Journal of Hydrology 541 (2016) 824–838

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Temporal clustering of floods in Germany: Do flood-rich and flood-poor periods exist?

Bruno Merz [a,b,*], Viet Dung Nguyen [a], Sergiy Vorogushyn [a]

[a] GFZ German Research Centre for Geosciences, Section Hydrology, 14473 Potsdam, Germany
[b] Institute of Earth and Environmental Science, University of Potsdam, 14476 Potsdam, Germany

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 25 July 2016
Accepted 27 July 2016
Available online 5 August 2016
This manuscript was handled by Andras Bardossy, Editor-in-Chief

Keywords:
Climate variability
Flooding
Temporal clustering
Index of dispersion
Kernel occurrence rate

ABSTRACT

The repeated occurrence of exceptional floods within a few years, such as the Rhine floods in 1993 and 1995 and the Elbe and Danube floods in 2002 and 2013, suggests that floods in Central Europe may be organized in flood-rich and flood-poor periods. This hypothesis is studied by testing the significance of temporal clustering in flood occurrence (peak-over-threshold) time series for 68 catchments across Germany for the period 1932–2005. To assess the robustness of the results, different methods are used: Firstly, the index of dispersion, which quantifies the departure from a homogeneous Poisson process, is investigated. Further, the time-variation of the flood occurrence rate is derived by non-parametric kernel implementation and the significance of clustering is evaluated via parametric and non-parametric tests. Although the methods give consistent overall results, the specific results differ considerably. Hence, we recommend applying different methods when investigating flood clustering. For flood estimation and risk management, it is of relevance to understand whether clustering changes with flood severity and time scale. To this end, clustering is assessed for different thresholds and time scales. It is found that the majority of catchments show temporal clustering at the 5% significance level for low thresholds and time scales of one to a few years. However, clustering decreases substantially with increasing threshold and time scale. We hypothesize that flood clustering in Germany is mainly caused by catchment memory effects along with intra- to inter-annual climate variability, and that decadal climate variability plays a minor role.

1. Introduction

Only eleven years after the disastrous flood in August 2002 in the Elbe and Danube catchments in Central Europe (the most expensive natural disaster for Germany so far), the same catchments were hit by another exceptional flood. The June 2013 flood was even more severe in hydrological terms although damages were significantly lower (Merz et al., 2014b; Schröter et al., 2015). Another example for exceptional flooding occurring within rather short time periods is the Rhine flood in December 1993, which was followed by a hydro-meteorologically very similar event in January 1995. Such reoccurrences give rise to the hypothesis that floods in Central Europe are temporally organized in flood-rich and flood-poor periods.

Temporal clustering of floods may have considerable consequences for flood estimation, flood design and risk management (Merz et al., 2014a), and clustering of catastrophic events is an important issue for the insurance industry when modelling the pricing of insurance contracts (Khare et al., 2015). Flood design is typically based on the T-year flood, i.e. the flood discharge that has a 1/T probability of being reached or exceeded in a given year. Based on the usual iid (independent, identically distributed) assumption of flood frequency analysis, the T-year flood quantile is assumed constant in time. However, temporal clustering may introduce serial correlation in the flood time series and invalidate the independence assumption. Serial correlation may reduce the information content of the sample and increase the uncertainty for flood quantile estimation (Koutsoyiannis, 2005). Further, temporal variations in the frequency and magnitude of flooding may bias flood design. The relevance of this effect depends on the ratio of oscillation period and observation length. If the oscillation period is significantly smaller than the observation length that is used for flood estimation, the effects of clustering may be negligible for design purposes (Jain and Lall, 2001). On the other hand, decadal-scale fluctuations may significantly bias estimates that are based on 30 or 40 years of record (Hirschboeck, 1988). Hence, it is of

* Corresponding author at: GFZ German Research Centre for Geosciences, Section Hydrology, 14473 Potsdam, Germany.
E-mail address: bmerz@gfz-potsdam.de (B. Merz).

http://dx.doi.org/10.1016/j.jhydrol.2016.07.041
0022-1694/

utmost importance to understand not only if clustering exists but how clustering changes with the time scale.

Flood clustering is typically explained by linkages between flood frequency or magnitude and climate. There are well-organized modes of inter-annual, inter-decadal and lower-frequency climate variability (Barnston and Livezey, 1987). This variability may have a significant impact on the occurrence and magnitude of floods by changed atmospheric moisture uptake, transport and deposition (Hirschboeck, 1988). For example, ENSO (El Niño Southern Oscillation), with inter-annual variations in the range of two to seven years, has been linked to floods in Peru (Waylen and Caviedes, 1986), in the United States (Cayan et al., 1999; Jain and Lall, 2000, 2001; Sankarasubramanian and Lall, 2003), China (Lin et al., 2005; Zhang et al., 2007), Australia (Kiem et al., 2003). Ward et al. (2010, 2014) presented a global analysis of flood discharge sensitivities to ENSO based on observed and modelled river flows, respectively, suggesting complex sensitivity patterns, but significant correlation for catchments covering more than a third of the global land surface. Other climate modes, such as the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO), or the Pacific/North American Index (PNA) have been shown to lead to flood episodes of varying intensity as well (e.g., Pizaro and Lall, 2002; Bouwer et al., 2006; Kingston et al., 2006; Petrow et al., 2009; Delgado et al., 2012; Villarini et al., 2013).

There are a number of studies on temporal fluctuations in flood occurrence for Germany and neighbouring regions in the scientific literature on paleo and historical floods. (For a compilation of studies on flood changes for other European regions see Kundzewicz (2012) and Hall et al. (2014).) Swierczynski et al. (2013) reconstructed a 7100-year long flood record from varved lake sediments in Lake Mondsee in Austria. This record contains striking fluctuations in flood occurrence showing 18 flood-rich periods with durations between 30 and 50 years. Mudelsee et al. (2003) analysed historical flood data since CE 1500 for the Central European rivers Elbe and Oder. Significant variations in the occurrence rate of heavy floods during the past centuries were detected. For the same period, Sturm et al. (2001), Jacobeit et al. (2003) and Glaser et al. (2010) reconstructed flood occurrence from documentary evidence for several Central European rivers and found significant flood-rich and flood-poor periods. Phases of maximum flood activity in Bohemian rivers (Elbe River and others) since 1501 were concentrated in the latter part of the 16th century and in the 19th century (Brázdil et al., 2005). Flood-rich periods have been reported for rivers in France, for example for the Loire at Orléans and the Seine River at Paris (Brázdil et al., 2012). Schmocker-Fackel and Naef (2010a) identified four flood-rich periods (1560–1590, 1740–1790, 1820–1940, since 1970) for 14 Swiss catchments. At the river Rhine at the Swiss-German border, the highest number of summer floods since 1268 occurred in the period 1651–1750, with no severe winter floods since the late 19th century (Wetter et al., 2011). It can be summarised that the studies reconstructing historical floods for Germany and neighbouring regions typically conclude that flood-rich and flood-poor periods at the scale of several decades to centuries are a widespread and important phenomenon. It should be noted that most of these studies classified historical floods based on documentary evidence. In some instances, this classification builds on societal impacts which depend both on the magnitude of the flood and the exposure and vulnerability of the affected regions. Hence, the reconstruction of historical flood occurrences is not only associated with higher uncertainties compared to systematically recorded data, the derived flood frequencies might also depend on societal aspects.

Studies on temporal clustering of floods in Germany or neighbouring regions based on systematic data are rare. Based on 102 long-term records since 1900 from gauges across Europe, Mediero et al. (2015) found significant clustering of floods occurring in Atlantic and Continental regions covering northern and central Germany. Schmocker-Fackel and Naef (2010b) analysed a data set of 83 gauges in Switzerland, augmented with data on historical floods since 1850, and concluded that flood-rich periods alternated with flood-poor periods. Robson et al. (1998) and Robson (2002) analysed annual maxima data and peak-over-threshold flood data for a large number of catchments in UK and for different time periods and found fluctuations in flood occurrence and magnitude. In their review paper on flood regime changes in Europe, Hall et al. (2014, p. 2745) concluded that "... future flood change analyses of systematic data should actually focus on identifying flood-poor and flood-rich periods instead of only detecting whether trends exist...". We address this call by investigating temporal clustering of flood occurrence based on systematic data from 68 streamflow gauges across Germany.

A question, which has not received much attention in the flood clustering literature, is how to determine flood-rich and flood-poor periods. The majority of studies that investigate flood clustering, in particular studies analysing historical data, apply simple and subjective methods to decide whether flood-rich and flood-poor periods exist. Most often, a time series showing the number of flood occurrence within prefixed time windows, e.g. 10 or 30 years, is generated. Flood-rich periods are then identified visually or by simple rules, such as a period is considered as flood-rich if the number of floods is larger than the mean number plus one standard deviation. (An overview of methods used in studies on European floods in the period 1560–1810 is given in Schmocker-Fackel and Naef, 2010a.) To determine clustering objectively, we deploy objective measures and test the statistical significance of clustering.

A simple measure for clustering is the dispersion index which has been used in analysing clustering of storms (e.g., Mailier et al., 2006; Vitolo et al., 2009) and floods (e.g., Eastoe and Tawn, 2010; Mediero et al., 2015). It investigates the number of event counts in a time window and relates the variability of counts to the expected value. The dispersion index identifies times series where events do not occur randomly but are clustered in time. Another approach for quantifying clustering is the kernel occurrence rate estimation. It estimates the time variation of event counts as smooth function of time. It has been applied to analyse flood occurrence in Central Europe (Mudelsee et al., 2003, 2004) and Portugal (Silva et al., 2012). For both approaches, Monte Carlo simulation allows determining whether variations in flood occurrence deviate for a given significance level from the null hypothesis of time-constant occurrence rate.

In this paper, we address the following questions: (1) Is there significant temporal clustering in flooding in Germany? (2) Does the significance of clustering change with flood severity and time scale? (3) How can the significance of clustering be determined objectively? To answer these questions, we select flood times series from 68 gauges across Germany. To understand if clustering changes with flood severity, different peak-over-threshold (POT) time series are analysed by applying the dispersion index method and two variants of the kernel occurrence rate estimation. The three methods are deployed for different time scales to understand if the significance of clustering changes with time scale.

2. Study area and data

We study temporal clustering in flood occurrence at 68 streamflow gauges distributed across Germany with mean daily discharge observations (Fig. 1). The selection of the streamflow gauges was based on the following criteria: (1) complete coverage of Germany, (2) large number of gauges for a common time period, (3) medium- and large-scale catchments, (4) time series as long as possible, (5)

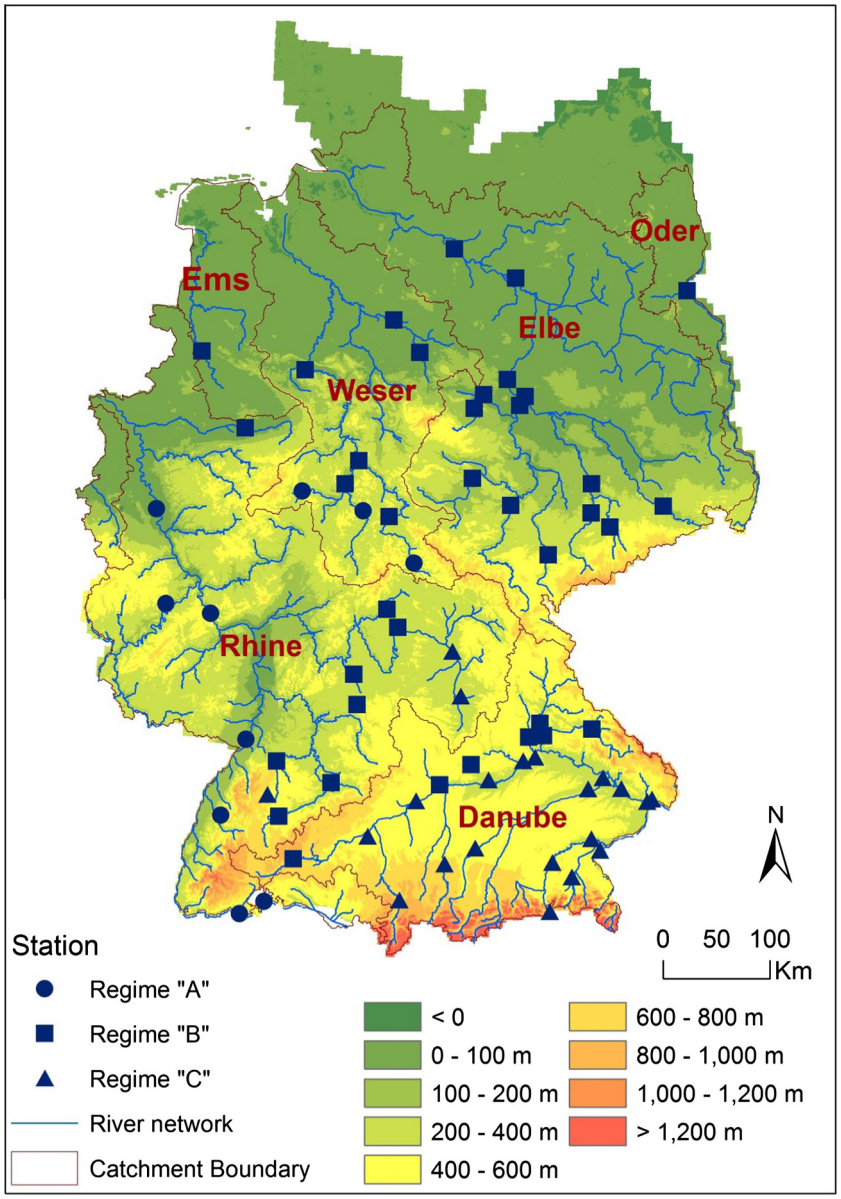

Fig. 1. Locations of 68 discharge gauges with daily streamflow for the hydrological years 1932–2005. Markers show the flood regime according to Beurton and Thieken (2009).

time series with little data gaps. Since some of these criteria conflict each other, a compromise had to be found.

Flood clustering is typically a consequence of climate variability. Since Germany has a substantial north-south gradient (in terms of topography and temperature) and west-east gradient (in terms of maritime and continental climate influence), flood regimes vary throughout Germany (Beurton and Thieken, 2009). Criterion (1) should ensure that the different flood regimes, which may be differently linked to climate variability, are covered (Fig. 1). Catchments in the western and central part of Germany have floods mostly in winter, and the annual maximum flow occurs rarely in other seasons. Gauges in north and east Germany are also dominated by winter floods, however, with a considerable share of spring and summer floods. Interestingly, the most extreme floods tend to be summer floods caused by widespread and intensive precipitation (Petrow et al., 2007). The southern region is dominated by summer floods, and flood generation typically involves snowmelt from the alpine Danube tributaries.

River catchments in Germany are hardly pristine, since Germany is densely populated and has a history of intensive water resources management. Land use changes are widespread, significant volumes of flood retention have been implemented in the last decades, and many rivers have experienced river training (e.g., Helms et al., 2002; Lammersen et al., 2002; Pfister et al., 2004; Vorogushyn and Merz, 2013). We have to assume that streamflow time series are contaminated by human interventions in the catchments and river systems, introducing noise into the time series and decreasing the signal-to-noise ratio. Criterion (2) is meant to work against this effect. Although local effects and anthropogenic influences, such as flood control measures, may strongly influence single gauges, the results on flood-rich and flood-poor periods of a large collection of gauges for the same time period should be little prone to such influences. This assumption is supported by the notion that climate impact operates on a large scale contrary to local anthropogenic influence (Blöschl et al., 2007). Criterion (3), i.e. selection of larger catchments, is also meant to reduce the

impact of human interventions on the results. Based on studies by Ihringer (1996), Michaud et al. (2001), Bronstert et al. (2002) and Pfister et al. (2004), we assume that land use changes and flood retention basins play a minor role for larger catchments. The catchment area varies between 135 and 144232 km^2 with a median of 4151 km^2. 62 out of 68 catchments (92%) are larger than 1000 km^2. A few catchments with smaller area have been selected in regions where no time series of larger catchments are available.

To study clustering at different time scales, very long time series would be desirable (criterion 4). We have selected time series with at least seven decades of continuous measurements. In addition, we have selected four gauges with much longer time series to verify the results for longer periods: Dresden/Elbe (1852–2011), Cologne/Rhine (1817–2011), Wasserburg/Danube (1826–2009) and Vlotho/Weser (1823–2005).

Finally, time series with little data gaps (criterion 5) have been chosen. In total 68 gauges for the period from 1932 to 2005 have been selected. Seven gauges have missing values for one complete year, and five gauges have data gaps of more than one year. These gaps were filled by correlating the daily time series to the station with the highest correlation coefficient. The runoff time series have been obtained from different water authorities in Germany. Since the data are part of the official hydrometric observation network of the water authorities in Germany, the observations are regularly checked and can be assumed reliable. Fig. 1 shows the location of the selected gauges, and Table 1 lists their main characteristics.

Time series of flood occurrence are derived by selecting the largest independent flood events above certain discharge thresholds (POT). To ensure independence, i.e. floods originating from different rainfall events, the extraction is based on the USWRC (1976) rule (Lang et al., 1999). This rule considers catchment area controlling the concentration time and additionally imposes a criteria that the intermediate flow between two peaks must drop below 75% of the lower of two neighbouring peaks:

$$\{dt > 5 \text{ days} + \log(0.3861 * A)\} \; AND \; \{Q_{min} < 0.75 * \min(Q_i, Q_{i+1})\} \quad (1)$$

where dt is the minimum time [days] between two independent peaks, A is the catchment area [km^2], Q_{min} is the minimum intermediate flow between two neighbouring peaks Q_i and Q_{i+1} [m^3/s].

Five thresholds are selected: POT3 (on average 3 events/year), POT1 (on average 1 event/year), POT03 (on average 1 event/3-years), POT05 (on average 1 event/5-years) and POT010 (on average 1 event/10-years). Besides the annual time series (1 November – 31 October), also seasonal, half-year time series are derived to analyse possible differences in clustering in winter- and summer-dominated flood regimes (winter: 1 November – 30 April; summer: 1 May – 31 October).

3. Methods

We apply two objective methods that have been reported in the flood clustering literature for determining whether flood-rich and flood-poor periods exist. They are based on the dispersion index (e.g., Mediero et al., 2015) and the kernel occurrence rate estimation with non-parametric significance testing (e.g., Silva et al., 2012), respectively. Because they take a different approach for addressing the significance of clustering, the application of both methods helps to understand how robust the results are. Further, we introduce a third method, namely kernel occurrence rate estimation with parametric significance testing. This method reframes the significance test by comparing the time-varying occurrence rate to the constant confidence interval that would be obtained assuming a homogeneous Poisson process. The field significance is tested for the dispersion index via the false discovery rate. These statistical tests are described in the following sections.

3.1. Method 1: Index of dispersion

The occurrence of floods can be interpreted as a realization of a point process. A point process which occurs randomly in time is a homogeneous Poisson process, i.e. the event occurrence at any time point is independent of the event occurrences at any previous time point. The degree of event clustering and departure from a homogeneous Poisson process can be characterised by the index of dispersion (D). It relates the variability of flood event counts to the expectation value of the counts:

$$D = \frac{\mathbb{V}ar(Z(T))}{\mathbb{E}(Z(T))} - 1, \quad (2)$$

where Z(T) is the series of event (POT) counts within a time window of length T, $\mathbb{V}ar(Z(T))$ is the variance of the flood counts and $\mathbb{E}(Z(T))$ is the expected value.

For a homogeneous Poisson, the index of dispersion is equal to zero. Negative values of D stand for under-dispersion and characterise a more regular pattern of event occurrence than a homogeneous Poisson process. Clustering would be indicated by positive D values which stand for over-dispersion.

To test whether a time series shows over-dispersion at a given significance level, a Monte Carlo simulation is performed (Eastoe and Tawn, 2010; Raschke, 2015). For an observed sample z containing the time series of event counts, the null hypothesis 'D = 0' is tested against the alternative hypothesis 'D > 0'. A large number of samples (1000) of the same length as the observed time series is generated by drawing randomly from the Poisson distribution with the occurrence rate equal to the observed mean. For each synthetic time series of flood counts, the dispersion index is calculated. The sampling distribution of the dispersion index is constructed, and the critical value at the predefined significance level is derived. Significant over-dispersion can be claimed in case the observed dispersion index is larger than the critical value.

The calculation of the dispersion index requires selecting the aggregation period T, i.e. the time period in which flood occurrences are counted to produce the time series of counts (Z). This choice determines the time scale at which the method quantifies clustering. If flood-rich and/or flood-poor periods have a certain duration, then this is detected by the window size of this duration. For the 68 catchment data set with a record length of 74 years, aggregation periods of 1, 2, 3, 5 and 7 years were chosen. Although it would be most interesting to quantify clustering at longer time scales, we feel that the total time series length does not allow longer aggregation periods. For an aggregation period of 7 years and a time series length of 74 years, there are 10 values to calculate the dispersion index. This is a low number, and the results for a single time series are associated with significant sampling uncertainty. However, the results are mainly discussed for the complete data set of 68 catchments, which reduces the sampling uncertainty. For the four time series with more than 160 years of observations, also larger aggregation periods up to 10 years were selected.

3.2. Method 2: Kernel occurrence rate estimation and non-parametric significance testing

The time-varying intensity of a point process can be estimated by non-parametric methods (e.g. Ellis, 1986; Diggle and Marron, 1988). A non-parametric kernel implementation estimates the occurrence rate as:

Table 1
Main characteristics of selected gauges.

#	Gauge name	River	Basin	Cat. area [km^2]	Specific mean flow [m^3 s^{-1} km^{-2}]	Flood regime[a]	Location	
							Lat	Long
1	Achstetten	B. Rot	Danube	264	0.0125	C	48.26	9.9
2	Kempten	Iller	Danube	955	0.0491	C	47.73	10.32
3	Fuerstenfeldbruck	Amper	Danube	1235	0.0191	C	48.18	11.26
4	Beuron	Donau	Danube	1309	0.0087	B	48.05	8.97
5	Chamerau	Regen	Danube	1357	0.0192	B	49.18	12.75
6	Seebruck	Alz	Danube	1388	0.0373	C	47.94	12.48
7	Eichstaett	Altmuehl	Danube	1400	0.0071	B	48.88	11.2
8	Landsberg	Lech	Danube	2295	0.0358	C	48.04	10.88
9	Regenstauf	Regen	Danube	2658	0.0141	B	49.13	12.13
10	Muenchshofen	Naab	Danube	4014	0.0093	B	49.24	12.08
11	Heitzenhofen	Naab	Danube	5426	0.0090	B	49.12	11.94
12	Burghausen	Salzach	Danube	6649	0.0377	C	48.16	12.84
13	Landau	Isar	Danube	8467	0.0200	C	48.68	12.69
14	Plattling	Isar	Danube	8839	0.0198	C	48.77	12.88
15	Oberaudorf	Inn	Danube	9712	0.0314	C	47.65	12.2
16	Dillingen	Donau	Danube	11,315	0.0143	C	48.57	10.5
17	Wasserburg[b]	Inn	Danube	11,980	0.0296	C	48.06	12.23
18	Eschelbach	Inn	Danube	13,354	0.0278	C	48.26	12.73
19	Donauwoerth	Donau	Danube	15,037	0.0127	B	48.71	10.8
20	Ingolstadt	Donau	Danube	20,001	0.0157	C	48.76	11.42
21	Kelheim	Donau	Danube	22,950	0.0145	C	48.92	11.87
22	Passau-Ingling	Inn	Danube	26,084	0.0283	C	48.56	13.44
23	Oberndorf	Donau	Danube	26,446	0.0134	C	48.95	12.02
24	Hofkirchen	Donau	Danube	47,496	0.0136	C	48.68	13.12
25	Achleiten	Donau	Danube	76,653	0.0186	C	48.58	13.51
26	Wegeleben	Bode	Elbe	1215	0.0069	B	51.89	11.19
27	Greiz	W. Elster	Elbe	1255	0.0084	B	50.66	12.2
28	Lichtenwalde	Zschopau	Elbe	1575	0.0138	B	50.89	13.02
29	Wechselburg	Z. Mulde	Elbe	2107	0.0126	B	51.01	12.77
30	Hadmersleben	Bode	Elbe	2758	0.0051	B	52.01	11.32
31	Camburg-Stoeben	Saale	Elbe	3977	0.0079	B	51.07	11.7
32	Oldisleben	Unstrut	Elbe	4174	0.0045	B	51.3	11.18
33	Golzern	V. Mulde	Elbe	5442	0.0114	B	51.25	12.78
34	Calbe	Saale	Elbe	23,719	0.0048	B	51.92	11.81
35	Dresden[b]	Elbe	Elbe	53,096	0.0061	B	51.06	13.74
36	Barby	Elbe	Elbe	94,060	0.0059	B	51.99	11.88
37	Magdeburg	Elbe	Elbe	94,942	0.0059	B	52.14	11.65
38	Wittenberge	Elbe	Elbe	123,532	0.0047	B	52.99	11.75
39	Neu-Darchau	Elbe	Elbe	131,950	0.0056	B	53.23	10.89
40	Rheine	Ems	Ems	3740	0.0054	B	52.29	7.43
41	Hohensaaten-Finow	Oder	Oder	109,564	0.0096	B	52.87	14.14
42	Altensteig	Nagold	Rhein	135	0.0194	C	48.58	8.61
43	Bad-Imnau	Eyach	Rhein	331	0.0093	B	48.41	8.77
44	Pforzheim-Wuerm	Wuerm	Rhein	418	0.0070	B	48.87	8.71
45	Schwaibach	Kinzig	Rhein	954	0.0241	A	48.39	8.03
46	Doerzbach	Jagst	Rhein	1030	0.0098	B	49.37	9.72
47	Nuernberg	Pegnitz	Rhein	1192	0.0093	C	49.46	11.05
48	Tauberbischofsheim	Tauber	Rhein	1584	0.0055	B	49.63	9.67
49	Bad-Kissingen	F. Saale	Rhein	1587	0.0076	B	50.18	10.07
50	Kesseler3	Lippe	Rhein	2003	0.0118	B	51.66	8.09
51	Plochingen	Neckar	Rhein	3995	0.0120	B	48.71	9.42
52	Pettstadt	Pegnitz	Rhein	7005	0.0075	C	49.84	10.94
53	Neuhausen	Rhein	Rhein	11,887	0.0310	A	47.68	8.63
54	Schweinfurt	Main	Rhein	12,715	0.0082	B	50.03	10.22
55	Rekingen	Rhein	Rhein	14,718	0.0299	A	47.57	8.33
56	Cochem	Mosel	Rhein	27,088	0.0116	A	50.14	7.17
57	Maxau	Rhein	Rhein	50,196	0.0250	A	49.04	8.31
58	Kaub	Rhein	Rhein	103,488	0.0159	A	50.09	7.77
59	Cologne[b]	Rhein	Rhein	144,232	0.0146	A	50.94	6.96
60	Meiningen	Werra	Weser	1170	0.0121	A	50.58	10.42
61	Schmittlotheim	Eder	Weser	1202	0.0158	A	51.16	8.9
62	Gross-Schwuelper	Oker	Weser	1734	0.0066	B	52.35	10.43
63	Rotenburg	Fulda	Weser	2523	0.0086	A	51.01	9.72
64	Gerstungen	Werra	Weser	3039	0.0102	B	50.96	10.07
65	Celle	Aller	Weser	4128	0.0065	B	52.62	10.06
66	Guntershausen	Fulda	Weser	6366	0.0089	B	51.23	9.47
67	Hann.-Muenden	Weser	Weser	12,442	0.0090	B	51.43	9.64
68	Vlotho[b]	Weser	Weser	17,618	0.0094	B	52.18	8.86

[a] Flood regime according to Beurton and Thieken (2009): A: predominantly winter floods; B: winter, spring and autumn floods; C: distinct summer floods.
[b] Gauges with time series of at least 160 years, which were additionally analysed.

$$\hat{\lambda}(t) = h^{-1}\sum_{i=1}^{n} \mathrm{K}\left(\frac{t - T_{obs}(i)}{h}\right) \quad (3)$$

where h is the bandwidth, K is the kernel function, $T_{obs}(i)$ represents the event time i in total n (observed) events and $\hat{\lambda}(t)$ signifies the estimated occurrence rate at time t. Among several types of commonly used kernel functions (e.g., uniform, triangle, biweight, Epanechnikov and Gaussian), the Gaussian kernel is selected because it yields a smooth occurrence rate estimation and allows to calculate $\hat{\lambda}(t)$ efficiently.

The selection of the bandwidth h is an important step, as it determines the smoothness of the occurrence rate and its bias and variance properties. There are methods for selecting the bandwidth, e.g. the cross-validation bandwidth selector of Brooks and Marron (1991) or Silverman's rule of thumb (Silverman, 1986). However, Ramesh and Davison (2002) conclude that the statistical literature on the choice of the bandwidth is somewhat inconclusive. As we are interested in clustering for different time scales, we apply the kernel estimator for different bandwidth values. This complies with the procedure in the case of method 1 where the dispersion index is calculated for different aggregation windows representing different time scales. To understand whether clustering changes for different time scales is not only important for flood design, as discussed in the introduction, it can also help to explain the cause of clustering. For example, if we find clustering at the decadal time scale, this has to be caused by decadal climate variation and not by catchment memory effects which are working on shorter time scales. Applying a range of bandwidth values follows Ramesh and Davison (2002) who recommend using different values in exploratory analysis in order to give different insights into the data.

Another important issue in kernel estimation is the boundary bias or edge effect (Cowling and Hall, 1996; Gasser and Müller, 1979). The absence of data beyond the two ends of the time series causes underestimation of the occurrence rate close to the boundaries. This effect can be reduced by generating pseudodata which extend the time series at both ends. Commonly used methods for pseudodata generation are the reflection and multiple-point rules. The former is adopted in the present paper (for details see Mudelsee, 2010).

To test whether the temporal variation of the occurrence rate deviates significantly from the null hypothesis ('the constant occurrence rate of a homogeneous Poisson process'), confidence intervals around the occurrence rate are constructed. Given the series of time points T_{org} derived from a time series, for example POT data, the following non-parametric bootstrap procedure is used to obtain confidence intervals (Cowling et al., 1996; Mudelsee et al., 2004):

- Step 1: Augment the original series T_{org} with size N_{org} with pseudodata using the reflection rule to obtain the extended series T_{aug} with size of N_{aug}.
- Step 2: From T_{aug}, sample with replacement a series T_{sam} with the size of N_{aug}. Then estimate the occurrence rate $\hat{\lambda}^*(t), t \in W_T = [T_{org}(1), T_{org}(N_{org})]$ (Eq. (3)) using the selected bandwidth h where W_T denotes the time window of the original series.
- Step 3: Repeat Step 2 until a large number of estimated rates $\hat{\lambda}^*(t)$ is available.
- Step 4: Derive the confidence interval at the predefined level (e.g., 90%, 95%) using the estimated $\hat{\lambda}^*(t)$ in step 3 with lower bound $\hat{\lambda}_L(t)$ and upper bound $\hat{\lambda}_U(t)$.

In method 2, clustering is tested by comparing the confidence intervals of the kernel estimation against the time-constant mean occurrence rate assuming a homogeneous Poisson process which has on average r floods per year and a time window of W_T. At a particular time $t \in W_T$, if $(\hat{\lambda}_L(t) - r)(\hat{\lambda}_U(t) - r)$ is positive, the estimated occurrence rate is said to differ significantly from the time-constant rate and hence the clustering is significant.

As an example, Fig. 2 shows the result for POT1 and a bandwidth of 10 years for the gauge Wasserburg/Danube. The time-varying confidence interval (grey area) of the flood occurrence rate is compared to the constant occurrence rate (red line). In case there is at least one period where the confidence interval is above or below the red line, there is significant clustering. This is the case for the given example: there is a significant flood-poor period in the 1860s and a significant flood-rich period in the 1950s. There is another situation in Fig. 2, at the end of the time series, where the method detects a significant flood-poor period. Since it cannot be ruled out that this is influenced by the pseudodata generation, the boundaries of the time series within one bandwidth are not further interpreted.

3.3. Method 3: Kernel occurrence rate estimation and parametric significant testing

We develop a variant of method 2 by reframing the significance test: The time-varying occurrence rate is compared to the constant

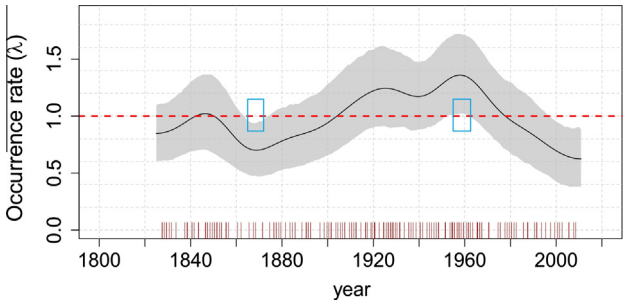

Fig. 2. Kernel occurrence rate estimates for POT1 and a bandwidth of 10 years for gauge Wasserburg/Danube. Method 2 (non-parametric significance test) compares the confidence interval (grey area) of the time-varying flood occurrence (black line) to the time-constant flood occurrence under the assumption of a homogeneous Poisson process (red line). Vertical bars denote time stamps of flood events. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

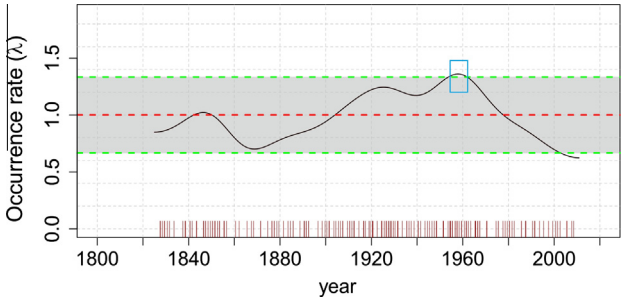

Fig. 3. Kernel occurrence rate estimates for POT1 and a bandwidth of 10 years for gauge Wasserburg/Danube. Method 3 (parametric significance test) compares the time-varying flood occurrence (black line) to the time-constant confidence interval under the assumption of a homogeneous Poisson process (green lines). Vertical bars denote time stamps of flood events. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

confidence interval that would be obtained assuming a homogeneous Poisson process with the occurrence rate λ. This is the same idea as the dispersion index based test. The observation is compared to the empirical distribution derived from the assumption of a homogeneous Poisson process. In this case, the inter-arrival time *IT* between event occurrences follows an exponential distribution with the parameter of $1/\lambda$. The confidence interval is constructed as follows:

- Step 1: Generate N samples $\{IT_i^*, i = 1 \dots N\}$ representing inter-arrival times using the exponential distribution exp($1/\lambda$) and construct the time series of flood occurrences $\{T_i^*, i = 1 \dots N\}$ by computing the cumulative sum of the inter-arrival times. N is the length of the observed time series.
- Step 2: Use Eq. (3) to estimate the occurrence rate $\hat{\lambda}^*(t), t \in W_T = [0, (N-1)\lambda]$ using the selected bandwidth h.
- Step 3: Repeat Steps 1 and 2 for a large number of realisations.
- Step 4: Derive two envelope curves which approximate the confidence interval (e.g., 90%, 95%) using the estimated $\hat{\lambda}^*(t)$ in Step 3.

The confidence bounds of the occurrence rate of a homogeneous Poisson process should be constant in time. However, edge effects may bias the estimation at both ends. Since the edge effects cannot be completely eliminated, the confidence bounds are estimated by averaging the curves obtained in Step 4 in the neighbourhood region W_r of the centre point c of W_T, where the estimation of the confidence level is considered to be robust. In practice, we chose $W_r = [c-h; c+h]$. The lower bound $\overline{\hat{\lambda}_L}$ and upper bound $\overline{\hat{\lambda}_U}$ of the confidence interval are hence derived. $\overline{\hat{\lambda}_L}$ and $\overline{\hat{\lambda}_U}$ depend on the occurrence rate of the homogenous Poisson process, the sample size N and the chosen bandwidth h. The significance of deviation from the homogenous Poisson process is assessed by comparing the time-varying estimate of the kernel occurrence rate $\lambda(t)$ of the observed data with the time-constant confidence interval $(\overline{\hat{\lambda}_L}, \overline{\hat{\lambda}_U})$. At a particular time $t \in W_T$, if $(\lambda(t) - \overline{\hat{\lambda}_L})(\lambda(t) - \overline{\hat{\lambda}_U})$ is positive, the estimated occurrence rate is said to differ significantly from the time-constant rate and hence the presence of a flood-rich or flood-poor period in the series can be considered as significant.

Fig. 3 illustrates how method 3 defines significant clustering. The time-varying flood occurrence rate (black line) is compared to the time-constant confidence interval (area between green lines). In case there is at least one period where the occurrence rate is above or below the confidence interval, clustering is significant. Method 3 identifies one significant situation, namely the flood-rich period in the 1950s, whereas method 2 additionally classifies the 1860s as flood-poor (see Fig. 2). This example already indicates that method 2 tends to classify a higher number of situations which are significantly above or below the expectation for a homogeneous Poisson process.

3.4. Field significance test

While local tests are used to derive significance statements for a single time series or a single location, field significance tests check the probability of obtaining a certain number of local significant results just by chance. We apply the false discovery rate (FDR) approach developed by Benjamini and Hochberg (1995) which has been shown robust in previous studies (Ventura et al., 2004; Wilks, 2006; Renard et al., 2008; Khaliq et al., 2009).

FDR controls the expected proportion of rejected local null hypotheses that are actually true. m local significance tests result in m corresponding p-values $\{p_i, i = 1 \dots m\}$ at the local significance level α_L. The field significance test is formulated based on a global significance level α_G which is the rate we are willing to allow of false rejections out of all rejections of local null hypothesis (H_o). α_L and α_G values are typically set to 5%, but they need not necessarily be equal. Assuming that the local tests are independent, the number of false detections can be controlled at the rate α_G by rejecting those local tests for which $p_{(i)}$ (denoting the i^{th} smallest of m p-values) is no greater than:

$$p_{FDR} = \max\left\{p_{(i)} : p_{(i)} \leqslant \frac{i\alpha_G}{m}, i = 1 \dots m\right\} \quad (4)$$

Field significance is claimed if at least one local null hypothesis is rejected. One advantageous feature of FDR is its robustness to spatial correlation, which is usually found for climatic and hydrological variables. The FDR procedure works with any statistical test based on p-values. Hence, FDR is suitable for testing field significance of clustering based on method 1 but not for methods 2 and 3, for which no p-values are defined.

4. Results

The results for the four long flood time series of gauges Dresden, Cologne, Wasserburg and Vlotho are presented in Section 4.1. These results are used to illustrate the differences between the three methods. The results for the 68 gauges data set for the common period 1932–2005 are given in Section 4.2. All significance tests, i.e. local tests and the field significance test, use a significance level of 5%. For each gauge, it is analysed whether clustering at the 5% significance level exists (1) for different POT thresholds (POT3, POT1, POT03, POT05), and (2) for different time scales (five aggregation periods for method 1 and four bandwidth values for methods 2 and 3).

First, the difference in the meaning of the parameters determining the time scale has to be stressed. For the index of dispersion (method 1), the aggregation period is equal to the time scale at which the clustering behaviour is characterised. In the case of the kernel occurrence rate estimator (methods 2 and 3), the bandwidth is related but not exactly equal to the time scale of clustering since the kernel estimator smooths the time series over a range larger than the bandwidth. This has to be considered when comparing the results of the different methods.

Another distinction between the dispersion index method and the kernel occurrence rate based methods is the fundamentally different approach for defining significant clustering. The dispersion index is a global metric, i.e. one number for the complete time series, whereas the kernel methods are based on a local view. They define clustering as significant if there is at least one flood-rich or flood-poor period in the time series which deviates sufficiently from the average expected occurrence rate. An advantage of the kernel based methods is that they identify the specific flood-rich and flood-poor periods – an information which is not directly obtained by the dispersion index method.

4.1. Temporal clustering in the four long flood time series

When calculating the dispersion index, there are T possibilities to select the starting point for the aggregation of event counts for an aggregation window of size T. Our analysis reveals that the dispersion index is sensitive to the starting point. This fact has not been addressed in the hydro-climatological literature so far. Hence,

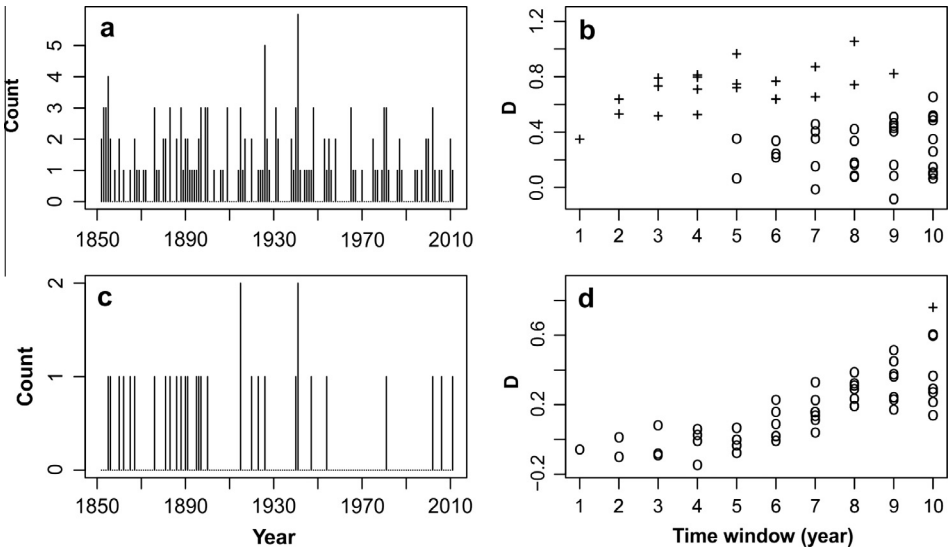

Fig. 4. Clustering analysis for POT1 (a, b) and POT05 (c, d) for the gauge Dresden/Elbe using the index of dispersion (method 1). Plots (a) and (c) show the time series of flood counts. Plots (b) and (d) show the index of dispersion as function of time scale. For each time scale T, the index is calculated T times. Significant clustering is denoted by '+'.

we calculate the dispersion index T times, shifting the start of the aggregation time window by one year. Fig. 4 exemplarily shows time series of flood counts for the gauge Dresden/Elbe for POT1 and POT05. In addition, the influence of the time scale on the dispersion index and its sensitivity to the start year of the aggregation is shown. In case of POT1, there is significant clustering for the time scales from 1 to 4 years, independent of the start year. Starting from the aggregation period of 5 years, an increasing fraction of non-significant values is found. At the time scale of 10 years, all 10 results are non-significant. A different picture arises for the POT05 series. With one exception (1 out of 10 results for the aggregation period of 10 years), all results are non-significant. In those cases where significant and non-significant results are obtained for a certain aggregation window, the overall significance for this aggregation window is calculated as the fraction of significant results. For example, the overall significance for POT05 and the aggregation window of 10 years is equal to 1/10.

Fig. 5 compares the results of the clustering analysis based on the kernel occurrence rate for selected thresholds and bandwidth values for the gauges Cologne/Rhine and Vlotho/Weser. Notably, the width of the confidence intervals depends on the bandwidth. Increasing the bandwidth leads to smaller confidence intervals. This is expected as the amplitude of the time-varying occurrence rate decreases with larger bandwidth. Further, the confidence interval is larger (in relative terms) for higher thresholds. This is explained by the higher sampling uncertainty for higher thresholds.

Fig. 6 summarizes the results for the four gauges for different time scales and POT thresholds. The results for methods 1 are rather diverse between the four gauges. Whereas at Cologne and Dresden stronger clustering is observed for lower thresholds, Vlotho shows the opposite result, i.e. significant clustering at higher thresholds but not at lower thresholds. No clustering is detected at gauge Wasserburg.

Compared to method 1, the kernel based methods 2 and 3 indicate more easily significant clustering. Clustering is found for almost all thresholds and bandwidth values. This difference between the methods is explained by their different definition of clustering, i.e. global view of method 1 versus local view of methods 2 and 3. In case of methods 2 and 3, clustering is identified if there is at least one flood-rich or flood-poor period which deviates significantly from the mean occurrence rate. Such an anomalous period is found in almost all of the 160-year long time series. Comparing methods 2 and 3 shows that method 3 is more conservative in identifying clustering, in particular concerning flood-poor periods. This reflects the fact that the width of the confidence bounds in method 2 is affected by the number of event counts in a certain period. In periods with below-average occurrence of floods, the confidence interval gets narrow, and hence, method 2 seems to be biased towards detecting flood-poor periods.

In terms of changes in clustering with threshold or time scale, the results are not unanimous. Both kernel based methods show a decreasing number of significant flood-rich/flood-poor periods with increasing time scale. In some cases this even leads to no anomalous period within the 160-year time series, and hence, no significant clustering. This effect of decreasing clustering with increasing time scale is not visible in the results of method 1. Comparing the results for different thresholds does not yield a clear pattern.

4.2. Temporal clustering in the 68-gauges data set

The three methods are applied to 68 gauges with time series from 1932 to 2005 considering different thresholds and aggregation window sizes or kernel bandwidth values, respectively. These results are summarised in Fig. 7.

All three methods indicate a significant deviation from the homogeneous Poisson process for many gauges in Germany. The fraction of gauges with significant over-dispersion or presence of flood-rich and flood-poor periods varies with the selected threshold, time scale and method. Method 2 detects the highest share of clustering. This effect is explained by the tendency of method 2 to detect many flood-poor periods due to the narrow confidence interval for periods with a low number of event occurrences. The dispersion index method gives the smallest number for the fraction of gauges with significant clustering. This is a consequence of the different approach of defining significance. The dispersion index method delivers a global value which considers the complete time series. In contrast, the kernel occurrence based methods signal significant clustering when there is at least one period in the time series with significant above- or below-average occurrence.

The field significance test for method 1 indicates that clustering is field-significant for almost all time scales and thresholds (Fig. 8).

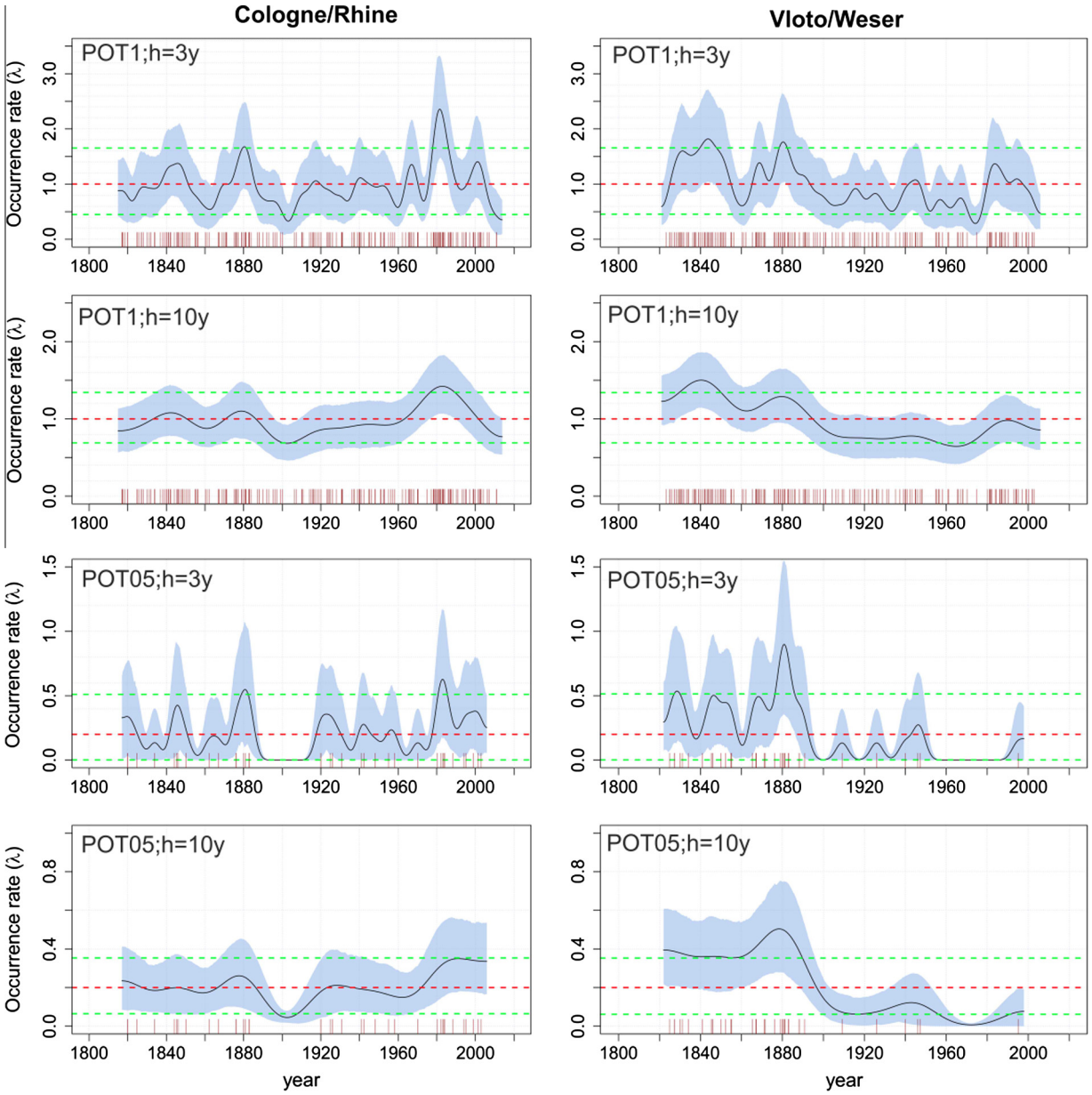

Fig. 5. Kernel occurrence rate estimates for Cologne/Rhine (left) and Vlotho/Weser (right) for different thresholds and bandwidth values; from top to bottom: POT1 – 3 years; POT1 – 10 years; POT05 – 3 years; POT05 – 10 years. Black line: time-varying flood occurrence; light blue area: 95% confidence interval (method 2: non-parametric significance test); red dotted line: time-constant flood occurrence under the assumption of homogeneous Poisson process; green dotted lines: 95% confidence interval (method 3: parametric significance test). Vertical bars denote time stamps of flood events. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Only for a few, typically higher thresholds the fraction of field-significant cases is below 100%.

The results of the dispersion index method (top panel in Fig. 7) show a strong decrease in clustering with increasing threshold. Whereas significant clustering is found for the annual time series in 53–78% of the gauges for POT3, this number drops below 10% for POT05. A similar tendency is observed for method 3, but the reduction of clustering with increasing threshold is much smaller. Method 2 shows only a relatively small decrease. Despite these differences in the quantitative results, which could again be explained by the different approaches of the methods, the overall pattern is the same: Floods above higher thresholds seem to cluster less.

In terms of changes in temporal clustering with time scale, there is again a consistent pattern across the three methods. Increasing the time scale decreases the fraction of gauges with significant clustering. For method 1, the fraction of significant gauges averaged over all thresholds for the annual time series decreases from 37% to 26% when the aggregation time window is increased from one year to seven years. Methods 2 and 3 show the same tendency, but the decline seems to be stronger. However, we should consider that these numbers are not directly comparable due to the different meaning of aggregation period and bandwidth and the fact that the longest time scale for methods 2 and 3 is somewhat larger than the longest aggregation period. When increasing the bandwidth from one to ten years, the fraction of gauges with significant clustering in annual time series drops from 100% to 26% for method 2 and from 92% to 10% for method 3. These results can be interpreted in a way that there is a higher chance of having

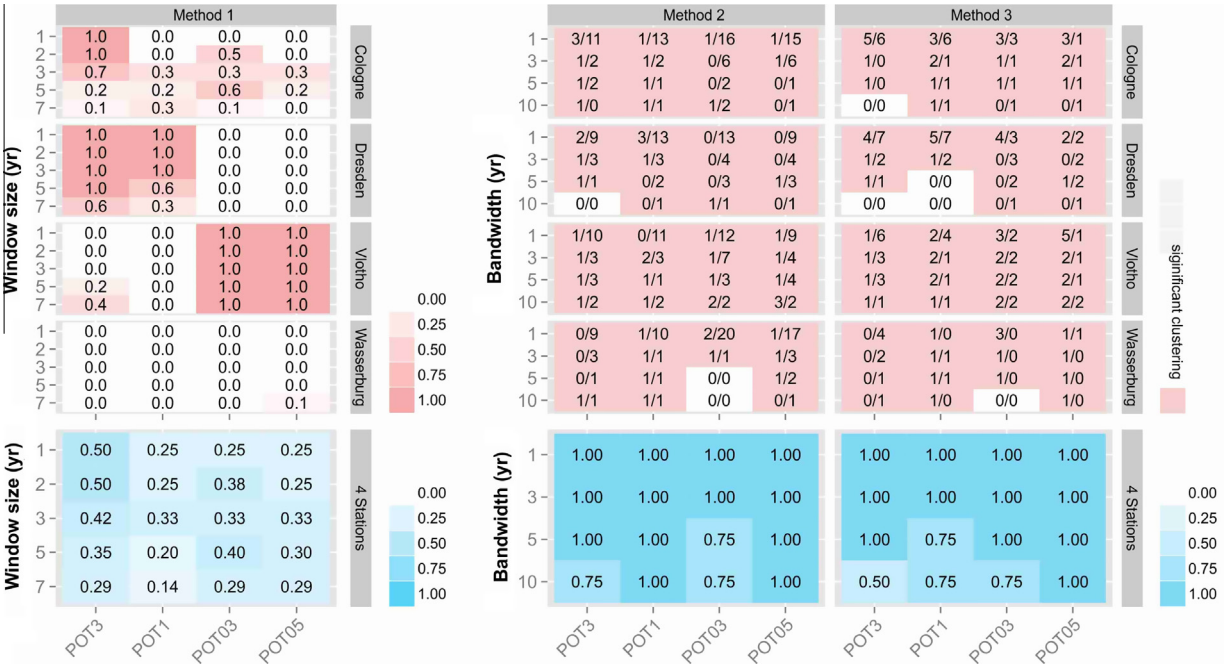

Fig. 6. Clustering results for the long time series for different POT thresholds and time scales. Left column: Method 1: Fraction of significant clustering for each gauge considering the shifting of the aggregation time window (upper pink panel), and fraction of cases with significant clustering (lower blue panel). Middle column: Method 2: Number of flood-rich/flood-poor periods detected (upper pink panel) and fraction of cases with significant clustering (lower blue panel). Right column: Method 3: Number of flood-rich/flood-poor periods detected (upper pink panel) and fraction of cases with significant clustering (lower blue panel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

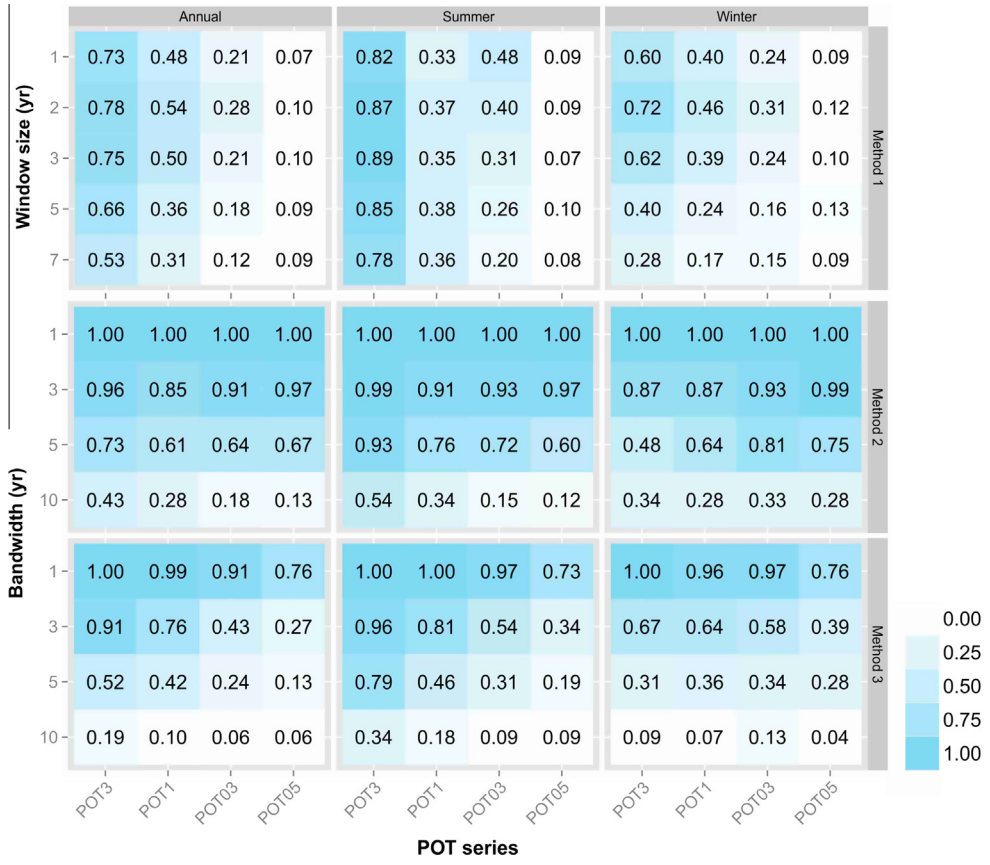

Fig. 7. Fraction of gauges with significant clustering at the 5% significance level for different POT thresholds and time scales for annual (left), summer (middle) and winter (right) data. Top: Method 1 (the sensitivity of the method to the starting year of the aggregation window is accounted for by calculating fractions of significant results); Middle: Method 2; Bottom: Method 3.

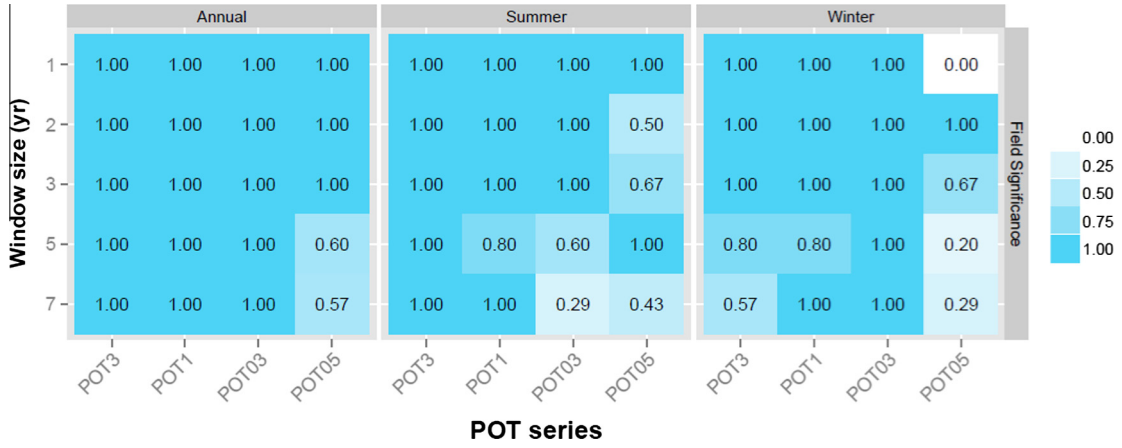

Fig. 8. Results for the field significance test at the 5% (local and global) significance level for method 1 (dispersion index) for different POT thresholds and aggregation window sizes. The numbers give the fraction of field-significant clustering, taking into account the variation of the results with shifting the starting year of the aggregation window.

flood clusters at a time scale of a few years rather than flood-rich and flood-poor decades.

Comparing the results for the annual time series with the seasonal results yields no clear differences. The overall pattern, i.e. decreasing clustering with increasing severity and time scale, and the order of magnitude of the results are the same for annual, winter and summer flood occurrences.

4.3. Spatial patterns of temporal clustering in Germany

Fig. 9 summarizes the spatial patterns of significant flood clustering across Germany for annual flood data and for different thresholds and time scales. We only show the results for method 3, however, the overall patterns are very similar for the three methods. As evident from the results in Section 4.2, temporal clustering decreases with increasing thresholds and increasing time scales. The clustering fades out stronger in the western parts of Germany (Rhine catchment) compared to the central and southern regions (Fig. 9). Particularly, in the central-eastern regions (Elbe catchment) and southern parts (Danube catchment), over-dispersion is persistent also at larger time scales (10 years). In these regions, particularly larger floods (POT03, POT05) seem to cluster at larger time scales.

Fig. 9 illustrates that a link between the flood regimes and the clustering behaviour is not obvious. However, at large (semi- or decadal) time scales the floods with regime "C" tend to be more clustered compared to the other two regimes while significant clustering of floods with regime "A" is almost absent.

5. Discussion

The application of the three methods to the four 160-year time series demonstrates that their results can diverge considerably. These deviations can be explained by their different approaches for indicating significant clustering. These examples also demonstrate the sensitivity of method 1 to the starting year of the aggregation window. Hence, we recommend that studies on temporal clustering do not rely on one method only.

The four very long time series do not allow answering the question whether clustering changes with time scale or flood severity. Hence, we quantify the temporal clustering for a large data set with 68 gauges. This, however, strongly decreases the length of our time series, and implies that our analysis is rather limited in terms of time scale and severity. For example, the question whether there is a consistent pattern in clustering at the multi-decadal time scale cannot be investigated with observations of 74 years length. Our highest severity threshold is POT05, i.e. one flood each five years on average. Hence, even our highest threshold contains floods which are not really extreme. These (data-related) limitations of our study need to be taken into account in the interpretation of the results, especially when comparing them to the analysis of much longer paleo and historic flood records.

The results of the three methods for the large data set with 68 gauges data suggest that temporal clustering of floods in Germany decreases with increasing flood severity and increasing time scale. This pattern is consistent for the three methods, despite their distinct approaches and differences in quantitative results. How can this decrease be explained? Temporal clustering of floods could be caused by climate and catchment memory effects. Inter-annual, inter-decadal and lower-frequency climate variability might modulate the occurrence and magnitude of floods (as discussed in the introduction). On the other hand, the probability and severity of floods might be larger for a wet catchment. Hence, the persistence of wet/dry catchment states over weeks, months or years might contribute to clustering at intra- to inter-annual time scales. This persistence is related to the hydrological characteristics of the catchment. Catchments with deep soils and/or saprolite zones, and mighty aquifers tend to have a higher catchment state persistence due to stronger memory effects. For example, Creutzfeldt et al. (2012) quantify the water storage changes by super-conducting gravimetry for a catchment with a very deep saprolite zone in south Germany. They show that catchment water storage is strongly depleted by the drought in 2003, and that this depletion is only recovered after several years. They argue that such long-term recovery of water storages may also exist at other locations or at larger scales, but may not be observed due to the lack of an appropriate monitoring technique.

Without further in-depth analysis, it is not possible to decipher the role of climate effects and catchment memory effects in relation to the decreasing clustering with increasing severity and time scale. One hypothesis is that, for short time scales, catchment memory effects contribute to the strong clustering along with possible (intra- to inter-annual) climatic effects. In case the catchment is in a wet state and the different storages are filled, a comparably small precipitation event might trigger a comparatively large flood. The probability of having smaller precipitation events is higher leading to higher clustering of smaller floods compared to stronger ones. Particularly for smaller floods the initial catchment wetness seems to play a stronger role than for severe floods (Merz and Plate, 1997). Although, the definition of independent POT values (Eq. (1)) reduces the effect of catchment storage and attempts to

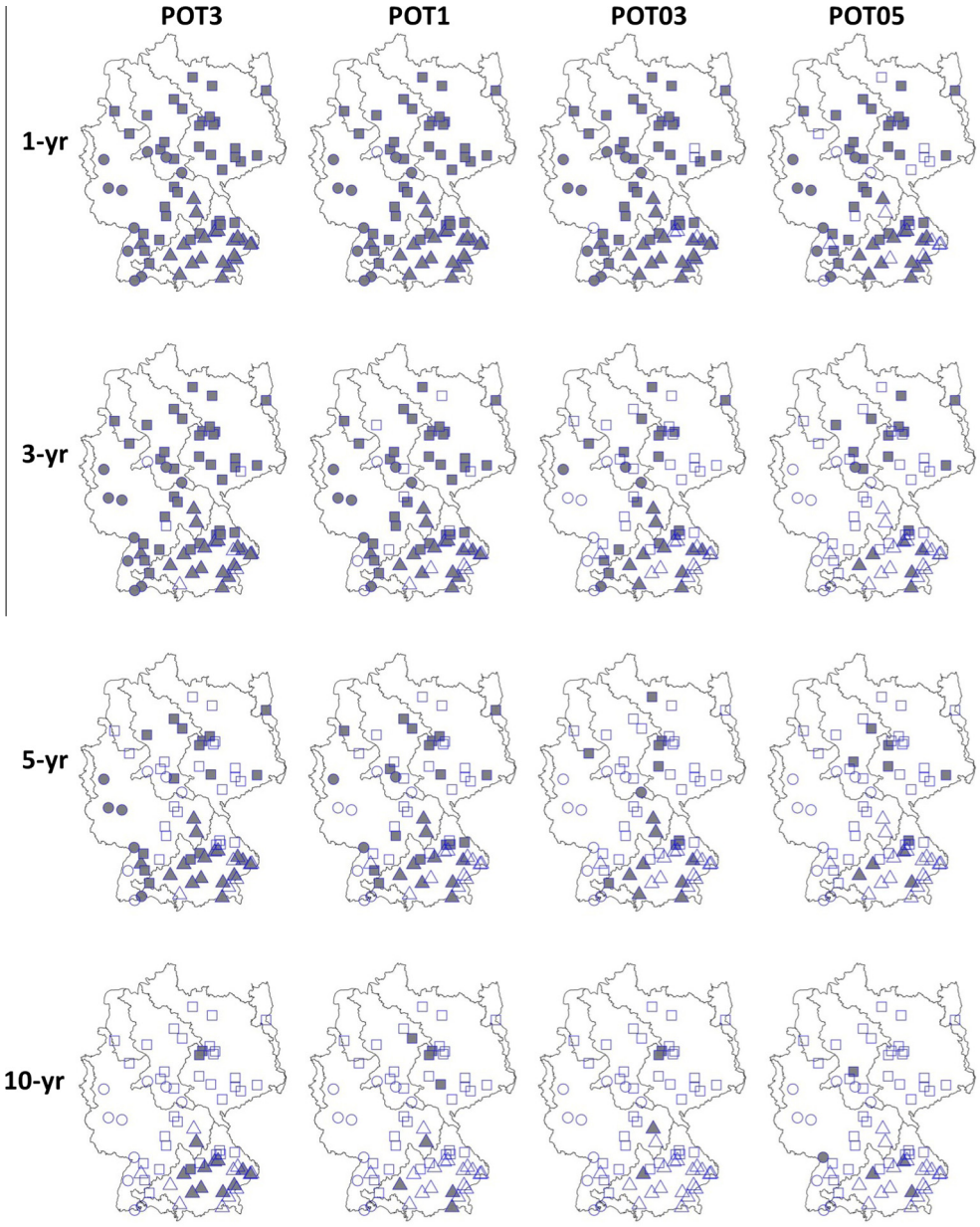

Fig. 9. Gauges with significant clustering in annual time series detected by method 3. Aggregation windows (1-year, 3-year, 5-year, 10-year) increase from top panel to bottom panel. POT thresholds (POT3, POT1, POT03, POT05) increase from left to right. Circles, squares and triangles denote the flood regime "A", "B", "C" (see Table 1), respectively. Filled markers indicate significant clustering at the 5% significance level, empty markers stand for non-significant clustering.

separate flood peaks from the neighbouring storm events, the long-term catchment memory is not fully eliminated from the POT series and seems to be reflected in the clustering behaviour.

This argument could also explain the decrease of clustering with severity. If catchment memory effects play an important role for generating temporal clustering, then we expect to see this effect in particular for POT time series with lower thresholds. The lower the threshold, the smaller are the inter-event arrival times, and the bigger the role of catchment memory effects. Further, smaller floods tend to be stronger influenced by the catchment conditions, and hence by catchment memory effects. Hence, this argument links the decrease in clustering for increasing time scale with the decrease clustering for increasing severity. This hypothesis is supported by the comparatively small evidence in the literature on strong linkages between flood occurrence in Germany and large-scale low-frequency climate variability. Although studies have found some relation of flooding in Germany to climate variability, these relations are typically weak (e.g. Caspary, 1995; Mudelsee et al., 2004; Petrow et al., 2007, 2009).

Our result does not support the findings of studies that reconstruct flood time series from historical sources. They typically conclude that flood-rich and flood-poor periods, at decadal time scales, are an important phenomenon for Germany and adjacent regions. One reason for this discrepancy could be that historical flood time series tend to look at more extreme floods than in our analysis with limited observation length. Based on our data set and analysis, we cannot rule out the existence of clustering for extreme floods in Germany. The statistical tests that we apply in combination with the rather short time series of 74 years might just not allow to detect clustering in extreme events, possibly on longer time scales. It could be the case that extreme floods are governed by large-scale, low-frequency climate variability and small floods by catchment memory effects, leading to clustering of both small and large floods but due to different reasons. To understand the role of

climate, process-based analyses seem to be necessary that investigate how and to which extent flood characteristics are governed by low-frequency changes in atmospheric moisture uptake, transport and deposition.

Our results on temporal clustering in the summer and winter seasons (Fig. 7) and the suggestion that clustering fades out more strongly in western Germany (Fig. 9) also require further in-depth analyses. The different clustering behaviour in western, central-eastern and southern Germany could correspond to the regions of different flood seasonality identified by Beurton and Thieken (2009). These regions are parts of larger homogenous regions in terms of flood behaviour (Atlantic, Continental and Alpine) covering a wider European domain with distinct clustering behaviour on the inter-annual scale as identified by Mediero et al. (2015). However, understanding the links between climate variability, catchment characteristics and temporal clustering of floods would require analysing flood generation processes which is beyond the scope of this study.

6. Conclusions

An analysis of clustering behaviour of floods in Germany was presented based on peak-over-threshold time series derived from the daily flow series at 68 gauges for the common period 1932–2005. The gauges cover a wide range of catchments located in three regions with distinct flood seasonality. In addition to the analysis for the common 74-years period, the clustering behaviour was investigated at four gauges with at least 160 years of record. Two methods for the analysis of clustering based on the index of dispersion and on kernel occurrence rate estimation were applied which detect the deviation from a homogenous Poisson process. For the latter, a non-parametric significance test was implemented. Additionally, a third method was proposed which combines the kernel occurrence rate estimation with parametric test. Whereas the index of dispersion takes a global view on the time series, the kernel-based methods identify flood-rich and flood-poor periods for which the occurrence rate significantly deviates from the expected value. We explored whether the significance of clustering changed with flood severity and time scale. Although such changes, if existent, would be important for flood design and risk management, this question had not been investigated in the flood literature, to the best of our knowledge.

In methodological terms, our analyses suggest:

1. The significance of clustering derived by the index of dispersion method is sensitive to the selection of the starting time point of the aggregation window. To the best of our knowledge, this sensitivity has not been reported in the hydro-meteorological literature. We, thus, suggest shifting the starting point and averaging the test outcomes in order to deliver a more robust result.
2. The three methods result in similar patterns of clustering behaviour. Also, the spatial patterns and seasonally stratified results share similarities. However, the specific results with respect to the share of gauges with significant clustering differ considerably. Hence, we propose to apply more than one method when analysing flood clustering. In particular, we recommend using complementary methods, such as the dispersion index based test to obtain a global view on clustering and kernel occurrence rate based methods, with a local view, to identify specific flood-poor and flood-rich periods.
3. The kernel occurrence rate estimation with non-parametric significance testing is sensitive to the event frequency and edge effects. The width of the confidence interval is smaller for periods with a lower number of events. This may lead to a much higher number of flood-poor periods compared to flood-rich periods within a time series. We, thus, suggest using the parametric significance test for the kernel-based occurrence rate estimation which we developed in this paper.

In terms of hydrological processes, our analyses suggest:

4. All methods suggest the presence of clustering in flood time series at a high fraction of gauges in Germany.
5. The number of gauges exhibiting significant clustering is decreasing with increasing flood severity threshold and time scale. We thus found no evidence of pronounced clustering for larger floods and for larger time scales. These results are however conditioned by the limited length of the available time series and the lower density of severe floods.
6. Spatial patterns of flood clustering behaviour in Germany are not strongly pronounced but seem to correspond to different regions of flood seasonality defined by Beurton and Thieken (2009) and to be part of larger European patterns (Mediero et al., 2015), where Germany sits at the joint of three large homogenous flood regions.
7. Clustering is very pronounced for smaller flood severities and smaller time scales. We hypothesize that this result could be explained by catchment memory effects along with intra-annual to inter-annual climate variability, where under generally wetter catchment conditions a relatively small precipitation event may lead to a flood above a selected threshold.

Our results do not support the findings of historical flood studies in Central Europe on the importance of flood clustering at the decadal time scale. Process-based analyses are necessary to investigate how and to which extent flood characteristics are governed by low-frequency changes in atmospheric moisture uptake, transport and deposition, and to explain this discrepancy.

Acknowledgements

We thank the Federal Agency for Cartography and Geodesy in Germany (BKG) for provision of the digital elevation model of Germany. We are grateful for providing the discharge data: Bavarian State Office of Environment (LfU), Baden-Württemberg Office of Environment, Measurements and Environmental Protection (LUBW), Brandenburg Office of Environment, Health and Consumer Protection (LUGV), Saxony State Office of Environment, Agriculture and Geology (SMUL), Saxony-Anhalt Office of Flood Protection and Water Management (LHW), Thüringen State Office of Environment and Geology (TLUG), Hessian Agency for the Environment and Geology (HLUG), Rhineland Palatinate Office of Environment, Water Management and the Factory Inspectorate (LUWG), Saarland Ministry for Environment and Consumer Protection (MUV), Office for Nature, Environment and Consumer Protection North Rhine-Westphalia (LANUV NRW), Lower Saxony Office for Water Management, Coast Protection and Nature Protection (NLWKN), Water and Shipping Management of the Fed. Rep. (WSV), prepared by the Federal Institute for Hydrology (BfG).

References

Barnston, A.G., Livezey, R.E., 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Mon. Weather Rev. 115, 1083–1126.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. 57B, 289–300.

Beurton, S., Thieken, A., 2009. Seasonality of floods in Germany. Hydrol. Sci. J. 54, 62–76. http://dx.doi.org/10.1623/hysj.54.1.62.

Blöschl, G., Ardoin-Bardin, S., Bonell, M., Dorninger, M., Goodrich, D., Gutknecht, D., Matamoros, D., Merz, B., Shand, P., Szolgay, J., 2007. At what scales do climate variability and land cover change impact on flooding and low flows? Invited

commentary. Hydrol. Process. 21, 1241–1247. http://dx.doi.org/10.1002/hyp.6669.
Bouwer, L.M., Vermaat, J.E., Aerts, J.C.J.H., 2006. Winter atmospheric circulation and river discharge in northwest Europe. Geophys. Res. Lett. 33, L06403. http://dx.doi.org/10.1029/2005GL025548.
Brázdil, R., Pfister, C., Wanner, H., Storch, H.V., Luterbacher, J., 2005. Historical climatology in Europe: the state of the art. Climatic Change 70, 363–430.
Brázdil, R., Chromá, K., Dobrovolný, P., Cernoch, Z., 2012. The tornado history of the Czech Lands, AD 1119–2010. Atmos. Res. 118, 193–204.
Bronstert, A., Niehoff, D., Bürger, G., 2002. Effects of climate and land-use change on storm runoff generation: present knowledge and modelling capabilities. J. Hydrol. 16, 509–529. http://dx.doi.org/10.1002/hyp.326.
Brooks, M.M., Marron, J.S., 1991. Asymptotic optimality of the least-squares crossvalidation bandwidth for kernel estimates of intensity functions. Stoch. Process. Appl. 38 (1), 157–165.
Caspary, H.J., 1995. Recent winter floods in Germany caused by changes in the atmospheric circulation across Europe. Phys. Chem. Earth 20, 459–462.
Cayan, D.R., Redmond, K.T., Riddle, L.G., 1999. ENSO and hydrological extremes in the western United States. J. Clim. 12, 2881–2893.
Cowling, A., Hall, P., 1996. On pseudodata methods for removing boundary effects in kernel density estimation. J. Roy. Stat. Soc. B 58 (3), 551–563.
Cowling, A., Hall, P., Phillips, M.J., 1996. Bootstrap confidence regions for the intensity of a Poisson point process. J. Am. Stat. Assoc. 91 (436), 1516–1524.
Creutzfeldt, B., Ferré, T., Troch, P., Merz, B., Wziontek, H., Güntner, A., 2012. Total water storage dynamics in response to climate variability and extremes: inference from long-term terrestrial gravity measurement. J. Geophys. Res. 117, D08112.
Delgado, J.M., Merz, B., Apel, H., 2012. A climate-flood link for the lower Mekong River. Hydrol. Earth Syst. Sci. 16, 1533–1541. http://dx.doi.org/10.5194/hess-16-1533-2012.
Diggle, P., Marron, J.S., 1988. Equivalence of smoothing parameter selectors in density and intensity estimation. J. Am. Stat. Assoc. 83 (403), 793–800.
Eastoe, E.F., Tawn, J.A., 2010. Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series. Water Resour. Res. 46, W02510. http://dx.doi.org/10.1029/2009WR007757.
Ellis, S.P., 1986. A limit theorem for spatial point processes. Adv. Appl. Prob. 18, 646–659.
Gasser, T., Müller, H.G., 1979. Kernel estimation of regression functions. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation. Springer, Berlin, pp. 23–68.
Glaser, R., Riemann, D., Schönbein, J., Barriendos, M., Brázdil, R., Bertolin, C., Camuffo, D., Deutsch, M., Dobrovolný, P., van Engelen, A., Enzi, S., Halícková, M., Koenig, S.J., Kotyza, O., Limanówka, D., Macková, J., Sghedoni, M., Martin, B., Himmelsbach, I., 2010. The variability of European floods since AD1500. Climatic Change 101, 235–256.
Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., Kjeldsen, T.R., Kriaučiūnienė, J., Kundzewicz, Z.W., Lang, M., Llasat, M.C., Macdonald, N., McIntyre, N., Mediero, L., Merz, B., Merz, R., Molnar, P., Montanari, A., Neuhold, C., Parajka, J., Perdigão, R.A.P., Plavcová, L., Rogger, M., Salinas, J.L., Sauquet, E., Schär, C., Szolgay, J., Viglione, A., Blöschl, G., 2014. Understanding flood regime changes in Europe: a state-of-the-art assessment. Hydrol. Earth Syst. Sci. 18, 2735–2772.
Helms, M., Büchele, B., Merkel, U., Ihringer, J., 2002. Statistical analysis of the flood situation and assessment of the impact of diking measures along the Elbe (Labe) river. J. Hydrol. 267, 94–114.
Hirschboeck, K.K., 1988. Flood hydroclimatology. In: Baker, V.R., Kockel, R.C., Patton, P.C. (Eds.), Flood Geomorphology. Wiley, NY.
Ihringer, J., 1996. Hochwasser aus ländlichen und städtischen Gebieten. Geowissenschaften 14 (12), 223–530.
Jain, S., Lall, U., 2000. Magnitude and timing of annual maximum floods: trends and large-scale climatic associations for the Blacksmith Fork River, Utah. Water Resour. Res. 36 (12), 3641–3651.
Jain, S., Lall, U., 2001. Floods in a changing climate: does the past represent the future? Water Resour. Res. 37, 3193–3205.
Jacobeit, J., Glaser, R., Luterbacher, J., Wanner, H., 2003. Links between flood events in central Europe since AD 1500 and large-scale atmospheric circulation modes. Geophys. Res. Lett. 30 (4), 21-1–21-4.
Khaliq, M.N., Ouards, T.B.M.J., Gachon, P., Sushama, L., St-Hilaire, A., 2009. Identification of hydrological trends in the presence of serial and cross correlations: a review of selected methods and their application to annual flow regimes of Canadian rivers. J. Hydrol. 368, 117–130. http://dx.doi.org/10.1016/j.jhydrol.2009.01.035.
Khare, S., Bonazzi, A., Mitas, C., Jewson, S., 2015. Modelling clustering of natural hazard phenomena and the effect on re/insurance loss perspectives. Nat. Hazards Earth Syst. Sci. 15, 1357–1370. http://dx.doi.org/10.5194/nhess-15-1357-2015.
Kiem, A.S., Franks, S.W., Kuczera, G., 2003. Multi-decadal variability of flood risk. Geophys. Res. Lett. 30, 1035. http://dx.doi.org/10.1029/2002GL015992.
Kingston, D.G., Lawler, D.M., McGregor, G.R., 2006. Linkages between atmospheric circulation, climate and streamflow in the northern North Atlantic: research prospects. Prog. Phys. Geogr. 30, 143–174.
Koutsoyiannis, D., 2005. Uncertainty, entropy, scaling and hydrological stochastics. 2. Time dependence of hydrological processes and time scaling. Hydrol. Sci. J. 50, 405–426.
Kundzewicz, Z.W. (Ed.), 2012. Changes in Flood Risk in Europe. IAHS Special Publication 10, ISBN 978-1-907161-28-5, p. 516.
Lammersen, R., Engel, H., van de Langemheen, W., Buiteveld, H., 2002. Impact of river training and retention measures on flood peaks along the Rhine. J. Hydrol. 267, 115–124.
Lang, M., Ouarda, T.B.M.J., Bobée, B., 1999. Towards operational guidelines for over-threshold modelling. J. Hydrol. 225 (1999), 103–117.
Lin, Z., Levy, J.K., Xu, X., Zhao, S., Hartmann, J., 2005. Weather and seasonal climate prediction for flood planning in the Yangtze River Basin. Stoch. Environ. Res Risk. Assess. 19, 428–437. http://dx.doi.org/10.1007/s00477-005-0007-4.
Mailier, P.J., Stephenson, D.B., Ferro, C.A.T., Hodges, K.I., 2006. Serial clustering of extratropical cyclones. Mon. Weather Rev. 134, 2224–2240.
Mediero, L., Kjeldsen, T.R., Macdonald, N., Kohnova, S., Merz, B., Vorogushyn, S., Wilson, D., Alburquerque, T., Blöschl, G., Bogdanowicz, E., Castellarin, A., Hall, J., Kobold, M., Kriauciuniene, J., Lang, M., Madsen, H., Onuşluel Gül, G., Perdigão, R. A.P., Roald, L.A., Salinas, J.L., Toumazis, A.D., Veijalainen, N., Þórarinsson, Ó., 2015. Identification of coherent flood regions across Europe by using the longest streamflow records. J. Hydrol. 528, 341–360.
Merz, B., Plate, E., 1997. An analysis of the effects of spatial variability of soil and soil moisture on runoff. Water Resour. Res. 33 (12), 2909–2922.
Merz, B., Aerts, J., Arnbjerg-Nielsen, K., Baldi, M., Becker, A., Bichet, A., Blöschl, G., Bouwer, L.M., Brauer, A., Cioffi, F., Delgado, J.M., Gocht, M., Guzzetti, F., Harrigan, S., Hirschboeck, K., Kilsby, C., Kron, W., Kwon, H.-H., Lall, U., Merz, R., Nissen, K., Salvatti, P., Swierczynski, T., Ulbrich, U., Viglione, A., Ward, P.J., Weiler, M., Wilhelm, B., Nied, M., 2014a. Floods and climate: emerging perspectives for flood risk assessment and management. Nat. Hazards Earth Syst. Sci. (NHESS) 14 (7), 1921–1942.
Merz, B., Elmer, F., Kunz, M., Mühr, B., Schröter, K., Uhlemann-Elmer, S., 2014b. The extreme flood in June 2013 in Germany. Houille Blanche - Revue internationale de l'eau 1, 5–10.
Michaud, J.D., Hirschboeck, K.K., Winchell, M., 2001. Regional variations in small basin floods in the United States. Water Resour. Res. 37 (5), 1405–1416.
Mudelsee, M., Börngen, M., Tetzlaff, G., Gründewald, U., 2003. No upward trends in the occurence of extreme floods in central Europe. Nature 425, 166–169.
Mudelsee, M., Börngen, M., Tetzlaff, G., Grünewald, U., 2004. Extreme floods in central Europe over the past 500 years: role of cyclone pathway "Zugstrasse Vb". J. Geophys. Res. 109, D23101.
Mudelsee, M., 2010. Climate Time Series Analysis: Classical Statistical and Bootstrap Methods. Springer, Dordrecht, p. 474 p.
Petrow, T., Merz, B., Lindenschmidt, K.-E., Thieken, A.H., 2007. Aspects of seasonality and flood generating circulation patterns in a mountainous catchment in south-eastern Germany. Hydrol. Earth Syst. Sci. 11, 1455–1468.
Petrow, T., Zimmer, J., Merz, B., 2009. Changes in the flood hazard in Germany through changing frequency and persistence of circulation patterns. Nat. Hazards Earth Syst. Sci. (NHESS) 9, 1409–1423 <www.nat-hazards-earth-syst-sci.net/9/1409/2009>.
Pfister, L., Kwadijk, J., Musy, A., Bronstert, A., Hoffmann, L., 2004. Climate change, land use change and runoff prediction in the Rhine-Meuse basins. River Res. Appl. 20, 229–241. http://dx.doi.org/10.1002/rra.775.
Pizaro, G., Lall, U., 2002. El Nino and floods in the U.S. West: what can we expect? EOS Trans., AGU 83 (32), 349–352.
Ramesh, N.I., Davison, A.C., 2002. Local models for exploratory analysis of hydrological extremes. J. Hydrol. 256, 106–119.
Raschke, M., 2015. Statistical detection and modelling of the over-dispersion of winter storm occurrence. Nat. Hazards Earth Syst. Sci. 15, 1757–1761. http://dx.doi.org/10.5194/nhess-15-1757-2015.
Renard, B., Lang, M., Bois, P., Dupeyrat, A., Mestre, O., Niel, H., Sauquet, E., Prudhomme, C., Parey, S., Paquet, E., Neppel, L., Gailhard, J., 2008. Regional methods for trend detection: assessing field significance and regional consistency. Water Resour. Res. 44, W08419. http://dx.doi.org/10.1029/2007WR006268.
Robson, A.J., Jones, T.K., Reed, D.W., Bayliss, A.C., 1998. A study of national trend and variation in UK floods. Int. J. Climatol. 18 (2), 165–182.
Robson, A., 2002. Evidence of trends in UK flooding. Philos. Trans. Roy. Soc. Lond. A 360, 1327–1343.
Sankarasubramanian, A., Lall, U., 2003. Flood quantiles in a changing climate: seasonal forecasts and causal relations. Water Resour. Res. 39, 1134. http://dx.doi.org/10.1029/2002WR001593.
Schmocker-Fackel, P., Naef, F., 2010a. Changes in flood frequencies in Switzerland since 1500. Hydrol. Earth Syst. Sci. 14, 1581–1594.
Schmocker-Fackel, P., Naef, F., 2010b. More frequent flooding? Changes in flood frequency in Switzerland since 1850. J. Hydrol. 381 (1–8), 2009. http://dx.doi.org/10.1016/j.jhydrol..09.022.
Schröter, K., Kunz, M., Elmer, F., Mühr, B., Merz, B., 2015. What made the June 2013 flood in Germany an exceptional event? A hydro-meteorological evaluation. Hydrol. Earth Syst. Sci. 19, 309–327.
Silva, A.T., Portela, M.M., Naghettini, M., 2012. Nonstationarities in the occurrence rates of flood events in Portuguese watersheds. Hydrol. Earth Syst. Sci. 16, 241–254. http://dx.doi.org/10.5194/hess-16-241-2012.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London, New York, p. 175 p.
Swierczynski, T., Lauterbach, S., Dulski, P., Delgado, J., Merz, B., Brauer, A., 2013. Mid- to late Holocene flood frequency changes in the northeastern Alps as recorded in varved sediments of Lake Mondsee (Upper Austria). Quatern. Sci. Rev. 80, 78–90.
Sturm, K., Glaser, R., Jacobeit, J., Deutsch, M., Brázdil, R., Pfister, C., Luterbacher, J., Wanner, H., 2001. Hochwasser in Mitteleuropa seit 1500 und Beziehung zur atmosphärischen Zirkulation. Petermanns Geogr. Mitt. 145, 14–23.

Ventura, V., Paciorek, C.J., Risbey, J.S., 2004. Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. J. Clim. 17, 4343–4356.

Villarini, G., Smith, J.A., Vitolo, R., Stephenson, D.B., 2013. On the temporal clustering of US floods and its relationship to climate teleconnection patterns. Int. J. Climatol. 33, 629–640. http://dx.doi.org/10.1002/joc.3458.

Vitolo, R., Stephenson, D.B., Cook, I.M., Mitchell-Wallace, K., 2009. Serial clustering of intense European storms. Meteorol. Z. 18, 411–424. http://dx.doi.org/10.1127/0941-2948/2009/0393.

Vorogushyn, S., Merz, B., 2013. Flood trends along the Rhine: the role of river training. Hydrol. Earth Syst. Sci. 17, 3871–3884. http://dx.doi.org/10.5194/hess-17-3871-2013.

Waylen, P.R., Caviedes, C.N., 1986. El Niño and annual floods on the north Peruvian Littoral. J. Hydrol. 89, 141–156.

Ward, P.J., Beets, W., Bouwer, L.M., Aerts, J.C.J.H., Renssen, H., 2010. Sensitivity of river discharge to ENSO. Geophys. Res. Lett. 37 (12), L12402. http://dx.doi.org/10.1029/2010GL043215.

Ward, P.J., Eisner, S., Flörke, M., Dettinger, M.D., Kummu, M., 2014. Annual flood sensitivities to El Niño Southern Oscillation at the global scale. Hydrol. Earth Syst. Sci. 18, 47–66. http://dx.doi.org/10.5194/hess-18-47-2014.

Wetter, O., Pfister, C., Weingartner, R., Luterbacher, J., Reist, T., Trösch, J., 2011. The largest floods in the High Rhine basin since 1268 assessed from documentary and instrumental evidence. Hydrol. Sci. J. 56, 733–758. http://dx.doi.org/10.1080/02626667.2011.583613.

Wilks, S.D., 2006. On "field significance" and false discovery rate. J. Appl. Meteorol. Climatol. 45, 1181–1189.

Zhang, Q., Xu, Ch., Jiang, T., Wu, Y., 2007. Possible influence of ENSO on annual maximum streamflow of the Yangtze River, China. J. Hydrol. 333, 265–274.