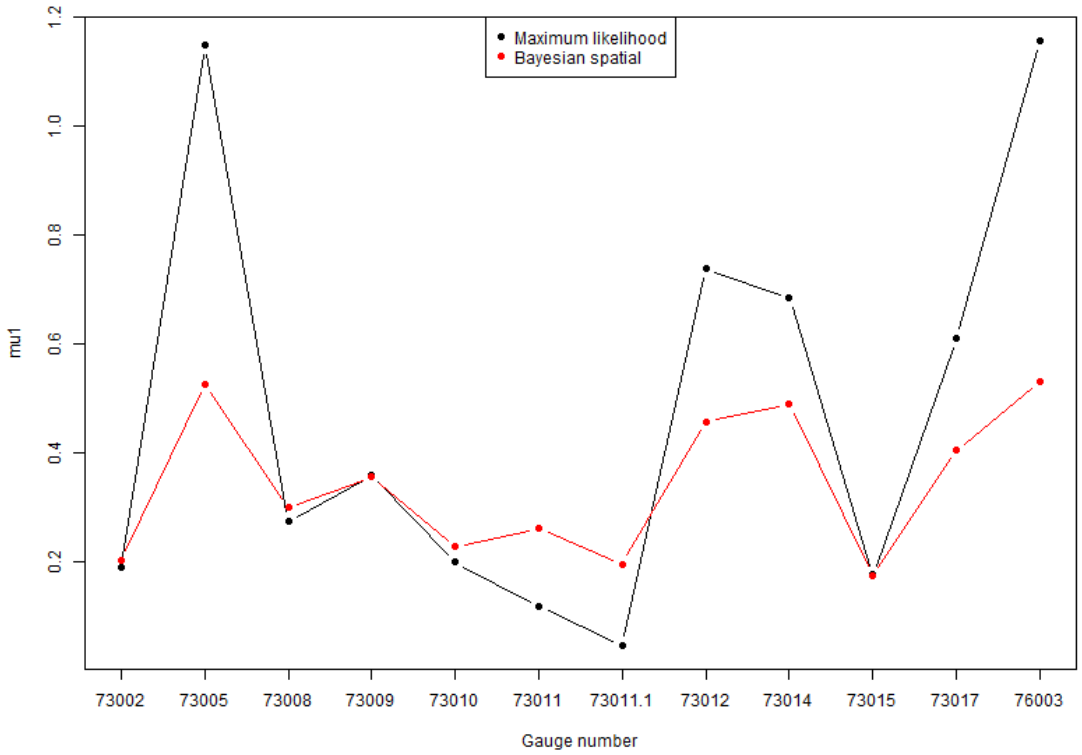


Figure D-1 Comparison of trend estimates in the location parameter of a GEV model: maximum likelihood estimates made separately at each gauge versus posterior means from the Bayesian spatial model

This model appears to provide a promising way to boost the detection of non-stationarity by removing some of the ‘noise’.

Uncertainty

A logical next step would be to explore the effect of the spatial analysis on parameter uncertainty. A common misconception with spatial pooling approaches is the perceived reduction in uncertainty gained from extra information used in the analysis. This would be true if the method pooled over independent sources of information, but annual maximum river flows are likely to be spatially correlated (see the discussion of the simulation study carried out to evaluate the performance of the various non-stationary pooling methods discussed in Appendix C). An assumption of independence is required for Bayesian hierarchical models, but it means that resulting confidence widths are likely to be underestimated. There are 2 potential approaches for deriving correct confidence intervals:

1. A spatial nonparametric bootstrap – first, the trend at each gauge is estimated and the residuals are calculated. These residuals are resampled and the trend added back in, before the model is refitted to the resampled data. Repeating this action multiple times builds up the sampling distribution of the trend parameter, from which confidence intervals can be extracted.

2. A likelihood correction - Sharkey and Winter (2019) estimate a measure of spatial dependence in the data that is used to correct confidence intervals for the false assumption of independence over locations. This measure, referred to as a magnitude adjustment, scales the independence likelihood to match the asymptotic properties of the 'true' likelihood,

which is unknown. This results in a widening of confidence intervals and a more correct representation of parameter uncertainty.

Further work on spatial statistics would involve implementing one of these methods to quantify uncertainty. However, considerable work would be needed to develop an approach that could be readily applied by practitioners, and it may be preferable to concentrate efforts on developing a non-stationary framework for the FEH pooling method.

Appendix E: Testing approaches to applying climate change adjustments

E.1 Introduction

This appendix considers existing guidance on making allowance for the effects of climate change on fluvial flood frequency estimates in the light of the current analysis of non-stationarity in annual maximum peak flow (AMAX) data in England and Wales.

The study of non-stationarity in flood peak data is complex, not least because of the impact of multiple and interacting factors such as land cover and land-use change, particularly urbanisation, hydraulic changes to river channels and the high degree of natural variability in the data. Francois and others (2019) discuss example catchments in the USA and the Netherlands where the influence of anthropogenic climate change and natural climate variability are difficult to disentangle. The influence of the latter is further confounded in the UK by so-called flood-rich and flood-poor periods. Prosdocimi and others (2015) used change in urban extent as well as a number of climate-based covariates to model changing flood regime over time in 2 catchments in England. Increasing urbanisation was shown to have a significant effect on high flows in one of the catchments, particularly in summer. Therefore, while anthropogenic climate change may be a major driver of non-stationarity in peak flow data, it is very important that global change impacts are attributed reliably and the risk of 'climatisation' is avoided by taking non-climatic factors into account (Wine and Davison 2019).

Existing guidance on climate change allowances for England

Agencies across the UK have been providing guidance on the potential impacts of climate change on floods for many years, so that these can be accounted for by flood management authorities and local planners aiming to reduce flood risk (Reynard and others, 2017). The most recent guidance adopts a regional risk-based approach (Environment Agency, 2016a, b), and is based on combining the UKCP09 climate projections (Murphy and others, 2009) with a sensitivity-based approach to modelling the impacts of climate change on peak flows (Kay and others, 2011; 2014).

The guidance for flood management authorities (Environment Agency, 2016a and Table E-1) provides a set of 5 numbers (lower, central, higher central, upper and H++) for each of 11 regions covering England (Figure E-1), for 3 future time slices (2020s, 2050s and 2080s). The 'lower', 'central' and 'upper' numbers represent the main range of estimated impacts of climate change on flood peaks from the UKCP09 projections. The H++ numbers represent plausible but unlikely high-end impacts of climate change. The guidance for flood risk assessments (Environment Agency, 2016b) is similar but without the 'lower' number, but the focus here is the guidance for flood management authorities.

The guidance (Environment Agency, 2016a) recommends that the 'central' estimate of change should be used to define the risk over the decision lifetime, with the 'upper' and 'lower' estimates provided to encourage consideration of the options required to manage the

fuller range of risk, for example, building flexibility into the plan to allow future adjustments if necessary (Reynard and others, 2017).

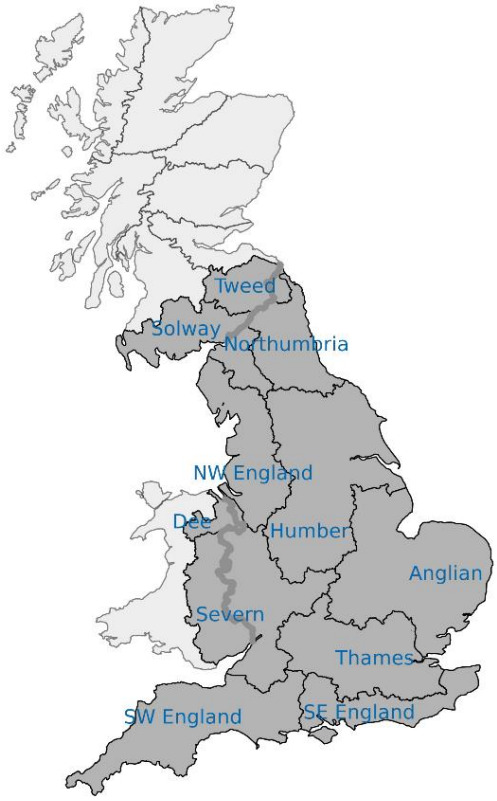


Figure E-1 The 11 regions covering England used in developing climate changes allowances (CCAs)

Table E-1 Regional guidance for England, for 3 time slices

	2020s	2050s	2080s
Solway			
H++	25	45	95
Upper (90th)	20	30	60
Higher Central (70th)	15	25	30
Central (50th)	10	20	25
Lower (10th)	5	10	10
NW England			
H++	25	45	95
Upper (90th)	20	35	70
Higher Central (70th)	20	30	35
Central (50th)	15	25	30
Lower (10th)	10	10	10
Dee			
H++	20	30	60
Upper (90th)	20	30	45
Higher Central (70th)	15	20	25
Central (50th)	10	15	20
Lower (10th)	5	5	5
Severn			
H++	25	45	90
Upper (90th)	25	40	70
Higher Central (70th)	15	25	35
Central (50th)	10	20	25
Lower (10th)	0	5	5
SW England			
H++	25	50	105
Upper (90th)	25	40	85
Higher Central (70th)	20	30	40
Central (50th)	10	20	30
Lower (10th)	5	5	10

	2020s	2050s	2080s
Tweed			
H++	20	35	75
Upper (90th)	20	25	45
Higher Central (70th)	15	20	25
Central (50th)	10	15	20
Lower (10th)	0	5	5
Northumbria			
H++	20	35	65
Upper (90th)	20	30	50
Higher Central (70th)	15	20	25
Central (50th)	10	15	20
Lower (10th)	5	5	10
Humber			
H++	20	35	65
Upper (90th)	20	30	50
Higher Central (70th)	15	20	30
Central (50th)	10	15	20
Lower (10th)	5	5	10
Anglian			
H++	25	40	80
Upper (90th)	25	35	65
Higher Central (70th)	15	20	35
Central (50th)	10	15	25
Lower (10th)	0	0	5
Thames			
H++	25	40	80
Upper (90th)	25	35	70
Higher Central (70th)	15	25	35
Central (50th)	10	15	25
Lower (10th)	-5	0	5
SE England			
H++	30	60	120
Upper (90th)	25	50	105
Higher Central (70th)	15	30	45
Central (50th)	10	20	35
Lower (10th)	-5	0	5

Issues when applying the guidance

Some of the issues that occur when applying climate change allowances for peak flows are listed below. Some of these issues are complex, and several are interrelated:

- Climate change allowances are derived from climate projections from a 1961 to 1990 baseline (typically using hydrological modelling for baseline period 1961 to 2001 (Kay and others, 2014)).
- The impacts for the 2020s time slice are based on the potential climate change between the baseline period and the period 2010 to 2039, therefore prompting the question of whether some of the climate change has 'already happened'. If so, is the application of the full allowance still valid?
- Even if a clear trend is apparent in the AMAX data for a particular catchment, it could be for a range of reasons other than climate change, including natural climate variability.
- The AMAX records for different catchments are very variable in length; some only have data for more recent years, while a small number of catchments have very long records. The Flood Estimation Handbook (FEH) statistical method for ungauged sites (Kjeldsen and others, 2008) pools the flood peak data from a network of hydrologically similar sites, and therefore the pooling group may contain data for different time periods with differing amounts of variability and trend.
- Applying the full climate change allowances immediately can result in large increases for the earliest time slices. However, the Environment Agency guidance does suggest applying a linear increase in the allowance for the period up to 2025 (see Figure E-2).
-

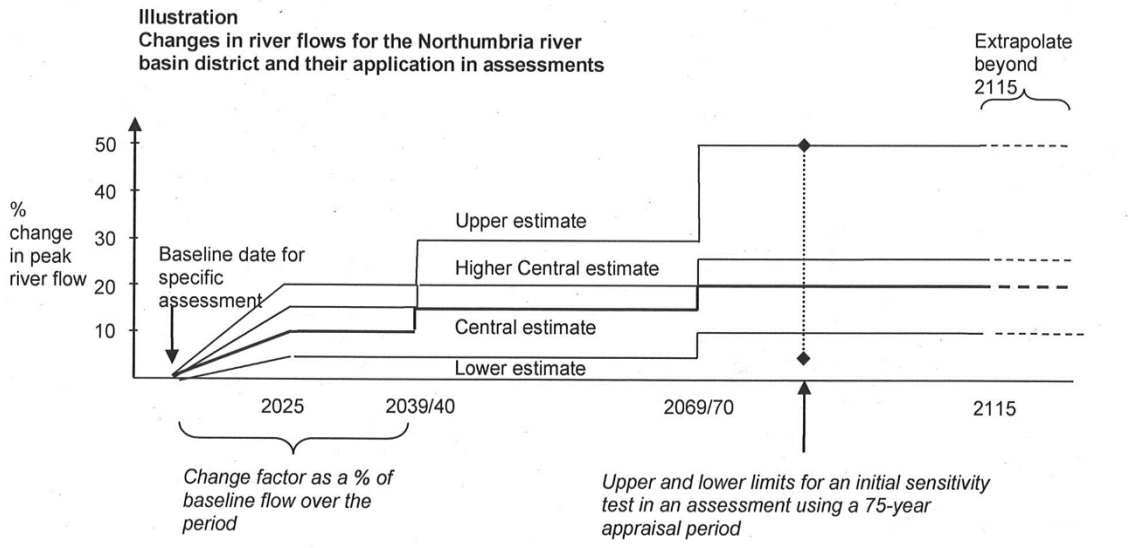


Figure E-2 Changes in river flows for the Northumbria river basin district and their application in assessments (from Environment Agency, 2016a)

Current methods of applying allowances

Current guidance on applying climate change allowances (CCAs) is somewhat open to interpretation. The allowances themselves were derived from climate projections from a 1961 to 1990 baseline. This appendix discusses various ways in which CCAs are, or could be, applied. It compares how different extrapolations to 2025, 2050, and 2080 are affected on a regional scale, depending on which baseline is chosen and whether the baseline is assumed to be stationary. The above issues are also investigated via a set of case studies.

E.2 Methods

CCA applications

CCAs and statistical extrapolation will be applied to 5 baselines:

- full record stationary (STFULL): stationary flood frequency estimates are calculated based on the whole period of record, then CCAs are applied. This is the method to which the other 4 will be compared
- 61 to 90 stationary (ST6190): stationary flood frequency estimates calculated based on 1961 to 1990, then applying CCAs
- 61 to 90 non-stationary (NST6190): non-stationary flood frequency estimates calculated based on 1961 to 1990
    - flood frequency estimates are based on the fitted parameter values as evaluated in 1990. CCAs will be applied to the 1990 estimate
- full record representative non-stationary (NSTREP): non-stationary flood frequency estimates calculated based on the whole period of record
    - flood frequency estimates are based on the fitted parameter values as estimated in 1990. CCAs will be applied to this 1990 estimate
- full record non-stationary (NSTEXT): non-stationary flood frequency estimates calculated based on the whole period of record. Extrapolations are used instead of applying CCAs, and so may report quite different patterns in change
    - the extrapolations to 2025, 2050 and 2080 are calculated without CCAs, using values of the fitted parameters (baseline location, trend in location, scale, shape) evaluated at the 3 horizons to calculate Q50

In all non-stationary cases, trend is introduced through a linear trend over time in the location parameter of the fitted distribution (generalised logistic distribution). For ST6190 and STFULL, GLO parameters are estimated using standard L-moment methods. For NSTEXT and NSTREP, the trend is computed using the Theil-Sen estimate based on the whole period of record. Non-stationary parameters are then estimated using maximum likelihood methods to generate ($\xi(t)$, α, κ) using the Theil-Sen estimate of trend for the location parameter. Scale and shape parameters are left stationary for the period of record. For NSTEXT, $\xi(2019)$ is used, and for NSTREP, $\xi(1990)$ is used. For NSTEXT, a constant trend out to 2080 is assumed for the purposes of this report.

For the purpose of example, only the 50-year and 100-year events are discussed here. In each case, Q50 (or Q100) is calculated for the 5 cases above for all of the 381 stations determined to be ‘suitable for non-stationary flood frequency analysis’; see Table 3 for a breakdown by region. Then the CCAs (central estimates) will be applied to obtain estimates for 2025, 2050 and 2080. For 1961 to 1990 baselines, estimates are given for values as in

1990. Preliminary work suggested that these results were generally invariant for this choice of year.

To compare these approaches, ST6190, NST6190, NSTREP and NSTEXT (NSTEXT extrapolated to 2025, 2050 and 2080) will be compared to STFULL. See Figure E-3 for an illustrative example of these different values of Q50. Percentage differences between STFULL and the alternatives will be calculated. These percentage differences will be summarised regionally, taking the median over each of the river basin districts.

NB: None of these methods is claimed to be 'the truth', only different possible ways of projecting future extremes.

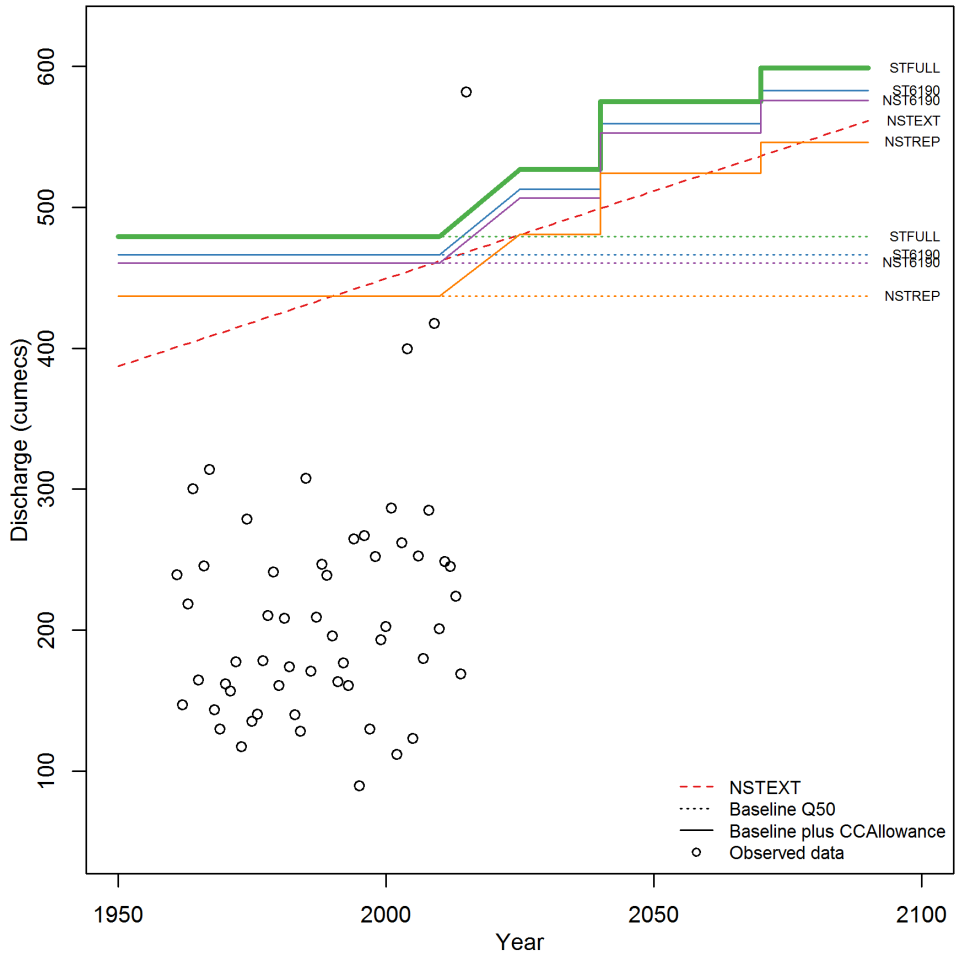


Figure E-3 Illustrative example of baseline periods and Q50 modified by CCAs using a station from the Environment Agency data set (Eamont at Udford). Bold green line indicates STFULL method, the most commonly used approach, to which everything else will be compared

Case study selection

A set of 3 catchments was selected from the data set constructed for the project (see details in Table E-2).

The 3 catchments were chosen for their long records and on the basis of trends found by Faulkner and others (2019). The AMAX data for Little Ouse showed a negative trend, with only the scale parameter changing over time. The record at Kennal displayed a positive trend, with only the scale parameter changing over time, while the data for the Eden showed a positive trend, with only the location parameter changing over time.

Table E-2 Details of case study catchments

Station number	Station name	Years of record	Region (Figure E-1)
33034	Little Ouse at Abbey Heath	48	Anglian
48007	Kennal at Ponsanooth	48	SW England
76005	Eden at Temple Sowerby	54	NW England

The characteristics of the 3 catchments are described below:

33034: Little Ouse at Abbey Heath

Predominantly arable, with the urban development of Thetford just upstream of the station (URBEXT = 0.0503), linked to groundwater abstraction for industry and agriculture, experiences some effluent returns. Fairly dry catchment (SAAR6190 = 607 mm). Small but significant ($p < 0.05$) negative trend in median peak flow. Rejected and missing data around 2000 to 2002. Reasonably large catchment of 688 $km^{2.}$

48007: Kennal at Ponsanooth

Small catchment (26.5 km^2) in SW Cornwall. No significant ($p > 0.05$) trend in the location parameter using peak flow data (1968 to present). Affected by exports from Stithians Reservoir 4 miles upstream, and abstraction for public water supply. Small urban extent (URBEXT = 0.0466) mostly grassland with high baseflow (BFIHOST = 0.74). Responsive to heavy rainfall (SAAR = 1294). Low FARL (0.867) due to 32.8% of catchment drained by Stithians Reservoir. High data quality.

76005: Eden at Temple Sowerby

Large, steep catchment in Cumbria (616 km^2) with little anthropogenic influence, especially above low flows. Subject to a number of highly extreme rainfall events in the last 20 years, leading to exceptional events. Nearly all rural (URBEXT = 0.0125) with no significant land use change and moderate rainfall average (SAAR = 1142 mm). No large water bodies affecting storage, and baseflow is moderate (BFIHOST = 0.47). Shows significant ($p < 0.05$) positive trend in peak flow according to Mann-Kendall tests.

Single-site analysis

Stationary flood frequency curves were fitted using the generalised logistic distribution to the AMAX data for each of the 3 case study catchments, and the existing climate change allowances were applied. The uplifted frequency estimates were then compared with those derived from a non-stationary frequency analysis. The importance of period of record was explored by fitting stationary flood frequency curves to data from 1961 to 1990 (reflecting the baseline used in the climate change impact modelling studies that underlie the existing guidance) and comparing them with stationary flood frequency curves fitted to all the AMAX data at each site.

E.3 Results

CCA applications

Figure E-4 and Figure E-5 illustrate the percentage difference described above for Q50 and Q100, with Figure E-6 restricting the Q50 differences to stations with a significant trend which is positive ($p < 0.05$). Estimates for different horizons (2025, 2050, and 2080) are in different columns; estimates for different baselines (ST6190, NST6190, NSTREP, NSTEXT) are in different rows. Note that positive percentage differences indicate that the alternative is higher than STFULL. Negative percentages indicate the alternative is lower than STFULL.

Figure E-4 and Figure E-5 show similar patterns. Compared to other alternatives, ST6190 seems to provide small percentage decreases in some places and even shows increases in places such as Wales. However, NSTREP showed more consistent small percentage decreases consistently across all regions. The Tweed region only contains one station, so should not be considered pivotal in the overall patterns. This is consistent across horizons, which is not surprising given the same CCAs used for STFULL, ST6190 and NSTREP. NST6190 (as evaluated in 1990) seems to be similar to the stationary equivalent, but all the regions show a greater percentage reduction. For NSTEXT, which does not have CCAs applied, there is, on average, less difference than seen for NST6190, and the overall spatial pattern is more consistent. The increase over time in difference between NSTEXT and STFULL is particularly clear.

Figure E-6 restricts the set of stations to those with a significant trend according to Mann-Kendall (p-value less than 0.05), and that trend is positive according to the Theil-Sen estimator. Table E-3 shows here that there are many fewer stations that satisfy this, and many regions contain very few stations with a significant and positive trend (for example, 9% of stations in the south-west have significant and positive trend). Figure E-6 shows a starkly different picture to Figure E-4. The ST6190 estimates actually produce larger estimates compared to STFULL in some westerly areas, though the massive difference in the south-west is only based on 7 stations, so this is not necessarily particularly representative of the region. For NST6190 the pattern of nationwide negative difference occurs in nearly all locations, and for the NSTEXT the difference is negative in all regions, except the single station in the Tweed region.

Finally, it should be pointed out that this linear extrapolation used by NSTEXT to 2050 and 2080 is not to be applied in practice, and is just generally indicative of one alternative way of considering future change by extrapolating historical changes. More informed future projections should be used, but the amount of difference between common practice and the alternatives suggest that there does need to be more guidance on how the CCAs developed from UKCP09 should be applied. This guidance should also be updated with the development of UKCP18.

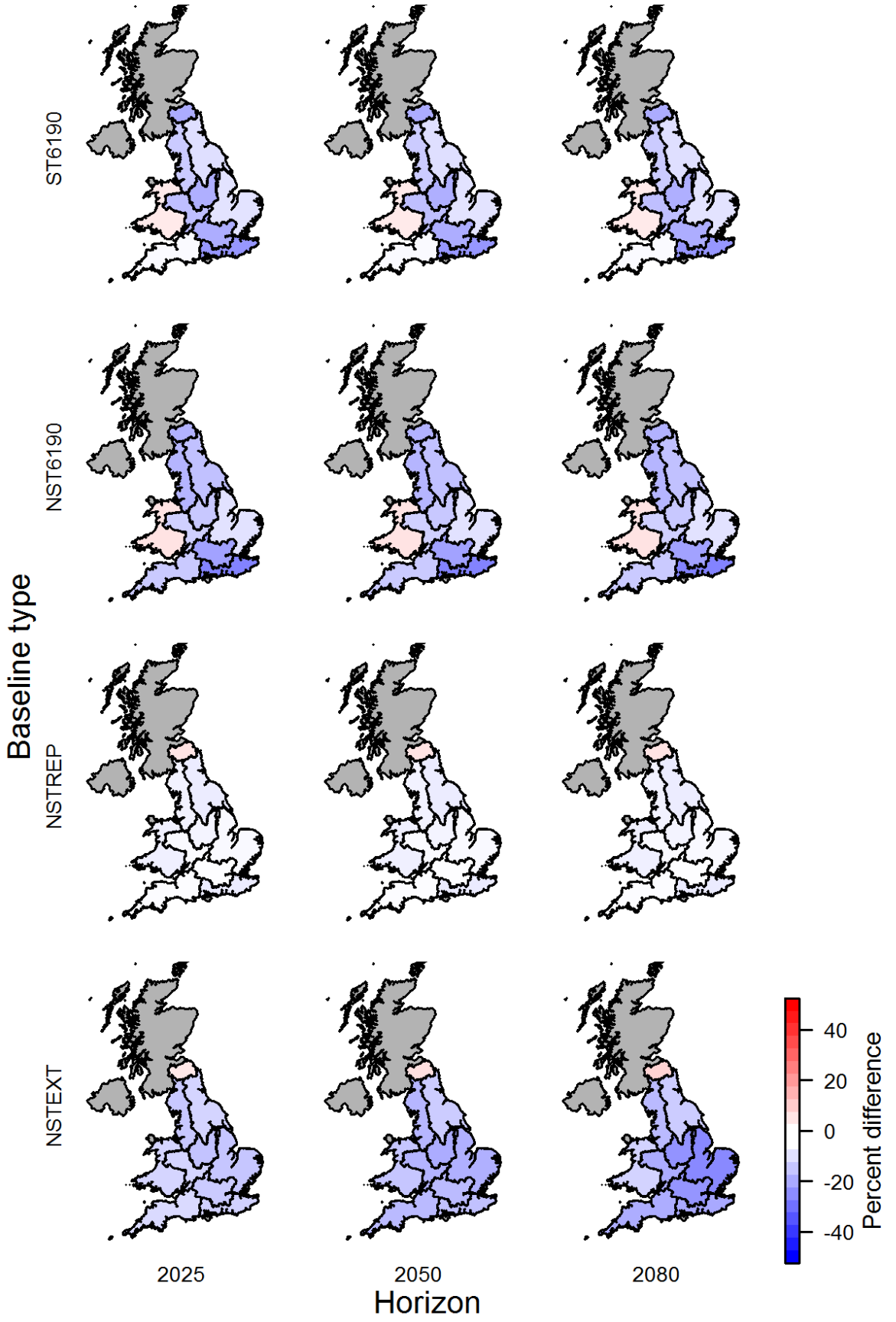


Figure E-4 Q50 percentage differences for various horizons and baseline calculations compared to STFULL (full record stationary estimates + CCAs)

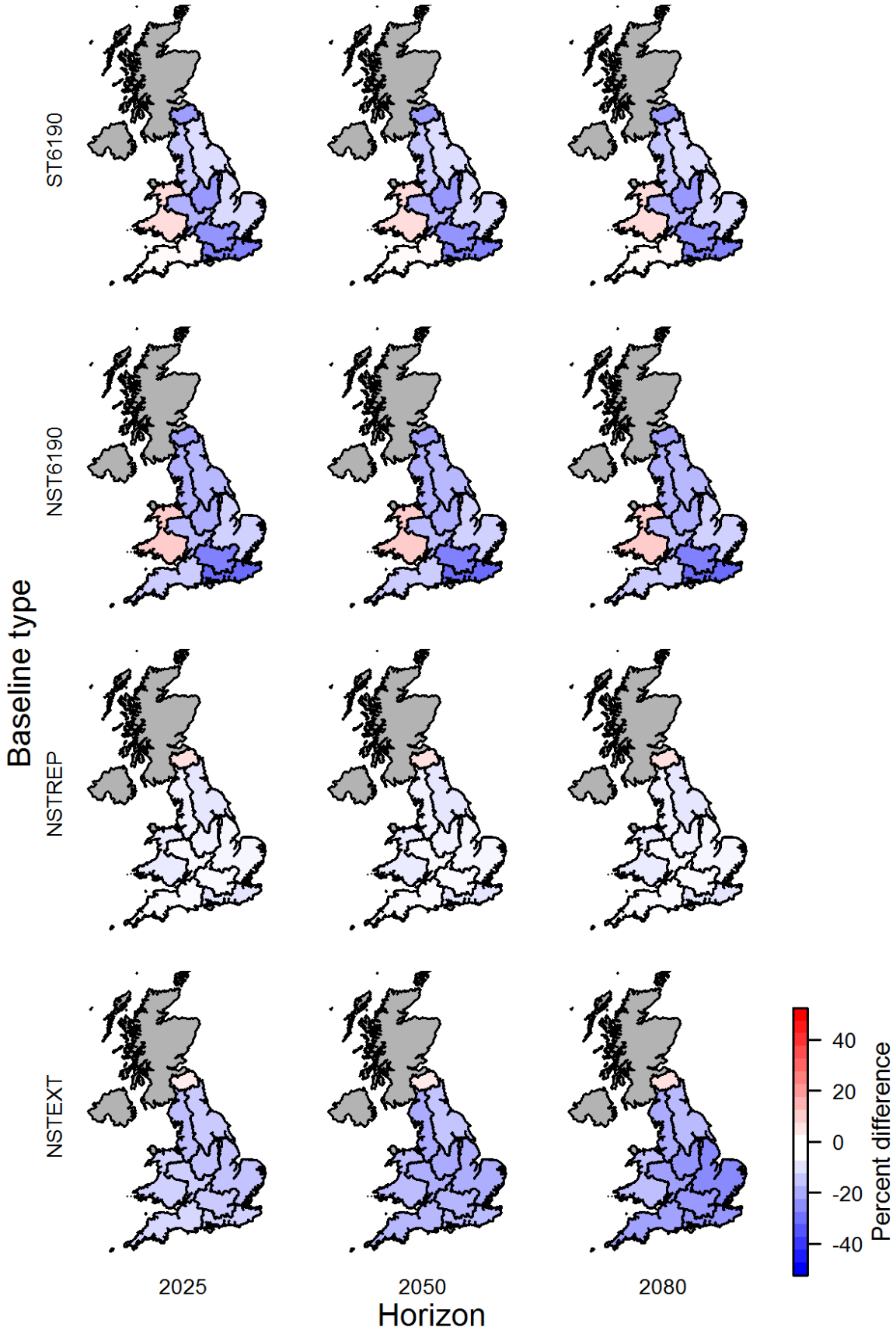


Figure E-5 Q100 percentage differences for various horizons and baseline calculations compared to STFULL (full record stationary estimates + CCAs)

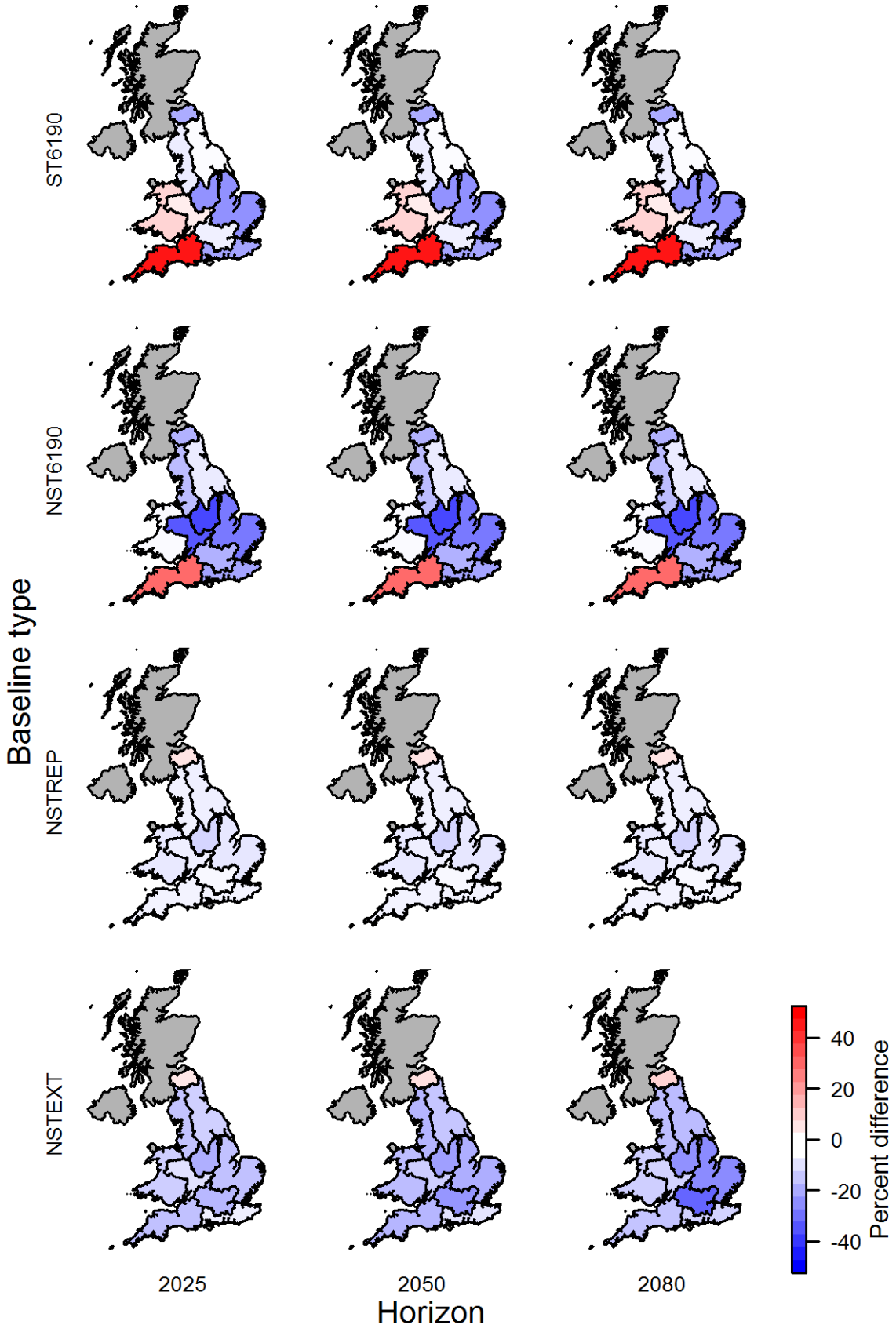


Figure E-6 Q50 percentage differences for various horizons and baseline calculations, restricted to stations with positive trend, compared to STFULL (full record stationary estimates + CCAs)

Figure E-7 Map of river district basins

Table E-3 Distribution of location of stations studied

Region	Number of stations	Number of stations with significant positive trend	Percentage of stations in region with significant positive trend (p<0.05)
Anglia	87	6	6.9%
North-east	62	10	16.1%
North-west	70	20	28.6%
South-east	20	3	15%
Severn	33	4	12.1%
South-west	78	7	9.0%
Trent	26	2	7.7%
Wales	30	5	16.7%
Thames	40	4	10%
Tweed	1	1	100%

Case study 1: Little Ouse at Abbey Heath

Figure E-8 shows the AMAX data for Little Ouse plotted as a time series (spanning 1968 to 2016), together with the effect of period of record on the estimate of QMED. The data show a negative trend, with QMED estimated over the 1961 to 1990 period (early in the AMAX record) indicated by a red line lying above the value of QMED estimated over the entire record. Also shown in Figure E-8 are non-stationary estimates of Q2, Q30, Q50 and Q100 fitted to all the AMAX data (using time as a covariate and allowing only the location parameter to change with time).

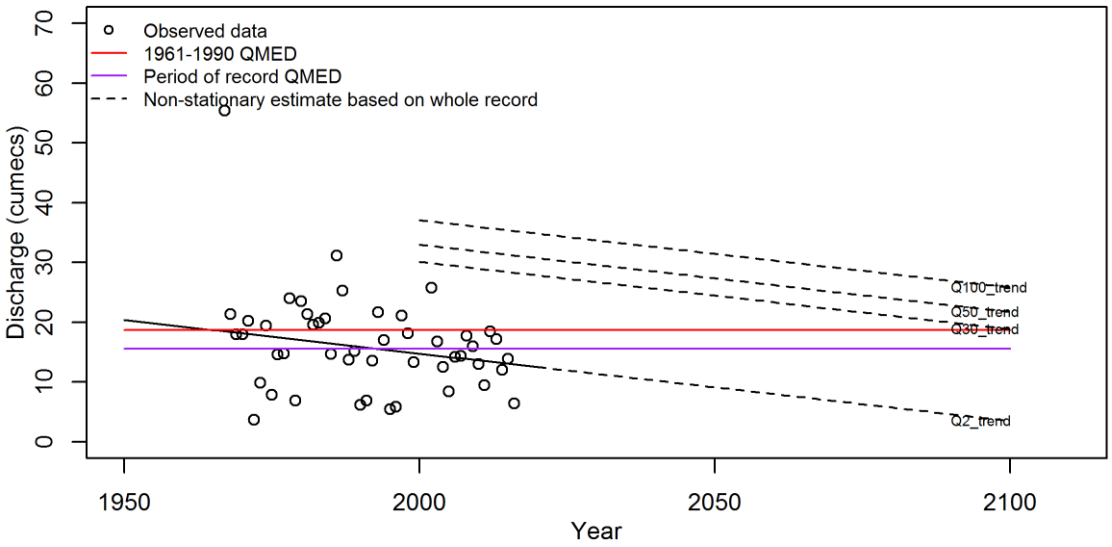


Figure E-8 AMAX data and QMED estimates for 1961 to 1990 and for the whole period of record (Little Ouse)

Stationary single-site flood frequency curves are shown in Figure E-9 fitted to data from 1961 to 1990 (in black) and for the full AMAX record (in mauve). The AMAX data points are plotted according to the Gringorten plotting position without accounting for any non-stationarity. The negative trend detected in the data is reflected in the position of the frequency curve for the full record which lies below that of the 1961 to 1990 curve. The currently recommended climate changes allowances (central) for the 2020s, 2050s and 2080s are plotted relative to the stationary curve based on the full period of record since this represents current best practice. Because of the negative trend in the data, it can be seen that the percentage uplifts for the 2080s time slice bring the frequency estimates roughly into line with the 1961 to 1990 frequency curve in this particular case. There is a single AMAX value that is plotted above both stationary frequency curves, and this represents the highest AMAX value recorded in 1968 at the beginning of the gauge record, which has a dominant influence on the trend in the data series.

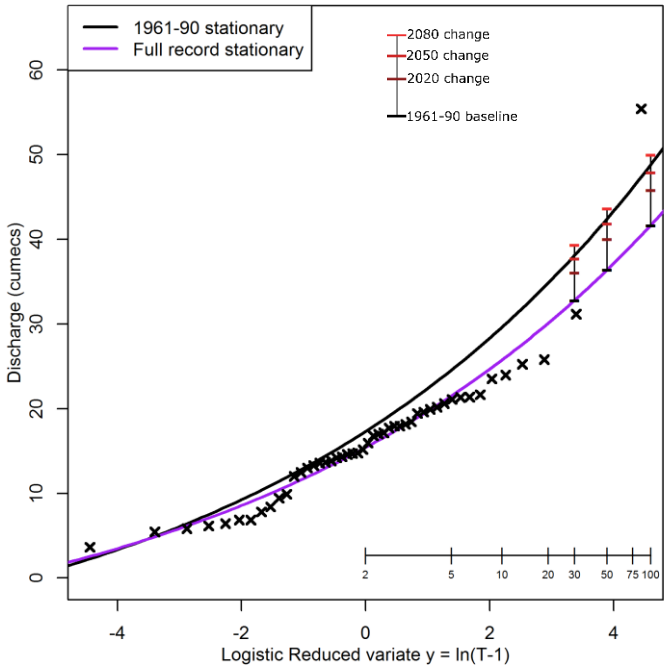


Figure E-9 Stationary flood frequency curves based on different periods of record showing climate change allowances (Little Ouse)

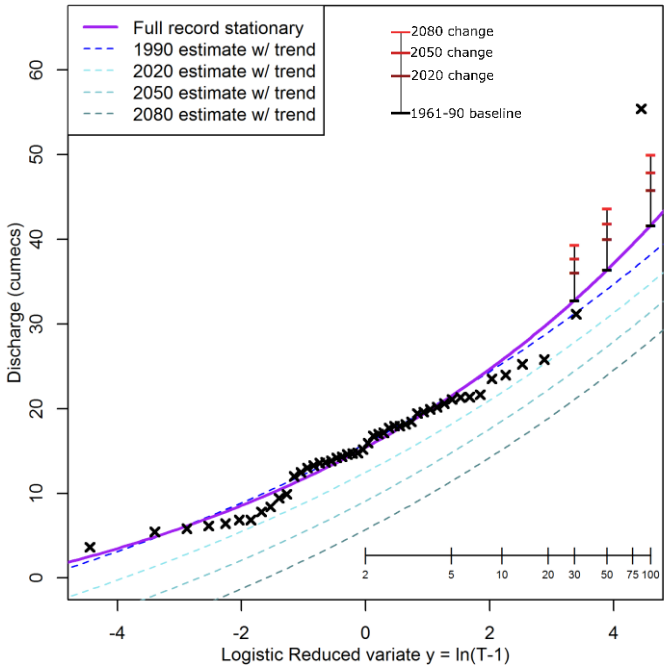


Figure E-10 Comparison of stationary and non-stationary models (Little Ouse)

Figure E-10 presents a comparison of stationary and non-stationary frequency curves for the Little Ouse. A single non-stationary model was fitted to the data set with a linear trend in the location parameter over the whole record. This has been extrapolated as far as 2080 (note that this is not advised in practice since the linear trend may not continue). The dashed lines show the evolution of the resultant non-stationary flood frequency curve over time, showing snapshots of it in 1990, 2020, 2050 and 2080. Since the trend is negative, the non-stationary

flood frequency curve moves ‘down’ the graph over time. Compared to the stationary model, it can also be seen that the non-stationary flood frequency curves are ‘flatter’, suggesting a shape parameter closer to zero. This is because the stationary model has to account for all the points at once equally, so has to fit both the new, smaller extremes with the older, larger ones using a single set of stationary GLO parameters. The non-stationary model can, in some sense, exchange ‘variance’ for ‘change over time’ in a way that the stationary distribution cannot. One can think of the non-stationary distribution being fitted by looking at the start, then the middle, then the end of the data; since there is less difference in the most extreme events over these shorter periods, the curve is flatter[9].

The final figure for this case study (Figure E-11) compares the non-stationary estimate for the 50-year return period (Q50) with the current climate change allowances applied to the stationary flood frequency curve. The negative trend causes the non-stationary estimates to be considerably lower than the stationary estimates, including the allowances for climate change for future time slices.

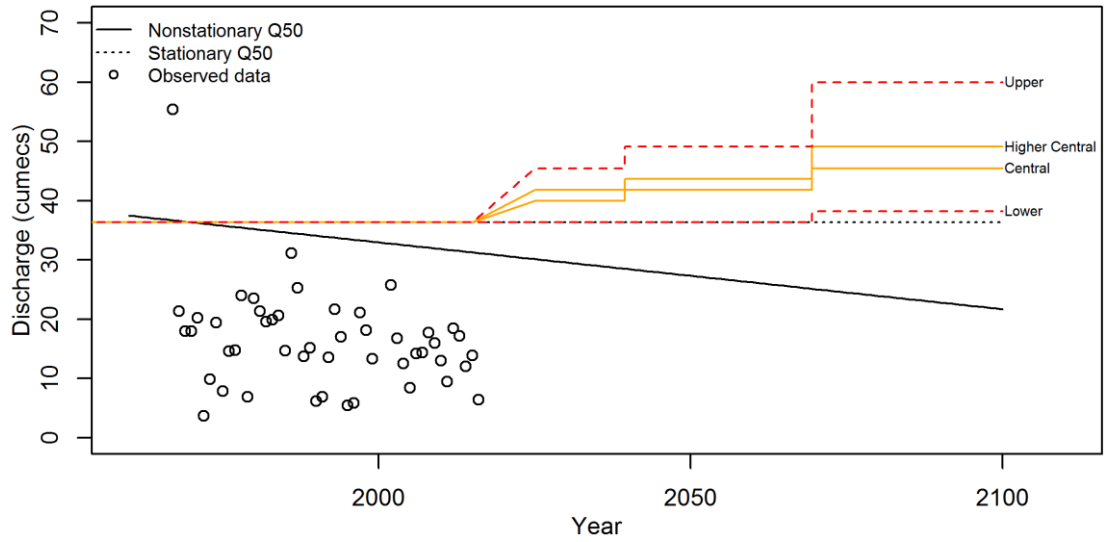


Figure E-11 Comparison of non-stationary Q50 estimates with stationary estimate plus climate change allowance (Little Ouse)

[9] Note that in theoretical and computational terms, the model is fitted all at once, not sequentially as this analogy suggests.

Case study 2: Kennal at Ponsanooth

Similar sets of figures are presented for the Kennal case study below. The AMAX data spans 1968 to 2016. This time the trend in the data is relatively small and there is little difference between the QMED estimate for the 1961 to 1990 baseline period and that for the complete period of record. However, the flood frequency curve for the baseline period lies below that of the full record for return periods from 2 to 50 years. Applying the climate change allowances increases the Q30, Q50 and Q100 estimates above the range of the observed data and above the non-stationary estimates for different time slices. The non-stationary Q50 estimates are broadly in line with the equivalent stationary estimates uplifted by the 'lower climate' change allowance for the 2080s.

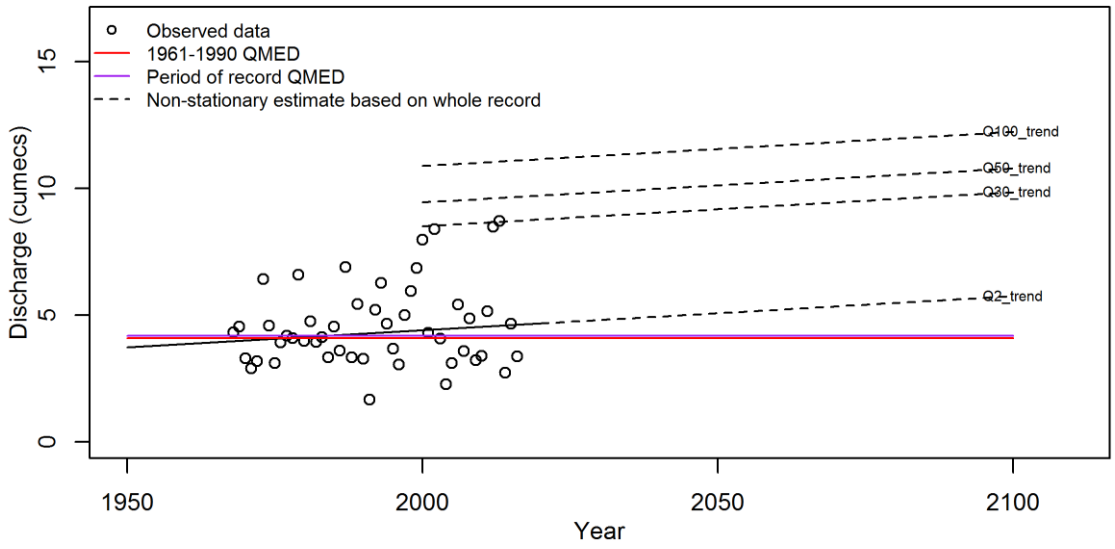


Figure E-12 AMAX data and QMED estimates for 1961 to 1990 and for the whole period of record (Kennal)

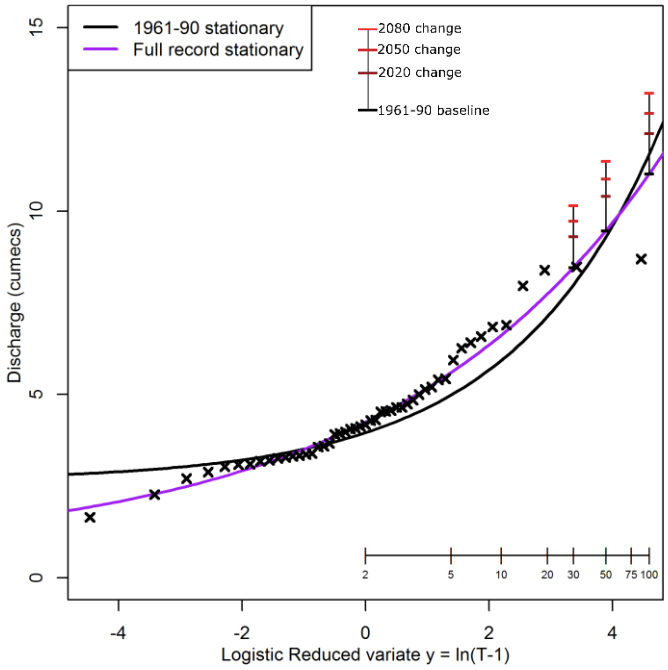


Figure E-13 Stationary flood frequency curves based on different periods of record showing climate change allowances (Kennal)

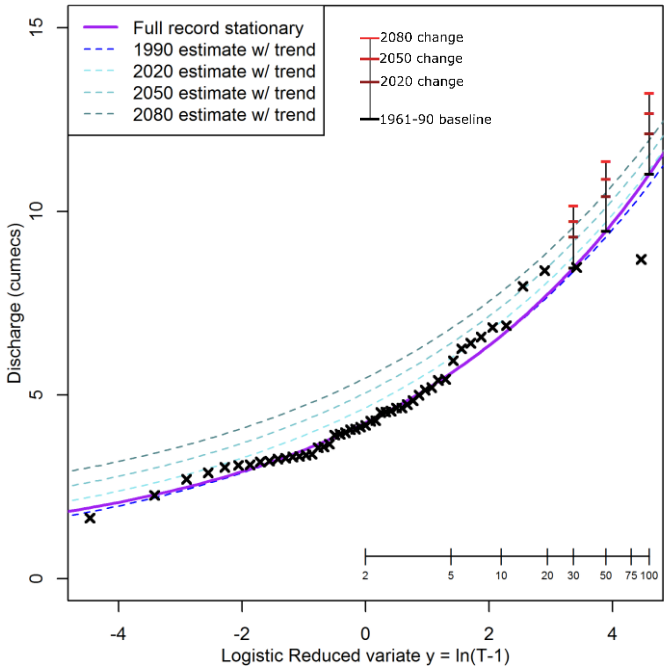


Figure E-14 Comparison of stationary and non-stationary models (Kennal)

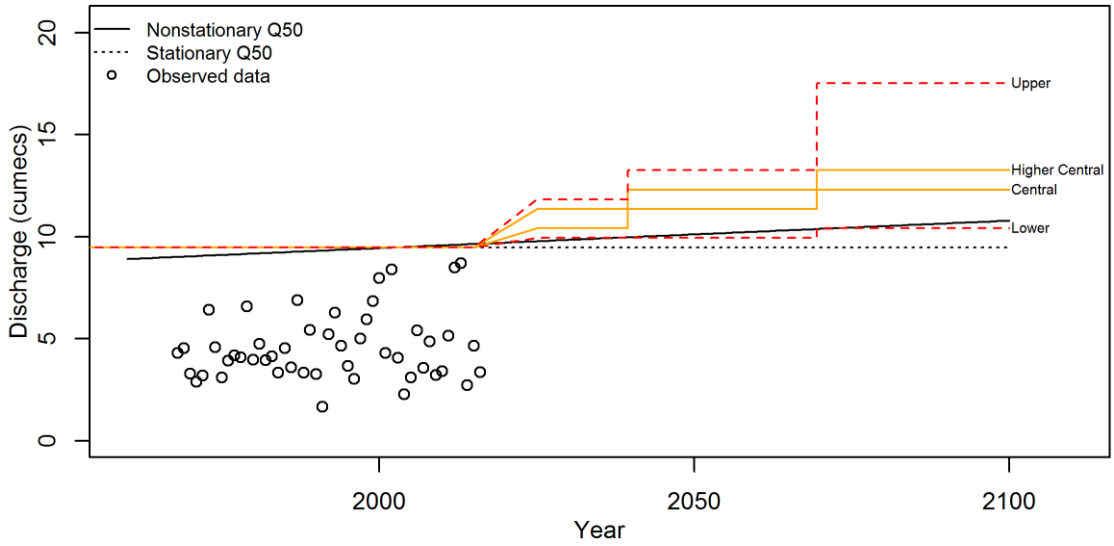


Figure E-15 Comparison of non-stationary Q50 estimates with stationary estimate plus climate change allowance (Kennal)

Case study 3: Eden at Temple Sowerby

Results for the third case study of the Eden at Temple Sowerby are presented in the figures below. The AMAX data spans 1964 to 2016. There is little difference between QMED values calculated over the 1961 to 1990 period and the full period of record. The very high AMAX values recorded in the catchment in recent years are largely responsible for the marked positive trend apparent in the time series. In this example, the highest observations are much closer to the stationary frequency estimates when the 'central' climate change allowances are added and they exceed the non-stationary estimates.

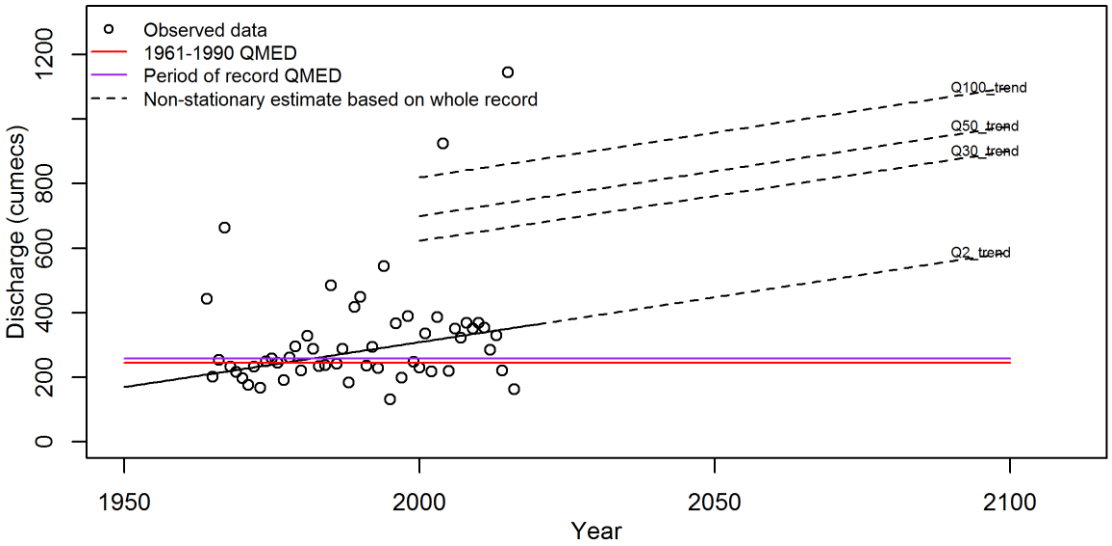


Figure E-16 AMAX data and QMED estimates for 1961 to 1990 and for the whole period of record (Eden)

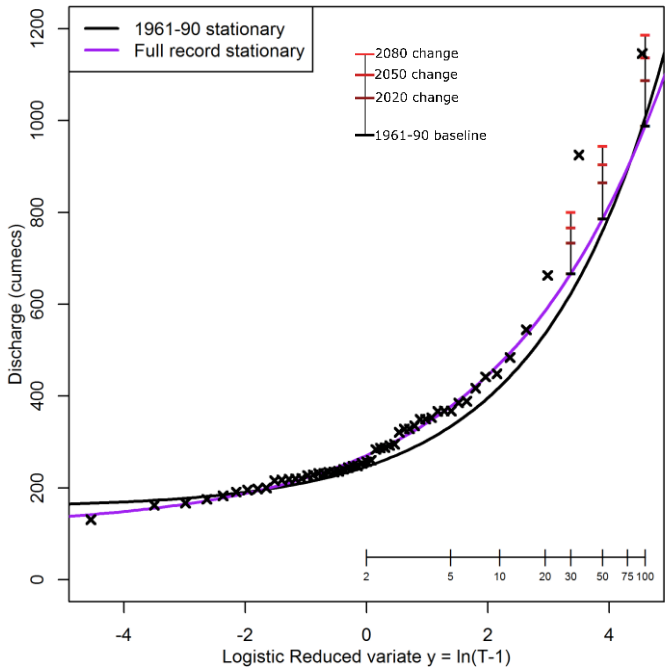


Figure E-17 Stationary flood frequency curves based on different periods of record showing climate change allowances (Eden)

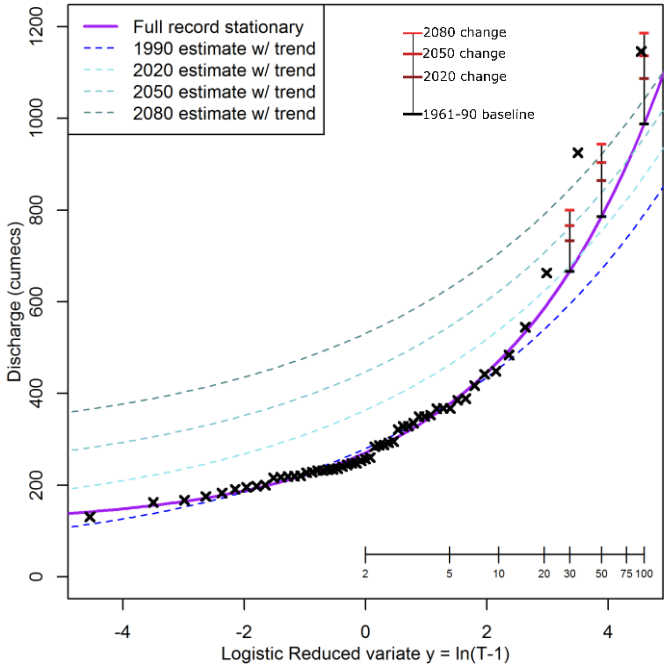


Figure E-18 Comparison of stationary and non-stationary models (Eden)

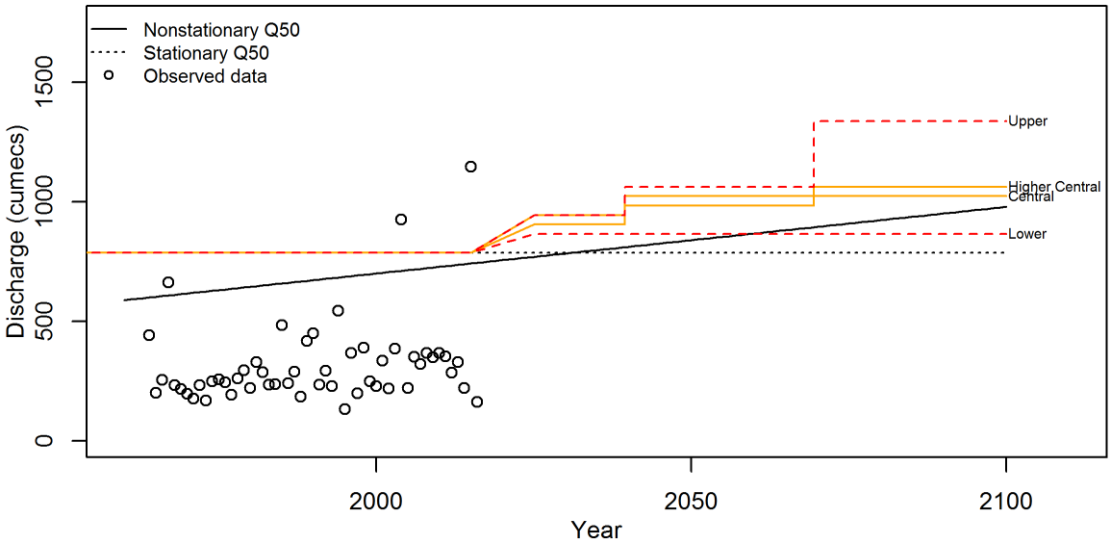


Figure E-19 Comparison of non-stationary Q50 estimates with stationary estimate plus climate change allowance (Eden)

Figure E-19 compares the stationary estimate of Q50 for the Eden catchment with the non-stationary Q50 estimate, assuming that the trend continues into the future. Despite the strength and significance of the trend, the non-stationary estimate lies below that of the stationary estimate with the 'central' allowance for climate change added into the 2080s period.

To assess the uncertainty associated with the flood frequency curves, and the effect that this uncertainty may have on the appropriateness of climate change allowances, 95% confidence

intervals[10] were obtained using non-parametric bootstrapping as developed in Yan and others (2017). Figure E-20 shows this confidence interval for the stationary flood frequency curve based on the whole period of record for the Eden catchment. Here, it can be seen that the confidence interval exceeds the climate change allowances by some margin, especially as the return period increases. This suggests that, although the project team's best estimate gives reasonable allowances, there is still some chance that these allowances will be exceeded, and that such extreme floods are possible (though unlikely) to occur during future engineering design lives.

Figure E-21 shows the 95% confidence interval for the non-stationary flood frequency curve as it appears in 2020. Here, it can be seen that the flatter curve gives a narrower confidence interval, but note that for 2050 and 2080, this whole confidence region will be lifted up to make more extreme flood frequency curves more plausible given the data. The key point to observe however is that, although the allowances are larger than the non-stationary estimates (the point estimates which give rise to the plotted flood frequency curves), the confidence interval greatly exceeds them, offering the possibility that the flood magnitudes in 2020 (or 2050/2080) may be much greater than predicted.

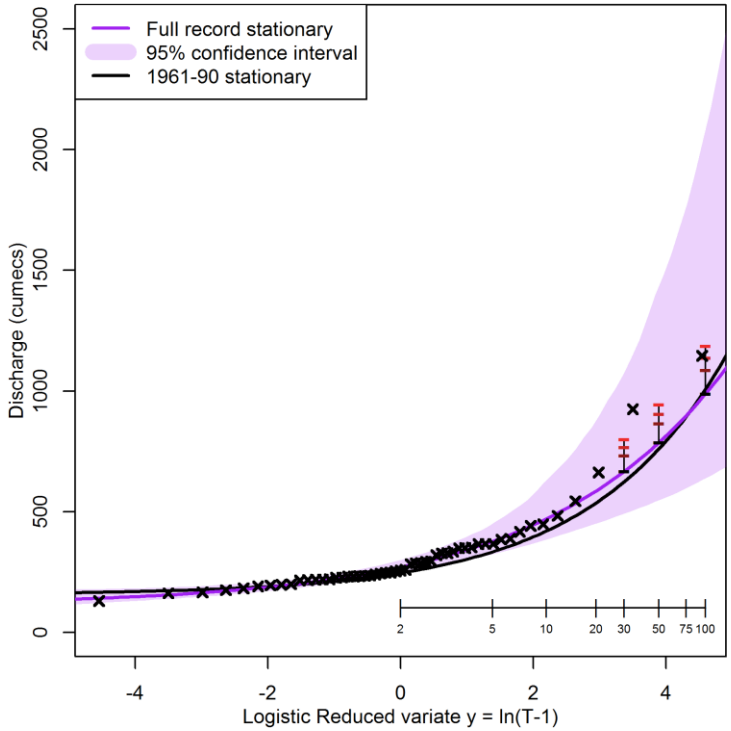


Figure E-20 Stationary flood frequency curve for the full record shown with 95% confidence interval (Eden)

[10] For reference, recall that a 95% confidence interval is a set of bounds which have a 95% probability of being around the 'true' values of the parameter of interest, assuming such a 'truth' exists. In other words, the true value has a 5% probability of being not contained by this band.

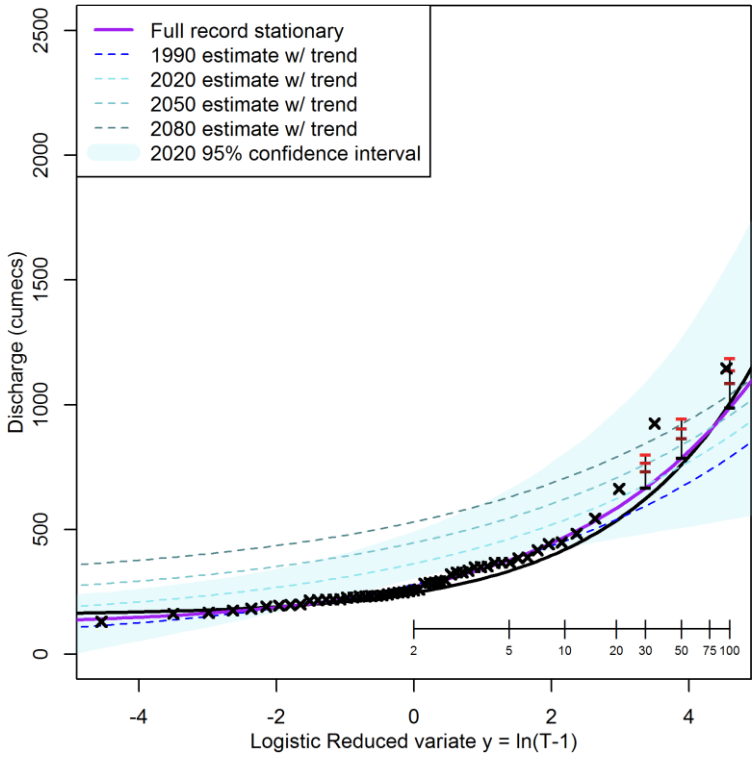


Figure E-21 95% Confidence interval for the non-stationary frequency curve as it appears in 2020 (Eden)

E.4 Discussion

The aim of the investigation presented here was to consider how best to apply climate change allowances to flood frequency estimates for catchments where non-stationarity has been detected in gauged flow records. There are some regional summaries of those methods that perform most similarly to stationary methods, and these have some interesting outcomes.

On average, using baselines other than the 61 to 90 stationary baseline in conjunction with the CCAs leads to smaller estimates of flow compared to STFULL. This is the same for both Q50 and Q100.

Restricting the analysis to only those stations with positive trend does not change this general observation, although the message is more mixed; some positive percentage differences are seen in the south-west and Wales.

Compared to STFULL, NSTEXT has greater negative differences for more distant horizons, suggesting the 2 approaches are diverging for estimating far into the future. However, compared to ST6190 and NST6190, NSTEXT gives more spatially consistent differences, which are also smaller on average for the near-future horizon.

South-east England stands out as a region for which there is the most variability between approaches. This could, however, be because of the greater impact of urbanisation, for example.

NSTREP shows high similarity with STFULL, and so could be considered as a method of compromise between ignoring more recent trends, while focusing on the period of record used to develop CCAs.

For a risk-averse (bigger flood estimates) approach, the present method of applying climate change allowances to a stationary estimate based on the whole period of record gives larger values of Q50 and Q100 on average in the northern regions, compared to the other methods examined except for NSTREP.

Using CCAs on the stationary 61 to 90 baseline is not recommended, as the more recent data is invaluable in giving a more accurate present-day picture.

However, the detailed analysis of only 3 case study catchments with different degrees and directions of trend cannot be easily generalised. Although it is very difficult to make general recommendations, the following comments may help practitioners choose an approach in cases where climate change rather than any other factor is believed to be a driver of non-stationarity.

No evidence has been found to suggest that the existing climate change allowances should be updated other than to apply the new UKCP18 probabilistic climate projections (see below).

To assess whether climate change has already started to affect flood frequency, QMED and/or a higher quantile estimated over the 1961 to 1990 baseline period should be compared to that of the full record. If the 2 differ substantially, the effect of applying the climate change allowances to the baseline and full record estimate should be explored. An alternative approach would be to add a percentage of the full climate change allowance to the stationary estimate from the complete period of record, although it is not clear how this proportion should be chosen.

Central (50th percentile) climate change allowances appear to be appropriate in the examples presented here where no negative trend is observed. However, the uncertainty associated with statistical flood frequency is high, as indicated by the plotted confidence intervals in Figure E-20 and Figure E-21.

For major engineering projects with a long design life, current CCAs give design levels much larger than those determined using a linear extrapolation of current trends. Current CCAs should be used as a risk-averse upper limit.

The existing guidance on climate change allowances for flood peaks (Environment Agency 2016 a,b) is in the process of being updated. The update will be based on a recent Environment Agency-funded project 'Providing more locally-appropriate information on potential impacts of climate change on flood peaks in England and Wales' (Kay and others, 2019), which used the sensitivity framework approach of the previous work (Kay and others, 2011, 2014) but with a national-scale grid-based hydrological model. The project applied the UKCP18 probabilistic projections for river-basin regions (Met Office Hadley Centre 2018), to assess the potential range of impacts of climate change on flood peaks on a 1 km grid across GB (for non-tidal 1 km cells with catchment area ≥ 100 km^2). This was done for 3 future 30-year time slices (2020s, 2050s, 2080s) and 4 emissions scenarios (RCP2.6, RCP4.5, RCP6.0, RCP8.5). As for the previous work, the baseline period for the projections was 1961 to 1990, with a longer baseline for the hydrological modelling (Oct 1961 to Sep 2001). The intention is for the outputs of the project to be made available via a web tool. However, the Environment Agency has not yet made any decisions on what information will be used, or in what form, regarding updates to guidance on flooding and climate change.

Appendix F: Investigating clusters of floods over time

F.1 Introduction

Background and scope

Better understanding of clustering and flood-rich or flood-poor periods may help attribute trends and therefore improve knowledge of whether and how they might continue. It could help determine whether the project team is relying on an unrepresentative period spanned by its peak flow data sets. Knowledge of clustering on a shorter timescale may help plan response and communication in the aftermath of flood incidents, and perhaps have implications for design of storage-based schemes.

Within the current project, the scope for investigating clustering has been limited to an initial exploratory analysis. The following tasks are included in the scope:

- literature review covering both clustering and also identifying flood-poor and flood-rich periods using longer-term sources of information
- data set development, screening long-term series of POT (peaks over threshold) data
- calculating clustering indices
- reporting

Evaluating temporal clustering - index of dispersion

A commonly used approach to investigating the presence of clustering of events in time series is the index of dispersion (*D*), defined as:

$$D = \frac{\mathbb{V}ar\big(Z(T)\big)}{\mathbb{E}\big(Z(T)\big)} - 1$$

where *Z(T)* is the series of POT counts, that is numbers of floods, within a time window of length *T*, $\mathbb{V}ar\big(Z(T)\big)$ is the variance of the flood counts and $\mathbb{E}\big(Z(T)\big)$ is the expected (mean) value.

This methodology interprets the occurrence of a flood event as a point process in which a randomly occurring event is independent of any events that may have occurred previously. This is referred to as a homogeneous Poisson process, and the degree to which a series of events conforms to the homogeneous Poisson can be evaluated using the index of dispersion.

In a time series where events occur more regularly than might be expected at random, *D* is negative. This is described as underdispersion. Time series with events that are more clustered than random exhibit overdispersion and *D* is positive.

For readers familiar with the FEH, *D* is used in the FEH (Volume 3, section 12.3.3), although defined without the minus 1, so FEH values of *D* will be 1 higher. The FEH calculated a UK-average value of 1.38 for *D* (that is, 0.38 using the definition above), using a 1-year time window.

The following section provides a summary of literature that has investigated temporal clustering of flood events, using the index of dispersion or alternative methodologies.

F.2 Literature review

Temporal clustering

Villarini and others (2012): On the temporal clustering of US floods and its relationship to climate teleconnection patterns. International Journal of Climatology.

Data used:

- 41 non-nested gauges in Iowa on the Missouri and Mississippi rivers. Maximum period of record 1950 to 2009 (30/41 gauges); shortest record is 38 years.
- POT defined flood events with thresholds selected to achieve an average of 2 events per year in the period March to October only; 15-day independence required. Assume stationarity based on previous trend investigations.
- Climate indices PNA, NAO used as covariates as 14- and 28-day means from daily values.
- Rainfall also used as a covariate for investigating 2 specific flood events

Temporal clustering methods:

- Assume homogeneous (Poisson) point process as the null hypothesis.
- Use Cox regression model with time-varying covariate (climate indices) describing the variation of rate of POT event occurrence from Poisson process.
- Utilise ‘survival’ R package.
- Interaction between covariates is allowed; AIC used to define formulation of covariates in final model.

Results:

- Time-varying model preferred at 27/41 stations.
- Climate index covariates in monthly form are best descriptors of variability.
- Cox model based on covariates can be used in a forecast sense to predict above/below normal floodiness.

Merz and others (2016): Temporal clustering of floods in Germany: Do flood-rich and flood-poor periods exist? Journal of Hydrology.

Data used:

- 68 catchments across Germany, 1932 to 2005 period; 4 gauges with > 160-year records.
- The minimum record length was 70 years, and even then, the authors concluded that was not enough information to detect whether there is significant clustering at the multi-decadal scale.
- POT defined flood events with thresholds selected to achieve an average of 1 and 3 events per year, then also 1 event every 3, 5 and 10 years.

Temporal clustering methods:

- Global view: Presence of temporal clustering - index of dispersion for varying time windows with Monte Carlo based significance testing [Method 1].
    - Time windows investigated: 1, 2, 3, 5 and 7 years for all series, and also 10 years for the 4-long series.
    - Field significance test (false discovery rate approach) also used to evaluate proportion of false rejections of null hypotheses.
    - All significance evaluated at the 5% level.
- Local view: Timing of temporal clustering - Time-variation of flood occurrence rate (kernel occurrence rate) with significance of clustering based on non-parametric [Method 2] and parametric [Method 3] tests
    - Time windows (as bandwidths) investigated: 1, 3, 5 and 10 years for all series.
    - Non-parametric significance evaluated using confidence intervals (CI) determined for the time-varying kernel occurrence rate through bootstrap approach. Departures of upper/lower CI bounds below/above Poisson occurrence rate indicates significant flood-poor/flood-rich clustering in the specified time window (bandwidth) at the indicated point in the time series.
    - Parametric significance evaluated again using CI defined for the Poisson occurrence rate (time-invariant). Departures of kernel occurrence rate beyond the lower and upper CI indicates significant flood-poor/flood-rich clustering in the specified time window (bandwidth) at the indicated point in the time series.

Other methodology notes:

- Seasonality of clustering also investigated; comparisons between annual, winter and summer results presented.
- Spatial variation of temporal clustering is also presented.

Comments on methodology:

- Different methods give different results; recommend using multiple measures.
- Method 1 cannot identify flood-rich or flood-poor periods, only the presence in the time series of clustering of events in time windows.
- Method 3 is more conservative at identifying clustering, especially for flood-poor periods.
- Method 2 may be biased towards finding flood-poor periods as the CI narrow in periods with below average occurrence of floods.

Results for Germany:

- Across Germany, method 1 finds clustering at most (>50%) gauges in the annual time series with the lowest threshold (3 events in one year), but at very few gauges (<=10%) at the higher threshold (1 event in 5 years). This signal is stronger/weaker in summer/winter.
- By methods 2 and 3, show >70% gauges show clustering on the interannual timescale regardless of the threshold. Also, the proportion of gauges showing clustering with either method 2 or 3 decreases with threshold and time window (bandwidth). Less difference is evident between winter and summer with methods 2 and 3.
- For low thresholds and short time windows, clustering is very pronounced. Hypothesised to be related to catchment memory effects and intra- and interannual climate variability. The paper postulates that catchments with deeper soils and more aquifer storage show more persistence. Small floods may cluster due to hydrological memory effects, over a short time scale, and large floods due to low-frequency climatic variations.

Gu and others (2016): Temporal clustering of floods and impacts of climate indices in the Tarim River basin, China. Global Planet. Change.

Review based on abstract only.

Data used:

- POT series for Tarim River basin, China.
- Climate indices AO and NAO used as covariates as monthly means.

Temporal clustering methods:

- Dispersion index to investigate interannual clustering of floods.
- Monthly frequency analysis to identify seasonal clustering of floods.
- Cox regression with climate indices as covariates to investigate intra-annual clustering of floods.

Results:

- Temporal clustering of floods is present on intra-annual timescales; flood events are not independent.
- Seasonal clustering of floods in June to August.
- Inter-annual clustering of floods more evident when pooling gauges regionally than at individual stations; regional floods are temporally clustered (over-dispersed) on inter-annual timescales.

Liu and Zhang (2017): Multi-temporal clustering of continental floods and associated atmospheric circulations. Journal of Hydrology.

Data used:

- 413 unregulated catchments across Australia for the period 1976 to 2010. Excluding 367 catchments with >10% missing data, and any years with >15% missing data. Missing data gap-filled using modelled flows.

- POT defined flood events with thresholds selected to achieve an average of 1, 2 and 3 events per year, then also 1 event every 2 years.
- Climate indices ENSO, IOD, IPO, SAM used as covariates as 14- and 28-day means from daily values.

Temporal clustering methods:

- Intra-annual:
  - Cox regression model with climate indices as covariates; AIC used to identify optimal set of covariates. Utilise 'survival' R package.
  - Monthly frequency analysis to identify the relative proportion of POT events by month to identify seasonal floods.
- Inter-annual:
  - Dispersion index for varying time windows with significance evaluated using the Lagrange multiplier statistic at 95% significance level.
- Time windows investigated: one to 5 years.
  - Kernel occurrence rate with significance of clustering based on a non-parametric test.
- Non-parametric significance evaluated using confidence intervals (CI) determined for the time-varying kernel occurrence rate through bootstrap approach. Departures of upper/lower CI bounds below/above Poisson occurrence rate indicates significant flood-poor/flood-rich clustering in the specified time window (bandwidth) at the indicated point in the time series.
- Cross-validation is used to identify appropriate time windows (bandwidths).

Other methodology notes:

- Present a summary table of studies investigating temporal clustering.

Results:

- Intra-annual:
  - Cox regression identifies that ENSO and SAM are most important covariates on sub-annual timescales. For AMAX and higher threshold, IOD and IPO are comparable.
  - Monthly frequency analysis picks out monsoon flooding in tropical north Australia in austral summer and austral winter/spring in southern Australia.
  - Results suggest the flood occurrence is non-independent and temporal clustering exists within one year.
- Inter-annual:
  - For Australia as a whole, temporal clustering is evident at the majority (>50%) gauges in all time windows for AMAX and lower thresholds. Regionally, temporal clustering is most evident in southern gauges.
  - Flood-poor periods are identified in most southern gauges during the 2000s and flood-rich during the early 1990s. Northern gauges tend to be opposite to this.

Liu and others (2017): Nonstationarity and clustering of flood characteristics and relations with the climate indices in the Poyang Lake basin, China. Hydrological Sciences Journal.

Data used:

- Daily streamflows at 6 gauges in Poyang Lake basin, China, covering the maximum period of 1952 to 2005. At one station, 2 years of missing data were filled through linear regression using neighbouring stations.
- POT, AMAX, seasonal maximum flow. POT defined to achieve an average of 2.4 to 3 events per year, depending on the station.
- Climate indices ENSO, NAO, IOD, PDO considered possible influential variables.

Temporal clustering methods:

- Monthly frequency analysis to identify the relative proportion of POT events by month to identify seasonal floods.
- Dispersion index with confidence intervals determined using bootstrapping and significance evaluated using the Lagrange multiplier statistic at 95% significance level.
- Kernel occurrence rate with significance of clustering based on a non-parametric test.

Other methodology notes:

- Trends investigated with the modified rank-based nonparametric Mann-Kendall test.
- Change points investigated with the rank-sum test.
- POT thresholds selected to ensure the distribution conforms to a Poisson distribution.
- Influence of climate modes of variability of analysed through Pearson correlation coefficients of annual/seasonal climate indices (lag 0 and lag 1-year) with magnitude, occurrence rate and timing of annual and seasonal floods.

Results:

- AMAX, autumn and winter seasonal maximum floods, and POT peak flows mainly exhibit an increasing tendency.
- Seasonal variations in trends are evident and trends vary between stations.
- Flood-rich periods identified during late 1960s to early 1970s and mid-1990s.
- A persistent increase in flood occurrence rates can be detected after the 1990s, especially in 1998.
- Dispersion index indicates presence of temporally clustered floods at all stations.
- Strong seasonal clustering of floods evident from the timing of AMAX and POT floods, with a peak April to July.
- Significant correlations between flood indices and climate indices exist.
- ENSO and IOD are the main climate indices having significant impacts on flood activities of the basin. ENSO has a significant impact on the flood occurrence rate and on annual maximum streamflow during spring.

Mallakpour and others (2017): On the use of Cox regression to examine the temporal clustering of flooding and heavy precipitation across the central United States. Global Planet. Change.

Review based on abstract only.

Data used:

- Daily streamflow and precipitation in central USA.
- Climate indices AO and PNA used as covariates.

Temporal clustering methods:

- Cox regression applied to both streamflow and precipitation with climate indices as covariates.

Results:

- Climate indices (either or both) influence temporal clustering at 78% of streamflow gauges; results are consistent for precipitation also.
- Conclusion is robust regardless of thresholds applied to define flood events.

Alternative methods of time series analysis

Wu and Lye (1994): Identification of temporal scaling behaviour of flood: A study of fractals. Fractals.

Review based on abstract only.

Data used:

- Daily flow records (107 years) at Yichan Station on the Yangtze River, China.

Methodology:

- Functional box counting procedure evaluating threshold exceedance.

Results:

- The number of threshold exceedances results in a power law behaviour for a specific range of time intervals, that is, threshold exceedance counts do not conform to the assumption of an independent Poisson process.
- Evaluation of saturation and break-points in the probability-scale-threshold can yield physically meaningful results.

Eastoe and Tawn (2010): Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series. Water Resources Research.

Data used:

- POT series for River Thames at Kingston derived from daily flow series assuming an average of 3.3 events per year.
- Daily rainfall data at a site in Oxford used to test dependence of flow data on covariates.

Methodology:

- Does not assume homogeneous (Poisson) point process as the null hypothesis due to empirical evidence to the contrary (the variance in POT counts is usually larger than the mean).
- Investigate other models to describe the expected POT distribution, for example, regression and mixed models, retaining some assumptions of the Poisson process, but modelling events as inhomogeneous with a time-varying parameter.

Results:

- Autocorrelation of an example POT series suggests dependence. Postulated that this is due to dependence in underlying covariates (climatic influences) rather than in the POT counts themselves.
- Inclusion of time-varying parameter and use of covariates in regression models improves model fit and therefore confidence in extrapolated return levels.

Clustering over the longer term: historical flood chronologies

Macdonald N and Sangster H. 2017 'High-magnitude flooding across Britain since AD 1750. Hydrology and Earth System Sciences, 21(3), 1631 to 1650

Data used:

- Detailed historical records, merged with systematic river flow measurements, spanning the period 1750 to 2014.
- Study covered 12 catchments across England, Wales and Scotland.

Results:

- discernible flood-rich periods were found at a national and catchment-scale
- these included 1765 to 1780, 1850s, late 1940s, mid-1960s and 2000 to present
- the recent flood-rich period (considering data up to 2014 only) is not unprecedented
- historical patterns of flooding were found to be linked to drivers including the NAO, the Atlantic Meridional Oscillation (AMO) and solar activity

Clustering over the short term: flood-rich seasons

Several recent years have seen multiple floods affecting the same locations in the UK. For example, 2012 saw repeat flooding in locations as widespread as Calderdale and Devon. Over the winter 2015 to 2016, a series of storms brought flood after flood to locations in Cumbria and elsewhere in north-west England. The paper below describes a method developed in response to these floods.

Towe R, Tawn J, Eastoe E and Lamb R. 2019 'Modelling the Clustering of Extreme Events for Short-Term Risk Assessment JABES. https://doi.org/10.1007/s13253-019-00376-0

Key points:

- develops a 'relative risk' measure to account for short-term increased likelihood of another flood occurring in the aftermath of a flood
- uses covariates and random effects to account for short-term non-stationarity of the flood-generating process

  This sort of research can potentially help equip flood risk managers to better manage sequences of floods. However, it has been excluded from the scope of the present study.

Key points from literature review

Summary of methodologies

In the literature reviewed above, the index of dispersion is frequently used to identify and quantify the degree of temporal clustering in POT series. Evaluating the dispersion index using several sizes of time window means the most dominant clustering timescales can be quantified. This index, however, can only identify the presence of, and degree of, clustering, but not the location of temporal clusters in the tested time series.

To identify relative flood-rich and flood-poor periods, the kernel occurrence rate analysis is frequently used in the literature above. Non-parametric confidence intervals, based on the bootstrap approach applied to the kernel occurrence rate estimate, is most frequently used to identify significant departures from the homogeneous Poisson occurrence rate. However, Merz and others (2016) showed that this method could be biased toward flood-poor periods as the confidence intervals are narrower during these periods due to a lower sample size. Therefore, using parametric confidence intervals, based on applying the bootstrap approach to the homogeneous Poisson occurrence rate, is also recommended. Merz and others (2016) also recommend using multiple methods to investigate temporal clustering as each method may give slightly different answers. Conclusions can then be drawn from robust commonalities between the methods and individual results, considering the caveats of each method.

Several studies investigate temporal clustering in different seasons, as the processes underlying the clustering may be different. The dispersion index or the monthly frequency analysis method can be used to investigate this.

Several studies also investigated the spatial variability of temporal clustering, frequently grouping individual stations and interpreting the results over the members of each region. This generally led to more robust signals of temporal clustering making flood-poor and flood-rich periods better defined.

The variability of flood frequency and attributing temporal clustering to atmospheric influences was frequently investigated using the Cox regression approach. In this methodology, climate indices or precipitation are used as time-varying covariates in a regression model, along with the time-varying flood occurrence rate. Selection of the optimal (set of) covariate(s) is tested, usually, using the AIC. Note this approach can be used to investigate intra-annual as well as inter-annual temporal clustering. Additionally, the Pearson correlation coefficient can be used to quantify the degree of similarity between a possible covariate (including at lagged timescales) and the flood frequency series. The benefit of the Cox regression model over the Pearson correlation coefficient is that it can also be used to predict likely future flood-rich or flood-poor periods.

Eastoe and Tawn (2010) note that a null hypothesis using the homogeneous Poisson is unwise as empirical evidence has shown it is generally not valid; POT counts are nearly always over-dispersed (variance larger than mean). This implies that POT series do exhibit clustering, the reasons for clustering in POT series require investigation. Eastoe and Tawn (2010) have suggested some approaches for characterising and investigating both POT and AMAX series where dependence in the series is attributed to covariates (for example, climatic influences). Merz and others (2016) have also suggested that of catchment memory effects influence clustering in POT series, which suggests POT series, and potentially AMAX series, are not comprised of independent events. Therefore, it may also be wise to approach the topic of temporal clustering, and flood frequency analysis, using extremal dependence approaches, or from a completely different angle, such as multifractal analysis, in order to test if conclusions really are robust.

Note on data quality for temporal cluster analysis

For the purposes of the temporal clustering analysis, only the most reliable and consistent records should be used at present. The full impact of the presence of missing data in records has not been investigated.

While in many of the studies summarised above, the POT series has been derived specifically for the study from daily flow records, the current study makes use of the POT series derived a priori. In some cases, the POT series for a given station may originate from a number of different sources, where the specification of the threshold used based on a desired POT frequency may not be consistent. The Merz and others (2016) study has demonstrated that measures describing temporal clustering can be sensitive to changes in the POT frequency, the record length, and to the starting point of the record. If there are artificially-introduced inconsistencies in the frequency of POT events in the records, the conclusions derived from the temporal cluster analysis will not necessarily be robust.

F.3 Summary of data review

The peaks over threshold (POT) data used in this study originate from the NRFA Peak Flow Dataset - Version 8. Since the FEH methods largely use AMAX data, relatively little effort has been put into improving the quality of the national POT data set in recent years. Problems with POT data can be more difficult to spot than those with AMAX: for example, it is obvious if an AMAX value is missing.

The steps carried out to screen the data set are described in the additional note. After this, a set of 16 stations with suitably long and reliable POT series was retained for analysis.

Any national-scale study using POT data would require a major effort on data quality.

F.4 Temporal clustering analysis

Using an R script with the NRFA Peak Flow Dataset V8 and the final selected station list as inputs, the POT files for final selection of stations were read in and the POT series restricted appropriately (rejected POT values and duplicated dates removed and series truncated, if needed).

Further to this, the percent complete for each water year for the remaining continuous water year period was determined. If any year had a percent complete below 69.5% (relaxed from 70% to accommodate a single deviation at one station) or if the record length did not exceed 60 years, the station was rejected. This resulted in an additional 2 stations being rejected from the temporal cluster analysis:

- station 27023 (Dearne at Barnsley Weir) - record length 53 years
- station 32003 (Harpers Brook at Old Mill Bridge) - record length 53 years

The stations for which temporal clustering results are presented are shown in Figure F-1.

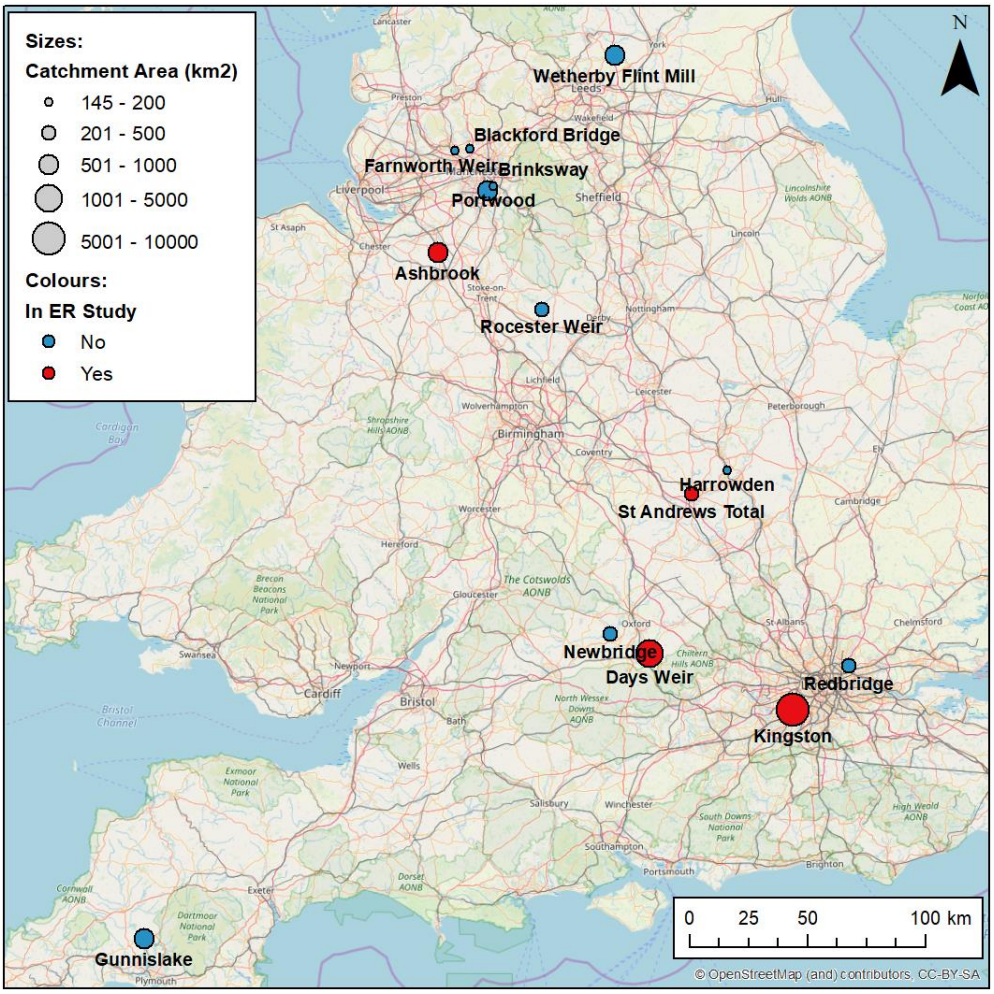


Figure F-1 Stations with temporal clustering results presented. Stations are symbolised with varying colours dependent on their use in an earlier unpublished study by Emma Raven[11], and by varying sized symbols dependent on their catchment area

For the current study, the temporal cluster analysis is limited to an investigation of the index of dispersion.

The methodology of the dispersion index analysis generally follows that of Merz and others (2016), whereby the time series of aggregated POT counts ($Z(T)$) is derived for multiple time windows (that is, multiple values of T) as a running sum of annual POT counts within the width of the specified time window (that is, the value of T). Note the use of a running sum may differ from the methodology used by Merz and others (2016), who appear to have constructed the series for each value of T as discrete block sums of POT counts.

Merz and others (2016) found that the dispersion index and associated significance can be sensitive to both the length of the underlying annual POT series, and also the starting point of the series. Therefore, following Merz and others (2016), considering this sensitivity is also incorporated into the current study. For each time window or value of T, the dispersion index is calculated T times, with the starting point of the series shifted by +1 year each time, such that the POT series is 1 year short with each integration in the range 1 to T.

The significance of all calculated dispersion index values is evaluated using Monte Carlo simulations, following Merz and others (2016). For each series ($Z(T)$), 1,000 comparable series of

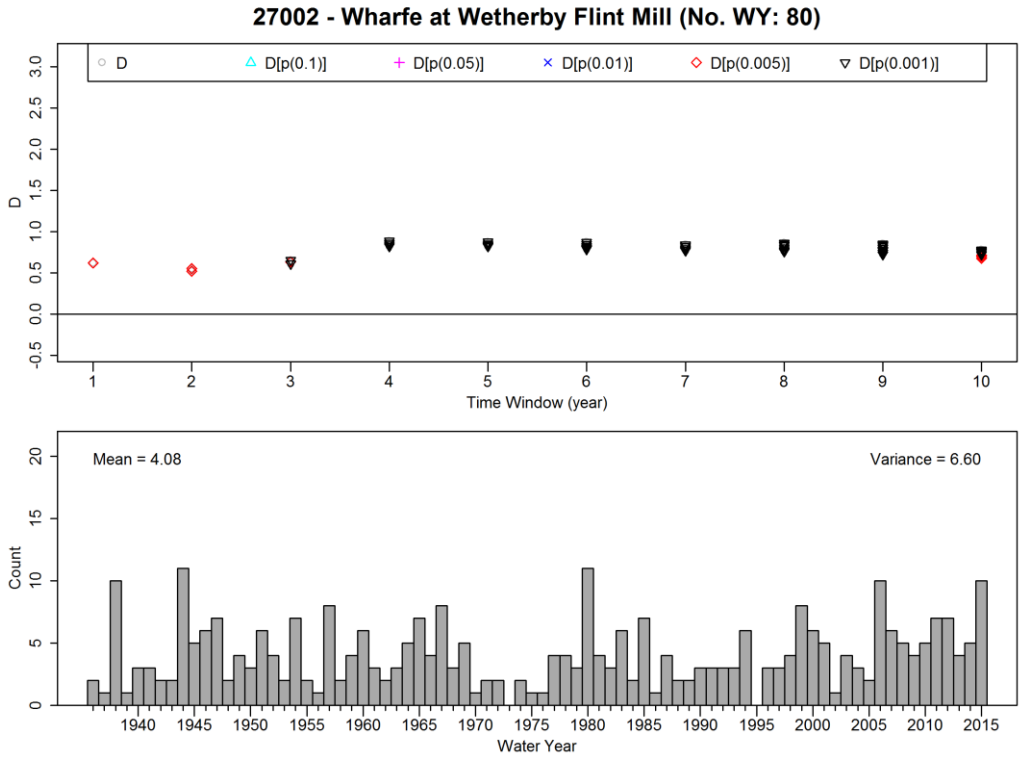
11 Raven EK, Vitolo R, Lane SN, Stephenson DB (unpublished): 'Characterising high river flow clustering in the UK'

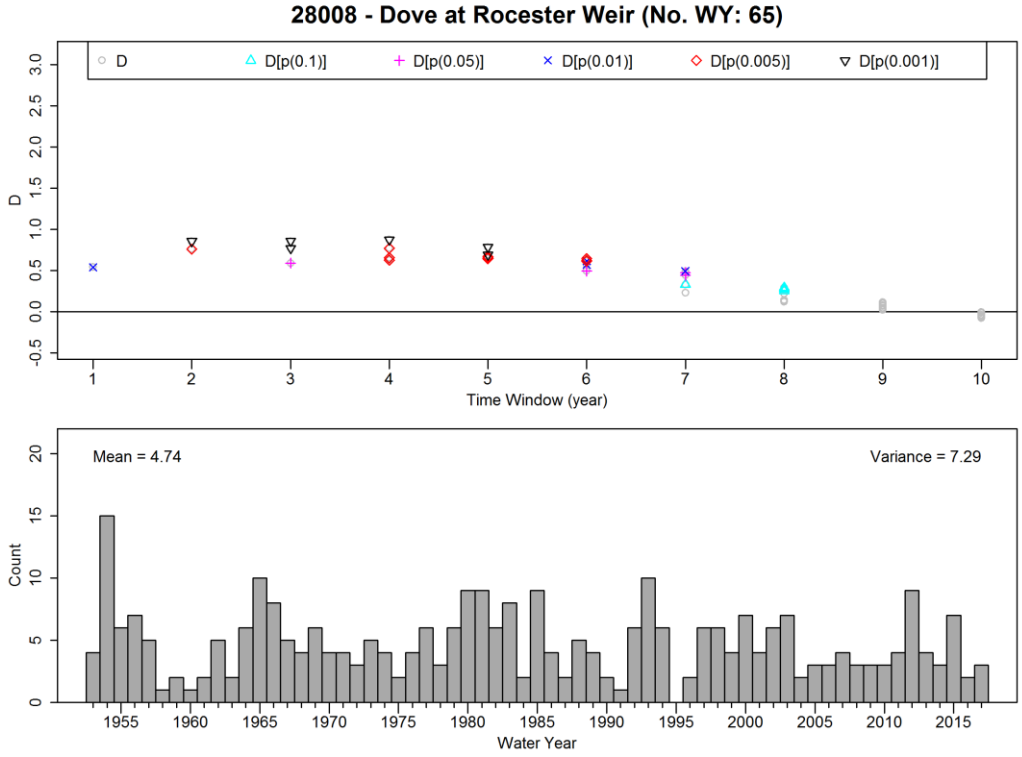
the same length as *Z(T)* are generated by drawing randomly from a Poisson distribution derived using the mean of *Z(T)*. The critical value of D associated with a specified significance level (several are used) is identified using the 1,000 samples, and the null hypothesis (*D*=0) rejected if *D* for *Z(T)* exceeds the critical value. In this situation, significant overdispersion is present in *Z(T)*, therefore events exhibit clustering within the specified time window (*T*) at some point in the series.

The time windows used to evaluate temporal clustering using the dispersion index by Merz and others (2016) were less than 7 years for series up to 74 years. For longer series (>160 years), a time window of 10 years was also used. As the appropriateness of window size the series length is at present a subjective decision, all window sizes in the range 1 to 10 years have been evaluated here, but caution should be used for all windows greater than, for example, 5 years.

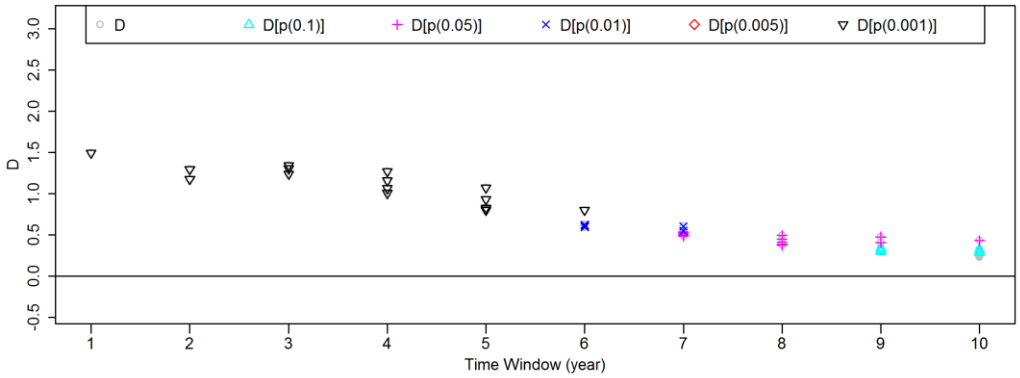
The index of dispersion is presented below as a function of window size for all stations for which the dispersion analysis was completed. For each window size, the various values of D for each starting point are shown, with the maximum significance level indicated by the type of symbol and colour. For reference, the annual POT count series used in the dispersion analysis (that is, not necessarily the full series available) is also presented for each station, with the mean and variance for that series annotated.

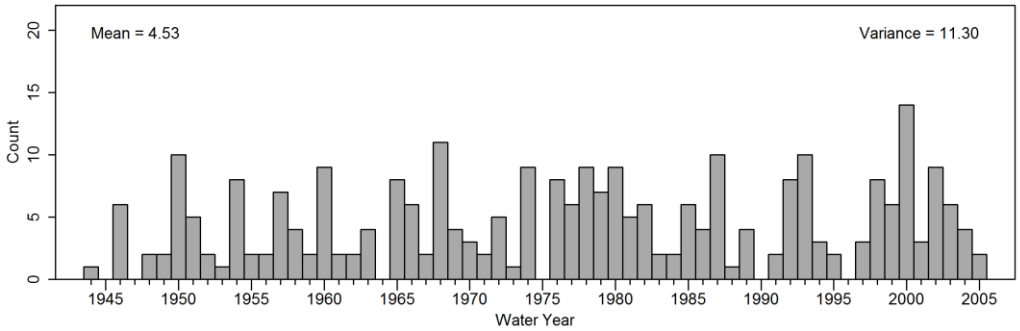
Information for each station relating to the dispersion analysis is summarised in Table F-1. Included in this summary is a note of any trends (non-stationarity) identified at the station based on other work in the current non-stationarity project. The variance to mean ratios in Table F-1 are calculated for the annual POT counts.


27002 - Wharfe at Wetherby Flint Mill (No. WY: 80)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 4.08
Variance = 6.60
Count
Water Year

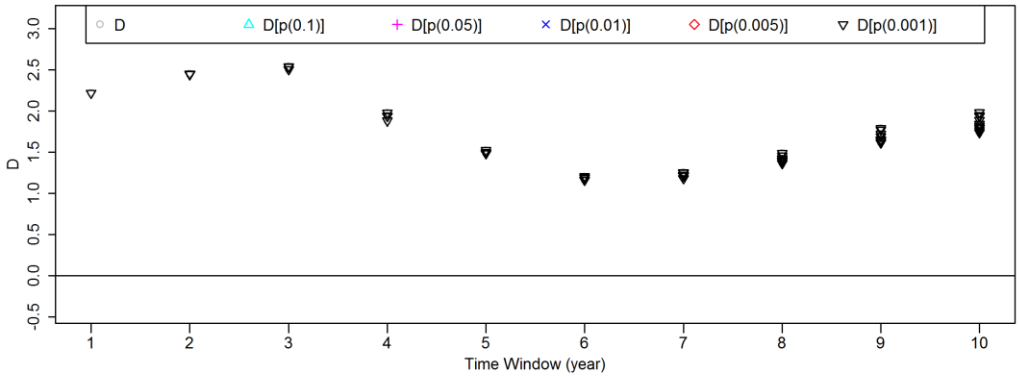

28008 - Dove at Rocester Weir (No. WY: 65)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 4.74
Variance = 7.29
Count
Water Year

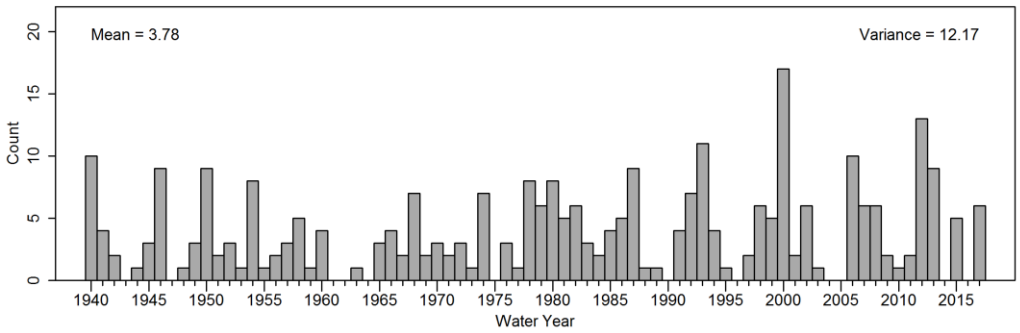
32004 - Ise Brook at Harrowden (No. WY: 62)

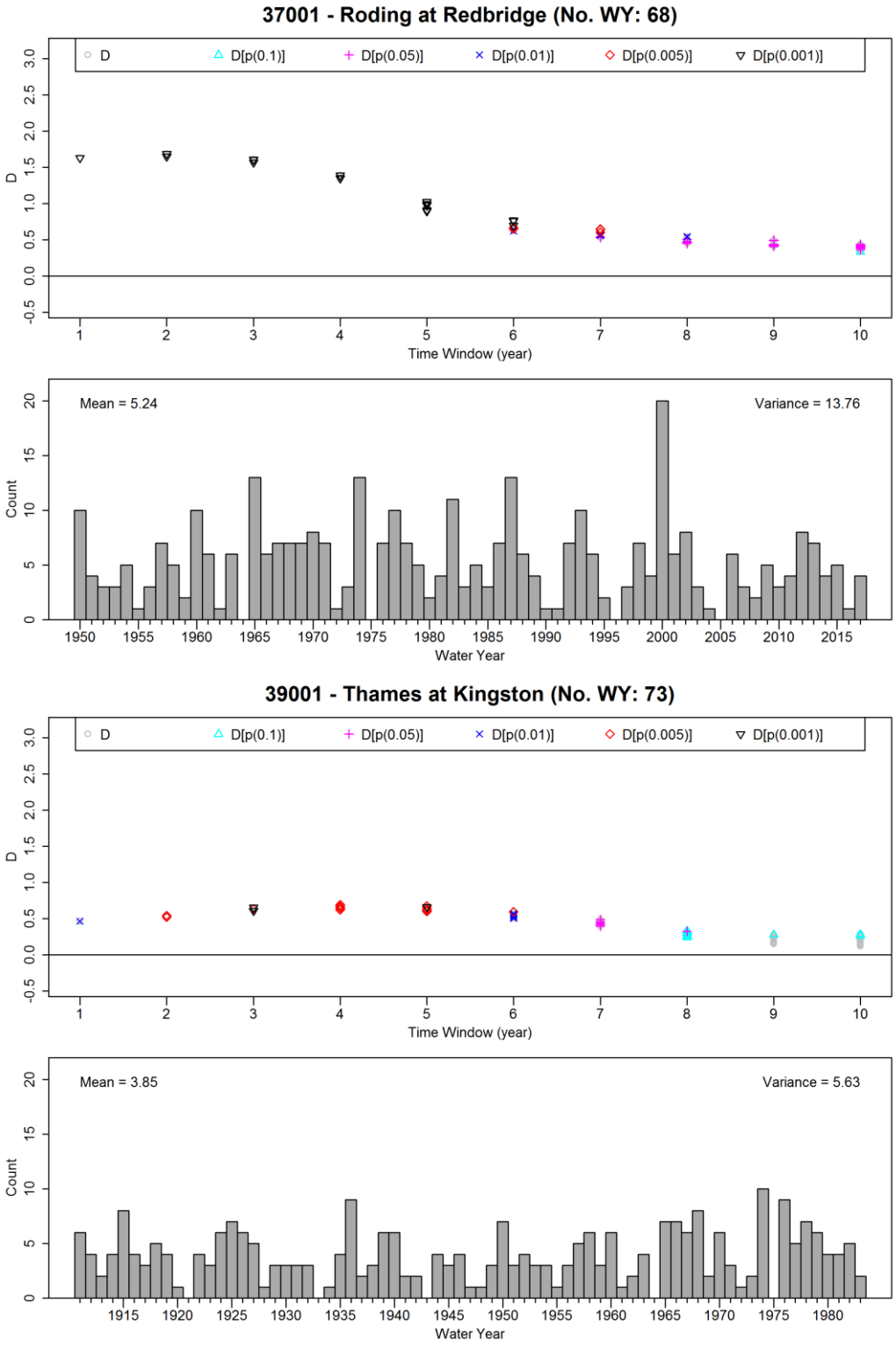




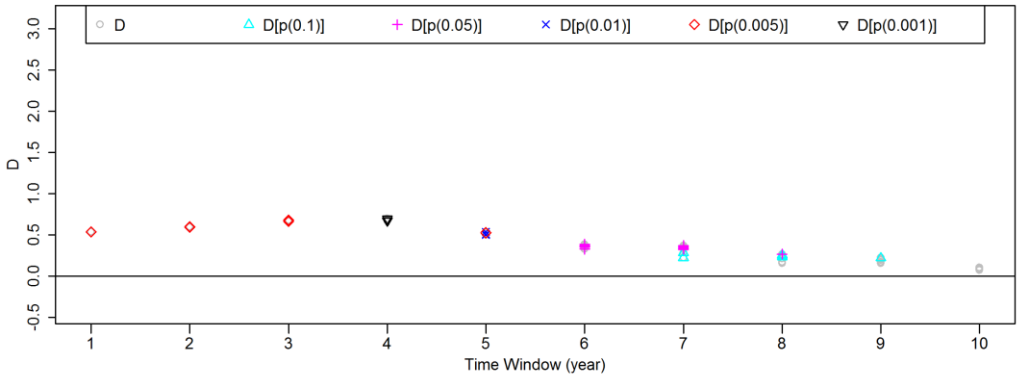
32007 - Nene/Brampton at St Andrews Total (No. WY: 78)

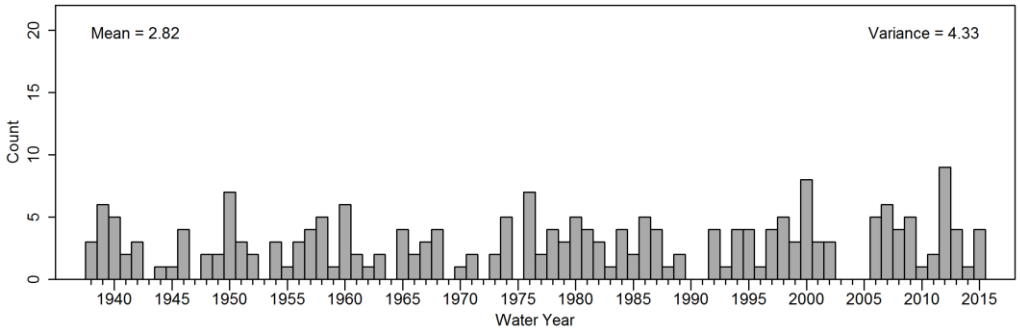




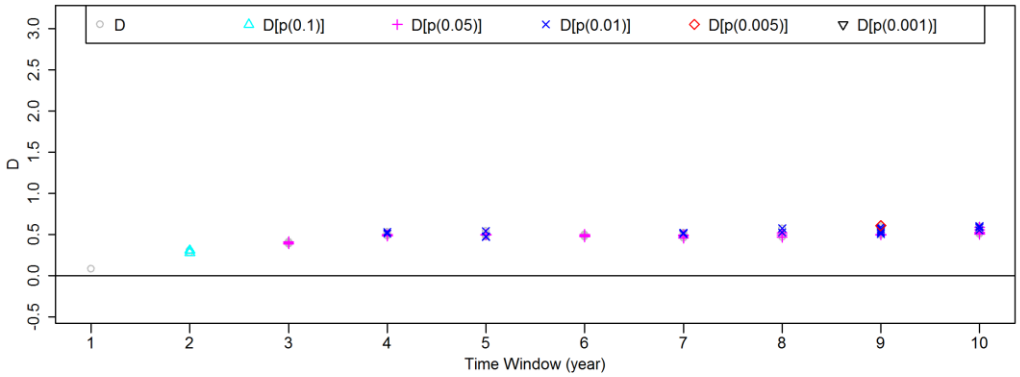

37001 - Roding at Redbridge (No. WY: 68)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 5.24
Variance = 13.76
Count
Water Year
39001 - Thames at Kingston (No. WY: 73)
Mean = 3.85
Variance = 5.63

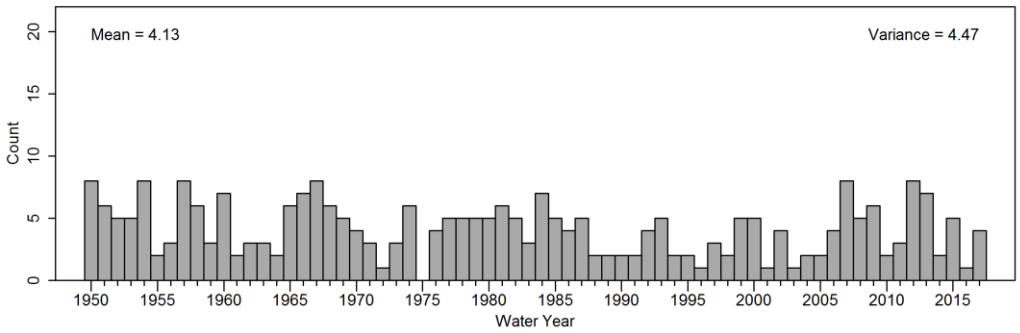
39002 - Thames at Days Weir (No. WY: 78)

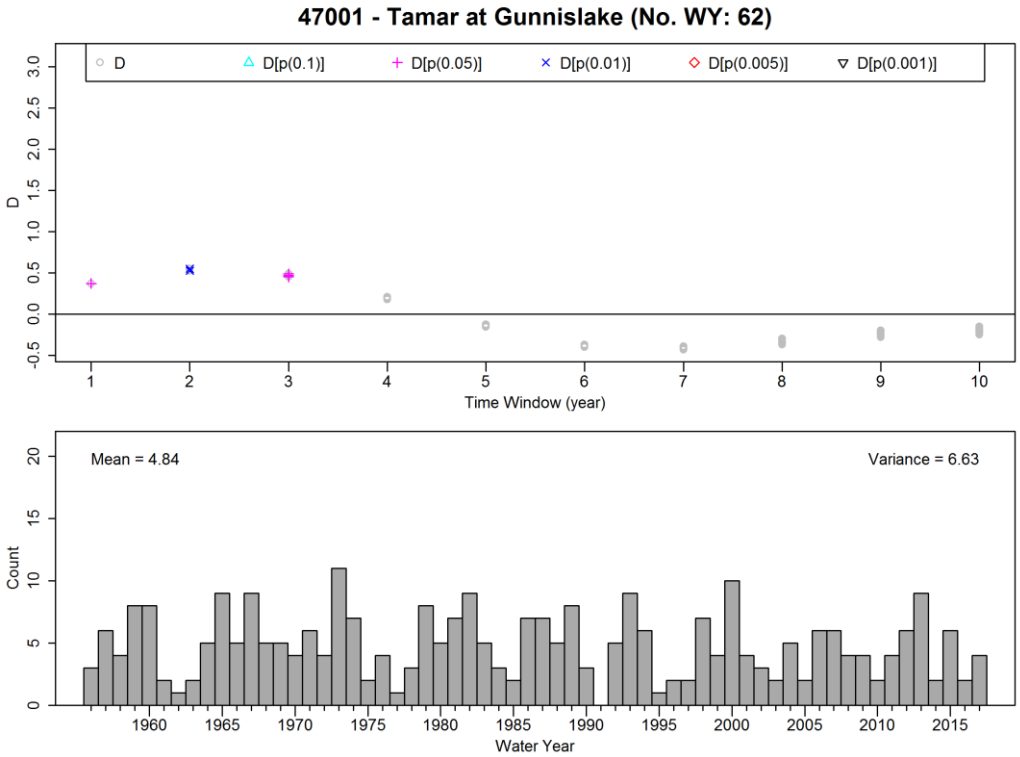


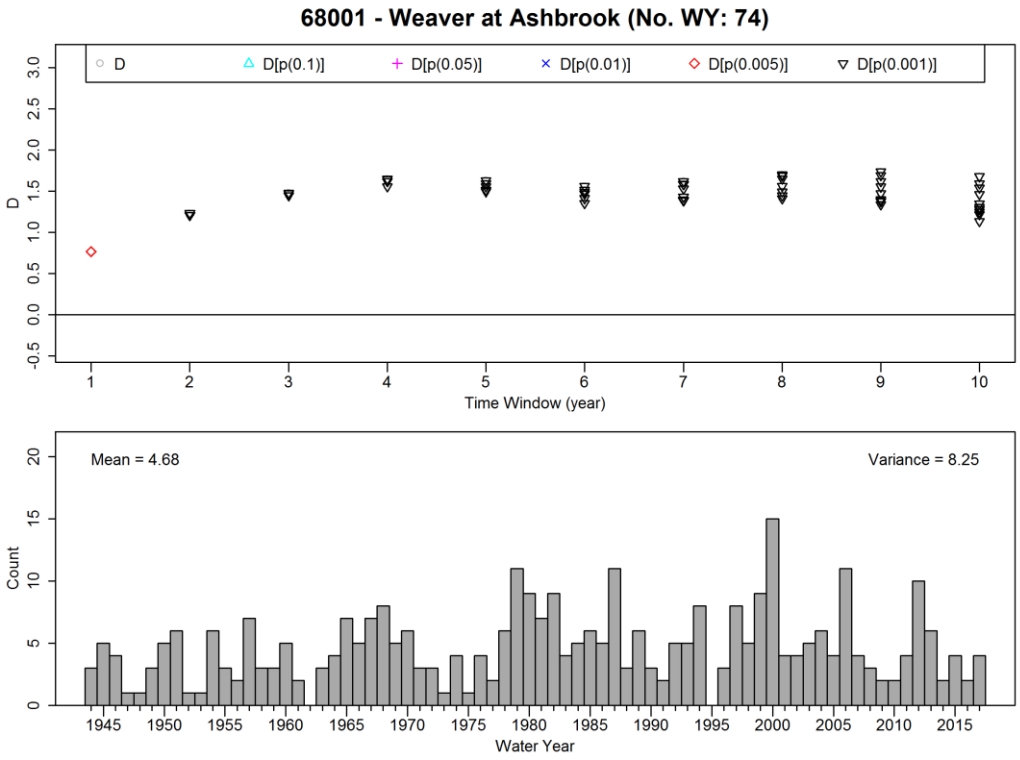


39006 - Windrush at Newbridge (No. WY: 68)

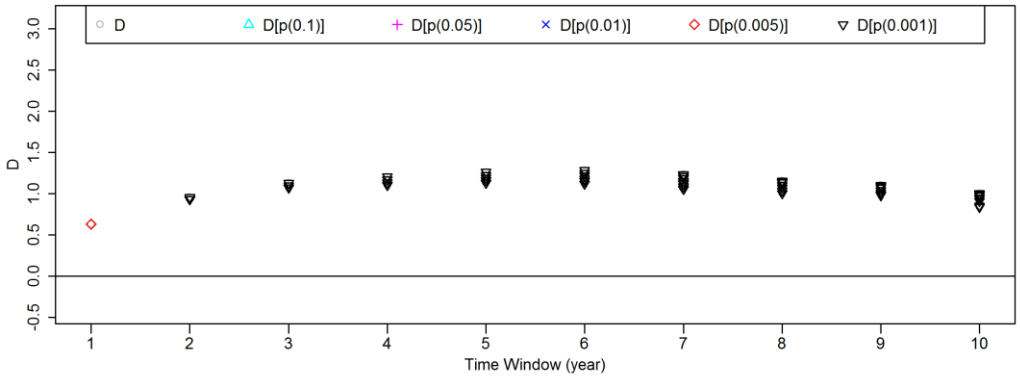


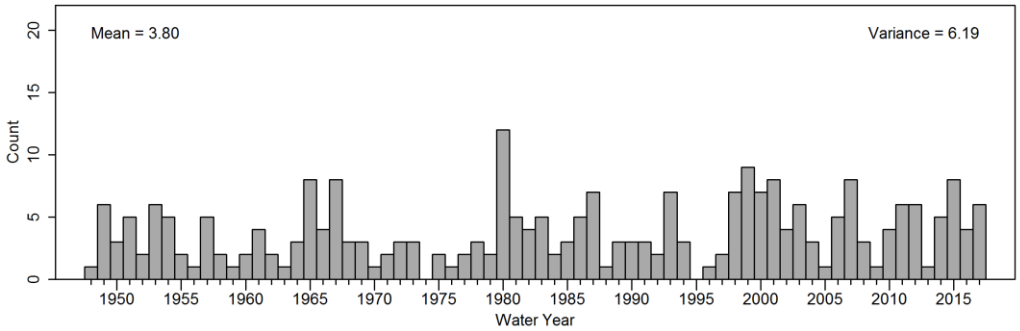



47001 - Tamar at Gunnislake (No. WY: 62)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 4.84
Variance = 6.63
Count
Water Year

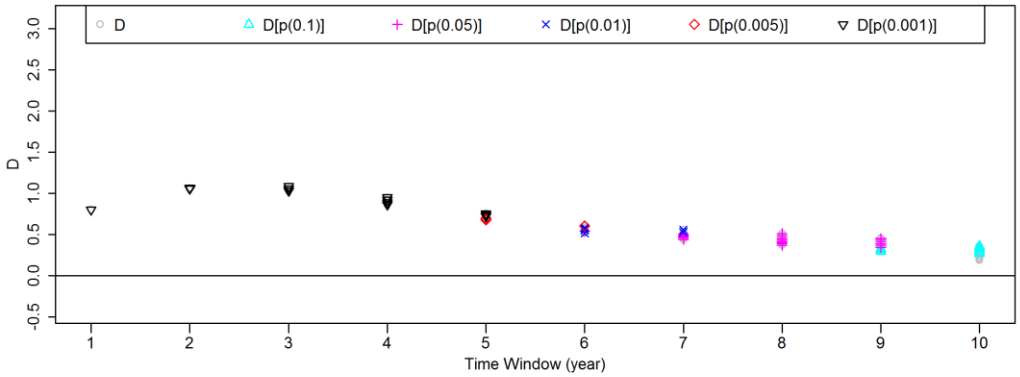

68001 - Weaver at Ashbrook (No. WY: 74)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 4.68
Variance = 8.25
Count
Water Year

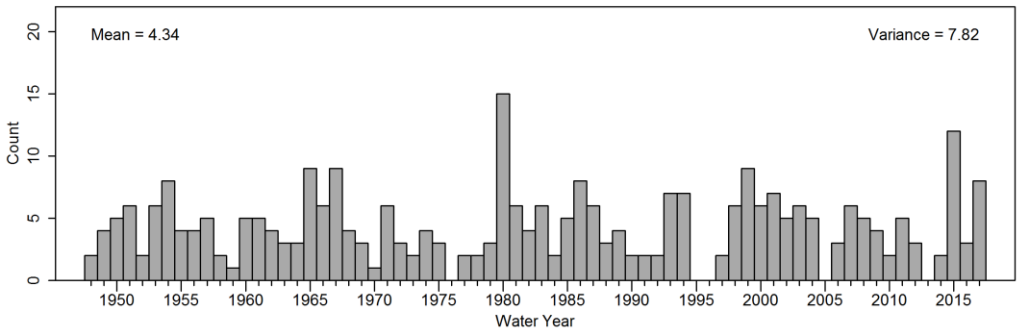
69023 - Roch at Blackford Bridge (No. WY: 70)

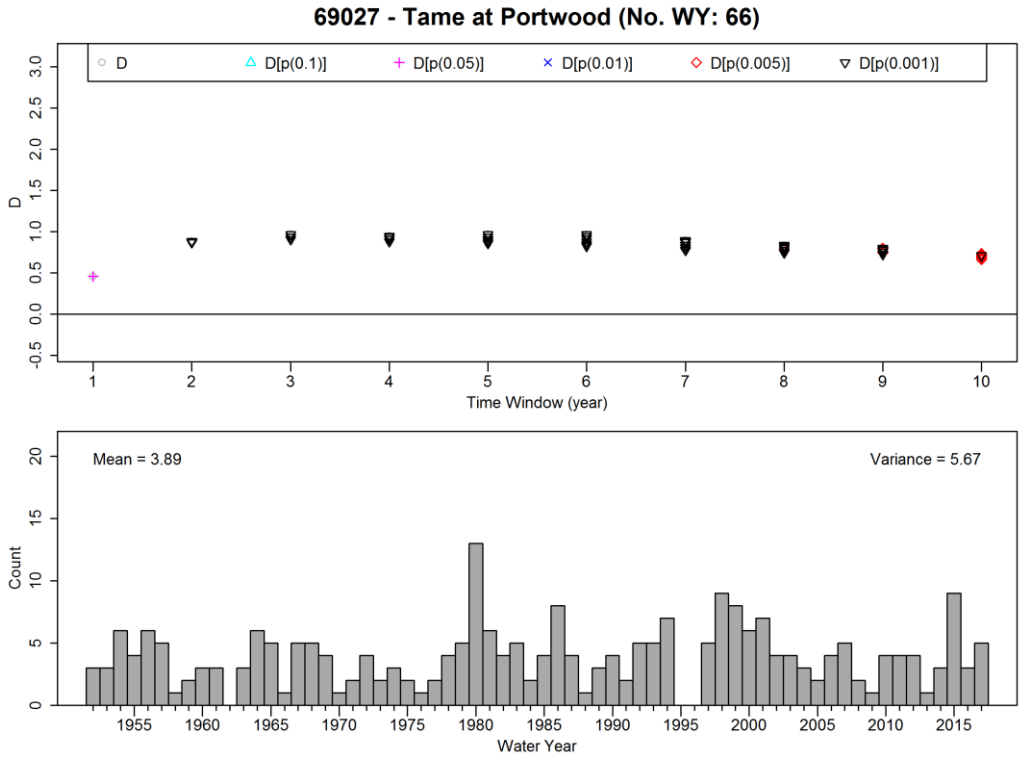




69024 - Croal at Farnworth Weir (No. WY: 70)






69027 - Tame at Portwood (No. WY: 66)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 3.89
Variance = 5.67
Count
Water Year

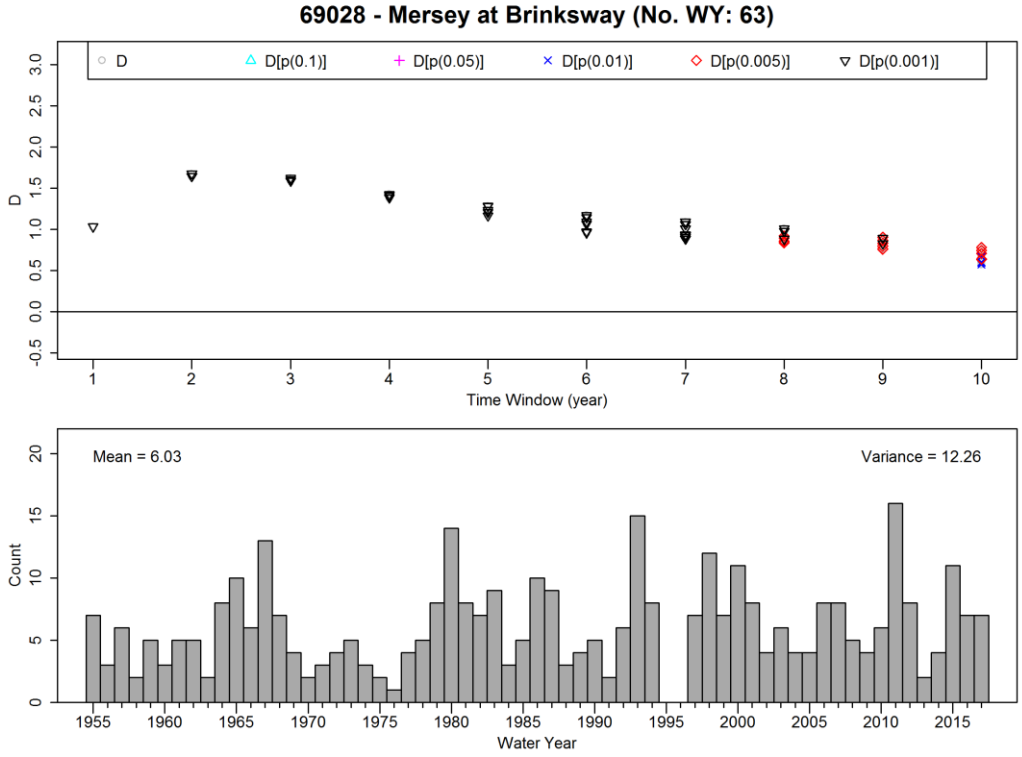

69028 - Mersey at Brinksway (No. WY: 63)
D
D[p(0.1)]
D[p(0.05)]
D[p(0.01)]
D[p(0.005)]
D[p(0.001)]
Time Window (year)
Mean = 6.03
Variance = 12.26
Count
Water Year

Table F-1 Summary of dispersion analysis results

Station ID	Station name	Trend	Variance to mean ratio	Record length (years)	Approx. peak *D*		Comments
					Value or range	Time windows	
27002	Wharfe at Wetherby Flint Mill	No significant trend identified.	1.62	80	0.5 to 1.0	4 to 6 years	Significance at 99.9% level for all windows > 3 years.
28008	Dove at Rocester Weir	No significant trend identified.	1.54	65	0.5 to 1.0	2 to 4 years	Significance at 99.9% level for 2 to 5 year windows, relatively low significance for windows > 7 years.
32004	Ise Brook at Harrowden	No significant trend identified.	2.49	62	1.0 to 1.5	one to 5 years	Significance at 99.9% level for one to 6 year windows, relatively low significance for windows > 8 years.
32007	Nene/Brampton at St Andrews Total	No significant trend identified.	3.22	78	1.5 to 2.5	2 to 4 years	Secondary peak for the largest windows; significance at 99.9% level for all windows.
37001	Roding at Redbridge	No significant trend identified.	2.63	68	1.0 to 1.5	one to 4 years	Significance at 99.9% level for one to 6 year windows, relatively low significance for windows > 8 years.
39001	Thames at Kingston	No significant trend identified.	1.46	73	0.5	2 to 6 years	Index of dispersion relatively low for all time windows; significance at 99.9% level only for 3 and 5 year windows, relatively low significance for windows > 7 years.
39002	Thames at Days Weir	No significant trend identified.	1.54	78	0.5	3 to 4 years	Index of dispersion relatively low for all time windows; significance at 99.9% level only for 4-year window, relatively low significance for windows > 7 years.
39006	Windrush at Newbridge	No significant trend identified.	1.46	68	0.5	4 to 10 years	Index of dispersion relatively low for all time windows, peaks at longest time windows but may not be appropriate; no significance at 99.9% level, significance mostly at 99% level, relatively low significance for windows < 3 years.
47001	Tamar at Gunnislake	No overall significant trend identified, but Mann-Kendall identified some negative trend for full period.	1.34	62	0.5	2 to 3 years	Index of dispersion is negative (no clustering) for windows > 3 years and low for other shorter windows; no significance at 99.9% level, significance mostly at 95% level.
68001	Weaver at Ashbrook	Mann-Kendall identified significant negative trend for long and full periods.	1.76	74	1.5 to 2.0	2 to 10 years	Index of dispersion is large for all windows > one year; significance at 99.9% level for all windows except one year.
69023	Roch at Blackford Bridge	Significant positive trends identified using Mann-Kendall for long and full periods and Pettitt change-point test.	1.63	70	1.0 to 1.5	2 to 10 years	Index of dispersion is large for all windows > one year; significance at 99.9% level for all windows except one year.
69024	Croal at Farnworth Weir	No significant trend identified.	1.80	70	0.5 to 1.0	2 to 4 years	Significance at 99.9% level for windows < 6 years.
69027	Tame at Portwood	No significant trend	1.46	66	0.5 to 1.0	2 to 10	Index of dispersion is relatively large for all windows > one year; significance at 99.9% level for all windows

Station ID	Station name	Trend	Variance to mean ratio	Record length (years)	Approx. peak *D*		Comments
					Value or range	Time windows	
		identified.				years	except one-year.
69028	Mersey at Brinksway	No significant trend identified.	2.03	63	1.5 to 2.0	one to 4 years	Index of dispersion relatively large for all time windows; significance at 99.9% level for one to 9 year windows.

Summary

Significant trend in AMAX flows identified:

- 68001 - Weaver at Ashbrook
- 69023 - Roch at Blackford Bridge

Dispersion index > 1:

- 32004 - Ise Brook at Harrowden
- 32007 - Nene/Brampton at St Andrews Total
- 37001 - Roding at Redbridge
- 68001 - Weaver at Ashbrook
- 69023 - Roch at Blackford Bridge
- 69028 - Mersey at Brinksway

Dispersion index small/negative:

- 39001 - Thames at Kingston
- 39002 - Thames at Days Weir
- 39006 - Windrush at Newbridge
- 47001 - Tamar at Gunnislake

Conclusions

Stations with large variance to mean ratio (>2) based on their annual POT counts also exhibit larger values for the peak dispersion index (>1), as anticipated. These stations are geographically dispersed, have a range of catchment sizes (186 to 622 km^2), with a range of catchment characteristics in terms of steepness and permeability.

The analysis above using varying time windows provides a step beyond this low order analysis to identify the timescales on which floods events cluster, which may not be readily identifiable by looking at the series of annual POT counts alone. However, it is not readily apparent how the dispersion index results can be interpreted in terms of the character of clustering in a given time window. For example, if clustering on the timescale of 4 years is indicated by a high D value, does this mean there is a cyclic pattern of flooding with a period of 4 years, or just that one period of 4 years in the data set exhibited larger POT counts?

This question could partially be addressed using the kernel occurrence rate method, which would highlight periods with higher or lower than expected flood counts.

The presence of temporal clustering of floods in a given POT series may also suggest that the flood peaks are not independent. Methods proposed by Eastoe and Tawn (2010) or the use of the Cox regression method could help identify the source of the dependence. Attribution of any dependence may then help to fit more appropriate models describing the extremes, and therefore design flows with smaller associated uncertainties.

F.5 Future work

There are several possible avenues for follow-up work. A more in-depth investigation could seek to answer some of the following questions:

- How does the degree of clustering vary with flood magnitude, with catchment type or location?
- Is clustering primarily due to hydrological or meteorological causes?
    - What information do we have on the typical duration of flood-rich or flood periods? (If we think that a flood-rich period started in the late 1990s, can we estimate when it might end?)
- How much more likely is a T1-year flood to occur within a period of N months after a T2-year flood?
    - How does this relative risk measure vary with catchment type or location?
    - Are there some types of floods or meteorological drivers for which it is possible to predict a greater likelihood of more flooding in the short-term? For example, if another event like June 2012 occurred, could flood risk managers have any knowledge to help them prepare for successive floods?
- How can the probability of a sequence of floods as opposed to a single event be quantified?

Here, the project team offer some specific suggestions as a follow-up to the work carried out within this investigation.

Data quality:

- Identify POT extraction methodology used for each data source across the period of record. Add specific information regarding POT methodology for time periods and data sources to NRFA station pages.
- Investigate obvious step changes in POT frequency, possibly associated with changes in data source, but requires confirmation.
- Investigate reasons for rejection of AMAX and POT records and cross-reference between the two data sets.
- Attempt to re-derive the POT records using a consistent method across period of record at each station. This would take considerable effort at stations where the full period of record is not availably digitally.
- Attempt to maximise the length of records to be used in temporal clustering analysis, considering the degree to which missing data may decrease the robustness of any results.

Methodology:

- Investigate impact missing data may have on effectiveness of temporal clustering analysis.
- Add confidence intervals to dispersion analysis plots for each time window.
- For dispersion analysis, determine maximum time window appropriate given length of record; this will be station dependent.
- Perform kernel occurrence rate analysis using both non-parametric and parametric confidence interval significance testing to identify flood-poor and flood-rich periods at each station.

- Evaluate the results from the dispersion analysis, and subsequent kernel occurrence rate analysis in the context of catchment location, size and geology, and through neighbourhood comparisons with nearby stations.
- Following the kernel occurrence rate analysis, evaluate the timing of flood-poor and flood-rich periods in terms of catchment rainfall, antecedent conditions, and correspondence with climate indices describing modes of atmospheric variability relevant to the UK (for example, NAO, EA).
- Perform Cox regression applied to both streamflow and precipitation with climate indices as covariates to further investigate the attribution of clustering of floods into flood-poor and flood-rich periods.
- Extend the Cox regression to build flood clustering prediction models to identify future periods of temporal clustering of floods.
- Evaluate the impact of non-stationarity (long-term trends) on temporal clustering results.
- Investigate the applicability of the assumption of the Poisson process for POT evaluation, and, by extension, for flood frequency analysis of annual maximum flows. The presence of temporal clustering of floods implies the events are not necessarily independent. This is particularly relevant for very slowly draining catchments and those that are groundwater-influenced.

Additional note: Details of data review

Notes on AMAX files

The AMAX files in the NRFA data set include the list of AMAX dates and flow values (and sometimes the level) for each year where an AMAX has been determined. The files also include a list of years where the AMAX has been rejected (for whatever reason, possibly due to too much missing data in the year, a poor rating or other gauge issues).

Note the dates listed in the AMAX files are for the actual date the AMAX value was recorded; no time values for the observation are present. This presents an issue due the fact that AMAX values are derived for water days (9am to 9am); if the AMAX value occurs before 9am on 1 October, the date in the AMAX file will put the AMAX value in the incorrect water year.

An example of this occurring is for station 69024 (Croal at Farnworth Weir) for the AMAX in the 1966 to 1967 water year.

Notes on POT files

The POT files include the list of POT dates and flow values (and sometimes the level) for each POT in the record. The files also include a list of POT values that have been rejected (for whatever reason), a list of periods where there is missing data, the start and end dates of the record, and the flow threshold used to define the POT series.

Inconsistencies have been identified between the noted data gaps and the presence (or lack) of POT values.

Examples of missing data being declared but POT values present are:

- Station 37001 (Roding at Redbridge) for the 1969 to 1970 and 1970 to 1971 water years. This issue has been corrected in v8 of the data set.
- Station 55002 (Wye at Belmont) for the 1956 to 1957 and 1957 to 1958 water years where all data are listed as missing but one POT exists for each year, which is also the AMAX for that year, neither of which are rejected.

One example of potentially missing POT values but (virtually) no missing data being declared is for station 73010 (Leven at Newby Bridge) for the water year period 1969 to 1970 to 1974 to 1975. During this period only one year has around 10% missing data, but only one POT value is reported in each year. Each of these POT values is also the AMAX value for the year. The source of the POT records during this period is listed as ‘Mfiche’ on the NRFA.

Potential inconsistencies have also been identified between the AMAX values listed as rejected and POT values, which are the AMAX value for the year but are not rejected in the POT file. Sometimes this may occur due to the proportion or location of missing data in a water year, making the AMAX unrepresentative for the year, but the POT is still valid.

Examples where the record is listed as 100% complete, the AMAX is rejected, and the POT is not rejected are:

- 1953 to 1954 water year for the station 38022 (Pymmes Brook at Edmonton Silver Street)
- 1955 to 1956 water year for the station 47001 (Tamar at Gunnislake)
- 1955 to 1956 water year for the station 54008 (Teme at Tenbury)
- 1936 to 1937 water year for the station 68001 (Weaver at Ashbrook)

- 1942 to 1943 and 1951 to 1952 water years for the station 69027 (Tame at Portwood)

Potential issues relating to multiple POT values recorded on the same day (with different flow values) or POT values recorded on sequential days have also been identified. In some cases, such as for very small, flashy catchments, these may represent true independent POT events. However, duplicated POT dates and sequential POT dates have been identified in very large catchments, where they are much less likely to represent independent events.

Examples of small catchments with duplicated POT dates:

- station 38007 (Canons Brook at Elizabeth Way; 21.4 km^2) for POT values on 21 November 1974 and 15 October 1987
- station 69024 (Croal at Farnworth Weir; 145 km^2) for POT values on 10 December 1965 and 11 March 1981
- station 39093 (Brent at Monks Park; 117.6 km^2) for POT values on 9 separate occasions
- station 38022 (Pymmes Brook at Edmonton Silver Street; 42.6 km^2) on 10 separate occasions

Examples of large catchments with duplicated POT dates:

- station 55007 (Wye at Erwood; 1,282.1 km^2) for POT values on 18 December 1965

Initial station selection

An initial subset of all NRFA peak flow stations was derived from the following lists of stations:

- a selection of 20 POT series in the UK that was used by Emma Raven in an (unpublished) study of flood clustering and impacts on reinsurance at Durham University. All these series start before 1940, and all were closely examined for errors and anomalies
- additional long record stations suggested by Peter Spencer of the Environment Agency
- those previously identified as having long records (>60 years) in this project

These lists were combined, and duplicates removed. This was used as the station list for filtering the NRFA Peak Flow Dataset.

Initial record length assessment

Using an R script with the NRFA Peak Flow Dataset V8 and the initial station list as inputs, the AMAX and POT files for each of the initial selection of stations were read in and evaluated.

Using the AMAX files, the following are determined:

- total number of AMAX values in the record
- number of AMAX values that are rejected
- number of AMAX values that are not rejected

Using the POT files, the following are determined:

- total number of POT values in the record

- number of water years represented in the POT record (for water years with at least one POT)
- average number of POT per water year (for water years with at least one POT)
- number of POT that are rejected
- total number of POT values in the record that are not rejected
- number of water years represented in the POT record (for water years with at least one POT) after rejected POT values are removed
- average number of POT per water year (for water years with at least one POT) after rejected POT values are removed
- percent complete for each water year over the length of the record

By linking the information retrieved from both the AMAX and POT files the following are determined:

- total number of POT values in the record after rejected POT values and POT values in years where the AMAX value has been rejected are removed
- number of water years represented in the POT record (for water years with at least one POT) after rejected POT values and POT values in years where the AMAX value has been rejected are removed
- average number of POT per water year (for water years with at least one POT) after rejected POT values and POT values in years where the AMAX value has been rejected are removed
- number of AMAX values that are less than the threshold used to define the POT series
- number of AMAX values that are present in the POT series
- number of the AMAX values (after removing rejected AMAX values) with years with at least 90, 75, 50 and 10% complete daily data

Not every station has a POT series, therefore, for some stations only the AMAX files are evaluated.

Using the information determined above, the initial station selection was filtered further to retain only those stations where POT records exist, the number of water years represented in the POT record (for water years with at least one POT) after rejected POT values are removed is greater than 60, and the number of the AMAX values (after removing rejected AMAX values) with years with at least 90% complete daily data is greater than 60. After these filters are applied, 28 stations remain.

POT data quality assessment

Using an R script, the AMAX and POT files for secondary selection of stations were evaluated and the POT and AMAX series plotted. The AMAX and POT data sets were also cross-referenced with one another to identify any matching records rejected in the other data set. The presence of any duplicated dates in the POT series was also noted.

Data quality plots were produced for each station, and visually assessed. An example is shown below. Each plot contained the following:

Top panel:

- the number of POT values (y-axis) in each water year (after rejected POT values and duplicated POT dates are removed; x-axis) with a 10-

year moving average applied. The bars are centred on the mid-point of the water year, with the starting water year identified

Bottom panel:

- grey bars show the percent complete of daily data, as determined based on the data gaps in the POT files. The height of the bar indicates the percent complete, with the scale on the right-hand y-axis
- blue diamond points show the AMAX values with red diamonds indicating AMAX values identified as rejected in the AMAX files
- blue crosses show the POT values with red crosses indicating POT values identified as rejected in the POT files
- for both the AMAX and POT data, points are located on the x-axis based on the date they occurred. The starting water year is indicated, identified at the mid-point of the water year

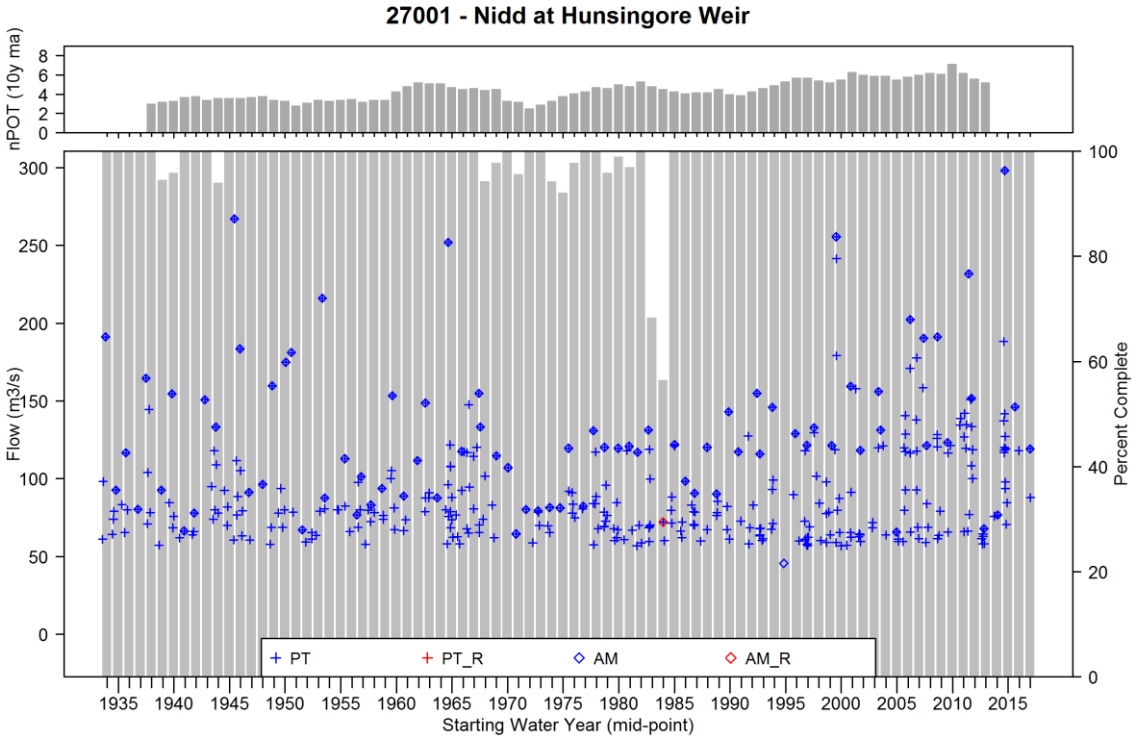


Figure F-2 Example POT data quality assessment plot

For the purposes of the temporal clustering analysis, only the most reliable and consistent records were retained. The impact missing data may have on the POT series, and any subsequent analysis, is not fully known at present. Therefore, the current study has imposed strict criteria on the quality of data used in the temporal cluster analysis.

Based on the information interpreted from the data quality plots, and comments on the NRFA website, 12 further stations were filtered out as unsuitable for temporal cluster analysis.

For the retained 16 stations, the data quality plots and comments on the NRFA website also resulted in truncation of the time series for some stations to exclude periods with suspect or missing data.

Appendix G: Maps of non-stationary model results

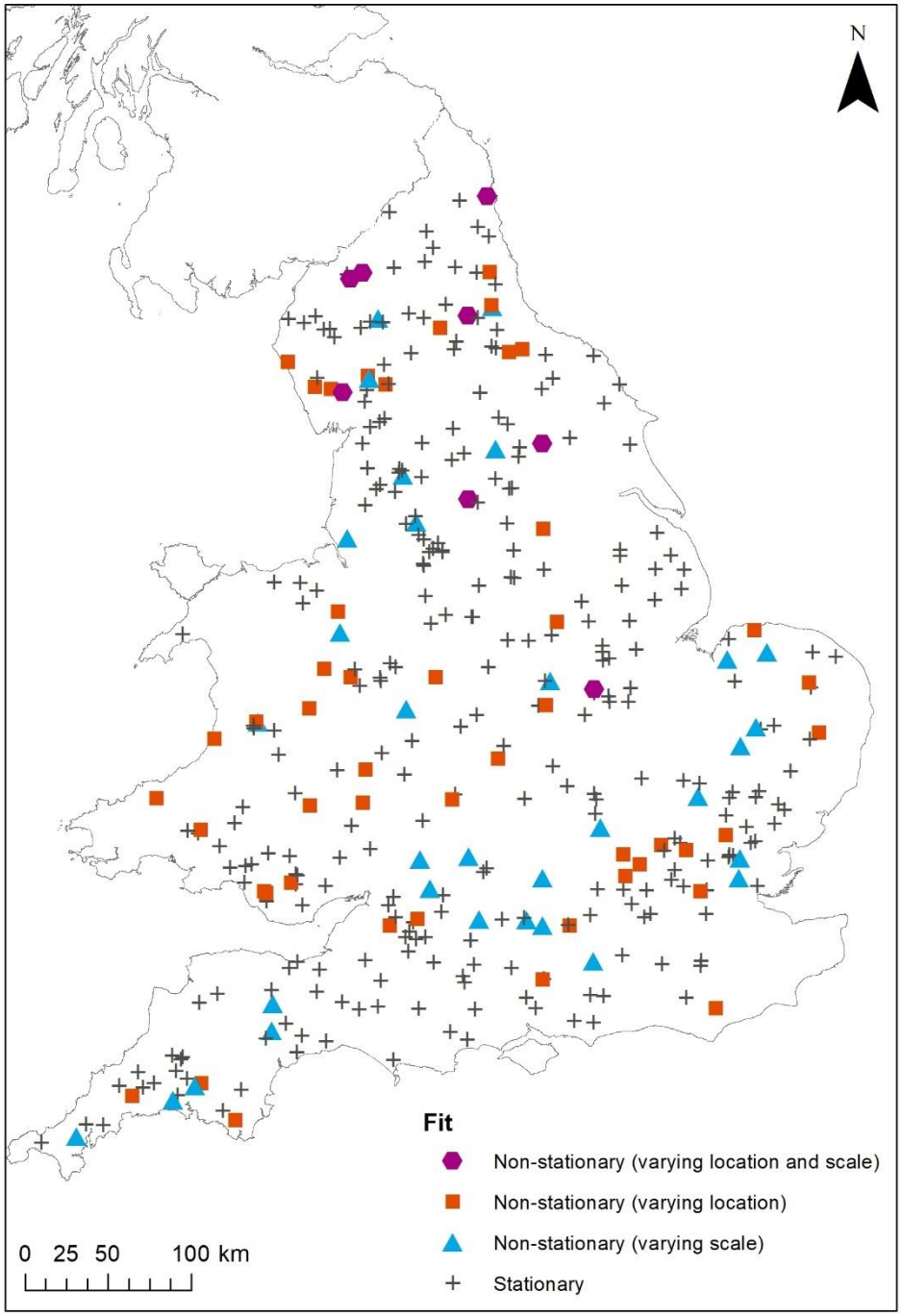


Figure G-1 Spatial distribution of preferred model fit

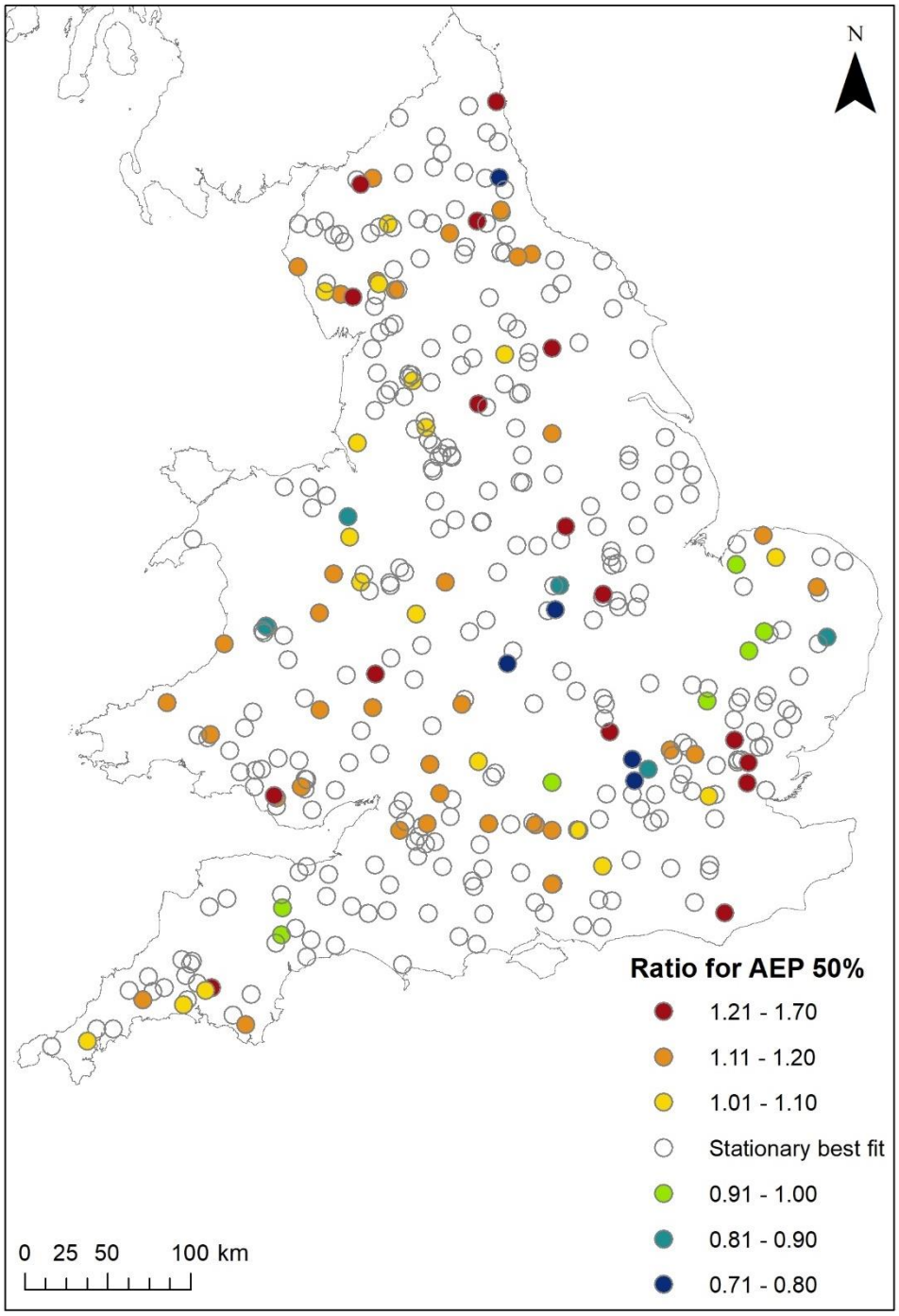


Figure G-2 Ratios of estimate from preferred model (GEV) to estimate from stationary GEV model, AEP 50%

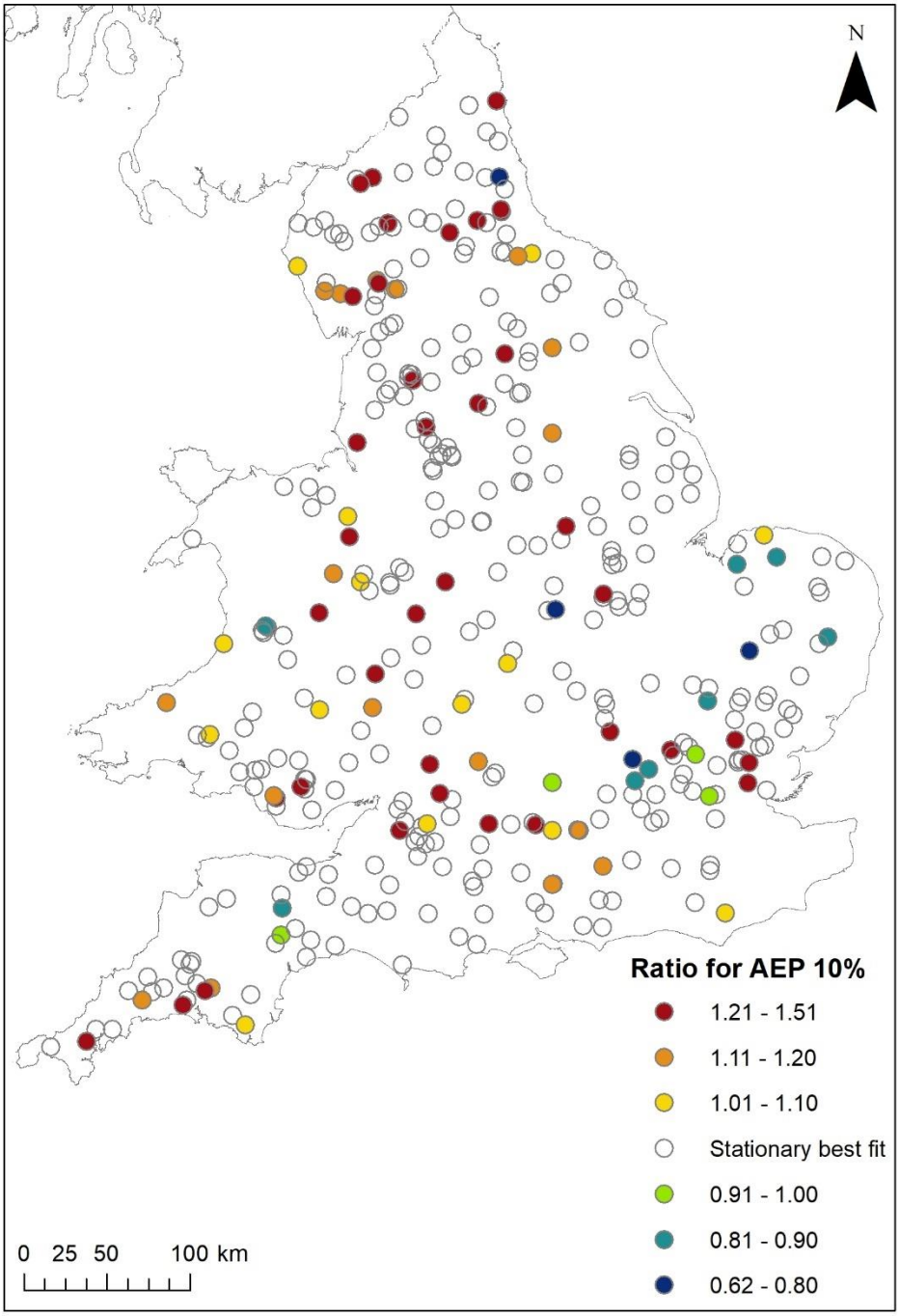


Figure G-3 Ratios of estimate from preferred model (GEV) to estimate from stationary GEV model, AEP 10%

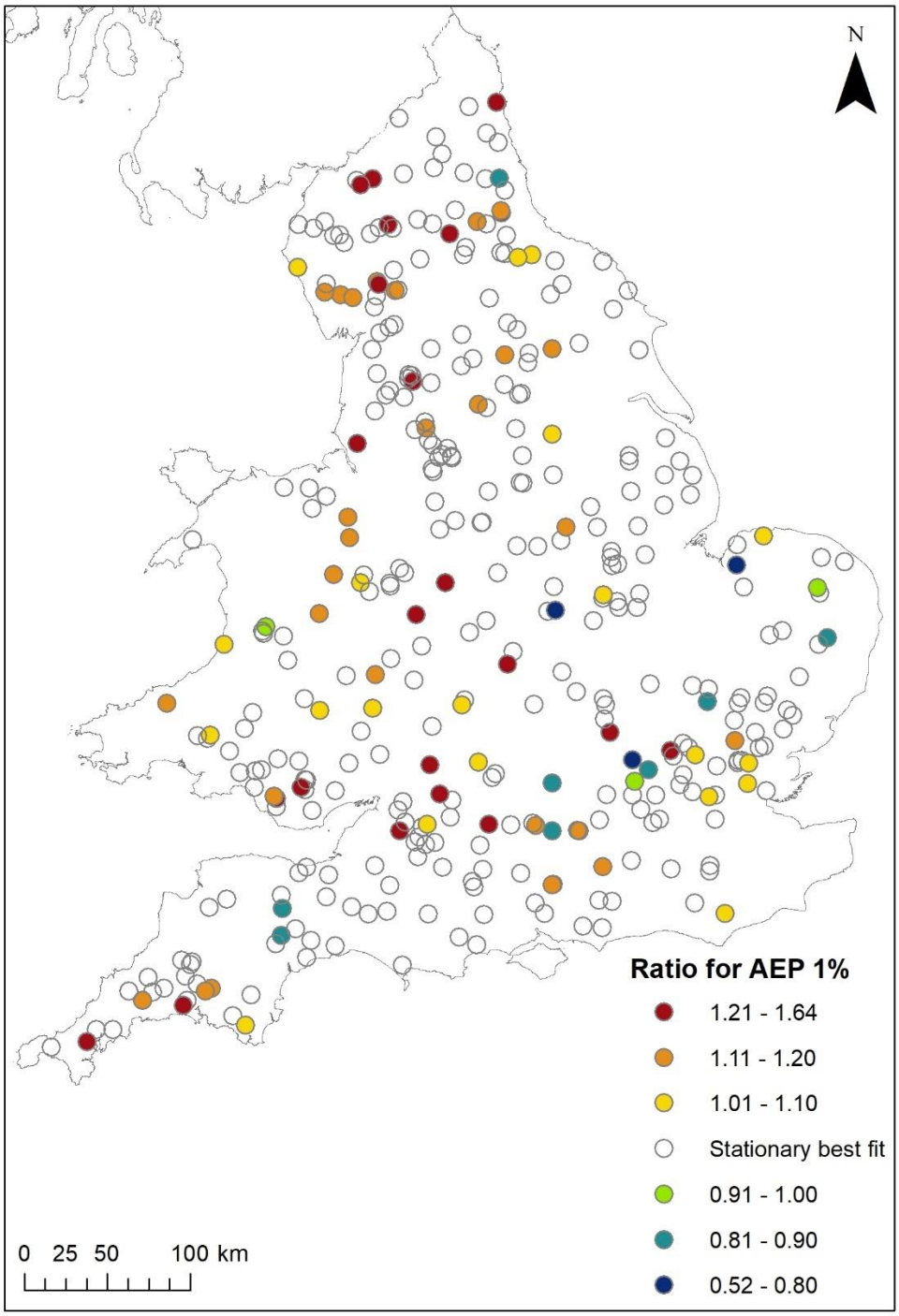


Figure G-4 Ratios of estimate from preferred model (GEV) to estimate from stationary GEV model, AEP 1%

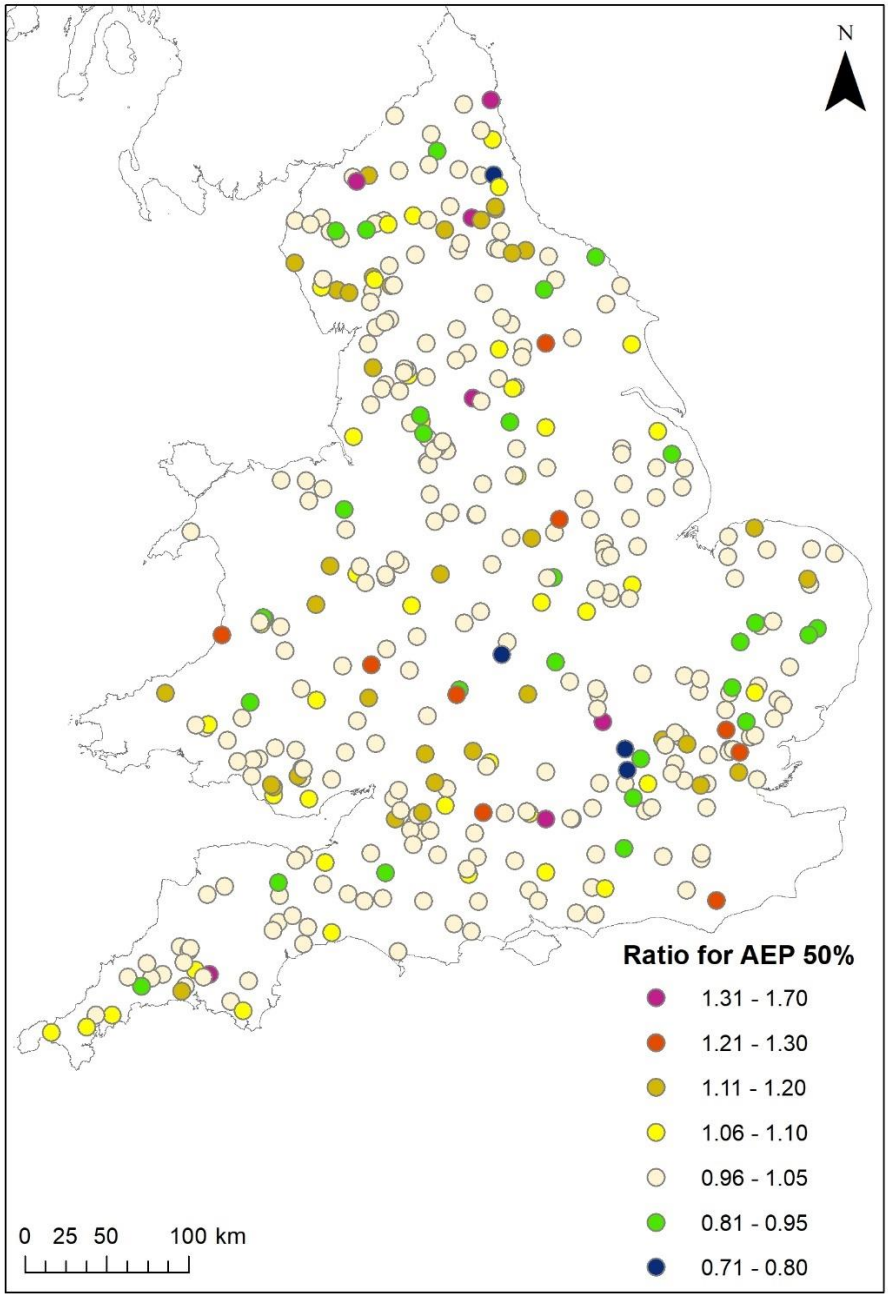


Figure G-5 Ratios of estimate from preferred model (GEV), considering water year as covariate, to estimate from FEH: AEP 50%

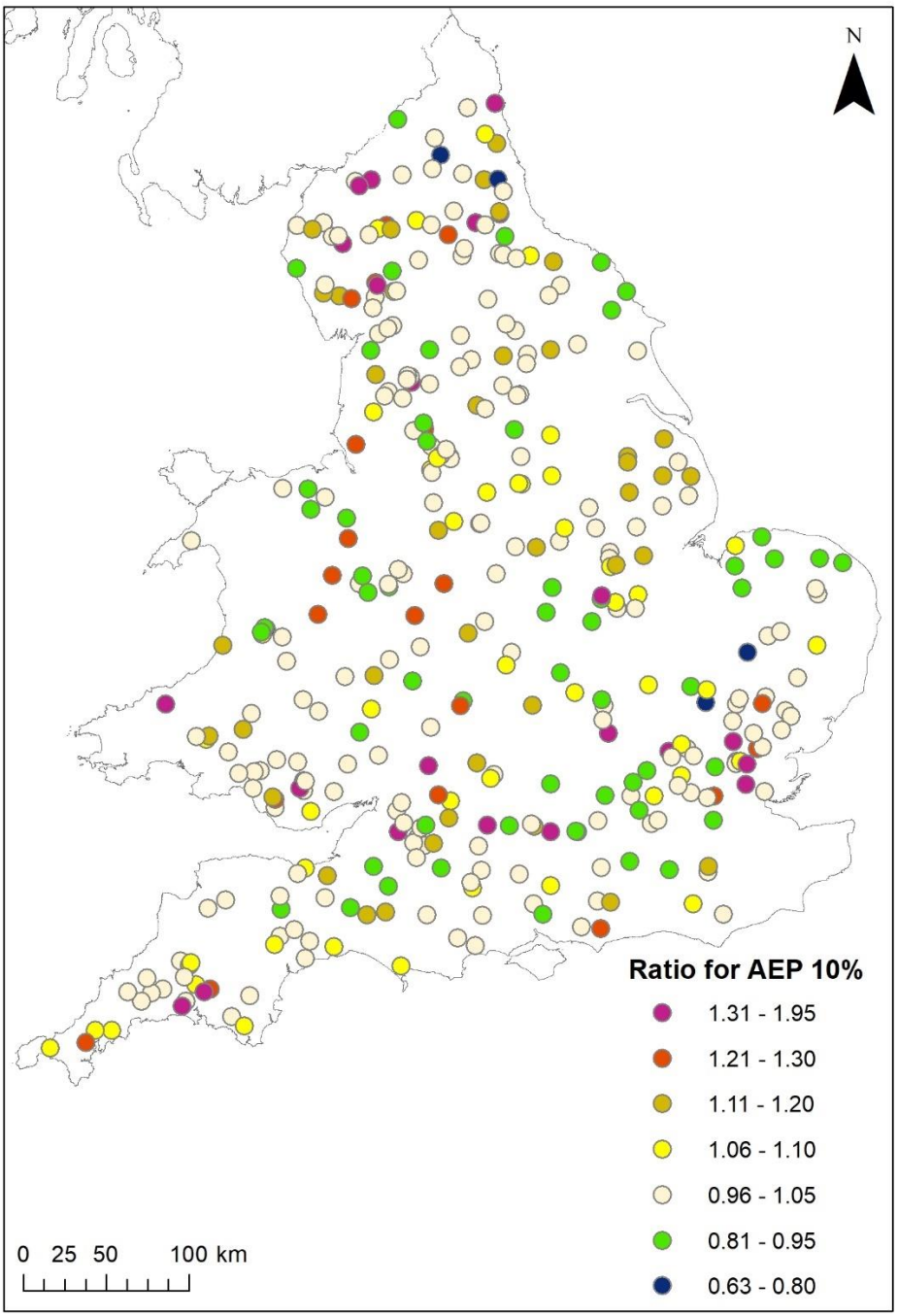


Figure G-6 Ratios of estimate from preferred model (GEV), considering water year as covariate, to estimate from FEH: AEP 10%

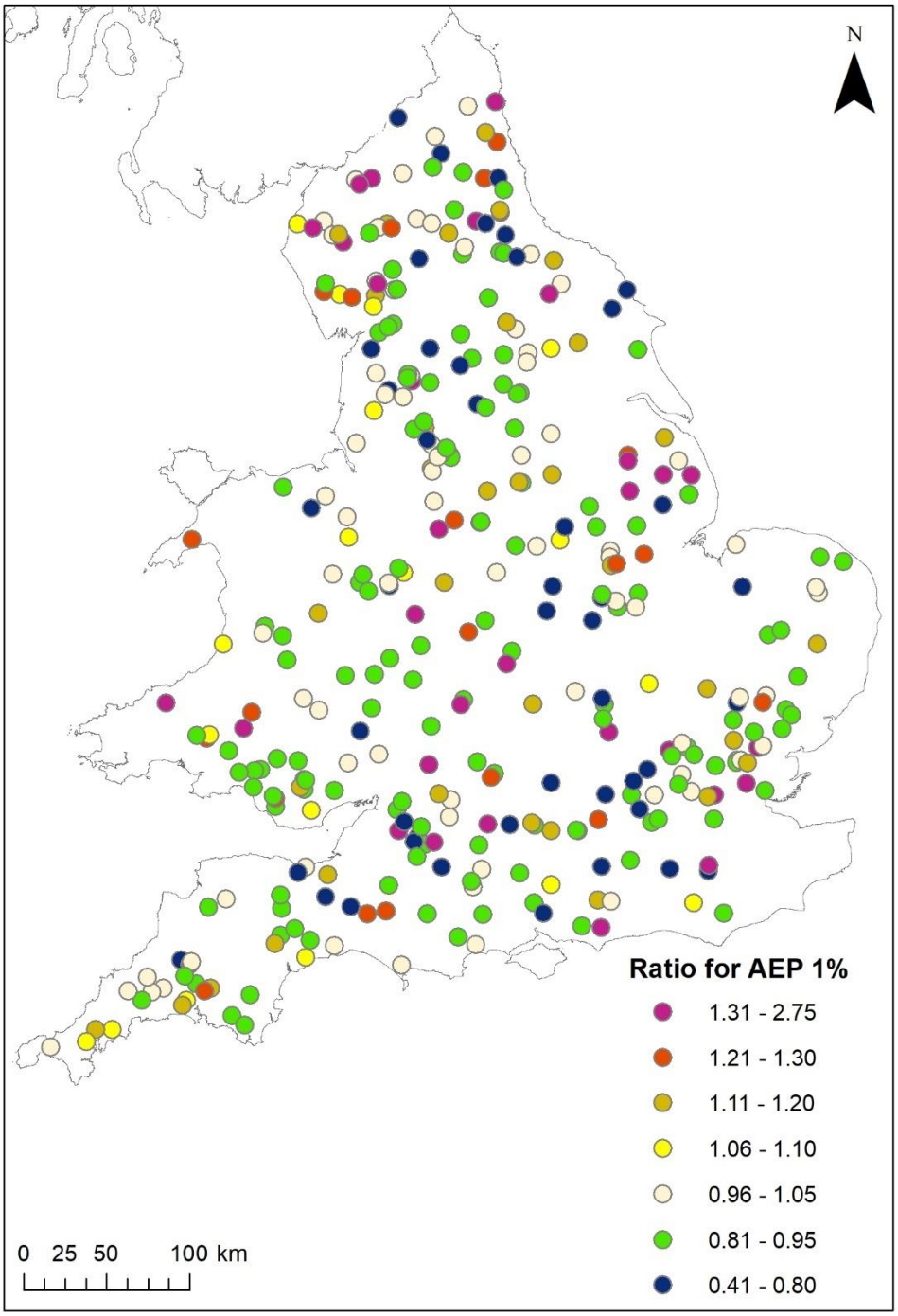


Figure G-7 Ratios of estimate from preferred model (GEV), considering water year as covariate, to estimate from FEH: AEP 1%

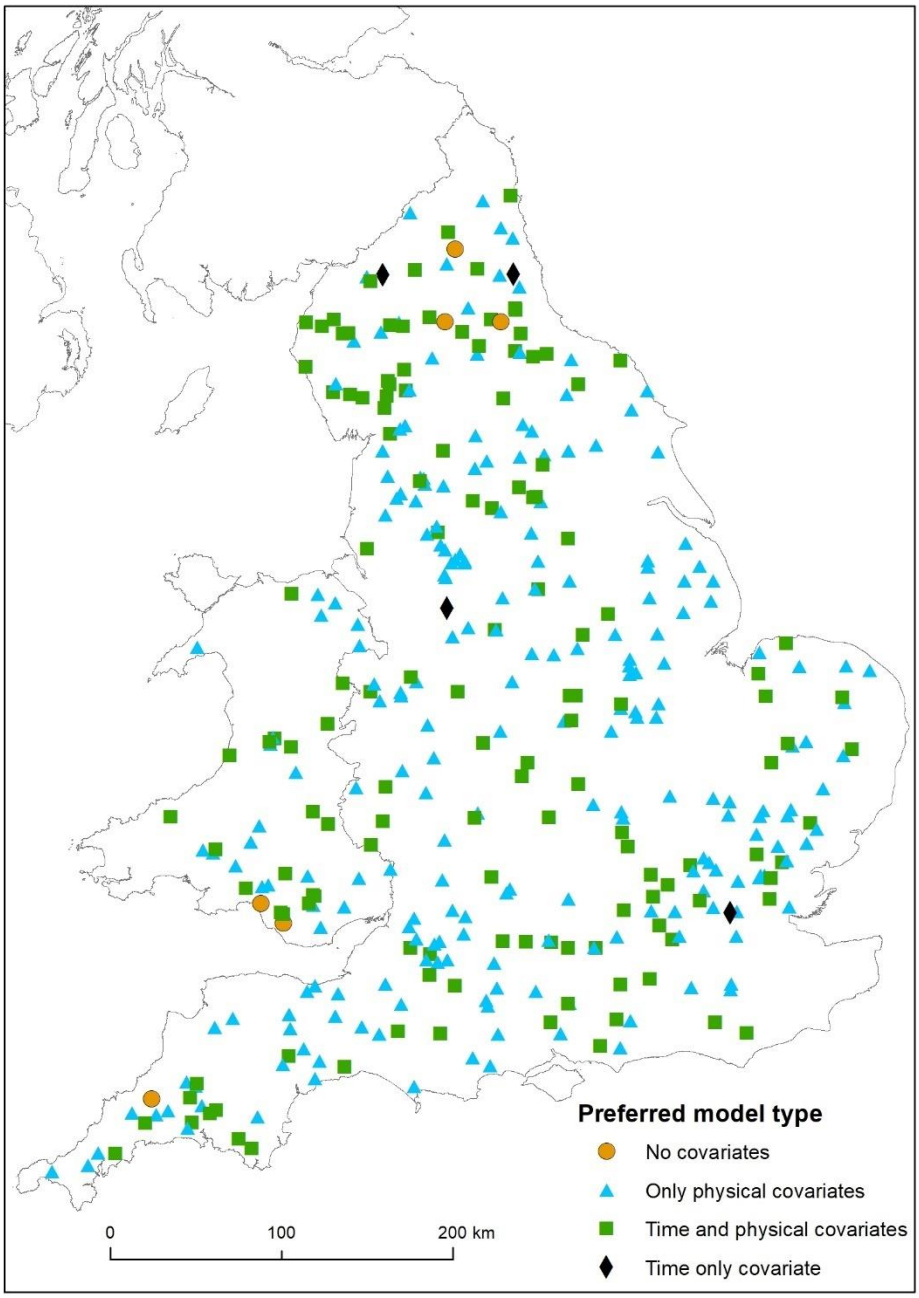


Figure G-8 Type of covariates chosen (by lowest BIC) at each gauge

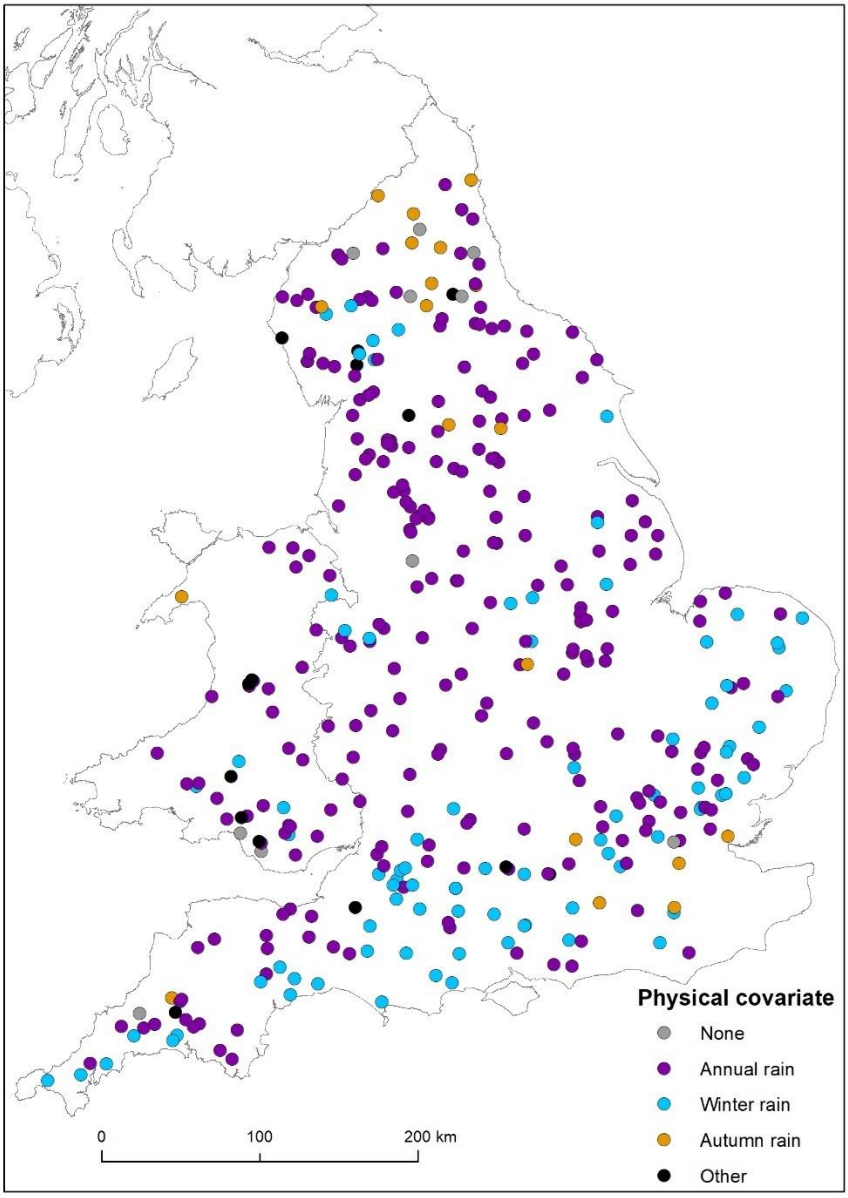


Figure G-9 Best-fitting physical covariate chosen at each gauge

'None' is used when the best-fitting model is stationary.

Appendix H: Project data set

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
21032	Glen	Kirknewton	45	1961 to 2009	Yes	Yes	No - data quality limited
22001	Coquet	Morwick	54	1963 to 2016	Yes	Yes	Yes
22006	Blyth	Hartford Bridge	57	1960 to 2016	Yes	Yes	Yes
22007	Wansbeck	Mitford	52	1962 to 2013	Yes	Yes	Yes
22009	Coquet	Rothbury	42	1973 to 2016	Yes	Yes	Yes
23001	Tyne	Bywell	61	1956 to 2016	Yes	Yes	Yes
23003	North Tyne	Reaverhill	37	1980 to 2016	Yes	Not enough data	Yes
23004	South Tyne	Haydon Bridge	58	1959 to 2016	Yes	Yes	Yes
23006	South Tyne	Featherstone	51	1966 to 2016	Yes	Yes	Yes
23007	Derwent	Rowlands Gill	53	1964 to 2016	Yes	Yes	Yes
23008	Rede	Rede Bridge	48	1968 to 2016	Yes	Yes	Yes
23011	Kielder Burn	Kielder	46	1970 to 2016	Yes	Yes	Yes
23017	Team	Team Valley	40	1974 to 2016	Yes	Not enough data	Yes
24001	Wear	Sunderland Bridge	60	1957 to 2016	Yes	Yes	Yes
24003	Wear	Stanhope	58	1958 to 2015	Yes	Yes	Yes
24004	Bedburn Beck	Bedburn	59	1959 to 2018	Yes	Yes	Yes
24005	Browney	Burnhall	60	1954 to 2014	Yes	Yes	Yes
24008	Wear	Witton Park	43	1974 to 2016	Yes	Yes	Yes
24009	Wear	Chester le Street	40	1977 to 2016	Yes	Not enough data	Yes
25001	Tees	Broken Scar	61	1956 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
25003	Trout Beck	Moor House	27	1990 to 2016	Yes	Not enough data	Yes
25004	Skerne	South Park	59	1957 to 2016	Yes	Yes	Yes
25005	Leven	Leven Bridge	48	1959 to 2007	Yes	Yes	Yes
25006	Greta	Rutherford Bridge	57	1960 to 2016	Yes	Yes	Yes
25008	Tees	Barnard Castle	50	1964 to 2016	Yes	Yes	Yes
25009	Tees	Low Moor	47	1969 to 2016	Yes	Yes	Yes
25012	Harwood Beck	Harwood	48	1969 to 2016	Yes	Yes	Yes
25018	Tees	Middleton in Teesdale	46	1971 to 2016	Yes	Yes	Yes
25019	Leven	Easby	39	1971 to 2016	No - gaps	Yes	Yes
25020	Skerne	Preston le Skerne	44	1973 to 2016	Yes	Yes	Yes
26003	Foston Beck	Foston Mill	57	1959 to 2016	Yes	Yes	Yes
27001	Nidd	Hunsingore Weir	82	1934 to 2016	Yes	Yes	Yes
27002	Wharfe	Flint Mill Weir	81	1936 to 2016	Yes	Yes	Yes
27006	Don	Hadfields Weir	61	1956 to 2016	Yes	Yes	Yes
27007	Ure	Westwick Lock	62	1955 to 2016	Yes	Yes	Yes
27009	Ouse	Skelton	131	1886 to 2016	Yes	Yes	Yes
27010	Hodge Beck	Bransdale Weir	41	1936 to 1976	Yes	Not enough data	Yes
27021	Don	Doncaster	59	1958 to 2016	Yes	Yes	Yes
27023	Dearne	Barnsley Weir	64	1953 to 2016	Yes	Yes	Yes
27025	Rother	Woodhouse Mill	55	1961 to 2016	Yes	Yes	No - data quality limited
27026	Rother	Whittington	57	1960 to 2016	Yes	Yes	Yes
27028	Aire	Armley	57	1960 to 2016	Yes	Yes	Yes
27029	Calder	Elland	47	1971 to 2017	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
27030	Dearne	Adwick	54	1963 to 2016	Yes	Yes	No - data quality limited
27031	Colne	Colne Bridge	53	1964 to 2016	Yes	Yes	Yes
27032	Hebden Beck	Hebden	51	1965 to 2016	Yes	Yes	Yes
27033	Sea Cut	Scarborough	52	1965 to 2016	Yes	Yes	Yes
27034	Ure	Kilgram Bridge	50	1967 to 2016	Yes	Yes	Yes
27041	Derwent	Buttercrambe	44	1973 to 2016	Yes	Yes	Yes
27043	Wharfe	Addingham	44	1973 to 2016	Yes	Yes	Yes
27051	Crimple	Burn Bridge	45	1972 to 2016	Yes	Yes	Yes
27052	Whitting	Sheepbridge	41	1976 to 2016	Yes	Yes	Yes
27053	Nidd	Birstwith	41	1976 to 2016	Yes	Yes	No - data quality limited
27055	Rye	Broadway Foot	37	1977 to 2016	Yes	Not enough data	Yes
27059	Laver	Ripon	40	1977 to 2016	Yes	Not enough data	Yes
27065	Holme	Queens Mill	38	1979 to 2016	Yes	Not enough data	No - data quality limited
27073	Brompton Beck	Snainton Ings	36	1980 to 2016	Yes	Not enough data	Yes
27079	Calder	Methley	29	1988 to 2016	Yes	Not enough data	No - data quality limited
27080	Aire	Lemonroyd	32	1985 to 2016	Yes	Not enough data	Yes
27081	Oulton Beck	Farrer Lane	31	1986 to 2016	Yes	Not enough data	Yes
27084	Eastburn Beck	Crosshills	30	1988 to 2017	Yes	Not enough data	Yes
27088	Calder	Mytholmroyd	28	1989 to 2016	Yes	Not enough data	No - data quality limited
27092	Esk	Briggswath	42	1976 to 2017	Yes	Yes	Yes
28003	Tame	Water Orton	62	1955 to 2016	Yes	Yes	Yes
28008	Dove	Rocester Weir	64	1953 to 2016	Yes	Yes	No - data quality limited
28009	Trent	Colwick	59	1958 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
28010	Derwent	Longbridge Weir	82	1935 to 2016	Yes	Yes	No - data quality limited
28011	Derwent	Matlock Bath	45	1958 to 2002	Yes	Not enough data	No - data quality limited
28018	Dove	Marston on Dove	56	1961 to 2016	Yes	Yes	No - data quality limited
28019	Trent	Drakelow Park	56	1959 to 2016	Yes	Yes	Yes
28022	Trent	North Muskham	49	1968 to 2016	Yes	Yes	Yes
28023	Wye	Ashford	53	1964 to 2016	Yes	Yes	Yes
28024	Wreake	Syston Mill	50	1966 to 2016	Yes	Yes	Yes
28026	Anker	Polesworth	50	1967 to 2016	Yes	Yes	No - data quality limited
28027	Erewash	Sandiacre	50	1965 to 2016	Yes	Yes	Yes
28031	Manifold	Ilam	47	1970 to 2016	Yes	Yes	Yes
28039	Rea	Calthorpe Park	45	1972 to 2016	Yes	Yes	Yes
28040	Trent	Stoke-On-Trent	50	1967 to 2016	Yes	Yes	Yes
28046	Dove	Izaak Walton	47	1970 to 2016	Yes	Yes	Yes
28047	Oldcotes Dyke	Blyth	47	1970 to 2016	Yes	Yes	No - data quality limited
28049	Ryton	Worksop	47	1970 to 2016	Yes	Yes	Yes
28053	Penk	Penkridge	45	1971 to 2016	Yes	Yes	Yes
28056	Rothley Brook	Rothley	45	1972 to 2016	Yes	Yes	Yes
28060	Dover Beck	Lowdham	41	1972 to 2016	Yes	Yes	Yes
28061	Churnet	Basford Bridge	40	1976 to 2016	Yes	Yes	Yes
28070	Burbage Brook	Burbage	56	1925 to 1981	Yes	Not enough data	No - data quality limited
28082	Soar	Littlethorpe	44	1970 to 2016	Yes	Yes	Yes
28085	Derwent	St Mary's Bridge	32	1985 to 2016	Yes	Not enough data	Yes
28086	Sence	South Wigston	31	1986 to 2016	Yes	Not enough data	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
29001	Waithe Beck	Brigsley	57	1960 to 2016	Yes	Yes	Yes
29002	Great Eau	Claythorpe Mill	54	1963 to 2016	Yes	Yes	Yes
29003	Lud	Louth	50	1967 to 2016	Yes	Yes	Yes
29004	Ancholme	Bishopbridge	49	1968 to 2016	Yes	Yes	Yes
29009	Ancholme	Toft Newton	43	1974 to 2016	Yes	Yes	Yes
30001	Witham	Claypole Mill	58	1959 to 2016	Yes	Yes	Yes
30003	Bain	Fulsby Lock	55	1962 to 2016	Yes	Yes	Yes
30004	Lymn	Partney Mill	55	1962 to 2016	Yes	Yes	Yes
30005	Witham	Saltersford Total	33	1984 to 2016	Yes	Not enough data	Yes
30006	Slea	Leasingham Mill	32	1984 to 2016	Yes	Not enough data	Yes
30011	Bain	Goulceby Bridge	49	1966 to 2016	Yes	Yes	Yes
30013	Heighington Beck	Heighington	41	1976 to 2016	Yes	Yes	Yes
30014	Pointon Lode	Pointon	45	1972 to 2016	Yes	Yes	Yes
30015	Cringle Brook	Stoke Rochford	38	1979 to 2016	Yes	Not enough data	Yes
30017	Witham	Colsterworth	39	1978 to 2016	Yes	Not enough data	Yes
31004	Welland	Tallington Total	46	1967 to 2012	Yes	Yes	Yes
31005	Welland	Tixover	46	1971 to 2016	Yes	Yes	Yes
31010	Chater	Fosters Bridge	50	1967 to 2016	Yes	Yes	Yes
31021	Welland	Ashley	42	1970 to 2012	Yes	Yes	Yes
31023	West Glen	Easton Wood	45	1972 to 2016	Yes	Yes	Yes
31025	Gwash South Arm	Manton	39	1978 to 2016	Yes	Not enough data	Yes
31026	Egleton Brook	Egleton	36	1978 to 2013	Yes	Not enough data	Yes
32002	Willow Brook	Fotheringhay	59	1938 to 1997	Yes	Not enough data	No - data quality limited

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
32004	Ise Brook	Harrowden	72	1944 to 2016	Yes	Yes	No - data quality limited
32007	Nene/Brampton	St Andrews Total	77	1940 to 2016	Yes	Yes	No - data quality limited
32008	Nene/Kislingbury	Dodford	48	1969 to 2016	Yes	Yes	Yes
32010	Nene	Wansford	51	1963 to 2016	Yes	Yes	Yes
33007	Nar	Marham	35	1981 to 2016	Yes	Not enough data	Yes
33014	Lark	Temple	57	1960 to 2016	Yes	Yes	Yes
33018	Tove	Cappenham Bridge	54	1963 to 2016	Yes	Yes	Yes
33019	Thet	Melford Bridge	57	1960 to 2016	Yes	Yes	Yes
33021	Rhee	Burnt Mill	55	1962 to 2016	Yes	Yes	Yes
33022	Ivel	Blunham	52	1965 to 2016	Yes	Yes	No - data quality limited
33023	Lea Brook	Beck Bridge	53	1963 to 2016	Yes	Yes	No - data quality limited
33024	Cam	Dernford	53	1963 to 2016	Yes	Yes	No - data quality limited
33027	Rhee	Wimpole	52	1965 to 2016	Yes	Yes	No - data quality limited
33031	Broughton Brook	Broughton	44	1970 to 2016	Yes	Yes	Yes
33032	Heacham	Heacham	49	1966 to 2016	Yes	Yes	Yes
33034	Little Ouse	Abbey Heath	48	1967 to 2016	Yes	Yes	Yes
33037	Bedford Ouse	Newport Pagnell Total	48	1969 to 2016	Yes	Yes	Yes
33039	Bedford Ouse	Roxton	45	1972 to 2016	Yes	Yes	Yes
33044	Thet	Bridgham	38	1979 to 2016	Yes	Not enough data	Yes
33051	Cam	Chesterford	48	1969 to 2016	Yes	Yes	Yes
33052	Swaffham Lode	Swaffham Bulbeck	43	1973 to 2016	Yes	Yes	No - data quality limited
33054	Babingley	Castle Rising	41	1976 to 2016	Yes	Yes	Yes
33055	Granta	Babraham	39	1978 to 2016	Yes	Not enough data	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
33057	Ouzel	Leighton Buzzard	37	1976 to 2016	Yes	Not enough data	Yes
33058	Ouzel	Bletchley	37	1978 to 2016	Yes	Not enough data	Yes
34001	Yare	Colney	59	1958 to 2016	Yes	Yes	Yes
34003	Bure	Ingworth	57	1959 to 2016	Yes	Yes	Yes
34005	Tud	Costessey Park	56	1961 to 2016	Yes	Yes	Yes
34006	Waveney	Needham Mill	38	1979 to 2016	Yes	Not enough data	Yes
34007	Dove	Oakley Park	41	1966 to 2006	Yes	Yes	Yes
34008	Ant	Honing Lock	49	1966 to 2016	Yes	Yes	Yes
34011	Wensum	Fakenham	51	1966 to 2016	Yes	Yes	Yes
34012	Burn	Burnham Overy	51	1966 to 2016	Yes	Yes	Yes
35003	Alde	Farnham	56	1961 to 2016	Yes	Yes	No - data quality limited
35008	Gipping	Stowmarket	50	1964 to 2016	Yes	Yes	Yes
36003	Box	Polstead	56	1961 to 2016	Yes	Yes	Yes
36004	Chad Brook	Long Melford	50	1967 to 2016	Yes	Yes	Yes
36006	Stour	Langham	42	1962 to 2003	Yes	Not enough data	Yes
36007	Belchamp Brook	Bardfield Bridge	52	1965 to 2016	Yes	Yes	Yes
36008	Stour	Westmill	57	1960 to 2016	Yes	Yes	No - data quality limited
36009	Brett	Cockfield	50	1967 to 2016	Yes	Yes	No - data quality limited
36010	Bumpstead Brook	Broad Green	50	1967 to 2016	Yes	Yes	Yes
36012	Stour	Kedington	50	1967 to 2016	Yes	Yes	Yes
37001	Roding	Redbridge	67	1950 to 2016	Yes	Yes	Yes
37003	Ter	Crabbs Bridge	52	1963 to 2016	Yes	Yes	No - data quality limited
37005	Colne	Lexden	57	1960 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
37006	Can	Beach's Mill	55	1962 to 2016	Yes	Yes	Yes
37007	Wid	Writtle	51	1964 to 2016	Yes	Yes	Yes
37008	Chelmer	Springfield	51	1965 to 2016	Yes	Yes	Yes
37009	Brain	Guithavon Valley	55	1962 to 2016	Yes	Yes	Yes
37010	Blackwater	Appleford Bridge	55	1962 to 2016	Yes	Yes	Yes
37011	Chelmer	Churchend	51	1963 to 2016	Yes	Yes	No - data quality limited
37012	Colne	Poolstreet	53	1964 to 2016	Yes	Yes	No - data quality limited
37013	Sandon Brook	Sandon Bridge	51	1963 to 2016	Yes	Yes	Yes
37014	Roding	High Ongar	54	1963 to 2016	Yes	Yes	Yes
37016	Pant	Copford Hall	52	1965 to 2016	Yes	Yes	Yes
37017	Blackwater	Stisted	48	1969 to 2016	Yes	Yes	Yes
37018	Ingrebourne	Gaynes Park	47	1970 to 2016	Yes	Yes	Yes
37019	Beam	Bretons Farm	52	1965 to 2016	Yes	Yes	Yes
37020	Chelmer	Felsted	47	1970 to 2016	Yes	Yes	Yes
37031	Crouch	Wickford	41	1976 to 2016	Yes	Yes	Yes
37033	Eastwood Brook	Eastwood	42	1974 to 2016	Yes	Yes	Yes
38001	Lee	Feildes Weir	40	1977 to 2016	Yes	Not enough data	No - data quality limited
38002	Ash	Mardock	76	1939 to 2016	Yes	Yes	Yes
38003	Mimram	Panshanger Park	65	1952 to 2016	Yes	Yes	Yes
38004	Rib	Wadesmill	58	1959 to 2016	Yes	Yes	Yes
38007	Canons Brook	Elizabeth Way	67	1950 to 2016	Yes	Yes	Yes
38018	Upper Lee	Water Hall	46	1971 to 2016	Yes	Yes	Yes
38020	Cobbins Brook	Sewardstone Road	44	1971 to 2014	Yes	Yes	No - data quality limited

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
38021	Turkey Brook	Albany Park	46	1971 to 2016	Yes	Yes	Yes
38022	Pymmes Brook	Edmonton Silver Street	61	1954 to 2016	Yes	Yes	Yes
39001	Thames	Kingston	134	1882 to 2016	Yes	Yes	Yes
39002	Thames	Days Weir	78	1938 to 2015	Yes	Yes	Yes
39003	Wandle	South Wimbledon	53	1964 to 2016	Yes	Yes	No - data quality limited
39005	Beverley Brook	Wimbledon Common	56	1961 to 2016	Yes	Yes	Yes
39006	Windrush	Newbridge	67	1950 to 2016	Yes	Yes	No - data quality limited
39007	Blackwater	Swallowfield	64	1953 to 2016	Yes	Yes	Yes
39010	Colne	Denham	65	1952 to 2016	Yes	Yes	Yes
39011	Wey	Tilford	45	1972 to 2016	Yes	Yes	Yes
39012	Hogsmill	Kingston upon Thames	62	1955 to 2016	Yes	Yes	No - data quality limited
39014	Ver	Hansteads	61	1956 to 2016	Yes	Yes	Yes
39019	Lambourn	Shaw	55	1962 to 2016	Yes	Yes	Yes
39020	Coln	Bibury	54	1963 to 2016	Yes	Yes	Yes
39022	Loddon	Sheepbridge	52	1965 to 2016	Yes	Yes	Yes
39023	Wye	Hedsor	53	1964 to 2016	Yes	Yes	Yes
39025	Enborne	Brimpton	32	1986 to 2017	Yes	Not enough data	Yes
39026	Cherwell	Banbury	50	1966 to 2017	Yes	Yes	Yes
39028	Dun	Hungerford	49	1968 to 2016	Yes	Yes	Yes
39033	Winterbourne Stream	Bagnor	55	1962 to 2016	Yes	Yes	Yes
39036	Law Brook	Albury	50	1967 to 2016	Yes	Yes	Yes
39037	Kennet	Marlborough	45	1972 to 2016	Yes	Yes	Yes
39042	Leach	Priory Mill Lechlade	45	1972 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
39049	Silk Stream	Colindeep Lane	41	1978 to 2016	Yes	Yes	No - data quality limited
39052	The Cut	Binfield	56	1957 to 2016	Yes	Yes	Yes
39053	Mole	Horley	56	1961 to 2016	Yes	Yes	No - data quality limited
39056	Ravensbourne	Catford Hill	43	1974 to 2016	Yes	Yes	No - data quality limited
39057	Crane	Cranford Park	33	1985 to 2017	Yes	Not enough data	Yes
39086	Gatwick Stream	Gatwick Link	42	1975 to 2016	Yes	Yes	Yes
39088	Chess	Rickmansworth	43	1974 to 2016	Yes	Yes	Yes
39089	Gade	Bury Mill	43	1974 to 2016	Yes	Yes	Yes
39090	Cole	Inglesham	41	1976 to 2016	Yes	Yes	Yes
39096	Wealdstone Brook	Wembley	41	1975 to 2016	Yes	Yes	Yes
40005	Beult	Stilebridge	42	1958 to 2000	Yes	Not enough data	No - data quality limited
40007	Medway	Chafford / Colliersland Bridge	56	1960 to 2016	Yes	Yes	Yes
40010	Eden	Penshurst / Vexour Bridge	56	1961 to 2016	Yes	Yes	Yes
40012	Darent	Hawley	27	1990 to 2016	Yes	Not enough data	Yes
40016	Cray	Crayford	27	1990 to 2016	Yes	Not enough data	No - data quality limited
41005	Ouse	Gold Bridge	56	1960 to 2016	Yes	Yes	Yes
41011	Rother	Iping Mill	50	1967 to 2016	Yes	Yes	Yes
41012	Adur E Branch	Sakeham	50	1967 to 2016	Yes	Yes	No - data quality limited
41015	Ems	Westbourne	50	1967 to 2016	Yes	Yes	Yes
41016	Cuckmere	Cowbeech	50	1967 to 2016	Yes	Yes	Yes
41020	Bevern Stream	Clappers Bridge	48	1969 to 2016	Yes	Yes	No - data quality limited
41022	Lod	Halfway Bridge	47	1970 to 2016	Yes	Yes	Yes
41023	Lavant	Graylingwell	44	1971 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
42007	Alre	Drove Lane Alresford	44	1974 to 2017	Yes	Yes	Yes
42008	Cheriton Stream	Sewards Bridge	46	1970 to 2016	Yes	Yes	No - data quality limited
42009	Candover Stream	Borough Bridge	47	1970 to 2016	Yes	Yes	Yes
42010	Itchen	Highbridge & Allbrook Total	59	1958 to 2016	Yes	Yes	Yes
42011	Hamble	Frogmill	45	1972 to 2016	Yes	Yes	Yes
42012	Anton	Fullerton	44	1973 to 2016	Yes	Yes	Yes
42014	Blackwater	Ower	41	1976 to 2016	Yes	Yes	No - data quality limited
43003	Avon	East Mills Total	46	1965 to 2016	No - gaps	Yes	Yes
43005	Avon	Amesbury	52	1965 to 2016	Yes	Yes	Yes
43006	Nadder	Wilton	50	1966 to 2016	Yes	Yes	Yes
43007	Stour	Throop	44	1973 to 2016	Yes	Yes	Yes
43008	Wylye	South Newton	51	1966 to 2016	Yes	Yes	Yes
43009	Stour	Hammoon	49	1968 to 2016	Yes	Yes	Yes
43012	Wylye	Norton Bavant	48	1969 to 2016	Yes	Yes	Yes
43014	East Avon	Upavon	46	1970 to 2016	Yes	Yes	Yes
43017	Allen	Upavon	47	1970 to 2016	Yes	Yes	Yes
43018	Allen	Walford Mill	43	1974 to 2016	Yes	Yes	Yes
44004	Frome	Dorchester Total	47	1969 to 2016	Yes	Yes	No - data quality limited
44006	Sydling Water	Sydling St Nicholas	43	1969 to 2016	No - gaps	Yes	No - data quality limited
44009	Wey	Broadwey	40	1975 to 2016	Yes	Not enough data	Yes
45001	Exe	Thorverton	61	1956 to 2016	Yes	Yes	Yes
45002	Exe	Stoodleigh	56	1960 to 2016	Yes	Yes	Yes
45003	Culm	Wood Mill	55	1962 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
45004	Axe	Whitford	53	1964 to 2016	Yes	Yes	Yes
45005	Otter	Dotton	56	1961 to 2016	Yes	Yes	Yes
45008	Otter	Fenny Bridges	43	1974 to 2016	Yes	Yes	Yes
45009	Exe	Pixton	51	1966 to 2016	Yes	Yes	Yes
45012	Creedy	Cowley	52	1965 to 2016	Yes	Yes	Yes
46003	Dart	Austins Bridge	59	1958 to 2016	Yes	Yes	Yes
46005	East Dart	Bellever	53	1964 to 2016	Yes	Yes	No - data quality limited
46006	Erme	Ermington	43	1974 to 2016	Yes	Yes	Yes
46007	West Dart	Dunnabridge	27	1990 to 2016	Yes	Not enough data	No - data quality limited
46008	Avon	Loddiswell	27	1990 to 2016	Yes	Not enough data	Yes
46013	Bovey	Bovey Parke	13	2004 to 2016	No - gaps	Not enough data	No - data quality limited
47001	Tamar	Gunnislake	61	1956 to 2016	Yes	Yes	Yes
47004	Lynher	Pillaton Mill	56	1961 to 2016	Yes	Yes	Yes
47005	Ottery	Werrington Park	33	1985 to 2017	Yes	Not enough data	Yes
47006	Lyd	Lifton Park	29	1988 to 2016	Yes	Not enough data	Yes
47007	Yealm	Puslinch	54	1962 to 2016	Yes	Yes	No - data quality limited
47008	Thrushel	Tinhay	28	1989 to 2016	Yes	Not enough data	Yes
47009	Tiddy	Tideford	48	1969 to 2016	Yes	Yes	Yes
47014	Walkham	Horrabridge	44	1973 to 2016	Yes	Yes	Yes
47015	Tavy	Ludbrook	36	1981 to 2016	Yes	Not enough data	Yes
47020	Inny	Bealsmill	28	1988 to 2016	Yes	Not enough data	Yes
48001	Fowey	Trekeivesteps	48	1969 to 2016	Yes	Yes	Yes
48003	Fal	Tregony	54	1961 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
48004	Warleggan	Trengoffe	48	1969 to 2016	Yes	Yes	Yes
48005	Kenwyn	Truro	49	1968 to 2016	Yes	Yes	Yes
48007	Kennal	Ponsanooth	49	1968 to 2016	Yes	Yes	Yes
48011	Fowey	Restormel	34	1983 to 2016	Yes	Not enough data	Yes
49001	Camel	Denby	53	1964 to 2016	Yes	Yes	Yes
49002	Hayle	St Erth	60	1957 to 2016	Yes	Yes	Yes
49003	De Lank	De Lank	51	1966 to 2016	Yes	Yes	Yes
50001	Taw	Umberleigh	59	1958 to 2016	Yes	Yes	Yes
50002	Torridge	Torrington	56	1960 to 2016	Yes	Yes	Yes
50005	West Okement	Vellake	43	1974 to 2016	Yes	Yes	No - data quality limited
50006	Mole	Woodleigh	52	1965 to 2016	Yes	Yes	No - data quality limited
50007	Taw	Taw Bridge	44	1973 to 2016	Yes	Yes	No - data quality limited
51001	Doniford Stream	Swill Bridge	51	1966 to 2016	Yes	Yes	Yes
51003	Washford	Beggearn Huish	50	1966 to 2016	Yes	Yes	Yes
52003	Halsewater	Halsewater	56	1961 to 2016	Yes	Yes	No - data quality limited
52004	Isle	Ashford Mill	55	1962 to 2016	Yes	Yes	Yes
52005	Tone	Bishops Hull	56	1961 to 2016	Yes	Yes	Yes
52006	Yeo	Pen Mill	55	1962 to 2016	Yes	Yes	Yes
52007	Parrett	Chiselborough	51	1966 to 2016	Yes	Yes	Yes
52009	Sheppey	Fenny Castle	50	1964 to 2016	Yes	Yes	Yes
52010	Brue	Lovington	34	1983 to 2016	Yes	Not enough data	Yes
52014	Tone	Greenham	51	1966 to 2016	Yes	Yes	No - data quality limited
52016	Currypool Stream	Currypool Farm	47	1970 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
52017	Congresbury Yeo	Iwood	42	1973 to 2016	Yes	Yes	No - data quality limited
53004	Chew	Compton Dando	57	1958 to 2015	Yes	Yes	Yes
53005	Midford Brook	Midford	56	1961 to 2016	Yes	Yes	Yes
53006	Frome (Bristol)	Frenchay	56	1961 to 2016	Yes	Yes	Yes
53007	Frome (Somerset)	Tellisford	56	1961 to 2016	Yes	Yes	Yes
53008	Avon	Great Somerford	54	1963 to 2016	Yes	Yes	Yes
53009	Wellow Brook	Wellow	51	1966 to 2016	Yes	Yes	Yes
53013	Marden	Stanley	47	1970 to 2016	Yes	Yes	Yes
53017	Boyd	Bitton	44	1973 to 2016	Yes	Yes	Yes
53018	Avon	Bathford	48	1969 to 2016	Yes	Yes	Yes
53023	Sherston Avon	Fosseway	41	1976 to 2016	Yes	Yes	Yes
53025	Mells	Vallis	38	1979 to 2016	Yes	Not enough data	Yes
53026	Frome (Bristol)	Frampton Cotterell	39	1978 to 2016	Yes	Not enough data	Yes
53028	By Brook	Middlehill	36	1981 to 2016	Yes	Not enough data	Yes
53029	Biss	Trowbridge	34	1983 to 2016	Yes	Not enough data	Yes
54001	Severn	Bewdley	94	1923 to 2016	Yes	Yes	Yes
54002	Avon	Evesham	80	1937 to 2016	Yes	Yes	Yes
54004	Sowe	Stoneleigh	38	1979 to 2016	Yes	Not enough data	Yes
54005	Severn	Montford	65	1952 to 2016	Yes	Yes	Yes
54007	Arrow	Broom	40	1977 to 2016	Yes	Not enough data	No - data quality limited
54008	Teme	Tenbury	61	1956 to 2016	Yes	Yes	Yes
54011	Salwarpe	Harford Hill	59	1958 to 2016	Yes	Yes	No - data quality limited
54012	Tern	Walcot	58	1959 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
54014	Severn	Abermule	49	1968 to 2016	Yes	Yes	Yes
54016	Roden	Rodington	40	1977 to 2016	Yes	Not enough data	Yes
54018	Rea Brook	Hookagate	55	1962 to 2016	Yes	Yes	Yes
54019	Avon	Stareton	54	1962 to 2016	Yes	Yes	No - data quality limited
54020	Perry	Yeaton	54	1963 to 2016	Yes	Yes	Yes
54022	Severn	Plynlimon flume	38	1971 to 2008	Yes	Yes	Yes
54024	Worfe	Burcote	46	1969 to 2016	Yes	Yes	Yes
54025	Dulas	Rhos-y-pentref	48	1969 to 2016	Yes	Yes	Yes
54027	Frome	Ebley Mill	47	1970 to 2016	Yes	Yes	Yes
54028	Vyrnwy	Llanymynech	48	1969 to 2016	Yes	Yes	Yes
54029	Teme	Knightsford Bridge	47	1970 to 2016	Yes	Yes	Yes
54034	Dowles Brook	Oak Cottage	46	1971 to 2016	Yes	Yes	No - data quality limited
54036	Isbourne	Hinton on the Green	44	1972 to 2016	Yes	Yes	Yes
54038	Tanat	Llanyblodwel	45	1972 to 2016	Yes	Yes	No - data quality limited
54040	Meese	Tibberton	44	1973 to 2016	Yes	Yes	Yes
54041	Tern	Eaton upon Tern	43	1972 to 2014	Yes	Yes	Yes
54044	Tern	Ternhill	45	1972 to 2016	Yes	Yes	No - data quality limited
54052	Bailey Brook	Ternhill	43	1972 to 2016	Yes	Yes	No - data quality limited
54057	Severn	Haw Bridge	45	1972 to 2016	Yes	Yes	Yes
54091	Severn	Hafren Flume	34	1975 to 2008	Yes	Yes	Yes
54092	Hore	Hore Flume	36	1973 to 2008	Yes	Yes	No - data quality limited
54102	Avon	Lilbourne	32	1985 to 2016	Yes	Not enough data	No - data quality limited
54106	Stour (Warks)	Shipston	31	1986 to 2016	Yes	Not enough data	No - data quality limited

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
54114	Avon	Warwick	42	1972 to 2016	Yes	Yes	Yes
54906	Stour (Worcs)	Kidderminster Gilt Edge	33	1952 to 1984	Yes	Not enough data	No - data quality limited
55002	Wye	Belmont	52	1966 to 2017	Yes	Yes	Yes
55007	Wye	Erwood	78	1938 to 2016	Yes	Yes	Yes
55008	Wye	Cefn Brwyn	58	1951 to 2008	Yes	Yes	Yes
55012	Irfon	Cilmery	48	1966 to 2016	Yes	Yes	No - data quality limited
55013	Arrow	Titley Mill	49	1966 to 2016	Yes	Yes	Yes
55014	Lugg	Byton	49	1966 to 2016	Yes	Yes	No - data quality limited
55016	Ithon	Disserth	42	1972 to 2016	Yes	Yes	No - data quality limited
55018	Frome	Yarkhill	49	1967 to 2016	Yes	Yes	No - data quality limited
55021	Lugg	Butts Bridge	46	1969 to 2016	Yes	Yes	Yes
55023	Wye	Redbrook	47	1969 to 2016	Yes	Yes	Yes
55025	Llynfi	Three Cocks	46	1970 to 2016	Yes	Yes	Yes
55026	Wye	Ddol Farm	48	1969 to 2016	Yes	Yes	Yes
55029	Monnow	Grosmont	44	1973 to 2016	Yes	Yes	Yes
55033	Wye	Gwy flume	33	1973 to 2008	Yes	Yes	Yes
55034	Cyff	Cyff flume	34	1973 to 2008	Yes	Yes	No - data quality limited
56001	Usk	Chainbridge	56	1957 to 2016	Yes	Yes	Yes
56002	Ebbw	Rhiwderin	58	1957 to 2016	Yes	Yes	Yes
56012	Grwyne	Millbrook	42	1971 to 2016	Yes	Yes	No - data quality limited
57004	Cynon	Abercynon	56	1961 to 2016	Yes	Yes	Yes
57005	Taff	Pontypridd	50	1967 to 2016	Yes	Yes	Yes
57006	Rhondda	Trehafod	48	1968 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
57007	Taff	Fiddlers Elbow	44	1973 to 2016	Yes	Yes	Yes
57008	Rhymney	Llanedeyrn	45	1972 to 2016	Yes	Yes	No - data quality limited
57009	Ely	St Fagans	43	1974 to 2016	Yes	Yes	Yes
57015	Taff	Merthyr Tydfil	39	1978 to 2016	Yes	Not enough data	Yes
58001	Ogmore	Bridgend	34	1983 to 2016	Yes	Not enough data	Yes
58002	Neath	Resolven	39	1978 to 2016	Yes	Not enough data	Yes
58005	Ogmore	Brynmenyn	48	1969 to 2016	Yes	Yes	Yes
58006	Mellte	Pontneddfechan	46	1971 to 2016	Yes	Yes	Yes
58007	Llynfi	Coytrahen	47	1970 to 2016	Yes	Yes	Yes
58008	Dulais	Cilfrew	45	1972 to 2016	Yes	Yes	Yes
58012	Afan	Marcroft Weir	37	1979 to 2016	Yes	Not enough data	Yes
59001	Tawe	Ynystanglws	43	1973 to 2016	Yes	Yes	Yes
59002	Loughor	Tir-y-dail	50	1967 to 2016	Yes	Yes	Yes
60001	Tywi	Ty Castell	45	1972 to 2016	Yes	Yes	Yes
60002	Cothi	Felin Mynachdy	56	1961 to 2016	Yes	Yes	Yes
60003	Taf	Clog-y-Fran	48	1964 to 2011	Yes	Yes	No - data quality limited
60006	Gwili	Glangwili	49	1968 to 2016	Yes	Yes	Yes
60007	Tywi	Dolau Hirion	45	1972 to 2016	Yes	Yes	Yes
60009	Sawdde	Felin-y-cwm	38	1970 to 2007	Yes	Yes	Yes
60012	Twrch	Ddol Las	27	1990 to 2016	Yes	Not enough data	No - data quality limited
61002	Eastern Cleddau	Canaston Bridge	43	1974 to 2016	Yes	Yes	No - data quality limited
61003	Gwaun	Cilrhedyn Bridge	48	1968 to 2016	Yes	Yes	No - data quality limited
62001	Teifi	Glanteifi	60	1959 to 2018	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
63001	Ystwyth	Pont Llolwyn	56	1961 to 2016	Yes	Yes	Yes
64001	Dyfi	Dyfi Bridge	54	1962 to 2015	Yes	Yes	No - data quality limited
64006	Leri	Dolybont	40	1974 to 2013	Yes	Yes	No - data quality limited
65001	Glaslyn	Beddgelert	49	1967 to 2016	Yes	Yes	No - data quality limited
65005	Erch	Pencaenewydd	45	1972 to 2016	Yes	Yes	Yes
65006	Seiont	Peblig Mill	32	1984 to 2016	Yes	Not enough data	No - data quality limited
65007	Dwyfor	Garndolbenmaen	43	1974 to 2016	Yes	Yes	No - data quality limited
66001	Clwyd	Pont-y-Cambwll	42	1973 to 2016	Yes	Yes	No - data quality limited
66004	Wheeler	Bodfari	43	1974 to 2016	Yes	Yes	Yes
66005	Clwyd	Ruthin Weir	43	1972 to 2018	Yes	Yes	Yes
66006	Elwy	Pont-y-Gwyddel	43	1974 to 2016	Yes	Yes	Yes
66011	Conwy	Cwmlanerch	52	1964 to 2016	Yes	Yes	No - data quality limited
67006	Alwen	Druid	57	1960 to 2016	Yes	Yes	No - data quality limited
67008	Alyn	Pont-y-Capel	53	1964 to 2016	Yes	Yes	Yes
67009	Alyn	Rhydymwyn	61	1956 to 2016	Yes	Yes	Yes
67010	Gelyn	Cynefail	27	1990 to 2016	Yes	Not enough data	No - data quality limited
67015	Dee	Manley Hall	47	1969 to 2016	Yes	Yes	Yes
68001	Weaver	Ashbrook	80	1937 to 2016	Yes	Yes	No - data quality limited
68003	Dane	Rudheath	38	1979 to 2016	Yes	Not enough data	No - data quality limited
68005	Weaver	Audlem	48	1969 to 2016	Yes	Yes	No - data quality limited
68007	Wincham Brook	Lostock Gralam	54	1960 to 2015	Yes	Yes	No - data quality limited
68018	Dane	Congleton Park	50	1967 to 2016	Yes	Yes	Yes
69007	Mersey	Ashton Weir	59	1958 to 2016	Yes	Yes	No - data quality limited

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
69008	Dean	Stanneylands	50	1966 to 2016	Yes	Yes	Yes
69012	Bollin	Wilmslow	50	1967 to 2016	Yes	Yes	Yes
69015	Etherow	Compstall	48	1968 to 2016	Yes	Yes	Yes
69017	Goyt	Marple Bridge	47	1969 to 2016	Yes	Yes	Yes
69020	Medlock	London Road	48	1969 to 2016	Yes	Yes	Yes
69023	Roch	Blackford Bridge	69	1948 to 2016	Yes	Yes	Yes
69024	Croal	Farnworth Weir	69	1948 to 2016	Yes	Yes	Yes
69025	Irwell	Manchester Racecourse	76	1941 to 2016	Yes	Yes	Yes
69027	Tame	Portwood	65	1952 to 2016	Yes	Yes	Yes
69028	Mersey	Brinksway	62	1955 to 2016	Yes	Yes	Yes
69032	Alt	Kirkby	39	1977 to 2016	Yes	Not enough data	Yes
69041	Tame	Broomstairs	50	1967 to 2016	Yes	Yes	Yes
69044	Irwell	Bury Ground	44	1973 to 2016	Yes	Yes	Yes
70004	Yarrow	Croston Mill	42	1975 to 2016	Yes	Yes	Yes
71001	Ribble	Samlesbury	56	1960 to 2016	Yes	Yes	Yes
71004	Calder	Whalley Weir	47	1970 to 2016	Yes	Yes	Yes
71006	Ribble	Henthorn	49	1968 to 2016	Yes	Yes	Yes
71008	Hodder	Hodder Place	48	1969 to 2016	Yes	Yes	Yes
71009	Ribble	New Jumbles Rock	47	1970 to 2016	Yes	Yes	Yes
71010	Pendle Water	Barden Lane	45	1972 to 2016	Yes	Yes	Yes
71011	Ribble	Arnford	47	1970 to 2016	Yes	Yes	Yes
71013	Darwen	Ewood	44	1973 to 2016	Yes	Yes	Yes
71014	Darwen	Blue Bridge	42	1975 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
72004	Lune	Caton	49	1968 to 2016	Yes	Yes	Yes
72005	Lune	Killington	48	1969 to 2016	Yes	Yes	Yes
72007	Brock	upstream of A6	39	1978 to 2016	Yes	Not enough data	Yes
72009	Wenning	Wennington	46	1971 to 2016	Yes	Yes	Yes
72011	Rawthey	Brigflatts	48	1968 to 2015	Yes	Yes	Yes
72014	Conder	Galgate	50	1966 to 2016	Yes	Yes	Yes
72015	Lune	Lunes Bridge	39	1979 to 2016	Yes	Not enough data	Yes
72807	Wenning	Hornby	60	1956 to 2016	Yes	Yes	Yes
73002	Crake	Low Nibthwaite	53	1962 to 2016	Yes	Yes	Yes
73005	Kent	Sedgwick	50	1968 to 2017	Yes	Yes	Yes
73008	Bela	Beetham	48	1969 to 2016	Yes	Yes	Yes
73009	Sprint	Sprint Mill	48	1969 to 2016	Yes	Yes	Yes
73010	Leven	Newby Bridge	78	1939 to 2016	Yes	Yes	Yes
73011	Mint	Mint Bridge	48	1969 to 2016	Yes	Yes	Yes
73012	Kent	Victoria Bridge	39	1978 to 2016	Yes	Not enough data	No - data quality limited
73014	Brathay	Jeffy Knotts	40	1975 to 2016	Yes	Not enough data	No - data quality limited
73017	Kent	Bowston	36	1981 to 2016	Yes	Not enough data	No - data quality limited
74001	Duddon	Duddon Hall	50	1967 to 2016	Yes	Yes	Yes
74002	Irt	Galesyke	48	1968 to 2016	Yes	Yes	No - data quality limited
74003	Ehen	Bleach Green	44	1973 to 2016	Yes	Yes	No - data quality limited
74006	Calder	Calder Hall	44	1973 to 2016	Yes	Yes	Yes
74007	Esk	Cropple How	43	1974 to 2016	Yes	Yes	No - data quality limited
74008	Duddon	Ulpha	44	1973 to 2016	Yes	Yes	Yes

Gauge	River	Location	Record length	Record period	Included in trend tests?	Included in split sample tests?	Included in non-stationary model fitting?
75001	St Johns Beck	Thirlmere Reservoir	44	1974 to 2016	Yes	Yes	Yes
75002	Derwent	Camerton	57	1960 to 2016	Yes	Yes	Yes
75003	Derwent	Ouse Bridge	50	1967 to 2016	Yes	Yes	Yes
75004	Cocker	Southwaite Bridge	51	1966 to 2016	Yes	Yes	Yes
75005	Derwent	Portinscale	44	1972 to 2015	Yes	Yes	Yes
75007	Glenderamackin	Threlkeld	31	1986 to 2016	Yes	Not enough data	No - data quality limited
75009	Greta	Low Briery	46	1971 to 2016	Yes	Yes	Yes
75017	Ellen	Bullgill	41	1976 to 2016	Yes	Yes	No - data quality limited
76003	Eamont	Udford	55	1961 to 2015	Yes	Yes	Yes
76004	Lowther	Eamont Bridge	55	1962 to 2016	Yes	Yes	Yes
76005	Eden	Temple Sowerby	53	1964 to 2016	Yes	Yes	Yes
76007	Eden	Sheepmount	52	1966 to 2017	Yes	Yes	Yes
76008	Irthing	Greenholme	50	1967 to 2016	Yes	Yes	Yes
76010	Petteril	Harraby Green	45	1970 to 2015	Yes	Yes	Yes
76014	Eden	Kirkby Stephen	46	1971 to 2016	Yes	Yes	Yes
76015	Eamont	Pooley Bridge	41	1976 to 2016	Yes	Yes	Yes
76017	Eden	Great Corby	58	1959 to 2016	Yes	Yes	No - data quality limited
77001	Esk	Netherby	42	1961 to 2002	Yes	Not enough data	No - data quality limited

Would you like to find out more about us or about your environment?

Then call us on

03708 506 506 (Monday to Friday, 8am to 6pm)

email

enquiries@environment-agency.gov.uk

or visit our website

www.gov.uk/environment-agency

incident hotline **0800 807060** (24 hours)

floodline **0345 988 1188 / 0845 988 1188** (24 hours)

Find out about call charges (www.gov.uk/call-charges)

Environment first: Are you viewing this on screen? Please consider the environment and only print if absolutely necessary. If you are reading a paper copy, please don't forget to reuse and recycle if possible.