INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. (2012)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.3458

RMetS
Royal Meteorological Society

On the temporal clustering of US floods and its relationship to climate teleconnection patterns

Gabriele Villarini,[a,b]* James A. Smith,[a] Renato Vitolo[b,c] and David B. Stephenson[c]
[a] *Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA*
[b] *Willis Research Network, London, UK*
[c] *College of Engineering, Mathematics, and Physical Sciences, University of Exeter, Exeter, UK*

ABSTRACT: This article examines whether the temporal clustering of flood events can be explained in terms of climate variability or time-varying land-surface state variables. The point process modelling framework for flood occurrence is based on Cox processes, which can be represented as Poisson processes with randomly varying rate of occurrence. In the special case that the rate of occurrence is deterministic, the Cox process simplifies to a Poisson process. Poisson processes represent flood occurrences which are not clustered. The Cox regression model is used to examine the dependence of the rate of occurrence on covariate processes. We focus on 41 stream gauge stations in Iowa, with discharge records covering the period 1950–2009. The climate covariates used in this study are the North Atlantic Oscillation (NAO) and the Pacific/North American Teleconnection (PNA). To examine the influence of land-surface forcing on flood occurrence, the antecedent 30 d rainfall accumulation is considered. In 27 out of 41 stations, either PNA or NAO, or both are selected as significant predictors, suggesting that flood occurrence in Iowa is influenced by large-scale climate indices. Antecedent rainfall, used as a proxy for soil moisture, plays an important role in driving the occurrence of flooding in Iowa. These results point to clustering as an important element of the flood occurrence process. Copyright © 2012 Royal Meteorological Society

KEY WORDS temporal clustering; flood; Cox regression; Iowa; NAO; PNA

Received 5 September 2011; Revised 30 January 2012; Accepted 5 February 2012

1. Introduction

The idea that rainfall cells cluster into mesoscale rain areas can be dated back to the work of Le Cam (1961). Because of the clustered nature of rainfall (Kavvas and Delleur, 1975; Gupta and Waymire, 1979; Smith and Karr, 1983), it is generally not appropriate to describe the occurrence of storms as a Poisson process. The need to reproduce the observed clustered behaviour of storm occurrences has resulted in advances in stochastic modelling of rainfall (Kavvas and Delleur, 1981; Waymire and Gupta, 1981; Waymire *et al*., 1984; Ramirez and Bras, 1985; Smith and Karr, 1985; Rodriguez-Iturbe *et al*., 1987; Istok and Boersma, 1989; consult Onof *et al*. (2000) for an overview). Models describing clustering of flood occurrence have also been developed (Cervantes *et al*., 1983; Smith and Karr, 1986; Kavvas, 1987; Futter *et al*., 1991).

Recent studies have also examined clustering in the occurrence of extratropical and tropical storms (Mailier *et al*., 2006; Vitolo *et al*., 2009; Villarini *et al*., 2010). These studies have pointed to clustering as a significant feature of the storm occurrence process. More generally, the question of whether hydrometeorological events (e.g. heavy rainfall, floods, winter storms, heat waves) cluster is not only of high scientific interest but has also large economic repercussions (Mailier *et al*., 2006; Vitolo *et al*., 2009).

The point process modelling framework (Cox and Isham, 1980; Karr, 1991) for flood occurrence is based on Cox processes, which can be represented as Poisson processes with randomly varying rate of occurrence. In the special case that the rate of occurrence is deterministic, the Cox process simplifies to a Poisson process. Poisson processes represent flood occurrences which are not clustered. Among the models proposed and developed over the past 30 years, the Cox regression model provides a powerful statistical framework to check whether the rate of occurrence of a counting process depends on covariate processes. Smith and Karr (1986) introduced the Cox regression model for flood occurrences in which the rate of occurrence depends on time-varying covariate processes. Despite its power and utility, the Cox regression model has received little attention by the hydrometeorological community (Smith and Karr, 1983, 1986; Futter *et al*., 1991; Maia and Meinke, 2010).

In this study, we examine the temporal clustering of flood events in Iowa. We focus on Iowa because

* Correspondence to: G. Villarini, Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA.
E-mail: gvillari@princeton.edu

Copyright © 2012 Royal Meteorological Society

flood events are responsible for large economic damage, as exemplified by the multi-billion dollar losses resulting from the 1993 and 2008 floods (Kunkel *et al.*, 1994; Otto, 2009). Iowa also represents an interesting study site, with almost half of the state draining into the Missouri to the west, and to the Mississippi to the east. We build on the work by Smith and Karr (1986), by performing flood frequency analysis using the Cox regression model, and including time-varying predictors related to climate variability and/or time-varying land-surface state variables, like soil moisture.

The main issues of this study focus on:

1. characterization of clustering of the flood occurrence process;
2. examination of the time-varying climate and land-surface state variables that could be useful in modelling the rate of occurrence of flooding;
3. assessment of the utility of the Cox regression model in flood frequency studies

This article is organized as follows. In the next section, we present the Cox regression model, followed by Section 3 in which we describe the flood peak data and the predictors. Section 4 presents the results of our analyses, while Section 5 summarizes the main points of this study.

2. Cox regression model

A marked point process model for flood occurrence and magnitude can be written as:

$$\{T_{ij}, X_{ij}; i = 1, \ldots n;\ \ j = 1, \ldots, M_i\} \tag{1}$$

where n is the number of years of record, M_i is the number of flood peaks (see Section 3 for their definition) during year i, T_{ij} is the time of the j^{th} flood during year i and X_{ij} is the magnitude of the j^{th} flood during year i.

We can describe the point process using a counting process representation

$$N_i(t) = \sum_{j=1}^{M_i} 1(T_{ij} \leq t) \tag{2}$$

for $t \in [0, T]$, where 0 represents time 0 during the year and T represents the ending time for the year.

Thinning by event magnitude x, we can rewrite Equation 2 as (Figure 1):

$$N_i^x(t) = \sum_{j=1}^{N_i(t)} 1(X_{ij} > x) \tag{3}$$

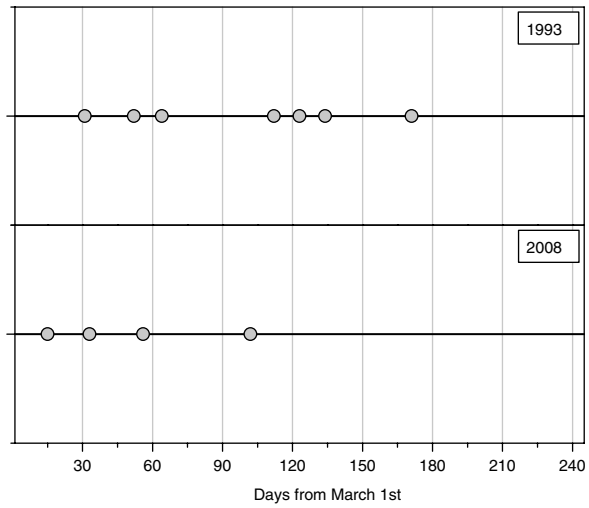

Figure 1. Point process representation of peak discharge for the Turkey River for the years 1993 and 2008.

$\{N_i^x(t), t \in [0, T]\}$ is a Poisson process provided that counts in discrete intervals are independent and the distribution of counts is Poisson, i.e.

$$Pr\{N_i^x(t) = k\} = \frac{\exp\left\{-\int_0^t \lambda(u)\ du\right\}\left[\int_0^t \lambda(u)\ \mathrm{d}u\right]^k}{k!} \tag{4}$$

where $(\lambda(u), u \in [0, T])$ is a non-negative function representing the time-varying rate of occurrence of the process. If there is no seasonality in the occurrence process, then $(\lambda(u), u \in [0, T])$ simplifies to a constant λ and the occurrence process is a homogeneous Poisson process.

We use the term clustered to mean that the occurrence process is not Poisson. A Poisson process with seasonally varying rate of occurrence may exhibit a concentration of flood occurrences during a particular time of year, but this does not represent clustering, in which the occurrence of an event contains information about the subsequent rate of occurrence of events. By applying an appropriate transformation to the time axis, the rate of occurrence of a non-homogeneous Poisson process reduces to a constant and the time of arrival would be described by an exponential distribution. Therefore, the occurrence process for homogeneous and non-homogeneous Poisson processes is still Poisson, and they represent special cases of Cox processes.

Cox processes, also known as doubly stochastic Poisson processes, are an important family of 'clustered' point processes, which can be viewed as Poisson processes with a randomly varying rate of occurrence (Kingman, 1964; Cox, 1972; Grandell, 1976; Karr, 1991). Applications of Cox processes to modelling occurrence processes for rainfall and floods include Smith and Karr (1983, 1985, 1986) and Futter *et al.* (1991). $N_i^x(t)$ is a Cox process provided that there is a stochastic process $\{\lambda(u); u \in [0, T]\}$, such that, counts in discrete intervals are conditionally independent given $\{\lambda(u); u \in [0, T]\}$ and the conditional distribution of counts, given

Copyright © 2012 Royal Meteorological Society

$\{\lambda(u); u \in [0, T]\}$, is Poisson with:

$$P\{N_i^x(t) = k|\lambda(u), u \le t\} = \frac{\exp\left\{-\int_0^t \lambda(u)\mathrm{d}u\right\}\left[\int_0^t \lambda(u)\mathrm{d}u\right]^k}{k!} \quad (5)$$

In a Cox process, an event can be relatively more likely (or less likely) to be followed by additional events depending on the distributional properties of $\{\lambda(u); u \geq 0\}$. Although the conditional distribution of counts given $\lambda(u)$ is Poisson, the unconditional distribution is not Poisson. Hence, in a realization of a Cox process the counts will exhibit random bursts of activity or inactivity, exceeding the variability of a Poisson process.

The Cox regression model was introduced by Cox (1972) and represents Cox processes for which the rate of occurrence process has a specific functional dependence on covariate processes. Denote the j^{th} covariate process for the i^{th} year as $Z_{ij}(t)$, with $i = 1, \ldots, n$ and $j = 1, \ldots, m$. For the i^{th} year, the rate of occurrence process (also known as the conditional intensity function or hazard function) is given by:

$$\lambda_i(t) = \lambda_0(t) \exp\left[\sum_{j=1}^{m} \beta_j Z_{ij}(t)\right] \quad (6)$$

where $\lambda_0(t)$ is called the *baseline hazard* and is a non-negative function of time, and β_j is the coefficient for the j^{th} covariate. The baseline hazard is not parametrically specified and the covariates are linearly related to the hazard function, implying that the model falls in the semiparametric family (Cox, 1972; Andersen *et al*., 1992).

For 2 years i and i', we can write the hazard ratio as:

$$\frac{\lambda_i(t)}{\lambda_{i'}(t)} = \frac{\lambda_0(t)\exp[Z_i(t)\beta]}{\lambda_0(t)\exp[Z_{i'}(t)\beta]} = \frac{\exp[Z_i(t)\beta]}{\exp[Z_{i'}(t)\beta]} \quad (7)$$

with the hazard ratio which is independent of time, making the Cox model a proportional-hazards model.

Estimation of the β coefficients is performed using the partial likelihood function for the case of no ties (among others, see Cox 1975; Gill, 1984; Andersen and Gill, 1982; and Smith and Karr, 1986):

$$\mathcal{L}(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left(\frac{\exp[Z_i(t_i)\beta]}{\sum_j \exp[Z_j(t_j)\beta]}\right)^{\mathrm{d}N_i(t)} \quad (8)$$

where $N_i(t)$ is the number of events for year i over the time interval [0, t] and $dN_i(t)$ represents the increment of N_i over a small time interval around t (Figure 1). In a Cox regression model, the random bursts of activity/inactivity in the count process are explicitly driven by the covariate processes Z_{ij} through Equation (6). In our application, clustering of the counts is induced by the external physical processes represented by the covariates. In the case of ties (e.g. peaks occurring on the same day but in different years) we use Efron's approximation because it is quite accurate and computationally efficient (Therneau and Grambsch, 2000). For a more extensive discussion on handling ties, the interested reader is pointed to Section 3.3 of Therneau and Grambsch (2000).

The baseline hazard function is often viewed as a nuisance parameter in applications. In our case, however, its specification is important in order to completely specify the time-varying rate of occurrence of floods (Smith and Karr, 1986). The mean rate of occurrence is

$$m(t) = \frac{d}{dt}E[N_i(t)] \quad (9)$$

and we denote its estimator $\hat{m}(t)$. If $\beta_j = 0$ for all j, then $\hat{m}(t)$ is the estimator of the Poisson intensity function $\lambda_0(t)$:

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i(t) = \hat{m}(t) \quad (10)$$

If $\beta_j \neq 0$, we can then estimate $\lambda_0(t)$ given β_j and $Z_{ij}(t)$ from the moments estimator:

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\lambda}_0(t)\exp\left[\sum_{j=1}^{m} Z_{ij}(t)\beta_j\right] = \hat{m}(t) \quad (11)$$

We can obtain the estimate $\hat{\lambda}_0(t)$ at time t so that the equality in Equation (11) is verified (the only unknown is $\hat{\lambda}_0(t)$). The approach we follow is to first smooth $\hat{m}(t)$ (computed as in Equation (10)) using local polynomial regression (loess function (Cleveland, 1979), with a span of 0.5), and then solve Equation (11) for $\hat{\lambda}_0(t)$.

In addition to the covariates Z_j $(j = 1, \ldots, m)$, we also include interaction terms between covariates. To decide which predictors should be included in the final model, we use a stepwise method, penalizing more complex models with respect to the Akaike Information Criterion (AIC; Akaike, 1974). To check whether the final model describes the data adequately, we use two different diagnostics (Therneau and Grambsch, 2000), the scaled Schoenfeld and Dfbeta residuals. The former are used to assess violations of the assumption of proportional hazards, and plotting these residuals against time provides indication about the presence of linear trends in the covariates. Different possible transformations are available (e.g. based on the rank of the event times, on the Kaplan–Meier estimate of the survival function). Dfbeta residuals provide information about the changes in the regression coefficients when one observation is removed, and they are used to assess the presence of influential observations.

All the calculations are performed in `R` (R Development Core Team, 2008) using the freely available `survival` package (Therneau and original R port by Thomas Lumley, 2009).

Copyright © 2012 Royal Meteorological Society

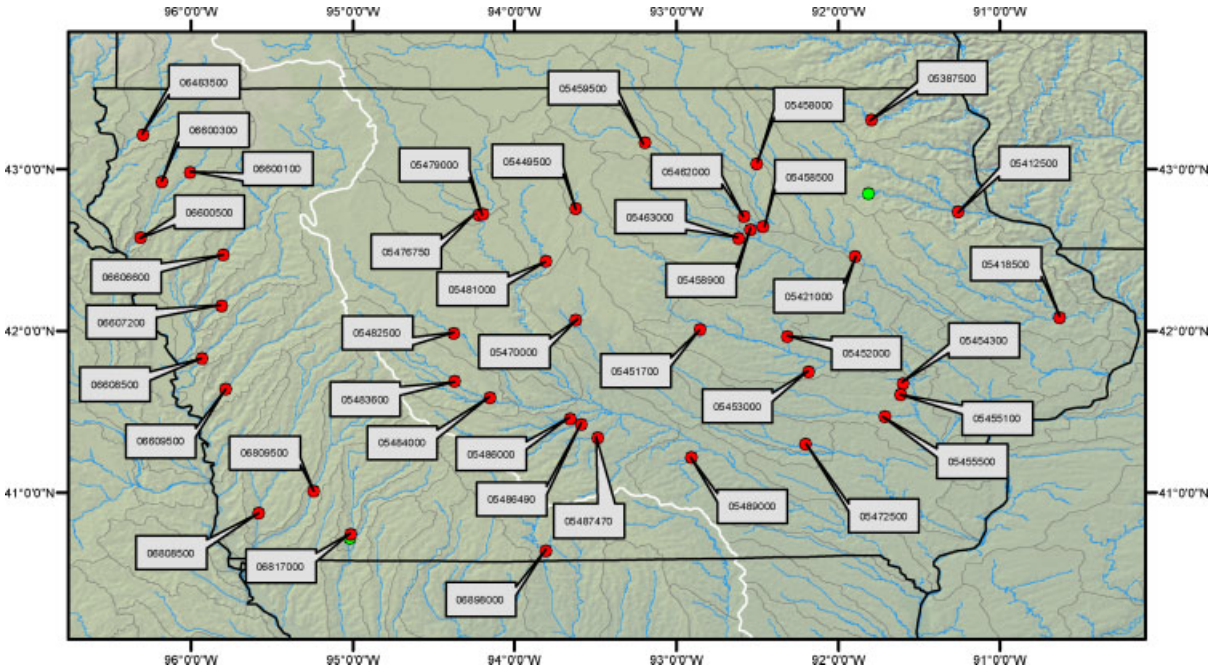

Figure 2. Map showing the location of the 41 stations included in this study (red circles; the number associated with each station represents the corresponding USGS ID number). The white line represents the divide between the Missouri River (to the west) and the Mississippi River (to the east). The green circles represent the location of the rain gauges. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

3. Data

In this study, we consider 41 mostly non-nested (no station is downstream of another) US Geological Survey (USGS) stations over Iowa (Figure 2), with a drainage area ranging from 254 to 6475 km^2, with a median value of 1733 km^2. The period of record is 1950–2009. Out of 41 stations, 1 has 38 years of data, 2 have 39 years, 1 has 45 years, 7 stations have between 54 and 59 years, and 30 stations have 60 years of data. Flood events are obtained by thresholding discharge observations. Different approaches have been proposed for threshold selection (Davison and Smith, 1990; Lang *et al*., 1999; Coles, 2001). In this study, we select the threshold so that there are, on average, two peaks per year. To avoid having peaks exceeding the threshold coming from the same rainfall event (this problem is more significant for larger drainage areas), we consider a window of 15 d (centred on the day of the peak) during which only one peak over the selected threshold is allowed (see also Lang *et al*. (1999) for a review). For every year, the peaks are selected during the period March–October (see also Smith and Karr, 1983), since this is the period of the year with the highest frequency of flood peaks (consult Villarini *et al*. (2011a, 2011b) for results concerning seasonality of flood peaks and heavy rainfall over Iowa).

Villarini *et al*. (2011a) performed analyses on the annual maximum flood peak distribution over the central United States. Over Iowa, few stations exhibited a statistically significant abrupt change in the mean and/or variance of the flood peak distribution and in only one station a statistically significant monotonic trend. Moreover, recent analyses on heavy rainfall in the upper Midwest United States do not point to significant increasing or decreasing trends in annual maximum daily rainfall (Villarini *et al*., 2011b). On the basis of these recent results, we can conclude that stationarity is a reasonable working assumption for this study. Therefore, description of the rate of occurrence as a function of covariates mostly indicates violation of the independence assumption, pointing to clustering of flood events (Smith and Karr, 1983; Karr, 1991).

Covariates used in this study are the North Atlantic Oscillation (NAO; Barnston and Livezey, 1987; Hurrell, 1995; Hurrell and Van Loon, 1997; Ambaum *et al*., 2001) and the Pacific/North American Teleconnection (PNA; Wallace and Gutzler, 1981; Leathers *et al*., 1991). Time series of these covariates are available at a daily time step from the Climate Prediction Center (CPC) covering the same period as the discharge data (1950–2009). Their values are computed based on the methodology described in Barnston and Livezey (1987). We focus on the values of these climate indices averaged over the previous 14 d (we refer to them as 'NAO14' and 'PNA14') and 28 d (we refer to them as 'NAO28' and 'PNA28'), to capture longer time scales of influence of these indices.

We selected these climate indices not only because of the availability of daily data covering a long time period but also because of the link between PNA and NAO and hydrometeorological variables over the Midwest United States. Barnston and Livezey (1987) found NAO to be a major mode of interannual variability for the Northern Hemisphere for all seasons, while PNA is an important mode mostly during early spring and fall-winter. These two indices have been shown to describe temperature and rainfall variability over the United States (Leathers *et al*., 1991). Leathers *et al*. (1991) found PNA to be

Copyright © 2012 Royal Meteorological Society

correlated with temperature during winter, spring, and autumn, while the correlation with precipitation is weaker than that found for temperature. Over the central United States and during the June–August months, Barlow *et al*. (2001) found a significant correlation between streamflow and monthly rainfall and the North Pacific sea surface temperatures (SSTs). Similarly, Ting and Wang (1997) found a significant correlation between summertime precipitation over the Great Plains and North Pacific SST (see Ting and Wang, 1997 for a discussion about the link between PNA and Pacific SST). Bates *et al*. (2001) found a significant relation between eastern Pacific SST and extreme springtime rainfall over the central United States. Recently, Coleman and Budikova (2010) highlighted the importance of PNA and NAO in the 1993 and 2008 Midwest floods.

In addition to examining whether PNA and/or NAO are significant predictors, as additional covariate for two stream gauge stations (Nodaway River and Turkey River; see Figure 2 for their location) we include the antecedent 30 d rainfall accumulation (normalized by 100 mm to have all the predictors on the same scale) based on the daily time series from the two rain gauges shown in Figure 2. This predictor provides information about the antecedent soil moisture conditions, allowing a comparison of the influence of atmospheric and land-surface forcings on flood occurrence.

4. Results

4.1. PNA and NAO as covariates

We used the Cox regression model to examine the relation between the occurrence of flood events in Iowa and climate indices. We have summarized our results in Figure 3 and the values of the coefficients in Table I. Using a stepwise method for covariate selection and AIC as penalty criterion, we found that for 27 out of

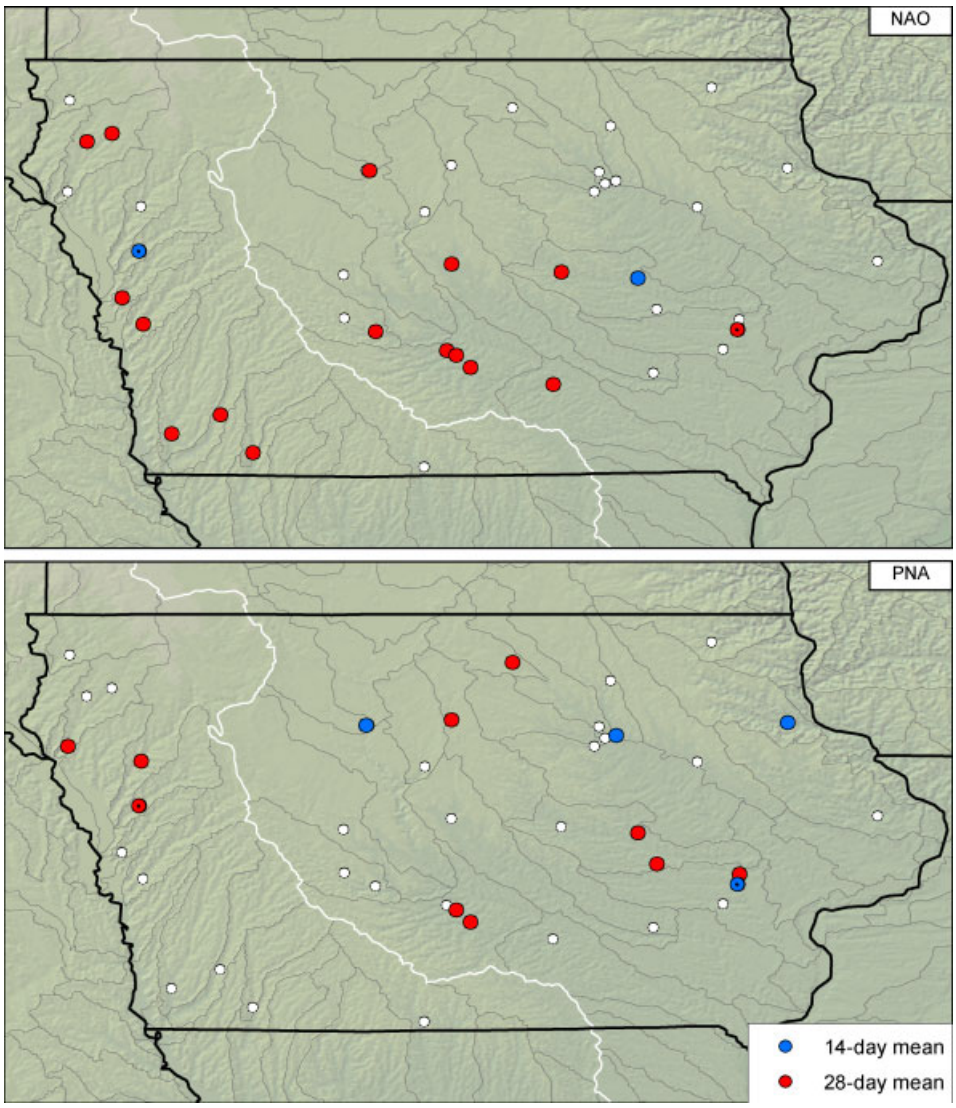

Figure 3. Map showing the stations for which NAO (top panel) and PNA (bottom panel) are retained as important covariates in the final models (white circle: the climate index is not retained as an important predictor; blue circle: value of the climate index averaged over 14 d prior to a given day; red circle: value of the climate index averaged over 28 d prior to a given day; the circles with the dot at the centre indicate that the final model includes an interaction term). This figure is available in colour online at wileyonlinelibrary.com/journal/joc

Copyright © 2012 Royal Meteorological Society

Table I. Summary statistics of the Cox regression model using climate indices as predictors. See Figure 2 for gauge location. In columns 2–7, the first value is the point estimate, while the one in parenthesis is the standard error. Model selection is performed using AIC as penalty criterion.

USGS ID	NAO14	PNA14	NAO28	PNA28	NAO28:PNA14	NAO14:PNA28
05412500	–	−0.21 (0.15)	–	–	–	–
05449500	–	–	–	0.32 (0.18)	–	–
05451700	–	–	−0.43 (0.19)	–	–	–
05452000	−0.39 (0.15)	–	–	0.30 (0.17)	–	–
05453000	–	–	–	0.35 (0.18)	–	–
05454300	–	–	–	0.27 (0.18)	–	–
05455100	–	–	−0.54 (0.25)	0.42 (0.23)	0.72 (0.39)	–
05458500	–	−0.22 (0.15)	–	–	–	–
05459500	–	–	–	0.38 (0.18)	–	–
05470000	–	–	−0.30 (0.20)	–	–	–
05476750	–	0.29 (0.16)	–	–	–	–
05479000	–	–	−0.36 (0.19)	–	–	–
05484000	–	–	−0.38 (0.19)	–	–	–
05486000	–	–	−0.50 (0.19)	–	–	–
05486490	–	–	−0.34 (0.19)	0.46 (0.18)	–	–
05487470	–	–	−0.39 (0.19)	0.47 (0.18)	–	–
05489000	–	–	−0.37 (0.19)	–	–	–
06600100	–	–	−0.35 (0.19)	–	–	–
06600300	–	–	−0.43 (0.23)	–	–	–
06600500	–	–	–	0.28 (0.18)	–	–
06606600	–	–	–	0.35 (0.18)	–	–
06607200	–	–	−0.43 (0.19)	0.37 (0.17)	–	0.74 (0.28)
06608500	–	–	−0.51 (0.19)	–	–	–
06609500	–	–	−0.34 (0.19)	–	–	–
06808500	–	–	−0.48 (0.19)	–	–	–
06809500	–	–	−0.63 (0.19)	–	–	–
06817000	–	–	−0.35 (0.19)	–	–	–

41 stations a model in which the rate of occurrence λ was a linear function of NAO and/or PNA (via a logarithmic link function) was preferred to a model with constant λ. While the values of NAO and PNA averaged over a 14 d period are important covariates in only one and three cases respectively, NAO and PNA averaged over a 28 d period are important predictors in 17 and 11 stations, respectively. Finally, there are two stations for which an interaction term is included in the final model (one between PNA28 and NAO14, and one between PNA14 and NAO28). For 22 stream gauges, only one covariate is included in the final model, while both PNA and NAO are included in five cases. NAO28 exhibits the clearest spatial structure, and it tends to be mostly significant in western and central Iowa (Figure 3, top panel). PNA is generally more widespread over the entire state. Of all the predictors, those averaged over a 28 d period are the ones that are the most frequently selected in this study. These results suggest that the average state a month prior to the occurrence of a flood event is more important that the state of the atmosphere 14 d prior to the flood.

For all basins, we have examined the scaled Schoenfeld and Dfbeta residuals (Figure 4). For each covariate and the global model (if more than one covariate was included in the final model), we tested the scaled Schoenfeld residuals for the presence of linear trends (see Figure 4 (left panels) for two examples). In the vast majority of the cases, the proportional-hazard assumption is valid at the 5% significance level. Examination of the Dfbeta residuals (see Figure 4 (right panels) for two examples) does not suggest that any of the observations was particularly influential. These results support the selection of these models. The values of NAO14 and NAO28 (PNA14 and PNA28) are negative (positive), implying that large (small) values of PNA and NAO would result in a reduced (increased) rate of occurrence of flooding over these catchments in Iowa. Therefore, in 27 out of 41 stations we are able to describe the rate of occurrence λ as a function of NAO and/or PNA, suggesting that the occurrence of flooding in Iowa is influenced by large-scale climate indices.

We also fitted a stratified Cox model to examine which, if any, of these covariates were important when we pooled stream gauge stations together. This model is an extension of the Cox model, in which the data are divided into strata, and each stratum shares the same coefficients of the covariates with other strata, but each has an individual baseline hazard function. In this case, each stratum is represented by a stream gauge station. We examined two grouping schemes. In the first one, we have pooled together all the 41 stations. In this case, NAO28 and PNA28 are important covariates according to AIC.

Copyright © 2012 Royal Meteorological Society

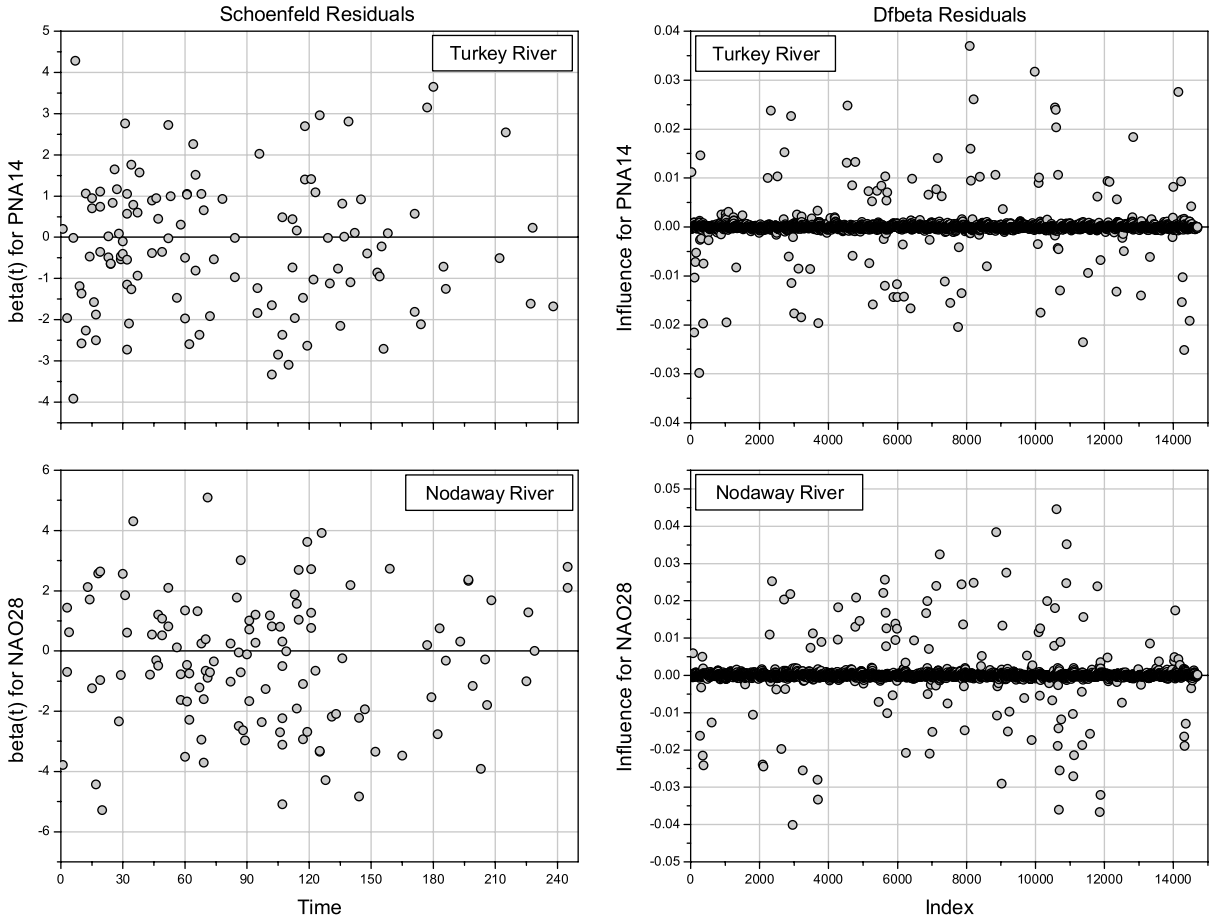

Figure 4. Schoenfeld (left panels) and Dfbeta (right panels) residuals for the Turkey River (upper panels) and the Nodaway River (bottom panels).

As a second scheme, we divided Iowa into four quadrant (north and south of 42°N, and east and west of 93.5 °W), with 8 stations in the southeast quadrant and 11 in each of the other three. Even in these cases, we found that PNA28 and NAO28 were included in the final model for three of the four quadrants, with the exception of the northeast one, for which PNA14 and NAO28 were included. These results provide supporting evidence of the importance of NAO and PNA in describing the occurrence of flooding over Iowa.

Another important outcome of these analyses concerns the hypothesis of independence of flood events. For the 27 catchments for which NAO and/or PNA are significant covariates, we found that the occurrence of flood events is not independent, as generally assumed, but exhibits temporal clustering.

4.2. PNA, NAO, and antecedent rainfall as covariates

Based on the results in Figure 3, PNA14 was retained as a significant predictor for the Turkey River, while NAO28 was included in the final model for the Nodaway River. On the other hand, when we added antecedent rainfall as a covariate, this additional predictor was selected as important for both catchments. For the Turkey River, the rate of occurrence is a function of PNA14 and antecedent rainfall, while for the Nodaway River it depends only on antecedent rainfall. In both cases, the coefficient of this additional predictor is positive, confirming that larger values of antecedent rainfall result in a larger rate of occurrence of flood events. This is consistent with the physical processes at play, because an increased antecedent rainfall would result in larger soil moisture, decreasing the infiltration capability of the soil, and resulting in an increased runoff.

After computing the baseline hazard function $\lambda_0(t)$ for each station (Figure 5), we can compute the rate of occurrence for any given year i (Figure 6). For reference, we also include $\hat{m}(t)$, which represents the behaviour of the corresponding inhomogeneous Poisson process (Figure 6, top panels). For the Turkey River, $\lambda_i(t)$ oscillates around $\hat{m}(t)$, with overall increasing rates in March–April and June–July. The Nodaway River exhibits a more regular pattern, with overall larger values of $\lambda_i(t)$ in May and June. In both cases, the larger values of the occurrence rate match well with the measured frequency of POT events (Figure 6, bottom panels). For the Turkey River, most of the events occur in March–April, with a second period of enhanced activity during June–July. For the Nodaway River most of the flood events are concentrated in the May–June period. Figure 6 illustrates that some years have a rate of occurrence that is much larger than in other years.

Because of the large socioeconomic impact of the 1993 and 2008 flood events, we focus on the rate

Copyright © 2012 Royal Meteorological Society

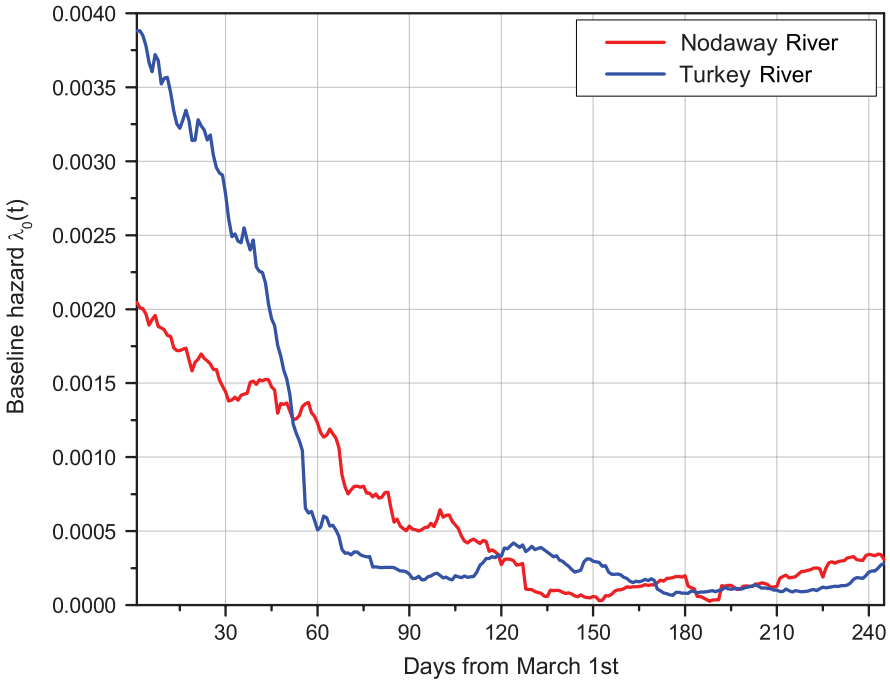

Figure 5. Baseline hazard function $\lambda_0(t)$ for the Turkey River and Nodaway River. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

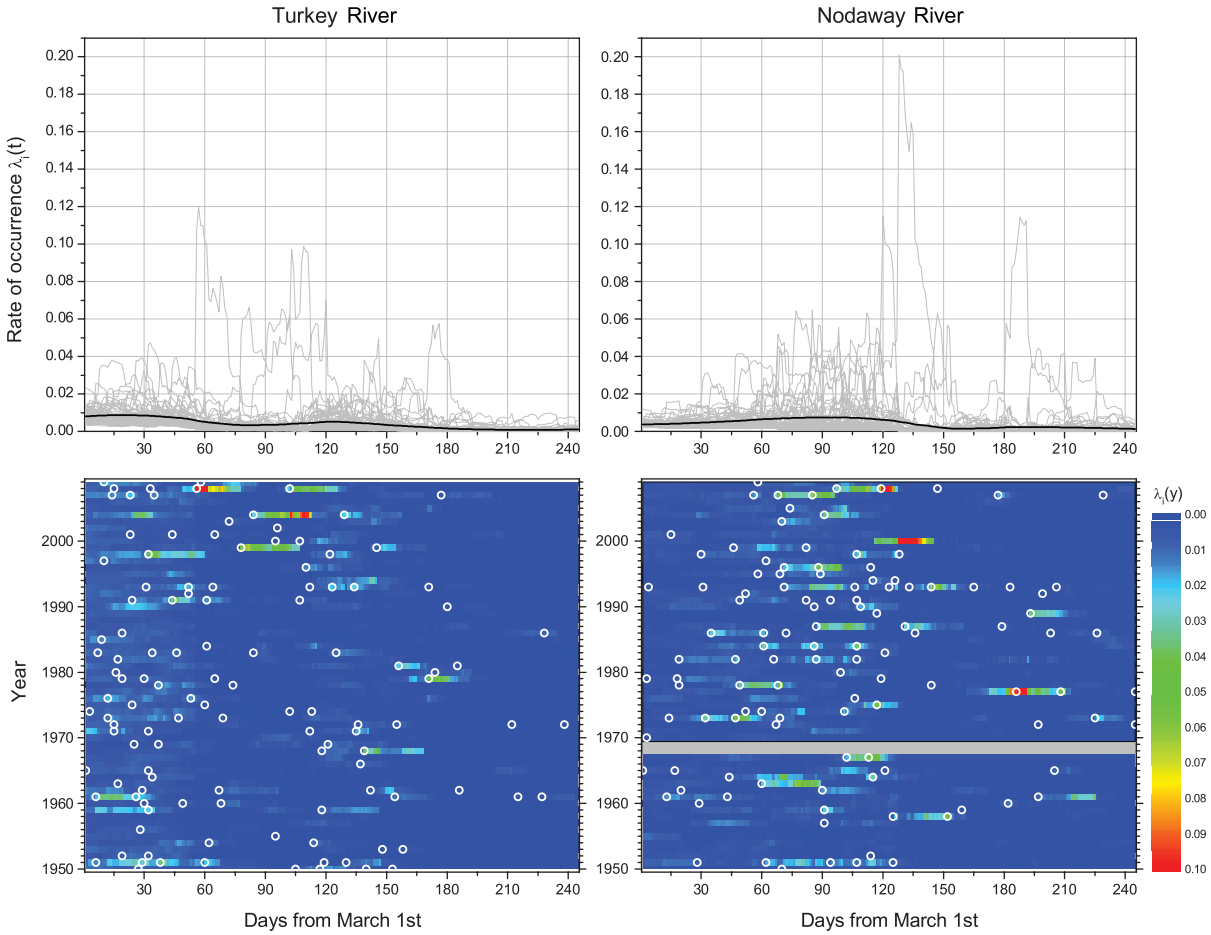

Figure 6. Plot of the $\lambda_i(t)$ for the Turkey River (left panels) and the Nodaway River (right panels). The solid black line in the top panels represents $\hat{m}(t)$. In the bottom panels, the white circles indicate the occurrence of a POT event from the data. Missing years in the bottom-right panel are represented by the grey band. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

Copyright © 2012 Royal Meteorological Society

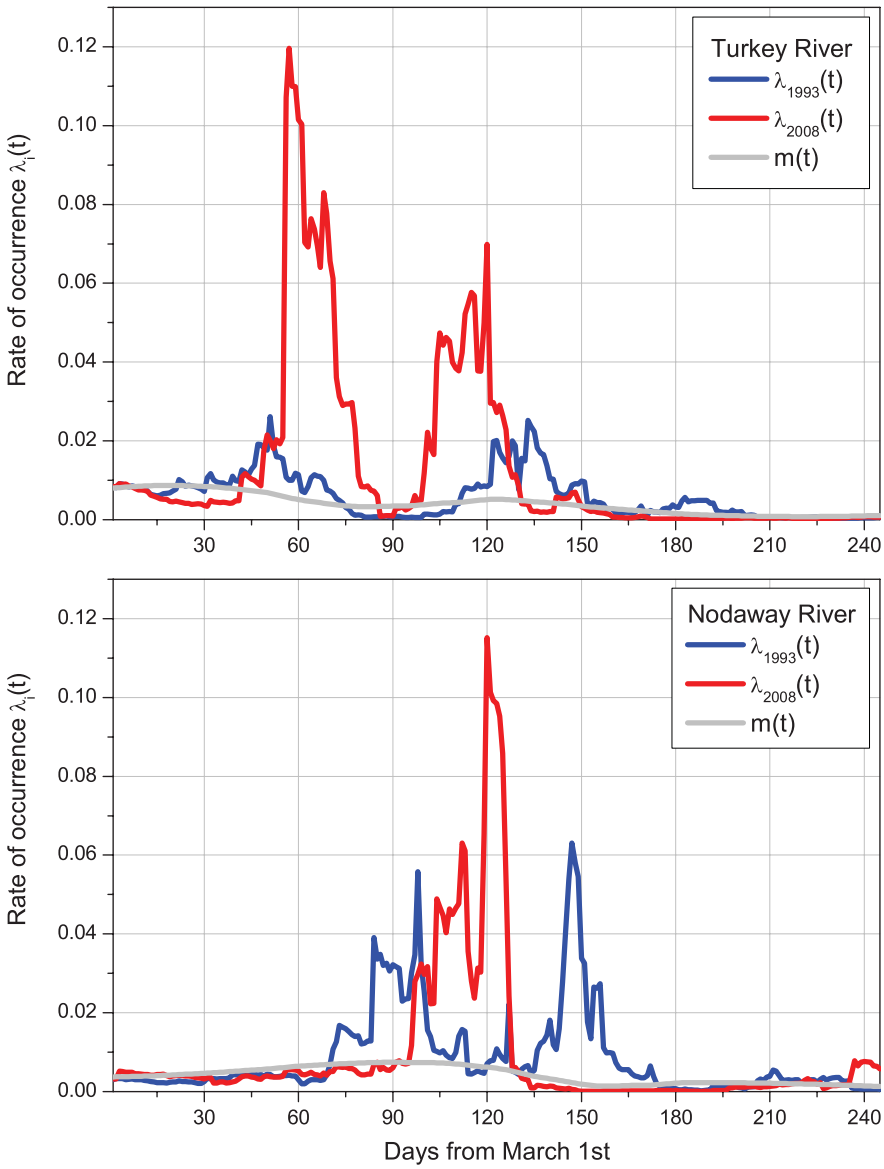

Figure 7. Plot of the $\lambda_i(t)$ for the Turkey River (top panel) and the Nodaway River (bottom panel) for the years 1993 and 2008. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

of occurrence for these 2 years (Figure 7). These plots highlight similarities and differences between catchments and years. In 2008, $\lambda_{2008}(t)$ for the Turkey River exhibits two peaks, one in late April–early May and one in June, corresponding to some of the largest monthly rainfall values recorded at the corresponding rain gauges. The behaviour for 1993 is different, with values that are larger than average but not as large as 2008, resulting from smaller monthly rainfall values. We can also highlight seasonal differences between these 2 years. In 2008 a series of flood peaks occurred during April, and then another series of events in June, with smaller discharge values from July to the end of the water year. In 1993, while the absolute magnitude of the events was smaller, the events were more spread over the season and they also occurred later (July and August); a similar behaviour is exhibited by the $\lambda_{1993}(t)$. For the Nodaway River, in 2008 the peak did not occur in the spring, rather the ratio of occurrence to $\hat{m}(t)$ was much larger in June. This is consistent with the discharge and rainfall time series, which exhibited a single heightened discharge and rainfall period in June. The results for 1993 are different, with larger values of $\lambda_{1993}(t)$ spread over a larger time span (from May to August), reflecting the more widespread rainfall accumulations.

Once we estimated the model coefficients β and the baseline hazard function (each station has its own model), we were able to use the values of the covariates for a certain day of year of interest to assess how much larger than $\hat{m}(t)$ the rate of occurrence $\lambda_i(t)$ was (Figure 8). For the Turkey River, the rate of occurrence in April–May and June was more than an order of magnitude larger than $\hat{m}(t)$. For the Nodaway River, the rate of occurrence was average or even below average, with the exception of June and early July, during which the rate of occurrence was much larger than $\hat{m}(t)$. It is worth mentioning

Copyright © 2012 Royal Meteorological Society

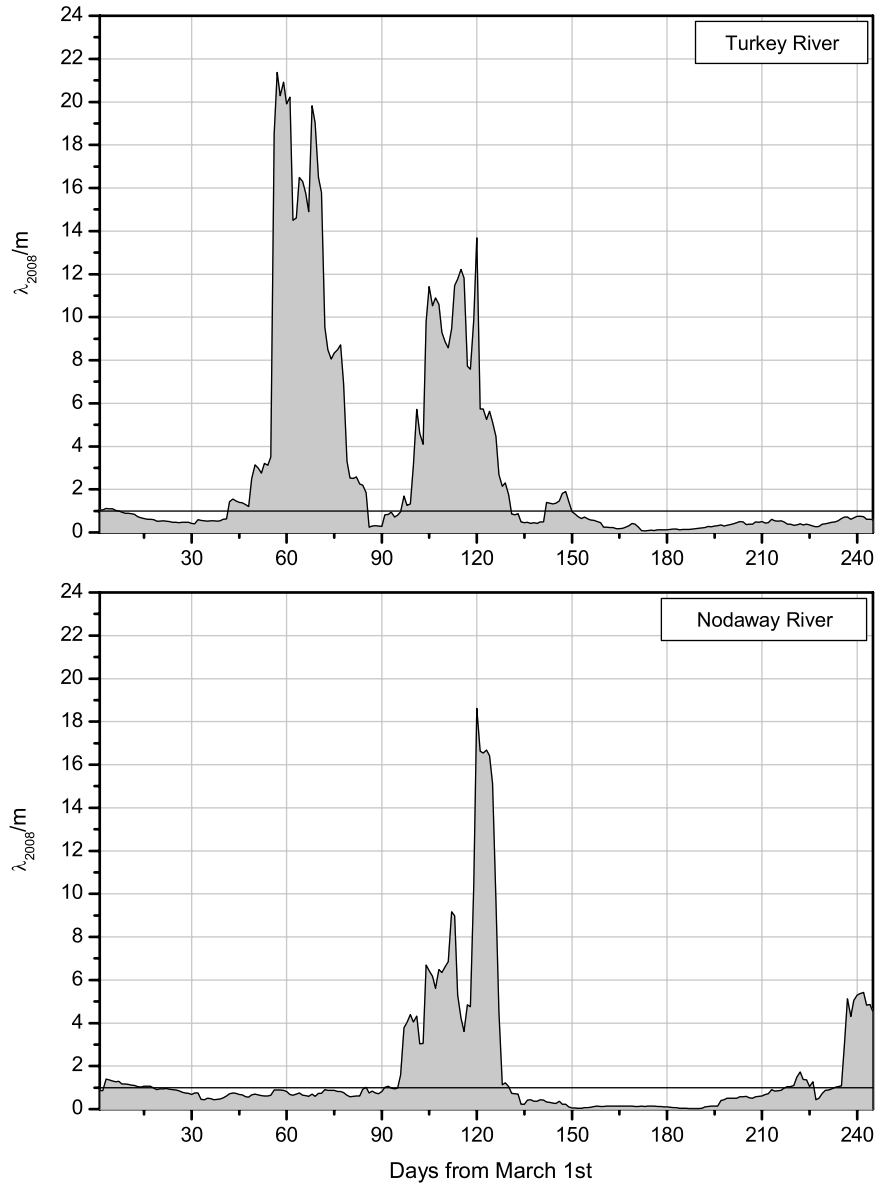

Figure 8. Plot of the ratio of $\lambda_{2008}(t)$ and $\hat{m}(t)$ for the Turkey River and the Nodaway River. A ratio value of 1 ($\lambda_{2008}(t)$ equal to $\hat{m}(t)$) is indicated by the black horizontal line.

that this ratio is independent of the span selected for the smoothing of the rate of occurrence $\hat{m}(t)$ and the individual $\lambda_i(t)$.

5. Discussion and conclusions

In this study, we used discharge observations from 41 USGS stream gauge stations over Iowa covering the period 1950–2009 to develop point process data sets of flood occurrence. We used the Cox regression model to examine the influence of climate indices and soil moisture on flood occurrence. The main findings of this study were as follows:

1. Two climate indices (NAO and PNA) were used as possible covariates to describe the rate of occurrence of floods in Iowa. In 27 of 41 stations, NAO and/or PNA were selected as significant predictors. For NAO, the sign of the coefficient was always negative, implying that larger (smaller) values would result in smaller (larger) values of the rate of occurrence. On the other hand, the sign of the PNA coefficient was generally positive. These results suggest that flood occurrence in Iowa is influenced by large-scale climate indices. On the basis of the results of the at-site and state-wide modelling, we suggest using a model in which the rate of occurrence depends on monthly averaged NAO and PNA, if a single set of predictors was to be selected for Iowa.
2. We selected NAO and PNA to represent the influences of both the Atlantic and Pacific Oceans on atmospheric conditions in Iowa as their daily time series were available from 1950. Iowa, and the Midwest United States in general, is influenced by both the Atlantic and Pacific Oceans, complicating the detection of

Copyright © 2012 Royal Meteorological Society

significant climate signals. Future studies should investigate the possibility of incorporating additional climate indices.

3. We examined the impact of antecedent rainfall (as a surrogate for soil moisture) on the occurrence of floods for two catchments (Turkey River and Nodaway River). In both cases, this additional predictor was selected as highly significant and included in the final model. The results for these basins point to the role of the land-surface component in driving the occurrence of flooding in Iowa. We did not model the rate of occurrence as a function of antecedent rainfall for all the catchments in the study area due to the sparse coverage by rain gauges with a long record. Future studies should extend this work to other catchments with long rainfall time series and for other parts of the country trying to assess whether the land-surface or the atmospheric component is the main driver of flood events. Moreover, because these climate indices could be used to describe heavy rainfall over this area, it would be important to examine the degree of collinearity among these predictors and its impact of the modelling results.
4. Dependence of the rate of occurrence process on covariate processes points to clustering as an important element of the flood occurrence process.
5. The Cox regression model can be used to assess the time-varying rate of occurrence of floods. Having the capability to forecast the predictors, this model could be used to forecast whether a certain period will be much more active than the norm, improving the preparedness to flooding. This could provide predictive skill out to several months in advance, which could be useful for decision makers.
6. We have shown that Cox regression represents a very powerful statistical model to describe the occurrence of flood events. However, while its use is widespread in biostatistics, its use by the hydrometeorological community has been very limited (Smith and Karr, 1983, 1986; Futter *et al*., 1991; Maia and Meinke, 2010). The Cox regression approach is also applicable to other extreme events (e.g. storms, heat waves) as well as floods, and would provide valuable information about the timing of extreme events and their modulation by climate. Availability of appropriate statistical software (as the `R` routines used in this study) should result in more widespread application of these techniques.

Acknowledgements

This publication was supported by the Willis Research Network and by a subcontract from Rutgers University, Department of Environmental Sciences, under Agreement Number G11AP20215 from the US Geological Survey. The authors would like to thank Dr Thernau and Dr Lumley for making the `survival` package (Therneau and original R port by Thomas Lumley, 2009) freely available in `R` (R Development Core Team, 2008), and two reviewers for useful comments on a previous version of the article.

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723.

Ambaum MHP, Hoskins BJ, Stephenson DB. 2001. Artic Oscialltion or North Atlantic Oscillation. *Journal of Climate* **14**: 3495–3507.

Andersen PK, Gill RD. 1982. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**(4): 1100–1120.

Andersen PK, Borgan O, Gill RD, Keiding N. 1992. *Statistical Models Based on Counting Processes*. Springer-Verlag: New York.

Barlow M, Nigam S, Berberry EH. 2001. ENSO, Pacific decadal variability, and U.S. summertime precipitation, drought, and stream flow. *Journal of Climate* **14**: 2105–2128.

Barnston AG, Livezey RE. 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review* **115**(1–2): 1083–1127.

Bates GT, Hoerling MP, Kumar A. 2001. Central U.S. springtime precipitation extremes: teleconnections and relationships with sea surface temperature. *Journal of Climate* **14**: 3751–3766.

Cervantes JE, Kavvas ML, Delleur JW. 1983. A cluster model for flood analysis. *Water Resources Research* **19**(1): 209–224.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**: 829–836.

Coleman JSM, Budikova D. 2010. Atmospheric aspects of the 2008 Midwest floods: a repeat of 1993? *International Journal of Climatology* **30**: 1645–1667.

Coles S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer: London.

Cox DR. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society – Series B* **34**: 187–220.

Cox DR. 1975. Partial likelihood. *Biometrika* **62**(2): 269–276.

Cox DR, Isham V. 1980. *Point Processes, Monographs on Statistics and Applied Probability*, vol. 12. Chapman & Hall: New York.

Davison AC, Smith RL. 1990. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B* **52**: 393–442.

Futter MR, Mawdsley JA, Metcalfe AV. 1991. Short-term flood risk prediction: a comparison of the Cox regression model and a conditional distribution model. *Water Resources Research* **27**(7): 1649–1656.

Gill RD. 1984. Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association* **79**(386): 441–447.

Grandell J. 1976. *Doubly Stochastic Poisson Processes*. Springer: Berlin.

Gupta VK, Waymire E. 1979. A stochastic kinematic study of subsynoptic space-time rainfall. *Water Resources Research* **15**(3): 637–644.

Hurrell JW. 1995. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**(5224): 676–679.

Hurrell JW, Van Loon H. 1997. Decadal variations in climate associated with the North Atlantic Oscillation. *Climatic Change* **36**(3–4): 301–326.

Istok JD, Boersma L. 1989. A stochastic cluster model for hourly precipitation data. *Journal of Hydrology* **106**: 257–285.

Karr AF. 1991. *Point Processes and Their Statistical Inference*. Dekker: New York.

Kavvas ML. 1987. Some new perspectives on the probabilistic modeling of floods. *Journal of Hydrology* **92**: 315–331.

Kavvas ML, Delleur JW. 1975. The stochastic and chronologic structure of rainfall sequences – Application to Indiana. *Technical Report 57*, Indiana Water Resources Research Center.

Kavvas ML, Delleur JW. 1981. A stochastic cluster model of daily rainfall sequences. *Water Resources Research* **17**(4): 1151–1160.

Kingman JFC. 1964. On doubly stochastic Poisson processes. *Mathematical Proceedings of the Cambridge Philosophical Society* **60**: 923–930.

Kunkel KE, Changnon SA, Angel JR. 1994. Climatic aspects of the 1993 Upper Mississippi River basin flood. *Bulletin of the American Meteorological Society* **75**: 811–822.

Copyright © 2012 Royal Meteorological Society

Lang M, Ouarda TBMJ, Bobeé B. 1999. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology* **225**: 103–117.

Le Cam LM. 1961. A stochasitc description of precipitation. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **3**: 517–528.

Leathers DJ, Yarnal B, Palecki MA. 1991. The Pacific/North American teleconnection pattern and United States climate. Part I: regional temperature and precipitation associations. *Journal of Climate* **4**: 517–528.

Maia AHN, Meinke H. 2010. Probabilistic methods for seasonal forecasting in a changing climate: Cox-type regression models. *International Journal of Climatology* **30**: 2277–2288.

Mailier PJ, Stephenson DB, Ferro CAT. 2006. Serial clustering of extratropical cyclones. *Monthly Weather Review* **134**(8): 2224–2240.

Onof C, Chandler RE, Kakou A, Northrop P, Wheater HS, Isham V. 2000. Rainfall modelling using Poisson-cluster processes: a review of developments. *Stochastic Environmental Research and Risk Assessment* **14**: 384–411.

Otto D. 2009. Economic losses from the floods? In *Watershed Year: Anatomy of the Iowa Floods of 2008*, Mutel CF (ed). University of Iowa Press: Iowa City, 139–146.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0.

Ramirez JA, Bras RL. 1985. Conditional distribution of Neyman-Scott models for storm arrivals and their use in irrigation scheduling. *Water Resources Research* **21**(3): 317–330.

Rodriguez-Iturbe I, Cox DR, Isham V. 1987. Some models for rainfall based on stochastic point process. *Proceedings of the Royal Society of London. Series A* **A410**: 269–288.

Smith JA, Karr AF. 1983. A point process model of summer season rainfall occurrences. *Water Resources Research* **19**(1): 95–103.

Smith JA, Karr AF. 1985. Statistical inference for point process models of rainfall. *Water Resources Research* **21**(1): 73–79.

Smith JA, Karr AF. 1986. Flood frequency analysis using the Cox regression model. *Water Resources Research* **22**(6): 890–896.

Therneau T and original R port by Thomas Lumley. 2009. *survival: Survival analysis, including penalised likelihood.* R package version 2.35–8.

Therneau TM, Grambsch PM. 2000. *Modeling Survival Data: Extending the Cox Model*. Springer: New York.

Ting M, Wang H. 1997. Summertime U.S. precipitation variability and its relation to Pacific sea surface temperature. *Journal of Climate* **10**: 1853–1873.

Villarini G, Vecchi GA, Smith JA. 2010. Modeling of the dependence of tropical storm counts in the North Atlantic Basin on climate indices. *Monthly Weather Review* **138**(7): 2681–2705.

Villarini G, Smith JA, Baeck ML, Krajewski WF. 2011a.. Examining flood frequency distributions in the Midwest U.S. *Journal of the American Water Resources Association* **43**(3): 447–463.

Villarini G, Smith JA, Baeck ML, Vitolo R, Stephenson DB, Krajewski W. 2011b.. On the frequency of heavy rainfall for the Midwest of the United States. *Journal of Hydrology* **400**(1–2): 103–120.

Vitolo R, Stephenson DB, Cook IM, Mitchell-Wallace K. 2009. Serial clustering of intense European storms. *Meteorologische Zeitschrift* **18**(4): 411–424.

Wallace JM, Gutzler DS. 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review* **109**: 784–812.

Waymire E, Gupta VK. 1981. The mathematical structure of rainfall representation, 1. A review of the stochastic rainfall models. *Water Resources Research* **17**(5): 1261–1272.

Waymire E, Gupta VK, Rodriguez-Iturbe I. 1984. Spectral theory of rainfall intensity at the meso-β scale. *Water Resources Research* **20**: 1453–1465.

Copyright © 2012 Royal Meteorological Society